The heritability curve and its applications to human non-Mendelian traits

Francesca Azzolini

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN

The heritability curve and its applications to human non-Mendelian traits

Francesca Azzolini

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 08.12.2023

Year: 2023

Title: The heritability curve and its applications to human non-Mendelian traits

Name: Francesca Azzolini

Print: Skipnes Kommunikasjon / University of Bergen

Acknowledgements

First, I would like to thank Hans Skaug for suggesting a research topic during my year as a researcher and for helping me turn it into a full-fledged PhD project. I am grateful for his insightful supervision both during my time in Bergen and my time in Ålesund.

I give my thanks to Geir Berentsen for his critical guidance during the early stages of my PhD.

I thank the researchers with whom I coauthored the papers in this thesis; it was a pleasure working with them.

For turning Bergen from an unknown city to a beautiful place to live in, I thank Erlend, Eugenia, Irene, Isaac, Mirjam, Morten, Stefano, Tommy and Wietse.

For spending quality time with me despite the distance, especially during the pandemic, I thank Daniel, Denis, Emanuele, Francesca, Jacopo, Silvia, and Tiziano.

For welcoming me into a warm and positive environment at NTNU, I thank my colleagues in Ålesund.

For the insightful conversations, I thank Martina and Hans.

For being my first and greatest supporters, I thank my parents and my family.

For his love, I thank Andrea.

Abstract in English

Heritability is an important measure to describe a non-Mendelian trait; it measures in which proportion the value of the trait is affected by genetic material.

In classical biometric models, the heritability is measured as a constant over the entire trait range. A constant heritability is seen as too reductive in more modern approaches, which instead work towards defining a continuous heritability, dependent on the trait value.

In this manuscript we define a heritability curve, a parametric measure of the heritability, based on a local definition of the correlation. Using a Gaussian mixture as the distribution underlying the data, we construct an explicit formula for the heritability curve. We study its properties and its use, applying it then on real human data.

For the estimation of the Gaussian mixture parameters, we use an algorithm based on automatic differentiation. This allows us to compute derivatives faster and improve the precision of the algorithm. We lastly study an Hamiltonian Monte Carlo algorithm to initialize the parameters in the optimization process.

Abstract in Norwegian

Arvbarhet er et viktig mål for å beskrive en ikke-mendelsk egenskap; den måler i hvilken andel verdien av egenskapen påvirkes av genetisk materiale. I klassiske biometriske modeller er arvbarheten konstant over hele egenskapsområdet. Dette blir sett på som for reduktivt i mer moderne tilnærminger. I stedet defineres kontinuerlig arvbarhet, som avhenger av verdien av egenskapen. I dette manuskriptet definerer vi en *arvbarhetskurve*, et parametrisk mål på arvbarheten, basert på en lokal definisjon av korrelasjon.

Ved å anta at fordelingen som ligger til grunn for dataene er en *Gaussian mixture*, konstruerer vi en eksplisitt formel for *arvbarhetskurven*. Vi studerer egenskapene og bruken av den, og bruker den deretter på registerdata fra mennesker.

For å estimere *Gaussian mixture* parametrene bruker vi en algoritme basert på automatisk differensiering. Det gjør at vi kan beregne derivater raskere og forbedre presisjonen til algoritmen. Vi studerer til slutt en *Hamiltonian Monte Carlo*-algoritme for å initialisere parametrene i optimaliseringsprosessen.

Contents

Chapter 1

Motivation

Genetic traits can be broadly separated into two categories: Mendelian traits and non-Mendelian traits. Mendelian traits are discrete and follow Mendel's law of inheritance; this means that the phenotype of a individual is almost entirely determined by a pair of alleles in the genome, one inherited by the mother, one by the father. In this context alleles are defined as either "dominant" or "recessive", with the phenotype associated with the dominant allele always manifesting when a dominant allele is present in the genotype, and the phenotype associated with the recessive allele manifesting only when a person inherits the recessive allele from both parents. One example of a Mendelian trait is albinism [Schalock et al., 2010]. The allele that determines most forms of albinism is recessive, so a person will be affected by albinism only if they receive a recessive allele from both parents.

Mendelian traits are overall easier to predict: by knowing the genome of both parents, one can straightforwardly calculate the probability of a Mendelian phenotype to manifest in their children.

Non-Mendelian traits, on the other hand, are continuous traits and as such more complex to model. They are often affected by both several genes and their interaction with each other, and by external environmental factors. An example of a Non-Mendelian trait is birth weight: it is a continuous value (for humans, usually ranging between 2.5 kilograms and 4.5 kilograms), it has a genetic component (there exists a positive correlation with the birth weight of the parents), and it is also affected by external factors (such as the diet of the mother, her smoking habits, and the number of gestation weeks [Kramer, 1987]).

The heritability is a measure of the dependence of a non-Medelian trait on genetics and it is a difficult value to quantify. There are currently two general approaches to

estimating the heritability of a trait. Thanks to modern technology, it is possible to generate a complete mapping of all human chromosomes, which makes it more feasible to identify the loci which influence specific non-Mendelian traits; for example, Yang et al. [2010] uses GWAS (*genome-wide association study*) to study the heritability of human height. Despite the technological advancement, this type of project is still quite challenging and time-consuming.

The second approach is based more on statistical models than biological experiments; instead of looking at the genes separately, it defines a latent variable which is responsible for all the genetic effects. Using well-known formulas that make use of family structures and the proportions of the genetic material that the family members share, it estimates both genetic and environmental effect on a trait. This is the approach that we follow in this manuscript.

The heritability is classically calculated as a constant over the entire range of the data, that is as a single value independent on the trait value. Recent studies (e.g. Logan et al. [2012] and Williams [2020]) have suggested that there might be more nuance in this measure, and that it might be affected by the trait value itself; for example, low values of a trait could be less heritable than higher ones.

This nuance is particularly interesting when applied to questions into the medicine field. Let us consider again the example of birth weight. A very low or very high value of weight at birth can cause severe medical conditions that can have immediate and possibly lasting consequences to the child's health (see, e.g. Barker and Osmond [1986], Reyes and Manalich [2005], and Palatianou et al. [2014]). If such values of birth weight depend mostly on the environment, an effort can be made to prevent them and hence to reduce the risks to the child's health.

This manuscript introduces and studies a continuous measure of the heritability, called heritability curve, which combines the classical definition the heritability coefficient via biometrial equations and the correlation curve, a local measure of the correlation, first defined in Bjerve and Doksum [1993].

Chapter 2

Notation

2.1 Classical biometrical models

Both genome and environment play a role in the value that a non-Mendelian trait assumes. We can further divide genetic component and environmental component depending on how they interact with each other and the level of randomness of this effect. While a more refined separation can be made, in this manuscript we consider the following four latent variables: additive genetic component (A), dominant genetic component (D), shared environmental component (C), and residual (or random) environmental component (E).

Given a measurement $\boldsymbol{Y}$ of a non-Mendelian trait for a family, for each jth family member we write

$$Y_j = \mu + A_j + C_j + D_j + E_j, \tag{2.1}$$

where μ is the overall mean and A_j, C_j, D_j, and E_j are mutually independent with mean 0 and variances σ_A^2, σ_C^2, σ_D^2, and σ_E^2 respectively. The total variance is then $\sigma^2 = \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2$.

The heritability coefficient (in a narrow sense) is then defined as

$$a^2 = \frac{\sigma_A^2}{\sigma^2},$$

that is, the proportion of the variance which is ascribed to the additive genetic effect. The heritability coefficient (in a broad sense) is instead defined as

$$a^2 + d^2 = \frac{\sigma_A^2}{\sigma^2} + \frac{\sigma_D^2}{\sigma^2},$$

that is, the proportion of the variance which is ascribed to both genetic components. It is classically less studied than a^2. From now on when we talk about heritability we mean a^2, the heritability in the narrow sense.

As mentioned above, A is a latent variable, with an unknown variance. In order to estimate the heritability coefficient, we use the quantities that we can measure, that is the correlation between family members of the trait in question. Then, making use of theoretical equations which describe the proportions of shared environment and shared genetic material among family members, we express the heritability coefficient as a linear combination of Pearson correlations within the family.

These equations depend on the family structure, that is, which family members we are taking measurements of. The two family structures that are mentioned in this manuscript are twin pairs (with a distinction between monozygotic and dizygotic pairs) and family trios made of father, mother, and child. Other popular family structures in biometrical models include non-twins siblings, cousins, and the parents' siblings in the study. Overall, "closer" family members (such as a parent and a child or a pair of siblings) share a larger portion of genetic material and are generally affected from the environment in a similar way. As such, these smaller, closer family structures can provide a good insight in the heritability of a trait. Moreover, measurements of small families are easier to collect.

Outside the nuclear family, the proportions of shared genetic material between family members become small enough to be difficult to properly analyze. For example, a parent and a child share, on average, half of the genetic material. When looking at a grandparent and their grandchild, this proportion is already reduced to one fourth and can be difficult to capture. This is not to say, however, that larger family structures should not be analyzed: as we will highlight later in this section, a model with more family members allows for a more refined and detailed subdivision of the latent variables describing environmental and genetic effects. Our choice to study only twin data and mother-father-child trios was uniquely determined by the data we had access to, and the research described in this manuscript can be generalized and refined by analysing a dataset with information about larger family structures.

Biometrical models allow us to express the heritability coefficient as a linear combination of the Pearson correlations of pairs within a family. To define the proper formula for the heritability, we make use of path analysis, which models latent and visible variables and their interactions in a diagram. As an illustrative example, Figure 2.1 shows the path model of twin data.

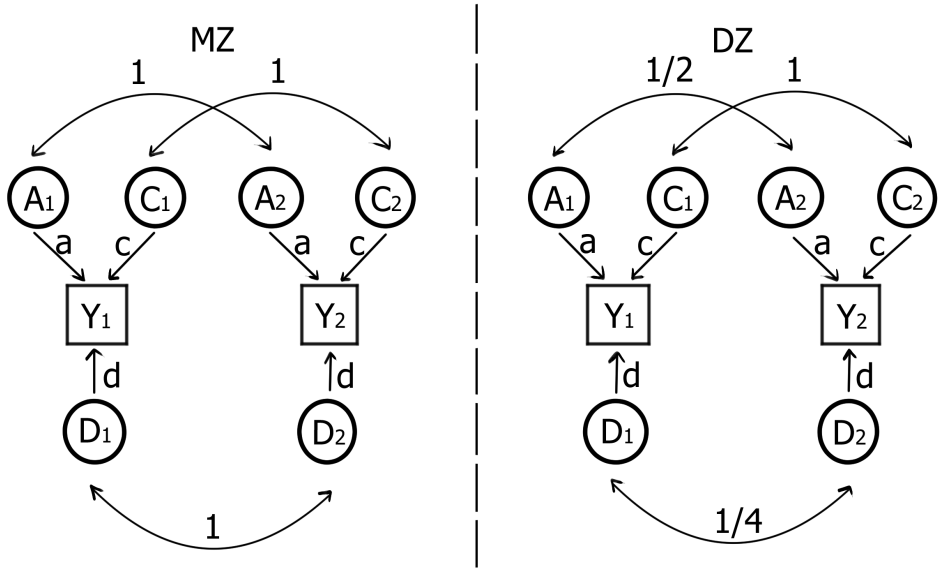


Figure 2.1: Path analysis for a family structure composed of monozygotic twins (on the left) and dizygotic twins (on the right). Y_j represents the measurement of a trait, while A_j, C_j, and D_j are the latent variables as defined in Section 2.1. The variable E is hidden in the figure, since there are no connections between it and the other latent variables.

In the path model, we use the notation from equation (2.1): Y is the trait measurement - that is, the measurable variable - while A, C, and D are the latent variables. The variable E is usually not represented in path models, since the random environmental effect has no significant relationship with the other latent variables. While the latent variables A, C, and D can take different values for the family members (denoted with the indexes 1 and 2 in Figure 2.1), their effect over the phenotype should be consistent over all family members (so, for both twins, in both monozygotic and dizygotic twins). This is represented by the arrow connecting each latent variable to the trait value, which is labeled with the same symbol for all members of the family (in Figure 2.1, for example, for each twin the arrow labeled "a" connects the value of the latent variable A_i to the trait measurement Y_i). The labels of these arrows are not casual; "a" is effectively the square root of the heritability coefficient, and we will see later in this section how to extract a formula for a^2 from this path model.

The arrows that determine the path model are those connecting latent variables with each other. These paths link the family members together, allowing us to extrapolate the equations which define the proportions of the variance dependent on the different latent variables. Let us consider the left side of Figure 2.1, which shows the path model

for a pair of monozygotic twins. Monozygotic twins develop from one single zygote, which is then split into two embrios. Because of that, monozygotic twins share the totality of the genetic material. In our model, the genetic component is divided into two latent variables, A and D, both of which are identical within a twin pair. In the path model we visualize it by connecting A_1 and A_2 with an arrow labeled with the value 1, and by doing the same to the variables D_1 and D_2.

When it comes to the shared environmental effect when studying birth weight, it makes sense to assume that monozygotic twins share the totality of the environment. Any noise and random difference between the two twins is captured by the variable E, the random environment.

When we look at dizygotic twins, on the right side of Figure 2.1, we expect them to share the totality of the shared environment as well. On the other hand, dizygotic twins are born from two separate fecundated eggs, and as such they do not share the totality of the genetic material. The standard assumption is that they share about half of the additive genetic material, and one fourth of the dominant genetic material.

This difference in shared genetic material between monozygotic and dizygotic twins is the key to creating our model. We can measure the correlation of the trait for monozygotic twins and for dizygotic twins (that we will denote as monozygotic correlation and dizygotic correlation, from now on) and any non-random difference between the two can only be attributed to a genetic effect, since the environment affects monozygotic and dizygotic twins equally.

To quantify these proportions, we make use of the path model again. We can see the monozygotic correlation as a path, made of arrows, that connects the values Y_1 and Y_2 in the left side of the diagram. We use the path that connects the variables in the diagram to express the correlation in terms of the proportion of the variance of the latent variables. All of the arrows in the path model are oriented: they can either point towards a direction (e.g. A to Y) or be bidirectional (e.g. A_1 to A_2). When generating a path between two variables, it is always allowed to follow an arrow along its orientation. It is also allowed to follow an arrow (or a sequence of arrows) against their orientation, until the orientation changes. Moving against the orientation, then along it and then against it again is not allowed (see Wright [1921]). For example, we might wish to find the connection between the variables Y_1 and Y_2. We can follow the arrow labeled "a" against its orientation, then move from A_1 to A_2 along the bidirectional arrow, and lastly follow the other arrow labeled "a" along its orientation to reach Y_2. The rule regarding the orientation of the arrows prevents getting stuck in a loop; for example, we cannot start from Y_1, move to Y_2 through the A variables, and then come back to Y_1 through

the D variables because we would follow the arrows against its orientation, then along it, and then against it again.

To describe the path that joins Y_1 and Y_2 through the additive genetic component we multiply the labels of the arrows to obtain $a \times 1 \times a = a^2$. That is not the only allowed path: passing through the shared environment gives a contribution of $c \times 1 \times c = c^2$, and passing through the dominant genetic component adds a contribution of $d \times 1 \times d = d^2$. To calculate the total contribution, we sum the three effects and obtain

$$\rho^{(MZ)} = a^2 + c^2 + d^2. \tag{2.2}$$

Applying the same rules to the dizygotic plot, we obtain the equation

$$\rho^{(DZ)} = \frac{1}{2}a^2 + c^2 + \frac{1}{4}d^2. \tag{2.3}$$

We have now found two equations which link a measurable quantity (the correlation coefficient) with a linear combination of our unknowns (the proportions of variance a^2, c^2, and d^2). The goal is to derive from this system of equations a formula for the proportions of variance, expressed in terms of the correlation coefficients.

Finding a unique solution for this system is currently impossible, because we have more unknowns than equations. In order to derive a unique expression for each of the four proportions of variance we require four distinct, non-trivial equations that relate them to each other and to some measurable quantity - in our case, the correlation coefficients. In our situation, we have three such equations: the two correlation equations (2.2) and (2.3), and the equality

$$a^2 + c^2 + d^2 + e^2 = 1,$$

which is always true due to the definition of the quantities a^2, c^2, d^2, and e^2 as proportions of the total variance. If we solve this three-equations system, we can express the proportions of the total variance as follows:

$$\begin{aligned} a^2 &= 2\left(\rho^{(MZ)} - \rho^{(DZ)}\right) + \frac{3}{2}d^2, \\ c^2 &= 2\rho^{(DZ)} - \rho^{(MZ)} - \frac{1}{2}d^2, \\ e^2 &= 1 - \rho^{(MZ)} - 2d^2. \\ d^2 &= d^2 \end{aligned} \tag{2.4}$$

We express a^2, c^2, and e^2 as functions of d^2. The system has infinite solutions, one for each value of d^2 between zero and one. Notice that we chose to keep d^2 as free parameter, but any other parameter could be chosen in its stead and the result would be analogous.

To have a system of equations that can be used for practical applications, given a dataset comprised of only monozygotic and dizygotic twin pairs, we are forced to simplify the model. It is usually done by assuming that one of the proportions of the total variance is zero. In literature, the two most-used models for twin data are the ACE and the ADE model, which assume no dominant genetic effect and no shared environmental effect, respectively. The formulas for the ACE and ADE models can be derived either from the path model of Figure 2.1, by pretending that either D or C is not present; or by substituting in Equation (2.4) d^2 with 0 or c^2 with 0, respectively. The resulting expressions for the non-zero proportions of variance are listed below, in Equation (2.5) and (2.6), respectively.

$$\begin{aligned} a^2 &= 2(\rho^{(MZ)} - \rho^{(DZ)}), \\ c^2 &= 2\rho^{(DZ)} - \rho^{(MZ)}, \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{2.5}$$

$$\begin{aligned} a^2 &= 4\rho^{(DZ)} - \rho^{(MZ)}, \\ d^2 &= 2(\rho^{(MZ)} - 2\rho^{(DZ)}), \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{2.6}$$

Equation (2.5) is also referred as Falconer's equation (see Falconer and Mackay [1983]). These equations are widely used, but they rely on the Pearson correlation coefficient to be computed.

The correlation coefficient measures the intensity of the linear dependence between two variables. As such, it has a statistical meaning only when the variables are linearly related, and is a misleading measure otherwise. The data that we analyze in this manuscript do not satisfy said condition: indeed, they have what is often called a "pear shape" (see Figure 2.2a), which does not satisfy the homoskedasticity condition. In the next section we introduce a tool that allows us to overcome this issue and is the fundamental stepping stone for this manuscript.

2.2 Correlation curve and heritability curve

The correlation coefficient is a very concise and elegant way to describe the dependence between two variables, hence statisticians have defined several equivalent mathematical concepts which can be applied to non-linearly correlated variables. Spearman's ρ, for

example, is a rank correlation coefficient - that is, it measures the statistical dependence of the rank values of the two variables. In practice, it evaluates how suitable a monotonic function is to describe the relationship between the two variables (see e.g. Dodge [2008]). In the field of information theory, the mutual information (MI) is used as measure of dependence between two variables (see MacKay et al. [2003]). In particular, it is used to determine the amount of information about the first variable that is gained by only looking at the second variable.

In Bjerve and Doksum [1993], the authors propose a non parametric measure of correlation for non-linear variables, called correlation curve. The correlation curve is constructed by generalizing the classical definition of the correlation coefficient. Let us assume that the pair (Y_1, Y_2) is distributed following a bivariate normal distribution $N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Then, we define the correlation coefficient as

$$\rho = \frac{\sigma_1 \beta_{2|1}}{\sigma_2}, \tag{2.7}$$

where $\beta_{2|1}$ is the slope of the regression of Y_2 over Y_1. The variance of Y_2 can be rewritten as expression of the variance of Y_1, the regression slope, and the conditional variance of Y_2 over Y_1:

$$\sigma_2^2 = \left(\sigma_1 \beta_{2|1}\right)^2 + \sigma_{2|1}^2,$$

where $\sigma_{2|1}^2 = \sigma_2^2 \left(1 - \rho^2\right)$. Equation (2.7) can be rewritten as

$$\rho = \frac{\sigma_1 \beta_{2|1}}{[(\sigma_1 \beta_{2|1})^2 + \sigma_{2|1}^2]^{1/2}}. \tag{2.8}$$

This formula is true under the normality assumption, but the goal of Bjerve and Doksum [1993] is to find a general expression that can be applicable to any pair of variables (Y_1, Y_2), independently of their distribution.

For this purpose, the correlation curve was defined as follows:

$$\rho(y) = \frac{\sigma_1 \beta_{2|1}(y)}{[(\sigma_1 \beta_{2|1}(y))^2 + \sigma_{2|1}^2(y)]^{1/2}} \tag{2.9}$$

where the slope of the regression line is substituted by the derivative of the conditional mean of Y_2 over Y_1 at $Y_1 = y$ ($\beta(y) = \mu'(y)$, with $\mu(y) = \mathrm{E}(Y_2|Y_1 = y)$) and the residual variance is computed conditionally on the same value $Y_1 = y$.

As mentioned in the previous section, our goal is to find a replacement measurement for the Pearson correlation in Equations (2.5) and (2.6) that can be applied whenever the data does not satisfy the linearity assumption. To achieve this goal, we can estimate the

correlation curve for each pair of family members in the family structure that we analyze. In the twin data example, we estimate the monozygotic and dizygotic correlation curves (denoted with $\rho^{(MZ)}(y)$ and $\rho^{(DZ)}(y)$ respectively). In the mother-father-child trios example (which is explored in ARTICLE I), on the other hand, we estimate mother-father, mother-child, and father-child correlation curves.

When estimating any correlation curve, we are making a choice regarding which variable we are conditioning on. In some contexts, this choice can be quite straightforward. When applying this concept to family data, instead, this choice is not at all obvious: one would expect a level of symmetry in the measure of the correlation curve, especially when talking about twin data. For this reason, we impose the two following exchangeability conditions:

$$\begin{aligned} \text{Var}(Y_2|Y_1 = y) &= \text{Var}(Y_1|Y_2 = y), \\ \text{E}(Y_2|Y_1 = y) &= \text{E}(Y_1|Y_2 = y). \end{aligned} \tag{2.10}$$

We will expand more on these two conditions in the next section, when we compute the correlation curve for data following a multivariate Gaussian mixture distribution.

Under these conditions, we can drop the indices from $\sigma^2_{2|1}(y)$ and $\beta_{2|1}(y)$ in the correlation curve. We further use the symbol τ_1 to denote the marginal standard deviation, to avoid confusing it with the conditional variance. The formula becomes

$$\rho(y) = \frac{\tau_1 \beta(y)}{[(\tau_1 \beta(y))^2 + \sigma^2(y)]^{1/2}} \tag{2.11}$$

We can now replace the Pearson correlations in Equations (2.5) and (2.6) with the corresponding correlation curves. The result is a system of equations whose solutions are functions of the trait value y.

$$\begin{aligned} a^2(y) &= 2(\rho^{(MZ)}(y) - \rho^{(DZ)}(y)), \\ c^2(y) &= 2\rho^{(DZ)}(y) - \rho^{(MZ)}(y), \\ e^2(y) &= 1 - \rho^{(MZ)(y)}. \end{aligned} \tag{2.12}$$

$$\begin{aligned} a^2(y) &= 4\rho^{(DZ)}(y) - \rho^{(MZ)}(y), \\ d^2(y) &= 2(\rho^{(MZ)}(y) - 2\rho^{(DZ)}(y)), \\ e^2(y) &= 1 - \rho^{(MZ)}(y). \end{aligned} \tag{2.13}$$

The curve defined as $a^2(y)$ is what we call the heritability curve, a local measure of the heritability of a trait. The functions $c^2(y)$, $d^2(y)$, and $e^2(y)$ are local measures of shared environment, dominant genetic effect, and random environment, respectively. These newly defined curves are linear combinations of the correlation curves; as such, the biggest challenge remains calculating the correlation curves themselves.

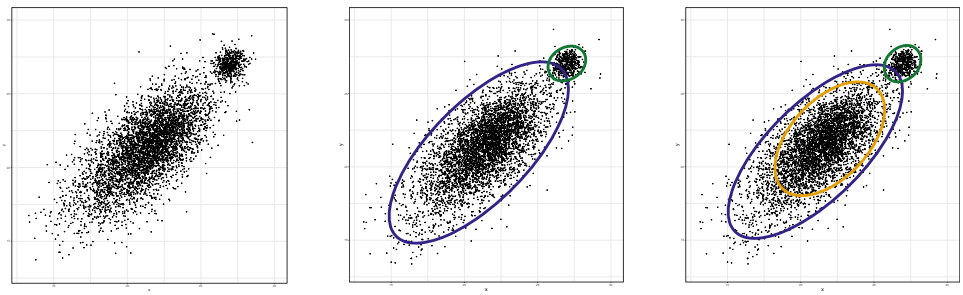
In practice, to calculate a correlation curve we require the following steps:

1. choose a distribution that well describes the dataset;
2. estimate the distributions's parameters;
3. if needed, compute the marginals for each pair of family members;
4. calculate conditional mean and conditional variance for each such pair;
5. calculate an explicit formula for the derivative of each conditional mean.

In this manuscript the goal is to find an explicit formula of the correlation curves, without resorting to a numerical approximation. Admitting approximations would simplify some of these steps, especially when calculating the derivatives of the conditional means, and it is a promising alternative path to the one we pursued.

The task at hand can be more or less complicated, depending greatly on the distribution that is chosen to fit the data. Clearly, a Gaussian distribution would return the easiest solution: one can easily see that, if we assume that the distribution is normal, the correlation curve is simplified back to the correlation coefficient (a proof of this statement is provided in Article I). The purpose of this manuscript, however, is to work with non-normal data and to apply the concept of a continuous, non-constant heritability to said dataset.

For this purpose, we directed our interest towards a distribution that maintains some useful properties of the Gaussian distribution, but guarantees more flexibility: a Gaussian mixture.



(a) Example of pear-shaped data, with two visible distinct clusters

(b) Two Gaussian components fitted on pear-shaped data

(c) Three Gaussian components fitted on pear-shaped data

Figure 2.2: Simulated data with a "pear shape" and two Gaussian mixture fitting, respectively with 2 and 3 components.

2.3 Gaussian mixtures

A mixture distribution is a weighted sum of multiple distributions, each with its set of parameters. To guarantee that a mixture is indeed a distribution, the weights must be all positive and sum up to one. Each distribution in a mixture is called component (or kernel). A mixture with m components, then, is a weighted sum of m distributions.

A Gaussian mixture distribution is a mixture distribution whose components are normal distributions (each with its distinct means, standard deviations and correlation coefficients). In the most general sense, the density of a multivariate Gaussian mixture distribution has the form

$$\sum_{k=1}^{m} p_k N_n(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where N_n describes the density of an n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_k$. Figure 2.3 is an example of a one-dimensional Gaussian mixture. In this manuscript we work with two-dimensional data (twin pairs) and three-dimensional data (mother-father-child trios), so we restrict our interest to $n = 2$ and 3.

Gaussian mixtures fit rather well on the datasets studied in this manuscript. As mentioned in Subection 2.1, BMI and birthweight data often have a "pear shape" similar to Figure 2.2a. We can imagine this "pear shape" as two (or more) separate clusters of data, each distributed following a normal multivariate distribution ((see Figure 2.2b and 2.2c). A Gaussian mixture is then able to capture these separate clusters into a single distribution model.

When computing the correlation curve, picking a Gaussian mixture distribution as underlying distribution is a very convenient choice. The conditional mean and conditional variance of a Gaussian mixture can easily be calculated from their definition, due to the properties of a mixture distribution (detailed calculations can be found in Article I).

Finding an exact expression for the densities of the marginal distributions can still be a challenge; for this reason we impose some restrictions on the parameters of each Gaussian component. In particular we require that, within each component, all family members share the same mean and same standard deviation. This means that, for each

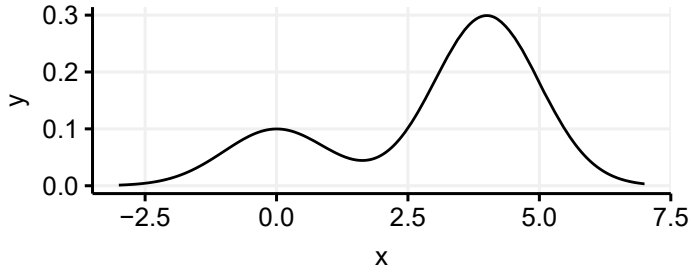


Figure 2.3: Density curve of a one-dimensional Gaussian mixture with two components, with mean values $\mu_1 = 0,\ \mu_2 = 4$, standard deviations $\sigma_1 = \sigma_2 = 1$, and weights $p_1 = 0.25,\ p_2 = 0.75$.

$k \in 1, ..., m$ (in the 2-dimensional example),

$$\boldsymbol{\mu}_k = (\mu_k, \mu_k)$$
$$\boldsymbol{\Sigma}_k = \sigma_k^2 \begin{pmatrix} 1 & \rho_k^{(1,2)} \\ \rho_k^{(2,1)} & 1 \end{pmatrix}$$

This is not an unreasonable assumption: the pear shape of the data already suggests a tendency of centering the means of the components along the diagonal $Y_1 = Y_2$, which corresponds to the condition on the mean vector. Similar remarks can be made about the standard deviations.

Article I shows how to derive the marginal densities when imposing these restrictions.

2.4 Gaussian mixtures and multimodality

Gaussian mixture distributions are often multimodal, with a mode corresponding to each component (see, for example, Figure 2.3). A certain level of caution is required when estimating the parameters of a mixture distribution.

To describe some issues that arise from multimodality, we use the Gaussian mixture from Figure 2.3 as example. The mixture has two components, that we label in ascending order according to the mean (that is, we assume $\mu_1 < \mu_2$). This choice is, however, completely arbitrary; if we label the component with the lowest mean as "second component" and the one with the highest mean as "first component" we are describing the

exact same distribution.

This is a common issue with mixture distributions: estimation algorithms often identify two sets of parameter estimates that are identical, except that the labels of the components are switched, and treat them as separate solutions. This makes the optimization more difficult, because multiple solutions are the optimal one. For this reason, we introduce a parametrization which imposes an order on the components, *de facto* collapsing all the different equivalent optimizations into a single one. We define the parameters α through the following relations:

$$\begin{aligned}\mu[1] &= \exp(\alpha[1]) \\ \mu[i] &= \mu[i-1] + \exp(\alpha[i]) \text{ for every } i = 2, ..., m\end{aligned}$$

The means of the components are now ordered from smallest to largest.

The estimation model that we use in this manuscript is a maximum likelihood method. The most popular algorithm for optimizing mixture distributions, called the EM algorithm, is also a maximum likelihood model, and is proven to converge (albeit slowly) to a maximum of the likelihood function (see Wu [1983]). It cannot, however, distinguish between local and global maxima. This means that, even when the algorithm converges, we are not certain to have found the optimal parameters of the mixture distribution. This is an issue that we encounter with our model as well.

Both the EM algorithm and the algorithm that we present in this manuscript make use of initial values to begin the exploration of the parameter space, and fine tuning the choice of said initial values is a way to reduce the probability of incurring in local maxima, instead of the global one. This subject is discussed at length in Article III, which proposes an HMC algorithm to explore the space of initial values.

Chapter 3

The algorithm

The common thread in the three articles presented in this thesis is the algorithm that we wrote to optimize the parameters of the Gaussian mixture which fits the data. The algorithm is written in the language R [R Core Team, 2020], and interpolates a file written in the language C++ [Stroustrup, 1995] to calculate the derivatives of the likelihood functions, which are then used to improve the performance of the optimization functions in R.

In this chapter we present an introduction to automatic differentiation, which allows to calculate derivatives in an efficient way. We then present the package `TMB` [Kristensen et al., 2016], which allows R to import and read a C++ file.

3.1 Automatic differentiation

Having access to differential information about the objective function (usually first and second order derivatives) can strongly improve the quality of the optimization process. When computing derivatives through an algorithm there are two factors that must be kept in mind: accuracy and efficiency. Being able to calculate the exact derivative of a function is *de facto* useless, if the computation time is so long that the algorithm becomes unusable. On the other hand, an approximation can be evaluated quickly, but the results may not be as precise.

The opposite sides of this spectrum are represented by symbolic derivation and numerical derivation, respectively. Symbolic derivation is used in the field of computer algebra, and works uniquely on symbolic mathematical expressions. To compute a derivative this approach uses well-known rules for derivation (such as the product rule and the chain

rule) and obtains an exact formula of the derivative.

Numeric differentiation, in contrast, approximates the value of the derivative of a function at a specific point. It does so starting from the formula

$$\frac{f(x+h) - f(x)}{h}$$

which, for a small enough value of h, approximates the slope of the tangent of the function f at a point x - that is, it approximates the value $f'(x)$.

Automatic differentiation, shortened to AD (see, e.g. Neidinger [2010]) was developed as a computational tool for estimating derivatives in opposition to numerical approximation and symbolic differentiation. It makes use of the chain rule and Laplace approximation to create an algorithm whose output is a precise and efficient measure of derivatives and partial derivatives up to the third order.

To explain how AD works, we use a classical example: the function

$$y = f(x_1, x_2) = x_1 x_2 + \sin(x_1).$$

We express $f(x_1, x_2)$ as a chain of simpler functions, that we denote with w_j. Together, they form an evaluation trace, that starts with the inputs x_1 and x_2 and outputs y.

$$\begin{aligned} w_1 &= x_1 \\ w_2 &= x_2 \\ w_3 &= w_1 w_2 \\ w_4 &= \sin(w_1) \\ w_5 &= w_3 + w_4 \end{aligned} \tag{3.1}$$

To better visualize the logical dependencies of the functions w_j, we represent the evaluation trace in the form of a graph (Figure 3.1). For example, the function w_4 depends only on w_1, while w_5 is a function of both w_4 and w_3. This decomposition of $f(x_1, x_2)$ is useful for computing the value of a function at a point (x_1, x_2) by identifying the terms that repeat themselves, especially in more complex functions; more importantly, it highlights the dependencies that will quicken significantly the computations of the derivatives.

There are two main approaches to AD: forward accumulation and reverse accumulation

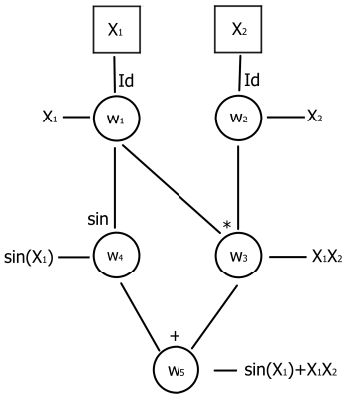


Figure 3.1: Graph of the evaluation trace of the function $f(x_1, x_2) = x_1 x_2 + \sin(x_1)$.

(see e.g. Linnainmaa [1976]). The first one consists in computing the derivative of the terms together with the values of the function itself, in one single (forward) sweep. The second method, instead, consists in performing first a "forward sweep" that computes the value of the function and in computing the derivative in a second sweep, from the end backwards. In mathematical terms, this means that forward accumulation will estimate the partial derivatives $\frac{\partial w_i}{\partial x_j}$ starting from $i = 1$, while the reverse accumulation will estimate the derivatives $\frac{dy}{dw_i}$ starting from $i = 5$. We present the two methods applied to the example function.

For forward accumulation, we require to compute the partial derivatives of each w_j with respect to both x_1 and x_2. For this example, we only show the partial derivatives with respect to x_1, with the computations with respect to x_2 being analogous.

$$\begin{aligned}
\frac{\partial w_1}{\partial x_1} &= \frac{\partial x_1}{\partial x_1} = 1 \\
\frac{\partial w_2}{\partial x_1} &= \frac{\partial x_2}{\partial x_1} = 0 \\
\frac{\partial w_3}{\partial x_1} &= \frac{\partial w_1 w_2}{\partial x_1} = \frac{\partial w_1}{\partial x_1} w_2 + \frac{\partial w_2}{\partial x_1} w_1 = w_2 \\
\frac{\partial w_4}{\partial x_1} &= \frac{\partial \sin(w_1)}{\partial x_1} = \cos(w_1) \frac{\partial w_1}{\partial x_1} = \cos(w_1) \\
\frac{\partial w_5}{\partial x_1} &= \frac{\partial (w_3 + w_4)}{\partial x_1} = \frac{\partial w_3}{\partial x_1} + \frac{\partial w_4}{\partial x_1} = w_2 + \cos(w_1)
\end{aligned} \tag{3.2}$$

As it is shown in Equations 3.2, the decomposition of the function $f(x_1, x_2)$ into the simpler functions w_j has a direct advantage: the single partial derivatives are easy

to compute, and by virtue of the definitions of the functions w_j we can build from previous computations in the chain to greatly simplify those of the latter derivatives. For example, the term $\frac{\partial w_1}{\partial x_1}$ is the first derivative that is computed, and storing it simplifies the computations of the terms $\frac{\partial w_3}{\partial x_1}$ and $\frac{\partial w_4}{\partial x_1}$. This is only amplified for more complex functions.

