A Multivariate Heterogeneous Variance Components Model for Multi-Environment Studies with Locational Genetic Effects

Ozge Karadag Atas (✉ ozgekaradag@hacettepe.edu.tr)
Hacettepe University: Hacettepe Universitesi https://orcid.org/0000-0002-2650-1458

Research Article

Keywords: multivariate variance components, mixed modelling, genetic variance components, location effect, heritability

Posted Date: May 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-241521/v1

A Multivariate Heterogeneous Variance Components Model for Multi-Environment Studies with Locational Genetic Effects

Özge Karadağ

Department of Statistics, Hacettepe University, 06800, Ankara, Turkey

ozgekaradag@hacettepe.edu.tr

Abstract

In this paper a multivariate heterogeneous variance components model is developed, which allows for determining 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 is evaluated through an extensive simulation study and comparison of models are conducted by heritability estimations. Simulation study reveals that the developed method can well control the locational heterogeneity and under the developed model the heritability estimations are close to desired proportions. A real plant breeding data set is used for illustration.

Keywords: multivariate variance components, mixed modelling, genetic variance components, location effect, heritability

Introduction

In last decades, there has been a growing interest in multivariate analysis due to the available extensive large-scaled data. Especially, in the field of interdisciplinary sciences such as statistical genetics, 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 also should be taken into consideration by a multivariate approach. Linear mixed model (LMM) has been widely used to analyse the relationship between a trait and the genetic random factors, and multivariate linear mixed model (mvLMM) is an extension of LMM for modelling multiple traits simultaneously.

In LMM approach, the certain sources of variation are assessed by random effect variables and the variance-covariance parameters of such variables, referred to as variance components. Since, they are useful for explaining the total variation, estimation of variance components have become an essential part of the modelling process in most of the 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.

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 serve for understanding the genetic ethology of the characteristics of interest.

Although variance components models have a long history, advanced computational strategies have been still developed for estimating variance parameters accurately (Henderson 1953; Smith and Graser 1986; Searle et al. 1992; Lynch and Walsh 1998). Especially, in mvLMM, due to the large dimensions of matrices, most of the extensions are presented as to overcome computational difficulties. For instance, Gilmour et al. (1995) develop a computationally convenient and extensive algorithm based on average information matrix for the estimation of variance parameters and Lee and Van der Werf (2006) discuss the efficiency of direct use of the variance covariance matrix with a general complex pedigree. Recently, an extension to the mvLMM based on genomic information is developed by combing the direct average information algorithm with an eigen-decomposition of genomic relationship matrix (Lee and Van der Werf 2006).

In addition the computational solutions, multivariate variance components models are need to be extended as to accounting 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 cause necessity of additional variance components. For instance, in multiple environment population studies, due to the interplay of genes and environment, it is misleading to assume that the genetic background is common across the different locations (Covarrubias-Pazaran 2016). In such study designs, 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, it is focused on 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 rely on genotype simulations. A multivariate heterogeneous variance components model is proposed taking the location specific variance components into consideration.

Multivariate Variance Components Model

Consider a study with multiple traits $Y_{11}, Y_{12}, \ldots, Y_{1n}, \ldots, Y_{m1}, Y_{m2}, \ldots, Y_{mn}$. Let $\boldsymbol{Y} = vec[\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m]$ be the response vector of m traits, where $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ and Y_{ij} denotes 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

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{ZU} + \boldsymbol{\epsilon}, \tag{1}$$

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

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{X}_m \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & Z_m \end{bmatrix} \tag{2}$$

The diagonal elements of $\boldsymbol{X}$ and $\mathbf{Z}$ are identical in themselves, $\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_m = \boldsymbol{X}_* \in \mathbb{R}^{n\times P}$ and $\boldsymbol{Z}_1 = \cdots = \boldsymbol{Z}_m = \boldsymbol{Z}_* \in \mathbb{R}^{n\times Q}$ and off-diagonal elements of $\boldsymbol{X}$ and $\mathbf{Z}$ are null matrices.

