Disentangling Pleiotropy along the Genome using Sparse Latent Variable Models

L. L. Janss
Center for Quantitative Genetics and Genomics, Aarhus University, Denmark

ABSTRACT: Bayesian models are described that use latent variables to model covariances. These models are flexible, scale up linearly in the number of traits, and allow separating covariance structures in different components at the trait level and at the genomic level. Multi-trait version of the BayesA (MT-BA) and Bayesian LASSO (MT-BL) are described that model heterogeneous variance and covariance over the genome, and a model that directly models multiple genomic breeding values (MT-MG), representing different genomic covariance structures. The models are demonstrated on a mouse data set to model the genomic covariances between body weight, feed intake and feed efficiency.
Keywords: genomic variance; genomic correlation; latent variables; Bayesian

Introduction

Multi-trait models are essential tools for the animal breeder. The classical pedigree-based multi-trait model separates phenotypic covariances in genetic and environmental covariances, and this is crucial information to predict correlated selection responses and to optimally weigh traits in a selection index. The genomic era holds the promise to further disentangle genetic (or then: *genomic*) correlations by identifying groups of SNPs that create different correlation patterns. The holy grail in genomic breeding would be to construct genomic predictors with correlated responses that deviate from the overall genetic correlation structure. This, however, needs sophisticated models that allow for variable variance/covariance structures in SNP effects, or that group SNPs according to different genomic covariance patterns. These will most likely be Bayesian models.

Multi-trait models have been considered in various QTL mapping (e.g., Lund et al. 2003, Meuwissen and Goddard, 2004) and association mapping (Ferreira and Purcell, 2009) applications, but limited work has been done to develop multi-trait whole-genome mapping and prediction models. Calus and Veerkamp (2011) developed a multi-trait variable selection model, but with constant covariance for SNP effects; other multi-trait genomic models are based on GBLUP, which has implicit assumptions of equal variance and covariance contributed by every SNP. Hence, limited, if any, work has been done to develop genomic models that infer different covariance patterns over the genome or that group SNPs according to different covariance patterns.

Here, Bayesian multi-trait models are described that use latent variables and hierarchical models to flexibly model and disentangle covariances. To my knowledge, these particular Bayesian latent variable models have not been considered before, although recursive models (Varona et al., 2007) and state-space models (Piepho and Ogutu, 2007; Forni et al., 2009) are in a similar spirit. Notably, state-space models use the same idea to introduce latent variables ('random effects') to model spatial or longitudinal covariance structures. Bayesian models that use latent variables to model covariance structures between traits, however, have not yet been described.

The aim of this study is to, firstly, present the general ideas of modeling covariances between traits using latent variables. Subsequently, genomic models are described that model heterogeneous covariance. Multi-trait versions of the BayesA and the Bayesian LASSO model are described, and a model that directly estimates genomic values that represent different covariance structures. Earlier versions of the models described here were used in Sørensen et al. (2012) and Krag et al. (2013) to estimate genomic correlations using bivariate models. Bouwman et al. (2014) used a 14-trait Bayesian polygenic model based on the same latent variable technique. However, this earlier work used Gaussian distributions in the latent variables, making them close to frequentist multivariate models. Here the latent variable models are extended to use heterogeneous and sparse latent variables, allowing to model heterogeneous covariances over the genome and to disentangle pleiotropy.

Materials and Methods

Modeling covariance using latent variables. Consider the model for three traits y_1, y_2, y_3:

$$\begin{aligned} y_1 &= X_1 b_1 + r_1 s + e_1 \\ y_2 &= X_2 b_2 + r_2 s + e_2 \\ y_3 &= X_3 b_3 + r_3 s + e_3 \end{aligned} \qquad (1)$$

with common ingredients $X_i b_i$ to model "fixed effects" using b_i~uniform, and e_i are residuals, with $e_i \sim N(0, I\sigma_i^2)$. The uncommon ingredients are the terms $r_i s$, where both r_i and s are model parameters, with r_i scalar and s a vector of the length of the y_i (an extension allowing for different lengths of y_i and missing data follows). The vector s is referred to here as latent variable (or latent vector), and also can be thought of as a vector of random effects. It is crucial to note that there is only one s vector, which is implied in the model for all three traits. The purpose of the latent vector s is to model covariances between the traits, and the 'regression coefficients' r_i determine the sizes of these covariances. Because the latent vector s models the covariances, the model residuals are taken as uncorrelated between traits, and σ_i^2 can be interpreted as the remaining variance in trait i not correlated to the other traits. The distributional assumptions for r_i and s are r_i~uniform and $s \sim N(0, I)$. The modeled variances and covariance are worked out as var(y_i) =var($r_i s + e_i$) = $r_i^2 + \sigma_i^2$, and