We now show an example of reverse accumulation. First, we perform a "forward sweep" to estimate the values of all w_i's following their definition in Equation 3.1; they will be used later. Then, to estimate the derivative of the function $f(x_1, x_2)$, we compute the derivatives $\frac{dy}{dw_i}$ (also called adjoints) for every i starting from $i = 5$. At each step we move backward through the graph, and we use the chain rule to express the adjoint $\frac{dy}{dw_i}$ in terms of the adjoints $\frac{dy}{dw_j}$, for $j > i$, and the derivatives of the functions w_i's.

$$
\begin{aligned}
\frac{dy}{dw_5} &= \frac{dy}{dy} = 1 \\
\frac{dy}{dw_4} &= \frac{dy}{dw_5}\frac{dw_5}{dw_4} = 1\frac{d(w_3 + w_4)}{dw_4} = 1 \\
\frac{dy}{dw_3} &= \frac{dy}{dw_5}\frac{dw_5}{dw_3} = 1\frac{d(w_3 + w_4)}{dw_3} = 1 \\
\frac{dy}{dw_2} &= \frac{dy}{dw_3}\frac{dw_3}{dw_2} = 1\frac{d(w_1 w_2)}{w_2} = w_1 \\
\frac{dy}{dw_1} &= \frac{dy}{dw_4}\frac{dw_4}{dw_1} + \frac{dy}{dw_3}\frac{dw_3}{dw_1} = 1\frac{d\left(\sin(w_1)\right)}{dw_1} + 1\frac{d\left(w_1 w_2\right)}{dw_1} = \cos(w_1) + w_2
\end{aligned}
$$

Recall that $w_1 = x_1$, $w_2 = x_2$, and $y = f(x_1, x_2)$. The last two steps of the algorithm, then, express the partial derivatives $\frac{\partial f(x_1,x_2)}{\partial x_1}$ and $\frac{\partial f(x_1,x_2)}{\partial x_2}$ in terms of the simple functions w_i's, which we estimated in the "forward sweep".

Notice that in the forward accumulation example we have not shown the entire process: in order to have the complete algorithm, we would need to compute the partial derivatives of each w_j with respect to x_2 as well. This extra step is instead not required in the reverse accumulation, as explained above. Extra steps are required for reverse accumulation if the function $f(\boldsymbol{x})$ has multiple outputs: for example, assume that our function is defined such that $f : (\boldsymbol{x}) \mapsto (y_1, y_2)$. In that case, we would need to compute both $\frac{dy_1}{dw_j}$ and $\frac{dy_2}{dw_j}$ for each j. Overall, given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, forward accumulation is preferable when $m \gg n$, while reverse accumulation is preferable when $n \gg m$. Whenever the function is scalar, then, reverse accumulation is the fastest approach.

The main issue with reverse accumulation is the need to store the intermediate steps that are required to compute the last derivatives.

3.2 TMB

The language R has inbuilt several optimization functions that can use differential information to improve the quality of the estimations.

The default derivation functions in R use symbolic differentiation. To implement automatic differentiation we use the package `TMB` (Template Model Builder), which allows to use a template in C++ to define the objective function. It uses the C++ packages `CppAD` [Bell, 2017] and `Eigen` [Jacob, 2006] to use automatic differentiation and to perform linear algebra operations, respectively.

In practice, when using `TMB`, we write two separate codes, one using R language and one using C++ language. In the R code, we import a dataset, define the parameters of the objective function, and assign an initial value to each. We then compile the C++ file, which reads in the data and the parameters from the R code and defines the objective function. Using automatic differentiation, C++ computes the gradient and the hessian of the objective function, which can then be used back in the R code, where the optimization phase takes place.

In the Appendix we describe in details a generalized version of the two codes used to optimize the negative log likelihood function of a Gaussian mixture distribution.

Chapter 4

Further research

4.1 EM algorithm

A popular approach when fitting a mixture is to assume that each point in the dataset is generated by one specific mixture component. The problem can be phrased as estimating the parameters of m separate Gaussian distributions, and then associating each point in the dataset to one such distribution. The weights p_k are then derived by calculating the proportion of points associated to each density. In this context, we have two separate sets of unknowns: the parameters describing each component of the mixture (that is $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ when working with Gaussian mixtures), and the latent variable which identifies which component each data point "belongs to". Under this assumption, a method to fit a mixture was developed as an iterative algorithm, made of two steps: the first step assumes the parameters to be known and estimates the latent variable, while the second step keeps the values of the latent variable fixed and estimates the parameters. This is the Estimation-Maximization (EM) algorithm [Dempster et al., 1977] applied to the estimation of Gaussian mixtures. The algorithm is defined for any statistical model that includes latent variables and has applications in a variety of fields. Relevant to our research, the EM algorithm is considered the standard approach for maximum likelihood estimation applied to finite mixture models.

In Dempster et al. [1977] it is proven that each iteration of the EM algorithm increases the log likelihood, guaranteeing a convergence to a maximum which is, although, not necessarily the global maximum of the log likelihood function. This is a common issue among maximum likelihood methods, including the one presented in this manuscript. A longer discussion about this issue and a proposed solution can be found in Article III.

	Parameter			
Component	μ	σ	ρ	p
1	(22, 23)	(1,0.5)	0.7	0.6
2	(22.2, 23.1)	(1,0.5)	0.5	0.3
3	(22.4, 23.2)	(1,0.5)	0.2	0.1

Table 4.1: Parameters generating the simulated dataset used in Section 4.1

While the EM algorithm always converges to a local maximum of the log likelihood, it is notoriously slow. This issue is well known and addressed through several approaches (see e.g. McLachlan and Peel [2000] for a review).

To start the first estimation step, the algorithm requires initial values for the parameters of the mixture. The choice of initial values affects the speed to which the algorithm converges and which maximum is reached. The initialization of the EM algorithm is widely discussed, especially in the context of Gaussian mixtures (for example, in Baudry and Celeux [2015], Biernacki et al. [2003], and Melnykov and Melnykov [2012]). The number of components is often treated as a hyperparameter that is trained using mathematical criteria such as AIC or BIC value.

The goal of this section is to compare the EM algorithm to the TMB approach when estimating Gaussian mixture parameters. We will use a dataset simulated from a multivariate Gaussian mixture with three components to compare the two methods.

An important remark: the code as we describe it in the Appendix contains strong restrictions on the parameters of the mixture that cannot be translated in the EM algorithm. In particular, this holds for the conditions $\mu_i^k = \mu_j^k$ and $\sigma_i^k = \sigma_j^k$ for each component k, and each family member i, j. We introduced these conditions to simplify the calculation of the correlation curve, so they are not necessary in this specific example. To make the comparison between the two models as fair as possible, we rewrote the C++ and R codes to remove these restrictions.

We generate a simulated dataset from a Gaussian mixture with three components; the parameters are listed in Table 4.1. We choose a dataset with components that are close to each other, which are usually more difficult to estimate correctly. We use two criteria to compare the two models: the time the algorithm takes to converge to a solution, and the negative log likelihood value of the estimated solution.

We first run both algorithms choosing as initial values the real parameters defining the generating Gaussian mixture (listed in Table 4.1). The results of this approach are collected in Table 4.2 (upper half). The TMB algorithm is significantly slower than the EM algorithm, but the negative log likelihood it returns is slightly better.

True values

Algorithm	Conv. time	Negative log likelihood
EM	1.18882	193265.3
TMB	69.18763	193262.0

Ten random seeds

Algorithm	Conv. time		Negative log likelihood			
	mean	std. dev.	mean	std. dev.	min	max
EM	25.23881	9.12835	193308.5	31.32351	193279.2	193368.2
TMB	167.90230	18.16970	193262.6	0.57158	193262.0	193263.4

Table 4.2: Convergence time (in seconds) and negative log likelihood for the EM and the TMB algorithm. The upper half of the table shows the results obtained by choosing as initial values the true parameter values. The lower half of the table shows the distribution of the results obtained by repeating the initialization through `shortemcluster` with ten random seeds.

We then used the function `shortemcluster` to generate initial values derived by a short EM run. This function is often used in an EM algorithm to establish initial values which are in a reasonable range for the parameters. The function is affected by a random seed, so we repeated the process for ten different seeds, obtaining ten separate sets of initial values, which we used to initialize both the EM and the TMB algorithm. Lastly, we collected convergence time and negative log likelihood for all ten attempts, and studied their distributions calculating mean and standard deviation. For the negative log likelihood we also recorded minimum and maximum value. The results are collected in Table 4.2 (lower half).

The EM algorithm is faster with random initial values as well, but it does not settle into one single solution, and the effect of the different initial values is seen in the larger standard deviation of the negative log likelihoods. The TMB negative log likelihood, on the other hand, varies very little; moreover, among the ten attempts, the maximum TMB negative log likelihood is lower than the minimum EM negative log likelihood.

Overall, we observe that EM is a faster algorithm, with consistently good results. TMB is slower, but outperforms EM in terms of accuracy, both with accurate initial values and random ones (albeit within a reasonable range from the true parameters).

4.2 An analysis of the number of components

In Article I and Article II, we utilize AIC and BIC values to choose the best number of components, m, for the fitted Gaussian mixture. If AIC and BIC provided conflicting results, we relied on the latter. Primarily, our goal was to select the model with the

	Parameter					
Component	μ_F	μ_M	σ	ρ_{MZ}	ρ_{DZ}	p
1	21.705	23.575	1.94	0.74	0.31	0.70
2	24.495	26.365	2.72	0.34	- 0.19	0.26
3	27.995	29.825	4.67	0.38	-0.22	0.04

Table 4.3: Parameter estimates used to generate a three-component mixture in Section 4.2.

lowest complexity, while still capturing the nuances of the data. The more conservative BIC was the better criterion to reach this goal.

In both articles, however, the difference in BIC values among some models was almost insignificant. An important question then arose: how much would the choice between models - that is, between number of components in a mixture - affect the subsequent analysis, in particular, in relation to the shape of the correlation and heritability curve? In this section we explore this issue, plotting heritability curves for Gaussian mixtures with a variety of components, fitting the same simulated dataset.

We generate a dataset based on the parameters estimated on the twin data of Article II. According to the BIC value, a Gaussian mixture model with three components was preferable, and the subsequent optimization returned the parameter estimates collected in Table 4.3.

We separately fitted the dataset with Gaussian mixtures with 2, 3, 4, 5, and 10 mixture components. We also fitted the dataset with a Gaussian distribution for comparison. We calculated AIC and BIC values for all the models described above. As expected, both criteria agree that the three components mixture fits the dataset best. In Table 4.4 we collected, for all models, the difference between their AIC and BIC values and those of the best fitting model (that is, $m = 3$). We denote these differences with the symbol ΔAIC ad ΔBIC.

The improvement from a simple Gaussian distribution to a mixture is very obvious, especially compared to the relatively smaller improvements between the other models. The difference in AIC values between $m = 3,\ 4$, 5, and 10 is quite small, and while not as trivial, the difference in BIC between $m = 3,\ 4$, and 5 is also relatively small.

To check the solidity of our results, we plot the heritability curves for all the models described in this section (Figure 4.1).

We observe that, regardless of the number of components, all heritability curves follow the same pattern (first decreasing for the majority of the trait range, and then increasing again until around the 0.975 quantile). The obvious exception is the heritability "curve"

m	ΔAIC	ΔBIC
1	2938.5	2866.9
2	238.9	203.1
3	**0**	**0**
4	6.4	42.2
5	5.1	76.7
10	3.9	254.5

Table 4.4: Difference in AIC and BIC values between Gaussian mixtures models with 1, 2, 3, 4, 5, and 10 components, and the best fitting model ($m = 3$).

for $m = 1$, which is the constant heritability coefficient along the entire data range by definition.

As the number of components increases, so does the wiggliness of the heritability curve. It is particularly noticeable in the curve for $m = 10$, which also exhibits a big drop in the left-hand tail (most likely due to high variation and the asymptotic behavior of heritability curves, explained thoroughly in Article I).

This result shows that, if we had chosen a different number of components as our model (for example by using the AIC value instead of the BIC value as selection criterion in Article I), the overall conclusions drawn about the heritability curve and its shape would be unchanged.

4.3 Other non-constant measures of the heritability

4.3.1 Quantile regression

While this manuscript focuses on the heritability curve, its properties and its applications, a big effort has been put into comparing it with existing non-constant measures of the heritability - namely, quantile regression.

Quantile regression (introduced in Koenker and Bassett Jr [1978]) is a regression model which estimates the conditional quantiles of the response variable, instead of its conditional mean (the approach of classical linear regression). Quantile regression is a powerful tool, since it requires no condition on the distribution of the dependent variable and, relevantly to our research, outputs a dynamic, non-constant regression curve. It has recently been used to measure the heritability for BMI in Williams [2020].

Since quantile regression is a valid alternative to solving the problem we wanted to

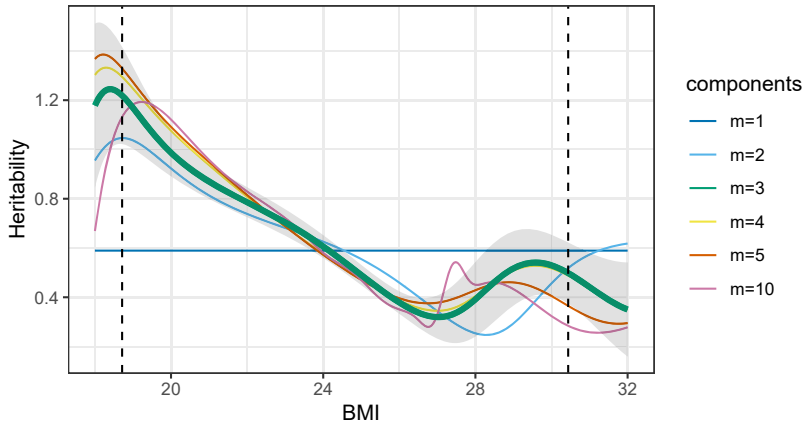


Figure 4.1: Heritability curves of female simulated data, generated assuming an underlying Gaussian mixture distribution. The hyperparameter m varies from one (that is, assuming an underlying Gaussian distribution) to five. The gray area is a 95% confidence interval for the heritability curve for $m = 3$ components, the best model according to AIC and BIC value. 95% of the data is contained between the two vertical lines.

tackle, we decided to inspect it and compare it to the heritability curve in Article II, using twin data. Section 7 in the Supplementary Material of Article II goes into details on the similarities and the differences between the results of the two methods, concluding that this topic should be explored more in depth.

4.3.2 Non-parametric models

As mentioned in Section 2.2, our goal when approaching the issue of a non-constant heritability was to define a fully parametric method, with an explicit formula for both correlation curve and heritability curve. This approach is very interesting from a mathematical and theoretical point of view, but can be very difficult to apply to real data. In order to generate a heritability curve on our data, for example, we have made strong assumptions about the Gaussian mixture underlying the data, as described in Section 2.3. While these assumptions are reasonable of the data we inspect in Article I and Article II, they restrict significantly the pool of data that we can calculate the heritability curve of.

For this reason, in Article II we explored a non-parametric version of the heritability curve. In particular, this version does not compute the symbolic derivative of the conditional mean, and uses numerical methods instead. A longer discussion of this method can be found in Section 6 of the Supplementary Material of Article II.

Chapter 5

The Articles

Article I, Heritability curves: a local measure of heritability in family models is a mostly theoretical article which introduces the heritability curve and its properties. It describes the framework used in this manuscript, studies the asymptotic behavior of the correlation curve when applied on Gaussian mixture, and presents two examples: one with measurements of BMI in pairs of twins, and one with measurements of birth weight in mother-father-child family trios.

Article II, The heritability of BMI varies across the range of BMI — a heritability curve analysis in a twin cohort shows the application of the heritability curve on a large twin cohort dataset and draws conclusions about the heritability of BMI. In the Supplementary Material we further explore the method, by comparing it to another non-constant measure of heritability, quantile regression, and by defining a non-parametric version of the curve.

Article III, Exploring the likelihood surface in multivariate Gaussian mixtures using Hamiltonian Monte Carlo, is a theoretical article which applies Hamiltonian Monte Carlo methods to initialize the TMB algorithm, with the goal to optimize the parameter estimates. The article highlights the struggles in applying Hamiltonian Monte Carlo methods to multivariate Gaussian mixtures and offers three approaches to increase the chance of finding realistic parameter estimates.

Appendix: The code

We show the code for a dataset of monozygotic and dizygotic twin pairs (the type of data illustrated in both ARTICLE I and ARTICLE II). For simplicity, we assume that all twin pairs are male. We later give a brief explanation on how we deal with sex differences in the data.

We first explain the R code, keeping in mind that the flow of the algorithm moves between R and C++ as described in Subsection 3.2.

The number of components of the Gaussian mixture we fit is a hyperparameter, that must be decided in advance. We provide the code for a generic, fixed value of m.

R code

As mentioned in Subsection 2.4, to initialize the algorithm we require to input an initial value for each variable of the objective function. Following the restrictions and reparametrizations described in Subsection 2.3 and Subsection 2.4 the parameters are: m reparametrized means α's; m standard deviations σ's; m monozygotic correlation coefficients $\rho^{(MZ)}$'s; m dizygotic correlation coefficients $\rho^{(DZ)}$'s; $m-1$ weights δ's, with the last weight defined as $1 - \sum_{i=0}^{m-1} \delta_i$. To impose the last condition we first define a vector of length m and then apply the function `delta.n2w` (defined below), which returns a vector of length $m - 1$.

To provide a general example, we do not choose the numerical initial values for the parameters, using a mock vector in its stead (for example, the vector of length m `(alpha.1,...,alpha.m)` as initialization for the α's). When applied to a real data, these mock vectors should be replaced with some appropriate initial values (see ARTICLE III for a longer discussion on the choice of initial values).

```
initial_alpha <- (alpha.1,...,alpha.m)
initial_sigma <- (sigma.1,...,sigma.m)
initial_rhoMZ <- (rhoMZ.1,...,rhoMZ.m)
initial_rhoDZ <- (rhoDZ.1,...,rhoDZ.m)
initial_delta <- (delta.1,...,delta.m)

delta.n2w <- function(m, delta){
foo <- log(delta/delta[1])
tdelta <- as.vector(tail(foo, m - 1))
return(tdelta)
}
```

We then collect all the initial values into a list, called `parameters`. Instead of optimizing the standard deviations, we first apply the logarithm; in this way we guarantee a positive value of the standard deviations (which are calculated later as the exponential of $\log(\sigma)$'s, the parameters which are optimized).

In a separate list, `dat`, we then collect all the known information, that is the data in matrix form (divided between monozygotic and dizygotic twin pairs for convenience) and the known number of components of the mixture distribution. We also generate a vector of points that we will use later to estimate correlation curves and heritability curve. The vector should collect equidistant points into the medium range of values that the trait we are studying can take.

```
data_mat_MZ <- data.matrix(Data_MZ)
data_mat_DZ <- data.matrix(Data_DZ)
dat <-list(data_mat_MZ = data_mat_MZ, data_mat_DZ = data_mat_DZ,
                 m = m, y_points = y_points)
parameters <- list(alpha = initial_alpha,
                   log_sigma  = log(initial_sigma),
                   rhoMZ = initial_rhoMZ, rhoDZ = initial_rhoDZ,
                   tdelta = delta.n2w(m, initial_delta))
```

The functions that allow us to read and compile a C++ code are `compile` and `dyn.load`. The first function runs the C++ code and generates a .DLL file - that is, a dynamically loaded library. The .DLL file is loaded (through the function `dyn.load`) and permits to execute the functions that are defined in the C++ file (namely, as we will see in the next section, the negative log likelihood).

```
library(TMB)

compile("Cpp_file.cpp")
dyn.load(dynlib("Cpp_file"))
```

The function `MakeADFun`, from the package `TMB`, reads the .DLL file and generates the functions required to calculate the objective function defined in the C++ file and its derivatives. To generate these functions it also requires the data in the list `dat` and the initial values for the parameters of the objective function, contained in the list `parameters`.

```
obj <- MakeADFun(data = dat, parameters = parameters,
                 DLL="Cpp_file")
```

Lastly, we optimize the parameters of the objective function using a default R optimizer, `nlminb`. The objective function, its parameters, and its first and second derivatives are collected in the object `obj` generated by `MakeADFun`.

```
## Optimization
opt_nlminb <- nlminb(obj$par, obj$fn, obj$gr,obj$he,
                     control = list(iter.max = 7000,
                                    eval.max = 3000))
```

Notice that the code described above is general in terms of the number of components. To test several m's one can create an automated loop, while setting initial values vectors with an appropriate length at each iteration.

C++ code

The following code is written in C++ language. Notice that C++ indexes from 0, rather than 1 (as R does).

As a first step, we read the TMB package with the function `include`.

```
#include <TMB.hpp>
```

The purpose of the C++ code is to define the objective function as a class that can be derived. The entire code is contained in the following command:

```
template <class Type>
Type objective_function<Type>::operator() () {

...

}
```

In the R file we constructed two lists - one containing data, another containing parameters. We import them in the C++ code with two separate functions (`DATA` for data, `PARAMETER` for parameter).

```
DATA_MATRIX(data_mat_MZ);
DATA_MATRIX(data_mat_DZ);
DATA_INTEGER(m);
DATA_VECTOR(y_points);

PARAMETER_VECTOR(alpha);
PARAMETER_VECTOR(log_sigma);
PARAMETER_VECTOR(rho);
PARAMETER_VECTOR(tdelta);
```

We now generate the mean vector `mu`, the standard deviation vector `sigma`, and the weight vector `delta` from the imported parameters `alpha`, `log_sigma`, and `tdelta` using the appropriate functions. When defining the log likelihood of the Gaussian mixture distribution, it is more convenient to work directly on these parameters.

```
// Generate the vector mu
vector<Type> mu(m);
mu(0) = exp(alpha(0));
for(int i = 1; i < m; i++){
   mu(i) = mu(i - 1) + exp(alpha(i));
}
// Generate the vector sigma
vector<Type> sigma = log_sigma.exp();
// Generate the vector delta
vector<Type> delta(m);
delta(0) = Type(1); // set first element to one
if(m > 1){
   delta.tail(m - 1) = exp(tdelta);
```

```
    delta = delta/delta.sum();
}
```

We construct covariance matrices and mean vectors for each Gaussian component, starting from the initial values read from the R file. We follow the constraints described in Section 2.3. We create a vector of Gaussian components (`mvdnormMZ`) which has length m and whose each element is a Gaussian component, created through a loop. We show the process for monozygotic twins.

```
//Create uncentered mixing densities and mean vectors
vector<vector<Type> > mu_vec(m);
matrix<Type> covMZ(2,2);
matrix<Type> I(2,2);
matrix<Type> U(2,2);
U = U.setOnes();
I = I.setIdentity();
matrix<Type> V = U - I;
using namespace density;
vector<MVNORM_t<Type> > mvdnorm(m);

for(int i = 0; i < m; i++) {
    //covariance matrix of component i
   covMZ =  sigma2(i)*I + sigma2(i)*rhoMZ(i)*V;
   //mean vector of component i
   vector<Type> mu_vec_i(2);
   mu_vec_i(0) = mu(i);
   mu_vec_i(2) = mu(i);
   mu_vec(i) = mu_vec_i;
   // uncentered density of component i
   mvdnormMZ(i) = MVNORM_t<Type>(covMZ);
}
```

Repeating the process switching `rhoMZ` with `rhoDZ` generates the object `mvdnormDZ`, which represents the contribution of the dizygotic twins.

Now we define the objective function - that is, the negative log likelihood. Through a loop we compute the likelihood for each data point in the dataset, and add it to the quantity `nll`. Monozygotic and dizygotic twin pairs are divided in two separate datasets, so we show the code only for the former.

Notice that in the loop over the Guassian components we take the exponential of `mvdnorm` and we then take the negative logarithm of the value obtained. This is because `mvdnorm` is by default defined as minus the logarithm of a Gaussian, to streamline the calculation of the negative log likelihood. Since we are not calculating the negative log likelihood of a Gaussian distribution, but that of a Gaussian mixture, the logarithm must be applied after summing all mixture components together; for this reason we require the extra step explained above.

```
// Evaluate negative log-likelihood
Type nll = 0;
int n_mz = data_mat_MZ.rows();   // number of MZ pairs

for (int i = 0; i < n_mz; i++){
   // Row no. i
   vector<Type> Y(2);
   Y(0) = data_mat_MZ(i,0);
   Y(1) = data_mat_MZ(i,1);
    // Evaluate mixture for obs no. i
   Type prob_i;
   for(int j = 0; j < m; j++){
      prob_i += delta(j)*exp(-mvdnormMZ(j)(Y - mu_vec(j)));
   }
   // Contribution
   nll -= log(prob_i);
}
```

We repeat the same process with dizygotic twins, adding their contribution to `nll`.

The code returns the negative log likelihood, which is then optimized by the R function `nlminb`. We can also return the values of the estimated parameters using the function `ADREPORT`.

```
return nll;
ADREPORT(mu);
ADREPORT(sigma);
...
```

Other than defining the negative log likelihood function, we use the C++ file to define the heritability curve. Using `ADREPORT`, we can then return the estimated value of the curve to the R file.

To define the heritability curve, we first must calculate the monozygotic and dizygotic curve. Again, the process is analogous for the two zygosities, so we illustrate the process only for monozygotic twins.

We are assuming that the underlying distribution is a Gaussian mixture following the restrictions detailed in Section 2.3. Under these assumptions, it is possible to obtain an explicit formula of the correlation curve. In the code we refer to the notation and decomposition of the correlation curve as defined in ARTICLE I, Section 3.

We initialize the correlation curve, a vector which has the same length as the vector `y_points` spanning over the range of the trait studied.

```
int n_points= y_points.size();
vector<Type> cor_curve_MZ(n_points); // correlation curve
```

All the following calculations are performed inside a loop over the points in `y_points`. In the loop we initialize and define the conditional probabilities $p_k^*(y) = P(\delta = k|Y_2 = y)$. Recall that m is the number of components, and whenever we loop over m we are considering the contributions of all mixture components.

```
vector<Type> p_star(m);
vector<Type> pk_norm(m);

for(int k = 0; k < m; k++){
   pk_norm(k) = delta(k)*dnorm(y_points(i), mu(k), sigma(k), false);
}

for(int k = 0; k < m; k=){
   p_star(k) = pk_norm(k)/pk_norm.sum();
}
```

We then initialize and define the conditional mean $\mu(y)$, the conditional variance $\sigma^2(y)$, and the derivative of the conditional mean $\beta(y)$, evaluated at each point `y_points(i)`. We divide the formula of $\beta(y)$ in three pieces to simplify the computations.

```
// Conditional mean
vector<Type> CM_vec(m);
for(int k = 0; k < m; k++){
   CM_vec(k) = p_star(k)*(mu(k) + rho(k)*(y_points(i) - mu(k)));
}
Type CM = CM_vec.sum();
```

```
// Conditional variance
vector<Type> CV_vec(m);
for(int k = 0; k < m; k++){
   CV_vec(k) = sigma(k)*sigma(k)*(1 - rhoMZ(k)*rhoMZ(k))*p_star(k)+
        p_star(k)*(mu(k) + rhoMZ(k)*(y_points(i) - mu(k))- CM)*
        (mu(k) + rhoMZ(k)*(y_points(i) - mu(k)) - CM);
}
Type CV = CV_vec.sum();

// Derivative of conditional mean
vector<Type> Beta_A_vec(m);
vector<Type> Beta_B_vec(m);
vector<Type> Beta_C_vec(m);

for(int k = 0; k < m; k++){
   Beta_A_vec(k) = p_star(k)*(rhoMZ(k) - (mu(k) + rhoMZ(k)*
               (y_points(i) - mu(k)))*(y_points(i) - mu(k))/
               (sigma(k)*sigma(k)));
   Beta_B_vec(k) = p_star(k)*(mu(k) + rhoMZ(k)*(y_points(i) - mu(k)));
   Beta_C_vec(k) = p_star(k)*(y_points(i) - mu(k))/(sigma(k)*sigma(k));
}
Type Beta = Beta_A_vec.sum() + Beta_B_vec.sum()*Beta_C_vec.sum();
```

One last component that we require to compute the correlation curve is the marginal standard deviation τ_1.

```
Type mu_bar = (delta*mu).sum();
Type marginal_variance = (delta*sigma2).sum() +
               (delta*(mu - mu_bar)(mu - mu_bar).sum();
Type marginal_sd = sqrt(marginal_variance_M);
```

We can then define the monozygotic correlation curve at the point `y_points(i)`:

```
Type sigma_Beta = marginal_sd*Beta;
Type sigma_Beta_squared = sigma_Beta*sigma_Beta;
cor_curve_MZ(i) = sigma_Beta/sqrt(sigma_Beta_squared + CV);
```

We can repeat the process for dizygotic twins, and obtain the dizygotic correlation curve. With both monozygotic and dizygotic correlation curves, we can finally define the heritability curve (we use the formula of $a^2_{ACE}(y)$ in an ACE model in this context).

```
vector<Type> her_curve(n_points);
for(int i = 0; i < n_points; i++){
   her_curve(i) = 2*(cor_curve_MZ(i) - cor_curve_DZ(i));
}
```

The curves $a^2_{ADE}(y)$, $c^2(y)$, $d^2(y)$, and $e^2(y)$ are linear combinations of the monozygotic and dizygotic correlation curves and can be easily computed analogously to $a^2_{ACE}(y)$.

Adding a sex covariate

In the data studied in Article II we have information about the sex of the twin pairs (only same-sex twin pairs were collected in the dataset). Following the model explained in Article II, we assume that male and female pairs share all parameters, except for the mean vector. We define a parameter γ such that $\boldsymbol{\mu_M} = \boldsymbol{\mu_F} + \gamma$. (In the articles this parameter is called β, but to avoid confusing it with the derivative of the conditional mean $\beta(y)$ we change the name in this explanation).

We define the new (scalar) parameter `gamma`, initialized in the R code and passed to the C++ code through the list `parameters`.

We also need to provide information about the sex of each pair; we do it with two binary vectors `sexMZ` and `sexDZ`, containing a 0 in the positions corresponding to female twin pairs, and a 1 in the positions corresponding to male twin pairs. These vectors are passed to C++ through the list `data`.

In the C++ code, instead of working with the parameter `mu`, we define the following two vectors:

```
vector<Type> mu_F(m);
vector<Type> mu_M(m);
for(int k = 0; k < m; k++){
   mu_F(i)= mu(i)-0.5*gamma;
   mu_M(i)= mu(i)+0.5*gamma;
}
```

We then define the objects `mu_vec_F` and `mu_vec_M` using the appropriate mean parameter. When computing the contributions of monozygotic and dizygotic twins to the negative log likelihood, we condition on the vector `sexMZ` and `sexDZ`; for example

```
if (sexMZ(i) == 1) {
   for(int j = 0; j < m; j++){
      prob_i += delta(j)*exp(- mvdnorm_DZ(j)(Y - mu_vec_M(j)));
   }
}
```

and so forth.

We compute two different heritability curves for male and female data, simply repeating the steps described in the Appendix section "C++ code" separately with the parameter `mu_F` and `mu_M`.

Bibliography

D. J. P. Barker and C. Osmond. Infant mortality, childhood nutrition, and ischaemic heart disease in england and wales. *The Lancet*, 327(8489):1077–1081, 1986.

J.-P. Baudry and G. Celeux. Em for mixtures: Initialization requires special care. *Statistics and computing*, 25:713–726, 2015.

B. Bell. *CppAD: a package for C++ algorithmic differentiation*, 2017. URL `http://www.coin-or.org/CppAD`.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

S. Bjerve and K. Doksum. Correlation curves: Measures of association as functions of covariate values. *The Annals of Statistics*, pages 890–902, 1993.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Y. Dodge. *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

D. S. Falconer and T. F. C. Mackay. *Quantitative genetics*. Longman London, 1983.

G. Jacob, B; Guennebaud. *Eigen: a C++ linear algebra library*, 2006.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

M. S. Kramer. Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the world health organization*, 65(5):663, 1987.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, 2016.

S. Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160, 1976.

J. A. R. Logan, S. A. Petrill, S. A. Hart, C. Schatschneider, L. A. Thompson, K. Deater-Deckard, L. S. DeThorne, and C. Bartlett. Heritability across the distribution: An application of quantile regression. *Behavior genetics*, 42:256–267, 2012.

D. J. C. MacKay et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley New York, 2000.

V. Melnykov and I. Melnykov. Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395, 2012.

R. D. Neidinger. Introduction to automatic differentiation and matlab object-oriented programming. *SIAM review*, 52(3):545–563, 2010.

M. E. Palatianou, Y. V. Simos, S. K. Andronikou, and D. N. Kiortsis. Long-term metabolic effects of high birth weight: a critical review of the literature. *Hormone and Metabolic Research*, 46(13):911–920, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

L. Reyes and R. Manalich. Long-term consequences of low birth weight. *Kidney International*, 68:S107–S111, 2005.

P. C. Schalock, J. T. S. Hsu, and K. A. Arndt. *Lippincott's primary care dermatology*. Lippincott Williams & Wilkins, 2010.

B. Stroustrup. *The C++ Programming Language*, 1995.

P. T. Williams. Quantile-dependent heritability of computed tomography, dual-energy x-ray absorptiometry, anthropometric, and bioelectrical measures of adiposity. *International journal of obesity*, 44(10):2101–2112, 2020.

S. Wright. Correlation and causation. 1921.

C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

Chapter 6

Scientific results

Article I

6.1 Heritability curves: A local measure of heritability in family models

Geir D. Berentsen, Francesca Azzolini, Hans J. Skaug, Rolv T. Lie, Håkon K. Gjessing

Statistics in Medicine, **40.6**, 1357-1382 (2021)

Heritability curves: A local measure of heritability in family models

Geir D. Berentsen, Francesca Azzolini, Hans J. Skaug, Rolv T. Lie, Håkon K. Gjessing

Abstract

This paper introduces a new measure of heritability which relaxes the classical assumption that the degree of heritability of a continuous trait can be summarized by a single number. This measure can be used in situations where the trait dependence structure between family members is non-linear, in which case traditional mixed effects models and covariance (correlation) based methods are inadequate. Our idea is to combine the notion of a correlation curve with traditional correlation-based measures of heritability, such as Falconer's formula. For estimation purposes, we use a multivariate Gaussian mixture, which is able to capture non-linear dependence and respects certain distributional constraints. We derive an analytical expression for the associated correlation curve, and investigate its limiting behaviour when the trait value becomes either large or small. The result is a measure of heritability that varies with the trait value. When applied to birth weight data on Norwegian mother–father–child trios, the conclusion is that low and high birth weight are less heritable traits than medium birth weight. On the other hand, we find no similar heterogeneity in the heritability of Body Mass Index (BMI) when studying monozygotic and dizygotic twins.

1 Introduction

Biometrical modeling of family trait correlations has a very long tradition, going back at least to Ronald Fisher [Fisher, 1919] and Sewall Wright [Wright, 1920, 1921], and being developed into an extensive modeling framework over the years [Bulmer, 1985, Neale, 2002], with openly available software tools, such as OpenMx [Neale et al., 2016]. For a continuous trait Y, such as weight or height, the basic idea is that trait variability –

or more precisely, the variance of the measured trait, $\mathrm{Var}(Y)$ – can be decomposed into genetic and environmental components, each explaining a portion of the observed trait variance. Thus, the concept of *heritability* can, loosely, be defined as the proportion of trait variance explained by genetic components, with environmental influences assumed to explain the rest [Hopper, 2002]. As an example, the most common twin model, known as the ACE model, decomposes the trait Y into additive genetic effects (A), common (shared) environment (C), and residual (random) environment (E). In terms of variances, we commonly define quantities a^2, c^2, and e^2 as the *proportions* of trait variances explained by the components A, C, and E, respectively. Thus, assuming that no other effects are present, we have $a^2 + c^2 + e^2 = 1$.

