A multivariate heterogeneous variance components model for multi-environment studies with locational genetic effects

Özge Karadag

Accepted: 9 August 2021 / Published online: 23 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

In this paper, a multivariate heterogeneous variance components model was developed which allows for determination of location specific variance components in the analysis of multiple related traits. In addition to spatial heterogeneity, genetic similarities are also considered by assigning genetic variance components. The performance of the developed model was evaluated through an extensive simulation study and comparison of models was conducted by heritability estimations. The simulation study reveals that the developed method can control the locational heterogeneity well and the heritability estimations are close to desired proportions for the developed model. A real plant breeding data set was used for illustration.

Keywords Multivariate variance components · Mixed modelling · Genetic variance components · Location effect · Heritability

1 Introduction

Previously, several methods were developed to estimate genetic variance components using random effect models. These models allow estimation of genetic effects to be related more to familial relationships such as pedigree structures rather than environmental relatedness. A motivation for our study is that the heterogeneity arising from the locational clustering may cause underestimation of genetic variance components, especially in animal and plant breeding studies.

In recent decades, there is growing interest in multivariate analysis due to the availability of extensive large-scale data (Li et al. 2021; Sundararajan 2021; Rao et al. 2020). Especially in the field of interdisciplinary sciences such as statistical genetics, the multivariate approach allows for analysis of multiple traits affected by the interplay of both genes and environmental influences. Moreover, these multiple traits are usually related to each other in several ways; hence, the correlation between multiple variants should also be taken into consideration with a multivariate approach. The linear mixed model (LMM) was widely used to analyse the relationship between a trait and genetic random factors, and the multivariate linear mixed model (mvLMM) is an extension of LMM for modelling multiple traits simultaneously.

In the LMM approach, certain sources of variation are assessed by random effect variables and the variance–covariance parameters of these variables are referred to as variance components. Since they are useful for explaining the total variation, estimation of variance components has become an essential part of the modelling process in most applied science fields. Particularly in animal and plant breeding studies, the familial and environmental relationships necessitate the usage of LMM due to the naturally clustered structure of the population (Correddu et al. 2021).

Moreover, this clustered structure may often cause genetic diversity among the members of species owing to natural selection or geographic conditions, and the genetic variance components ensure the understanding of the genetic ethology of the characteristics of interest.

Although variance components models have a long history, advanced computational strategies are still being developed to estimate variance parameters accurately (Henderson 1953; Smith and Graser 1986; Searle et al. ...)
where \( Y \) Consider a study with multiple traits. Multivariate variance components model (1992; Lynch and Walsh 1998). Especially, due to the large dimensions of matrices in mvLMM, most of the extensions are presented to overcome computational difficulties. For instance, Gilmour et al. (1995) developed a computationally convenient and extensive algorithm based on average information matrix for the estimation of variance parameters, and Lee and Van der Werf (2006) discussed the efficiency of direct use of the variance covariance matrix with a general complex pedigree. Recently, an extension of mvLMM based on genomic information was developed by combining the direct average of the information algorithm with an eigen decomposition of the genomic relationship matrix (Lee and Van derWerf 2006).

In addition to computational solutions, multivariate variance components models need to be extended to account for several different sources of relatedness and heterogeneity. Especially in genetic analysis, most of the complex traits are affected by a collaboration of genetic and environmental factors, and this collaboration may require additional variance components. For instance, due to the multivariate heterogeneous variance components model is considered by assigning genetic variance components. Due to multiple-related traits across multiple locations. In addition, designing, specifying separate genetic variance components for each environment is a way to model the heterogeneity arising from the interaction of genetic and environmental factors.

In this paper, the focus is the estimation of heterogeneous variance components of mvLMM for the analysis of multiple-related traits across multiple locations. In addition to spatial heterogeneity, genetic similarities are also considered by assigning genetic variance components. Due to the genetic background and location clustered structure of the desired design, a complex multiple phenotype simulation is conducted that relies on genotype simulations. A multivariate heterogeneous variance components model is proposed by taking location-specific variance components into consideration.

