

Network Psychometrics

Sacha Epskamp, Gunter Maris, Lourens J. Waldorp and Denny Borsboom
University of Amsterdam

Contents

Introduction	1
Notation	2
Markov Random Fields	3
Parameterizing Markov Random Fields	4
The Ising model	5
The Ising model in Psychometrics	10
The relation between the Ising model and Item Response Theory	11
Estimating the Ising model	15
Example analysis	19
The interpretation of latent variables in psychometric models	21
Discussion	25

Introduction

In fact, statistical field theory may have even more to offer. It always struck me that there appears to be a close connection between the basic expressions underlying item-response theory and the solutions of elementary lattice fields in statistical physics. For instance, there is almost a one-to-one formal correspondence of the solution of the Ising model (a lattice with nearest neighbor interaction between binary-valued sites; e.g., Kindermann, Snell, et al. 1980, Chapter 1) and the Rasch model (Fischer 1974).

—Peter Molenaar (2003, p. 82)

In recent years, network models have been proposed as an alternative way of looking at psychometric problems (Van der Maas et al. 2006; Cramer et al. 2010; Borsboom and Cramer 2013). In these models, psychometric item responses are conceived of as proxies for variables that directly interact with each other. For example, the symptoms of depression

(such as loss of energy, sleep problems, and low self esteem) are traditionally thought of as being determined by a common latent variable (depression, or the liability to become depressed; Aggen, Neale, and Kendler 2005). In network models, these symptoms are instead hypothesised to form networks of mutually reinforcing variables (e.g., sleep problems may lead to loss of energy, which may lead to low self esteem, which may cause rumination that in turn may reinforce sleep problems). On the face of it, such network models offer an entirely different conceptualization of why psychometric variables cluster in the way that they do. However, it has also been suggested in the literature that latent variables may somehow correspond to sets of tightly intertwined observables (e.g., see the Appendix of Van der Maas et al. 2006), and as the above quote shows, Molenaar (2003) already suspected that network models in physics are closely connected to psychometric models with latent variables.

In the current chapter, we aim to make this connection explicit. As we will show, a particular class of latent variable models (namely, multidimensional Item Response Theory models) yields exactly the same probability distribution over the observed variables as a particular class of network models (namely, Ising models). In the current chapter, we exploit the consequences of this equivalence. We will first introduce the general class of models used in network analysis called Markov Random Fields. Specifically, we will discuss the Markov random field for binary data called the *Ising Model*, which originated from statistical physics but has since been used in many fields of science. We will show how the Ising Model relates to psychometrical practise, with a focus on the equivalence between the Ising Model and multidimensional item response theory. We will demonstrate how the Ising model can be estimated and finally, we will discuss the conceptual implications of this equivalence.

Notation

Throughout this chapter we will denote random variables with capital letters and possible realizations with lower case letters; vectors will be represented with bold-faced letters. For parameters, we will use boldfaced capital letters to indicate matrices instead of vectors whereas for random variables we will use boldfaced capital letters to indicate a random vector. Roman letters will be used to denote observable variables and parameters (such as the number of nodes) and Greek letters will be used to denote unobservable variables and parameters that need to be estimated.

In this chapter we will mainly model the random vector $\mathbf{X}^\top = [X_1 \ X_2 \ \dots \ X_P]$ containing P binary variables that take the values 1 (e.g., correct, true or yes) and -1 (e.g., incorrect, false or no). We will denote a realization, or *state*, of \mathbf{X} with $\mathbf{x}^\top = [x_1 \ x_2 \ \dots \ x_p]$. Let N be the number of observations and $n(\mathbf{x})$ the number of observations that have response pattern \mathbf{x} . Furthermore, let i denote the subscript of a random variable and j the subscript of a different random variable ($j \neq i$). Thus, X_i is the i th random variable and x_i its realization. The superscript $-(\dots)$ will indicate that elements are removed from a vector; for example, $\mathbf{X}^{-(i)}$ indicates the random vector \mathbf{X} without X_i : $\mathbf{X}^{-(i)} = [X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_P]$, and $\mathbf{x}^{-(i)}$ indicates its realization. Similarly, $\mathbf{X}^{-(i,j)}$ indicates \mathbf{X} without X_i and X_j and $\mathbf{x}^{-(i,j)}$ its realization. An overview of all notations used in this chapter can be seen in Appendix B.

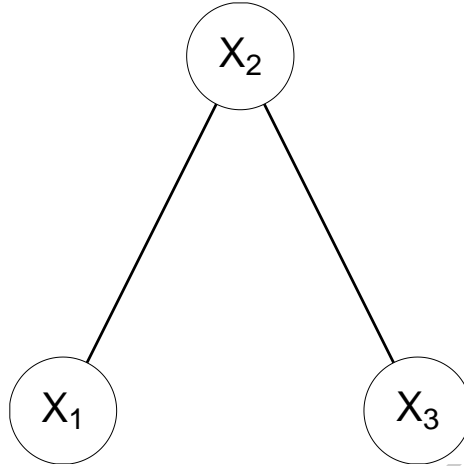


Figure 33.1. Example of a PMRF of three nodes, X_1 , X_2 and X_3 , connected by two edges, one between X_1 and X_2 and one between X_2 and X_3 .