To separate genetic variance from environmental variance, family data are needed. Genetic correlations between family members decrease in more distant relationships, thus providing contrasts from which the genetic components can be estimated. For instance, in the classical ACE twin design, the additive genetic correlation in monozygotic twin pairs is assumed to be 1, whereas the corresponding correlation, or degree of shared genetic influence, is assumed to be 1/2 in dizygotic twin pairs. In addition, it is frequently assumed that the amount of shared environment is the same in dizygotic twins as is monozygotic twins. The quantities a and c above can also be seen as the degree to which the underlying genes A and shared environmental C are being "expressed" in the phenotype of each individual. Thus, the monozygotic twin pair phenotype correlation will be $\rho^{(MZ)} = a^2 + c^2$, and $\rho^{(DZ)} = \frac{1}{2}a^2 + c^2$ for the dizygotic twin pairs. As a consequence, the difference $\frac{1}{2}a^2$ between monozygotic and dizygotic twin pair correlations is ascribed to genes alone, providing an estimate of the heritability a^2.

The ACE model is very specific in its assumption of additive genetic effects, as well as independent, additive contributions from the environment. In the biometrical modeling literature, a wide range of variants and extentions have been developed. Using family structures of increasing complexity, numerous different effects can be identified, such as additive genetic effects, dominant genetic effects, X-chromosome effects, effects of maternal genes on the fetus during pregnancy, effects of mitochondrial genes, gene-gene interactions, gene-environment interactions, etc. [Neale, 2002, Hopper and Visscher, 2002, Gjessing and Lie, 2008]. Extending the family structures used for modeling is in general challenging since genetic correlations between more distant relatives quickly drop to nearly undetectable levels, and assumptions about how environmental factors are shared within larger families become harder to verify [Gjessing and Lie, 2008]. Still, with a steady increase in registry-based population studies with large sample sizes and available data on environmental covariates, such modeling has become feasible.

Common to practically all models in the field is that the degree of heritability is assumed

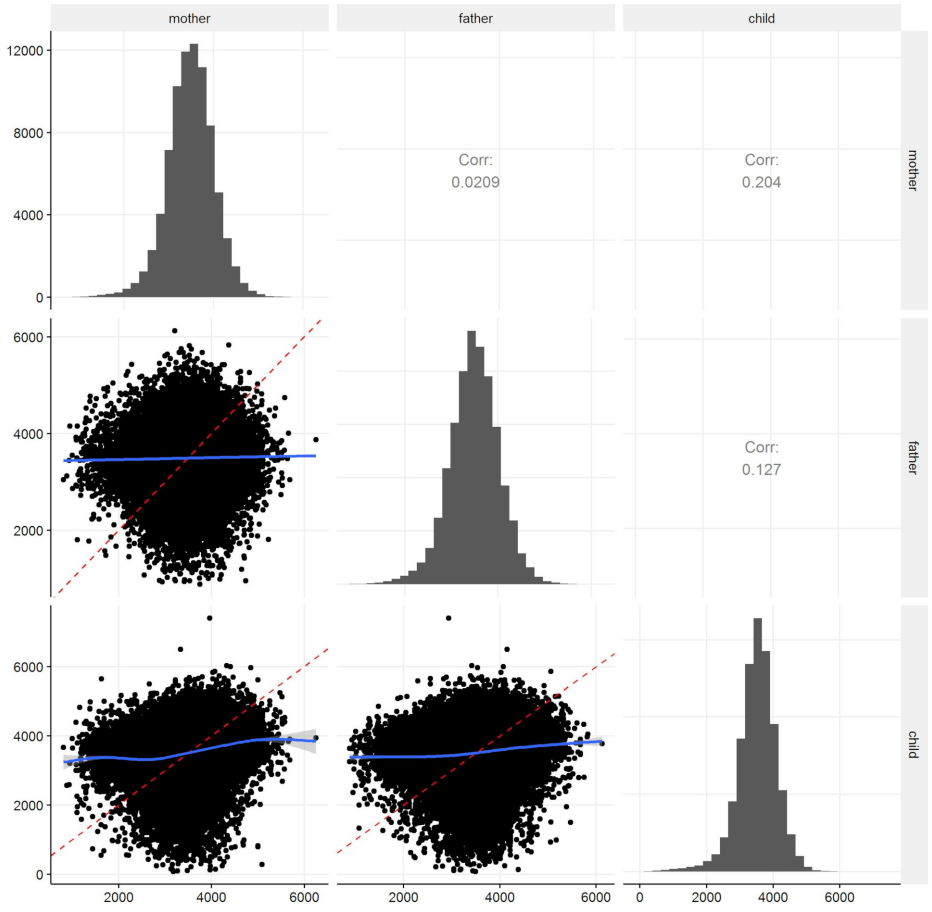


Figure 1: Birth weights (gram) for 81,144 mother–father–child trios from the Norwegian Birth Registry. Diagonal: histograms of marginal birth weights. Lower triangle: pairwise scatter plots with estimated nonparametric regression line (blue) and identity line (dashed red), where $y = x$. Upper triangle: pairwise empirical correlation.

constant across the full range of the phenotype. For instance, the estimated proportion a^2 of variance explained by additive genes is assumed to be the same whether the phenotype Y is small, close to its mean, or large. It seems clear, however, that for instance rare but dramatic environmental influences on the phenotype may occasionally cause the phenotype to deviate strongly from its mean value, much more than would be expected under "normal" circumstances. Below, we illustrate our models of heritability using a child's birth weight (BW) as phenotype. While the birth weight distribution is close to a normal distribution, it has a heavier tail to the left (Figure 1); this may indicate a higher proportion of low birth weight children than what would be expected

from many minor genetic and environmental components adding up during pregnancy.

This simple observation may suggest that the degree of heritability of birth weight can differ in the different ranges of weight; perhaps the lowest BW values are caused by "rouge" environmental factors that act more strongly than genetic effects in the tail, or maybe they are caused by rare, recessive genes that only occasionally excert a strong negative influence on BW.

These observations motivate us to look for differences in heritability across the range of the trait value Y. The existing methods for investigating such differences are almost exclusively based on regression methods. In their seminal work [DeFries and Fulker, 1985], DeFries and Fulker evaluate the degree of regression to the mean for co-twins of probands from strata in the tails of a continuous trait distribution. The idea is that if the trait is heritable, then we should observe DZ co-twins with a higher degree of regression to the mean compared to the MZ co-twins. This approach is known as DeFries-Fulker (DF) extremes analysis for twins. Later, a formal test was developed to examine whether the heritability of the trait for probands in the selected strata was equal or different to the unselected population [DeFries and Fulker, 1988]. This methodology was extended by Cherny et al. [Cherny et al., 1992a] by considering interaction effects between the heritability of the trait and the realized value of the trait for the proband. This approach can be used to detect linear and quadratic changes in heritability as the trait value changes. These methods all have the drawback of only providing a rough description of how the heritability varies with the trait value. The DF approach requires the researcher to select a cut-off point (a low or high trait value) for choosing the strata; the result can thus be misleading if the heritability changes smoothly as the trait value vary. Conversely, if there exists a point in the trait distribution where the heritability jumps and then stabilize again, the Cherny approach will only model this change by a linear or quadratic curve.

These drawbacks were addressed in Logan et al. [2012a] using quantile regression; by using the extended DF extremes analysis [LaBuda et al., 1986] as the quantile regression equation, the authors obtain a heritability measure for each quantile of the trait distribution. Consequently, their method results in a heritability measure for each value of the trait $Y = y$, corresponding to a specific quantile of the distribution.

However, in the present paper we introduce an approach based on localizing traditional genetic models. Informally, this means making sense of estimating, for instance, the additive genetic effect as a function of the phenotype; i.e. to define meaningfully $a^2(y)$ as the proportion of phenotype variance explained by additive genetic effects, conditional on $Y = y$. Such a definition may seem self-contradictory since one conditions on the

variable whose variance is being decomposed. Nevertheless, it is fully possible to make sense of this concept, and we show in this paper how to develop *heritability curves*, such as $a^2(y)$. This definition thus provides a "local" measure of heritability, depending on the phenotype value.

As for the ACE twin model, all standard biometrical models rely on the phenotype correlations between family individuals to estimate the variance components that determine heritability. Our starting point for developing a local measure of heritability is thus a local measure of dependence between family members; more specifically, we need a local measure of correlation. There are several local measures proposed in the literature, such as the local Gaussian correlation [Tjøstheim and Hufthammer, 2013], the dependence function [Holland and Wang, 1987], and the correlation curve [Bjerve and Doksum, 1993]. We base our approach on the correlation curve [Bjerve and Doksum, 1993] $\rho(y)$, which can be defined as a measure of locally explained variance, and thus fits the framework of heritability as a proportion of explained variance. The correlation curve is similar to the traditional Pearson's correlation in that it takes values between minus one and one, and the square $\rho^2(y)$ is a measure of locally explained variance. In a bivariate Gaussian distribution, the correlation curve is constant (independent of y), and equal to the standard Pearson correlation. In contrast to the Pearson correlation the local correlation of a bivariate relationship depends on direction; for a bivariate random variable (Y_1, Y_2), the locally explained variance of Y_2 conditional on $Y_1 = y$ may differ from the locally explained variance of Y_1 conditional on $Y_2 = y$.

With phenotype measurements on, for instance, a mother (Y_1) and her child (Y_2), it may seem reasonable, for instance, to study the distribution of a child phenotype conditionally on the maternal phenotype. However, most biometrical models are formulated in terms of genetic and environmental factors *shared* by the two family members, thus assuming a form of exchangeability between the two. This is particularly clear in twin pairs, where conditioning one twin on the other twin is unnatural. In the model of Logan et al. [2012b] this assignment was done randomly, while Cherny et al. [1992b] explored both a random assignment and a double-entry approach. However, the population value of the correlation curve can be derived from the joint distribution of two variables. If the joint distribution is exchangeable, so that (Y_1, Y_2) has the same bivariate distribution as (Y_2, Y_1), the correlation curve is invariant to which variable we condition on, i.e. whether we measure the locally explained variance of Y_1 conditional on Y_2 or vice versa. This means that the role of the mother and child in the above interpretation can be interchanged.

The correlation curve may be estimated parametrically or non-parametrically from observed values of a bivariate distribution (Y_1, Y_2) by conditioning on either $Y_1 = y$ or

$Y_2 = y$. However, our approach is instead to first model the bivariate distribution as a Gaussian mixture distribution, where the mixture distribution is restricted in such a way as to be exchangeable. From the mixture distribution, the correlation curve can be derived explicitly. We estimate the distribution by maximum likelihood, and by allowing a sufficient number of components, a mixture distribution is very flexible and fits a wide range of distributional shapes. Having obtained the parameters of the mixture distribution, the correlation curve can be derived from its explicit expression by plugging in the estimated parameters.

The paper is structured as follows. In Section 2, we define a standard mixed-effect model for continuous traits, and structure it for two specific family models: twin pairs and mother–father–child trios. Following a standard twin approach [Falconer, 1960], and models for family trios [Magnus et al., 2001, Lunde et al., 2007], we derive expressions for the heritability estimates in both family structures. In Subsections 2.2 and 2.3, we explain the concept of correlation curves, and extend the traditional definition of heritability to the heritability curve, which depends on the trait value y. In Section 3, we introduce and analyze a Gaussian mixture [McLachlan and Peel, 2000] for bivariate phenotype distributions, parameterized to be exchangeable. We then study the limiting behaviour of the correlation curve for large and small phenotype values under this model in Subsection 3.1. Lastly, in Subsection 3.2, we discuss the estimation of the correlation curve for the twin-pairs and the mother–father–child trios models. Section 4 provides two applications of this approach. Namely, the first application is the analysis of BMI values for twin pairs collected in the dataset "twinData", found in the R-package "OpenMx" [Neale et al., 2016]; the second one is the analysis of birth weight data of mother–father–child trios from the Medical Birth Registry of Norway. For both family structures we compute AIC and BIC values to select the best-fitting mixture models, and explore the resulting distributions and heritability curves. Proofs are provided in an appendix.

2 Development of Heritability curves

2.1 Traditional models for twins and family trios

We first provide a basic description of how traditional biometrical models can be set up in some generality, and in particular for twins and family trios. While there are numerous ways of building, parametrizing, and interpreting such models, our approach is fairly standard, and in a form that supports our development of heritability curves. Let Y_{ij} be the trait value of individual j in a family i, and consider the mixed-effect

model (see e.g. McCulloch and Neuhaus [2001])

$$Y_{ij} = \mu + \beta^t x_{ij} + A_{ij} + C_{ij} + D_{ij} + E_{ij}, \tag{1}$$

where A_{ij}, C_{ij}, D_{ij} and E_{ij} represent additive genetic, common environmental, dominant genetic, and residual environmental random effects, respectively (see e.g. Falconer [1960]). We assume the four components A_{ij}, C_{ij}, D_{ij} and E_{ij} to be mutually independent, with mean 0 and variances σ_A^2, σ_C^2, σ_D^2 and σ_E^2. The inclusion of the term $\beta^t x_{ij}$ (fixed effects) allows the average phenotype level to depend on covariates. Note that this model assumes no gene-environment interaction. In traditional biometrical modelling (see e.g. Gjessing and Lie [2008]) the random effects are assumed to be normally distributed with expectation 0, i.e. $A_{ij} \sim N(0, \sigma_A^2)$, $C_{ij} \sim N(0, \sigma_C^2)$, $D_{ij} \sim N(0, \sigma_D^2)$ and $E_{ij} \sim N(0, \sigma_E^2)$. The assumption of normality is seen as natural based on the central limit theorem if Y is the result of numerous small, independent genetic and environmental effects that add up to produce the trait value. Under the above assumptions the total variance of the trait is given by

$$\sigma^2 = Var(Y_{ij}) = \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2. \tag{2}$$

We define $a^2 = \sigma_A^2/\sigma^2$, $c^2 = \sigma_C^2/\sigma^2$, $d^2 = \sigma_D^2/\sigma^2$, and $e^2 = \sigma_E^2/\sigma^2$ as the proportions of the total variance that derive from each of the four genetic and environmental components. Note that

$$a^2 + c^2 + d^2 + e^2 = 1,$$

i.e. the contributions from all components sum to one. Thus, in a model including A, C, and E, excluding dominant effects, one may quantify the genes-versus-environment contribution to trait variability as a^2. This proportion is often referred to as *heritability* and can be interpreted as how strongly the genetic effect A_{ij} contributes to the trait value. The heritability based on the additive genetic component is often referred to as *narrow sense heritability*. Some models may also include dominant genetic effects, and in such cases one may refer to $a^2 + d^2$ as the *broad sense heritability* [Khoury et al., 1993].

From independent observations of Y_{ij} alone, it is not possible to identify the individual variance components σ_A^2, σ_C^2, σ_D^2, and σ_E^2 in (2), only the total variance σ^2. In order to make the individual variances identifiable, one has to consider data on family members, for which the Y's are correlated due to shared genetic material and environment. We focus on two basic family structures — mother–father–child trios and twin pairs — in the following. As is well known, these family structures are quite restricted in the number of

effects they allow to be estimated, and assumptions have to be made about what genetic and environmental effects to include in each model. In the following, we will present the specific models that will serve as illustrations when developing heritability curves.

2.1.1 Twins

Perhaps the best known biometrical model is the ACE model for twins, complemented by the alternative ADE model. While the expressions for twin correlations in these models are very well known, we state them here as a starting point for the heritability curves.

Let Y_{ij} be the trait value of twin j ($j = 1, 2$) in twin-pair i. Let $\rho^{(MZ)}$ and $\rho^{(DZ)}$ be the phenotype correlations $\text{cor}(Y_{i1}, Y_{i2})$ for MZ and DZ twins, respectively. Both ACE and ADE models include the additive genetic component A. For MZ-twins $\text{cor}(A_{i1}, A_{i2}) = 1$, while for DZ-twins $\text{cor}(A_{i1}, A_{i2}) = 1/2$. In the standard ACE model, the correlation for the common environmental effect is assumed to be $\text{cor}(C_{i1}, C_{i2}) = 1$ in all twin pairs; thus, one makes the common assumption of DZ twins sharing their environment to the same degree as the MZ twins. In the alternative ADE one assumes $\text{cor}(D_{i1}, D_{i2}) = 1$ for MZ twins and $\text{cor}(D_{i1}, D_{i2}) = 1/4$ for DZ twins. In both models, residual environmental effects are assumed to be independent.

Since the basic twin models utilize only the $\rho^{(MZ)}$ and $\rho^{(DZ)}$ phenotype correlations, they allow estimating two parameters. In addition, e^2 can be estimated from $e^2 = 1 - a^2 - c^2 - d^2$. The ACE model assumes $d^2 = 0$, and thus the parameters a^2, c^2, and e^2 can be identified; the ADE model assumes $c^2 = 0$, and thus the parameters a^2, d^2, and e^2 can be identified.

For the ACE model, it follows from the above that

$$\begin{aligned} \rho^{(MZ)} &= a^2 + c^2, \\ \rho^{(DZ)} &= \frac{1}{2}a^2 + c^2. \end{aligned}$$

For the ADE model, the equations are

$$\begin{aligned} \rho^{(MZ)} &= a^2 + d^2, \\ \rho^{(DZ)} &= \frac{1}{2}a^2 + \frac{1}{4}d^2. \end{aligned}$$

The simplest approach to estimating a^2, c^2, and d^2 is by moment estimators, i.e. to solve this set of equations, using empirical values for $\rho^{(MZ)}$ and $\rho^{(DZ)}$, and use $e^2 =$

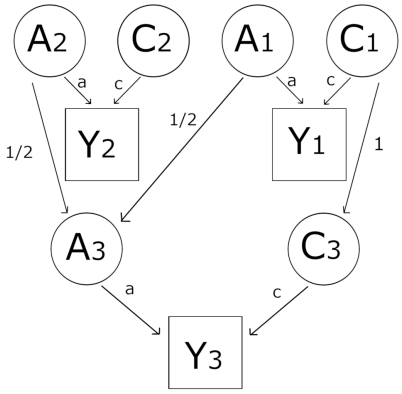


Figure 2: Path diagram representing the birth weight of mother Y_1, father Y_2, and child Y_3 (represented as squares). The traits are determined by the unobserved genotype values (A) and environmental values (C) (shown as circles), as well as the independent residual environmental values (E) (not shown).

$1 - a^2 - c^2 - d^2$ to estimate e^2. The resulting solutions for the ACE model are the celebrated formulas of Falconer: [Falconer, 1960]

$$\begin{aligned} a^2 &= 2(\rho^{(MZ)} - \rho^{(DZ)}), \\ c^2 &= 2\rho^{(DZ)} - \rho^{(MZ)}, \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{3}$$

For the ADE model, the corresponding set of solutions are

$$\begin{aligned} a^2 &= 4\rho^{(DZ)} - \rho^{(MZ)}, \\ d^2 &= 2(\rho^{(MZ)} - 2\rho^{(DZ)}), \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{4}$$

Without further assumptions, an informal choice between the ACE and ADE models is often made based on whether empirically $\rho^{(MZ)} < 2\rho^{(DZ)}$ or not. If this is the case, the ACE model is a natural choice; otherwise, the ADE model can be used.

2.1.2 Mother-father-child trios

Let Y_{ij} be the observed trait value of individual j in nuclear family trio i. We let $j = 1, 2, 3$ correspond to the mother, father, and child, respectively. A phenotype correlation between mother and father may signify, for instance, assortative mating, inbreeding, or

social homogamy among the parents. However, the correlation is typically low, and we will here assume it is zero [Magnus et al., 2001]. There are thus only two correlations that provide information: the mother-child and father-child correlations. There are numerous ways of parametrizing correlations in nuclear families [Magnus et al., 2001, Pawitan et al., 2004, Lunde et al., 2007, Gjessing and Lie, 2008, Rabe-Hesketh et al., 2008], but being restricted to two correlations means that these cannot be separated. In our setting, we assume, for additive autosomal genes, that $\text{cor}(A_{i1}, A_{i3}) = \text{cor}(A_{i2}, A_{i3}) = 1/2$, and that $\text{cor}(A_{i1}A_{i2}) = 0$ for the parents. Also, we assume that mother and child share an environmental component, but no such sharing between father and child, leading to $\text{cor}(C_{i1}, C_{i3}) = 1$ and $\text{cor}(C_{i2}, C_{i3}) = 0$. Thus,

$$\Sigma_A = \sigma_A^2 \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{bmatrix}, \quad \Sigma_C = \sigma_C^2 \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \text{ and } \Sigma_E = \sigma_E^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

are the covariance matrices for the vectors (A_{i1}, A_{i2}, A_{i3}), (C_{i1}, C_{i2}, C_{i3}), and (E_{i1}, E_{i2}, E_{i3}), respectively.

A graphical representation of the above model is displayed in a path diagram in Figure 2.

Under the above assumptions the vectors (Y_{i1}, Y_{i2}, Y_{i3}) are *i.i.d.* multivariate normal with mean

$$(\mu + \beta^t x_{i1}, \mu + \beta^t x_{i2}, \mu + \beta^t x_{i3}) \tag{5}$$

and covariance matrix

$$\Sigma = \Sigma_A + \Sigma_C + \Sigma_E = (\sigma_A^2 + \sigma_C^2 + \sigma_E^2) \begin{bmatrix} 1 & 0 & \frac{1}{2}a^2 + c^2 \\ 0 & 1 & \frac{1}{2}a^2 \\ \frac{1}{2}a^2 + c^2 & \frac{1}{2}a^2 & 1 \end{bmatrix}, \tag{6}$$

where a^2, c^2, and e^2 are defined as above. Again, the unknown values can simply be estimated by the methods of moments by matching the correlation matrix (6) to its empirical counterpart, and solve for a^2, c^2 and e^2 under the constraint $a^2 + c^2 + e^2 = 1$. The solution is given by the following equations

$$\begin{aligned} a^2 &= 2\rho^{(FC)} \\ c^2 &= \rho^{(MC)} - \rho^{(FC)} \\ e^2 &= 1 - \rho^{(MC)} - \rho^{(FC)}, \end{aligned} \tag{7}$$

where $\rho^{(MC)}$ and $\rho^{(FC)}$ are the mother-child and father-child correlations, respectively.

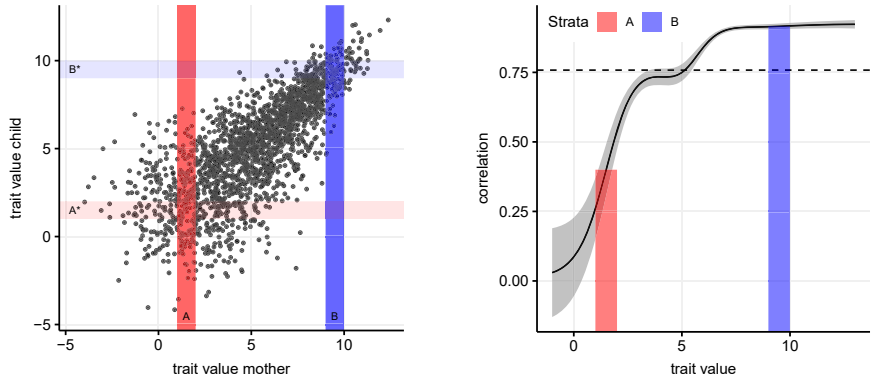
We will, in the following, use these solutions, and those for the ADE twin model, to obtain local versions of a^2, c^2, d^2, and e^2. Note that in both cases, the underlying assumption is that the covariance (correlation) matrix completely characterizes the dependence structure between traits in a family and can be decomposed as in e.g. (6).

2.2 Correlation curves for non-linear bivariate relationships

We now explain the concept of local correlation curves, following the approach of Bjerve and Doksum [Bjerve and Doksum, 1993]. To illustrate the principle of localization, we use simulated data from a hypothetical phenotype, as seen in Figure 3a.

We consider two strata (A and B) consisting of all mother-child pairs for which the mother's trait $Y_1 = y_1$ falls within two intervals (interval A and B) on the x-axis. The corresponding correlation curve is shown in Figure 3b; as a function of y_1 (horizontal axis) it is smaller in stratum A than in stratum B. This indicates that the mother-child association is stronger in stratum B compared to stratum A. In a non-parametric regression setting, this would mean that the child's trait can be predicted by the mother's trait with higher precision in stratum B than in stratum A. For both strata, an increase in the mother's trait is associated with an increase in the child's trait since the correlation curve is positive. Since the correlation curve is continuous, the location argument y_1 can be seen as the center of infinitesimal intervals from which strata such as A and B can be constructed, while the value of the correlation curve is a measure of dependence for the corresponding strata. A constant correlation curve indicates that the dependence properties are constant across these strata, while a varying correlation curve indicates strata that differ in their dependence properties.

If the joint distribution is exchangeable, so that (Y_1, Y_2) has the same bivariate distribution as (Y_2, Y_1), the correlation curve is invariant to which variable we condition on, i.e. whether we measure the locally explained variance of Y_1 conditional on Y_2 or vice versa. This means that the role of the mother and child in the above interpretation can be interchanged, and the dependence structure in strata A^* and B^* in Figure 3a) is similar to the dependence structure in strata A and B; the correlation curve $\rho(y)$ as a function of y thus represents a measure of the mother-child trait dependence when either the mother or the child has trait value equal to y. In the next section, we show more precisely how $\rho(y)$ is defined in terms of locally explained variance.



(a) Simulated data from an exchangeable Gaussian mixture, and the definition of strata.

(b) Estimated correlation curve for the data displayed in panel (a), and 95% pointwise confidence intervals are shown in grey. The height of the bars displays the average value of the correlation curve within strata A and B.

Figure 3: Illustration of the concept of a correlation curve and the role of exchangeability using simulated data. Strata A and B include all mother-child pairs for which the mother's trait value falls in the intervals $[1, 2]$ and $[9, 10]$, respectively. Strata A^* and B^* include all mother-child pairs for which the child's trait value falls in the same intervals.

2.2.1 Standard correlation curves for bivariate relationships

Let (Y_1, Y_2) be random variables from a bivariate continuous distribution, and define $\tau_1^2 = \mathrm{Var}(Y_1)$, $\tau_2^2 = \mathrm{Var}(Y_2)$, and $\rho = \mathrm{cor}(Y_1, Y_2)$. Further, define $\mu(y) = \mathrm{E}(Y_1|Y_2 = y)$ and $\sigma^2(y) = \mathrm{Var}(Y_1|Y_2 = y)$ as functions of y. Assuming that $\mu(y)$ is differentiable, define $\beta(y) = \mu'(y)$, i.e. the slope of the (typically non-linear) regression curve $\mu(y)$ when Y_1 is regressed on Y_2. Recall that in a standard linear regression context, $\mu(y)$ is a linear function of y, where the slope $\beta_{1|2} := \beta(y)$ and the conditional variance $\sigma^2_{1|2} := \sigma^2(y)$ are both constant.

By the law of total variance,

$$\mathrm{Var}(Y_1) = \mathrm{Var}(\mathrm{E}(Y_1|Y_2)) + \mathrm{E}(\mathrm{Var}(Y_1|Y_2)),$$

and it thus seems natural to define in general

$$\text{Proportion of } \mathrm{Var}(Y_1) \text{ explained by } Y_2 = \frac{\mathrm{Var}(\mathrm{E}(Y_1|Y_2))}{\mathrm{Var}(\mathrm{E}(Y_1|Y_2)) + \mathrm{E}(\mathrm{Var}(Y_1|Y_2))}.$$

In the case of linear regression, $\mathrm{Var}(\mathrm{E}(Y_1|Y_2)) = \tau_2^2\beta_{1|2}^2$ and $\mathrm{E}(\mathrm{Var}(Y_1|Y_2)) = \sigma_{1|2}^2$, and the proportion of explained variance can thus be written

$$\frac{(\tau_2\beta_{1|2})^2}{(\tau_2\beta_{1|2})^2 + \sigma_{1|2}^2} = \left(\frac{\tau_2\beta_{1|2}}{\tau_1}\right)^2 = \rho^2, \tag{8}$$

which is the usual formula for explained variance in a linear regression.

We want to define a "local" variant of ρ^2, describing the proportion of explained variance when $Y_2 = y$, thus to define $\rho^2(y)$ as a function of y. To this end, (8) is a natural starting point, and the extension to a non-linear setting would thus be to allow both $\beta(y)$ and $\sigma^2(y)$ to depend on y. This leads to the definition

$$\rho(y) = \frac{\tau_2\beta(y)}{\left[(\tau_2\beta(y))^2 + \sigma^2(y)\right]^{1/2}}, \tag{9}$$

where we recall that $\tau_2^2 = \mathrm{Var}(Y_2)$, $\beta(y) = \frac{d}{dy}\,\mathrm{E}(Y_1|Y_2 = y)$, and $\sigma^2(y) = \mathrm{Var}(Y_1 \mid Y_2 = y)$.

Indeed, this is the formula developed by Bjerve et al. [Bjerve and Doksum, 1993] and Doksum et al. [Doksum et al., 1994]. As pointed out by Bjerve et al., the correlation curve should not be confused with the conditional correlation obtained by applying the usual correlation formula to the conditional distribution of (Y_1, Y_2) given $Y_2 = y$, which would always be zero. It should also be noted that while τ_2 is kept fixed in (9), the denominator $(\tau_2\beta(y))^2 + \sigma^2(y)$ is no longer necessarily equal to $\tau_1^2 = \mathrm{Var}(Y_1)$ from the original distribution. In fact, for a *fixed* $y = y_0$, it corresponds to $\mathrm{Var}(Z_1)$ from a hypothetical bivariate distribution (Z_1, Z_2) where $\mathrm{Var}(Z_2) = \tau_2^2$ and $\mathrm{Var}(Z_1)$ is determined from having a linear regression of Z_1 on Z_2 with constant slope $\beta(y_0)$ and constant conditional variance $\mathrm{Var}(Z_1|Z_2) = \sigma^2(y_0)$.

2.2.2 Correlation curves for symmetric bivariate relationships

In our setting, we are interested in relationships between pairs of family members, for example, a pair of twins or a child and a parent. We denote the pair's respective trait values by Y_1 and Y_2. At first glance, it may seem natural to ask about the explained variation of a child trait Y_1, conditional on its parental value Y_2. However, this is less natural for twins, who are from the same generation. Indeed, most biometrical models assume that the positive correlation between the trait values is generated by shared genes and shared environment; the sharing is symmetrical between family members, and the generational aspect is only used to compute the degree of relatedness. That is, in pairs of

family members, the two members should be exchangeable, so that (Y_1, Y_2) and (Y_2, Y_1) have the same bivariate distribution. Clearly, this means that when applying (9) in a heritability setting, it would be reasonable to expect that Y_1 conditional on Y_2 should provide the same answers as Y_2 conditional on Y_1. While exchangeability is obviously not the case for general bivariate distributions, we achieve pairwise exchangeability by a corresponding restriction of our parametric models for the bivariate distributions, as described later. When including covariates, the assumption of pairwise exchangeability should apply to the residuals, i.e. the mean-adjusted traits $Y_1 - \beta^t x_1$ and $Y_2 - \beta^t x_2$.

Note that it would suffice to assume that, for all y,

$$\begin{aligned} \tau_1^2 = \mathrm{Var}(Y_1) = \mathrm{Var}(Y_2) = \tau_2^2 =: \sigma, \\ \mathrm{E}(Y_1 \mid Y_2 = y) = \mathrm{E}(Y_2 \mid Y_1 = y) =: \mu(y), \\ \mathrm{Var}(Y_2 \mid Y_1 = y) = \mathrm{Var}(Y_1 \mid Y_2 = y) =: \sigma^2(y), \end{aligned} \tag{10}$$

since this would imply that (9) would be invariant to the direction of conditioning. However, the models presented in this paper all imply full pairwise exchangeability. We do *not*, however, ask for full exchangeability of the multivariate outcome distribution; for instance, a mother-father-child trio would clearly not have the same trivariate distribution as a child-father-mother trio. Nevertheless, the pairwise exchangeability implies that all family members have the same marginal distributions. The appropriateness of the exchangeability assumptions will be addressed in the Discussion.

2.3 Heritability curves

Assuming $\rho(y)$ to be well defined for the joint distribution of the two family members, we are interested in the degree to which the value of $\rho(y)$ can be attributed to heritability on one side, and to environment on the other. In particular, we are interested in knowing how these contributions vary with y.

Definition 1 (Heritability curve for the twin ADE model)**.** *Assume the exchangeability property (10) holds for both MZ and DZ bivariate distributions. Adopting the moment equations* (4)*, we define the heritability curve by*

$$a^2(y) = 4\rho^{(DZ)}(y) - \rho^{(MZ)}(y), \tag{11}$$

where $\rho^{(MZ)}(y)$ and $\rho^{(DZ)}(y)$ are the correlation curves of MZ and DZ twins calculated according to (9). Similarly, (4) allows local versions of the dominance effect

$$d^2(y) = 2\left[(\rho^{(MZ)}(y) - 2\rho^{(DZ)}(y)\right] \tag{12}$$

and residual environment

$$e^2(y) = 1 - \rho^{(MZ)}(y) \tag{13}$$

to be defined.

Note that with Equation (11), a trait value can in principle display a non-linear association within both MZ and DZ twins, but have constant local heritability $a^2(y)$ due to a canceling effect in $4\rho^{(DZ)}(y) - \rho^{(MZ)}(y)$.

We similarly define the heritability curve for family trios by adopting the genetic model described in Section 2.1.2 locally.

Definition 2 (Heritability curve for an ACE model of mother-father-child trios)**.** *Assuming the exchangeability property* (10)*, let* $\rho^{(MC)}(y)$ *and* $\rho^{(FC)}(y)$ *be correlation curves (9) for mother-child and father-child relationships, respectively. The heritability curves* $a^2(y)$*,* $c^2(y)$*, and* $e^2(y)$ *are then given by*

$$a^2(y) = 2\rho^{(FC)}(y) \tag{14}$$

$$c^2(y) = \rho^{(MC)}(y) - \rho^{(FC)}(y) \tag{15}$$

$$e^2(y) = 1 - \rho^{(MC)}(y) - \rho^{(FC)}(y) \tag{16}$$

We next define a parametric class of multivariate densities for family data that can easily be fit by maximum likelihood, allows for non-linear dependence, and admits an analytical expression for the correlation curve (9).

3 Correlation and heritability curves for Gaussian mixtures

Throughout this paper we denote by $\phi_d(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ a d dimensional Gaussian density, evaluated at $\boldsymbol{y} = (y_1, \ldots, y_d)$, and with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We will only use $d = 1, 2, 3$.

Consider the observed trait vector $\boldsymbol{y} = (y_1, y_2)$ for a pair of family members. We assume that it follows a m-component Gaussian mixture with density

$$\sum_{k=1}^{m} p_k \phi_2(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{17}$$

where $\sum_{k=1}^{m} p_k = 1$. The mean and covariance structure of the the kth mixture compo-

nent is taken to be

$$\boldsymbol{\mu}_k = (\mu_k, \mu_k), \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_k^2 & \sigma_k^2\rho_k \\ \sigma_k^2\rho_k & \sigma_k^2 \end{pmatrix}, \tag{18}$$

where $\rho_k \in (-1, 1)$ is the correlation parameter. The components of the mixture are ordered such that $\sigma_1 \leq \cdots \leq \sigma_m$. If $\sigma_q = \sigma_{q+1} = \cdots = \sigma_m$ for some $q < m$, then we order the components in ascending order with respect of the means, i.e. $\mu_q < \cdots < \mu_m$. Note that under the above constraints on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, the exchangeability condition (10) is satisfied. In addition, Y_1 and Y_2 have the same marginal distribution, with marginal density

$$g(y) = \sum_{k=1}^{m} g_k(y) \tag{19}$$

as the sum over the individual (weighted) components $g_k(y) := p_k\phi_1(y; \mu_k, \sigma_k^2)$. The (total) marginal mean, marginal variance, and correlation are given by

$$\mu = \sum_{k=1}^{m} p_k\mu_k, \quad \sigma^2 = \sum_{k=1}^{m} p_k\left[\sigma_k^2 + (\mu_k - \mu)^2\right] \quad \text{and} \quad \rho = \sigma^{-2}\sum_{k=1}^{m} p_k\left[\rho_k\sigma_k^2 + (\mu_k - \mu)^2\right]. \tag{20}$$

We next derive local versions of μ and σ. Let δ be a latent variable with $P(\delta = k) = p_k$, $k = 1, \ldots m$, showing which mixture component is realized. From Bayes' rule, it follows that the distribution of $\delta \mid Y_2 = y$ is given as

$$p_k^*(y) := P(\delta = k \mid Y_2 = y) = \frac{g_k(y)}{g(y)}. \tag{21}$$

Also, by the assumed normality of each mixture component, it follows that

$$\mu_k(y) := \mathrm{E}(Y_1 \mid Y_2 = y, \delta = k) = \mu_k + \rho_k \cdot (y - \mu_k),$$

i.e. $\mu_k(y)$ is a line with slope ρ_k, going through the point (μ_k, μ_k). By the law of total expectation,

$$\begin{aligned} \mu(y) &:= \mathrm{E}(Y_1|Y_2 = y) = \mathrm{E}\left[\mathrm{E}\left(Y_1 \mid Y_2 = y, \delta\right) \mid Y_2 = y\right] \\ &= \sum_{k=1}^{m} p_k^*(y)\mu_k(y). \end{aligned} \tag{22}$$

Similarly, by the law of total variance:

$$\begin{aligned}\sigma^2(y) &:= \mathrm{Var}(Y_1|Y_2 = y) \\ &= \mathrm{E}\left[\mathrm{Var}\left(Y_1 \mid Y_2 = y, \delta = k\right) \mid Y_2 = y\right] \\ &\quad + \mathrm{Var}\left[\mathrm{E}\left(Y_1 \mid Y_2 = y, \delta = k\right) \mid Y_2 = y\right] \\ &= \mathrm{E}\left(\sigma_\delta^2(1-\rho_\delta^2) \mid Y_2 = y\right) + \mathrm{Var}\left(\mu_\delta(y) \mid Y_2 = y\right) \\ &= \sum_{k=1}^{m} p_k^*(y)\left[\sigma_k^2(1-\rho_k^2) + \left[\mu_k(y) - \mu(y)\right]^2\right].\end{aligned} \tag{23}$$

We are now ready to give the expression for $\beta(y) = \mu'(y)$, to be used in the correlation curve (9) for the mixture distribution.