### 2 Multivariate variance components model

Consider a study with multiple traits \( Y_{i1}, Y_{i2}, \ldots, Y_{im}, Y_{m1}, Y_{m2}, \ldots, Y_{mn} \). Let \( Y = vec[Y_1, Y_2, \ldots, Y_m] \) be the response vector of \( m \) traits, where \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im}) \) and \( Y_i \) denote the response for trait \( i \) of individual \( j, i = 1, 2, \ldots, m \) and \( j = 1, 2, \ldots, n \). Then, the matrix notation of mvLMM is

\[
Y = X\beta + ZU + \epsilon,
\]

where \( X \in \mathbb{R}^{nm \times Pm} \) is the design matrix for \( P \) fixed effects and \( Z \in \mathbb{R}^{nm \times Qm} \) is the design matrix for \( Q \) random effects.

\[
X = \begin{bmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_m \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_m \end{bmatrix}
\]

(2)

The diagonal elements of \( X \) and \( Z \) are identical in themselves, \( X_1 = \cdots = X_m = X \in \mathbb{R}^{n \times P} \) and \( Z_1 = \cdots = Z_m = Z \in \mathbb{R}^{n \times Q} \) and off-diagonal elements of \( X \) and \( Z \) are null matrices.

\[
\beta = vec[\beta_1, \beta_2, \ldots, \beta_m] \text{ is a vector of fixed effects coefficients where } \beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iP})'.\quad \epsilon = vec[\epsilon_1, \epsilon_2, \ldots, \epsilon_m]' \text{ denotes the random error vector and }\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{im})' \text{ is the vector of individual residual effects with } \epsilon_i \sim N(0, R_i) \text{ where } R_i = I_n \sigma^2_{e_i} \text{ represents the residual variance–covariance matrix of trait } i \text{ and the term }\sigma^2_{e_i} \text{ is the residual variance of trait } i.
\]

The random effects are \( U = vec[u_1, u_2, \ldots, u_m] \) and \( u_i = (u_{i1}, u_{i2}, \ldots, u_{iQ})' \). \( u_i \) is assumed to be normally distributed with zero mean and the variance–covariance matrix of \( u_i \) is \( D_i \in \mathbb{R}^{Q \times Q} \).

For multiple traits, the multivariate distribution can be assumed to be

\( Y \sim MVN(X\beta, V) \)

(3)

where \( V \) is the variance–covariance matrix of all observations.

\[
V = \begin{bmatrix} Z_iDZ_i & \cdots & Z_iD_{im}Z_i & R_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ Z_{im}DZ_i & \cdots & Z_{im}D_{im}Z_i & R_{im} \end{bmatrix}
\]

(4)

Due to the multivariate structure, the random effects covariance and the residual covariance between traits \( i \) and \( t \) are denoted by \( D_{it} \) and \( R_{it} = I_n \sigma^2_{e_i} \), respectively \((i = 1, 2, \ldots, m; t = 1, 2, \ldots, m \text{ and } i \neq t) \). The term \( \sigma^2_{e_i} \) is the residual covariance between traits \( i \) and \( t \).

In the mixed modelling approach, the random part of the model may have several components reflecting different grouping effects such as treatment effect, common environmental effect or serial effect for repeated measurements. Especially in genetic studies, a random term is usually defined to consider the genetic similarities. In this case, including genetic background effects, the components of the random vector for trait \( i \) can be split into two parts; genetic and non-genetic random effects. For simplicity, considering a single component for the random genetic effects on all individuals as a vector of total genetic value, \( g_i = (g_{i1}, g_{i2}, \ldots, g_{im})' \), then \( u_i = (g_{i1}, u_{i1}, u_{i2}, \ldots, u_{i(Q-1)})' \) and the variance–covariance matrix for trait \( i \) can be written as

\[
\begin{bmatrix} Z_iDZ_i & \cdots & Z_iD_{im}Z_i & R_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ Z_{im}DZ_i & \cdots & Z_{im}D_{im}Z_i & R_{im} \end{bmatrix}
\]

\[
\begin{bmatrix} D_{i1} & \cdots & D_{i(m-1)} \cdots D_{im} & R_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ D_{m1} & \cdots & D_{m(m-1)} \cdots D_{im} & R_{1m} \end{bmatrix}
\]

\[
\begin{bmatrix} D_{i1} & \cdots & D_{i(m-1)} \cdots D_{im} & R_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ D_{m1} & \cdots & D_{m(m-1)} \cdots D_{im} & R_{1m} \end{bmatrix}
\]
where $G_l$ is the genetic relatedness matrix, and $E_{i1}, \ldots, E_{iS}$ are the variance–covariance matrices of $S = (Q - 1)$ random effect components other than genetic background. Here, $G_l = K\sigma^2_{g_l}$ and $\sigma^2_{g_l}$ denote the genetic variance component related to trait $i$. $K \in \mathbb{R}^{n \times n}$ is the genomic relationship or kinship matrix. Similarly, the genetic component of $D_{il}$ is $G_{il} = K\sigma^2_{g_{il}}$ and the term $\sigma^2_{g_{il}}$ is the genetic covariance between traits $i$ and $l$.

