Published in final edited form as:

Behav Genet. 2014 May ; 44(3): 232–239. doi:10.1007/s10519-013-9631-9.

Multivariate Genetic Analyses in Heterogeneous Populations

Gitta Lubke^{1,2} and Daniel McArtor¹

¹University of Notre Dame

²VU University Amsterdam

Abstract

Martin & Eaves (1977) proposed a multivariate model for twin and family data in order to investigate potential differences in the genetic and environmental architecture of multivariate phenotypes. The general form of the model is the independent pathway model, which differentiates between genetic and environmental influences at the item level, and therefore permits the decomposition to differ across items. A restricted version is the common pathway model, where the decomposition takes place at the factor level. The paper has spurred numerous studies, and evidence for differences in genetic and environmental architecture has been established for personality and several other psychiatric phenotypes by showing a better fit of the independent pathway model compared to the common pathway model. We show that genome-wide association studies (GWAS) that use an aggregate score computed from multiple questionnaire items as a univariate phenotype implicitly assume a similar structure as the common pathway model. It has been shown that in case of a differential genetic and environmental architecture, multivariate GWAS methods can outperform the univariate GWAS approach. However, current multivariate methods rely on the assumptions of phenotypic and genetic homogeneity, that is, item responses are assumed to have the same means and covariances, and genetic effects are assumed to be the same for all subjects. We describe a distance-based regression technique that is designed to account for subgroups in the population, and that therefore can account for differential genetic effects. A first evaluation with simulated data shows a substantial increase of power compared to univariate GWAS.

Introduction

Behavioral and psychiatric traits are usually not directly observable, and are typically measured with multiple questionnaire items or symptom scales that cover the different aspects of the behavior of interest. Common factor analysis is a popular statistical tool that relates multiple observed items to a small number of factors that represent the unobserved traits. The factor model is a linear regression model where a factor functions as an unobserved predictor of multiple observed items, thereby causing the items to covary. In the genetic analysis of multivariate data, one of the dominant questions is how much of the covariation of observed items is due to genetic and/or environmental factors. Genetic and environmental factors are thought to contribute to the covariances of the items additively;

however, the pattern of regression coefficients or loadings is not necessarily the same for the genetic and the environmental factors. It is possible that a subset of items is more heavily influenced by shared genetic factors whereas a different subset of items covaries mainly because of environmental influences. Figure 1a illustrates a scale consisting of 5 items where G (genetic effects) and E (environmental effects and error) have a different loading pattern. The result is that the proportion of variances and covariances of the observed items that is due to genetic and/or environmental factors is item specific.

The possibility of differential genetic and environmental influences is at the very core of the seminal paper by Martin & Eaves (1977) in which they propose a multivariate model for twin and family data. The model is designed to estimate the genetic and environmental contributions to the covariation of a set of observed items. Importantly, the decomposition of observed variances and covariances is achieved at the item level by having all items load directly on genetic as well as environmental factors. Martin & Eaves' model therefore provides a means to test whether the genetic factor loadings are collinear¹ with the environmental loadings. The test is achieved by comparing models with and without constraining the loadings to be collinear. Collinearity of the two loading matrices implies that a much more constrained model fits the data, namely one in which the observed items load on a common "psychometric" factor, which is in turn influenced by genetic and environmental factors. In later literature, the item-level decomposition model has been called the independent pathway or biometric model whereas the factor level decomposition model has been termed the common pathway or psychometric model (Kendler et al., 1987, McArdle & Goldsmith, 1990). The common pathway model is illustrated in Figure 1b.

Note that since the item covariances in the common pathway model are modeled by a common psychometric factor, the effects of the genetic and environmental factors on the observed items are completely moderated by the psychometric factor, and are therefore necessarily multiples of the phenotypic loadings (see Figure 1b, the phenotypic loadings are either multiplied by g or by e to obtain the loadings of items on the genetic and environmental factors). Note also that in this model the genetic and environmental loadings are not only mutually collinear but also by design collinear with the phenotypic loadings. In other words, the phenotypic factor structure mirrors the genetic as well as the environmental factor structure.

