A Finite Mixture Distribution Model for Data Collected from Twins

Michael C. Neale

Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, USA

Most analyses of data collected from a classical twin study of monozygotic (MZ) and dizygotic (DZ) twins assume that zygosity has been diagnosed without error. However, large scale surveys frequently resort to questionnaire-based methods of diagnosis which classify twins as MZ or DZ with less than perfect accuracy. This article describes a mixture distribution approach to the analysis of twin data when zygosity is not perfectly diagnosed. Estimates of diagnostic accuracy are used to weight the likelihood of the data according to the probability that any given pair is either MZ or DZ. The performance of this method is compared to fully accurate diagnosis, and to the analysis of samples that include some misclassified pairs. Conventional analysis of samples containing misclassified pairs yields biased estimates of variance components, such that additive genetic variance (A) is underestimated while common environment (C) and specific environment (E) components are overestimated. The bias is non-trivial; for 10% misclassification, true values of Additive genetic: Common environment: Specific Environment variance components of .6: .2: .2 are estimated as .48: .29: .23, respectively. The mixture distribution yields unbiased estimates, while showing relatively little loss of statistical precision for misclassification rates of 15% or less. The method is shown to perform quite well even when no information on zygosity is available, and may be applied when pair-specific estimates of zygosity probabilities are available.

The majority of investigations into the causes of variation in human populations begin with a classical twin study. Twin pairs who have been raised in the same home are diagnosed as MZ or DZ, and the correlations of the two types of pair are compared. As intimated by Galton (1865) and later formalized (Merriman, 1924; Siemens, 1924), greater average similarity of MZ than DZ twins is taken as support for the hypothesis that genetic factors cause individual differences in the population. Several formulae exist to compute an estimate of the relative impact of genetic and environmental factors (Holzinger, 1929; Vandenberg, 1966) but these have important limitations (Jinks & Fulker, 1970; Neale, in press). Consequently, most modern analyses use a model-fitting approach (Eaves et al., 1978; Neale & Cardon, 1992) to obtain maximum likelihood estimates of variance components.

Most methods for the genetic analysis of twin data rely on the accurate diagnosis of zygosity, but a few analyses of twin data have been conducted without zygosity information (Scarr-Salapatek, 1971). An estimate of the MZ correlation was obtained by assuming that the observed correlation for same-sex pairs was generated by DZ pairs with equal sample size and equal correlation to those of the opposite sex pairs, along with MZ pairs whose sample size and correlation can be obtained by subtraction, via z-transformed correlations. Assumptions of this earlier method required that the number of same-sex DZ pairs was equal to the number of opposite-sex pairs (i.e., no gender effects on study participation) and that sex-limitation was absent. In practice, these assumptions are unlikely to be met because gender bias in participation is frequently observed (Lykken, 1978). Furthermore, this approach has been shown to be of limited use beyond simple univariate heritability estimation, and to be dramatically short of statistical power when comparing groups (Eaves & Jinks, 1972). A more general method that can be used for multivariate, longitudinal, and other more complex genetically informative study designs would be useful.

Apart from rare cases where there is extensive genetic marker data collected from every pair in the sample, perfectly accurate diagnosis of zygosity is almost never achieved in practice. In large studies it is prohibitively expensive to genotype every twin pair, so questionnaire-based methods are frequently used. While some questionnaires have been demonstrated to be accurate in up to 95% of cases (Jackson et al., 2001; Kasriel & Eaves, 1976; Lykken, 1978; Nichols & Bilbro, 1966), the effects of this level of misclassification have not been explored systematically. It is possible that misclassification has biased the results of almost all studies of twins. Meta-analyses may be particularly susceptible if the degree of misclassification varies between the studies being examined.

The goal of this article is to present a new method to analyze twin data in the presence of known rates of misclassification. The method is appropriate when average misclassification rates are known for MZ and DZ pairs. The two rates do not have to be equal, as it is usually found that misclassification of MZ pairs as DZ is more common than the reverse. Furthermore, the method may be applied on a case-by-case basis, so that depending on the response pattern of the twin pair, individual zygosity probabilities may be calculated and used. Heath et al. (2003) recently described a latent-class approach to the diagnosis of zygosity which

Address for correspondence: Michael C. Neale, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Box 980710, Richmond, VA 23298, USA. Email: neale@hsc.vcu.edu

yields estimates suitable for this purpose. Finally, the method may be applied in meta-analyses, multivariate or longitudinal analyses, and studies of extended pedigrees.

