Explaining heritability through genetic effects on covariance

Marida Ianni-Ravn, Andy Dahl

August 18, 2023

0.1 Introduction

Understanding the relationships between disease symptoms has been a longstanding pursuit in medical research. Unraveling these correlations not only contributes to our interpretation of disease pathophysiology but also has significant implications for diagnosis, treatment, and patient care.

One intriguing idea is that genetics influences the correlation structures among traits and symptoms, particularly within the context of complex diseases. A recent method, referred to as "correlation by individual level product" (CLIP), aims to find variants affecting the covariance between traits by treating the "individual-level product" (the normalised pairwise product of trait values) as a phenotype (Lea et al. 2019). This approach found and was able to replicate a set of "correlation QTLs" affecting correlations between whole blood-derived gene expression data. This provides some support that trait correlation structures might indeed exhibit inter-individual variability.

This observation raises a compelling hypothesis: a portion of the heritability associated with compound diseases could be attributed to genetic effects on the covariance of symptoms, rather than solely on the individual symptoms themselves.

This observation leads us to speculate about the potential implications within the context of a complex disorder like Major Depressive Disorder (MDD). The criteria for diagnosing MDD encompass the presence of specific symptoms set out in the DSM-5 criteria (Regier, Kuhl, and Kupfer 2013). These include persistent low mood, loss of interest or pleasure, notable changes in appetite or weight, persistent fatigue or loss of energy, and recurrent thoughts of death or suicide. To meet the criteria for an MDD diagnosis, patients must endure several of these symptoms for a specified duration. Consider the scenario where specific genetic variants predispose certain individuals to experience fatigue as a response to weight loss, while this correlation may not be as pronounced in others. The heightened susceptibility of these individuals to exhibit correlated symptoms could subsequently elevate their overall risk of receiving a diagnosis of MDD.

Despite the insights gleaned from CLIP, a comprehensive model for how individual-level variation might influence trait inheritance or disease risk has not been proposed. This gap not only constrains our intuition and limits the development of accurate estimation procedures. Moreover, while evidence supporting the existence of correlation QTLs has surfaced in the context of genetic

effects on gene expression data, such support has yet to extend to other traits.

In this project, we propose a model to explain for the inheritance of trait correlations. We show via simulation that this model can explain disease heritability, even in the case when there are no direct genetic effects on symptoms. We lastly propose an algorithm for estimating the heritability of covariance.

All the code for this project is available at the github repo `covariance-QTLs`.

1 The model

Suppose a disease diagnosis involves some function f on a set of p symptoms X. We are interested in assessing to what extent we can find heritability in the disease, purely due to the covariance between symptoms (rather than genetic effects on the symptoms directly).

To investigate this, we have a model where each individual has their own covariance matrix (Σ_i), which is fully determined by one parameter w. w represents the covariance between the symptoms: $\Sigma_i := w_i J + (1 - w_i)I$, where J is a $(p \times p)$ matrix of ones and I is the identity matrix, also $(p \times p)$. In other words, the covariance between each pair of symptoms is w_i.

Genetic effects occur on w: $w = \beta G + E$. Therefore,

$$w \sim MVN(0_n, h_w^2 K + (1 - h_w^2) I_n) \tag{1}$$

Where h_w^2 is the heritability of the covariance. Given w_i, we define individual-level symptom covariance matrices Σ_i: **[no need for sigma2g terms yeah?]**

$$\Sigma_i := w_i J + (1 - w_i) I \tag{2}$$

[need to define J, maybe even I too]

In practise, we truncate w values to lie between -1 and 1 (which is required in order for the symptom covariance matrices to be valid.

Each individual's symptoms are drawn from a multivariate normal with the respective covariance matrix:

$$X_{i,} | \Sigma_i \sim MVN(0_p, \Sigma_i) \tag{3}$$

Each individual's symptoms are independent given w. Finally, the disease is a function of the symptoms.

$$y_i = f(X_{i,}) \tag{4}$$

1.0.1 The likelihood

Under the model specified above, the likelihood function is

$$\begin{aligned}\mathcal{L}(w|X,K,h_w^2) &= P(X,w|K,h_w^2)\\ &= P(X|w)P(w|K,h_w^2)\\ &= (\prod_i \phi(X_i,0,w_iJ+(1-w_i)I)) * \phi(w,0,h_w^2K+(1-h_w^2)I)\end{aligned}$$

This gives the log likelihood

$$-2\ell(w|X,K,h_w^2) = \sum_{i=1}^{n}(p\ln 2\pi + \ln|w_iJ+(1-w_i)I| + X_{i,}^T(w_iJ+(1-w_i)I)^{-1}X_{i,}) \tag{5}$$

$$+ n\ln 2\pi + \ln|h_w^2K+(1-h_w^2)I| + (w^T(h_w^2K+(1-h_w^2)I)^{-1}w) \tag{6}$$

Here, part is independent between individuals, and reflects the covariance structure of an individual's symptoms. There is no kinship or heritability in this part. If we were to only consider this part, the maximum likelihood estimate of w_i would be $\tilde{X_{i,}}\tilde{X_{i,}}^T$, where the tilde indicates centred and scaled data.