Eaves et al. (1977) applied their model to 40 items measuring impulsivity, and showed that the genetic and environmental loading matrices were indeed collinear. Using the same method, Heath & Martin (1990) showed that for psychoticism as measured by the P-scale of Eysenck's Personality Questionnaire this was not the case. Establishing evidence for non-collinearity of loading matrices is a matter of statistical power, and the two studies differed considerably in sample size ($N=588$ twin pairs in the impulsivity study vs. 2,903 pairs in the psychoticism study). A more recent study of the comorbidity of different personality disorders also provided evidence for a differential genetic and environmental architecture (Kendler et al., 2008). In addition to personality phenotypes, differences in genetic and

¹Collinearity means that the two loading matrices are multiples of each other, and therefore impose the same pattern of interrelations between items.

environmental influences on multivariate outcomes have been established for disorders such as depression (Kendler et al. 1987), obsessive compulsive disorder (Iervolino et al., 2011), and autism spectrum disorder (Ronald et al. 2006), among others. These results show that the phenotypic factor structure does not coincide with the genetic factor structure for a number of psychiatric phenotypes.

Twin and family models such as the multivariate model proposed in Martin & Eaves (1977) utilize the expected degree of genetic similarity between different types of family members (MZ, DZ twins, parents, sibs) across the whole genome to achieve the decomposition into genetic and environmental variance components depicted in panels A and B of Figure 1. Twin and family models thus target the totality of genetic effects. However, the question of whether or not the phenotypic structure of interrelations between items is mirrored by the structure of genetic effects is similarly relevant when investigating the effects of single nucleotide polymorphisms (SNPs).

GWAS: Assuming a common pathway model

In genome-wide association studies (GWAS) of behaviors measured with multivariate instruments, the individual items are commonly aggregated into a univariate score to ease the analysis of millions of SNPs. The aggregation is commonly done by summing the item scores, which can be represented by a single factor model where all loadings are fixed to one, and all measurement errors are fixed to zero. In a GWAS the aggregate score is then used as a univariate phenotype and is regressed on each SNP separately in a simple linear regression model. Conceptually, the approach therefore reflects the common pathway model, and assumes that the effects of the SNP on the individual items are completely moderated by the psychometric factor, and that therefore the regression coefficients of the items on the SNP are collinear with the loadings of the model used to aggregate the items. This is illustrated in Figure 2. Since a sum score is computed with equal weights for all items, a GWAS using a sum score implicitly assumes that SNPs have the same effect on all items.

Note the similarity between Figure 1b and Figure 2. The collinearity argument is the same as in the common pathway model with G in Figure 1b being replaced by a single SNP in Figure 2. Since the genetic factor is limited to a single SNP, all other genetic effects are now included in the second factor. The interpretation of the second contributor to the variance of the psychometric factor is therefore different, instead of environmental influences and error denoted by E in Figure 1b it also comprises all other genetic effects, hence the notation GE in Figure 2.

The power of the univariate GWAS model to detect SNP effects is expected to be optimal if there is in fact a single factor underlying the aggregated items, and if the vector of regression coefficients of items on the SNP is collinear with the loadings on the single factor. As outlined in Martin & Eaves (1977), collinearity is a testable hypothesis, and the possibility of deviations from the more constrained common pathway scenario where the SNP effect is completely mediated by the psychometric factor has spurred the development of multivariate alternatives to the univariate GWAS model (e.g., O'Reilley et al., 2012, van der Sluis, 2013).

It is likely that the advantages of the different methods depend at least to some degree on the mechanism leading to non-collinearity of genetic regression coefficients and phenotypic loadings. This paper focuses on two different scenarios that may cause the genetic structure to deviate from the phenotypic structure, and evaluates a new method that has the potential to provide better power to detect effects under those conditions.

Two mechanisms causing a decrease of power in the univariate GWAS model

There are two main mechanisms that imply that regression coefficients of the items on the SNP are not collinear with the loadings of the phenotypic model used to aggregate items. First, it is possible that a SNP affects only a subset of the interrelated items. This is likely to occur if the set of observed items measure aspects of a behavior that belong to somewhat different biological pathways. The different proposed multivariate alternatives to the commonly used univariate GWAS model mainly focus on this scenario (e.g., MultiPhen, O'Reilley et al., 2012, Bates, van der Sluis et al., 2013). In a recent simulation study comparing several multivariate methods, van der Sluis et al. (2013) found considerable gains in power if analyses use the item level data rather than the sum score.