In single-environment studies, $\sigma^2_{g_l}$ and $\sigma^2_{g_{il}}$ are assumed to also be single. However, in genetic studies, heterogeneity can occur based on an environmental indicator, such as location. In this case, the distribution of a genetic component is expressed by location-specific variances.

### 2.1 Extension to the multivariate variance components model

In this study, the multivariate variance components model is typically expanded to include location-specific variance components by specifying a heterogeneous variance structure for the trait vector, at specific levels of a random factor.

Compared to the standard mvLMM, the proposed method has the important advantage of accounting for several different sources of relatedness and heterogeneity which occur in genetic analysis. Moreover, the developed model allows modelling of spatial trends to provide more flexibility.

Here, a multi-environment design with $L$ different locations is considered. Regarding the locational heterogeneity, the genetic component of the mvLMM is assumed to have location-specific variance $\sigma^2_{g_{il}}$ denoting the genetic variance component for environment $l$ ($l = 1, 2, \ldots, L$). Then, the genetic variance–covariance matrix across $L$ locations for trait $i$ can be written as

$$G_l = K \begin{bmatrix} \sigma^2_{g_{i1}} & \cdots & \sigma^2_{g_{iL}} \\ \vdots & \ddots & \vdots \\ \sigma^2_{g_{L1}} & \cdots & \sigma^2_{g_{LL}} \end{bmatrix}$$

(6)

where $\sigma^2_{g_{il}}$ is the covariance term reflecting genetic effects in two different environments $l$ and $b$ ($l = 1, 2, \ldots, L; b = 1, 2, \ldots, L$ and $l \neq b$). For an mvLMM accounting for heterogeneous variances, the variance components vector can be decomposed as

$$\Theta = \left( \sigma^2_{g_{i1}}, \ldots, \sigma^2_{g_{il}}, \ldots, \sigma^2_{g_{iL}}, \ldots, \sigma^2_{g_{il}}, \ldots, \sigma^2_{g_{i2}}, \ldots, \sigma^2_{g_{i1}} \right)^T$$

(7)

To illustrate the location-specific variance components, a multiple trait design in four locations is considered ($L = 4$). Regarding the locational heterogeneity, the genetic component of the mvLMM is assumed to have location-specific variance $\sigma^2_{g_{il}}$ denoting the genetic variance component for environment $l$ ($l = 1, 2, \ldots, L$).

The log likelihood function of the mvLMM is

$$L = -\frac{1}{2} \left[ \ln|V| + \ln|X'V^{-1}X| + Y'\Sigma Y \right]$$

(8)

where $\Sigma = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$.

### 2.2 Estimation of heterogeneous variance components

In this study, a multivariate Newton–Raphson iterative algorithm was used to obtain residual/restricted maximum likelihood estimates (REML) for variance components. REML is often solved by updating variance components based on Hessian matrix or Fisher’s information matrix consisting of second derivatives of the log likelihood function (Searle et al 1992; Lynch and Walsh 1998). For a more efficient computation, the REML method is implemented via the average of the observed Hessian and the expected Fisher information matrices (Gilmour et al 1995; Lee and Van der Werf 2006).

In the average information Newton–Raphson (AI-NR) algorithm, the REML estimates are obtained with

$$\Theta^{k+1} = \Theta^{(k)} - \left( \frac{\partial L}{\partial \Theta} \right)^{-1} \frac{\partial^2 L}{\partial \Theta^2} \Theta^{(k)}$$

(9)

where $\Theta$ is the vector of variance components decomposed to Eq. (7) and $\Lambda$ is the average information matrix consisting of second derivatives of the log likelihood function $L$.