Part (6) is calculated from the whole ensemble of w values, and represents the constraints of kinship.

The gradient of the log likelihood with respect to w is:

$$\begin{aligned}-2\nabla_w(\ell) &= \nabla_w(\sum_{i=1}^{n}(\ln|w_iJ+(1-w_i)I| + X_{i,}^T(w_iJ+(1-w_i)I)^{-1}X_{i,}) + \nabla_w\left(w^T(h_w^2K+(1-h_w^2)I)^{-1}w\right)\\ &= \sum_{i=1}^{n}[\frac{(p-1)pw_i}{(w_i-1)(w_i(p-1)+1)}\\ &+ \frac{1}{(1-w_i)^2}X_{i,}^T\left(I-\left(\frac{1+w_i^2(p-1)(p+1)}{(w_i(p-1)+1)^2}\right)JX_{i,}\right)]\\ &+ 2(h_w^2K+(1-h_w^2)I)^{-1}w\end{aligned}$$

We usually think of collecting only one vector of phenotypes per individual. However, repeated observations of symptoms over time could be used in the first part of the gradient (X_i, being a $(p \times t)$ matrix instead of a column vector).

1.1 Simulating symptoms from the model

We set out to show by simulation that running GWAS on a compound disease recovers genetic effects on the covariance. In order to do this, we simulated compound disease inheritance from our model. We chose to begin with a disease made up of $p = 2$ symptoms, in a scenario where there are no genetic effects on the symptoms, and the covariance heritability h_w^2 is 1. The simulation proceeds as follows:

- Simulate a matrix of genotypes and scale
- Draw covariance effect sizes for all l loci from $\mathcal{N}\left(0, h_2^2/lI\right)$
- Give each individual their covariance weight $w = \beta G + E$ where $E \sim \mathcal{N}\left(0, (1-h^2)I\right)$
- For each individual, draw symptoms as $(L\sqrt{D}\beta)G + L\sqrt{D}E$ where L and D are the Cholevsky factors of $\Sigma_i = w_i J + (1-w_i)I$
- For each individual, assign disease: $Y = f(X)$, where $f(X) = 1_{\geq 2 \text{ symptoms past threshold}}$

We then estimated effect sizes of each locus on the disease, as well as on the individual-level product of symptoms $((X_1 - \mu_1) * (X_2 - \mu_2))/\sqrt{Var(X_1) * Var(X_2)}$ and the true covariance weights w.

As expected, there was little correlation between the true covariance effect sizes and the estimated effect sizes on symptoms. However, we found strong correlation between the true covariance effect sizes and the estimated effect sizes for the disease. This indicates that a proportion of disease heritability can be explained by effects on the covariance, even when the symptoms are not heritable.

These simulations are carried out in the notebook `GWAS_on_covQTLs.Rmd`.

1.2 A gradient descent algorithm for estimating the covariance heritability

Here we set out an algorithm which, given individual-level symptom data, can estimate the individual-level covariance weights and heritability of the covariance. This algorithm assumes that the heritability of the symptoms is zero.

Let's separate out the gradient into

- $\nabla_{w_i, cov}\ell = -\frac{1}{2}\left(\frac{pw_i}{(w_i^2-1)} + \frac{1}{(w_i-1)^2}X_{i,}^T\left(I - \frac{w_i^2+p-1}{(1+w_i)^2(p-1)}J\right)X_{i,}\right)$: calculation requires only one individual's w_i and symptoms $X_{i,}$.

- $\nabla_{w,rel}\ell = -(h_w^2 K + (1-h_w^2)I)^{-1}w$: calculation requires the kinship matrix, h_{cov}^2, and the whole vector of w.

Algorithm 1 Estimating covariance heritability by gradient descent

$w_i \leftarrow (X_i, -\mu)(X_i, -\mu)^T \forall i$ ▷ Set w as the ILP
$h_w^2 \leftarrow \texttt{greml}(w)$
while not converged **do**
  for $i \in \{1, ..., n\}$ **do**
    $w_i \leftarrow w_i + \eta \times \nabla_{w_i,cov}\ell$ ▷ Update each w entry at a time
    $w_i \leftarrow w_i + \eta \times \nabla_{w,rel}\ell_i$ ▷ Update based on the vector of $w|h_2^2, K$
  end for
  $h_w^2 \leftarrow \texttt{greml}(w)$ ▷ Estimate a new h_w^2
end while

The notebook `gradient_descent.Rmd` implements this algorithm on symptoms simulated from the model.