Markov Random Fields

A network, also called a graph, can be encoded as a set G consisting of two sets: V , which contains the nodes in the network, and E , which contains the edges that connect these nodes. For example, the graph in Figure 33.1 contains three nodes: $V = \{1, 2, 3\}$, which are connected by two edges: $E = \{(1, 2), (2, 3)\}$. We will use this type of network to represent a *pairwise Markov random field* (PMRF; Lauritzen 1996; Murphy 2012), in which nodes represent observed random variables¹ and edges represent (conditional) association between two nodes. More importantly, the absence of an edge represents the Markov property that two nodes are conditionally independent given all other nodes in the network:

$$X_i \perp\!\!\!\perp X_j \mid \mathbf{X}^{-\{i,j\}} = \mathbf{x}^{-\{i,j\}} \iff (i, j) \notin E \quad (33.1)$$

Thus, a PMRF encodes the independence structure of the system of nodes. In the case of Figure 33.1, X_1 and X_3 are independent given that we know $X_2 = x_2$. This could be due to several reasons; there might be a causal path from X_1 to X_3 or vice versa, X_2 might be the common cause of X_1 and X_3 , unobserved variables might cause the dependencies between X_1 and X_2 and X_2 and X_3 , or the edges in the network might indicate actual pairwise interactions between X_1 and X_2 and X_2 and X_3 .

Of particular interest to psychometrics are models in which the presence of latent common causes induces associations among the observed variables. If such a common cause model holds, we cannot condition on any observed variable to completely remove the association between two nodes (Pearl 2000). Thus, if an unobserved variable acts as a common cause to some of the observed variables, we should find a fully connected clique in the PMRF that describes the associations among these nodes. The network in Figure 33.1, for example, cannot represent associations between three nodes that are subject to the

1. Throughout this chapter, nodes in a network designate variables, hence the terms are used interchangeably.

influence of a latent common cause; if that were the case, it would be impossible to obtain conditional independence between X_1 and X_3 by conditioning on X_2 .

Parameterizing Markov Random Fields

A PMRF can be parameterized as a product of strictly positive potential functions $\phi(x)$ (Murphy 2012):

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_i \phi_i(x_i) \prod_{\langle ij \rangle} \phi_{ij}(x_i, x_j), \quad (33.2)$$

in which \prod_i takes the product over all nodes, $i = 1, 2, \dots, P$, $\prod_{\langle ij \rangle}$ takes the product over all distinct pairs of nodes i and j ($j > i$), and Z is a normalizing constant such that the probability function sums to unity over all possible patterns of observations in the sample space:

$$Z = \sum_{\mathbf{x}} \prod_i \phi_i(x_i) \prod_{\langle ij \rangle} \phi_{ij}(x_i, x_j).$$

Here, $\sum_{\mathbf{x}}$ takes the sum over all possible realizations of \mathbf{X} . All $\phi(x)$ functions result in positive real numbers, which encode the *potentials*: the preference for the relevant part of \mathbf{X} to be in some state. The $\phi_i(x_i)$ functions encode the node potentials of the network; the preference of node X_i to be in state x_i , regardless of the state of the other nodes in the network. Thus, $\phi_i(x_i)$ maps the potential for X_i to take the value x_i regardless of the rest of the network. If $\phi_i(x_i) = 0$, for instance, then X_i will never take the value x_i , while $\phi_i(x_i) = 1$ indicates that there is no preference for X_i to take any particular value and $\phi_i(x_i) = \infty$ indicates that the system always prefers X_i to take the value x_i . The $\phi_{ij}(x_i, x_j)$ functions encode the pairwise potentials of the network; the preference of nodes X_i and X_j to both be in states x_i and x_j . As $\phi_{ij}(x_i, x_j)$ grows higher we would expect to observe $X_j = x_j$ whenever $X_i = x_i$. Note that the potential functions are not identified; we can multiply both $\phi_i(x_i)$ or $\phi_{ij}(x_i, x_j)$ with some constant for all possible outcomes of x_i , in which case this constant becomes a constant multiplier to (33.2) and is cancelled out in the normalizing constant Z . A typical identification constraint on the potential functions is to set the marginal geometric means of all outcomes equal to 1; over all possible outcomes of each argument, the logarithm of each potential function should sum to 0:

$$\sum_{x_i} \ln \phi_i(x_i) = \sum_{x_i} \ln \phi_{ij}(x_i, x_j) = \sum_{x_j} \ln \phi_{ij}(x_i, x_j) = 0 \quad \forall x_i, x_j \quad (33.3)$$

in which \sum_{x_i} denotes the sum over all possible realizations for X_i , and \sum_{x_j} denotes the sum over all possible realizations of X_j .