For mvLMM, the $\Lambda$ matrix is directly derived from $V$ as

$$\Lambda = \frac{1}{2} \left[ \Sigma \frac{\partial^2 V}{\partial \sigma^2} \Sigma \frac{\partial^2 \Sigma}{\partial \sigma^2} \right]$$

(10)

In a multiple trait design, considering observations collected in $L$ environments, the variance covariance matrix $V$ can be rewritten as

$$V = \begin{bmatrix} Z_1D_1Z_1' + I\sigma^2_{i1} & \cdots & Z_1D_1mZ_1' + R_1m \\ \vdots & \ddots & \vdots \\ Z_lD_lmZ_l' + R_1m & \cdots & Z_mD_mZ_m' + I\sigma^2_{m1} \end{bmatrix}$$

(11)
3 Data simulation

In the simulation scenario design, the number of multiple traits was assigned as three and number of samples was assigned as \( n = 1000 \). Multiple traits are simulated as the sum of genetic background effect, locational effect and noise effects as described in Meyer and Birney (2018), and covariate effects other than the genetic and locational effects are considered to be nuisances. Genetic background effects were simulated based on kinship estimates. For the kinship estimates, 500 bi-allelic single nucleotide polymorphisms (SNP) are simulated for 1000 samples and the level of genotype is generated relying on probability in a binomial distribution with uniformly sampled reference allele frequencies (0.05, 0.1, 0.3, 0.4 and 0.5).

In order to reveal the effects of location-specific variance components, a location indicator was simulated as a categorical variable and influenced multiple traits with different proportions; \( w = 0, 0.3, 0.5, 0.8 \) and 1. Due to the multivariate structure of the design, three different levels of correlation \( q = 0.25, 0.5 \) and 0.75 between traits were considered and the correlated effect was simulated from multivariate normal distribution (MVN) with the considered level of correlation. The proportion of genetic variance \( \sigma^2_g \) is set to 0.30. The correlation structures for some remarkable scenarios were visualized by heat maps, as given in Fig. 1.

In the modelling process of the simulation study, the heritability estimations were obtained under mvLMM and heterogeneous mvLMM with location-specific variance components. Estimation of heritability, \( h^2 \), relies on the partitioning of observed variation into unobserved genetic and environmental components (Wray and Visscher 2008). For the mvLMM with a general variance component, the heritability is estimated as the proportion of genetic variance to total variance, \( h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \). However, in a multi-environment design with \( L \) different locations, heritability depends on the proportion of total genetic variance in \( L \) locations and can be estimated as

\[
h^2 = \frac{\sum_{i=1}^n \sigma^2_{gi}}{\sum_{i=1}^n \sigma^2_{gi} + \sigma^2_e} \tag{12}
\]

The results in Table 1 indicate the proximity of the heritability estimation to the true proportion (30%) of genetic variance.

In Table 1, \( w = 0 \) indicates the absence of locational heterogeneity and heritability estimations are similar and about 0.30. When using mvLMM with location-specific variance components, as the proportion associated with location increases, the heritability estimations are closer to 0.30. Our results also show that the dependence level of multiple traits do not affect estimations remarkably.

4 Application to real data

In this section, the two variance components models were fitted to safflower (\textit{Carthamus tinctorius} L.) data collected from three different locations (Bolu, Haymana and...
Yenimahalle) in the central Anatolian region, Turkey to examine the performance of the developed model. The oil yield and length were considered as dependent variables. The correlation between yield and length in three different environments is illustrated in Fig. 2.

The REML estimations were obtained from mvLMM and heterogeneous mvLMM with location-specific variance components by using the AI-NR algorithm. The model results are summarized in Table 2.

In standard mvLMM, six components \( \sigma^2_{\text{Field}} \), \( \sigma^2_{\text{Length}} \), \( \sigma^2_{\text{Field} \times \text{Length}} \), \( \sigma^2_{\text{Field} \times \text{Loc}} \), \( \sigma^2_{\text{Length} \times \text{Loc}} \), \( \sigma^2_{\text{Field} \times \text{Length} \times \text{Loc}} \) were estimated, whereas the number of components was doubled for heterogeneous mvLMM by including separate components for each environment. For instance, the genetic variance component of yield was represented by \( \sigma^2_{\text{Field} \times \text{Loc}} \) in standard mvLMM; however, in heterogeneous mvLMM, the genetic variance component of yield has its own sub-components of \( \sigma^2_{\text{Field} \times \text{Loc}}^{(\text{Bolu})} \), \( \sigma^2_{\text{Field} \times \text{Loc}}^{(\text{Haymana})} \), and \( \sigma^2_{\text{Field} \times \text{Loc}}^{(\text{Yenimahalle})} \) for Bolu, Haymana and Yenimahalle, respectively.

5 Discussion

In this paper, a multivariate heterogeneous variance components model with location-specific variance components was developed, which allowed for determination of location-specific variance components. The simulation results show that the heritability estimations are closer to the desired proportions using the developed mvLMM with location-specific genetic variance components as the locational heterogeneity increases. Based on real data results, the developed heterogeneous mvLMM with location-specific variance components fits the data better in multi-environmental designs. Thus, our method can control for locational heterogeneity better compared to an mvLMM with a general variance component.