Second, it is possible that a SNP affects items differentially in subgroups of subjects. As an example, Lamers et al. (2012) showed in a recent study that there are considerable differences in biological correlates between individuals with melancholic depression and individuals with atypical depression. It is likely in such a case that at least a subset of SNPs relate differentially to depression items in these two subgroups. SNP effects in the univariate GWAS model are modeled as fixed effects, and therefore estimate the SNP effect averaged over subgroups. To our knowledge, no methods have been evaluated that target this type of differential SNP effects, and that can be applied in a genome-wide context. The COSA-MDMR approach outlined below is a promising method, which is designed to take differential SNP effects into account as well as clustering of subjects. COSA-MDMR can be used genome-wide.

The two mechanisms are important to consider because both can decrease the statistical power of the univariate GWAS analysis. The aggregate score used in GWAS represents the common content of all items. If a SNP affects only a few aspects of a multivariately measured behavior, then this effect is diluted in the aggregate score. The extent depends on how much the affected items deviate from the common core and measure a more specific aspect of the phenotype. Neglecting the population heterogeneity described in the second scenario can cause loss of power depending on the relative size of the cluster(s) affected by the SNP because the association estimated in the univariate GWAS model is an unweighted average effect size across clusters.

COSA-MDMR

The COSA-MDMR approach combines the Clustering Objects on Subsets of Attributes (COSA) algorithm developed by Friedman & Meulman (2004) with regression methods for distance matrices that have been developed in ecology (Anderson 2001, McArdle &

Anderson, 2001), and that have recently been used under the name Multivariate Distance Matrix Regression for the analysis of expression data (Wessel et al., 2007, Zappala & Schork, 2006, Schork & Zappala, 2012). COSA-MDMR is a framework that differs conceptually and statistically from factor analytic methods. Rather than relying on an apriori specified parametric model that relates items to underlying factors, COSA is an entirely data-driven algorithm that computes distances between all pairs of subjects based on the observed item level data. COSA distances account for heterogeneity, and can be regressed on SNPs using MDMR.

COSA—The COSA algorithm² computes optimized, non-Euclidean distances between pairs of individuals, and extends existing methods for distance calculation. In the simplest case, pairwise distances for multivariate data can be computed as Euclidean distances by summing the squared differences between two subjects on each of the items, and taking the square root of this sum. The resulting distances are based on the assumption that the population is homogeneous and all items are equally important because each item has the same weight for all subjects. The weights are not optimized. Several algorithms have been developed that optimize the weights by allowing items to have a differential impact on the pairwise distances. COSA provides a further extension by permitting the weights to also vary across subgroups of subjects. As a result, sets of items can contribute to the pairwise distances differently for different subgroups of subjects. The optimization problem is conceptually somewhat comparable to estimating cluster specific loadings in factor mixture models. Suppose individuals cluster on different aspects of a disorder. In that case items that measure aspects of the phenotype that characterize a given cluster will contribute more heavily to the distances between subjects within that cluster. COSA distances aggregate item level information, however, this is achieved while preserving information regarding clusters in the population as well as the relative importance of items within the clusters. COSA has been evaluated with empirical and simulated data, and provides superior performance compared to standard clustering or feature detection algorithms in case of population heterogeneity (Friedman & Meulman, 2004, but see Witten & Tibshirani, 2010). An example of an application to empirical data in a genetic context is Draisma et al.'s use of COSA in the analysis of lipids data (Draisma et al., 2008).