$\text{cov}(y_i, y_j) = \text{cov}(r_i s + e_i,\ r_j s + e_j) = r_i r_j$. This can be collected as:

$$var\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{bmatrix} r_1^2 & r_1 r_2 & r_1 r_3 \\ r_1 r_2 & r_2^2 & r_2 r_3 \\ r_1 r_3 & r_2 r_3 & r_3^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \quad (2)$$

The covariance structure that is modeled by the r_i parameters has the same form as the elements of a spectral decomposition and is also used in factor analytic models.

The model (1) lends itself well for a Bayesian MCMC-based implementation. The inclusion of the latent variable factorizes the likelihood, which allows updating location parameters within each trait without the need to consider correlations across traits. The conditional posterior distributions for b_i, r_i and s all break down to univariate Gaussian distributions.

Missing data in the traits can be accommodated by introducing additional design matrices, extending the model (1) for trait *i* to:

$$y_i = X_i b_i + r_i Z_i s + e_i \quad (3)$$

where now s will have a row for every record that has at least one of the traits observed, and Z_i will match s to the observed data in y_i. In this approach, there is no need to estimate (impute) missing data, which is required when parameterizing in variance-covariance matrices and sampling variance-covariance matrices from inverse Wishart distributions, e.g. as in Wang et al. (1994).

Covariance between random effects: MT-rrBLUP model. Covariance between random effects can be modeled by introducing latent variables in the expectation of the random effect, or by writing a hierarchical model. This can be used to develop a multi-trait rrBLUP (MT-rrBLUP, ridge regression or random regression BLUP) model. Because the single trait rrBLUP model has constant variance for all SNPs, the logical assumption in the MT-rrBLUP model is to also assume constant covariance for all SNPs. This is analogous to the assumptions that are implicit in single trait and multi-trait GBLUP models. The MT-rrBLUP model is specified by extending (3) with SNP effects:

$$y_i = X_i b_i + r_i Z_i s + W a_i + e_i \quad (4)$$

where W is a matrix with SNP covariates (centered and possibly scaled), and a_i are SNP effects for trait *i*. SNP effects across traits are correlated by writing hierarchical models for the SNP effects:

$$a_i = v_i u + a_i^* \quad (5)$$
$$a_i^* \sim N(0, I\sigma_{ai}^2), u \sim N(0, I)$$

where now u is a latent variable that models the covariance between SNP effects across traits. Similar to the models at the trait level (1), (5) has 'residual SNP effects' a_i^* that are taken as uncorrelated across traits. Analogous to (2) it can be worked out that this model models $v_i^2 + \sigma_{ai}^2$ variance for the SNP effects for trait *i*, and $v_i v_j$ covariance between the SNP effects for trait *i* and *j*. Bivariate versions of this model were used in Sørensen et al (2012) and Krag et al (2013) to estimate genomic correlations.

Heterogeneous genomic covariance models: MT-BA and MT-BL. Heterogeneous variance-covariance models for SNP effects are developed by considering heterogeneous variance in the latent variable u, and optionally also in the 'residual SNP effects' a_i^* from (5). Here the approach known from the BayesA (Meuwissen et al., 2001) and Bayesian LASSO (Park and Casella, 2008) models is used by introducing a SNP-specific variance for every level in u. This develops the multi-trait BayesA (MT-BA) and multi-trait Bayesian LASSO (MT-BL) model by modifying (5) to:

$$u_i \sim N(0, \tau_i^2)$$
$$\tau_i^2 \sim \chi^{-2}(sc_i, df_i) \quad \text{MT-BA (6)}$$
or $$\tau_i^2 \sim Exp(\lambda) \quad \text{MT-BL (7)}$$

The BayesA and Bayesian LASSO only differ in the assumed distribution for these SNP-specific variances: in (6) a scaled-inverse chi-square with scale sc_i and degrees of freedom df_i is used, and in (7) an exponential distribution with rate λ. Because the variance explained by the latent variable is already modeled by the regression parameters v_i in (5), scale in (6) and rate in (7) can be taken known and are set to 1. The degrees of freedom parameter in (6) can be set to control the spread of the individual SNP variance around the common scale, and here a value of 5 was used.