We assume that every node has a potential function $\phi_i(x_i)$ and nodes only have a relevant pairwise potential function $\phi_{ij}(x_i, x_j)$ when they are connected by an edge; thus, two unconnected nodes have a constant pairwise potential function which, due to identification above, is equal to 1 for all possible realizations of X_i and X_j :

$$\phi_{ij}(x_i, x_j) = 1 \quad \forall x_i, x_j \iff (i, j) \notin E. \quad (33.4)$$

From Equation (33.2) it follows that the distribution of \mathbf{X} marginalized over X_k and X_l , that is, the marginal distribution of $\mathbf{X}^{-\langle k, l \rangle}$ (the random vector \mathbf{X} without elements X_k

and X_l), has the following form:

$$\begin{aligned} \Pr(\mathbf{X}^{-(k,l)} = \mathbf{x}^{-(k,l)}) &= \sum_{x_k, x_l} \Pr(\mathbf{X} = \mathbf{x}) \\ &= \frac{1}{Z} \prod_{i \notin \{k,l\}} \phi_i(x_i) \prod_{\langle ij \notin \{k,l\} \rangle} \phi_{ij}(x_i, x_j) \\ &\quad \sum_{x_k, x_l} \left(\phi_k(x_k) \phi_l(x_l) \phi_{kl}(x_k, x_l) \prod_{i \notin \{k,l\}} \phi_{ik}(x_i, x_k) \phi_{il}(x_i, x_l) \right), \end{aligned} \quad (33.5)$$

in which $\prod_{i \notin \{k,l\}}$ takes the product over all nodes except node k and l and $\prod_{\langle ij \notin \{k,l\} \rangle}$ takes the product over all unique pairs of nodes that do not involve k and l . The expression in (33.5) has two important consequences. First, (33.5) does not have the form of (33.2); a PMRF is *not* a PMRF under marginalization. Second, dividing (33.2) by (33.5) an expression can be obtained for the conditional distribution of $\{X_k, X_l\}$ given that we know $\mathbf{X}^{-(k,l)} = \mathbf{x}^{-(k,l)}$:

$$\begin{aligned} \Pr(X_k, X_l | \mathbf{X}^{-(k,l)} = \mathbf{x}^{-(k,l)}) &= \frac{\Pr(\mathbf{X} = \mathbf{x})}{\Pr(\mathbf{X}^{-(k,l)} = \mathbf{x}^{-(k,l)})} \\ &= \frac{\phi_k^*(x_k) \phi_l^*(x_l) \phi_{kl}(x_k, x_l)}{\sum_{x_k, x_l} \phi_k^*(x_k) \phi_l^*(x_l) \phi_{kl}(x_k, x_l)}, \end{aligned} \quad (33.6)$$

in which $\phi_k^*(x_k) = \phi_k(x_k) \prod_{i \notin \{k,l\}} \phi_{ik}(x_i, x_k)$ and $\phi_l^*(x_l) = \phi_l(x_l) \prod_{i \notin \{k,l\}} \phi_{il}(x_i, x_l)$. Now, (33.6) *does* have the same form as (33.2); a PMRF *is* a PMRF under conditioning. Furthermore, if there is no edge between nodes k and l , $\phi_{kl}(x_k, x_l) = 1$ according to (33.4), in which case (33.6) reduces to a product of two independent functions of x_k and x_l which renders X_k and X_l independent; thus proving the Markov property in (33.1).

The Ising model

The node potential functions $\phi_i(x_i)$ can map a unique potential for every possible realization of X_i and the pairwise potential functions $\phi_{ij}(x_i, x_j)$ can likewise map unique potentials to every possible pair of outcomes for X_i and X_j . When the data are binary, only two realizations are possible for x_i , while four realizations are possible for the pair x_i and x_j . Under the constraint that the log potential functions should sum to 0 over all marginals, this means that in the binary case each potential function has one degree of freedom. If we let all X 's take the values 1 and -1 , there exists a conveniently loglinear model representation for the potential functions:

$$\begin{aligned} \ln \phi_i(x_i) &= \tau_i x_i \\ \ln \phi_{ij}(x_i, x_j) &= \omega_{ij} x_i x_j. \end{aligned}$$

The parameters τ_i and ω_{ij} are real numbers. In the case that $x_i = 1$ and $x_j = 1$, it can be seen that these parameters form an identity link with the logarithm of the potential functions:

$$\begin{aligned} \tau_i &= \ln \phi_i(1) \\ \omega_{ij} &= \ln \phi_{ij}(1, 1). \end{aligned}$$

These parameters are centered around 0 and have intuitive interpretations. The τ_i parameters can be interpreted as *threshold parameters*. If $\tau_i = 0$ the model does not prefer to be in one state or the other, and if τ_i is higher (lower) the model prefers node X_i to be in state 1 (-1). The ω_{ij} parameters are the *network parameters* and denote the pairwise interaction between nodes X_i and X_j ; if $\omega_{ij} = 0$ there is no edge between nodes X_i and X_j :

$$\omega_{ij} \begin{cases} = 0 & \text{if } (i, j) \notin E \\ \in \mathbb{R} & \text{if } (i, j) \in E \end{cases}. \quad (33.7)$$

The higher (lower) ω_{ij} becomes, the more nodes X_i and X_j prefer to be in the same (different) state. Implementing these potential functions in (33.2) gives the following distribution for \mathbf{X} :

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left(\sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j \right) \quad (33.8)$$

$$Z = \sum_{\mathbf{x}} \exp \left(\sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j \right),$$

which is known as the Ising model (Ising 1925).