Proposition 1. *Define*

$$d_k(y) := -(y - \mu_k)/\sigma_k^2.$$

Then,

$$\beta(y) = \sum_{k=1}^{m} p_k^*(y)\left[\rho_k + \left(\mu_k(y) - \mu(y)\right) d_k(y)\right], \tag{24}$$

where $p_k^*(y)$ is given by (21).

proof. See Appendix.

Notice that when there is only a single mixture component ($m = 1$), yielding a bivariate Gaussian distribution, the above expressions reduce to $\sigma = \sigma_1$, $\mu(y) = \mu_1$, $\sigma^2(y) = \sigma_1^2(1-\rho_1^2)$ and $\beta(y) = \rho_1$. Inserting these expressions in (9) we get a constant correlation curve, $\rho(y) = \rho_1$ for every y. Hence, if $m = 1$ the heritability curve $a^2(y)$, given by (11) or (14), reduces to the ordinary heritability coefficient a^2.

3.1 Properties of the correlation curve under a Gaussian mixture

It is of interest to investigate the asymptotic behaviour of $\rho(y)$ as $y \to \pm\infty$ under the mixture (17) since this can be used to evaluate the asymptotic behaviour of the heritability curve $a^2(y)$, which in general will depend on the family design. We state the result in the following theorem, which also includes the limit behaviour of $\beta(y)$ and $\sigma^2(y)$.

Intuitively, a one-dimensional mixture distribution is asymptotically dominated in the

tails by the component with the largest variance; if two or more components all share the largest variance, the sizes of the mean values come into play, with the component with the smallest mean value dominating when $y \to -\infty$, and the largest when $y \to +\infty$. While this in itself is fairly obvious, we here use it to develop the resulting asymptotic behavior of $\beta(y)$, $\sigma^2(y)$, and $\rho(y)$.

We consider the following two cases: Recall the ordering $\sigma_1^2 \leq \cdots \leq \sigma_m^2$, and define $q = \min\{l : \sigma_l^2 = \sigma_m^2\}$. We define Case I as $q = m$. For the alternative, Case II, where $q < m$, our conventions is that the mean values are then ordered such that $\mu_q < \mu_m$. To simplify the notation, define the constant K as follows:

$$\begin{aligned} &\text{Case I } (q = m), \quad y \to \pm\infty, \quad K := m, \\ &\text{Case II } (q < m), \quad y \to -\infty, \quad K := q, \\ &\text{Case II } (q < m), \quad y \to +\infty, \quad K := m. \end{aligned}$$

Theorem 3.1.1. *The asymptotic behavior of $\beta(y)$, $\sigma^2(y)$, and $\rho(y)$, given by (24), (23), and (9), are*

$$\begin{aligned} \lim_y \beta(y) &= \rho_K, \\ \lim_y \sigma^2(y) &= \sigma_K^2(1-\rho_K^2), \\ \lim_y \rho(y) &= \tilde{\rho}_K := \frac{\sigma\rho_K}{[\sigma^2\rho_K^2 + \sigma_K^2(1-\rho_K^2)]^{1/2}}. \end{aligned} \tag{25}$$

The global variance σ^2 is defined as in (20).

proof See Appendix.

Theorem 3.1.1 shows that $\beta(y)$, $\sigma^2(y)$, and $\rho(y)$ all stabilize to finite limits as $y \to \pm\infty$, and their behaviour is determined by the variance and correlation of mixture component K, in addition to the global variance σ^2. In Case I we have that the asymptotic correlation is the same in both tails, as exemplified in Figure 4a) where $K = 3$ and $\tilde{\rho}_3 \approx 0.5$.

Identical correlations in both tails may seem unmotivated for family data. Still, within the data range the correlation curve will be determined by all of the mixture components, in accordance with (9), which allows for different behaviour in the tails.

Case II, on the other hand, allows for different asymptotic correlation in the left and right tail, with the differences being the use of ρ_n versus ρ_m in (25).

Theorem 3.1.1 is further illustrated in Figure 4 showing the limiting behaviour of $\beta(y)$,

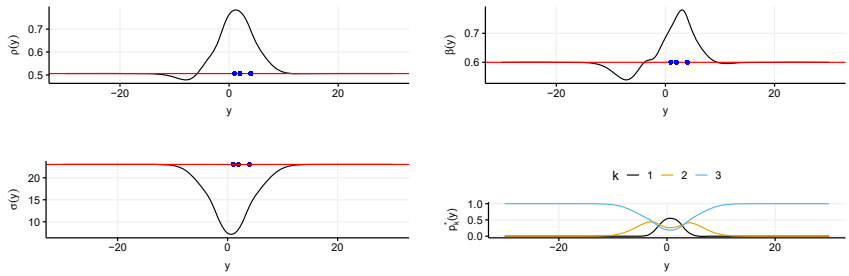


Figure 4: Illustration of the asymptotic tail behaviour (red line) of the correlation curve and its builing blocks under a $m = 3$ component mixture model: (a) $\rho(y)$, (b) $\beta(y)$, (c) $\sigma^2(y)$, and (d) $p_k^*(y)$. The mixture model has parameters $(\sigma_1, \sigma_2, \sigma_3) = (2, 4, 6)$, $(\mu_1, \mu_2, \mu_3) = (1, 2, 4)$, $(\rho_1, \rho_2, \rho_3) = (0.7, 0.8, 0.6)$ and $(p_1, p_2, p_3) = (0.3, 0.3, 0.4)$.

$\sigma^2(y)$, and $\rho(y)$ for a three-component mixture under Case I. Note that the limiting correlation satisfies $\tilde{\rho}_3 < \min(\rho_1, \rho_2, \rho_3)$ for the parameter values used in the figure. This is counter-intuitive because the posterior probability $p_3^*(y)$ approaches 1 in the tails (upper left panel), but still the limiting correlation is not simply ρ_3. The peak in correlation around $\mu_2 = 2$ is reasonable as the second component has the highest ρ.

3.1.1 The case of equal σ_k's

It is worth studying the special case that $\sigma_1 = \sigma_2 = \cdots = \sigma_m$, with their common value denoted by σ_0. This is Case II of Theorem 3.1.1 with $q = 1$. From (20) we get $\sigma^2 = \sigma_0^2 + \sigma_\mu^2$, where

$$\sigma_\mu^2 = \sum_{k=1}^{m} p_k(\mu_k - \mu)^2, \tag{26}$$

which is the variance due to differences in locations of mixture components. Recall the convention that the mixture components are ordered such that $\mu_1 < \mu_2 < \cdots < \mu_m$. We are now ready to state the following corollary to Theorem 3.1.1.

Corollary 3.1.2. *When $\sigma_1 = \cdots = \sigma_m$ the asymptotic behavior of $\rho(y)$, given by (9), is*

$$\lim_{y\to-\infty} \rho(y) = \rho_1\sqrt{\frac{1+\gamma}{1+\gamma\rho_1^2}} \quad \textit{and} \quad \lim_{y\to\infty} \rho(y) = \rho_m\sqrt{\frac{1+\gamma}{1+\gamma\rho_m^2}}, \tag{27}$$

where $\gamma = \sigma_\mu^2/\sigma_0^2$ is the ratio of between and within-component variance in the Gaussian mixture.

The limiting correlations always exceed (in absolute value) ρ_1 and ρ_m, respectively. When $\gamma \to \infty$, i.e. the mixture components gets increasingly spread out, both limits

approach 1 in absolute value.

3.2 Estimation

In this section we explain how to fit Gaussian mixtures to family data. On one hand, they are fully parametric distributions, which can be exploited in estimation and inference. On the other hand, allowing the number of mixture components m to grow, mixtures become increasingly flexible, which allows us to view them also as nonparametric tools. In particular, Gaussian mixtures seem well suited to model small perturbations from Gaussianity.

First, let $\boldsymbol{y} = (y_1, y_2, y_3)$ denote the trait vector for the mother-father-child trio, which is assumed to have the following mixture density:

$$\sum_{k=1}^{m} p_k \phi_3(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,.$$

Here $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ are structured in the following way:

$$\boldsymbol{\mu}_k = (\mu_k, \mu_k, \mu_k), \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_k^2 & \sigma_k^2\rho_k^{(MF)} & \sigma_k^2\rho_k^{(MC)} \\ \sigma_k^2\rho_k^{(MF)} & \sigma_k^2 & \sigma^2\rho_k^{(FC)} \\ \sigma_k^2\rho_k^{(MC)} & \sigma_k^2\rho_k^{(FC)} & \sigma_k^2 \end{pmatrix}, \tag{28}$$

where we use superscripts on the ρ's to denote relationship. Integrating the above joint density with respect to any one of the three family members (y_1, y_2, or y_3) will result in the bivariate Gaussian mixture (17) from which we defined the correlation curve. The reason for performing joint estimation, rather than pairwise, is to optimally utilize the information contained in mother-father-child trios. Note that the three marginals are identical by construction, although the joint distribution is not exchangeable unless $\rho_k^{(MF)} = \rho_k^{(MC)} = \rho_k^{(FC)}$ for $k = 1, \ldots, m$.

Given n such trios, the parameters (μ_k, σ_k, ρ_k, p_k) can be estimated by maximizing the following log-likelihood:

$$\log L = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{m} p_k \phi_3(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \Sigma_k) \right]. \tag{29}$$

Once the parameters are estimated, the heritability curve $a^2(y)$ can be obtained via the correlation curves as described in Definition 2.

For twins, consider first a dizygotic pair with trait vector $\boldsymbol{y} = (y_1, y_2)$. The likelihood

contribution from $n^{(MZ)}$ such pairs is:

$$\log L^{(MZ)} = \sum_{i=1}^{n^{(MZ)}} \log \sum_{k=1}^{m} p_k \phi_2(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{30}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are structured as in (18). The likelihood contribution of $n^{(DZ)}$ dizygotic twin pairs, $\log L^{(DZ)}$, is defined analogously using the same number m of mixture components. The only parameters that differ between the MZ and DZ cases are the correlation parameters ρ_k in (18). The fact that p_k, μ_k, and σ_k are shared across the MZ and DZ mixtures, calls for using a combined log-likelihood $\log L = \log L^{(MZ)} + \log L^{(DZ)}$. Once the parameters are estimated, the heritability curve $a^2(y)$ can be obtained via the correlation curves as described in Definition 1.

Both of the log-likelihoods (29) and (30) will be maximized using the R-package TMB [Kristensen et al., 2016]. In TMB the (negative) log-likelihood is implemented as a C++ function, which is compiled and linked into the R session, where the standard function minimizer `nlminb` is employed. In addition, TMB calculates the gradient and Hessian (1st and 2nd order derivatives) of the log-likelihood by Automatic Differentiation [Kristensen et al., 2016]. Such derivative information can substantially speed up the minimizer and make it more robust. Finally, TMB uses derivatives to calculate the approximate standard deviation of any interest quantity, as a function of the parameters, using the delta method. This feature of TMB will be used to estimate pointwise

m	no. of parameters	AIC	BIC
1	4	259.4	227.6
2	9	8.0	0
3	14	2.8	18.5
4	19	6.5	46.0
5	24	0	63.3

Table 1: Model comparison for the twin BMI data, where m is the number of mixture components and $5m-1$ is the number of parameters in the model. AIC and BIC values are relative to the best fitting models (respectively, $m = 5$ and $m = 2$).

Parameters	$k = 1$	$k = 2$	Global
μ_k	21.20	22.20	21.39
σ_k	0.63	1.26	0.88
$\rho_k^{(MZ)}$	0.75	0.70	0.78
$\rho_k^{(DZ)}$	0.28	−0.04	0.30
p_k	0.81	0.19	

Table 2: Parameter estimates for the chosen Gaussian mixture ($m = 2$) for the twin data. The mixture components are ordered according to the value of σ_k. The global quantities, μ, σ, $\rho^{(MZ)}$ and $\rho^{(DZ)}$ are calculated from (20).

confidence intervals of correlation and heritability curves.

For the purpose of selecting the number of mixture components, m, we calculate both of the criteria AIC $= -2\log(L) + 2Q$ and BIC $= -2\log(L) + \log(n)Q$ for each candidate model, where Q is the number of parameters and $\log(L)$ is obtained either from (29) or (30). Contributing to Q is the total number of p_k's, μ_k's, σ_k's, and ρ_k's, but due to the constraint $\sum_{k=1}^{m} p_k = 1$ there are only $m-1$ free p_k's. Hence, for the trio likelihood (29) we have $Q = 6m-1$, while for the twin likelihood (30), with different ρ_k for MZ and DZ twins, we have $Q = 5m-1$. It is clear that for $\log(n) > 2$, BIC will be more conservative than AIC, in the sense of favoring smaller values of m. As will be shown below, the correlation curve tends to be more unstable (fluctuating) for larger values of m. For this reason we will use BIC as our model selection criterion, but we will still report AIC as a comparison.

4 Applications

4.1 BMI of twins

We use the "twinData" dataset found in the R-package "OpenMx" [Neale et al., 2016]. As our response, we take BMI measurements (around age 18) for $n^{(MZ)} = 534$ monozygotic and $n^{(DZ)} = 328$ dizygotic female-female twin pairs. Table 1 compares models in the range $1 \leq m \leq 5$, and it is seen that the pure bivariate Gaussian model ($m = 1$) fits considerably worse than any of the mixture models ($m > 1$). The lowest AIC and BIC values occur for $m = 5$ and $m = 2$, respectively, but it is seen that AIC is almost indecisive between models with $m > 1$. Due to its heavier penalization, $\log\left(n^{(MZ)} + n^{(DZ)}\right) = \log(862) = 6.8$, of the number of parameters, BIC more clearly favours $m = 2$. According to our decision to base model selection on BIC, we choose the model with $m = 2$.

Table 2 shows the parameter estimates. The first mixture component is dominating with $p_1 = 0.81$. For MZ twins there is high correlation (ρ_k) within in each of the two components, while for DZ twins ρ_2 is close to zero. The (global) correlations for the mixtures as a whole, matches exactly the empirical Pearson correlations, which are 0.78 (MZ) and 0.30 (DZ), respectively.

Figure 5 displays the estimated correlation curve for both MZ and DZ twins, using the parameter values from Table 2. Also shown are 95% confidence intervals calculated using the delta method. Both correlation curves are fairly flat within the center 90%

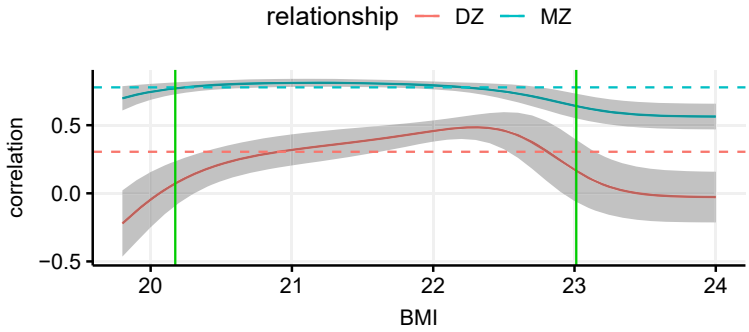


Figure 5: Estimated monozygotic (MZ) and dizygotic (DZ) twins correlation curves for the BMI data, with pointwise 95% confidence intervals (in grey). The dashed lines display the (overall) Pearson correlation within MZ and DZ twin pairs, respectively. The vertical green lines represent the 0.05 and 0.95 quantiles of the data.

data range (represented by the two vertical green bars), while they both drop for low and high BMI. This yields (Figure 6) an estimated heritability curve $a^2(y)$ that does not differ significantly (except maybe around $y = 22.3$) from the classical heritability coefficient (4).

The TMB (R and C++) code used to produce the parameter estimates in Table 2 plots in Figure 6 is available from `https://github.com/skaug/Supplementary`.

4.2 Birth weight of family trios

To illustrate the family trio analyses, we used birth weights of $n = 81,144$ complete mother–father–child trios. The data originally derived from the Medical Birth Registry of Norway, where the birth weight variables were added some random noise and rounded off to guarantee anonymity. The same data with some additional restrictions on parity, plurality, etc. were previously described and analyzed elsewhere [Magnus et al., 2001]. The data were restricted to all births (mother, father, and child) taking place within the years 1967–1998. Due to Norwegian ethical and legal restrictions, Norwegian data used in this study are available upon request to the Medical Birth Registry of Norway, the Norwegian Institute of Public Health. URL: https://www.fhi.no/hn/helseregistre-og-registre/mfr. Requests for data access can be directed to Datatilgang@fhi.no¡mailto:Datatilgang@fhi.no¿.

We did not have information about the gender of the child; hence, we performed a

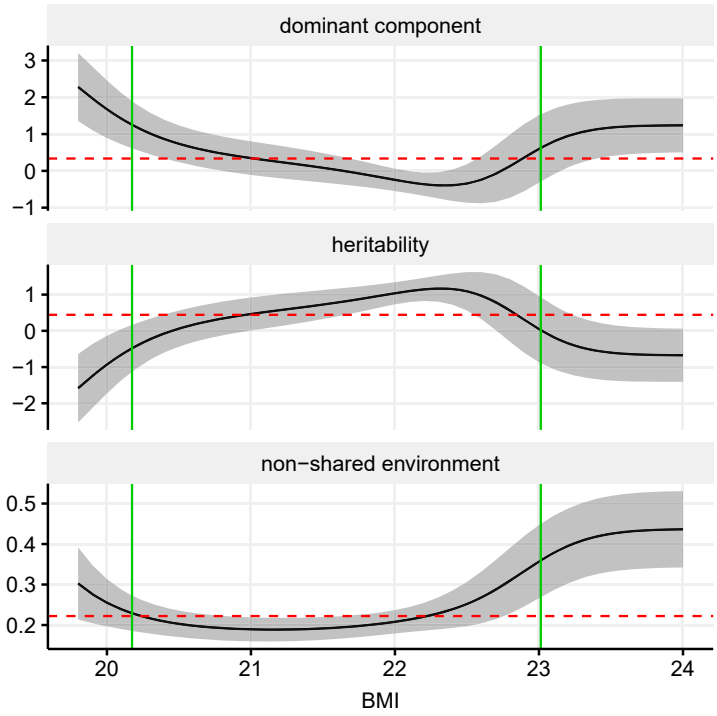


Figure 6: Estimated dominant genetic component $d^2(y)$, heritability curve $a^2(y)$, and environment curve $c^2(y)$ for the BMI data under the ADE model (Definition 1), with pointwise 95% confidence intervals (in grey). The red dashed lines display the classical estimates of dominant component, heritability, and environment, given by (4). The vertical green lines represent the 0.05 and 0.95 quantile in data.

standardization of the data. We assumed a 50% sex ratio in the offspring, and introduced the quantity $D \triangleq \frac{1}{2}(\bar{y}_M - \bar{y}_F)$, where $\bar{y}_M$ is the mean of the birth weights of mothers, and $\bar{y}_F$ is the mean of the birth weights of fathers. We hence added D to the father's weight and subtracted it to the mother's weight; in this way, the average among mothers and fathers is the same, and close (25g deviation) to the average in the offspring. This standardization is of little consequence to the end result.

Figure 1 summarizes the marginal and bivariate properties of the data. The marginal distributions are close to a Gaussian shape, but the left tail of the child birth weights is slightly heavier than the right tail. As suggested in the Introduction, this may be indicative of strong but rare factors dominating in producing the lowest birth weigths, which is what we will confirm in our analyses of local heritability below.

m	no. parameters	Δ AIC	Δ BIC
1	5	14848	14749
2	11	1148	904.4
3	17	480.4	292.5
4	23	132.1	0
5	29	109.7	33.5
6	35	36.3	16.0
7	41	0	35.5

Table 3: Model comparison for family trios, where m is the number of mixture components. The total number of (free) parameters is $6m - 1$, counting all p_k, μ_k, σ_k, $\rho_k^{(MC)}$, $\rho_k^{(FC)}$ and $\rho_k^{(MF)}$. AIC and BIC values are relative to the lowest one, represented in red.

Parameters	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Global
μ_k	3516	3687	3093	2243	3493
σ_k	440.5	572.9	690.5	1116	555.0
$\rho_k^{(MC)}$	0.240	0.143	-0.189	-0.826	0.123
$\rho_k^{(FC)}$	0.134	0.053	-0.254	-0.845	0.201
$\rho_k^{(MF)}$	-0.011	-0.084	-0.289	0.750	0.068
p_k	0.636	0.231	0.126	0.007	

Table 4: Parameter estimates and standard deviations for the Gaussian mixture ($m = 4$) fit to the mother–father–child trios. The mixture components are ordered according to the value of σ_k. The global quantities, μ, σ, $\rho^{(MC)}$, $\rho^{(FC)}$ and $\rho^{(MF)}$ are calculated from (20).

The scatter plots are roughly symmetric around the identity line, which is consistent with the exchangeability assumption made in Section 2.2. It should be noted, however, that the left hand tail of the marginal distributions is somewhat heavier in the children than in the parents; this is likely because parents are selected by the fact that they have children; it is known that individuals born with low birth weight have somewhat reduced fertility later in life. We have, however, not taken this into consideration in our model.

From the non-parametric regression (blue curve), it is clear that there is no association between mother and father, which is reflected in the low Pearson correlation of 0.0209. For the two relationships involving the child, the non-parametric regression curve indicates a non-linear relationship, particularly for mother-child. For birth weights less than 3000g there seems to be a low association, while for larger birth weights the association is increasing.

The Gaussian mixture (17) was fit by maximum likelihood for $m = 1, \ldots, 7$. We computed both AIC and BIC values for this model. According to the BIC criterion, the

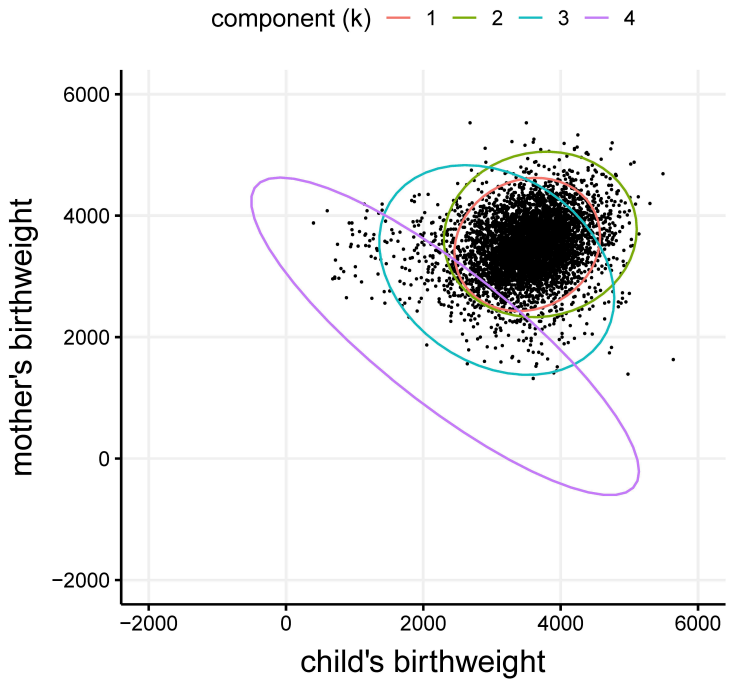


Figure 7: Birth weight (gram) of a random subset of 5000 mother-child pairs taken from Figure 1. Also shown are 95% level curves (ellipses) for each of the $m = 4$ mixture components in Table 4, i.e. each elipse include 95% of the probability mass for that bivariate normal component.

best fitting mixture has $m = 4$ components (see Table 3). Parameters estimates for this model are given in Table 4. Figure 7 shows the underlying mother-child pairs, overlaid by the five mixture components.

The mother-child distribution is pear-shaped relative to a bivariate normal distribution, with more spread around the identity line ($y_1 = y_2$) for small birth weights. The mixture model adapts to this shape by assigning negative ρ_k's to its two components ($k = 3, 4$) with the smallest μ_k. The remaining two components ($k = 1, 2$), which together constitute 87% of the probability mass, form a bivariate distribution that is hard to distinguish visually from a Gaussian distribution. The estimates of global correlation for the mixture in Table 3, closely match the corresponding empirical Pearson correlations given in Figure 1 for MC, FC and MF pairs. It is seen to fit the empirical marginals fairly well, and to posses a heavier left hand tail.

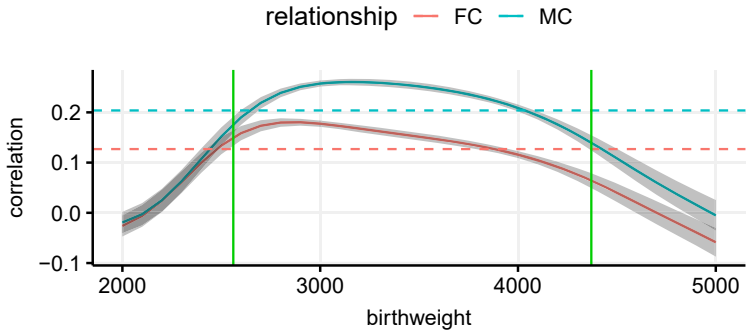


Figure 8: Estimated mother-child (MC) and father-child (FC) correlation curves for the Norwegian Birth Registry data, with pointwise 95% confidence intervals (in grey). The dashed lines display the (overall) Pearson correlation within MC and FC pairs, respectively. The vertical green lines represent the 0.05 and 0.95 quantiles of the data.

Figure 8 shows the two estimated correlation curves $\rho^{(FC)}(y)$ and $\rho^{(MC)}(y)$, which are the components going into $a^2(y)$, $c^2(y)$, and $e^2(y)$, given respectively by (14)–(16). Overall, the Pearson correlation and the correlation curves for MF exceed those for FC. Both curves exceed their respective Pearson correlations in the center of the data, while they decrease for both low and high birth weights. The FC curve has its maximum somewhat to the left of the maximum of the MC curve. As a robustness check, we also computed the local Gaussian correlations [Tjøstheim and Hufthammer, 2013] between mother and child as displayed in Figure 9. These exhibit the same behaviour as the correlation curve; large values in the center of the data which are decreasing towards both tails. Figure 10 shows heritability and environment curves. The overall conclusion is that variation in birth weight is mostly attributable to environment, which was also seen in previous publications [Magnus et al., 2001, Lunde et al., 2007, Gjessing and Lie, 2008], and is reflected in the classical measures of heritability $a^2 = 0.246$ and environment $c^2 = 0.754$, and the variation in the corresponding curves.

Recall that, under the assumed model (7) the heritability curve $a^2(y)$ is completely determined by the FC correlation curve $\rho^{(FC)}(y)$. Since the FC correlation curve exceeds the Pearson FC correlation in the center of the data, the heritability curve also exceeds the classical heritability measure in the same region.

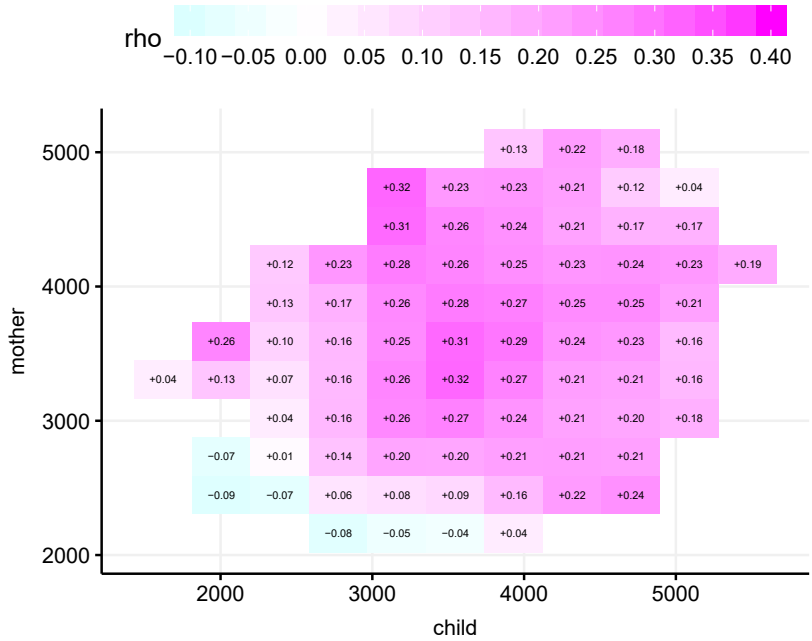


Figure 9: Estimated local Gaussian correlation between mother and child. Note that this correlation measure has two location arguments (y_1 and y_2).

5 Discussion and conclusion

We have provided closed-form expressions for the correlation curve for exchangeable bivariate Gaussian mixtures. To our knowledge, this result is new and should be useful generally in situations where exchangeability can be assumed. Since differences in mean values may accounted for using a linear predictor like (5), it is only exchangeability of the residuals, or the weaker condition (10), that is required. In the context of our family data, the exchangeability assumption is rather reasonable for twin data. In nuclear families, it is less obvious that parents and children have the exact same marginal distribution even when using covariates to adjust for systematic generational differences. With our generational birth weight data, we observe that the left hand tail in the parental distribution is smaller than among the children. As discussed in Subsection 4.2, this may well be a selection phenomenon; somebody born with a very low birth weight is less likely to become a parent, and are thus possibly under-represented in our data file. For instance, increased mortality among the smallest newborns is thought to lead to a selection pressure on the birth weight distribution over generations [Cavalli-Sforza and

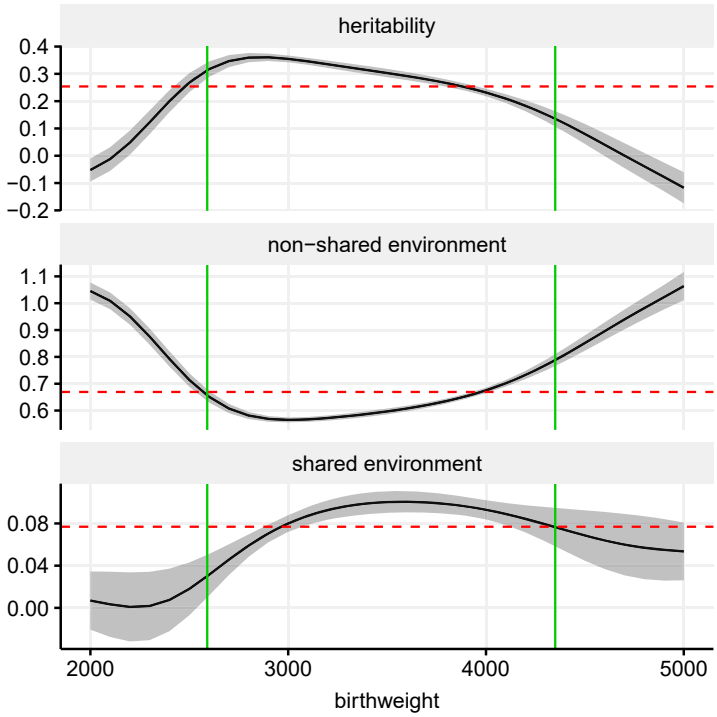


Figure 10: Estimated heritability curve $a^2(y)$, environment curve $c^2(y)$, and residual environment $e^2(y)$ for the Norwegian Birth Registry data under the ACE model (Definition 2), with pointwise 95% confidence intervals (in grey). The red dashed lines display the classical estimates of heritability and environment, i.e. empirical versions of (7). The vertical green lines represent the 0.05 and 0.95 quantiles of the data.

Bodmer, 1999].

A restriction of our model is that we have applied it only in situations with simple family structures where moment estimators of the heritability are explicit. In larger family structures, several pairwise relationships may provide information about the same heritability parameters. For instance, family trios with sibling data add the sibling correlation as a source of information [Lunde et al., 2007]. We will not discuss that issue further, but note that if pairwise correlation curves are estimated from larger data structures, weighted least squares estimation may provide a way of combining them into a common estimate of heritability curves [Gjessing and Lie, 2008].

In our twin BMI example, we chose the ADE model for the estimation since for the estimated overall correlations, $\rho^{(MZ)} > 2\rho^{(DZ)}$. However, as seen in Figure 6, there

are values for y (the BMI) where the estimated $d^2(y)$ drops below zero. This indicates that in this region, the ACE model might be more appropriate. Note that there is no difficulty in letting the local heritability curves switch from an ADE model to an ACE model locally. In particular, we see that when $\rho^{(MZ)} = 2\rho^{(DZ)}$, both (3) and (4) provide the same estimates for a^2 and e^2, and both c^2 and d^2 are estimated as zero. The estimated heritability curves would thus still be continuous if switching from one model to another.

The choice of Gaussian mixtures was made due to their flexibility, in the spirit of non-parametric estimation. Our approach is pragmatic in the sense that we have not attempted to interpret individual mixture components as sub-populations. One reason for this is the negative estimates for some of the ρ_k seen in both Table 2 and 4, which would be hard to interpret biologically.

On the other hand, Gaussian mixtures are fully parametric models, which allows us to use the standard parametric toolbox. For instance, covariates can easily enter the mean, as in (5), and it would also be straight forward to formulate model in which the σ_k were affect by family level covariates. A further benefit of having a parametric model is that we can select model complexity (m) based on standard AIC or BIC criteria.

The parametric structure is also the basis for the results about the tail behaviour of the correlation curve in Theorem 3.1.1. While the center of the distribution may have sufficient data to allow stable non-parametric estimation of the heritability, the estimates in the tails are more dependent on the model structure. This is both a strength and a weakness of the mixture model. The heritability curves converge to constant values in the tails, which makes the estimates more stable; on the other hand, those estimates depend on the dominant mixture components in the tails, and the number and placement of mixture components may not always be clear cut.

There are also well known problems with Gaussian mixtures. Among these are local maxima on the likelihood surface [Baudry and Celeux, 2015], which can be explored by using different initial values for the numerical optimization. We avoided the classical "label switching" problem by constraining the parameters of the mixture (σ's and μ's), but have nevertheless observed some sensitivity of the parameter estimates in Table 4. Although we cannot guarantee that we have found the global optimum of the likelihood surface, the choice of model complexity (m) seems to be robust to the choice of initial values. Similarly, the shape of the correlation curves (and consequently heritability and environment curves) are quite stable. A related problem is that of singlularity of the Fisher information matrix which can occur for mixture models [Drton and Plummer, 2017]. This could potentially affect the validity of AIC and BIC criteria, as well as

the standard deviations based on the observed Fisher information that have been used throughout this paper. Such standard deviations are produced automatically by TMB, and are very convenient in an exploratory phase, but we recommend that they are validated by simulation (parametric bootstrap).

6 Acknowledgements

This research was supported by Research Council of Norway grant 225912/F50 “Health Registries for Research” and the Centres of Excellence funding scheme (Grant 262700).