$\boldsymbol{\beta} = vec[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_m]$ is a vector of fixed effects coefficients where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{iP})'$. $\boldsymbol{\epsilon} = vec[\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_m]$ denotes the random error vector and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{in})'$ is the vector of individual residual effects with $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i)$ where $\boldsymbol{R}_i = \boldsymbol{I}_n \sigma^2_{\varepsilon_i}$ representing the residual variance-covariance matrix of trait i and the term $\sigma^2_{\varepsilon_i}$ is the residual variance of trait i.

The random effects $\boldsymbol{U} = vec[\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m]$ and $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \dots, \boldsymbol{u}_{iQ})'$. $\boldsymbol{u}_i$ is assumed to be normally distributed with zero mean and the variance-covariance matrix of $\boldsymbol{u}_i$ is $\boldsymbol{D}_i \in \mathbb{R}^{Q\times Q}$**.**

For multiple traits, the multivariate distribution can be assumed as

$$\boldsymbol{Y} \sim MVN(\boldsymbol{X\beta}, \boldsymbol{V}) \tag{3}$$

where $\boldsymbol{V}$ is the variance-covariance matrix of all observations**.**

$$\boldsymbol{V} = \begin{bmatrix} \mathbf{Z}_*\mathbf{D}_1\mathbf{Z}_*' + \mathbf{R}_1 & \cdots & \mathbf{Z}_*\mathbf{D}_{1m}\mathbf{Z}_*' + \mathbf{R}_{1m} \\ \vdots & \ddots & \vdots \\ \mathbf{Z}_*\mathbf{D}_{1m}\mathbf{Z}_*' + \mathbf{R}_{1m} & \cdots & \mathbf{Z}_*\mathbf{D}_m\mathbf{Z}_*' + \mathbf{R}_m \end{bmatrix} \tag{4}$$

Due to multivariate structure, the random effects covariance and the residual covariance between traits i and t are denoted by $\boldsymbol{D}_{it}$ and $\boldsymbol{R}_{it} = \boldsymbol{I}_n\sigma^2_{\varepsilon_{it}}$, respectively ($i = 1,2,\dots,m$; $t = 1,2,\dots,m$ and $i \neq t$). The term $\sigma^2_{\varepsilon_{it}}$ is the residual covariance between traits i and t.

In 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 for considering the genetic similarities. In this case including genetic background effects, the components of the random vector for trait i can split into two parts, genetic and non-genetic random effects. For simplicity, considering a single component for the random genetic effects over all individuals as a vector of total genetic value, $\boldsymbol{g}_i = (g_{i1}, g_{i2}, \dots, g_{in})'$, then $\boldsymbol{u}_i = \left(\boldsymbol{g}_i, \boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \dots, \boldsymbol{u}_{i(Q-1)}\right)'$ and the variance-covariance matrix for trait i can be written as

$$\boldsymbol{D}_{\boldsymbol{i}} = \begin{bmatrix} \boldsymbol{G}_{\boldsymbol{i}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_{i1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{E}_{iS} \end{bmatrix} \tag{5}$$

where $\boldsymbol{G}_{\boldsymbol{i}}$ is the genetic relatedness matrix and $\boldsymbol{E}_{i1}$, ..., $\boldsymbol{E}_{iS}$ are the variance-covariance matrices of $S = (Q-1)$ random effect components other than genetic background. Here, $\boldsymbol{G}_{\boldsymbol{i}} = \boldsymbol{K}\sigma^2_{g_i}$ and $\sigma^2_{g_i}$ denotes the genetic variance component related to trait i. $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the genomic relationship or kinship matrix. Similarly, the genetic component of the $\boldsymbol{D}_{it}$ is $\boldsymbol{G}_{\boldsymbol{it}} = \boldsymbol{K}\sigma^2_{g_{it}}$ and the term $\sigma^2_{g_{it}}$ is the genetic covariance between traits i and t.