Table 33.1

Probability of all states from the network in Figure 33.1.

x_1	x_2	x_3	Potential	Probability
-1	-1	-1	3.6693	0.3514
1	-1	-1	1.1052	0.1058
-1	1	-1	0.4066	0.0389
1	1	-1	0.9048	0.0866
-1	-1	1	1.1052	0.1058
1	-1	1	0.3329	0.0319
-1	1	1	0.9048	0.0866
1	1	1	2.0138	0.1928

For example, consider the PMRF in Figure 33.1. In this network there are three nodes (X_1, X_2 and X_3), and two edges (between X_1 and X_2 , and between X_2 and X_3). Suppose these three nodes are binary, and take the values 1 and -1 . We can then model this PMRF as an Ising model with 3 threshold parameters, τ_1, τ_2 and τ_3 and two network parameters, ω_{12} and ω_{23} . Suppose we set all threshold parameters to $\tau_1 = \tau_2 = \tau_3 = -0.1$, which indicates that all nodes have a general preference to be in the state -1 . Furthermore we can set the two network parameters to $\omega_{12} = \omega_{23} = 0.5$. Thus, X_1 and X_2 prefer to be in the same state, and X_2 and X_3 prefer to be in the same state as well. Due to these interactions, X_1 and X_3 become associated; these nodes also prefer to be in the same state, even though they are independent once we condition on X_2 . We can then compute the non-normalized potentials $\exp \left(\sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j \right)$ for all possible outcomes of \mathbf{X} and finally divide that value by the sum over all non-normalized potentials to compute the probabilities of each possible outcome. For instance, for the state $X_1 = -1, X_2 = 1$ and $X_3 = -1$, we

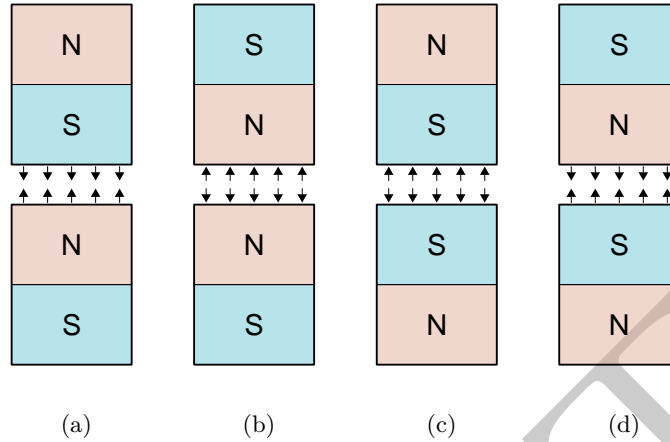


Figure 33.2. Example of the effect of holding two magnets with a north and south pole close to each other. The arrows indicate the direction the magnets want to move; the same poles, as in (b) and (c), repulse each other and opposite poles, as in (a) and (d), attract each other.

can compute the potential as $\exp(-0.1 + 0.1 + -0.1 + -0.5 + -0.5) \approx 0.332$. Computing all these potentials and summing them leads to the normalizing constant of $Z \approx 10.443$, which can then be used to compute the probabilities of each state. These values can be seen in Table 33.1. Not surprisingly, the probability $P(X_1 = -1, X_2 = -1, X_3 = -1)$ is the highest probable state in Table 33.1, due to the the threshold parameters being all negative. Furthermore, the probability $P(X_1 = 1, X_2 = 1, X_3 = 1)$ is the second highest probability in Table 33.1; if one node is put into state 1 then all nodes prefer to be in that state due to the network structure.

The Ising model was introduced in statistical physics, to explain the phenomenon of magnetism. To this end, the model was originally defined on a field of particles connected on a lattice. We will give a short introduction on this application in physics because it exemplifies an important aspect of the Ising model; namely, that the interactions between nodes can lead to synchronised behaviour of the system as a whole (e.g., spontaneous magnetisation). To explain how this works, note that a magnet, such as a common household magnet or the arrow in a compass, has two poles: a north pole and a south pole. Figure 33.2 shows the effect of pushing two such magnets together; the north pole of one magnet attracts to the south pole of another magnet and vice versa, and the same poles on both magnets repulse each other. This is due to the generally tendency of magnets to align, called *ferromagnetism*. Exactly the same process causes the arrow of a compass to align with the magnetic field of the Earth itself, causing it to point north. Any material which is ferromagnetic, such as a plate of iron, consists of particles that behave in the same way as magnets; they have a north and south pole and lie in some direction. Suppose the particles can only lie in two directions: the north pole can be up or the south pole can be up. Figure 33.3a shows a simple 2-dimensional representation of a possible state for a field of 4×4 particles. We can encode each particle as a random variable, X_i , which can take the values -1 (south pole is up) and 1 (north pole is up). Furthermore we can assume that