Two key features of the proposed method are of interest. First, the bias of the estimates of the A, C and E components that occurs with different rates of misclassification is compared between the conventional and mixture distribution methods. Second, the 95% confidence intervals of these parameter estimates are compared. Some broadening of the confidence intervals is expected to occur when the mixture distribution method is used.

Method

Finite mixture distributions (Everitt & Hand, 1981) provide a suitable mathematical model for the analysis of imperfectly classified data. This approach has been used in linkage analyses of quantitative traits, where the classification of sibling pairs into those sharing zero, one or two alleles identical-by-descent at a given genetic locus is subject to error (Eaves et al., 1996; Fulker & Cherny, 1996; Neale, 2003). In the present case, the population of twin pairs consists of two classes, MZ and DZ, whose zygosity may not be known.

Suppose that twin pairs are assessed on a trait which is assumed to follow a normal distribution in the population, and a bivariate normal in twin pairs. Under a simple model of twin resemblance (Neale & Cardon, 1992), the predicted covariance of MZ twin pairs is given by:

$$\Sigma_{MZ} = \begin{bmatrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix} \quad (1)$$

where a^2 c^2 and e^2 are the additive genetic, common environment and random environment variance components, respectively. The corresponding matrix for DZ twin pairs is:

$$\Sigma_{DZ} = \begin{bmatrix} a^2 + c^2 + e^2 & .5a^2 + c \\ .5a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix} \quad (2)$$

The likelihood of a vector of scores from MZ pairs is given by the bivariate normal probability density function:

$$L(MZ) = \left| 2 \pi \Sigma_{MZ} \right|^{-k/2}$$
$$\exp \left\{ -\frac{1}{2} (x_i - \mu)' \Sigma_{MZ}^{-1} (x_i - \mu) \right\}$$

in which μ is the (2×1) column vector of estimated population means of the $k = 2$ observed scores of the twins, x_1 and x_2, and $|\Sigma_{MZ}|$ and Σ_{MZ}^{-1} denote the determinant and inverse of the matrix Σ_{MZ}, respectively. The likelihood of observed scores of DZ pairs, L(DZ) is computed in an analogous fashion by substituting Σ_{DZ} for Σ_{MZ}.

When zygosity is not known exactly, the likelihood may be computed as a weighted sum of L(MZ) and L(DZ). The weights used are the probabilities $p(MZ)$ and $p(DZ) = 1-p(MZ)$ which are derived from available information to diagnose zygosity. These weights would typically be different for pairs initially identified (with error) as MZ

from those identified as DZ. More generally, the weights may be different for all pairs $i = 1 \dots N$ in the sample:

$$L_i = p(MZ_i) L_i(MZ) + p(DZ_i) L_i(DZ).$$

Simulation

To explore the effects of misclassification on the parameter estimates and confidence intervals on parameters, data were simulated under a standard genetic model. The trait was generated by additive genetic (a^2), common environment (c^2) and specific environment (e^2) variance components that constituted 60, 20 and 20% of the variance, respectively. Data on 1000 pairs of MZ and 1000 pairs of DZ twins were generated using SAS (SAS, 1999), which was also used to implement the misclassification process. The percentage of pairs misclassified was set at 1, 5, 10, 15 or 50%. In many samples MZ pairs are more frequently misclassified than DZ, so some tests were run with higher MZ error rates. Each simulation was repeated 500 times.

For each simulated dataset, the data were analyzed using three methods. Under the "True" method, the assigned zygosity was set to the true zygosity (no misclassification) and the weights were assigned as $p(MZ) = 1$ for MZ pairs and $p(DZ) = 1$ for DZ pairs. Under the "Conventional" method, the pairs were analyzed with the same weights, but the assigned zygosity was incorrect in a proportion of cases. Under the "Mixture" method, the data were analyzed using the known population misclassification rates for MZ and DZ twins as weights for the preliminarily diagnosed pairs.

Parameter estimates and likelihood-based confidence intervals (Neale & Miller, 1997; Venzon & Moolgavkar, 1988) were obtained with the freely available Mx package (Neale et al., 1999). Mx scripts for the analyses are available on the Mx website[1] in the examples section.