A Proofs

Proof of Proposition 1 Let $g(y)$, $g_k(y)$, $p_k^*(y)$ etc. be defined as in Section 3. First, note that

$$\frac{g_k'(y)}{g_k(y)} = d_k(y).$$

Furthermore, define

$$d(y) := \sum_{i=1}^{m} p_i^*(y) d_i(y),$$

i.e. the weighted average of the $d_i(y)$'s. Then

$$\frac{g'(y)}{g(y)} = \frac{\sum_{i=1}^{m} d_i(y) g_i(y)}{g(y)} = d(y).$$

For any fraction $s(y) = a(y)/b(y)$ of differentiable functions, note that the chain rule can be written as $\frac{s'(y)}{s(y)} = \frac{a'(y)}{a(y)} - \frac{b'(y)}{b(y)}$. Thus,

$$\frac{p_k^{*'}(y)}{p_k^*(y)} = \frac{g_k'(y)}{g_k(y)} - \frac{g'(y)}{g(y)} = d_k(y) - d(y).$$

Recall from (22) that $\mu(y) = \mathrm{E}\left[Y_1 \mid Y_2 = y\right] = \sum_{i=1}^{m} p_i^*(y)\mu_i(y)$ is the conditional expectation,

$$\begin{aligned}
\beta(y) = \mu'(y) &= \sum_{i=1}^{m} \left(p_i^*(y)\mu_i'(y) + p_i^{*'}(y)\mu_i(y) \right) \\
&= \sum_{i=1}^{m} p_i^*(y) \left(\rho_i + \mu_i(y) \left(d_i(y) - d(y) \right) \right) \\
&= \sum_{i=1}^{m} p_i^*(y) \left(\rho_i + \left(\mu_i(y) - \mu(y) \right) \left(d_i(y) - d(y) \right) \right) \\
&= \sum_{i=1}^{m} p_i^*(y) \left(\rho_i + \left(\mu_i(y) - \mu(y) \right) d_i(y) \right),
\end{aligned}$$

where we make use of $\sum_{i=1}^{m} p_i^*(y) \left(d_i(y) - d(y) \right) = 0$ and $\sum_{i=1}^{m} p_i^*(y) \left(\mu_i(y) - \mu(y) \right) = 0$.

A.0.1 Proof of Theorem 3.1.1 - asymptotic behavior of $\beta(y)$, $\sigma^2(y)$, and $\rho(y)$

For two functions $a(y)$ and $b(y)$, as $y \to \infty$ (or $-\infty$), we use the standard notation that $a(y) \sim b(y)$ means $\lim_{y\to\infty} a(y)/b(y) = 1$, and $a(y) \ll b(y)$ means $\lim_{y\to\infty} a(y)/b(y) = 0$. Our proofs below follow mostly from standard theory on asymptotic behavior of real functions[Bender and Orszag, 2013].

Asymptotic behavior of mixture components For one mixture component $g_k(y)$, the asymptotic behavior when $y \to \pm\infty$ is

$$g_k(y) \sim C_k \exp\left(\frac{\mu_k}{\sigma_k} y - \frac{1}{2\sigma_k^2} y^2\right),$$

for a constant C_k. Comparing two components $g_k(y)$ and $g_l(y)$ with $\sigma_k^2 < \sigma_l^2$, we clearly have

$$g_k(y) \ll g_l(y) \quad \text{as} \quad y \to \pm\infty \tag{31}$$

since the y^2-term dominates the asymptotics. If $\sigma_k^2 = \sigma_l^2$, assume that $\mu_k < \mu_l$. Then

$$g_k(y) \ll g_l(y) \quad \text{as} \quad y \to +\infty, \tag{32}$$

and

$$g_l(y) \ll g_k(y) \quad \text{as} \quad y \to -\infty. \tag{33}$$

Let $a_k(y)$ be non-zero polynomial functions in y for $k = 1, \ldots, m$. Since polynomials are asymptotically dominated by exponentials of polynomials, the products $g_k(y)a_k(y)$ are asymptotically ordered in the same way as in (31), (32), and (33) above.

Asymptotic behavior of mixtures Recall the definition of K in Theorem 3.1.1. The results above apply directly to the sum $\sum_{k=1}^m g_k(y)a_k(y)$, which will asymptotically follow the dominant term with $k = K$. I.e.,

$$\sum_{k=1}^{m} g_k(y)a_k(y) \sim g_K(y)a_K(y).$$

In particular, for the full density we get

$$g(y) = \sum_{i=1}^{m} g_i(y) \sim g_K(y).$$

Similarly, if $k \neq K$,

$$p_k^*(y)a_k(y) = \frac{g_k(y)a_k(y)}{g(y)} \to 0, \tag{34}$$

and

$$p_K^*(y)a_K(y) \sim a_K(y).$$

Conditional mean $\mu(y)$ Applying the above results to μ, we obtain

$$\mu(y) = \sum_{k=1}^{m} p_k^*(y)\mu_k(y) \sim \mu_K(y) \sim \rho_K \cdot y.$$

Furthermore, letting $a_k(y) = \rho_k + (\mu_k(y) - \mu(y))\, d_k(y)$, we get

$$\beta(y) = \sum_{k=1}^{m} p_k^*(y) a_k(y) \sim a_K(y).$$

However, by 34,

$$(\mu_K(y) - \mu(y))\, d_K(y) = \sum_{k=1}^{m} p_k^*(y)(\mu_K(y) - \mu_k(y)) d_K(y) \to 0$$

since the K'th term vanishes. It follows that

$$\beta(y) \sim a_K(y) \to \rho_K.$$

Conditional variance $\sigma^2(y)$ For the conditional variance,

$$\begin{aligned}\sigma^2(y) &= \sum_{k=1}^{m} p_k^*(y) \left[\sigma_k^2(1-\rho_k^2) + [\mu_k(y) - \mu(y)]^2\right] \\ &\sim \sigma_K^2(1-\rho_K^2).\end{aligned}$$

Correlation curve $\rho(y)$ Finally, the result for the correlation curve $\rho(y)$ follows directly from the results for $\sigma^2(y)$ and $\beta(y)$.

References

J.-P. Baudry and G. Celeux. Em for mixtures. *Statistics and computing*, 25(4):713–726, 2015.

C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer Science & Business Media, Mar. 2013. ISBN 978-1-4757-3069-2. Google-Books-ID: xz0mBQAAQBAJ.

S. Bjerve and K. Doksum. Correlation curves: Measures of association as functions of covariate values. *Ann. Statist.*, 21(2):890–902, 06 1993. doi: 10.1214/aos/1176349156. URL `https://doi.org/10.1214/aos/1176349156`.

M. G. Bulmer. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, 1985. ISBN 978-0-19-857633-4.

L. L. Cavalli-Sforza and W. F. Bodmer. *The Genetics of Human Populations*, pages 612–614. Courier Corporation, Jan. 1999. ISBN 978-0-486-40693-0.

S. Cherny, L. Cardon, D. W. Fulker, and J. DeFries. Differential heritability across levels of cognitive ability. *Behavior genetics*, 22(2):153–162, 1992a.

S. S. Cherny, L. R. Cardon, D. W. Fulker, and J. C. DeFries. Differential heritability across levels of cognitive ability. *Behavior Genetics*, 22(2):153–162, 1992b. URL `http://www.springerlink.com/index/U04615LJ476532V6.pdf`.

J. C. DeFries and D. W. Fulker. Multiple regression analysis of twin data. *Behavior genetics*, 15(5):467–473, 1985.

J. C. DeFries and D. W. Fulker. Multiple regression analysis of twin data: Etiology of deviant scores versus individual differences. *Acta Geneticae Medicae et Gemellologiae: Twin Research*, 37(3-4):205–216, 1988.

K. Doksum, S. Blyth, E. Bradlow, X.-L. Meng, and H. Zhao. Correlation Curves as Local Measures of Variance Explained by Regression. *Journal of the American Statistical Association*, 89(426):571–582, June 1994. ISSN 0162-1459. doi: 10.1080/01621459.1994.10476782. URL `https://doi.org/10.1080/01621459.1994.10476782`.

M. Drton and M. Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380, 2017.

D. S. Falconer. *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London, 1960.

R. A. Fisher. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 52(2):399–433, 1919. ISSN 2053-5945, 0080-4568. doi: 10.1017/S0080456800012163. Publisher: Royal Society of Edinburgh Scotland Foundation.

H. K. Gjessing and R. T. Lie. Biometrical modelling in genetics: are complex traits too complex? *Statistical Methods in Medical Research*, 17(1):75–96, 2008. doi: 10.1177/0962280207081241. URL `https://doi.org/10.1177/0962280207081241`. PMID: 17855744.

P. W. Holland and Y. J. Wang. Dependence function for continuous bivariate densities. *Communications in Statistics - Theory and Methods*, 16(3):863–876, 1987. doi: 10.1080/03610928708829408. URL `https://doi.org/10.1080/03610928708829408`.

J. L. Hopper. Heritability. In R. Elston, J. Olson, and L. Palmer, editors, *Biostatistical Genetics and Genetic Epidemiology*, Wiley reference series in biostatistics, pages 371–372. Wiley, West Sussex, UK, 2002.

J. L. Hopper and P. M. Visscher. Genetic Correlations and Covariances. In R. Elston, J. Olson, and L. Palmer, editors, *Biostatistical Genetics and Genetic Epidemiology*, Wiley reference series in biostatistics, pages 327–331. Wiley, West Sussex, UK, 2002.

M. J. Khoury, T. H. Beaty, and B. H. Cohen. *Fundamentals of Genetic Epidemiology*. Oxford University Press, 1993. ISBN 978-0-19-505288-6.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, 2016.

M. C. LaBuda, J. DeFries, D. W. Fulker, and D. Rao. Multiple regression analysis of twin data obtained from selected samples. *Genetic epidemiology*, 3(6):425–433, 1986.

J. A. Logan, S. A. Petrill, S. A. Hart, C. Schatschneider, L. A. Thompson, K. Deater-Deckard, L. S. DeThorne, and C. Bartlett. Heritability across the distribution: An application of quantile regression. *Behavior genetics*, 42(2):256–267, 2012a.

J. A. Logan, S. A. Petrill, S. A. Hart, C. Schatschneider, L. A. Thompson, K. Deater-Deckard, L. S. DeThorne, and C. Bartlett. Heritability Across the Distribution: An Application of Quantile Regression. *Behavior genetics*, 42(2):256–267, 2012b. URL `http://www.springerlink.com/index/61N4NQ7R15X08544.pdf`.

A. Lunde, K. K. Melve, H. K. Gjessing, R. Skjærven, and L. M. Irgens. Genetic and Environmental Influences on Birth Weight, Birth Length, Head Circumference, and Gestational Age by Use of Population-based Parent-Offspring Data. *American Journal of Epidemiology*, 165(7):734–741, Apr. 2007. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwk107. URL `http://aje.oxfordjournals.org/content/165/7/734`.

P. Magnus, H. K. Gjessing, A. Skrondal, and R. Skjærven. Paternal contribution to birth weight. *Journal of Epidemiology & Community Health*, 55(12):873–877, 2001. ISSN 0143-005X. doi: 10.1136/jech.55.12.873. URL `http://jech.bmj.com/content/55/12/873`.

C. E. McCulloch and J. M. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

G. McLachlan and D. Peel. *Finite mixture models*. Wiley New York, 2000.

M. C. Neale. Twin Analysis. In R. Elston, J. Olson, and L. Palmer, editors, *Biostatistical Genetics and Genetic Epidemiology*, Wiley reference series in biostatistics, pages 206–217. Wiley, West Sussex, UK, 2002.

M. C. Neale, M. D. Hunter, J. N. Pritikin, M. Zahery, T. R. Brick, R. M. Kirkpatrick, R. Estabrook, T. C. Bates, H. H. Maes, and S. M. Boker. Openmx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, 81(2):535–549, 2016.

Y. Pawitan, M. Reilly, E. Nilsson, S. Cnattingius, and P. Lichtenstein. Estimation of genetic and environmental factors for binary traits using family data. *Statistics in Medicine*, 23:449–465, 2004.

S. Rabe-Hesketh, A. Skrondal, and H. Gjessing. Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, 64(1):280–288, 2008.

D. Tjøstheim and K. O. Hufthammer. Local gaussian correlation: A new measure of dependence. *Journal of Econometrics*, 172(1):33 – 48, 2013. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2012.08.001. URL `http://www.sciencedirect.com/science/article/pii/S0304407612001741`.

S. Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):320–332, 1920. Publisher: National Academy of Sciences.

S. Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

Article II

6.2 The heritability of BMI varies across the range of BMI - a heritability curve analysis in a twin cohort

Francesca Azzolini, Geir D. Berentsen, Hans J. Skaug, Jacob v.B. Hjelmborg, Jaakko A. Kaprio

International Journal of Obesity, **46.10**, 1786-1791 (2022)

The heritability of BMI varies across the range of BMI – a heritability curve analysis in a twin cohort

Francesca Azzolini[1], Geir D. Berentsen[2], Hans J. Skaug[3], Jacob v.B. Hjelmborg[4], and Jaakko A. Kaprio[5]

[1]Department of Mathematics, University of Bergen, Bergen,Norway.

Email address: francesca.azzolini.@uib.no

Address: Allégaten 41, 5007 Bergen, Norway

Phone number: 0047 40236764

[2]Department of Business and Management Science, NHH Norwegian School of Economics, Bergen, Norway

[3]Department of Mathematics, University of Bergen, Bergen,Norway.

[4]Department of Public Health, University of Southern Denmark, Odense, Denmark

[5]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

Competing interests: All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Abstract

The heritability of traits such as body mass index (BMI), a measure of obesity, is generally estimated using family, twin, and increasingly by molecular genetic approaches. These studies generally assume that genetic effects are uniform across all trait values, yet there is emerging evidence that this may not always be the case. This paper analyzes twin data using a recently developed measure of heritability called the heritability curve. Under the assumption that trait values in twin pairs are governed by a flexible Gaussian mixture distribution, heritability curves may vary across trait values. The data consist of repeated measures of BMI on 1506 monozygotic (MZ) and 2843 like-sexed dizygotic (DZ) adult twin pairs, gathered from multiple surveys in older Finnish Twin Cohorts. The heritability curve and

BMI value-specific MZ and DZ pairwise correlations were estimated, and these varied across the range of BMI. MZ correlations were highest at BMI values from 21 to 24, with a stronger decrease for women than for men at higher values. Models with additive and dominance effects fit best at low and high BMI values, while models with additive genetic and common environmental effects fit best in the normal range of BMI. Thus, we demonstrate that twin and molecular genetic studies need to consider how genetic effects vary across trait values. Such variation may reconcile findings of traits with high heritabilities and major differences in mean values between countries or over time.

1 Introduction

Twin and family studies of humans have provided evidence for genetic influences on anthropometric measures. One of the most studied phenotypes has been relative weight, the degree to which an individual is lean, of normal weight or has excess weight relative to their height. Body mass index (BMI), weight divided by height squared, is the most used measure in research due to the ease of its assessment and because among adults it is at most weakly correlated with height [Benn, 1971]. As excess weight is also associated with risk of cardiovascular and metabolic diseases [Must et al., 1999], BMI is also in widespread clinical use.

Early meta-analyses on twins based on published summary data have shown that the heritability of BMI is generally high. The estimates based on twins are consistent with the patterns of resemblance of other first-degree relationships [Maes et al., 1997, Elks et al., 2012]. These studies indicated that there is relatively little variation over age, but non-genetic familial influences seen in childhood and adolescence are largely absent in adults [Nan et al., 2012, Silventoinen et al., 2010]. By pooling individual data on height and weight from twin studies across the globe on over 140,000 twin pairs, Silventoinen et al. [2017] show that heritability of BMI decreases from young adulthood to old age, with relatively little differences by region or calendar time. Using the same resource with data from 87,782 twin pairs under the age of 20, Silventoinen et al. [2016] show that heritability of BMI was lowest in early childhood. Cross-sectional data from these twin and family studies, as well as from large molecular genetic studies [Khera et al., 2019] imply that genetic influences are fairly stable over the lifespan from early childhood onwards.

In contrast, longitudinal twin studies indicate that genetic influences do vary with age. Molecular genetic analyses suggest that different sets of genes act at different ages, both in childhood and among adults [Choh et al., 2014]. Twin analyses of children

and adolescents show that as the individual develops and grows, there are novel genetic influences coming into play at different ages [Silventoinen et al., 2008, 2011]; these may reflect both changes in lean mass, such as muscle and bone growth, and in fat mass. Among adults, whose growth has ended, changes in weight result mainly from changes in body fat. Twin models indicate that genetic effects on weight change are poorly correlated with the stable component of BMI [Ortega-Alonso et al., 2012, Hjelmborg et al., 2008]. Analyses of genetic risk scores at different ages support these results [Choh et al., 2014].

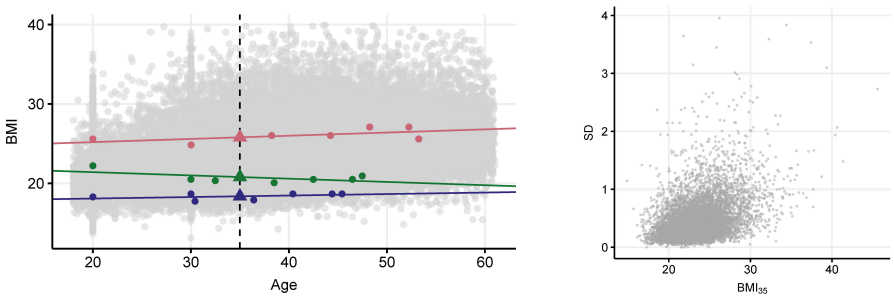
At different levels of BMI, the proportions of lean and fat mass differ on average, and hence it could be expected that genetic effects are not uniform across all BMI values. In an analysis of extreme leanness vs obesity, Riveros-McKay et al. [2019] found that the two traits were only partially correlated genetically ($rG = 0.49$). Using commingling analyses of BMI in MZ twin pairs, Price and Stunkard [1989] found twin correlations to be lower in overweight and obese twin pairs. This restricts to a truncated upper-tail of the BMI distribution. The authors did not have DZ pairs to derive heritability estimates at different levels of obesity. Studies of the genetics of BMI across the whole spectrum of BMI are rare; a recent study uses parent-offspring and sibpair relationship and quantile regression to estimate heritability of BMI at various BMI values [Williams, 2020]. The study finds increasing heritability with increasing BMI values, a result seen with other measures of fatness but not height. However, using family relationships or only MZ pairs can be challenging to distinguish between genetic and non-genetic familial effects contributing to the estimated heritability.

Recently Berentsen et al. [2020] extended the classical notion of heritability to that of a heritability curve, which allows the heritability to vary with the trait value. Using empirical data from the Finnish Twin Cohort, we demonstrate that there is variation in the contribution of genetic factors over the range of BMI values seen in a population sample.

2 Material and Methods

2.1 The dataset

The dataset we use in our analysis contains repeated BMI measurements on 4349 same sex twin pairs (1506 monozygotic and 2843 dizygotic) from the Finnish Twin Cohort [Hjelmborg et al., 2008], [Kaprio et al., 2019]. Each twin pair was asked to provide BMI measurements at different stages in life; for each pair we have up to 7 different values,



(a) Longitudinal BMI measurements (grey dots) pooled across individuals, with data and regression line highlighted for three individuals (green, blue, and pink). The triangles indicate the estimate of BMI_{35}.

(b) Regression standard deviation (SD) against estimated BMI_{35}.

Figure 1: Regression analysis used to estimate BMI_{35} (BMI at age 35) for each of 8598 individuals in the study.

at different ages and waves of data collections. Each such measurement is accompanied by the following information: the twin pair it belongs to, the wave number, the age at which the measurement is taken, the sex of the twins, and their zygosity. The data include information not only on weight at the current wave, but also recall of weight earlier in life.

Since the measurements were taken at different ages for different twin pairs, we use linear regression, separately on each individual, to obtain estimates of BMI at the reference age 35, which we denote with BMI_{35} (Figure 1). To reduce estimation uncertainty, we consider only twin pairs that have been measured three or more times. The resulting dataset contains 1493 monozygotic twin pairs (606 males and 887 females) and 2806 dizygotic twin pairs (1218 males and 1588 females). More details on the preprocess of the data can be found in the supplementary material.

2.2 Statistical methods

2.2.1 Heritability curve

In biometrical models, heritability is typically defined as the proportion of a trait variance attributed to genetic effects. Depending on the family structure of the data, the trait variance can be decomposed in several ways. The most commonly used biometrical model for twins is the ACE model, where it is assumed that the trait value can be decomposed into additive genetic effects (A), common (shared) environment (C),

and residual (random) environment (E). The proportion of trait variance explained by component A is often referred to as narrow-sense heritability. Another frequently used model for twins is the ADE model, where the C component in the ACE model is replaced by dominant genetic effects (D). The proportion of trait variance explained by the components A and D combined is then referred to as the broad sense heritability [Khoury et al., 1993].

Data on monozygotic and dizygotic twins provide contrasts from which the genetic variance can be separated from the environmental variance. For the ACE model, it is assumed that the amount of shared environment is the same for the two types of twins and that the amount of shared additive genetic effects is 100% and 50 % for mono- and dizygotic twins, respectively. If the correlation between the trait values of monozygotic twins is larger than the correlation between dizygotic twins, the difference is ascribed to the additive genetic effects alone, and the trait is heritable. The heritability can be quantified by comparing empirical correlations of monozygotic and dizygotic twins with the "expected" correlations implied by the model, and for the ACE model, this results in the well-known Falconer's formula [Falconer, 1960] for heritability:

$$\begin{aligned} a^2 &= 2(\rho^{(MZ)} - \rho^{(DZ)}), \\ c^2 &= 2\rho^{(DZ)} - \rho^{(MZ)}, \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{1}$$

Here a^2, c^2 and e^2 denote the proportions of the total variance explained by the components A, C, and E, respectively, while $\rho^{(MZ)}$ and $\rho^{(DZ)}$ denote the Pearson intraclass correlation of monozygotic and dizygotic twins, respectively. For the ADE model, the corresponding equations are given by

$$\begin{aligned} a^2 &= 4\rho^{(DZ)} - \rho^{(MZ)}, \\ d^2 &= 2(\rho^{(MZ)} - 2\rho^{(DZ)}), \\ e^2 &= 1 - \rho^{(MZ)}. \end{aligned} \tag{2}$$

The derivation of equations 1 and 2 can be found in the supplementary material.

Recently, the classical notion of heritability has been extended to that of a heritability curve [Berentsen et al., 2020] assuming that trait values in pairs are governed by a

Gaussian mixture distribution (see section 2.2.2). This allows the heritability to vary with the trait value, resulting in a curve $a^2(y)$ that potentially varies for different trait values y. The heritability curves are derived based on the same type of variance decomposition as for ACE and ADE model, but conditionally on a given phenotypic value. In this way, the heritability curve measures the heritability as a function of the trait itself, and would not be expected to be constant over the whole phenotypic range. The conditioning on a phenotype value is done via local correlations curves [Bjerve and Doksum, 1993]. Rather than comparing the ordinary Pearson correlation between phenotype values of monozygotic and dizygotic twins, we do the same type of comparison (e.g. using Falconer's formula under the ACE model) using correlation curves $\rho_{MZ}(y)$ and $\rho_{DZ}(y)$, evaluated at different values of BMI. Note that this procedure also provides curves $c^2(y)$ (or $d^2(y)$ for the ADE model) and $e^2(y)$ allowing the other components in the biometrical model to vary with the trait value as well. When there is no variation with trait value, the heritability curve reduces to the classical heritability coefficient.

2.2.2 Gaussian mixtures

Classical heritability models assume a bivariate Gaussian distribution for pair of traits in twins, typically with a positive correlation. Under this assumption the heritability curve reduces to the classical heritability coefficient [Berentsen et al., 2020], and does not provide any additional insight. Bivariate Gaussian mixtures are a more flexible class of bivariate distributions and underlie the implementation of the heritability curve of Berentsen et al. [2020]. A Gaussian mixture is a weighted sum of Gaussian kernels (Figure 2). The number m of Gaussian kernels is a data driven parameter, and for this purpose we use the BIC criterion [Berentsen et al., 2020]. Note that with $m = 1$ the mixture reduces to an ordinary bivariate Gaussian distribution. Each bivariate Gaussian kernel has three parameters: mean, variance and correlation, i.e. mean and variance are assumed identical across the twin individuals, yielding in total 3 times m unknown parameters. Monozygotic and dizygotic twins are allowed to have different correlations, which introduces m additional correlation parameters. Finally, there are m weight parameters, but due to a sum-to-one constraint, only $m - 1$ of these need to be estimated. The total of $Q = 5m - 1$ parameters are estimated by maximum likelihood [Berentsen et al., 2020]. Formulae exist for the marginal mean, variance and correlation, referred to as "global", in terms of the kernel-specific parameters.

Covariates can be introduced into both the mean, variance, or covariance part of the model. In our study sex is the only covariate, and we consider three different configurations of the model:

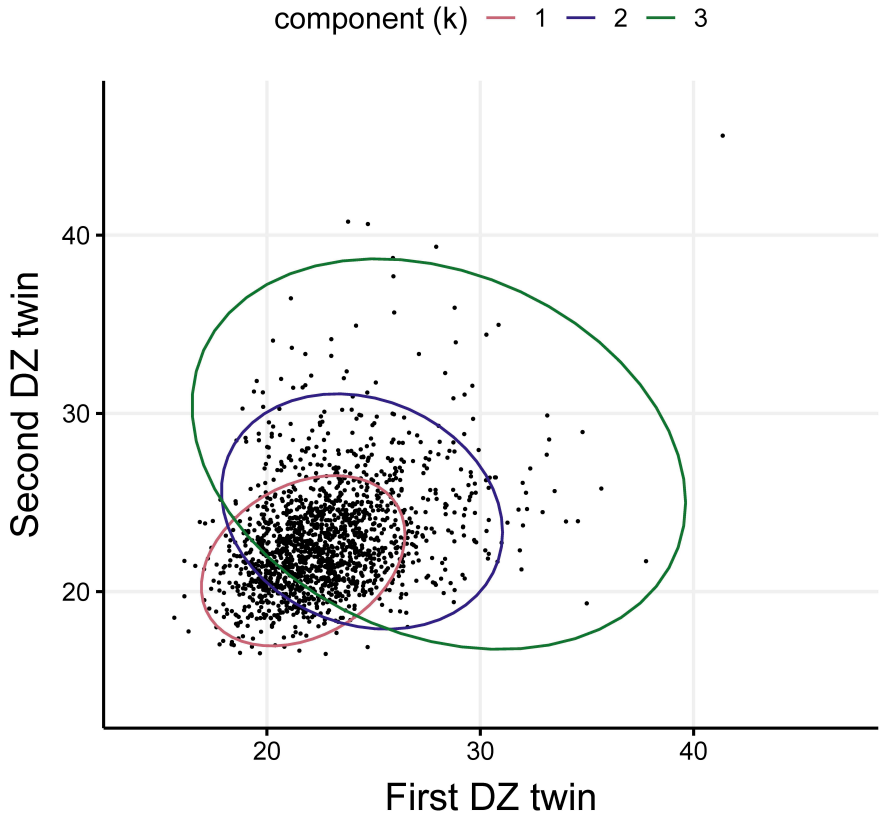


Figure 2: BMI_{35} for pairs of female dizygotic twins, obtained by regression analysis. The three ellipses represent 0.95 probability regions for the three Gaussian kernels of the fitted mixture distribution. The parameter values associated with each kernel can be found in Table 2.

1. "Stratified" in which fully separate Gaussian mixtures are fitted for males and females, except that they are constrained to have the same value of m,

2. "Mean" in which only the mean is sex specific. The mean of each Gaussian kernel for males is right-shifted by the same amount from females.

3. "Mean+covariance" in which m mean and correlation parameters are sex specific, but the variance is assumed equal by sex.

A mathematical description of the model is provided in supplementary material. The

supplement also contains details about how the parameters of the Gaussian mixture were estimated from data using the software package TMB [Kristensen et al., 2016].

2.2.3 Biometrical model selection

The choice between the ACE and ADE model relies traditionally on the relationship between the (empirical) Pearson correlations $\rho^{(MZ)}$ and $\rho^{(DZ)}$. More specifically, the sign of the quantity $2\rho^{(DZ)} - \rho^{(MZ)}$ indicates which model is more appropriate. Under the ACE model, this quantity corresponds to the proportion of variance explained by the shared environment, c^2. By contrasting equations (1) and (2) we obtain the relationship $d^2 = -2c^2$ with the proportion of variance explained by the dominant genetic effects, d^2. Consequently, if $2\rho^{(DZ)} - \rho^{(MZ)} < 0$, i.e. c^2 is negative, the ACE model is (most likely) misspecified. Vice versa, if $2\rho^{(DZ)} - \rho^{(MZ)} > 0$, d^2 is negative and the ADE model is misspecified.

In the context of correlation curves, the ACE model may be suitable in some part of the phenotypic range, while the ADE model may be more appropriate in the remaining part. Adopting the above procedure, we can then switch between ACE and ADE model according to the sign of the quantity $2\rho^{(DZ)}(y) - \rho^{(MZ)}(y)$; the ACE model is preferred when $2\rho^{(DZ)}(y) - \rho^{(MZ)}(y) > 0$, and the ADE model otherwise [Berentsen et al., 2020].

3 Results

The longitudinal BMI measurements used in the regression analysis are shown in Figure 1a. The distribution of uncertainties in fitted BMI_{35} is mostly confined to the interval from 0 to 2, but some twin pairs have higher uncertainty (Figure 1b). The bivariate distribution of BMI_{35} within twins deviates from normality for DZ females and is having a pear shape with less association for large BMI values (Figure 2). The same is true for males, and to a lesser extent for MZ twins (Supplementary material). The mixture model can accommodate this pear shape, by using $m = 3$ individual Gaussian components (Figure 2). The fact that the mixture distribution fits data much better than a bivariate Gaussian distribution ($m = 1$) is clear from a comparison of BIC values (Table 1).

The best fitting covariate model is that in which sex affects only the mean of the response (Table 1). However, the difference in terms of BIC between this model and the stratified model or the model with a sex effect also in the correlation structure is not large. The

m	Mean		Mean+covariance		Stratified	
	Q	ΔBIC	Q	ΔBIC	Q	Δ BIC
1	5	1211.03	7	1204.01	8	1125.78
2	10	472.95	12	485.87	18	56.92
3	15	0*	17	12.46	28	17.30
4	20	0.65	22	12.33	38	56.57
5	25	11.72	27	24.59	48	94.98

Table 1: Model selection by the BIC criterion among three candidate models for sex effect (columns) and the number of mixture components (m). Q represents the number of parameter estimated, and ΔBIC shows BIC relative to the best fitting model (*) across the table. Color red highlights the model with lowest BIC within each column.

latter two are both more flexible in their ability to fit the distributional shape of data but are being penalized by the BIC criterion for having more parameters than the selected model. The selected model has $m = 3$ mixture components. The BIC values in Table 1 are for the entire dataset (male/female and MZ/DZ). In the stratified model, this amounts to adding the BIC values computed separately for males and females. In an additional analysis, where males and females were allowed to have a different value of m it was found that the best fitting values were $m = 2$ for males and $m = 3$ for females, but the total BIC was not lower than the selected model in Table 1. A more in depth analysis of the different models can be found in the supplementary material.

The parameter estimates for the mixture model (Table 2; column "Global") show that mean BMI is higher for males than for females by $\mu_{\text{male}} - \mu_{\text{female}} = 1.86$ units. The global correlation is expectedly stronger in MZ twins than in DZ twins ($\rho_{\text{MZ}} = 0.70$ versus $\rho_{\text{DZ}} = 0.34$). By constraint of the chosen model (Table 1), correlations are the same for males and females. For the three individual Gaussian components of the mixture, components $k = 2, 3$ have a negative correlation for DZ twins, which is contributing to the pear shape of the overall mixture (Figure 2). Component $k = 3$ has the largest standard deviation (σ) and represents $100 \times p_3 = 4\%$ of the data (Table 2). This part of the data includes twin pairs which may be classified as outliers in the sense of having a strong negative association in BMI across the twins (Figure 2).

By construction of the selected model, the shape of the correlation curve for males is identical to that for females, but is right-shifted by an amount $\mu_{male} - \mu_{female} = 1.87$ (Figure 3). While this might seem like a strong restriction, it is worth noting that our model is preferred by the BIC criterion over the two other models in Table 1, which both allow for more flexibility in the correlation curves. Twin [Silventoinen et al., 2017] and molecular genetic studies of BMI [Khera et al., 2019] have shown very little evidence of sex-specific genetic variance or genes expressed only in one sex. In contrast, the distribution of fat differs between men and women, and is affected by genetic factors.

Parameters	$k=1$	$k=2$	$k=3$	Global
μ_{male}	23.57	26.36	29.82	24.55
μ_{female}	21.70	24.49	27.96	22.68
σ	1.94	2.72	4.67	2.84
ρ_{MZ}	0.74	0.34	0.38	0.69
ρ_{DZ}	0.31	-0.19	-0.22	0.36
p	0.70	0.26	0.04	

Table 2: Parameter estimates for the chosen $m = 3$ component Gaussian mixture assuming a sex effect in the mean (μ). Additional parameters are standard deviation (σ), monozygotic and dizygotic correlation (ρ), and mixture weights p. Each column ($k = 1, 2, 3$) corresponds to different mixture components. The final column refers the parameter value for the mixture as a whole.

We observe a drop in the correlation curve for high BMI in both monozygotic and dizygotic twins (Figure 3). The correlations increase before dropping, but this pattern is more noticeable in dizygotic twins (it increases up to a BMI value of about 24 for female data and 26 for male data).

The property that the male correlation curve is, by construction, just a right-shifted version of the female one carries over to the heritability curves. Hence, we only discuss female heritability in the following. Figure 4 displays curves obtained using both ACE and ADE genetic models. The panel labeled "common" contains the dominant genetic component (which appears in the ADE model) and the shared environment (which appears in the ACE model). The curve for the residual environment is the same in both models.

The Pearson correlations are $\rho^{(MZ)} = 0.74$ for monozygotic twins and $\rho^{(DZ)} = 0.42$ for dizygotic twins. The quantity $2\rho^{(DZ)} - \rho^{(MZ)}$ is positive; hence, if we opt to use one single model for the whole dataset range, the ACE model is most appropriate. However, as discussed in Section 2.2.3 we can instead switch between ACE and ADE model depending on the sign of the quantity $2\rho^{(DZ)}(y) - \rho^{(MZ)}(y)$. The dashed line in Figure 4 indicates the preferred (combined) model. We define the "mid range" of BMI values (21–27) as the region where the ACE model is preferred and "low/ high range" BMI values as the region where the ADE model is preferred.

The mid range BMI values are governed by additive genetic effects (A) and to some extent the shared environment (C). The residual environment (E) plays a larger role for the upper mid-range BMI values. The heritability curve a^2 steadily decreases while the BMI values increase, starting from its highest value of 0.78 for the BMI value 21. The shared environment curve c^2, instead, displays a convex shape, increasing together with the BMI up until it reaches its maximum value of 0.25 around a BMI value of 24 and

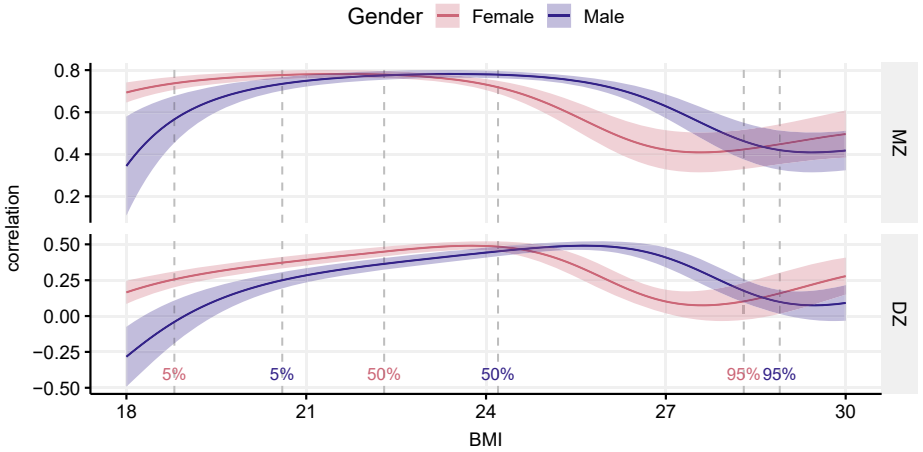


Figure 3: Correlation curves for male and female data for monozygotic (top) and dizygotic (bottom) twins, with pointwise 95% confidence bands (shaded regions). Vertical dashed lines show empirical quantiles (not model dependent) separately by sex, but pooled over MZ and DZ twins.

decreasing after. By construction, c^2 is zero at both extremes.