Directly modeling multiple genomic values explaining different covariance structures: MT-MG model. An alternative approach, compared to the previous models, is to use latent variables in the trait models that represent genomic values, and in a hierarchical model match SNP effects to each of these genomic values. By constraining the signs of the regression parameters on these genomic values, each genomic value can be forced to explain a particular covariance pattern, and the model will map the SNPs that can be associated with that covariance pattern. For 3 traits y_1, y_2, y_3, and 2 vectors of genomic values g_1, g_2, this is the model:

$$y_1 = X_1 b_1 + v_{11} Z_1 g_1 + v_{21} Z_1 g_2 + r_1 Z_1 s + e_1$$
$$y_2 = X_2 b_2 + v_{12} Z_2 g_1 + v_{22} Z_2 g_2 + r_2 Z_2 s + e_2 \quad (8)$$
$$y_3 = X_3 b_3 + v_{13} Z_3 g_1 + v_{23} Z_3 g_2 + r_3 Z_3 s + e_3$$

The terms X_i, b_i, r_i, s, e_i are the same as in (1) and (3), and the $r_i Z_i s$ terms are included to model environmental covariances, allowing for missing data in the traits. Also the match of genomic values to records accounts for missing data in the traits by inserting the same Z_i design matrices introduced in (3). The vectors of genomic values have hierarchical models using all SNPs in each:

$$g_1 = W a_1 + g_1^* \quad (9)$$
$$g_2 = W a_2 + g_2^*$$

where $W a_i$ are again genotype covariates and SNP effects as in (4), except here a_i represents SNP effects for genomic vector g_i, and not for a particular trait. Indirectly, these SNP effects relate to traits, depending on the regression parameters v_{ij} in (8). The signs of the regression parameters v_{ij} are constraint to force g_1 and g_2 to explain different covariance patterns across traits. In the sequel, where such a 3-trait model with 2 genomic values is used, the constraints were $v_{11} > 0, v_{12} > 0$ so that g_1 explains a positive covariance between trait 1 and 2, and $v_{21} > 0, v_{22} < 0$ so that g_2 explains a negative covariance between trait 1 and 2. In this particular application, v_{13}, v_{23} were unconstraint, but in principle more vectors of genomic values can be added in order to also model different covariance patterns with trait 3.

In (8) the 'residual genomic values' g_i^* are confounded with the latent variable s that models residual covariance, and therefore var(g_i^*) was constraint to a small value. In this model it is convenient to scale the traits to have variance 1, and var(g_i^*) was set to 1/100. Some minimum variance is needed in these residual terms for the MCMC machinery to work. Further in (8) it is attractive to choose a sparse shrinkage or mixture distribution on the SNP effects a_i so that the model can map different sets of SNPs in each of the genomic vectors. As in (6-7) there is no need to model SNP variances at this level, and a Bayesian LASSO or Power LASSO with known rate can be chosen, or a mixture model with known proportions and variances. In the following application, a Bayesian Power LASSO (Gao et al., 2013) was chosen with rate parameter 1 and power parameter 0.5.

Estimating total explained genomic variance and covariance. To compute genomic explained variance for trait i, var(Wa_i) from (4) is evaluated at every MCMC cycle and the estimate of the posterior mean of genomic variance is the mean of these var(Wa_i) values. This is a generic way of evaluating genomic variance that can be used irrespective of the distributional assumptions used for the SNP effects. Because Wa_i represents genomic values (SNPs times SNP effects), this approach effectively computes the genomic values per MCMC cycle, and evaluates the genomic variance as the variance of genomic values, all per MCMC cycle. In an analogous fashion, the genomic variance in the MT-MG model is evaluates as var($v_{1i}Z_ig_1 + v_{2i}Z_ig_2$), or as var($v_{1i}Z_ig_1$) + var($v_{2i}Z_ig_2$) + 2cov($v_{1i}Z_ig_1, v_{2i}Z_ig_2$).

Data. The described multitait models are demonstrated on a data set that was previously described and analysed using univariate models by Ehsani et al. (2012). The data consists of a mouse F2 cross from inbred lines that were extreme in body weight and fatness. The traits analysed were Body Weight (BW), Feed Intake (FI) and Feed Efficincy (FE). Ehsani et al. (2012) showed that this mouse cross segregates for some large QTL affecting BW on chromosomes 1,2,9,10,11, affecting FI on chromosomes 2,7, and affecting FE on chromosomes 6,11,12. In contrast to the analyses by Ehsani et al. (2012) here more data was used: Ehsani et al. made a subset of 440 animals which also had gene expression data; in the current analysis no gene expression data was considered and a larger data set with 1163 BW records and 748 FI and FE records was used. The marker data used were 1806 biallelic SNPs.