Results

Table 1 shows the average estimates of a^2, c^2 and e^2 and their average confidence intervals under the three misclassification conditions. Six features of this table are especially noteworthy. First, parameter estimates are close to the population values for the True analyses, as is expected when there is no zygosity classification error. Second, the mixture distribution analyses yield parameter estimates very close to the population values. Third, and especially interesting, the confidence intervals from the mixture distribution approach are very close to those from the True analyses, indicating that the loss of information incurred by specifying a mixture distribution is relatively slight, even when the misclassification rate is as high as 15%. Fourth, the Conventional analyses that ignore misclassification show substantial bias from the population estimates, with additive genetic variance underestimated and environmental sources, particularly common environment, over-estimated, except when the misclassification rate is low (1% for MZ twins). Fifth, the confidence intervals of the Conventional analyses span approximately the same range as those of the True analyses, and are just as biased as the Conventional

Table 1

Averaged Parameter Estimates (Est) and 95% Confidence Intervals (Low, High) from Three Analyses of Simulated Twin Data Sets. True is Without Misclassification, Conventional is Standard Analysis of Partly Misclassified Samples, and Mixture Uses the Mixture Distribution Method. Each Data Set Consists of 500 Replicates of Samples of 1000 MZ and 1000 DZ Pairs of Twins with Misclassification Rates of 15, 10 or 5%. Variance Components (VC) are a^2: Additive Genetic; c^2: Common Environment; and e^2: Specific Environment

Misclassification (%)			True			Conventional			Mixture		
MZ	DZ	VC	Est	Low	High	Est	Low	High	Est	Low	High
1	1	a^2	0.601	0.512	0.697	0.589	0.500	0.685	0.601	0.511	0.699
1	1	c^2	0.198	0.105	0.287	0.207	0.114	0.296	0.197	0.103	0.288
1	1	e^2	0.200	0.184	0.219	0.203	0.186	0.222	0.200	0.183	0.219
1	5	a^2	0.601	0.513	0.696	0.566	0.476	0.662	0.603	0.510	0.702
1	5	c^2	0.197	0.106	0.286	0.218	0.125	0.308	0.196	0.102	0.287
1	5	e^2	0.200	0.183	0.219	0.215	0.197	0.235	0.200	0.181	0.221
5	5	a^2	0.600	0.512	0.697	0.540	0.452	0.634	0.603	0.507	0.706
5	5	c^2	0.200	0.108	0.290	0.246	0.154	0.335	0.197	0.100	0.292
5	5	e^2	0.200	0.183	0.219	0.215	0.197	0.235	0.200	0.181	0.221
10	10	a^2	0.601	0.513	0.698	0.481	0.394	0.574	0.608	0.504	0.718
10	10	c^2	0.199	0.107	0.289	0.290	0.199	0.378	0.192	0.090	0.291
10	10	e^2	0.199	0.183	0.218	0.230	0.211	0.251	0.199	0.178	0.221
15	5	a^2	0.601	0.515	0.694	0.481	0.391	0.577	0.606	0.504	0.714
15	5	c^2	0.199	0.110	0.285	0.274	0.182	0.364	0.193	0.095	0.289
15	5	e^2	0.199	0.182	0.219	0.244	0.224	0.267	0.199	0.176	0.224
15	15	a^2	0.601	0.512	0.697	0.420	0.334	0.512	0.612	0.501	0.728
15	15	c^2	0.199	0.107	0.289	0.335	0.245	0.422	0.188	0.082	0.292
15	15	e^2	0.199	0.183	0.218	0.245	0.224	0.267	0.198	0.176	0.223
50	50	a^2	0.600	0.512	0.696	0.020	0.001	0.091	0.643	0.480	0.786
50	50	c^2	0.199	0.107	0.289	0.634	0.561	0.698	0.161	0.042	0.299
50	50	e^2	0.199	0.183	0.218	0.344	0.319	0.370	0.191	0.161	0.230

parameter estimates. Sixth, even when there is essentially no information about zygosity (misclassification is 50% for both MZ and DZ) the mixture approach is still able to recover parameter estimates, albeit with less precision than when zygosity is diagnosed perfectly (e.g., .15 vs. .09 each side for a^2).