Low BMI values are overall governed by the genetic effects A and D. Dominant genetic effects (D) play a more pronounced role as the BMI values decrease, with its maximum value (0.73) reached at BMI value of 18, while the opposite trend can be seen with the additive heritability curve a^2. We also see a slight increase in environmental effects (E) as the BMI values decrease.

High BMI values are increasingly governed by environmental effects (E) (for a maximum value of 0.60) while broad sense heritability is decreasing (Figure 4). Interestingly, a change in type of genetic action is suggested by the curves; while additive genetic effect, the A component, is the suggested action for BMI values in the normal range, dominant genetic and weak epistatic effects (the D component) tend to govern the upper part of the BMI scale effects (D) ((for a maximum value of 0.52 around a BMI value of 28). The basis for this suggestion follows from decreasing within-pair correlation in BMI among MZ pairs with increasing BMI and similar decreasing within-pair correlation among DZ pairs, however with larger (or faster) decrease for the DZ pairs with increasing BMI. Implications of this suggested change in mode of genetic influence is discussed below.

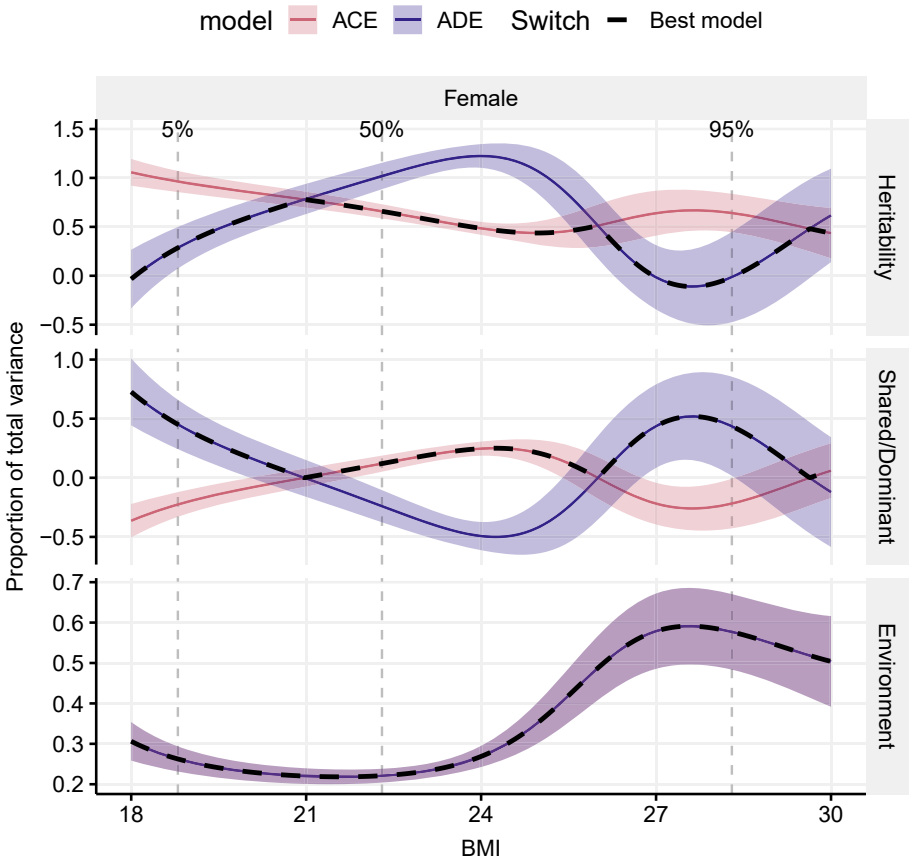


Figure 4: Decomposition of total variance into genetic and envirolmental effects by BMI, sex, and genetic model (ACE/ADE). Shaded areas indicate 95% pointwise confidence bands. The environmental part is identical for the ACE and ADE models, indicated by the pink color in the lower part. The dashed black line represents the combination of the ACE and ADE models.

4 Discussion

In the present analysis, we demonstrate that multiple mixture models account for the pairwise relationship of BMI rather than a single bivariate distribution assumed in prior analyses of twin data of BMI. Further, the 95% probability regions of the kernels of the distributions are shaped differently as seen in Figure 2. The majority of pairs are in a symmetric circular distribution, while the remainder are in distributions indicating greater within pair differences, possibly due to greater than average genetic differences

and/or specific environmental triggers affecting body weight development more in one twin than in the other. Such pairs discordant for BMI have proved very informative for the study of causes and consequences of obesity [van der Kolk et al., 2021], [Naukkarinen et al., 2012]. The existence of several types of bivariate distributions suggests that a single bivariate normal model of multifactorial inheritance with a polygenic component is not sufficient to account for the complexity of interplay of genetic and environmental factors in BMI, even though GWA studies have been highly successful in identifying hundreds of BMI-associated genes and accounting for about a fifth of the variance in BMI [Khera et al., 2019]. On the other hand, rare Mendelian variants and various obesity-related syndromes account for a relatively small proportion of variance in BMI [Kaur et al., 2017].

When we consider the resulting heritability curves, and associated curves of MZ and DZ correlations by level of BMI, we observe very high estimates of the contribution of genetic factors in the region of what is generally termed normal BMI [Seidell and Flegal, 1997]. As BMI comprises both lean (muscle, organs and bone) as well as fat mass, our results are consistent with the notion that in the absence of excess fat, body build is highly genetic determined. As BMI increases, the proportion of weight accounted for by fat mass increases, and the contribution of genetic variation decreases. This is consistent also with the rapid increase in obesity in global populations [GBD 2015 Obesity Collaborators, 2017] being due to environmental factors rather than changes in the gene pool over the past decades.

In the ADE model, the genetic effect is split into an additive genetic component (A) and a dominant genetic component (D). We can compare the curve $a^2(y)_{ACE}$, computed using the ACE model, with the sum $a^2(y)_{ADE} + d^2(y)$, as they both represent the total heritability of the model (the latter assuming independence of A and D). Hence, even though a first look at Figure 4 may suggest a contradiction between the heritability estimate in the ACE and ADE models, we should not compare $a^2(y)_{ACE}$ and $a^2(y)_{ADE}$. Instead, we can study the behavior of the separate components of the total heritability. In particular, the dominant genetic component plays a larger role on the tails, while the additive genetic component has a larger effect in the middle of the data range.

Noteworthy, as can be derived from the biometric model, the D component may reflect some evidence for epistatis besides the dominant effects of variants. Hence the heritability curves may shed light on values for which such action may take place.

For BMI, it is biologically plausible that the genetic and environmental components vary over the range of BMI. For example, very large or small values of BMI could be caused by "sporadic" environmental factors such as accidents or by rare genetic mutations whereas

the medium phenotype variation may be dominated by multiple common genetic factors. The heritability curves then provide insights to an evolutionary normal spectrum of BMI of which magnitude of genetic variants is observed. Expectedly, genetic action on BMI for values outside the normal spectrum may stem from different localizations of variants governing different mechanisms. Hence the curves relate to some combination of genotypic, environmental and epigenetic interactions, the broad-sense heritability and it becomes important to study how curves may change given observed genetic variants which will be a perspective for further studies of correlation curves.

In this paper, we use a Gaussian mixture distribution to fit the data. To test the soundness of this assumption, we fitted a non-parametric correlation curve and showed that it returns similar results to Figure 3 everywhere except for low BMI values for female dizygotic data, where it estimates a higher correlation. See supplementary material for more details.

Data Availability

The FTC data is not publicly available due to the restrictions of informed consent. However, the FTC data is available through the Institute for Molecular Medicine Finland (FIMM) Data Access Committee (DAC) (fimm-dac@helsinki.fi) for authorized researchers who have IRB/ethics approval and an institutionally approved study plan. To ensure the protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will depend on the nature of the requested data.

Author Contributions

FA, GDB, HJS and JBH contributed to the conception and design of the work. JBH and JAK provided the data material. FA carried out all statistical analyses. FA drafted the manuscript, except for sections Introduction and Discussion which were drafted by JAK. All authors participated in finalizing the manuscript, and gave final approval and agreed to be accountable for all aspects of work ensuring integrity and accuracy.

References

R. T. Benn. Some mathematical properties of weight-for-height indices used as measures of adiposity. *British journal of preventive & social medicine*, 25(1):42, 1971.

G. D. Berentsen, F. Azzolini, H. J. Skaug, R. T. Lie, and H. K. Gjessing. Heritability curves: A local measure of heritability in family models. *Statistics in Medicine*, 2020.

S. Bjerve and K. Doksum. Correlation curves: Measures of association as functions of covariate values. *Ann. Statist.*, 21(2):890–902, 06 1993. doi: 10.1214/aos/1176349156. URL `https://doi.org/10.1214/aos/1176349156`.

A. C. Choh, M. Lee, J. W. Kent, V. P. Diego, W. Johnson, J. E. Curran, T. D. Dyer, C. Bellis, J. Blangero, R. M. Siervogel, et al. Gene-by-age effects on bmi from birth to adulthood: The fels longitudinal study. *Obesity*, 22(3):875–881, 2014.

C. E. Elks, M. Den Hoed, J. H. Zhao, S. J. Sharp, N. J. Wareham, R. J. F. Loos, and K. K. Ong. Variability in the heritability of body mass index: a systematic review and meta-regression. *Frontiers in endocrinology*, 3:29, 2012.

D. S. Falconer. *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London, 1960.

GBD 2015 Obesity Collaborators. Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine*, 377(1):13–27, 2017.

J. v. Hjelmborg, C. Fagnani, K. Silventoinen, M. McGue, M. Korkeila, K. Christensen, A. Rissanen, and J. Kaprio. Genetic influences on growth traits of bmi: a longitudinal study of adult twins. *Obesity*, 16(4):847–852, 2008.

J. Kaprio, S. Bollepalli, J. Buchwald, P. Iso-Markku, T. Korhonen, V. Kovanen, U. Kujala, E. K. Laakkonen, A. Latvala, T. Leskinen, et al. The older finnish twin cohort—45 years of follow-up. *Twin Research and Human Genetics*, 22(4):240–254, 2019.

Y. Kaur, R. J. De Souza, W. T. Gibson, and D. Meyre. A systematic review of genetic syndromes with obesity. *Obesity Reviews*, 18(6):603–634, 2017.

A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177(3):587–596, 2019.

M. J. Khoury, T. H. Beaty, and B. H. Cohen. *Fundamentals of Genetic Epidemiology*. Oxford University Press, 1993. ISBN 978-0-19-505288-6.

K. Kristensen, A. Nielsen, C. W. Berg, H. J. Skaug, and B. M. Bell. Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, 2016.

H. H. M. Maes, M. C. Neale, and L. J. Eaves. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics*, 27(4):325–351, 1997.

A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz. The disease burden associated with overweight and obesity. *Jama*, 282(16):1523–1529, 1999.

C. Nan, B. Guo, C. Warner, T. Fowler, T. Barrett, D. Boomsma, T. Nelson, K. Whitfield, G. Beunen, M. Thomis, et al. Heritability of body mass index in pre-adolescence, young adulthood and late adulthood. *European journal of epidemiology*, 27(4):247–253, 2012.

J. Naukkarinen, A. Rissanen, J. Kaprio, and K. H. Pietiläinen. Causes and consequences of obesity: the contribution of recent twin studies. *International Journal of Obesity*, 36(8):1017–1024, 2012.

A. Ortega-Alonso, K. H. Pietiläinen, K. Silventoinen, S. E. Saarni, and J. Kaprio. Genetic and environmental factors influencing bmi development from adolescence to young adulthood. *Behavior genetics*, 42(1):73–85, 2012.

A. Price and A. J. Stunkard. Commingling analysis of obesity in twins. *Human Heredity*, 39(3):121–135, 1989.

F. Riveros-McKay, V. Mistry, R. Bounds, A. Hendricks, J. M. Keogh, H. Thomas, E. Henning, L. J. Corbin, U. S. S. Group, S. O'Rahilly, et al. Genetic architecture of human thinness compared to severe obesity. *PLoS genetics*, 15(1):e1007603, 2019.

J. C. Seidell and K. M. Flegal. Assessing obesity: classification and epidemiology. *British medical bulletin*, 53(2):238–252, 1997.

K. Silventoinen, K. H. Pietiläinen, P. Tynelius, T. I. A. Sørensen, J. Kaprio, and F. Rasmussen. Genetic regulation of growth from birth to 18 years of age: the swedish young male twins study. *American Journal of Human Biology: The Official Journal of the Human Biology Association*, 20(3):292–298, 2008.

K. Silventoinen, B. Rokholm, J. Kaprio, and T. I. A. Sørensen. The genetic and environmental influences on childhood obesity: a systematic review of twin and adoption studies. *International journal of obesity*, 34(1):29–40, 2010.

K. Silventoinen, J. Kaprio, and Y. Yokoyama. Genetic regulation of pre-pubertal development of body mass index: a longitudinal study of japanese twin boys and girls. *Behavior genetics*, 41(2):234–241, 2011.

K. Silventoinen, A. Jelenkovic, R. Sund, Y. Hur, Y. Yokoyama, C. Honda, J. v. Hjelmborg, S. Möller, S. Ooki, S. Aaltonen, et al. Genetic and environmental effects on body mass index from infancy to the onset of adulthood: an individual-based pooled analysis of 45 twin cohorts participating in the collaborative project of development of anthropometrical measures in twins (codatwins) study. *The American journal of clinical nutrition*, 104(2):371–379, 2016.

K. Silventoinen, A. Jelenkovic, R. Sund, Y. Yokoyama, Y. Hur, W. Cozen, A. E. Hwang, T. M. Mack, C. Honda, F. Inui, et al. Differences in genetic and environmental variation in adult bmi by sex, age, time period, and region: an individual-based pooled analysis of 40 twin cohorts. *The American journal of clinical nutrition*, 106 (2):457–466, 2017.

B. W. van der Kolk, S. Saari, A. Lovric, M. Arif, M. Alvarez, A. Ko, Z. Miao, N. Sahebekhtiari, M. Muniandy, S. Heinonen, et al. Molecular pathways behind acquired obesity: Adipose tissue and skeletal muscle multiomics in monozygotic twin pairs discordant for bmi. *Cell Reports Medicine*, 2(4):100226, 2021.

P. T. Williams. Quantile-dependent heritability of computed tomography, dual-energy x-ray absorptiometry, anthropometric, and bioelectrical measures of adiposity. *International Journal of Obesity*, 44(10):2101–2112, 2020.

Supporting Material for "The heritability of BMI varies across the range of BMI – a heritability curve analysis in a twin cohort"

1 Data preprocessing by linear regression

The purpose of this supplementary is to provide additional information about the analysis in "The heritability of BMI varies across the range of BMI – a heritability curve analysis in a twin cohort" (Azzolini, Berentsen, Skaug, Hjelmborg, Kaprio), which is referred to as "main text".

The main text studies the BMI of twins using the Finnish Twin Cohort [Kaprio et al., 2019]. The relevant variables to the analysis are listed in Table 1.

Since the measurements are taken at different ages for different twins, we need to preprocess the data to obtain a set of comparable values. For each twin, we interpolate in up to seven BMI measurements from the different waves using simple linear regression to obtain BMI estimates at age 35 (approximately the average age in the dataset).

We use the program R [R Core Team, 2020] for the computations. We show a simplified version of the code in appendix A.

Name	Description	Values
BMI	BMI measurements	$[13.11, 29.92]$
Age	Age of the twin pair at measurement	$[18, 61]$
Sex	Sex of the pair (only same-sex pairs are included)	1 (Male), 2 (Female)
Zygoty	Zygosity of the twin pair	1 (MZ), 2 (DZ)
Wave	Different measurements of the same twin pair	from 1 to 7
Tvparnr	ID number of the twin pair	from 1 to 7639
Twinnumber	Different twins in the same pair	1,2
Twinnid	summary of both Tvparnr and Twinnumber	from 11 to 76392

Table 1: List and description of the variables included in the dataset.

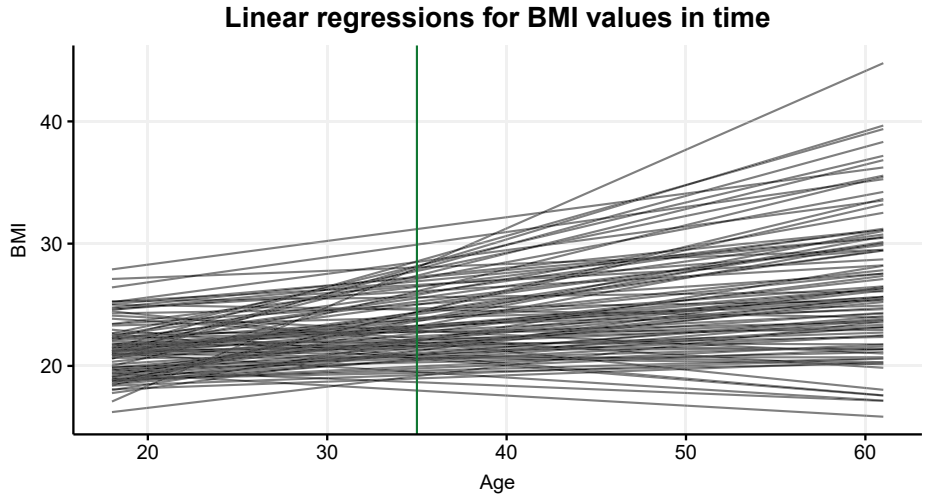


Figure 1: Regression lines fitted to BMI-at-age for 100 randomly selected individuals. The vertical green line is at age 35.

Figure 1 shows the regression lines of 100 randomly selected individuals from the real data, obtained through the process explained above.

2 Derivation of classical heritability formulas

In this section we show how to derive equations (2.1) and (2.2) in the main text.

Let Y_{ij} be the trait value of twin j $(j = 1, 2)$ in twin-pair i, and let $\rho^{(MZ)}$ and $\rho^{(DZ)}$ be the Pearson correlations $\text{cor}(Y_{i1}, Y_{i2})$ for monozygotic and dizygotic twins, respectively. Consider the mixed-effect model [McCulloch and Neuhaus, 2001]

$$Y_{ij} = \mu + \beta^t x_{ij} + A_{ij} + C_{ij} + D_{ij} + E_{ij}, \tag{1}$$

where A_{ij}, C_{ij}, D_{ij} and E_{ij} are mutually independent and follow a normal distribution with mean 0 and variances σ_A^2, σ_C^2, σ_D^2 and σ_E^2, respectively. The total variance of Y_{ij} is $\sigma^2 = Var(Y_{ij}) = \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2$. Note that this model assumes no gene-environment interaction. We define $a^2 = \sigma_A^2/\sigma^2$, $c^2 = \sigma_C^2/\sigma^2$, $d^2 = \sigma_D^2/\sigma^2$, and $e^2 = \sigma_E^2/\sigma^2$. By definition,

$$a^2 + c^2 + d^2 + e^2 = 1, \tag{2}$$

i.e. the contributions from all components sum to one.

The ACE and ADE models assume, respectively, that dominant genetic effects and shared environment do not affect the trait in study; in other words, d^2 and c^2 are assumed to be zero.

Mono- and dizygotic twins share 100% and 50% of the additive genetic effects, respectively. In formula, we write $\text{cor}(A_{i1}^{MZ}, A_{i2}^{MZ}) = 1$ and $\text{cor}(A_{i1}^{DZ}, A_{i2}^{DZ}) = 1/2$.

Monozygotic and dizygotic twins alike share the totality of the common environment, hence we make the common assumption $\text{cor}(C_{i1}, C_{i2}) = 1$. The dominant genetic component, instead, affects monozygotic and dizygotic twins differently; in particular, $\text{cor}(D_{i1}^{MZ}, D_{i2}^{MZ}) = 1$ and $\text{cor}(D_{i1}^{DZ}, D_{i2}^{DZ}) = 1/4$.

Traditional twin models utilize only the $\rho^{(MZ)}$ and $\rho^{(DZ)}$ phenotype correlations, which for the ACE model are

$$\begin{aligned} \rho^{(MZ)} &= \quad a^2 + c^2, \\ \rho^{(DZ)} &= \frac{1}{2}a^2 + c^2, \end{aligned} \tag{3}$$

while for the ADE model are given as

$$\begin{aligned} \rho^{(MZ)} &= a^2 + d^2, \\ \rho^{(DZ)} &= \frac{1}{2}a^2 + \frac{1}{4}d^2. \end{aligned} \tag{4}$$

Equations (2.1) and (2.2) in the main text can be easily derived from 2, 3, and 4.

3 Bivariate Gaussian mixtures

In Berentsen et al. [2020] are described in depth the advantages of using Gaussian mixture models as underlying distributions when constructing correlation curves. We summarize below the model that we use in this analysis, but we refer the interested reader to Berentsen et al. [2020] for more details.

The probability density of a m-component Gaussian mixture for a twin phenotype $\boldsymbol{y} = (y_1, y_2)$ is

$$\sum_{k=1}^{m} p_k \phi_2(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,. \tag{5}$$

The parameters $p_1, \ldots, p_m$ are non-negative values satisfying $\sum_{k=1}^{m} p_k = 1$, and

$\phi_2(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a bivariate normal density, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In the main text we refer to these densities as kernels.

To guarantee symmetry between the two twins, we impose some conditions on the mean vector and the covariance matrix:

$$\boldsymbol{\mu}_k = (\mu_k, \mu_k), \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_k^2 & \sigma_k^2 \rho_k \\ \sigma_k^2 \rho_k & \sigma_k^2 \end{pmatrix}, \tag{6}$$

where $\rho_k \in (-1, 1)$ is the correlation parameter and σ_k is the standard deviation. We further assume the p_k's, μ_k's, and σ_k's to be shared between monozygotic and dizygotic twins, with only the ρ_k's being different. This is a natural assumption as marginal BMI distributions of twins of the same zygosity are expected to be identical. Therefore, we estimate $m-1$ p_k's, m μ_k's, m σ_k's, and $2m$ ρ_k's. Furthermore, we introduce restrictions on the means to ensure identifiability of the model (See Section 4 for more details).

We select the most parsimonious number of mixture components using the criterion $\text{BIC} = -2\log(L) + \log(n)Q$ for each candidate model, where Q is the number of parameters and $\log(L)$ is the log likelihood function, as defined in [Berentsen et al., 2020, equation (3.14)].

3.1 Models with a sex covariate

The mean and covariance structures (6) allow to introduce covariates easily. In the main text, we briefly described how to include a sex effect in the model in three different ways. Below, we show the mathematical formulas.

The "Stratified" model consists in estimating two completely independent Gaussian mixtures with m components for male and female data, so that the parameters μ_k's, σ_k's, ρ_k's, and p_k's are all sex specific. The total number of parameters is $n = 2(5m-1)$.

The "Mean" model assumes a sex effect on μ_k's; we follow the parametrization from Berentsen et al. [2020], and introduce a sex effect term β_μ such that for each component k:

$$\boldsymbol{\mu}_k = (\mu_k + \beta_\mu x_i, \mu_k + \beta_\mu x_i), \tag{7}$$

where $x_i = 0.5$ for male data and $x_i = -0.5$ for female data. Hence, $\boldsymbol{\mu}_{male} = \boldsymbol{\mu}_{female} + \beta_\mu$. The total number of parameters is $n = 5m$.

The "Mean+covariance" model assumes a sex effect on both μ_k's and ρ_k's. In addition to assuming the structure 7 for μ_k's, we define the common term β_ρ such that for each component k:

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_k^2 & \sigma_k^2(\rho_k + \beta_\rho x_i) \\ \sigma_k^2(\rho_k + \beta_\rho x_i) & \sigma_k^2 \end{pmatrix}, \tag{8}$$

where $x_i = 0.5$ for male data and $x_i = -0.5$ for female data. To reflect the differences between monozygotic and dizygotic correlation coefficients, we introduce two different parameters, β_ρ^{MZ} and β_ρ^{DZ}. The total number of parameters is $n = 5m + 2$.

4 Implementation in TMB

To perform the analysis on the dataset, we use two files, written in two languages: R [R Core Team, 2020] and C++ [ISO/IEC, 2017]. We access and integrate the C++ code in the optimization process using the package TMB [Kristensen et al., 2016]. The code is accessible at https://github.com/skaug/Supplementary, under the repository Azzolini-etal-BMIvaries.

In the R file we read the dataset, we initialize the parameters, and we run the optimization function. This optimization function invokes the C++ script and minimizes the negative log likelihood function there defined.

The number of components of the Gaussian mixture is a hyperparameter that we select using BIC as criterion. We have already performed model selection on the dataset, so we only report the code for the best fitting model, with $m = 3$ components.

An issue that arises when Gaussian mixtures are estimated is label switching - that is, having several equivalent parameter estimates where the only difference is the order of the components. To avoid this problem, we force the means to be ordered from smallest to largest. To achieve this we reparameterize the model in terms of $\boldsymbol{\alpha}$ and we then construct the mean vector $\boldsymbol{\mu}$ as follows:

$$\mu_1 = e^{\alpha_1}$$

"Stratified"										
Male data										
Par	$k=1$	se	$k=2$	se	$k=3$	se	$k=4$	se	Global	se
μ_k	23.64	0.11	26.33	0.27	30.35	0.20			24.66	
σ_k	1.97	0.05	2.46	0.11	3.67	0.47			2.59	
$\rho_k^{(MZ)}$	0.72	0.02	0.27	0.14	0.24	0.24			0.69	
$\rho_k^{(DZ)}$	0.34	0.04	−0.29	0.12	−0.94	0.03			0.36	
p_k	0.75	0.05	0.24	0.05	0.02	0.01				
Female data										
Par	$k=1$	se	$k=2$	se	$k=3$	se	$k=4$	se	Global	se
μ_k	21.62	0.09	24.57	0.25	28.23	0.78			22.74	
σ_k	1.91	0.05	2.83	0.13	4.91	0.35			2.96	
$\rho_k^{(MZ)}$	0.75	0.02	0.34	0.09	0.41	0.20			0.69	
$\rho_k^{(DZ)}$	0.26	0.04	−0.22	0.08	−0.18	0.14			0.34	
p_k	0.67	0.04	0.29	0.03	0.04	0.01				
"Mean"										
Par	$k=1$	se	$k=2$	se	$k=3$	se	$k=4$	se	Global	se
μ_k	22.64	0.08	25.42	0.23	28.89	0.69			23.61	
σ_k	1.94	0.04	2.72	0.12	4.67	0.32			2.84	
$\rho_k^{(MZ)}$	0.74	0.02	0.34	0.08	0.38	0.15			0.69	
$\rho_k^{(DZ)}$	0.31	0.03	-0.19	0.07	-0.22	0.12			0.36	
p_k	0.70	0.03	0.26	0.03	0.04	0.01				
β_μ									1.86	0.06
"Mean+covariance"										
Par	$k=1$	se	$k=2$	se	$k=3$	se	$k=4$	se	Global	se
μ_k	21.41	0.24	23.17	0.14	26.15	0.91	29.68	0.10	23.57	
σ_k	1.51	0.10	2.02	0.05	2.94	0.13	5.02	0.39	2.79	
$\rho_k^{(MZ)}$	0.88	0.04	0.65	0.03	0.26	0.09	0.37	0.22	0.69	
$\rho_k^{(DZ)}$	0.51	0.10	0.16	0.05	-0.30	0.08	-0.14	0.16	0.36	
p_k	0.15	0.05	0.65	0.04	0.18	0.02	0.02	0.01		
β_μ									1.86	0.06
MZ β_ρ									-0.01	0.03
DZ β_ρ									0.11	0.05

Table 2: Parameter estimates for the best-fitting Gaussian mixtures for each covariate model. For each estimate, we present its standard error. The mixture components are ordered according to the value of σ_k. The global quantities, μ, σ, $\rho^{(MZ)}$ and $\rho^{(DZ)}$ are calculated from [Berentsen et al., 2020, equation (3.4)].

Global quantities			
Parameter	"Mean"	"Mean+covariance"	"Stratified"
male μ	24.54	24.50	24.66
female μ	22.68	22.64	22.74
male σ	2.84	2.79	2.59
female σ	2.84	2.79	2.96
male $\rho^{(MZ)}$	0.69	0.69	0.69
female $\rho^{(MZ)}$	0.69	0.70	0.69
male $\rho^{(DZ)}$	0.36	0.42	0.36
female $\rho^{(DZ)}$	0.36	0.31	0.34

Table 3: Comparison of the global quantities for the three covariate models for both sexes.

and, for every $2 \leq i \leq m$,

$$\mu_i = \mu_{i-1} + e^{\alpha_i}.$$

This guarantees that $\mu_1 < \mu_2 < ... < \mu_m$. This coding works if all mean components are expected to be positive (which is the case with BMI), but can easily be tweaked so that it admits negative values as well.

In the C++ file we read the data and the parameters and we define the negative log likelihood function that is optimized in the R file. Moreover, we construct the expressions for the correlation curves and estimate them at 500 points along the BMI range. These points are used to draw the plots of the curves.

To improve the precision of our estimates, we use both the gradient and the hessian matrix in the optimization function. We are able to compute the hessian thanks to automatic differentiation performed by the package TMB.

Due to privacy, we cannot share the dataset we worked on in this document. To present the performance of the code, we created a simulated dataset. The dataset follows a Gaussian mixture distribution with three components. As parameters we used the estimates we obtained from the analysis on the twin data (Table 2). Figure 2 shows the scatterplot of the simulated data.

5 Comparison between different covariate models

Figure 3 is an extension of Figure 2 from the main text. The ellipses describe 0.95 probability regions for the three Gaussian components of the best-fitting model. We observe that male and female dizygotic data are both pear-shaped while that is not so

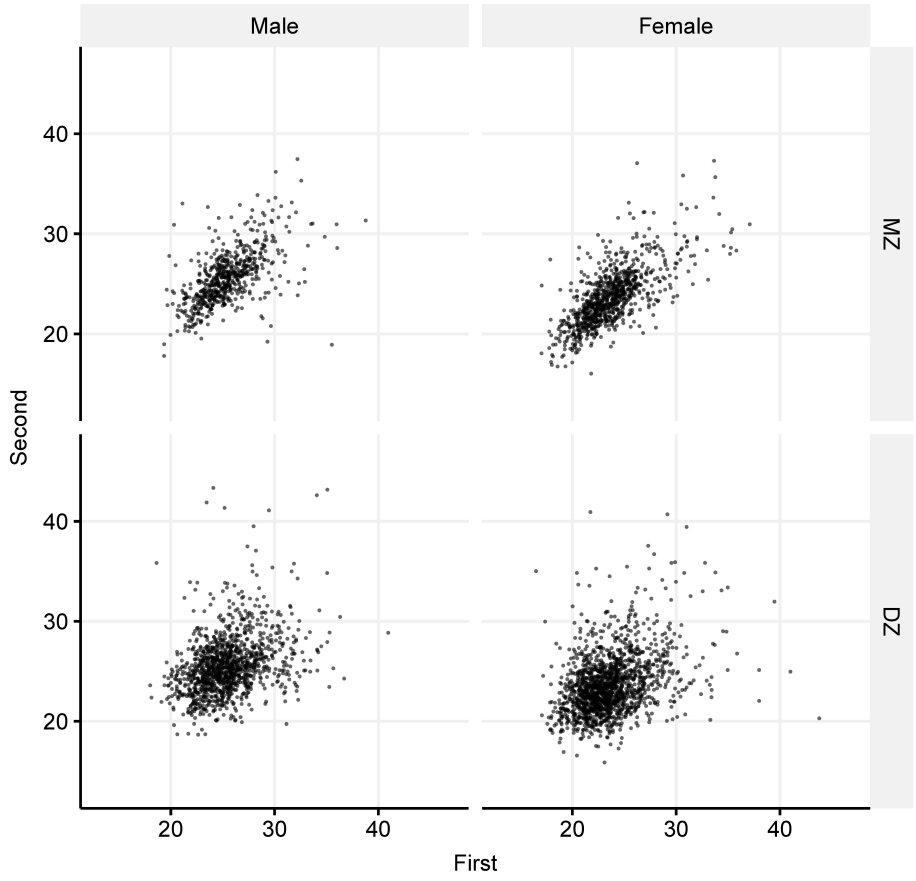


Figure 2: Scatterplot of the simulated dataset, divided by sex (male and female) and zygosity (monozygotic and dizygotic). The dataset follows a Gaussian mixture distribution, using as parameters the estimates from our analysis.

evident in monozygotic data. The correlation coefficients (Table 2) well capture this difference. It also shows the higher variability of female data, which is captured by the models with a sex covariate.

Table 2 contains the parameter estimates of the best fitting Gaussian mixture within each covariate model. They also contain the standard error estimate for each coefficient.

Female data have a larger variance in its right tail (as can be seen in Figure 3) compared to male data, both for monozygotic and dizygotic twins. This is captured better by the "stratified" model. We notice that both the weights p_2 and p_3 and the standard deviations σ_2 and σ_3 are larger for female data. The lower variability in the right tail

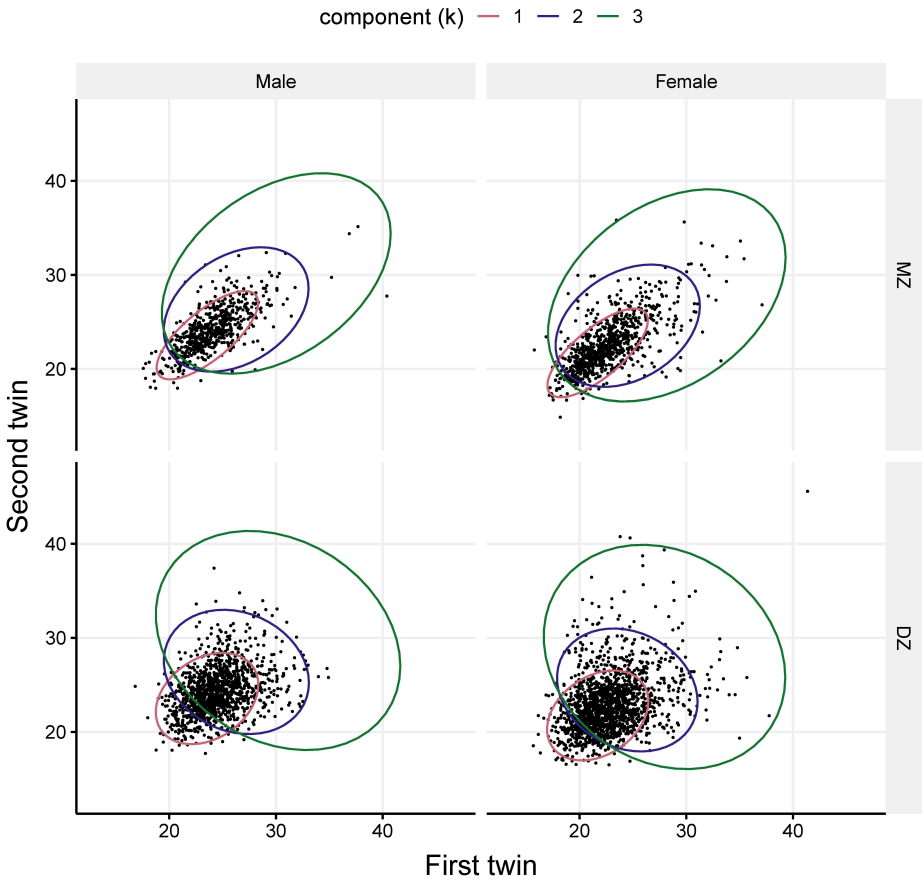


Figure 3: Ellipses representing the three Gaussian kernels, divided by sex and zygosity.

for male data, especially in dizygotic twins, is also reflected in the correlation coefficient $\rho_3^{(DZ)}$ which approaches -1, compared to the value of -0.18 for female data. Despite these differences being justified by the dataset, the BIC value still prefers the less flexible mean covariate model.

The values of μ_k^F and μ_k^M obtained through the stratified analysis are quite similar to the ones from the best fitting model (mean covariate), albeit slightly larger. The average difference between male and female mean components for stratified analysis is 1.97, slightly larger than the estimated parameter $\beta_\mu = 1.86$. This is not reflected in the global quantities (Table 3), which are very similar between the two different models.