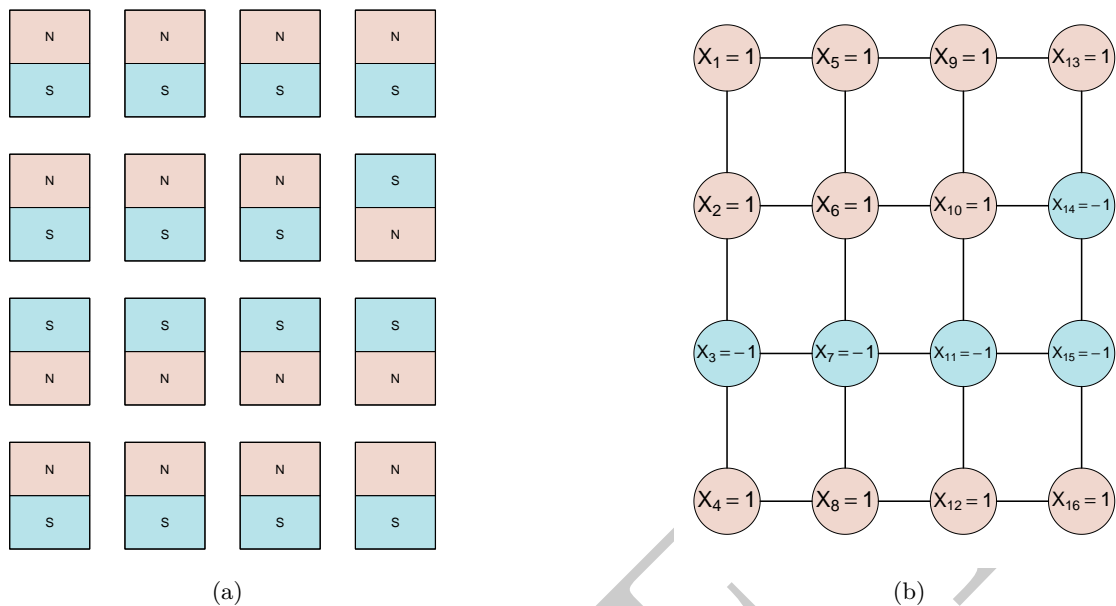


Figure 33.3. A field of particles (a) can be represented by a network shaped as a lattice as in (b). +1 indicates that the north pole is aligned upwards and -1 indicates that the south pole is aligned upwards. The lattice in (b) adheres to a PMRF in that the probability of a particle (node) being in some state is only dependent on the state of its direct neighbors.

the probability of X_i being in state x_i only depends on the direct neighbors (north, south east and west) of particle i . With this assumption in place, the system in Figure 33.3a can be represented as a PMRF on a lattice, as represented in Figure 33.3b.

A certain amount of energy is required for a system of particles to be in some state, such as in Figure 33.2. For example, in Figure 33.3b the node X_7 is in the state -1 (south pole up). Its neighbors X_3 and X_{11} are both in the same state and thus aligned, which reduces stress on the system and thus reduces the energy function. The other neighbors of X_7 , X_6 and X_8 , are in the opposite state of X_7 , and thus are not aligned, which increasing the stress on the system. The total energy configuration can be summarized in the *Hamiltonian* function:

$$H(\mathbf{x}) = - \sum_i \tau_i x_i - \sum_{\langle i,j \rangle} \omega_{ij} x_i x_j,$$

which is used in the Gibbs distribution (Murphy 2012) to model the probability of \mathbf{X} being in some state \mathbf{x} :

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{\exp(-\beta H(\mathbf{x}))}{Z}. \quad (33.9)$$

The parameter β indicates the inverse temperature of the system, which is not identifiable since we can multiply β with some constant and divide all τ and ω parameters with that same constant to obtain the same probability. Thus, it can arbitrarily be set to $\beta = 1$. Furthermore, the minus signs in the Gibbs distribution and Hamiltonian cancel out, leading to the Ising model as expressed in (33.8).

The threshold parameters τ_i indicate the natural disposition for particle i to point up or down, which could be due to the influence of an external magnetic field not part of the system of nodes in \mathbf{X} . For example, suppose we model a single compass, there is only one node thus the Hamiltonian reduces to $-\tau x$. Let $X = 1$ indicate the compass points north and $X = -1$ indicate the compass points south. Then, τ should be positive as the compass has a natural tendency to point north due to the presence of the Earth's magnetic field. As such, the τ parameters are also called external fields. The network parameters ω_{ij} indicate the interaction between two particles. Its sign indicates if particles i and j tend to be in the same state (positive; ferromagnetic) or in different states (negative; anti-ferromagnetic). The absolute value, $|\omega_{ij}|$, indicates the strength of interaction. For any two non-neighboring particles ω_{ij} will be 0 and for neighboring particles the stronger ω_{ij} the stronger the interaction between the two. Because the closer magnets, and thus particles, are moved together the stronger the magnetic force, we can interpret $|\omega_{ij}|$ as a measure for *closeness* between two nodes.