MDMR—Pairwise distances are most commonly used in cluster analyses. Gower (1966) provided the basis to use distances as an outcome in a regression. Pairwise distances can be arranged in an $N \times N$ distance matrix, where N is the number of subjects. If Euclidean distances are computed for a single outcome, then the sum of the diagonal elements of the distance matrix equals the sums of squares of the univariate outcome that is commonly used in standard least squares regression. This is due to the fact that a distance matrix computed for a single variable Y equals the outer product YY' , and the trace (=sum of diagonal elements) of an outer product equals the trace of an inner product $Y'Y$. The latter is the usual sums of squares. Gower (1966) provided a transformation of the distance matrix in case other distance metrics are used and/or multivariate items. The sum of diagonal elements of the transformed distance matrix can be decomposed into sums of squares due to

²COSA is freely available at <http://statweb.stanford.edu/~jhf/COSA.html>

regression and sums of squares due to error analogous to standard regression. The construction of a pseudo F-statistic is also analogous to regression. However, since it does not follow an F distribution, the statistic has to be tested in a permutation test. Distance matrix regression is often used in ecology due to the reliance on non-Euclidean distances. The method has been evaluated for the use in high dimensional data by Schork & Zapala (2012), who showed that power is comparable to other methods.

COSA and MDMR can be combined in a two-step process that takes into account potential clustering in the population as well as differential importance of phenotype items. In the first step, pairwise COSA distances between subjects are computed from a multivariate phenotype, and in the second step, these distances are regressed onto a SNP using MDMR. Efficient programming makes the procedure computationally feasible for genome-wide analyses. The next section covers a first small simulation study which aims at quantifying differences in power between univariate GWAS and COSA-MDMR in case only a subset of items is related to a SNP in only a subgroup of subjects.

Evaluation of COSA-MDMR using simulated data

The data-generating model is a two-class mixture model featuring a single SNP that affects a small subset of a total of 50 phenotype items in one of the classes (“affected class”). Specifically, the SNP affects either 5 or 20 out of the 50 phenotype items in the affected class. The remaining items are not associated with the SNP in either class. The size of the affected class is 10%, 20%, 30%, or 100% of the population, with the last case chosen to assess the performance of COSA-MDMR using a phenotype without population heterogeneity. Items are generated with moderate covariances that represent other genetic or environmental influences. This is achieved using an underlying factor in the data generating model.

The SNP explains 55%, 15%, 7.5%, or 5% of the variance in the affected items in the affected class. As such, the overall sum score effect size varies across settings, and is inversely related to the size of the affected class and the number of items affected. The sample size is fixed at 1000. The effect size settings are chosen such that the performance of standard sum score GWAS ranges between zero and 100% detection of the SNP for this sample size. Each simulation setting was replicated 100 times.

COSA can be called from R (Friedman & Meulman, 2004). MDMR and the pseudo F-test are easy to implement in R.³ The permutation test was carried out with 600 permutations, which is substantially larger than the number of permutations reported for MDMR (Schork & Zapala, 2012), but which seemed necessary to obtain stable p-values in a prior evaluation. Performance of the two methods was evaluated using the number of replications in which the SNP was statistically significant

Table 1 summarizes the results. In the conditions where 20 of the 50 items had a SNP effect, the methods perform equally. However, the advantage of COSA-MDMR becomes clear in conditions where only a small subset of the phenotype items display an effect. When only 5

³We are currently testing a parallelized implementation of MDMR, which will be available on request.

of 50 items are affected, COSA-MDMR outperforms GWAS by up to a factor of 4.86, with equal performance between the methods only when they each have virtually perfect power. In the most favorable conditions, COSA-MDMR had a 68% detection rate compared to 14% for the GWAS model.

Notably, this advantage extends even to situations without population heterogeneity. With the SNP explaining 5% of the variance in 5 of 50 phenotype variables for all observations in the sample, COSA-MDMR shows a perfect detection rate compared to 21% for the GWAS model. In these settings, COSA-MDMR either matches or outperforms traditional GWAS when only a small subset of the phenotype items is affected.

Both of these methods suffer from inadequate power when the size of the affected class, the number of affected items, and the SNP's effect size are small. Our ongoing research focuses on the behavior of COSA-MDMR under these problematic conditions. The COSA algorithm has a number of parameters that can be adjusted to better capitalize on the characteristics of the data, and additional simulation will provide insight in the extent to which these adjustment increase the power of COSA-MDMR. In addition, we plan to address in how far the current result can be generalized to smaller effects but larger samples.

