A major focus in studies of multivariate evolution is on trying to understand the biological and evolutionary forces that shape patterns of covariance between organismal traits. This is particularly true with respect to studies concerned with the topics of phenotypic integration and modularity (see Chernoff and Magwene 1999, for a review).

In a recent paper published in Systematic Biology on the topic of modularity and integration, Mitteroecker and Bookstein (2007) reviewed the use of “Wright-style” factor analysis to characterize and study patterns of morphological integration. These authors compared their preferred approach with a number of frameworks that previous investigators have employed to address similar questions. Among the methods that Mitteroecker and Bookstein comment on is an approach, described in Magwene (2001), for studying modularity and integration using Gaussian graphical models (GGMs). Mitteroecker and Bookstein (hereafter M&B) assert that the methods described in Magwene (2001) are unsuitable for the study of morphological integration because they are only applicable to modules with 2 variables. In their words “Magwene’s example...accidentally worked because all modules consisted of two variables only.”

Below I demonstrate that M&B’s assertion is incorrect and there is no inherent limitation in terms of the size of the modules that can be distinguished using the GGM approach. I discuss conditional independence relationships and briefly describe what might be termed a “Wright-style” approach based on graphical modeling. I also highlight a number of recent biological applications of graphical modeling.

**Graphical Models for Studying Morphological Integration**

Magwene (2001) proposed the use of GGMs as a statistical framework for characterizing the multivariable structure of phenotypic traits. GGMs (also known as covariance selection models or Markov Random Fields) are one of a family of related statistical methods, referred to generally as probabilistic graphical models. A probabilistic graphical model defines a family of probability distributions in terms of a graph. The nodes in the graph represent random variables, and the structure of the graph represents patterns of dependence and independence among the variables (Whittaker 1990; Edwards 1995; Lauritzen 1996; Jordan 2004). There are a variety of types of graphical models, including ones that employ directed edges (usually called Bayesian networks) as well as models that combine directed and undirected edges (e.g., Cox and Wermuth 1996). These varieties of graphical models have different properties, incorporate different assumptions about symmetries of interaction, and have different factorization properties (see Bishop 2007, for an overview). Shipley (2000) gives a good introduction to graphical modeling, with an emphasis on directed graphical models, aimed primarily at biologists.

A GGM is determined by identifying the pairs of variables that are independent, conditional on all other variables in the analysis. For multivariate normal data, this is equivalent to identifying pairs of variables, $x_i$ and $x_j$, which exhibit vanishing partial correlations, $\pi_{ij} = 0$, where $X \setminus \{i, j\}$ indicates the set of variables in $X$ other than $x_i$ and $x_j$ (Whittaker 1990). A variety of model-fitting approaches are available to identify the most probable model(s) or to test the fit of a graphical model to a data set (e.g., Whittaker 1990; Giudici and Green 1999; Drton and Perlman 2004).

Magwene (2001) proposed 2 operational definitions of phenotypic modularity based on the structure of graphical models. The strong definition proposed is to define a module as a clique in the conditional independence graph. The weak definition is simply a relaxation to include pseudocliques (most but not necessarily all variables are conditionally dependent). A simple verbal interpretation of the strong definition is that a module is a set of traits, all of which are mutually informative, conditional on all other traits under consideration. Cliques in an undirected conditional independence graph also form the basis of a factorization of a probability distribution (Lauritzen 1996; Bishop 2007).

M&B’s statement that “Magwene’s example...accidentally worked because all modules consisted of two variables only” implies that the structure of the estimated conditional independence graph will bear no relation to the generating structure of the data if there are “local factors” (see below) that affect more than pairs of variables. Figure 1 shows this to be false. I generated a simulated data set with 6 variables (traits) and 150
observations, with a covariance structure specified by the path model in Figure 1a (compare with Fig. 2 in M&B). The corresponding correlation and conditional (partial) correlation matrices, edge exclusion deviance and conditional independence matrices, and estimated graphical model are given in Figure 1b–d. As can be seen from this example, the GGM-based approach clearly finds the 2 modules (each consisting of 3 traits).

To further illustrate that the graphical modeling approach is applicable to both small and large examples, I generated a simulated data set, consisting of 30 variables with an underlying structure consisting of a size variable and 3 local factors (modules), using the model shown in Figure 4 of M&B. For this example, all path coefficients were set to unity, the variance of the latent size factor was given a value of 3, the variance of each of the latent local factors was 1, and each observed variables had an error variance of 1. I used this structure to generate a simulated data set consisting of 900 observations (30 times the number of variables). I estimated the partial correlation matrix and corresponding edge exclusion deviance values from the simulated data. Figure 2 shows a graphical representation of an adjacency matrix (i.e., edges in the graphical model) for all pairwise interactions that have an edge exclusion value greater than 3.84 (corresponding to a P value of 0.05). There are 25 misclassified edges (both false positives and false negatives), corresponding to an error rate of approximately 5.7%, but the graphical model clearly recovers the underlying modular structure of the generating functions.

As the examples given above demonstrate, there is no inherent limitation of the graphical modeling approach to finding modules with only 2 variables. The extent to which this holds for the general, case depends on the underlying generating function, and such factors as sample size, but in general, undirected graphical models are considered an excellent statistical framework for exploring covariance structure (e.g., Knuiman 1978; Krzanowski and Marriott 1995). The assertion that the GGM approach only works when the underlying structure involves just pairs of variables has no theoretical or mathematical basis. For further comparison of GGMs and factor analysis, see Knuiman (1978).