In single-environment studies, $\sigma^2_{g_i}$ and $\sigma^2_{g_{it}}$ are assumed as also single. However, in genetic studies, heterogeneity can be occurred by an environmental indicator, such as locations. In this case, the distribution of genetic component is expressed by location specific variances.

Here, a multi-environment design with L different locations is considered. For regarding the locational heterogeneity, the genetic component of the mvLMM is assumed as 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

$$\boldsymbol{G_i} = \boldsymbol{K}\begin{bmatrix} \sigma^2_{g_{i1}} & \cdots & \sigma^2_{g_{i1,L}} \\ \vdots & \ddots & \vdots \\ \sigma^2_{g_{iL,1}} & \cdots & \sigma^2_{g_{iL}} \end{bmatrix} \quad (6)$$

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

$$\boldsymbol{\Theta} = \left(\sigma^2_{g_{11}},\ldots,\sigma^2_{g_{1L}},\ldots,\sigma^2_{g_{m1}},\ldots,\sigma^2_{g_{mL}},\ldots,\sigma^2_{\varepsilon_1},\ldots,\sigma^2_{\varepsilon_m}\right)' \quad (7)$$

To illustrate the location-specific variance components, a multiple trait design over 4 locations is considered $(L = 4)$. For regarding the locational heterogeneity, the genetic component of the mvLMM is assumed as 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

$$\mathcal{L} = -\frac{1}{2}\left[\ln|\boldsymbol{V}| + ln\left|\boldsymbol{X'}^{\boldsymbol{V}^{-1}}\boldsymbol{X}\right| + \boldsymbol{Y'\Sigma Y}\right] \quad (8)$$

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

Estimation of Heterogeneous Variance Components

In this study, a multivariate Newton-Raphson iterative algorithm is used to obtain residual/restricted maximum likelihood estimates (REML) of 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

$$\boldsymbol{\Theta}^{(\mathrm{k+1})} = \boldsymbol{\Theta}^{(\mathrm{k})} - \left(\mathbf{AI}^{(\mathrm{k})}\right)^{-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}}\bigg|_{\boldsymbol{\Theta}^{\{\mathrm{k}\}}} \tag{9}$$

where Θ is the vector of variance components decomposed as in Equation (7) and AI is the average information matrix consisting of the second derivatives of the log likelihood function L.

For the mvLMM, AI matrix is directly derived from V as

$$\mathbf{AI} = \frac{\mathbf{1}}{\mathbf{2}}\left[\boldsymbol{Y}' \frac{\partial \boldsymbol{V}}{\partial \sigma_i^2} \boldsymbol{\Sigma} \frac{\partial \boldsymbol{V}}{\partial \sigma_t^2} \boldsymbol{\Sigma\Sigma Y}\right] \tag{10}$$

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

$$\boldsymbol{V} = \begin{bmatrix} \mathbf{Z}_*\mathbf{D}_1\mathbf{Z}_*' + \mathbf{I}\sigma_{\varepsilon_1}^2 & \cdots & \mathbf{Z}_*\mathbf{D}_{1\mathrm{m}}\mathbf{Z}_*' + \mathbf{R}_{1\mathrm{m}} \\ \vdots & \ddots & \vdots \\ \mathbf{Z}_*\mathbf{D}_{1\mathrm{m}}\mathbf{Z}_*' + \mathbf{R}_{1\mathrm{m}} & \cdots & \mathbf{Z}_*\mathbf{D}_{\mathrm{m}}\mathbf{Z}_*' + \mathbf{I}\sigma_{\varepsilon_m}^2 \end{bmatrix} \tag{11}$$