Discussion

Martin & Eaves (1977) emphasized the potential differences between genetic and environmental effects in the context of modeling multivariate data collected in twins and families. If the matrices of genetic and environmental loadings are collinear, the model reduces to the more constrained psychometric or common pathway model. In this model, observed items are regressed on a psychometric factor, which in turn is regressed on the genetic and environmental factors. If the factor loading matrices containing the regression coefficients are not collinear, decomposition into genetic and environmental effects needs to be achieved by directly regressing the items on genetic and environmental factors. The same principle applies when regressing multiple items on SNPs. Aggregating multiple items to obtain a univariate sum score prior to GWAS will only have optimal power if the SNP effects on the items are in fact collinear with the loadings used to aggregate items. Furthermore, estimation of a single fixed effect for each SNP is optimal only in the absence of heterogeneity among subjects.

Aggregate scores have the advantage that their reliability is higher than that of an individual item because variability due to specific item content and measurement error cancels out. To the extent that a GWAS is focused on finding SNPs related to the common core of the multiple items, aggregation is clearly useful, and this simplification of the phenotype can be warranted to increase power especially in meta-analyses. However, the gain in power can come at the cost of loss of information or potential distortion of information. It is likely that the cost-benefit trade-off depends on the degree of phenotype heterogeneity. As shown in previous simulations, and in our first evaluation of COSA-MDMR, deviations from collinearity of genetic effects and phenotypic loadings used to aggregate the items as well as clustering within a sample can cause a decrease of power when using aggregate scores in GWAS (van de Sluis et al., 2013, Manchia et al., 2013).

COSA was developed with the aim of improving the detection of important items among a large number of candidate items in case the population of subjects in which the features are observed exhibits a clustered structure (Friedman & Meulman, 2004). Similar to sum scores COSA aggregates multivariate item level data, however it provides the double advantage that information regarding heterogeneity in the population is preserved, and that items have cluster-specific weights. Just as univariate GWAS, the COSA-MDMR approach proposed in this paper does, however, not account for possible epistatic effects since the distances are regressed on one SNP at a time. Methods such as CART and MARS provide this option (York et al., 2006), and might be combined with COSA phenotypes by using an optimization criterion similar to the one used in MDMR. Distances produced by COSA are most commonly used in cluster analyses. Multivariate Distance Matrix Regression (MDMR) provides an analog to standard regression that uses distances as outcome variables. MDMR has been tested for high dimensional data (Schork & Zapala, 2012), and applied successfully in the area of ecologic research (Anderson, 2001, McArdle & Anderson, 2001), and gene expression data (Wessel et al., 2007, Zapala & Schork, 2006).

The simulation presented in this paper shows promising gains of power in a gene-finding context. We are currently focusing on a more thorough evaluation of COSA-MDMR by assessing its performance in more complex settings (e.g., multiple SNPs affect different items), and in case item level data are ordinal or binary. Ongoing work also aims to improve COSA's performance. COSA has several procedural parameters that govern aspects such as targeting specific item responses (e.g., high scores), which can improve the resolution of the clustering structure in the distance matrix. Furthermore, previous simulations have shown that COSA is especially suited to detect important items in very large item collections (Friedman & Meulman, 2004), making the COSA-MDMR procedure especially suitable for the analysis of comorbid disorders. Using COSA distances computed from sets of items measuring comorbid disorders would shift the focus of a genome-wide analysis to detecting SNPs related to subsets of items that characterize comorbidity. In addition to our ongoing simulations therefore that focus on quantifying the power of COSA when multiple SNPs affect different sets of items in varying proportions of affected subjects we also address the question in how far results are generalizable to smaller effects in larger samples.

In our evaluation of COSA-MDMR presented in this article, almost all conditions included clustered populations. In case the population is not clustered with respect to the phenotype of interest, COSA might lose power due to permitting differences between clusters (Witten & Tibshirani, 2010). In our simulation this did not seem to be the case, but we only considered the very simple situation where a single SNP affects a subset of items. In practice, it is advisable to consider using COSA-MDMR if theory or prior research indicates potential heterogeneity. Latent class and other mixture analyses have shown phenotypic clustering for a number of psychiatric phenotypes (e.g., Lamers et al., 2012, Klonsky & Olino, 2008), and COSA-MDMR is a promising method to investigate genetic effects that are not homogeneous across clusters.