Comparing the "mean+covariance" parameter estimates (Table 2) is slightly more dif-

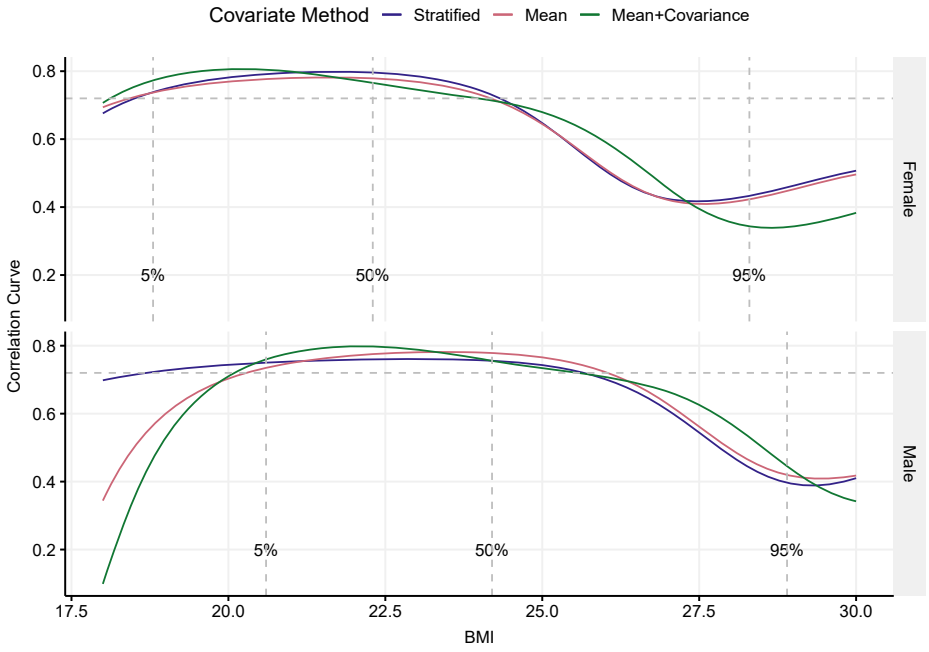


Figure 4: Correlation curves for monozygotic twins using the three different methods described in section 3.1. For each method, we plotted the correlation curve for the best model for both male and female BMI. The horizontal grey lines represent the Pearson global correlation $\rho^{(MZ)}$. The vertical dotted lines represent the 0.05, 0.5, and 0.95 quantiles of the data, divided by sex.

ficult, since the best fitting mixture has one component more than the other two best fitting models. Looking at the global quantities (Table 3), we see that they are pretty consistent with the other two models. Moreover, the coefficient β_μ is the same as in the "Mean" model.

The coefficient $\beta_\rho^{(MZ)}$ is not significant. The model does identify a more significant sex effect on the dizygotic correlation coefficient. The BIC values still favors the simpler "Mean" model.

5.1 Comparison of correlation curves

The differences in BIC values between the best-fitting mixtures among different covariate models is not very large (see Table 1 in main text). We also showed, in the above section, that the parameter estimates are relatively consistent between the different models. In

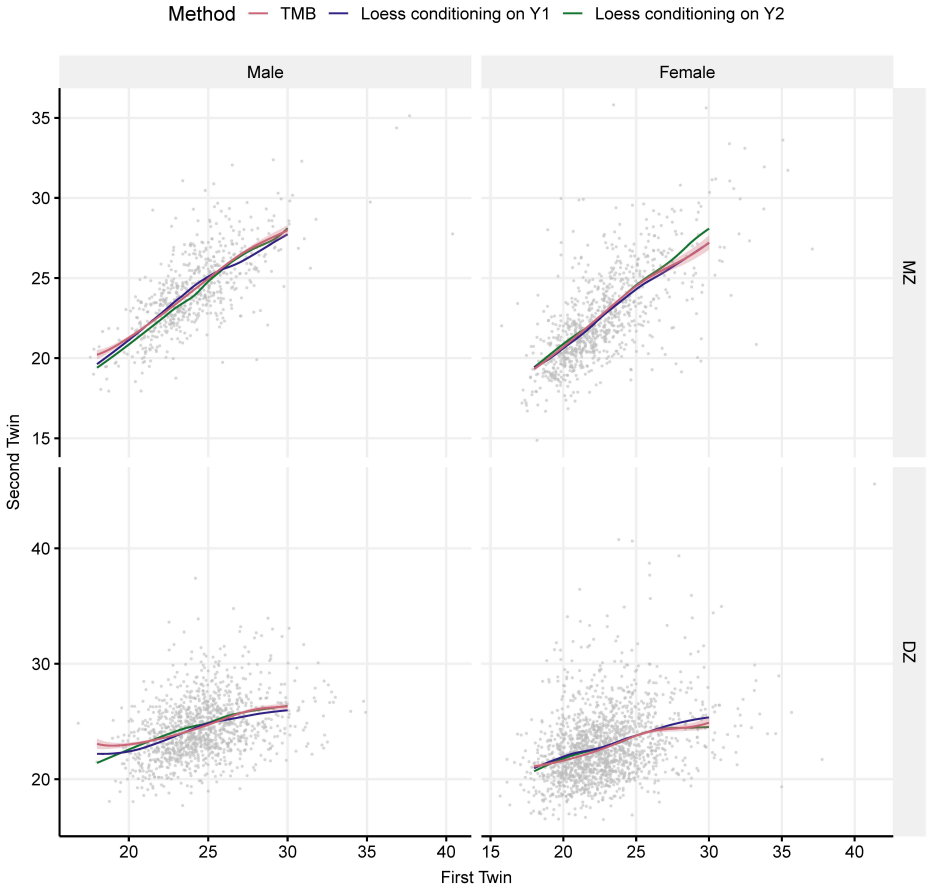


Figure 5: Comparison of estimated conditional mean curves based on the Gaussian mixture (TMB; red) with the two fully nonparametric estimates (blue and green) obtained from "loess". The estimation uncertainty is displayed only for the Gaussian mixture, and the underlying data are shown as grey dots.

this section we compare the monozygotic correlation curves obtained using the models from the previous section ("stratified", "Mean", "Mean+Covariance").

Figure 4 displays the estimated monozygotic correlation curves according to the different models. For each of the three methods, we plot a curve for male data and one for female data. The three vertical lines represent the 0.05, 0.50, and 0.95 quantiles of the male and female data separately.

We observe that the curves of the same sex follow the same broad shape. In general, "stratified" and "mean" models generate curves that are quite similar. We remark

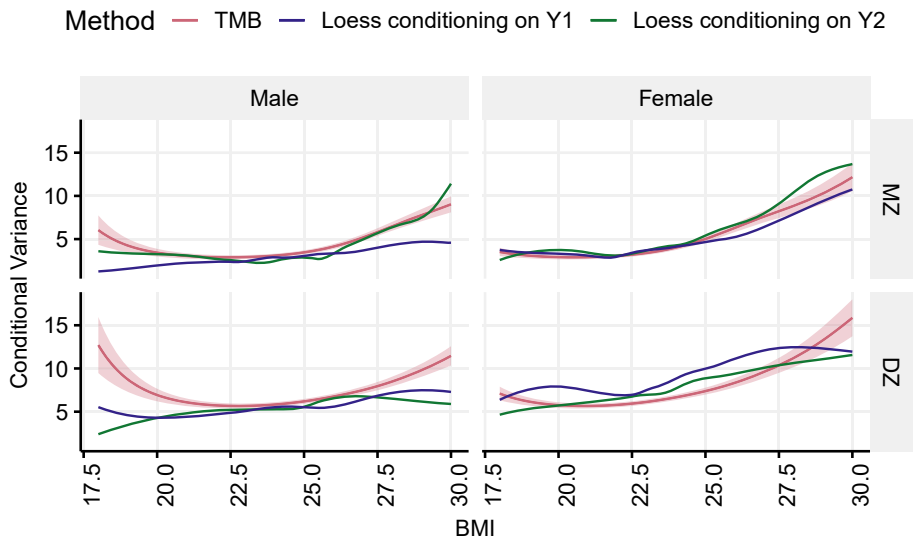


Figure 6: Comparison of estimated conditional variance curves based on the Gaussian mixture (TMB; red) with the two fully nonparametric estimates (blue and green) obtained from "loess". The estimation uncertainty is displayed only for the Gaussian mixture.

that, while the "mean" model generates male and female curves which are identical and shifted on the x-axis, the same is not true for the more flexible "stratified" model. This is particularly evident by looking at the distance between male and female curves in the left tail.

The curves generated by the "mean+covariance" model deviate slightly from the pattern above, especially in the right tail. This difference is possibly caused by the different number of components of the best fitting model.

6 Sensitivity Study

As a sensitivity study, we compared the estimated correlation curve based on the Gaussian mixture with a fully nonparametric method based on the smoother function "loess" in R. In particular, we computed nonparametric conditional means and variances, and these were used to obtain nonparametric correlation curves. In "loess" we first regressed $Y1$ on $Y2$, and secondly $Y2$ on $Y1$, and hence obtained two different estimated cor-

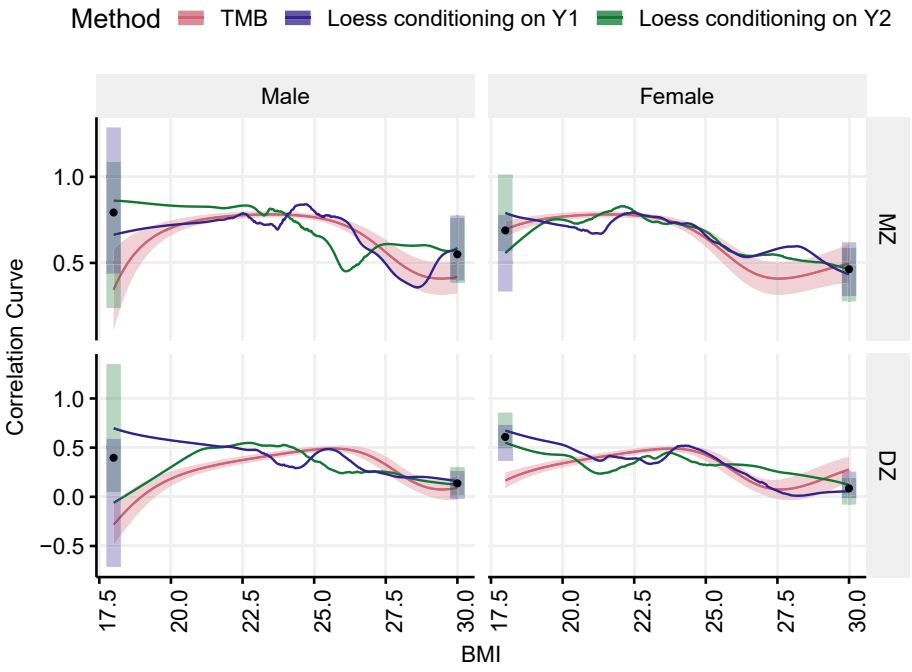


Figure 7: Comparison of estimated correlation curves based on the Gaussian mixture (TMB; red) with the two fully nonparametric estimates (blue and green) obtained from "loess". For the Gaussian mixture the estimation uncertainty is displayed for the full range of BMI values, while for the two nonparametric curves 95% confidence intervals are displayed only for BMI=18 and 30. The black dot represents the average of the pooled 100 bootstrap samples.

relation curves. Due to the arbitrary labeling of twin pair members these curves are estimates of the same quantity, and differ only due to estimation uncertainty. As we are estimating a conditional mean curve, we will refer to "regressing $Y1$ on $Y2$" as "conditioning on $Y2$". We show a simplified version of the code in appendix B.

Looking at Figure 5, 6 and 7 we see that, overall, female curves among different models seem more similar to each other compared to those of males. A similar observation can be made for dizygotic curves compared monozygotic ones. We remark that the sample sizes of the subsets are quite different; there are about 30% more female subjects than male subjects, and the size of the dizygotic subset is double the size of the monozygotic subset. The tail behaviour is sometimes discordant with the previous remark, especially in Figure 7. We investigate this phenomenon later.

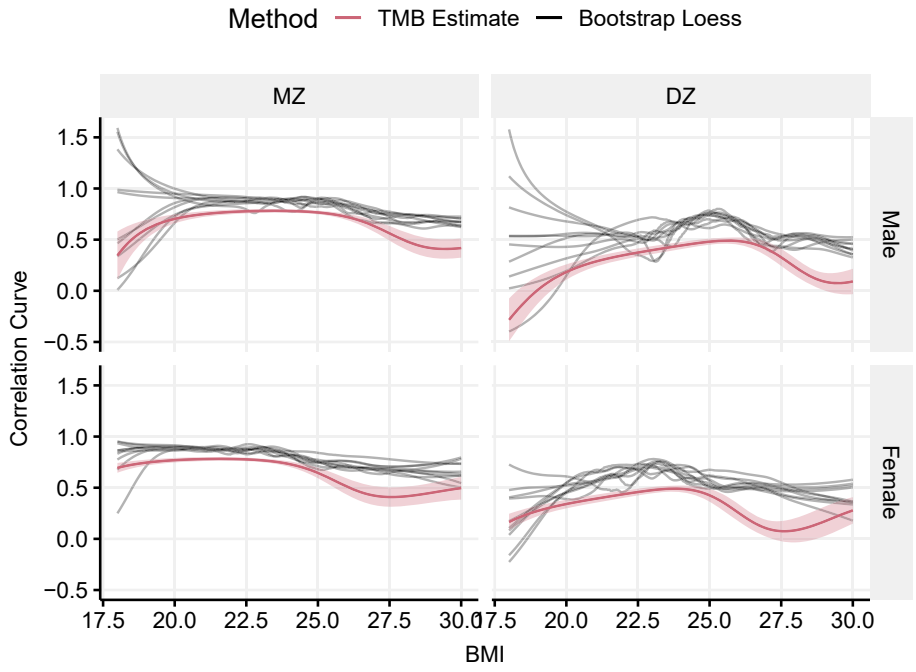


Figure 8: Fully nonparametric correlation curves estimated on simulated data (in black) compared to the parametric TMB estimate (in red). The estimation uncertainty is displayed only for the Gaussian mixture.

Figure 5 shows the conditional means obtained using TMB and "loess" conditioning on both twin pairs separately. The three curves identify the same pattern in the data, although they do not coincide perfectly along the entire BMI range. We observe a larger variance in the tails, which can be expected, given the sparsity of the dataset.

The two nonparametric conditional variance curves lie partly outside the pointwise 95% confidence intervals for the Gaussian mixture (Figure 6), particularly in the tails. There is also a discrepancy between the two nonparametric curves, and this discrepancy can be interpreted as estimation uncertainty, because the blue and green curves are the same estimator applied to two different datasets (switching the roles of $Y1$ on $Y2$).

To compute the derivative of the conditional mean curve needed for the correlation curve, we used the finite difference approximation

$$\beta(y) = \frac{\mu(y+h) - \mu(y-h)}{2h}, \tag{9}$$

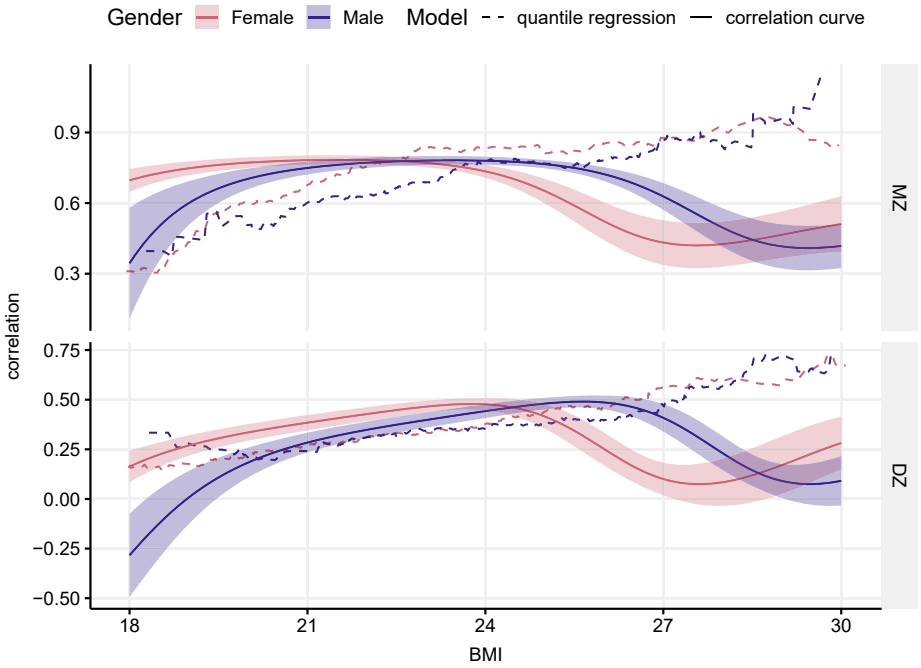


Figure 9: Comparison of quantile regression (dashed) with the correlation curve in Figure 3 of the main text (see caption of that figure).

for $h = 0.1$, where $\mu(y)$ is the conditional mean obtained from "loess". In Figure 7 we show the loess estimate of the correlation curve. Overall, they show the same results as the TMB estimate. The behaviour of the curves differ mostly in the tails. We investigate this, and in particular the left-hand tail for dizygotic females, in more detail below.

To this end we generated 100 bootstrap samples from the dataset, sampling twin pairs with replacement and randomly reassigning the order of the twin members. For each of the 100 bootstrap datasets we computed the nonparametric correlation curve, and subsequently calculated bootstrap mean and standard deviation. Since the largest difference between correlation curves is in the tails, Figure 7 displays the bootstrap results only for BMI=18 and 30. The 95% bootstrap confidence intervals are constructed using 1.96 standard deviations around the curve.

The confidence intervals for male data at BMI=18 are very wide; this is probably a reflection of the few data points around that value in the original dataset. Looking at the female dizygotic data at BMI=18, neither confidence intervals (blue or green) overlap with the TMB estimate. Hence, the differences cannot be described by uncertainty alone

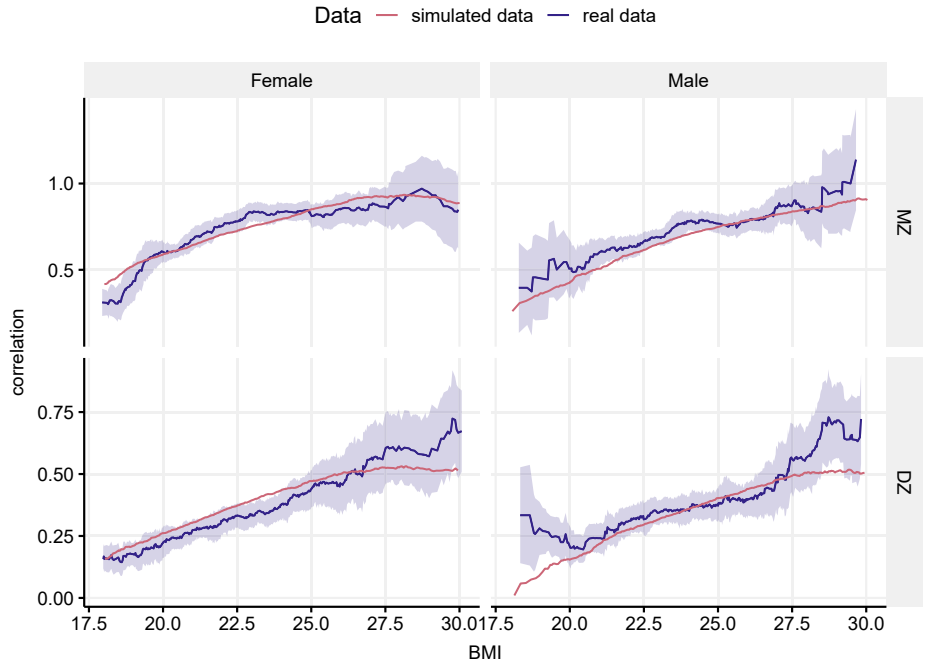


Figure 10: Quantile regression of correlation on the same data as Figure 3 of the main text, with a 95% confidence interval (in blue) and on simulated data (in red).

and may be attributable to either lack of flexibility of the Gaussian mixture (TMB) or to some bias in the nonparametric estimator.

To investigate the statistical properties of the nonparametric estimator, we conducted a simulation experiment in which data were simulated from the fitted Gaussian mixture (based on the real data). Because an analytic expression for the correlation curve is available for Gaussian mixtures [Berentsen et al., 2020] we can compare the nonparametric estimator with the "true" correlation curve. Figure 8 shows that the nonparametric estimator has an upwards bias throughout the BMI range. Strictly speaking, this conclusion is valid only when the true data generating mechanism is a Gaussian mixture, but it seems necessary to explore the properties of the nonparametric estimator further.

7 Comparison with quantile regression

An alternative to our correlation curve based method, which allows heritability to vary with BMI, is quantile regression [Williams, 2020, Mitchell et al., 2013, Beyerlein et al., 2011, Abadi et al., 2017, Rokholm et al., 2011]. Figure 9 compares the correlation curves with the quantile regression of correlation. The two methods agree fairly well in the lower half of the BMI range, but in the upper half the correlation curve decreases, while the quantile regression estimate keeps increasing (roughly speaking).

We next attempt to understand if this discrepancy is due to the use of a Gaussian mixture in the current analysis, or if it is caused by the correlation curve and quantile regression yielding fundamentally different measures of correlation. Figure 10 shows the quantile regression both for the real data (same as Figure 9) and simulated data. The simulated data were generated from the estimated Gaussian mixture of the current analysis (see main text), and a sample size of $100,000$ twin pairs was used in order to minimize estimation uncertainty. The curves for the simulated and real data of Figure 10 are fairly similar, indicating that the use of a Gaussian mixture is not the source of the difference we see in Figure 9. Hence, the conclusion is that the discrepancies in Figure 9 must be caused by the correlation curve and quantile regression being different measures of correlation, at least for the data in this study. This finding needs further investigation.

A simplified version of the code used to generate Figure 9 and 10 can be found in Appendix C.

A Data preprocessing, R code

Below, we show a simplified version of the preprocessing code and we apply it to a dummy dataset. The dataset contains four measurements of two twins belonging to the same pair.

```
 Dataset
##   Tvparnr Twinnumber Twinid Wave Age Sex Zygoti  BMI
## 1       1          1     11    1  20   1      1 21.3
## 2       1          1     11    2  30   1      1 22.5
## 3       1          1     11    3  37   1      1 23.2
## 4       1          1     11    4  47   1      1 25.1
## 5       1          2     12    1  20   1      1 21.6
```

```
## 6       1          2     12    2  30    1       1 22.4
## 7       1          2     12    3  37    1       1 22.7
## 8       1          2     12    4  47    1       1 23.6
library(tidyverse)
value<-c()
newdat<-data.frame(Age=35)
IDs<-unique(Dataset$Twinid)
for(j in IDs) {
   Subset<-Dataset%>%filter(Twinid==j)
   if(dim(Subset)[1]>2){
   lm_mod<-lm(BMI~Age, data = Subset)
   }
 P<-predict(lm_mod, newdata = newdat)
 value<-c(value, P)
}

value
       1        1
## 23.23226 22.68316
```

The code separates the dataset depending on the variable `Twinid`, checks if the number of waves is larger than two (to reduce uncertainty we only consider twin pairs with three or more measurements), and performs a linear regression on said dataset. The output `value` is the predicted BMI value at age 35 and forms the dataset used in the analysis.

B Nonparametric correlation curve, R code

The code below is a simplified version of the code used to compute the nonparametric correlation curve. Assume that the dataset called `Dataset` contains data from one single zygosity and one sex (for example, only male monozygotic twins). The code shows how we compute conditional variance and the derivative of the conditional mean using `loess`. We only condition on `Y2`, since conditioning on `Y1` follows the same procedure, only with the role of `Y1` and `Y2` switched.

The quantity `Fit_square` corresponds to $\mathbb{E}(Y1^2 \mid Y2)$. To compute the conditional variance, we then calculate $\mathbb{E}(Y1^2 \mid Y2) - \mathbb{E}(Y1 \mid Y2)^2$.

To compute the derivative of the conditional mean, we use equation 9 for $h = 0.1$.

```
library(tidyverse)
vec<-seq(18, 30, length.out =500)
newdata<-data.frame(First=vec,Second=vec,
                    Firstsq=vec*vec, Secondsq=vec*vec)
new_loess<-function(data, formula){
      loess(formula, data,
               control=loess.control(surface="direct"))
}
## conditional mean
CM<-Dataset%>%new_loess(First~Second)%>%
  predict(newdata=newdata)
## conditional variance
Fit_square<-Dataset%>%new_loess(Firstsq~Second)%>%
  predict(newdata=newdata)
CV<-Fit_square-CM*CM
## derivative of the conditional mean (beta)
h<-0.1
newdata_plus<-newdata+h
newdata_minus<-newdata-h
Fit_plus<-Dataset%>%new_loess(First ~Second)%>%
  predict(newdata=newdata_plus)
Fit_minus<-Dataset%>%new_loess(First ~Second)%>%
  predict(newdata=newdata_minus)
Beta<-(Fit_plus-Fit_minus)/(2*h)
## correlation curve
Sigma_Beta<-sd(Dataset$Second)*Beta
Sigma_Beta_squared<-Sigma_Beta*Sigma_Beta
Correlation_curve<-Sigma_Beta/sqrt(Sigma_Beta_squared+CV)
```

C Quantile regression

Below, we show a simplified version of the R code for quantile regression that was used to produce Figure 9 and 10. For brevity, we only show code for male, monozygotic twin pairs. The code for the other subpanels are very similar.

```
library(tidyverse)
library(quantreg)
## We select male, monozygotic data
MMZ_dat<-Dat%>%filter(Gender=="Male", Zygosity=="MZ")%>%
                select(First, Second)
## We switch the order of the twins in some twin pairs, at random
set.seed(31415)
index<-sample(size=nrow(MMZ_dat), x=c(1, 2),replace= TRUE)
MMZ_dat_s<-MMZ_dat
for(i in nrow(MMZ_dat)) {
   if(index[i]==2){
   MMZ_dat_s$First[i] = MMZ_dat$Second[i]
   MMZ_dat_s$Second[i] = MMZ_dat$First[i]
   }
}
## We define a vector of quantiles
quant<-seq(0.1, 0.9, length.out=500)
## We normalize the data and perform quantile regression
v.mmz<-as.matrix(MMZ_dat_s)
dfMMZ<-data.frame((v.mmz- mean(v.mmz))/sd(v.mmz))
modelMMZ <-  rq(Second ~ First , tau = quant, data = dfMMZ)
## We access to coefficients and standard errors, used in the plot
summary(modelMMZ)
## We transform the quantiles back to BMI scale
xMMZ<-quantile(x=dfMMZ[,1], prob=quant)*sd(v.mmz)+mean(v.mmz)
```

Quantile regression, just like `loess`, can produce different results whether we condition on `Y1` or `Y2`. To create a curve that can be compared to the correlation curve, for each twin pair we randomly choose whether to switch the position of the twins inside the pair. The resulting dataset `MMZ_dat_s` was then used in the quantile regression.

The object `modelMMZ` contains the coefficient estimate for the regression for each quantile in `quant`. The summary of the model allows us also to access to standard errors, which are used in the plot to produce the 95% confidence interval. We create the quantile vector by starting in the BMI scale so to have evenly-spaced points when we convert the results back to the BMI scale. Due to approximation in the `rq` function, we might lose some precision in the tails.

References

A. Abadi, A. Alyass, S. R. du Pont, B. Bolker, P. Singh, V. Mohan, R. Diaz, J. C. Engert, S. Yusuf, H. C. Gerstein, et al. Penetrance of polygenic obesity susceptibility loci across the body mass index distribution. *The American Journal of Human Genetics*, 101(6):925–938, 2017.

G. D. Berentsen, F. Azzolini, H. J. Skaug, R. T. Lie, and H. K. Gjessing. Heritability curves: A local measure of heritability in family models. *Statistics in Medicine*, 2020.

A. Beyerlein, R. von Kries, A. R. Ness, and K. K. Ong. Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. *Plos one*, 6(4):e19057, 2011.

ISO/IEC. Programming languages — c++. Draft International Standard N4660, March 2017. URL `https://web.archive.org/web/20170325025026/http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2017/n4660.pdf`.

J. Kaprio, S. Bollepalli, J. Buchwald, P. Iso-Markku, T. Korhonen, V. Kovanen, U. Kujala, E. K. Laakkonen, A. Latvala, T. Leskinen, et al. The older Finnish twin cohort—45 years of follow-up. *Twin Research and Human Genetics*, 22(4):240–254, 2019.

K. Kristensen, A. Nielsen, C. W. Berg, H. J. Skaug, and B. M. Bell. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21, 2016. doi: 10.18637/jss.v070.i05.

C. McCulloch and J. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

J. A. Mitchell, H. Hakonarson, T. R. Rebbeck, and S. F. A. Grant. Obesity-susceptibility loci and the tails of the pediatric BMI distribution. *Obesity*, 21(6):1256–1260, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

B. Rokholm, K. Silventoinen, L. Ängquist, A. Skytthe, K. O. Kyvik, and T. I. A. Sørensen. Increased genetic variance of BMI with a higher prevalence of obesity. *PloS one*, 6(6):e20816, 2011.

P. T. Williams. Quantile-dependent heritability of computed tomography, dual-energy X-ray absorptiometry, anthropometric, and bioelectrical measures of adiposity. *International Journal of Obesity*, 44(10):2101–2112, 2020.

Article III

6.3 Exploring the likelihood surface in multivariate Gaussian mixtures using Hamiltonian Monte Carlo

Francesca Azzolini, Hans J. Skaug

arXiv, https://doi.org/10.48550/arXiv.2308.14700 (2023)

Exploring the likelihood surface in multivariate Gaussian mixtures using Hamiltonian Monte Carlo

Francesca Azzolini[1] and Hans J. Skaug[1]

[1]Department of Mathematics, University of Bergen, Bergen,Norway.

Abstract

Multimodality of the likelihood in Gaussian mixtures is a well-known problem. The choice of the initial parameter vector for the numerical optimizer may affect whether the optimizer finds the global maximum, or gets trapped in a local maximum of the likelihood. We propose to use Hamiltonian Monte Carlo (HMC) to explore the part of the parameter space which has a high likelihood. Each sampled parameter vector is used as the initial value for quasi-Newton optimizer, and the resulting sample of (maximum) likelihood values is used to determine if the likelihood is multimodal. We use a single simulated data set from a three component bivariate mixture to develop and test the method. We use state-of-the-art HCM software, but experience difficulties when trying to directly apply HMC to the full model with 15 parameters. To improve the mixing of the Markov Chain we explore various tricks, and conclude that for the dataset at hand we have found the global maximum likelihood estimate.

1 Introduction

Mixture distributions are linear combinations of probability densities (called components), where the weights of the sum add to one. Mixtures occur naturally for datasets that are comprised of multiple populations, but more generally they are a flexible mechanism for generating probability distributions in dimension $r \geq 1$. The most popular mixture distribution, and the focus of this paper, is the Gaussian mixture, with density

$$f(\boldsymbol{x}) = \sum_{k=1}^{m} p_k N_r(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{x} = (x_1, \ldots, x_r)$, N_r is the r-dimensional Gaussian density with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ for $k = 1, ..., m$, and the p_k's are the weights of the mixture ($\sum_{k=1}^{m} p_k = 1$). The parameters of the mixture, which will be estimated by maximum likelihood, are $p_1, ..., p_m$, $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$.

The most widely used estimation method for Gaussian mixtures is the iterative Estimation-Maximization (EM) algorithm [Dempster et al., 1977]. The EM algorithm requires initialization of the parameters, and these initial values influence the performance, and potentially the result, of the algorithm. While it can be proven that the EM algorithm, if allowed enough iterations, will reach a local maximum of the likelihood function [Wu, 1983], the computational time might be too high from a practical perspective. Moreover, the EM algorithm does not necessarily find the global maximum, but only a local one. The choice of initial parameter values is crucial in reducing these issues, and several approaches to finding some suitable initial values have been proposed. The earliest idea was to perform a grid search on the initial values [Laird, 1978], while later approaches prefer to perform a pre-clustering of the data to identify the separate components. Many clustering methods have been proposed, from K-means and hierarchical clustering [Shireman et al., 2017], to a shorter run of the EM algorithm itself [Baudry and Celeux, 2015].

The dependence on initial values carries over to other algorithms for maximizing the likelihood. In the current paper we use a quasi-Newton algorithm in combination with automatic differentiation for numerical evaluations of gradient and Hessian of the log-likelihood [Berentsen et al., 2021, Azzolini et al., 2022]. Although this is a numerically robust and efficient estimation algorithm, it may be trapped in a local optimum of the likelihood, if present. To solve this problem we suggest to use a Hamiltonian Monte Carlo (HMC) [Duane et al., 1987] to sample from the parameter space, and to use the resulting samples as inital values for the quasi-Newton algorithm. This will amount to doing a grid search with an irregular grid, using a finer mesh in regions where the likelihood is high.

To illustrate this approach we use a single simulated dataset from the three component bivariate Gaussian mixture fitted in Azzolini et al. [2022]. The software Stan [Stan Development Team, 2019] is used, via the interface tmbstan [Monnahan and Kristensen, 2018], to perform the HMC sampling. We discuss different implementation tricks needed in order for the sampler to work properly.

In Section 2 we introduce the HMC algorithm and explain its advantages. We then present the simulated dataset used in the following sections and explain the code used to perform our tests. In Section 3 we show our preliminary results and the issues that

we encounter while running a off-the-shelf HMC sampling. In Section 4 we propose three changes to the sampler to fix its major issues and we highlight their advantages and drawbacks. We moreover use these three new approaches to explore the parameter space in search of new initial values. Lastly, in Section 5 we draw some conclusions about the dataset we analyzed and highlight the potential of this approach.

2 HMC and its implementation

2.1 Hamiltonian Monte Carlo

The HMC algorithm is a Markov chain Monte Carlo method which was conceived as an alternative to the Metropolis-Hastings algorithm with the goal of being more efficient in sampling from posterior distributions. HMC applies the laws of physics to constrain the sampler to regions of the parameter space with high posterior density, often referred to the "typical set". It does so by introducing a set of auxiliary parameters, that are referred to as "momenta", and using these to create a vector field that is aligned with the typical set. This vector field is generated following Hamilton's equations [Betancourt, 2017], which give the algorithm its name. Our goal is to use HMC to sample the parameter space in proportion to the value of the likelihood function, but of course, by adopting flat priors on all parameters the likelihood may be viewed as a posterior distribution.

Hamilton's equations are partial differential equations that relate position and momentum of particles in space to the total energy of the system. In our setting, the "positions of the particles in space" are the current values of the parameters in the parameter space. Paired with fictitious momenta, they live inside the so-called phase space, where we can generate Hamilton trajectories - that is, trajectories which follow the vector field generated by Hamilton's equations. The gradient of log-likelihood that is used to solve Hamilton's equations is the same as the one used by the quasi-Newton optimization algorithm.

The HMC method is an iterative algorithm that at each iteration selects a new momentum (chosen stochastically) and pairs it to the current value in the parameter space; this allows us to follow the Hamilton trajectories in the phase space for a predetermined amount of time. The momenta are then discarded, projecting the pair (position, momentum) back onto the parameter space, and the "position" estimate that is reached is the sample of said iteration. As we will invoke the HMC algorithm only via the interface tmbstan, we do not need to go in more detail, but a deeper dive into Hamilton's equations and their use in MCMC algorithms can be found in Betancourt [2017] and

Neal [1993].

Symplectic integrators [Donnelly and Rogers, 2005] are a category of approximators built specifically for estimating the solution of Hamilton's equations. Among these, in this paper we use the Leapfrog integrator [Betancourt, 2017]. Several choices must be made when implementing the Leapfrog integrator: among these, the number of steps between one sample and the other, and the size of each such step. The No-U-Turns (NUTS) algorithm [Hoffman et al., 2014] is a variation of the Leapfrog integrator which automatically optimizes the number of steps in each iteration.

Sampling from around the typical set is equivalent, if the distribution in analysis is a unimodal distribution, to sampling mostly around the mode. The log likelihood of mixtures, however, can often be multimodal. This proves to be an issue, that is further explored in Sections 3 and 4.

2.2 A simulated dataset

For the purpose of studying the ability of HMC to explore the parameter space of Gaussian mixtures, we use a simulated dataset which resembles the twin data in Azzolini et al. [2022]. The dataset contains $n = 1200$ bivariate observations (x_1, x_2), that we can interpret as measurements of some trait measured on twin pairs. Of these 1200 twin pairs, half are same-sex male pairs, and half are same-sex female pairs. We also divide these data by zygosity of the twin pairs: two thirds are dizygotic (DZ), while the last third are monozygotic (MZ).