2 Remaining Questions

- Can we find a maximum likelihood solution for w?
- Expected value of HE regression on y?
- How does the gradient descent algorithm work when the symptoms are also heritable directly?

3 Alternative models

3.1 The disease function

In our simulations, we chose a simple threshold model for disease, which takes two parameters: a threshold and a minimum number of symptoms. Altering these to parameters can affect disease heritability and prevalence. We provide the notebook `disease_function.Rmd` to explore this.

3.2 Inheriting covariance through Cholevsky factors

The LDL decomposition of the symptoms covariance matrix leads to the L matrix, which represents the connections between symptoms, and a D matrix, which holds the independent variances of the symptoms. In the notebook `effects_through_LDL.Rmd`, we give example implementations of both.

3.3 Deriving the log likelihood

Here we derive the gradient of the log likelihood with respect to w. Taking the gradient with respect to one entry of w:

$$-2\nabla_{w_j}(\ell) = \frac{\partial}{\partial w_j}(\ln|w_jJ + (1-w_j)I| + X_{j,}^T(w_jJ + (1-w_j)I)^{-1}X_{j,}) + \nabla_w(w^T(h_w^2K + (1-h_w^2)I)^{-1}w))_j$$

By the Sherman-Morrison formula, we can derive that $(w_jJ+(1-w_j)I)^{-1} = \frac{1}{(1-w_j)}\left(I - \frac{w_j}{1+w_j(p-1)}J\right)$. Tackling the first part of the partial derivative:

$$\begin{aligned}
-2\frac{\partial}{\partial w_j}(\ln|w_jJ + (1-w_j)I|) &= \operatorname{tr}\left((w_jJ + (1-w_j)I)^{-1}(J-I)\right) \\
&= \operatorname{tr}\left(\frac{1}{(1-w_j)}\left(I - \frac{w_j}{(1+w_j(p-1))}J\right)(J-I)\right) \\
&= \frac{1}{(1-w_j)}\operatorname{tr}\left(\left(I - \frac{w_j}{1+w_j(p-1)}J\right)(J-I)\right) \\
&= \frac{1}{(1-w_j)}\operatorname{tr}\left(-I + \left(1 - \frac{w_j(p-1)}{1+w_j(p-1)}\right)J\right) \\
&= \frac{p}{(1-w_j)}\left(\frac{-w_j(p-1)}{1+w_j(p-1)}\right) \\
&= \frac{(p-1)pw_j}{(w_j-1)(w_j(p-1)+1)}
\end{aligned}$$

Now tackling the second term:

$$\begin{aligned}
-2\frac{\partial}{\partial w_j}(X_{i,}^T(w_iJ + (1-w_i)I)^{-1}X_{i,}) &= -\frac{1}{(1-w_j)^2}X_{j,}^T\left(I - \frac{w_j}{1+w_j(p-1)}J\right)(J-I)\left(I - \frac{w_j}{1+w_j(p-1)}J\right)X_{j,} \\
&= \frac{1}{(1-w_j)^2}X_{j,}^T\left(I - \left(\frac{1+w_j^2(p-1)(p+1)}{(w_j(p-1)+1)^2}\right)J\right)X_{j,}
\end{aligned}$$

The last part of the gradient pushes the relationships in the vector of w to resemble the kinship matrix:

$$-2\nabla_w((w^T(h_w^2K + (1-h_w^2)I)^{-1}w)) = -2(h_w^2K + (1-h_w^2)I)^{-1}w$$

Therefore, the full log likelihood with respect to one entry of w is:

$$\begin{aligned}
-2\nabla_{w_j}(\ell) = & \frac{(p-1)pw_j}{(w_j-1)(w_j(p-1)+1)} \\
& + \frac{1}{(1-w_j)^2} X_{j,}^T \left(I - \left(\frac{1+w_j^2(p-1)(p+1)}{(w_j(p-1)+1)^2} \right) J \right) X_{j,} \\
& + [2(h_w^2 K + (1-h_w^2)I)^{-1} w]_{j,}
\end{aligned}$$

References

Lea, Amanda et al. (2019). "Genetic and environmental perturbations lead to regulatory decoherence". In: *Elife* 8, e40538.

Regier, Darrel A, Emily A Kuhl, and David J Kupfer (2013). "The DSM-5: Classification and criteria changes". In: *World psychiatry* 12.2, pp. 92–98.