References

- Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*. 2001; 26(1):32–46.
- Draisma HH, Reijmers TH, Bobeldijk-Pastorova I, Meulman JJ, Estourgie-Van Burk GF, Bartels M, Ramaker R, van der Greef J, Boomsma DI, Hankemeier T. Similarities and differences in lipidomics profiles among healthy monozygotic twin pairs. *Omics*. 2008; 12(1):17–31. [PubMed: 18266560]
- Eaves LJ, Martin NG, Eysenck SB. An application of the analysis of covariance structures to the psychogenetical study of impulsiveness. *British Journal of Mathematical and Statistical Psychology*. 1977; 30(2):185–197.
- Friedman JH, Meulman JJ. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2004; 66(4):815–849.
- Gower J. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966; 53(3):325–338.
- Heath AC, Martin NG. Psychoticism as a dimension of personality: A multivariate genetic test of Eysenck and Eysenck's psychoticism construct. *Journal of Personality and Social Psychology*. 1990; 58(1):111. [PubMed: 2308068]
- Iervolino AC, Rijdsdijk FV, Cherkas L, Fullana MA, Mataix-Cols D. A multivariate twin study of obsessive-compulsive symptom dimensions. *Archives of general psychiatry*. 2011; 68(6):637. [PubMed: 21646580]
- Klonsky ED, Olino TM. Identifying clinically distinct subgroups of self-injurers among young adults: a latent class analysis. *Journal of Consulting and Clinical Psychology*. 2008; 76(1):22. [PubMed: 18229979]
- Kendler KS, Heath AC, Martin NG, Eaves LJ. Symptoms of anxiety and symptoms of depression: same genes, different environments? *Archives of General Psychiatry*. 1987; 44(5):451. [PubMed: 3579496]
- Kendler KS, Aggen SH, Czajkowski N, Roysamb E, Tambs K, Torgersen S, Reichborn-Kjennerud T. The structure of genetic and environmental risk factors for DSM-IV personality disorders: a multivariate twin study. *Archives of General Psychiatry*. 2008; 65(12):1438. [PubMed: 19047531]
- Lamers F, Vogelzangs N, Merikangas KR, de Jonge P, Beekman ATF, Penninx BWJH. Evidence for a differential role of HPA-axis function, inflammation and metabolic syndrome in melancholic versus atypical depression. *Molecular psychiatry*. 2012
- Martin NG, Eaves LJ. The genetical analysis of covariance structure. *Heredity*. 1977; 38(1):79–95. [PubMed: 268313]
- Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, et al. The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS ONE*. 2013; 8(10):e76295.10.1371/journal.pone.0076295 [PubMed: 24146854]
- McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*. 2001; 82(1):290–297.
- McArdle JJ, Goldsmith HH. Alternative common factor models for multivariate biometric analyses. *Behavior Genetics*. 1990; 20(5):569–608. [PubMed: 2288547]
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PloS one*. 2012; 7(5):e34861. [PubMed: 22567092]
- Ronald A, Happe F, Bolton P, Butcher LM, Price TS, Wheelwright S, Plomin R. Genetic heterogeneity between the three components of the autism spectrum: a twin study. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2006; 45(6):691–699. [PubMed: 16721319]
- Schork NJ, Zapala MA. Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis. *Frontiers in Genetics*. 2012; 3:190. [PubMed: 23060897]
- van der Sluis S, Posthuma D, Dolan CV. TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS genetics*. 2013; 9.1:e1003235. [PubMed: 23359524]
- Wessel J, Zapala MA, Schork NJ. Accommodating pathway information in expression quantitative trait locus analysis. *Genomics*. 2007; 90(1):132–142. [PubMed: 17493783]

- Witten DM, Tibshirani R. A framework for feature selection in clustering. *Journal of the American Statistical Association*. 2010; 105(490):713–726. [PubMed: 20811510]
- York TP, Eaves LJ, van den Oord EJ. Multivariate adaptive regression splines: a powerful method for detecting disease-risk relationship differences among subgroups. *Stat Med*. 2006; 25(8):1355–67. [PubMed: 16100739]
- Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the national academy of sciences*. 2006; 103(51):19430–19435.