Results

Genomic variances and correlations. Table 1 presents estimates for genomic heritabilities (genomic explained variance as proportion of the total variance) and genomic correlations for the traits Body Weight (BW), Feed Intake (FI) and Feed Efficiency (FE) using the presented genomic models. The multi-trait versions of the BayesA (MT-BA) and Bayesian LASSO (MT-BL) model produce nearly identical estimates than the rrBLUP model; the latter is equivalent to a GBLUP model to estimate genomic explained variance (Yang et al., 2010). The MT-MG model captures less of the genomic variance and correlation.

Table 1. Genomic heritabilities (h2) and correlations (rG) for Body Weight (BW), Feed Intake (FI) and Feed Efficiency (FE) in diferent genomic models.

	rr BLUP	MT-BA	MT-BL	MT-MG
h^2 BW	0.37	0.37	0.37	0.25
h^2 FI	0.30	0.31	0.31	0.14
h^2 FE	0.32	0.33	0.33	0.24
r_G BW-FI	0.82	0.81	0.82	0.65
r_G BW-FE	0.63	0.62	0.63	0.49

Genomic covariance profiles from rrBLUP, MT-BA and MT-BL models. Figure 1 present genomic correlations estimated in 5-marker windows for the genomic correlation between Body Weight and Feed Intake obtained from the MT-BA model. The overall genomic correlation between these traits was 0.81, but locally in the genome this correlation varies from about 0.3 to 0.9. The result from the MT-BL model was very similar, while the result from the MT-rrBLUP showed a smaller spread in the local genomic correlations of about 0.5 to 0.9.

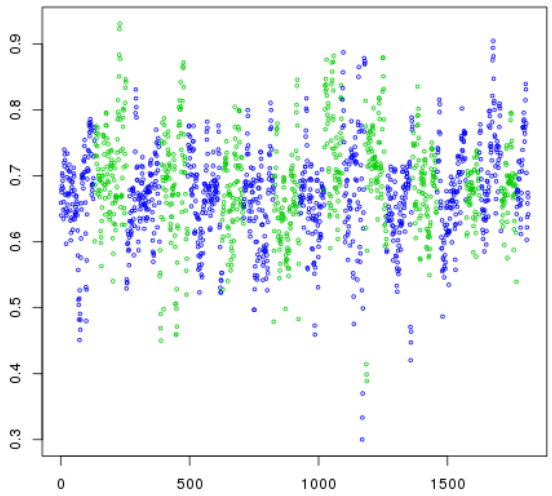

Figure 1. Genomic correlations between Body Weight and Feed Intake over the genome by marker (SNP number), obtained from the MT-BA model, smoothed in 5-marker windows and colored by chromosome.

De-constructing genomic covariance using the MT-MG model. Although the MT-MG model covered less of the genomic variance and covariance (Table 1), this model was able to construct two genomic breeding values showing different covariances with the traits. The correlations of these two genomic breeding values is graphically presented in Figure 2: the g_1 correlates positively to BW and FI (0.34 and 0.40, respectively), while g_2 correlates positively to BW but has approx. zero (slightly positive) correlation with FI (0.34 and 0.08, respectively). The correlations of these genomic values with

FE were -0.07 and 0.49, respectively. Hence, g_1 is the genomic value that predicts animals with higher body weight and higher feed intake (with equal or even slightly reduced efficiency), while g_2 is the genomic value that predicts animals with higher body weight without higher feed intake, i.e., the more efficient ones. Figure 3 presents SNP effects mapped in each of these two genomic values, showing that SNPs are often only mapped in one of the two genomic values.

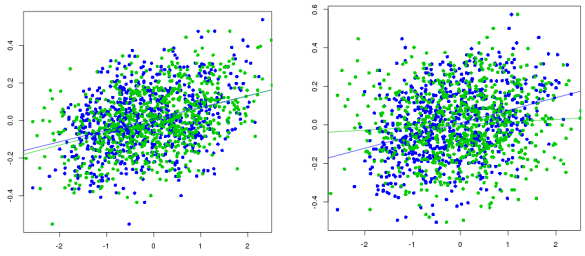

Figure 2. Deconstructed genomic values from the MT-MG model (vertical axes), left the g_1 and right the g_2 genomic value, versus the standardized phenotypes for Body Weight (blue) and Feed Intake (green) on the horizontal axes.