Data Simulation

In the simulation scenario design, the number of multiple traits is assigned as three and number of samples are 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 as nuisance. Genetic background effects are simulated based on kinship estimates. For the kinship estimates, 500 bi-allelic SNPs (Single Nucleotide Polymorphism) are simulated for 1000 samples and the level of genotype is generated as relying on the 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 is simulated as a categorical variable and as to influence the multiple traits with different proportions w =0, 0.3, 0.5, 0.8 and 1. Due to multivariate structure of design, three different level of correlation ρ =0.25, 0.5 and 0.75 between traits are considered and the correlated effect is simulated from multivariate normal distribution (MVN) with the considered level of correlation. Proportion of genetic variance σ_{g}^2 is set to 0.30. The correlation structures under some remarkable scenarios are visualized by heat maps as given in Figure 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, $h^2 = \sigma_{\mathrm{g}}^2/(\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2)$. However, in a multi-environment design with L different locations, the heritability depends on the proportion of total genetic variance over L locations and can be estimated as

$$h^2 = \frac{\sum_{l=1}^{m} \sigma_{\mathrm{g}l}^2}{\sum_{l=1}^{m} \sigma_{\mathrm{g}l}^2 + \sigma_\varepsilon^2} \tag{12}$$

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

Table 1. Heritability estimations

	Proportion associated with location	**Level of correlation**		
		$\rho = \mathbf{0.25}$	$\rho = \mathbf{0.5}$	$\rho = \mathbf{0.75}$
mvLMM with a general variance component	$\boldsymbol{w} = \mathbf{0}$	0.2741	0.2966	0.2834
	$\boldsymbol{w} = \mathbf{0.2}$	0.2707	0.2831	0.2416
	$\boldsymbol{w} = \mathbf{0.4}$	0.2561	0.2721	0.2311
	$\boldsymbol{w} = \mathbf{0.6}$	0.2489	0.2504	0.2341
	$\boldsymbol{w} = \mathbf{0.8}$	0.2313	0.2361	0.2229
	$\boldsymbol{w} = \mathbf{1}$	0.2329	0.2375	0.2266
mvLMM with location specific variance components	$\boldsymbol{w} = \mathbf{0}$	0.2895	0.2896	0.2756
	$\boldsymbol{w} = \mathbf{0.2}$	0.2891	0.2889	0.3016
	$\boldsymbol{w} = \mathbf{0.4}$	0.2995	0.2797	0.2967
	$\boldsymbol{w} = \mathbf{0.6}$	0.2930	0.2879	0.2954
	$\boldsymbol{w} = \mathbf{0.8}$	0.2964	0.2849	0.3028
	$\boldsymbol{w} = \mathbf{1}$	0.3013	0.2966	0.3016

In Table 1, $w = 0$ indicates the absence the 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 dependency level of multiple traits don't effect estimations remarkably.

Application to Real Data

In this section, to examine the performance of the developed model, the two variance components models were fitted to safflower (Carthamus tinctorius L.) data collected from 3 different locations (Bolu, Haymana and Yenimahalle) in central Anatolian region, Turkey. The oil yield and the length were considered as dependent variables. The correlation between yield and length across 3 different environments is illustrated by Figure 2.