Panel A: Independent pathway model

loading matrix relating items Y_1 - Y_5 to G and E

g_1	0
g_2	0
g_3	e_3
g_4	e_4
0	e_5

Note: the two vectors of loadings are not collinear unless $g_1=g_2=e_5=0$, and $g_3=c_1*e_3$ and $g_4=c_2*e_4$ where c_1 and c_2 are constants

Panel B: Common pathway model

loading matrix relating items Y_1 - Y_5 to G and E through P where P is standardized

g^*p_1	e^*p_1
g^*p_2	e^*p_2
g^*p_3	e^*p_3
g^*p_4	e^*p_4
g^*p_5	e^*p_5

Note: the two vectors of loadings are collinear since they are both multiples of the loadings of the items on

Figure 1.

Examples of independent and common pathway models

Note: Panel A shows an example of an independent pathway model where the genetic loadings differ from the environmental loadings. The loadings of the common pathway model in panel B are necessarily collinear. For reasons of simplicity, the decomposition of residual variances (depicted as single factors influencing each observed variable) is not included in the path diagrams.

GWAS of a multivariate phenotype is commonly carried out in two steps. First, items are aggregated into either a sum score by setting all loadings of Y_1 - Y_5 on P to 1 and all residuals to zero, or by fitting the phenotypic factor model, and saving factor scores. In the second step, P is considered observed, and a regression model of P on S is estimated. The variance of GE , which denotes environmental influences, error, and all other genetic effects influencing the aggregate phenotype, is estimated as residual variance in the GWAS model.

Figure 2.

GWAS model

Note: As in Figure 1, residual item variances are depicted as single factors influencing each observed variable.

Table 1

	GWAS	COSA-MDMR	Power gain factor
<i>Item-wise $R^2 = 0.55$</i>			
nAff = 100%, xAff = 5	1.00	0.97	0.97
nAff = 100%, xAff = 20	1.00	0.99	0.99
nAff = 30%, xAff = 5	0.31	0.95	3.06
nAff = 30%, xAff = 20	1.00	1.00	1.00
nAff = 20%, xAff = 5	0.14	0.68	4.86
nAff = 20%, xAff = 20	0.97	0.97	1.00
<i>Item-wise $R^2 = 0.15$</i>			
nAff = 100%, xAff = 5	0.35	0.97	2.77
nAff = 100%, xAff = 20	1.00	0.96	0.96
nAff = 30%, xAff = 5	0.06	0.28	4.67
nAff = 30%, xAff = 20	0.67	0.71	1.06
nAff = 20%, xAff = 5	0.05	0.11	2.20
nAff = 20%, xAff = 20	0.31	0.34	1.10
<i>Item-wise $R^2 = 0.075$</i>			
nAff = 100%, xAff = 5	0.23	0.98	4.26
nAff = 100%, xAff = 20	1.00	0.92	0.92
nAff = 30%, xAff = 5	0.06	0.24	4.00
nAff = 30%, xAff = 20	0.51	0.46	0.90
nAff = 20%, xAff = 5	0.05	0.07	1.40
nAff = 20%, xAff = 20	0.19	0.24	1.26
<i>Item-wise $R^2 = 0.05$</i>			
nAff = 100%, xAff = 5	0.21	1.00	4.76
nAff = 100%, xAff = 20	1.00	0.97	0.97
nAff = 30%, xAff = 5	0.05	0.13	2.60
nAff = 30%, xAff = 20	0.42	0.36	0.86
nAff = 20%, xAff = 5	0.05	0.05	1.00
nAff = 20%, xAff = 20	0.15	0.19	1.27

Note: nAff indicates percent of the population affected by the SNP, xAff is the number of phenotype items affected by the SNP, cell entries show the proportion of simulation replications where the SNP was statistically significant