Discussion

This article presents a method of analyzing data collected from twins when zygosity is known with less than full precision for some or all of the sample. Results of simulations indicate that the mixture distribution approach recovers the population values used for simulation, and does so even when the misclassification rate is as high as 50%, which is probably more than is common in twin studies. Better still, with misclassification rates as high as 15%, there is little loss of information; confidence intervals are only slightly broader than when no misclassification is present. This result indicates that statistical power will not be adversely affected by the specification of a mixture distribution. Failure to specify a mixture distribution when one exists can result in substantial bias of parameter estimates. With 10% misclassification of MZ and DZ pairs (N = 1000 pairs each), the average heritability estimate of .481 is outside the average 95% confidence intervals of the estimate of .608 provided by the mixture distribution analyses.

In the present era of relatively inexpensive genotyping, it is likely that misclassification of zygosity will be less common, especially in small to moderate sized samples of twins. The method presented here should still prove useful in larger studies where genotyping is impractical or too costly. Indeed, except when the cost of phenotyping is very high relative to genotyping, questionnaire-based zygosity assessment and analysis with the mixture distribution method may prove to be the more cost-effective strategy.

The observed bias in parameter estimates due to misclassification is non-trivial once misclassification rates reach 5%. This has implications for meta-analyses of twin studies (Hettema et al., 2001; Sullivan et al., 2000) as varying rates of misclassification are a likely source of discrepancy between studies. If zygosity diagnosis has become more precise in more recent studies, greater heritability and smaller effects of the shared environment may be observed in later versus earlier studies.

An interesting possibility for practical research is that it would be possible to use data collected from twins in which zygosity is completely unknown. Prior information about the ratio of MZ to same-sex DZ twin pairs could be used to weight those cases that have no direct data on zygosity. This approach may be most effective when the sample also includes some pairs with reasonably accurate zygosity diagnosis.

The analyses used here were conducted using simulated bivariate normal distributions for the twins' trait scores. The analysis of binary or ordinal data with this method can be tackled with a threshold model (Falconer, 1965). In principle, any analytic approach for which the likelihood can be written or approximated (via e.g., Monte Carlo Markov Chain methods) is amenable to the finite mixture distribution approach used here.

While the mixture distribution method is particularly suitable for studies of twins, it is not limited to this genetically informative design. Non-paternity is a documented problem for genetic research where the ostensible father is not the biological father (Neale et al., 2000). Rates vary, but 5% may be a reasonable estimate. In the absence of data such as genetic markers that unambiguously diagnose paternity, it would seem prudent to fit mixture distribution models to data collected from families, to avoid bias in parameter estimates derived from father–child resemblance.

Finally, although this article is limited to univariate analysis, it is equally appropriate for multivariate or longitudinal analyses. Furthermore, it seems likely that the loss of information due to misclassification would be even less when there are multivariate data. The likelihood of the observed phenotypic scores for the incorrect zygosity would be substantially less than that of the correct zygosity, and to a greater degree than would be expected in a univariate analysis. That is, the phenotypes would themselves be providing an indirect and partial zygosity diagnosis.

These results should be considered in the light of three potentially important limitations. First, the proposed method relies upon accurate estimates of the probabilities that pairs are MZ or DZ. Biased parameter estimates may be expected when these estimates are incorrect, as can be seen in the results of the Conventional analyses in Table 1. Incorrect application of $p(MZ) = 1$ or $p(MZ) = 0$ generates bias in the parameter estimates. Second, a single set of variance component proportions (A: C: E ratio of .6: .2: .2) was used for exploring the effects of misclassification. When the ratio was set to .2: .1: .7 relatively little bias was observed: for 5% misclassification, the average estimates from the conventional approach were .179: .115: 0.706 versus .202: .098: .700 for the mixture distribution analysis. Similar effects are to be expected across the range of heritability, with more bias in the conventional approach when the difference between the MZ and DZ correlations is greater. An architecture of only common and specific environment components would not incur bias from misclassification, because the predicted correlations of MZ and DZ twins are equal so misclassification would have no effect. Conversely, substantial genetic non-additivity, with MZ correlation much greater than the DZ would accrue considerable bias from misclassification.