While the inverse temperature β is not identifiable in the sense of parameter estimation, it is an important element in the Ising model; in physics the temperature can be manipulated whereas the ferromagnetic strength or distance between particles cannot. The inverse temperature plays a crucial part in the *entropy* of (33.9) (Wainwright and Jordan 2008):

$$\begin{aligned} \text{Entropy}(\mathbf{X}) &= \mathbb{E}[-\ln \Pr(\mathbf{X} = \mathbf{x})] \\ &= -\beta \mathbb{E} \left[-\ln \frac{\exp(-H(\mathbf{x}))}{Z^*} \right], \end{aligned} \quad (33.10)$$

in which Z^* is the rescaled normalizing constant without inverse temperature β . The expectation $\mathbb{E} \left[-\ln \frac{\exp(-H(\mathbf{x}))}{Z^*} \right]$ can be recognized as the entropy of the Ising model as defined in (33.8). Thus, the inverse temperature β directly scales the entropy of the Ising model. As β shrinks to 0, the system is “heated up” and all states become equally likely, causing a high level of entropy. If β is subsequently increased, then the probability function becomes concentrated on a smaller number of states, and the entropy shrinks to eventually only allow the state in which all particles are aligned. The possibility that all particles become aligned is called *spontaneous magnetization* (Lin 1992; Kac 1966); when all particles are aligned (all X are either 1 or -1) the entire field of particles becomes magnetized, which is how iron can be turned into a permanent magnet. We take this behavior as a particular important aspect of the Ising model; behavior on microscopic level (interactions between neighboring particles) can cause noticeable behavior on macroscopic level (the creation of a permanent magnet).

In our view, psychological variables may behave in the same way. For example, interactions between components of a system (e.g., symptoms of depression) can cause synchronised effects of the system as a whole (e.g., depression as a disorder). Do note that, in setting up such analogies, we need to interpret the concepts of closeness and neighbourhood less literally than in the physical sense. Concepts such as “sleep deprivation” and “fatigue” can be said to be close to each other, in that they mutually influence each other; sleep deprivation can lead to fatigue and in turn fatigue can lead to a disrupted sleeping rhythm. The neighborhood of these symptoms can then be defined as the symptoms that frequently co-occur with sleep deprivation and fatigue, which can be seen in a network as a cluster of

connected nodes. As in the Ising model, the state of these nodes will tend to be the same if the connections between these nodes are positive. This leads to the interpretation that a latent trait, such as depression, can be seen as a cluster of connected nodes (Borsboom et al. 2011). In the next section, we will prove that there is a clear relationship between network modeling and latent variable modeling; indeed, clusters in a network can cause data to behave as if it were generated by a latent variable model.

The Ising model in Psychometrics

In this section, we show that the Ising model is equivalent or closely related to prominent modelling techniques in psychometrics. We will first discuss the relationship between the Ising model and loglinear analysis and logistic regressions, next show that the Ising model can be equivalent to Item Response Theory (IRT) models which dominate psychometrics. In addition, we highlight relevant earlier work on the relationship between IRT and the Ising model.

To begin, we can gain further insight in the Ising model by looking at the conditional distribution of X_i given that we know the value of the remaining nodes: $\mathbf{X}^{(-i)} = \mathbf{x}^{(-i)}$:

$$\begin{aligned} \Pr(X_i | \mathbf{X}^{(-i)} = \mathbf{x}^{(-i)}) &= \frac{\Pr(\mathbf{X} = \mathbf{x})}{\Pr(\mathbf{X}^{(-i)} = \mathbf{x}^{(-i)})} \\ &= \frac{\Pr(\mathbf{X} = \mathbf{x})}{\sum_{x_i} \Pr(X_i = x_i, \mathbf{X}^{(-i)} = \mathbf{x}^{(-i)})} \\ &= \frac{\exp(x_i (\tau_i + \sum_j \omega_{ij} x_j))}{\sum_{x_i} \exp(x_i (\tau_i + \sum_j \omega_{ij} x_j))}, \end{aligned} \quad (33.11)$$

in which \sum_{x_i} takes the sum over both possible outcomes of x_i . We can recognize this expression as a *logistic regression* model (Agresti 1990). Thus, the Ising model can be seen as the joint distribution of response and predictor variables, where each variable is predicted by all other nodes in the network. The Ising model therefore forms a predictive network in which the neighbors of each node, the set of connected nodes, represent the variables that predict the outcome of the node of interest.