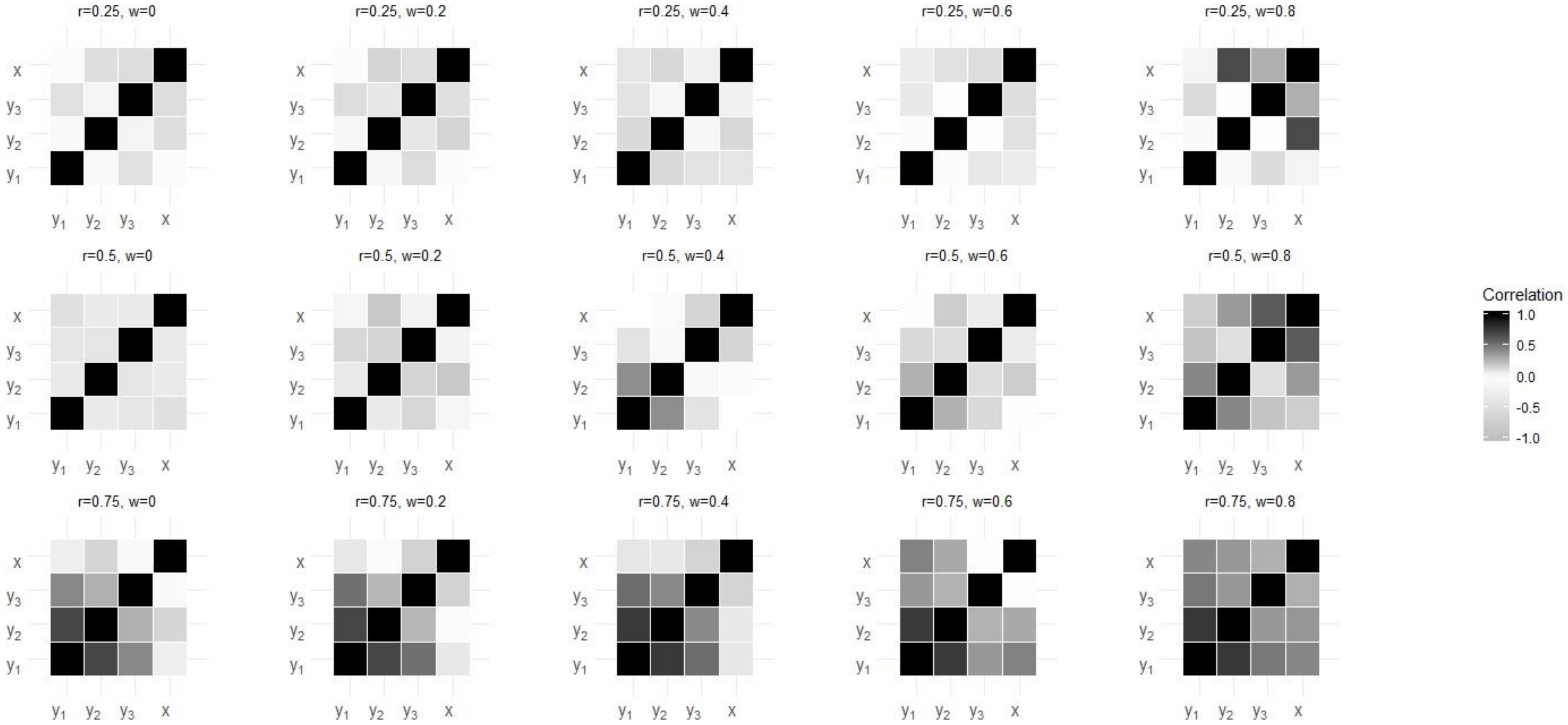

Figure 1. Illustration of trait by trait correlations of the simulated data

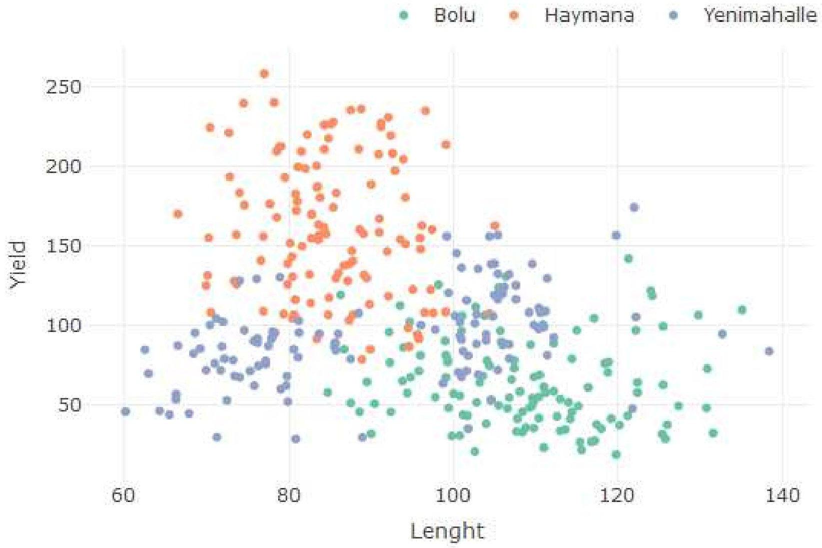

Figure 2. Correlation across three different environments

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

Table 2. Model results

	mvLMM with a general variance component	**mvLMM with location specific variance components**
Number of VC	6	12
AIC	653.73	236.06
BIC	662.88	245.21

VC: Variance Components

In standard mvLMM, six components ($\sigma^2_{g_{Yield}}$, $\sigma^2_{g_{Lenght}}$, $\sigma^2_{g_{Yield-Lenght}}$, $\sigma^2_{\mathrm{e}_{Yield}}$, $\sigma^2_{e_{Lenght}}$, $\sigma^2_{e_{Yield-Lenght}}$) were estimated, whereas for heterogeneous mvLMM the number of components were doubled as including separate components for each environment. For instance, in standard mvLMM, the genetic variance component of yield was represented by $\sigma^2_{g_{Yield}}$, however in heterogeneous mvLMM, the genetic variance component of yield had its own sub-components $\sigma^2_{g_{Yield(B)}}$, $\sigma^2_{g_{Yield(H)}}$ and $\sigma^2_{g_{Yield(Y)}}$ for Bolu, Haymana and Yenimahalle, respectively.

Discussion

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

Acknowledgement: The author special thanks to Prof. Dilek BAŞALMA (Ankara University, Faculty of Agriculture, Turkey) for her assistance with data validation.

References

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.

Lynch M, Walsh B (1998) Variance component estimation. In: Lynch M, Walsh B, editors. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer Associates, Inc. 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.

Searle SR, Casella G, McCulloch CE (1992) Variance components. New York (NY): John Wiley & Sons.

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.

Wray N, Visscher P (2008) Estimating trait heritability. Nature Education 1(1):29.