Third, the method is based on the assumption that zygosity diagnosis is not derived from the phenotypes under study. While this is likely to be the case for most traits in most studies, for studies of physical characteristics such as height or facial morphology, misclassification may be partially systematic. In this case, the more similar DZ twins would be more likely to be misclassified as MZ, which would reduce the DZ correlation. The effect of the contamination of MZ pairs with more similar DZ pairs is less clear, as it depends on how similar the misclassified DZ pairs are relative to the similarity of MZ pairs. Similarly, misclassification of the more dissimilar MZ pairs would reduce the apparent MZ correlation, and could have a variety of effects on the DZ correlation. In many cases, the impact of this sort of misclassification would be to bias estimates of genetic variance components upwards, and common environment downwards — the opposite of the random misclassification mechanism considered in this article.

Acknowledgments

Endnote

1   http://www.vcu.edu/mx

References

Eaves, L. J., & Jinks, J. L. (1972). Insignificance of evidence for differences in heritability of IQ between races and social classes. *Nature, 240,* 84–88.

Eaves, L. J., Last, K., Young, P. A., & Martin, N. G. (1978). Model fitting approaches to the analysis of human behavior. *Heredity, 41,* 249–320.

Eaves, L. J., Neale, M. C., & Maes, H. H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior Genetics, 26,* 519–526.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

Falconer, D. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annuals of Human Genetics, 29,* 51–76.

Fulker, D. W., & Cherny, S. S. (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics, 26,* 527–532.

Galton, F. (1865). Hereditary talent and character. *MacMillan's Magazine, 12,* 157–166, 318–327.

Heath, A. C., Nyholt, D. R., Neuman, R., Madden, P. A. F., Bucholz, K. K., Todd, R. D., et al. (2003). Zygosity diagnosis in the absence of genotypic data: An approach using latent class analysis. *Twin Research, 6,* 22–26.

Hettema, J., Neale, M., & Kendler, K. (2001). A review and meta-analysis of the genetic epidemiology of anxiety disorders. *American Journal of Psychiatry, 158,* 1568–1578.

Holzinger, K. J. (1929). The relative effect of nature and nurture influences on twin differences. *Journal of Educational Psychology, 20,* 245–248.

Jackson, R., Snieder, H., Davis, H., & Treiber, F. (2001). *Determination of twin zygosity: A comparison of DNA with various questionnaire indices, 4,* 12–18.

Jinks, J. L., & Fulker, D. W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin, 73,* 311–349.

Kasriel, J., & Eaves, L. (1976). The zygosity of twins: Further evidence on the agreement between diagnosis by blood groups

and written questionnaires. *Journal of Biosocial Sciences, 8,* 263–266.

Lykken, D. (1978). The diagnosis of zygosity in twins. *Behavior Genetics, 8,* 437–473.

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs, 33,* 1–58.

Neale, M. C. Individual fit, heterogeneity, and missing data in multigroup sem. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families.* Kluwer: Academic Press.

Neale, M. C., & Miller, M. M. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27,* 113–120.

Neale, M. (in press). Twin studies: Software and algorithms. In D. Cooper (Ed.), *Encyclopedia of the human genome.* London: Macmillan Publishers Ltd, Nature Publishing Group.

Neale, M., Boker, S., Xie, G., & Maes, H. (1999). *Mx: Statistical modeling* (5th ed.). Department of Psychiatry, Virginia Commonwealth University, Box 980126 Richmond VA.

Neale, M., Neale, B., & Sullivan, P. (2002). Non-paternity in linkage studies of extremely discordant sib-pairs. *American Journal of Human Genetics, 70,* 526–529.

Nichols, R., & Bilbro, W. (1966). The diagnosis of twin zygosity. *Acta Geneticae Medicae et Gemollologiae, 16,* 265–275.

SAS (1999). *SAS OnlineDoc, Version 8.* Cary, NC: SAS Institute Inc.

Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science, 174,* 1285–1295.

Siemens, H. (1924). *Die Zwillingspathologie.* Berlin: Springer-Verlag.

Sullivan, P., Neale, M., & Kendler, K. (2000). The genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry, 157,* 1552–1562.

Vandenberg, S. G. (1966). Contributions of twin research to psychology. *Psychological Bulletin, 66,* 327–352.

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics, 3,* 87–94.