The dataset is simulated from the Gaussian mixture with $m = 3$ components, which was the best fitting model in Azzolini et al. [2022]. Twins within a pair have identical means and standard deviations, e.g. $\boldsymbol{\mu}_k = (\mu_k, \mu_k)$ (for $k = 1, ..., 3$). MZ and DZ groups share all parameters, except for the correlation coefficient (that we denote with $\rho^{(MZ)}$ and $\rho^{(DZ)}$, respectively). We hence define the covariate matrices as

$$\boldsymbol{\Sigma}_k^{(MZ)} = \sigma_k^2 \cdot \begin{pmatrix} 1 & \rho_k^{(MZ)} \\ \rho_k^{(MZ)} & 1 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_k^{(DZ)} = \sigma_k^2 \cdot \begin{pmatrix} 1 & \rho_k^{(DZ)} \\ \rho_k^{(DZ)} & 1 \end{pmatrix}.$$

We also assume that male and female data are generated using the same parameters, except for the mean vector, where $\mu_k^M = \mu_k^F + \beta$, where β is a common parameter between all components. The true values of the parameters of the Gaussian mixture which generated the dataset can be found in Table 1.

2.2.1 The likelihood function

Since the MZ and DZ groups have different parameter values, we keep their contributions to the log likelihood separate. Let us denote with n^{MZ} the number of MZ pairs, and with n^{DZ} the number of DZ pairs. Then, the log likelihood corresponding to the model from which data are generated is:

$$\begin{aligned}\log L(\boldsymbol{\theta}) = &\sum_{i=1}^{n^{MZ}} \log \left\{ \sum_{k=1}^{3} p_k \boldsymbol{N}_2(\boldsymbol{x}_i; \boldsymbol{\mu} = \boldsymbol{\mu}_k + C_i\boldsymbol{\beta}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k^{MZ}) \right\} \\ &+ \sum_{i=1}^{n^{DZ}} \log \left\{ \sum_{k=1}^{3} p_k \boldsymbol{N}_2(\boldsymbol{x}_i; \boldsymbol{\mu} = \boldsymbol{\mu}_k + C_i\boldsymbol{\beta}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k^{DZ}) \right\},\end{aligned} \tag{1}$$

where

$$C_i = \begin{cases} 0, & \text{if the sex of the ith individual is female} \\ 1, & \text{if the sex of the ith individual is male} \end{cases}$$

and $\boldsymbol{\theta}$ is the parameter vector containing the 15 parameters that describe the likelihood: three means μ_k, three standard deviations σ_k, three MZ correlation coefficients $\rho_k^{(MZ)}$, three DZ correlation coefficients $\rho_k^{(DZ)}$, one sex covariate β, and two weights p_k (with the third one being defined as $p_3 = 1 - p_1 - p_2$).

When estimating these parameters we must constrain their ranges to only meaningful values, e.g. standard deviations must be non-negative. For this reason, when implementing a MLE code, we rather estimate $\log(\sigma)$ instead of σ. For the same reason we estimate only two of the three weights, and then apply a transformation that produces the third one and normalizes them to sum up to one.

A major issue with mixtures is label-switching, that is the randomness in assigning the label to the components. This means that two virtually identical mixture estimates can be treated as different because the labels of the components are switched around. To prevent this issue, we order the means from lowest to highest by reparametrizing them as a sum of exponentials: $\mu_1 = \exp(\alpha_1)$, $\mu_2 = \mu_1 + \exp(\alpha_2)$, and $\mu_3 = \mu_2 + \exp(\alpha_3)$.

The problem of maximizing the log likelihood is identical to that of minimizing the negative log likelihood. Since the software we work on are implemented to solve the latter problem, for the rest of the paper we will talk about negative log likelihood instead.

2.3 HMC using TMB and Stan

We use the C++ [ISO/IEC, 2017] code for the log-likelihood used in Azzolini et al. [2022] which is linked into the R package TMB [Kristensen et al., 2016]. The R routine nlminb is used to maximize the likelihood (1), and TMB is used to calculate both the gradient and Hessian matrix of the objective function using automatic differentiation. The use of both first and second order derivatives makes the quasi-Newton method that is built into nlminb numerically stable [Azzolini et al., 2022]. Throughout the paper we will refer to the maximum likelihood estimate as "nlminb". We use box constraints in nlminb to limit the parameter space. This comes in addition to the reparameterizations mentioned above.

Via the R-package tmbstan the objective function is sent to Stan, which executes the HMC sampling algorithm. Stan includes a variety of options for symplectic integrators and the number of steps in the integrators themselves. We pick the NUTS algorithm, and we set the step size to 0.95, which is the default value.

TMB has a parameter MAP which controls which parameters should be estimated, or fixed at particular values. This mechanism will be used to sample from reduces models.

3 Base HMC approach

As a first step we run the R and C++ codes to obtain an estimate of the parameters via the optimizer nlminb. The results are listed in Table 1. We use these estimates both as starting values for the HMC sampling and as comparison.

Nlminb does a good job estimating the parameters, as it can be seen by comparing its output to the true values. The parameters that nlminb struggles the most to estimate correctly are the correlation coefficients (see the standard errors in Table 1), especially the second and third component of the correlation vector.

We run tmbstan using the nlminb estimates as starting values. Notice that we must generate a new MakeADFun object with the nlminb estimates as initial values, that will be used as input for the HMC algorithm. We perform 1000 iterations, and set the warm-up iterations to 500 (as per default, half of the total amount). The seed is chosen randomly. The parameter adapt_delta is set to 0.95.

The overall behavior of the samples can be seen in the traceplot of Figure 1. A traceplot visualizes the development of the samples at each iteration. The warmup iterations are

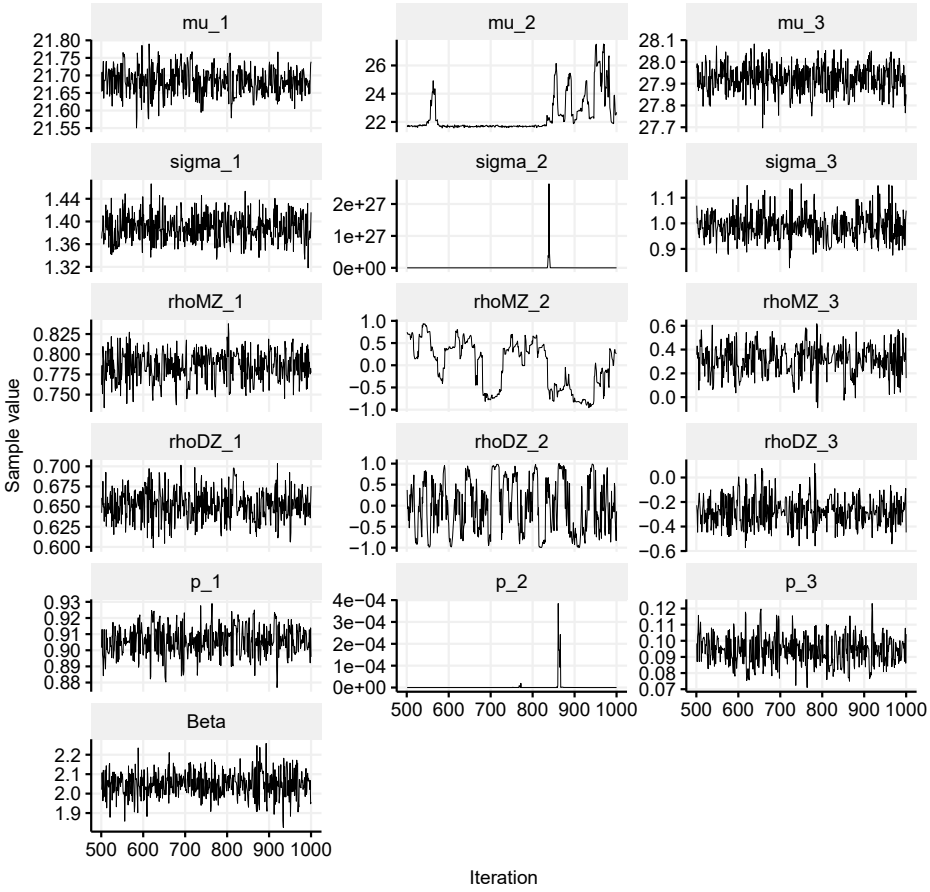


Figure 1: Traceplot of the samples produced by the HMC algorithm.

not displayed. To make the plot easier to interpret, we visualize the already transformed parameters (that is, for example, we visualize μ instead of α).

We can immediately notice how differently the estimates of the parameters of the first and third components behave compared to the second component. The behaviors of the latter are overall less stable, with spikes that go well beyond the normal range of the parameter (e.g. σ_2) or oscillating wildly between all admissible values (e.g. the correlation coefficients, which oscillate between -1 and 1). The parameter estimates of the first and third components seem overall more stable and converging to a sensible value for the parameter. To explore this behavior in more detail, we collect averages and standard deviations for each parameter estimate in Table 1.

Par	True	nlminb				HMC algorithm	
		first run		2-component runs		base approach	
		mean	std. dev.	mean	std. dev.	mean	std. dev.
μ_1	21	21.07	0.09	21.69	0.09	21.68	0.04
μ_2	23	23.09	0.16	/	/	22.42	1.32
μ_3	28	27.93	0.06	27.46	1.60	27.92	0.06
σ_1	1	1.03	0.04	1.58	0.73	1.39	0.02
σ_2	1	1.04	0.06	/	/	8.36e+24	1.22e+26
σ_3	1	0.97	0.05	1.00	0.11	0.99	0.05
$\rho_1^{(MZ)}$	0.7	0.69	0.04	0.80	0.05	0.79	0.02
$\rho_2^{(MZ)}$	0.5	0.46	0.10	/	/	0.00	0.54
$\rho_3^{(MZ)}$	0.3	0.34	0.14	0.35	0.13	0.31	0.12
$\rho_1^{(DZ)}$	0.4	0.46	0.05	0.67	0.07	0.65	0.02
$\rho_2^{(DZ)}$	0.3	0.14	0.14	/	/	0.00	0.61
$\rho_3^{(DZ)}$	-0.2	-0.30	0.11	-0.27	0.25	-0.28	0.12
p_1	0.60	0.63	0.05	0.87	0.17	0.91	0.01
p_2	0.30	0.27	0.05	/	/	0.00	0.00
p_3	0.10	0.09	0.01	0.13	0.17	0.09	0.01
β	2	1.97	0.06	2.05	0.02	2.05	0.07

Table 1: Mean and standard deviations of parameter estimates for simulated dataset. The column contains the following results: "True" contains the true values of the underlying distribution; "first run" contains nlminb estimates; "2-component runs" contains nlminb estimates under the two-component assumption described in Subsection 4.3; "base approach" contains average and standard deviations of the samples generated via a base HMC algorithm.

This table confirms the initial observations we made looking at the traceplot. The spike shown in the plot of σ_2 is translated in a very large average and standard deviation. On the other side, the mean parameters are comparable to the true values, although the second component has a relatively large uncertainty.

Notice that the estimate of the second weight of the mixture, p_2, is, with no uncertainty, zero. This is visible in the traceplot as well: most of the samples lie around a value of 0, with the only spike reaching a value of 0.0004. This means that, virtually, the samples that HMC has produced belong to a Gaussian mixture with only two components. While, in a general setting, this might be a sign that the initial assumption regarding the number of components might be wrong, we are aware that the mixture distribution generating the dataset has three distinct components.

Despite this being the case, HMC seems to strongly prefer a solution with fewer components than the number suggested by the initial values it is given. To remedy for the lack of one extra components, the first one compensates by having a larger mean value (very close to the average of the first two means of the generating mixture) and with a

larger standard deviation.

We can compare global quantities - that is, measures of the parameters of the global distribution, as defined in Berentsen et al. [2021] - of the true distribution and the average of the HMC samples. The global mean of the true distribution is the weighted sum of the three true means, that is 22.3. When estimating the global mean for the HMC sample, only the first and third component contribute, because the second weight is zero. This is true for the global quantities of all parameters. The average of the global quantities of the distributions sampled by HMC are collected in Table 2.

We notice that the averages of the global quantities obtained via HMC sampling are on a comparable scale to the global quantities of the true distribution, suggesting that HMC identifies the global distribution underlying the dataset, but struggles when attributing the correct values to the separate components.

There exists multiple criteria to judge whether a Markov Chain Monte Carlo method has converged or not. tmbstan provides automatically the Rhat convergence diagnostic, which compares within- and between estimates. Since we run tmbstan with a single chain, Rhat is not a recommended diagnostic. As alternative, one can look at the Geweke diagnostic [Geweke et al., 1991]. According to this diagnostic, several chains do not converge, among which many of the parameters describing the second component (the Z-scores are listed in Table 1 of the supplementary material). It is important to mention, however, that the purpose of this paper is not to verify the convergence of the algorithm; the goal is to find a better negative log likelihood value, and that can be found independently on whether the chains have converged or not. This will be further explored in Section 4.2.

Experiments show this behavior also for other simulations: in particular, when working with a dataset generated by a Gaussian mixture with two components, HMC tends to estimate one of the two weights to zero and hence de-facto revert to a Gaussian

Model	μ	σ	$\rho^{(MZ)}$	$\rho^{(DZ)}$
True value	22.30	2.33	0.92	0.87
nlminb	22.02	2.25	0.92	0.86
base HMC	22.24	2.24	0.92	0.86
loop HMC	22.54	2.33	0.92	0.86
MAP HMC	-	2.25	0.92	0.86
bounds HMC	22.53	2.33	0.92	0.86

Table 2: Global quantities of the different models studied in this paper. The first row lists the true quantities, calculated from the true values of the generating mixture, and should be treated as reference.

distribution.

This sampling depends on the random seed that is given to tmbstan. Other random seeds have produced samples that collapse three components into a single non-zero weighted component, whose parameters are comparable to the global quantities of the true distribution. A table collecting one such example can be found in the supplementary material (Table 2).

Another common occurrence is to estimate the means of two components (say, the first and the second), as identical: the sampler estimates α_2 as a very large negative number, hence $\exp(\alpha_2) \approx 0$, and $\mu_2 = \mu_1 + \exp(\alpha_2) \approx \mu_1 + 0 = \mu_1$.

Some specific (and relatively rare) random seeds have also produced the desired three distinct non-zero components; a longer discussion about this can be found in Subsection 4.1.1.

This model is clearly not optimized; a confirmation can be obtained by looking at the negative log likelihood. Using TMB, we estimate the negative log likelihood for each HMC sample. The HMC sampling method performs significantly worse than nlminb, reaching a minimum value of 4103.04 against the nlminb value of 4070.52 (see Table 3).

4 More refined approaches

4.1 The fixes

In this section we show three variations of the HMC sampling algorithm that successfully return samples from three distinct, non-zero Gaussian mixtures. The efficiency of these methods varies, and we present them starting with the least efficient. Each of these approaches comes with some restrictions that require to be discussed.

model	nlminb	base	loop	MAP	bounds
nll	4070.52	4103.04	4071.53	4071.32	4072.46

Table 3: minimum value of the negative log likelihood from the samples collected using all the approaches described in this paper. The nlminb value is reported as a comparison.

Par.	True val.	loop		MAP		bounds	
		mean	std. dev.	mean	std. dev.	mean	std. dev.
μ_1	21	21.06	0.09	-	-	21.06	0.11
μ_2	23	23.05	0.18	-	-	23.04	0.20
μ_3	28	27.93	0.06	-	-	27.92	0.06
σ_1	1	1.03	0.05	1.03	0.03	1.03	0.05
σ_2	1	1.07	0.07	1.05	0.04	1.07	0.07
σ_3	1	0.99	0.06	0.98	0.05	0.98	0.05
$\rho_1^{(MZ)}$	0.7	0.69	0.04	0.69	0.04	0.69	0.04
$\rho_2^{(MZ)}$	0.5	0.47	0.10	0.45	0.08	0.46	0.11
$\rho_3^{(MZ)}$	0.3	0.32	0.13	0.32	0.13	0.32	0.14
$\rho_1^{(DZ)}$	0.4	0.46	0.05	0.46	0.04	0.47	0.06
$\rho_2^{(DZ)}$	0.3	0.18	0.13	0.15	0.08	0.17	0.15
$\rho_3^{(DZ)}$	-0.2	-0.29	0.11	-0.29	0.11	-0.28	0.11
p_1	0.6	0.62	0.05	0.63	0.02	0.62	0.06
p_2	0.3	0.29	0.05	0.27	0.02	0.29	0.06
p_3	0.1	0.10	0.01	0.09	0.01	0.10	0.01
β	2	2.07	0.06	2.07	0.06	2.07	0.06

Table 4: Mean and standard deviation of the samples of the parameters under the three alternative approaches, from left to right: repeating the sampling 15 times; keeping the mean parameter fixed; setting boundaries on the parameters. The first column lists the true values as reference.

4.1.1 Trying different seeds for the random number generator

The first approach that we present relies on brute force. As mentioned in Section 2, tmbstan requires a random seed to explore the parameter space. Different initial random seeds can result in vastly different samples. While the majority of our experiments returned a non-optimal result (as described in Section 3), some random seeds produced a set of samples which belonged to a non-trivial three-component Gaussian mixture.

This first approach, then, simply consists in repeating the HMC sampling several times, with a different random seed each time. At the end of each completed sampling, we save the output only if the negative log likelihood is lower than the one obtained using the previous seed. In our example, we repeat the HMC sampling fifteen times, and we obtain at least one result with three distinct, non-zero components. The best results from this sampling are collected in Table 4. The averages of the parameters are comparable to the true values, and the standard deviations are reasonable and comparable to the standard errors obtained via TMB.

This approach relies on repeating multiple times an already lengthy process, and is the slowest among the three methods we suggest in this section. Moreover, there is a

component of randomness in this result as well: as we mentioned in Section 3, the most common results collapse two or even all three components into one, so fifteen random seeds might not be enough to produce one sample from three distinct components.

It is still important to discuss this result, because it proves that HMC can, potentially, identify three separate components, even though it struggles to do so.

4.1.2 Fixing the value of a subset of parameter

The second approach that we present consists in fixing some parameters to their nlminb values during the entire sampling process. In this way, HMC receives parameters describing distinct components, and hopefully it will sample the other parameters accordingly. We use the argument MAP in the MakeADFun object that is used as input in tmbstan.

When applying this approach, we must choose a subset of parameters to keep fixed. The (maybe obvious) choice of fixing the weights does not provide consistent distinct components: two means are often estimated as the same identical value, practically collapsing these two components into one.

While it is not reflected in the specific example shown in Table 1, the mean parameters tend to behave quite erratically in the base approach. Two such examples are shown in Table 2 of the supplementary material.

We show the results of this approach in Table 4. The mean parameters are not reported, since they are not sampled via HMC. This approach identifies three distinct components and the parameter estimates are comparable to the true values of the underlying distribution. Moreover, the standard deviations of the estimates from this approach are smaller compared to the other two presented approaches.

This algorithm uses nlminb estimates as reference for a set of parameters (the mean values in this specific case), that are not sampled in the HMC iterations. Note that, in this example, the nlminb estimates that we inherit are very similar to the real values, and this can strongly impact the results of this analysis.

Moreover, the smaller standard deviations seem to suggest that keeping some parameters fixed prevents the other parameters from assuming very unexpected values. This implies that the samples won't deviate much from the original nlminb estimates, and if those weren't the optimal ones, it would be very difficult for HMC to find a lower negative log likelihood.

There are other variations of this approach that we can explore: for example, one could

	α	$\log(\sigma)$	$\rho^{(MZ)}$	$\rho^{(DZ)}$	β	pre-p
lower	-5	-5	-1	-1	-5	-5
upper	5	5	1	1	5	5

Table 5: Lower and upper boundaries for the parameters. The same boundary was kept for the parameter in each component. The boundaries are defined around the parameters that are used in the function tmbstan.

use the argument MAP to fix only a subset of a parameter vector (in the example of a Gaussian mixture with three components, one could fix only the first two mean values). We performed some tests and the randomness of the initial seed plays a role on the "success" of the sampling process. As mentioned above, other parameters can be chosen as fixed, and the results can strongly vary depending on the initial seed. Overall, fixing all three mean parameters has proven to be the most consistent approach. A summary of these tests can be found in the supplementary material (Table 3).

4.1.3 Bounding the parameter space

The last approach that we present consists in setting boundaries in the space that HMC explores when collecting samples. As seen in Section 3, when the sample of the weight of a component is very close to zero the other parameters associated to that component are very unstable and tend towards extreme values (in Figure 1, it can be seen for σ_2).

Choosing the boundaries for the parameters is not a trivial feat. The only obvious choice relates to the correlation coefficients, which should always take a value between -1 and 1. For all the other parameters, the choice of bounds is not as straightforward: we want to allow HMC to explore the entirety of the relevant parameter space to avoid missing the best solution. In this case, we are advantaged by knowing the parameters of the generating Gaussian mixture. We use them as reference, but still give enough space for HMC to explore the parameter space.

The chosen upper and lower boundaries are listed in Table 5. Notice that the boundaries are set on the parameters that are read in MakeADFun and tmbstan (e.g. α, $\log(\sigma)$).

The results from this approach are shown in Table 4. The averages and standard deviations are very similar to those of the previous two approaches.

Preventing the sampler from fully exploring the parameter space by setting too restrictive boundaries can reduce the efficiency of this experiment. In general, a first analysis of the dataset can help choosing upper and lower boundaries, especially for the means and the standard deviations. The issues detailed in Section 3 are often accompanied

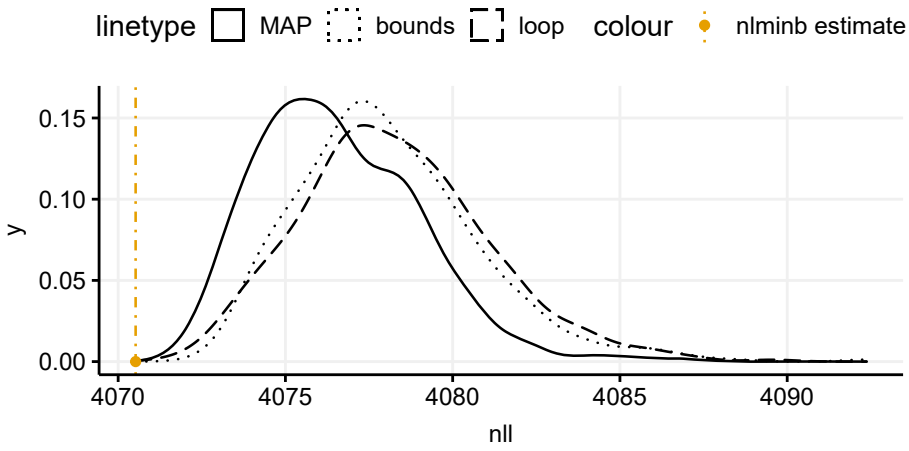


Figure 2: Density curves of the negative log likelihood of the HMC samples compared to the nlminb estimate.

with very large or very small values for μ's and σ's, and preventing these escalations without compromising a thorough exploration of the parameter space can be achieved.

Moreover, not all parameters require a boundary: the problematic ones are usually the mean and the standard deviation, so setting lower and upper bounds on these might be enough to prevent the issues discussed in this paper. The fewer parameters are bounded, however, the larger is the chance to not estimate three distinct components. As with the other approaches, the random seed can play an important role. The results of some experiments in this direction are collected in the Supplementary Material (Table 4).

4.2 Comparison

Looking at Table 4, we notice that in all three samples HMC identifies the three distinct components. The first weight is overall slightly overestimated, to the expense of the second weight. These two components are quite close to each other, so a margin of error is expected.

Overall, these three methods avoid the main issue that we encountered in Section 3 and provide accurate estimates of the parameters. Table 2 displays global values of these three approaches as well. The three models capture the overall shape of the distribution, and they all provide global values which are comparable to the true global values.

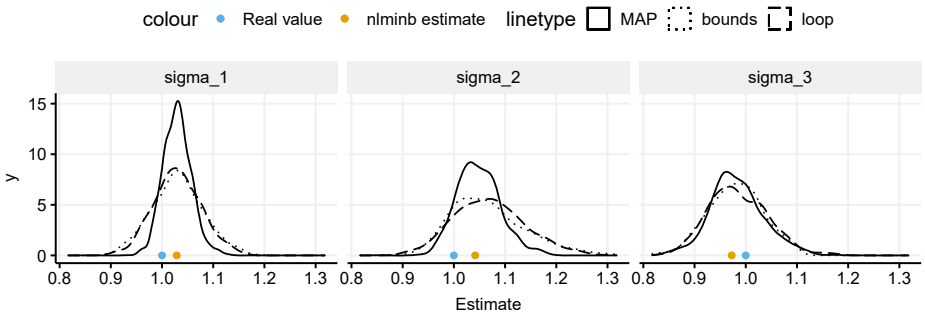


Figure 3: Density curves of the HMC samples compared to real values and nlminb estimates of the standard deviations of the three mixture components.

Figure 2 collects the density curves of the negative log likelihoods obtained from the samples of all three approaches. The density curve for the base HMC is not displayed, since it is on a different scale. While the HMC samples never find a better value than the nlminb result, there is clear improvement compared to the negative log likelihoods that were calculated from the base HMC approach.

The minimum values of the negative log likelihood for all approaches are listed in Table 3, compared to the nlminb result.

Figure 3 shows the densities of the three sigma parameters, for each model, plotted against the real values and the nlminb estimates. The densities are rather comparable, with the exception of the "MAP" approach which, as we already discussed, estimates the parameters with an overall smaller standard deviation.

4.3 Search for a better minimum

The goal of this paper is to find a procedure that can explore the parameter space thoroughly and efficiently to find initial values that lead to the global minimum of the negative log likelihood.

We wish, hence, to test the efficiency of the samples that we gathered by using them as initial values for the TMB algorithm and the optimizer nlminb.

We begin by using the samples gathered by the base HMC algorithm. As it can be expected, the extreme values of the weights and the standard deviations of the second component hinder the efficiency of TMB. Since the optimization begins from initial

components	min AIC	min BIC
1	8934	8959
2	8276	8276
3	8171	8247

Table 6: AIC and BIC values for fitted mixture models with different number of components. The first two rows list the minimum values among 500 fitted models (one for each non-warmup HMC sample). The third row lists AIC and BIC values for the single nlminb run described in Table 1, second column.

values which are orders of magnitude distant from the true values, most of the iterations end without converging and with worse negative log likelihoods. The few that converge do not find a better negative log likelihood.

The base approach substantially samples from a two-component Gaussian mixture, and this suggested mixture might be a better fit for our dataset.

To verify that our three-component mixture is indeed the better choice, we run TMB for a two component Gaussian mixture by using only the parameter samples of the first and third component. We calculate average and standard deviation for each parameter, and collect the results in Table 1. The parameter estimates are comparable to those of the first and second component of the HMC sampling, but with a larger standard deviation.

To compare the three-component and the two-component mixture models, we use the AIC and BIC values. In Table 6, we list the minimum AIC and BIC values of the fitted two-component mixtures and compare them to the values obtained from the first optimization via nlminb. The nlminb result still performs the best with a wide margin, despite the larger number of parameters. In the supplementary material (Table 2), we presented the estimates for one base case where only one component had a non-zero weight. To complete this analysis, we also test whether the estimates of that single non-zero component could find a better fit of the data. Among the three models we studied, the latter performed the worst. According to these criteria, hence, a three component mixture is the best fit for this data.

We now test the samples that we obtained by using the three successful approaches of Subsection 4.1. Following the same procedure, we run TMB using as initial values the non-warm-up samples of each iteration, for the three separate models. The results are consistent between the three different approaches: the code converges for each of these initial values. However, we do not obtain any new solution: the parameter estimates and negative log likelihood resulting from using any of these sets of initial values are identical to those that we found in the first run of TMB and subsequent optimization (that is, the estimates listed in Table 1).

5 Discussion

In all of the experiments carried out, HMC has not been able to find a lower negative log likelihood than the one found by the quasi Newton algorithm built into nlminb. While the base HMC approach evidently fails even at identifying the distribution underlying the dataset, adding some restraints to the parameter space allows the algorithm to return samples which are consistent with the target distribution. As our procedure involves restarting nlminb from every individual sample point we feel that we have provided evidence that the real global maximum likelihood has been found for this dataset. Arguably, we should have chosen a dataset which exhibited a multimodal likelihood to better illustrate the method, but we wanted a dataset with similar properties to that of Azzolini et al. [2022].

This exploration of the HMC algorithm has found that a base approach struggles with sampling from a multimodal distribution, and has a tendency to collapse some components of the mixture distribution. Of the three approaches we propose to fix this issue, the last one is the most promising. The first approach relies on randomness and it is quite time consuming, while the second one relies on trusting the outputs of other optimizers. On the contrary, one can set boundaries large enough to be safe that the main part of the parameter space is explored, but preventing the extreme estimates that we incur in the base case. We propose this approach as a tool for exploring the parameter space in search of the global minimum of the negative log likelihood.

When applying the tmbstan function, there are several options that can be chosen: the number of iterations, the maximal tree depth, the length of the leapfrog "jump". These choices can produce slower or faster processes, more or less efficient. This code, though, seems to have run into several issues with tmbstan. Indeed, our analysis was hindered by the simulation getting stuck into areas of the parameter space which would greatly slow down, or downright interrupt, the algorithm. This happened especially when the number of iteration was too large.

At a late stage in this work we became aware of the R package pdmphmc [Kleppe, 2023] which is designed to be a computationally fast and stable implementation of HMC. By imposing some extra priors, pdmphmc seems to generate samples with good mixing properties, and should be investigated in further detail.

6 Acknowledgements

Parts of this work have been done in the context of CEDAS (Center for Data Science, University of Bergen, Norway).

References

F. Azzolini, G. Berentsen, H. Skaug, J. Hjelmborg, and J. Kaprio. The heritability of BMI varies across the range of BMI: a heritability curve analysis in a twin cohort. *bioRxiv*, 2022.

J.-P. Baudry and G. Celeux. EM for mixtures: initialization requires special care. *Statist Comput. July*, 25(4):713–726, 2015.

G. D. Berentsen, F. Azzolini, H. J. Skaug, R. T. Lie, and H. K. Gjessing. Heritability curves: A local measure of heritability in family models. *Statistics in Medicine*, 40 (6):1357–1382, 2021.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

D. Donnelly and E. Rogers. Symplectic integrators: An introduction. *American Journal of Physics*, 73(10):938–945, 2005.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

J. F. Geweke et al. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis, 1991.

M. D. Hoffman, A. Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

ISO/IEC. Programming languages — C++. Draft International Standard N4660, March 2017. URL `https://web.archive.org/web/20170325025026/http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2017/n4660.pdf`.

T. S. Kleppe. *pdmphmc - numerical generalized randomized HMC processes for R*, 2023. URL `https://github.com/torekleppe/pdmphmc`.

K. Kristensen, A. Nielsen, C. W. Berg, H. J. Skaug, and B. M. Bell. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21, 2016. doi: 10.18637/jss.v070.i05.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

C. Monnahan and K. Kristensen. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PloS one*, 13(5), 2018. doi: 10.1371/journal.pone.0197954.

R. M. Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

E. Shireman, D. Steinley, and M. J. Brusco. Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior research methods*, 49(1):282–293, 2017.

Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 2.29*. 2019. URL `http://mc-stan.org/`.

C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

Supporting material for "Exploring the likelihood surface in multivariate Gaussian mixtures using Hamiltonian Monte Carlo"

Parameter	Z-score
α	(1.41, -0.69, 1.92)
$\log(\sigma)$	(-1.10, -2.18, 0.96)
$\rho^{(MZ)}$	(-1.04, 2.44, -0.35)
$\rho^{(DZ)}$	(-0.52, -0.03, -1.30)
$\mathrm{t}\delta$	(-0.67, 2.72)
β	0.48

Table 1: Z-scores of the Geweke diagnostic for all parameter samples for the HMC iteration described in Section 3 of the main article. Parameters α, $\log(\sigma)$, $\rho^{(MZ)}$ and $\rho^{(DZ)}$ are vectors of length three, one element associated to each component of the Gaussian Mixture in study. The vector $\mathrm{t}\delta$ has two elements (then converted into a vector of length three by imposing conditions on the sum of the weights). β is a scalar value. The convergence criterion is $|Z| \leq 1.28$. In red, the parameters that do not converge.

Parameter	True values	nlminb	HMC samples			
			Example 1		Example 2	
			average	std. dev.	average	std. dev.
μ_1	21	21.07	0.05	0.00	21.70	0.05
μ_2	23	23.09	0.06	0.10	21.70	0.05
μ_3	28	27.93	22.60	0.07	6.57e+95	1.30e+97
σ_1	1	1.03	40.05	762.43	4.42	0.25
σ_2	1	1.04	1.87e+21	3.81e+22	1.30	0.03
σ_3	1	0.97	2.26	0.04	5.54e+57	5.51e+58
$\rho_1^{(MZ)}$	0.7	0.69	0.07	0.60	0.96	0.01
$\rho_2^{(MZ)}$	0.5	0.46	-0.27	0.52	0.77	0.02
$\rho_3^{(MZ)}$	0.3	0.34	0.91	0.01	0.05	0.55
$\rho_1^{(DZ)}$	0.4	0.46	-0.01	0.66	0.94	0.01
$\rho_2^{(DZ)}$	0.3	0.14	0.04	0.59	0.64	0.03
$\rho_3^{(DZ)}$	-0.2	-0.30	0.86	0.01	0.00	0.58
β	2	1.97	2.16	0.15	2.12	0.09
p_1	0.60	0.63	0.00	0.00	0.21	0.09
p_2	0.30	0.27	0.00	0.00	0.79	0.02
p_3	0.10	0.09	1.00	0.00	0.00	0.00

Table 2: Real parameter values of the Gaussian mixture generating the simulated dataset (first column) and nlminb estimates (second column) The last four columns contain average and standard deviation of two samples generated via No U-Turns HMC. The warm-up samples are not included. Both examples do not identify three distinct components.

	number of fixed parameters							
	alpha			log sigma			pre-p	
Seed	3	2	1	3	2	1	2	1
950222	4072.0	4462.2	4103.3	6602.3	4338.2	6601.8	4072.1	NA
335738	4071.6	4071.8	4462.2	4273.5	6618.9	4103.2	7296.2	4196.8
133073	NA	NA	4462.2	NA	NA	4336.8	7296.2	4462.1
490112	4072.3	NA	NA	4273.6	NA	NA	4073.6	4072.4
60746	4072.2	4462.1	4463.0	6601.7	7480.8	4073.6	4367.3	4110.0
357948	4071.7	NA	4072.4	NA	NA	4103.4	NA	4283.0
227117	4071.5	4135.4	NA	NA	NA	NA	4365.8	NA
400075	4072.7	4462.4	NA	4072.4	4103.5	NA	4072.0	NA
936546	NA	4071.8	4071.9	NA	4103.4	4283.4	NA	NA
837627	4072.3	4462.5	4462.3	4273.4	4318.1	4461.1	4367.0	4277.3

Table 3: List of negative log likelihoods obtained running an HMC algorithm while keeping some parameters fixed to their nlminb estimate. We repeated the example with alpha, log(sigma) and pre-p. For each parameter vector, we attempted fixing all or some elements. The numbers 1, 2, and 3 in the third row indicate how many elements in the corresponding parameter vector were kept fixed during the HMC iteration.

Seed	Boundaries 1	Boundaries 2	Boundaries 3
950222	4102.6	4462.2	NA
335738	4073.0	4283.4	4271.8
133073	4101.0	4103.2	4103.8
490112	4073.0	4074.2	4276.6
60746	4073.3	4104.0	4073.5
357948	4073.2	4284.0	4284.6
227117	4072.8	4103.7	4103.5
400075	-Inf	4073.7	4406.8
936546	-Inf	4104.1	NA
837627	4271.9	4103.5	4072.0

Table 4: List of negative log likelihoods obtained by setting boundaries on all or some parameters during the HMC algorithm. We constructed three separate boundaries, called in the table "Boundaries 1", "Boundaries 2" and "Boundaries 3". The values chosen as boundaries are identical to the ones shown in the manuscript, Table 4. "Boundaries 1" imposes boundaries on all the parameters; "Boundaries 2" imposes boundaries on all parameters except for pre-p and beta "Boundaries 3" imposes boundaries only on mean and log(sigma).

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230844120 (print)
9788230859582 (PDF)