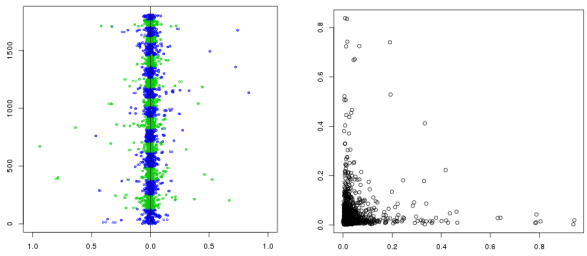

Figure 3. SNP effects mapped to the deconstructed genomic values from the MT-MG model: left graph inclusion probabilities for SNPs to map to g_1 towards the left, and to map to g_2 towards the right, colored by chromosome; right graph the same SNP inclusion probabilities plotted against each other.

Discussion and Conclusion

Further implementation of genomic selection in breeding programs will require extension of current univariate genomic evaluation models to multi-trait versions. A multi-trait version of the GBLUP model is in principle straightforward; this is the classical polygenic multi-trait model where the numerator relationship matrix **A** would be replaced by the genomic relationship matrix **G**. However, the commonly used **G**-matrices embed the implicit assumption of equal contribution of every SNP to the genomic variance. This is clear from deriving the GBLUP model from the rrBLUP model with constant variance for SNP effects (e.g., as in Yang et al., 2010). A multi-trait GBLUP model will embed the implicit assumption of equal contribution of every SNP to the covariance between traits. These GBLUP models may not be able to fully develop the potential of multi-trait genomic prediction. For many breeding applications it will be interesting to search for SNPs that show correlations between traits, which deviate from the overall correlation, or that show less strong correlation than the overall correlation. Models presented here allow identifying such SNPs. It is possible to extend the GBLUP models to use weighted **G**-matrices to use this information.

The data used in this study was an F2 between inbred lines, which does not allow resolving genomic effects to a very detailed level. In other types of populations, with finer LD structure, it will be interesting to run the same models with larger SNP densities. This is computationally feasible, e.g. Krag et al. (2013) ran similar models using close to 600K SNPs.

The Bayesian multi-trait models presented here are also in other settings quite competitive with REML approaches for estimating covariances. This was demonstrated by Bouwman et al. (2014) who used the same latent variable techniques to fit a 14-trait polygenic model.

Literature Cited

Bouwman, A.C., Valente, B.D., Janss, L.L.G., et al. (2014). Genet. Sel. Evol. 46:2.

Calus, M.P.L., Veerkamp, R.F. (2011). Genet. Sel. Evol. 43:26.

Ehsani, A., Sørensen, P., Pomp, D., et al. (2012). BMC Genomics 13:456.

Ferreira, M.A.R., and Purcell, S.M. (2009). Bioinformatics 25(1): 132-133.

Forni, S., Gianola, D., Rosa, G.J.M. (2009). J. Anim. Sci. 87: 3854-3864.

Gao. H., Su, G., Janss L. et al. (2013). J. Dairy. Sci. 96: 4678-4687.

Krag, K., Poulsen, N.A., Larsen, M.K., et al. (2013). BMC Genetics 14:79.

Lund, M., Sorensen, P., Guldbrantsen, B., et al. (2003). Genetics 163: 405-410.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001). Genetics 157: 1819-1829.

Meuwissen, T.H.E., Goddard, M.E. (2004). Genet. Sel. Evol. 36: 261-279.

Park, T., Casella, G. (2008). J. Amer. Stat. Assoc. 103: 681-686.

Piepho, H.P., Ogutu, J.O. (2007). The Amer. Stat. 61(3): 224-232.

Sørensen, L.P., Janss, L., Madsen, P., et al. (2012). Genet. Sel. Evol. 44:18.

Varona, L., Sorensen, D., Thompson, R. (2007). Genetics 177: 1791-1799.

Yang, J., Benyamin, B., McEvoy, B.P. et al. (2010). Nat. Genetics doi:10.1038/ng.608.

Wang, C.S., Rutledge, J.J., Gianola, D. (1994). Genet. Sel. Evol. 26: 91-115.