**Conditional Independence Relationships**

Though a fundamental notion in statistics, conditional relationships are not necessarily intuitively easy to grasp. Thomas and O’Quigley (1993) provided a geometric interpretation of partial correlation in terms of spherical geometry, and Li et al. (1975) provided a complementary interpretation of partial correlations in terms of path analysis. Here, I briefly outline the relevant geometry in terms of vector spaces, drawing on the description of Wickens (1995). If we adopt an object space representation of a multivariate data set, then \( p + 2 \) variables of interest can be represented as vectors in a space of dimension \( n \) (where \( n \) is the number of objects). We wish to consider the partial correlation of variables \( x \) and \( y \) conditioned on \( p \) vectors \( z_1, z_2, \ldots, z_p \). The vectors, \( z_i \), define a \( p \) dimensional subspace \( V_z \). The \( z \) vectors together with \( x \) and \( y \) span a space of dimension \( p + 2 \), which is composed of the subspace \( V_z \) and its orthogonal complement \( V_{z^*} \). The partial correlation of \( x \) and \( y \) conditioned on \( z \) is equivalent to the cosine of the angle between the projections of \( x \) and \( y \) onto \( V_{z^*} \) (Wickens 1995). Conditional independence obtains when the projected vectors are at right angles to each other.

The notion of conditional independence is a key feature not only of graphical models, but of factor analysis as well. In the classical single-factor model, all variables are independent conditional on an unobserved (latent) general factor (Raveh 1985). In the “Wright-style” factor analysis approach (Bookstein et al. 1985; Mitteroecker and Bookstein 2007), some variables are independent conditional on a general common factor (i.e., variables in different modules), whereas some variables are independent conditional on both a common factor and a local factor (variables within the same module). A key assumption of this factor analytic approach is that one knows the factor structure (and hence the conditional independence relationships) a priori.

**A “Wright-Style” Graphical Modeling Approach**

Mitteroecker and Bookstein stated that “…for any data with more than two variables per module this approach [Magwene 2001] must yield uninterpretable results.” The basis of this statement would seem to be a misunderstanding of the nature of conditioning implicit in the GGMs. M&B write:

As indicated by Equation (3), within module covariances conditional on those variables outside the module relate to the residual covariances in (6) after the common factors have been partialed out. But when correlations are conditional on variables from the same module, that partial correlation will be near zero because, per definition, the variables within one module are highly correlated and even the local factors are removed.

The statement given above implies that the approach described by Magwene (2001) is equivalent to conditioning on a common factor and then subsequently conditioning on the residual covariances within modules. This implication is incorrect. Unlike the algorithm for Wright-style factor analysis, in the GGM approach, there is no iterative estimation of the conditional covariances/correlations for common factors, followed by estimation of local factors. One can calculate partial correlations via an interactive formula, but that is not the same as partialing out a common factor and then subsequently partialing out local factors.
FIGURE 1. Estimation of a GGM for a simulated data set with a modular covariance structure. (a) The simulated data generated the covariance structure implied by the given path model, with a common factor, \( A \), and 2 more local factors, \( M \) and \( N \) (modules). (b) Correlation matrix for the simulated data (upper triangle) and partial correlation matrix (lower triangle). (c) Edge exclusion deviance matrix (upper triangle) and inferred conditional independence matrix (lower triangle). (d) Inferred graphical model for the simulated data set. See the text for simulation details and Magwene (2001) for details on calculations.

FIGURE 2. Adjacency matrix corresponding to a graphical model estimated for a simulated data set consisting of 30 variables. The “true” underlying model consists of a global size variable and 3 local variables, after Figure 4 of Mitteroecker and Bookstein (2007). The graphical model recovers the modular structure of the generating function without any a priori specification of the model structure. See text for further details.

What if we did as M&B’s statement suggests, and fit a single general factor and then fit a GGM model to the residual covariance matrix? Again we assume the presence of a common factor \( (A) \) and 2 local factors \( (M \) and \( N) \) as shown in the path diagram in Figure 1. Following the removal of the common factor \( (A) \), the residual covariance matrix has a block structure as indicated in M&B’s Equation 3. Because the covariance between blocks is 0, we can treat each of the blocks as separate single-factor models. If the single-factor model holds for each of the blocks (as it must in the models that M&B consider), one can show that none of the variables within each block will be conditionally independent (Salgueiro et al. 2007). Thus, the corresponding conditional independence graph based on the residual covariance structure described in M&B’s Equation 3 would be that shown in Figure 1d.

This suggests that a “Wright-style” graphical modeling approach, in which one estimates a common factor and then estimates the conditional independence graph from the residual covariance matrix, may be particularly useful for morphological data. The advantage of the graphical modeling approach in this case is that other than a general “size” factor, one need not make any assumptions about the structure of the local factors. Rather than proving fatal to a graphical modeling approach, the scenario that M&B outline seems to highlight the power and simplicity of the graphical modeling framework.

BIOLGICAL APPLICATIONS OF GRAPHICAL MODELS

Graphical models have seen wide application to biological problems in recent years. Both directed and undirected graphical models have been applied to the study of gene regulatory networks (e.g., Pe’er et al. 2001;
To summarize, GGMs and related methods are a robust set of statistical techniques appropriate for studying the covariance structure of multivariate data sets. GGMs have found wide application in both biological and nonbiological contexts and have been shown to be a useful tool for both exploratory and confirmatory studies of integration and modularity (e.g., Magwene 2001; Polanski and Franciscus 2006; Pie and Traniello 2007). The Wright-style factor analysis approach favored by Mitteroecker and Bookstein is a related and equally useful statistical tool. In the context of studies of integration, the latter would appear to be best suited to situations where one has strong a priori assumptions about covariance structure inherent in the data (Mitteroecker and Bookstein 2007).

**REFERENCES**


**First submitted 16 February 2008; reviews returned 14 April 2008; final acceptance 25 August 2008**

**Associate Editor: Norman MacLeod**