6 Conclusion

mvLMM is generally suitable for modelling multiple traits simultaneously, and multivariate variance components are useful for explaining the total variation. However, especially in complex experimental designs where heterogeneity can occur due to several indicators, standard variance components models have some limitations such as assuming single variance components. As an extension of the multivariate variance components model, the proposed method allows for location-specific variance components by specifying a heterogeneous variance structure for the

**Table 1** Heritability estimations

| Proportion associated with location | Level of correlation | \( \rho = 0.25 \) | \( \rho = 0.5 \) | \( \rho = 0.75 \) |
|------------------------------------|---------------------|-----------------|-----------------|-----------------|
| mvLMM with a general variance component | \( w = 0 \) | 0.2741 | 0.2966 | 0.2834 |
|                                    | \( w = 0.2 \)   | 0.2707 | 0.2831 | 0.2416 |
|                                    | \( w = 0.4 \)   | 0.2561 | 0.2721 | 0.2311 |
|                                    | \( w = 0.6 \)   | 0.2489 | 0.2504 | 0.2341 |
|                                    | \( w = 0.8 \)   | 0.2313 | 0.2361 | 0.2229 |
|                                    | \( w = 1 \)     | 0.2239 | 0.2375 | 0.2266 |
| mvLMM with location-specific variance components | \( w = 0 \)     | 0.2895 | 0.2896 | 0.2756 |
|                                    | \( w = 0.2 \)   | 0.2891 | 0.2889 | 0.3016 |
|                                    | \( w = 0.4 \)   | 0.2995 | 0.2797 | 0.2967 |
|                                    | \( w = 0.6 \)   | 0.2930 | 0.2879 | 0.2954 |
|                                    | \( w = 0.8 \)   | 0.2964 | 0.2849 | 0.3028 |
|                                    | \( w = 1 \)     | 0.3013 | 0.2966 | 0.3016 |

Fig. 2 Correlation in three different environments
trait vector. In addition to spatial heterogeneity, genetic similarities are also considered by assigning genetic variance components.

For future studies, the developed model can be extended to allow for modelling of longitudinal responses in later implementations. Moreover, future simulation studies based on the multivariate heterogeneous variance components model may offer insights about how to identify and interpret the parameters in more complex models. A more comparative study can be developed by including the addition of other multivariate modelling methods such as structural equation algorithms.

Acknowledgements The author is especially grateful to Prof. Dilek Başalma (Ankara University, Faculty of Agriculture, Turkey) for her assistance with data validation.

Authors' contributions ÖK idealized, supervised the research, analysed the data and approved the final draft.

Funding The author received no specific funding for this work.

Declarations

Conflict of interest No conflict of interest to declare.

Data Availability Due to its proprietary nature, supporting data cannot be made openly available.

References

Correddu F, Cesarani A, Dimauro C, Gaspa G, Macciotta NPP (2021) Principal component and multivariate factor analysis of detailed sheep milk fatty acid profile. J Dairy Sci 104(4):5079–5094
Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R Package sommer. PLoS ONE 11(6):1–15
Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51(4):1440–1450
Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226–252
Lee SH, van der Werf JH (2006) An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet Sel Evol 38:25–43
Lee SH, van der Werf JH (2016) MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32(9):1420–1422
Li P, Rao C, Goh M, Yang Z (2021) Pricing strategies and profit coordination under a double echelon green supply chain. J Clean Prod 278:123694
Lynch M, Walsh B (1998) Variance component estimation. In: Lynch M, Walsh B (eds) Genetics and analysis of quantitative traits. Sinauer Associates Inc, Sunderland, MA, p 980
Meyer VH, Birney E (2018) PhenotypeSimulator: a comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics 34(17):2951–2956
Rao C, Lin H, Liu M (2020) Design of comprehensive evaluation index system for P2P credit risk of “three rural” borrowers. Soft Comput 24:11493–11509
Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York (NY)
Smith SP, Graser H-U (1986) Estimating variance components in a class of mixed models by restricted maximum likelihood. J Dairy Sci 69:1156–1165
Sundararajan RR (2021) Principal component analysis using frequency components of multivariate time series. Comput Stat Data Anal 157:107164
Wray N, Visscher P (2008) Estimating trait heritability. Nat Edu 1(1):29

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.