Note that the definition of Markov random fields in (33.2) can be extended to include higher order interaction terms:

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_i \phi_i(x_i) \prod_{\langle ij \rangle} \phi_{ij}(x_i, x_j) \prod_{\langle ijk \rangle} \phi_{ijk}(x_i, x_j, x_k) \cdots,$$

all the way up to the P -th order interaction term, in which case the model becomes saturated. Specifying $\nu_{\dots}(\dots) = \ln \phi_{\dots}(\dots)$ for all potential functions, we obtain a log-linear model:

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp \left(\sum_i \nu_i(x_i) + \sum_{\langle ij \rangle} \nu_{ij}(x_i, x_j) + \sum_{\langle ijk \rangle} \nu_{ijk}(x_i, x_j, x_k) \cdots \right).$$

Let $n(\mathbf{x})$ be the number of respondents with response pattern \mathbf{x} from a sample of N respondents. Then, we may model the expected frequency $n(\mathbf{x})$ as follows:

$$\begin{aligned} \mathbb{E}[n(\mathbf{x})] &= N \Pr(\mathbf{X} = \mathbf{x}) \\ &= \exp\left(\nu + \sum_i \nu_i(x_i) + \sum_{\langle ij \rangle} \nu_{ij}(x_i, x_j) + \sum_{\langle ijk \rangle} \nu_{ijk}(x_i, x_j, x_k) \cdots\right), \end{aligned} \quad (33.12)$$

in which $\nu = \ln N - \ln Z$. The model in (33.12) has extensively been used in loglinear analysis (Agresti 1990; Wickens 1989)². In loglinear analysis, the same constraints are typically used as in (33.3); all ν functions should sum to 0 over all margins. Thus, if at most second-order interaction terms are included in the loglinear model, it is equivalent to the Ising model and can be represented exactly as in (33.8). The Ising model, when represented as a loglinear model with at most second-order interactions, has been used in various ways. Agresti (1990) and Wickens (1989) call the model the *homogeneous association* model. Because it does not include three-way or higher order interactions, the association between X_i and X_j —the odds-ratio—is constant for any configuration of $\mathbf{X}^{-(i,j)}$. Also, Cox (1972; Cox and Wermuth 1994) used the same model, but termed it the *quadratic exponential binary distribution*, which has since often been used in biometrics and statistics (e.g., Fitzmaurice, Laird, and Rotnitzky 1993; Zhao and Prentice 1990). Interestingly, none of these authors mention the Ising model.

The relation between the Ising model and Item Response Theory

In this section we will show that the Ising model is a closely related modeling framework of Item Response Theory (IRT), which is of central importance to psychometrics. In fact, we will show that the Ising model is equivalent to a special case of the multivariate 2-parameter logistic model (MIRT). However, instead of being hypothesised common causes of the item responses, in our representation the latent variables in the model are *generated* by cliques in the network.

In IRT, the responses on a set of binary variables \mathbf{X} are assumed to be determined by an set of M ($M \leq P$) latent variables $\boldsymbol{\Theta}^\top = [\Theta_1 \ \Theta_2 \ \dots \ \Theta_M]$. These latent variables are often denoted as *abilities*, which betrays the roots of the model in educational testing. In IRT, the probability of obtaining a realization x_i on the variable X_i —often called *items*—is modeled through item response functions, which model the probability of obtaining one of the two possible responses (typically, scored 1 for correct responses and 0 for incorrect responses) as a function of $\boldsymbol{\theta}$. For instance, in the Rasch (1960) model, also called the one parameter logistic model (1PL), only one latent trait is assumed ($M = 1$ and $\boldsymbol{\Theta} = \Theta$) and the conditional probability of a response given the latent trait takes the form of a simple logistic function:

$$\Pr(X_i = x_i \mid \Theta = \theta)_{1\text{PL}} = \frac{\exp(x_i \alpha (\theta - \delta_i))}{\sum_{x_i} \exp(x_i \alpha (\theta - \delta_i))},$$

² both Agresti and Wickens used λ rather than ν to denote the log potentials, which we changed in this chapter to avoid confusion with eigenvalues and the LASSO tuning parameter.

in which δ_i acts as a *difficulty parameter* and α is a common *discrimination parameter* for all items. A typical generalization of the 1PL is the Birnbaum (1968) model, often called the two-parameter logistic model (2PL), in which the discrimination is allowed to vary between items:

$$\Pr(X_i = x_i \mid \Theta = \theta)_{2\text{PL}} = \frac{\exp(x_i \alpha_i (\theta - \delta_i))}{\sum_{x_i} \exp(x_i \alpha_i (\theta - \delta_i))}.$$

The 2PL reduces to the 1PL if all discrimination parameters are equal: $\alpha_1 = \alpha_2 = \dots = \alpha$. Generalizing the 2PL model to more than 1 latent variable ($M > 1$) leads to the 2PL multidimensional IRT model (MIRT; Reckase 2009):

$$\Pr(X_i = x_i \mid \Theta = \boldsymbol{\theta})_{\text{MIRT}} = \frac{\exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}{\sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}, \quad (33.13)$$

in which $\boldsymbol{\theta}$ is a vector of length M that contains the realization of Θ , while $\boldsymbol{\alpha}_i$ is a vector of length M that contains the discrimination of item i on every latent trait in the multidimensional space. The MIRT model reduces to the 2PL model if $\boldsymbol{\alpha}_i$ equals zero in all but one of its elements.