Figures

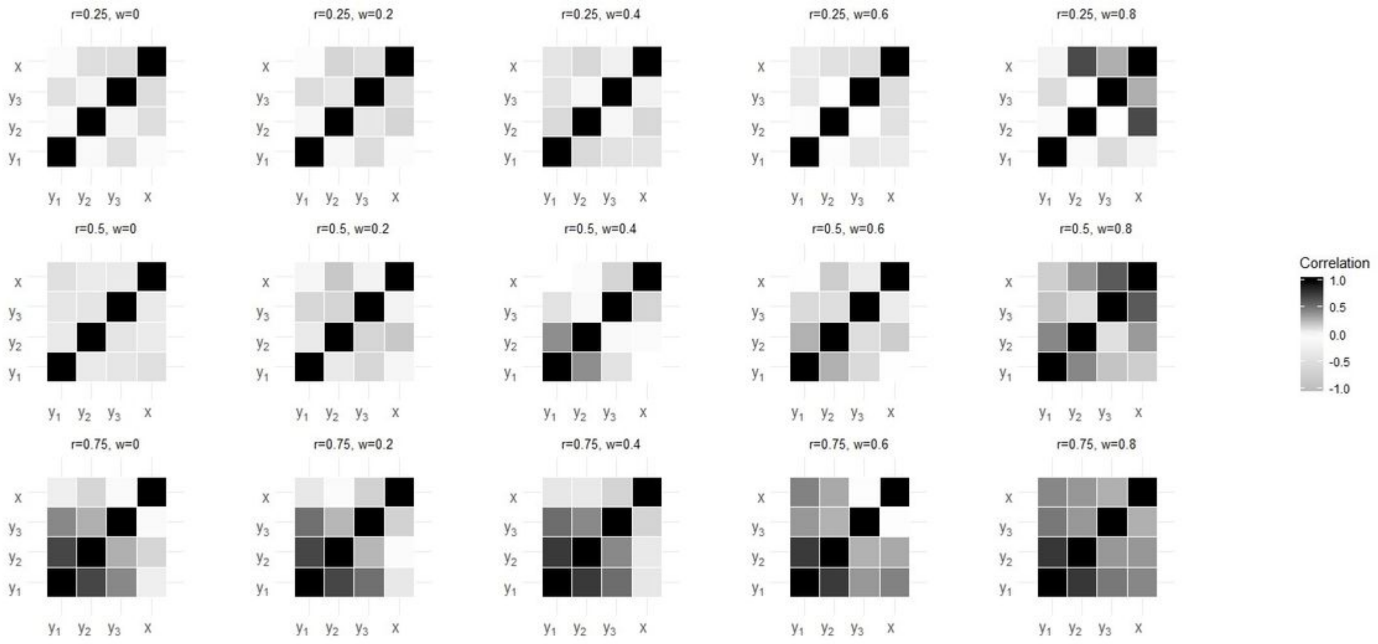

Figure 1

Illustration of trait by trait correlations of the simulated data

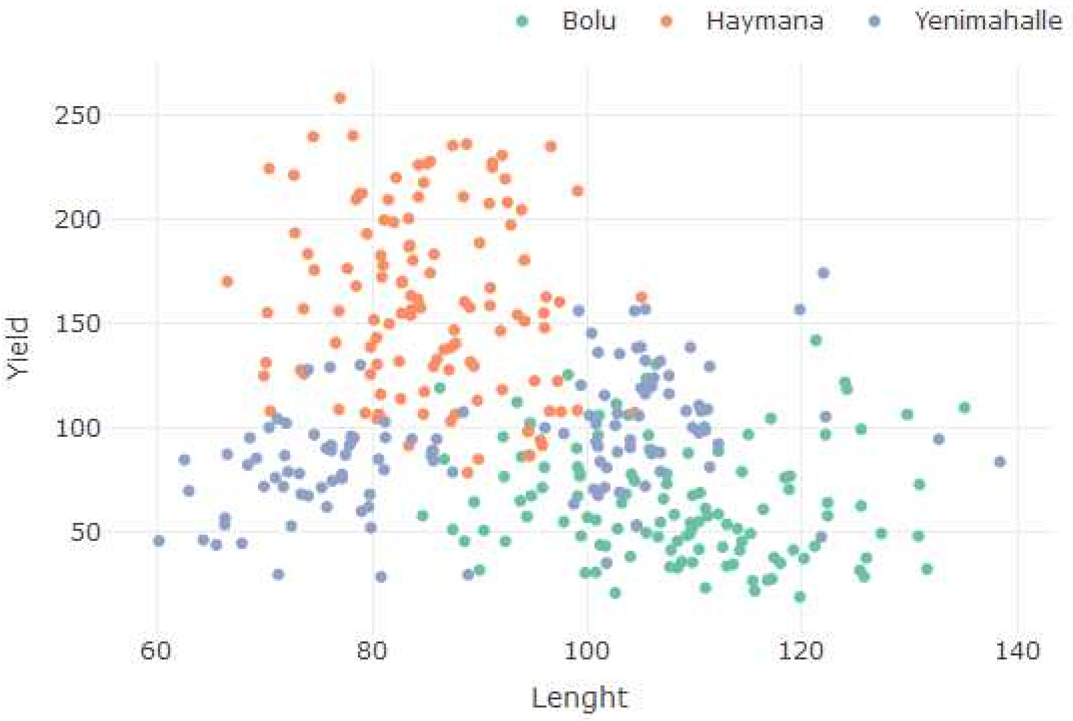

Figure 2

Correlation across three different environments