Because IRT assumes local independence—the items are independent of each other after conditioning on the latent traits—the joint conditional probability of $\mathbf{X} = \mathbf{x}$ can be written as product of the conditional probabilities of each item:

$$\Pr(\mathbf{X} = \mathbf{x} \mid \Theta = \boldsymbol{\theta}) = \prod_i \Pr(X_i = x_i \mid \Theta = \boldsymbol{\theta}). \quad (33.14)$$

The marginal probability, and thus the likelihood, of the 2PL MIRT model can be obtained by integrating over distribution $f(\boldsymbol{\theta})$ of Θ :

$$\Pr(\mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} f(\boldsymbol{\theta}) \Pr(\mathbf{X} = \mathbf{x} \mid \Theta = \boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (33.15)$$

in which the integral is over all M latent variables. For typical distributions of Θ , such as a multivariate Gaussian distribution, this likelihood does not have a closed form solution. Furthermore, as M grows it becomes hard to numerically approximate (33.15). However, if the distribution of Θ is chosen such that it is conditionally Gaussian—the posterior distribution of Θ given that we observed $\mathbf{X} = \mathbf{x}$ takes a Gaussian form—we *can* obtain a closed form solution for (33.15). Furthermore, this closed form solution is, in fact, the Ising model as presented in (33.8).

As also shown by Marsman et al. (2015) and in more detail in Appendix A, after reparameterizing $\tau_i = -\delta_i$ and $-2\sqrt{\lambda_j/2q_{ij}} = \alpha_{ij}$, in which q_{ij} is the i th element of the j th eigenvector of $\boldsymbol{\Omega}$ (with an arbitrary diagonal chosen such that $\boldsymbol{\Omega}$ is positive definite), the Ising model is equivalent to a MIRT model in which the posterior distribution of the latent traits is equal to the product of univariate normal distributions with equal variance:

$$\Theta_j \mid \mathbf{X} = \mathbf{x} \sim N\left(\pm \frac{1}{2} \sum_i \alpha_{ij} x_i, \sqrt{\frac{1}{2}}\right).$$

The mean of these univariate posterior distributions for Θ_j is equal to the weighted sumscore $\pm \frac{1}{2} \sum_i a_{ij} x_i$. Finally, since

$$f(\boldsymbol{\theta}) = \sum_{\mathbf{x}} f(\boldsymbol{\theta} | \mathbf{X} = \mathbf{x}) \Pr(\mathbf{X} = \mathbf{x}),$$

we can see that the marginal distribution of $\boldsymbol{\Theta}$ in (33.15) is a *mixture of multivariate Gaussian distributions with homogenous variance-covariance*, with the mixing probability equal to the marginal probability of observing each response pattern.

Whenever $\alpha_{ij} = 0$ for all i and some dimension j —i.e., none of the items discriminate on the latent trait—we can see that the marginal distribution of Θ_j becomes a Gaussian distribution with mean 0 and standard-deviation $\sqrt{1/2}$. This corresponds to complete randomness; all states are equally probable given the latent trait. When discrimination parameters diverge from 0, the probability function becomes concentrated on particular response patterns. For example, in case X_1 designates the response variable for a very easy item, while X_2 is the response variable for a very hard item, the state in which the first item is answered correctly and the second incorrectly becomes less likely. This corresponds to a decrease in entropy and, as can be seen in (33.10), is related to the *temperature* of the system. The lower the temperature, the more the system prefers to be in states in which all items are answered correctly or incorrectly. When this happens, the distribution of Θ_j diverges from a Gaussian distribution and becomes a bi-modal distribution with two peaks, centered around the weighted sumscores that correspond to situations in which all items are answered correctly or incorrectly. If the entropy is relatively high, $f(\Theta_j)$ can be well approximated by a Gaussian distribution, whereas if the entropy is (extremely) low a mixture of two Gaussian distributions best approximates $f(\Theta_j)$.

For example, consider again the network structure of Figure 33.1. When we parameterized all threshold functions $\tau_1 = \tau_2 = \tau_3 = -0.1$ and all network parameters $\omega_{12} = \omega_{23} = 0.5$ we obtained the probability distribution as specified in Table 33.1. We can form the matrix $\boldsymbol{\Omega}$ first with zeroes on the diagonal:

$$\begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix},$$

which is not positive semi-definite. Subtracting the lowest eigenvalue (-0.707) from the diagonal gives us a positive semi-definite $\boldsymbol{\Omega}$ matrix:

$$\boldsymbol{\Omega} = \begin{bmatrix} 0.707 & 0.5 & 0 \\ 0.5 & 0.707 & 0.5 \\ 0 & 0.5 & 0.707 \end{bmatrix}.$$

It's eigenvalue decomposition is as follows:

$$\mathbf{Q} = \begin{bmatrix} 0.500 & 0.707 & 0.500 \\ 0.707 & 0.000 & -0.707 \\ 0.500 & -0.707 & 0.500 \end{bmatrix}$$

$$\boldsymbol{\lambda} = \begin{bmatrix} 1.414 & 0.707 & 0.000 \end{bmatrix}.$$

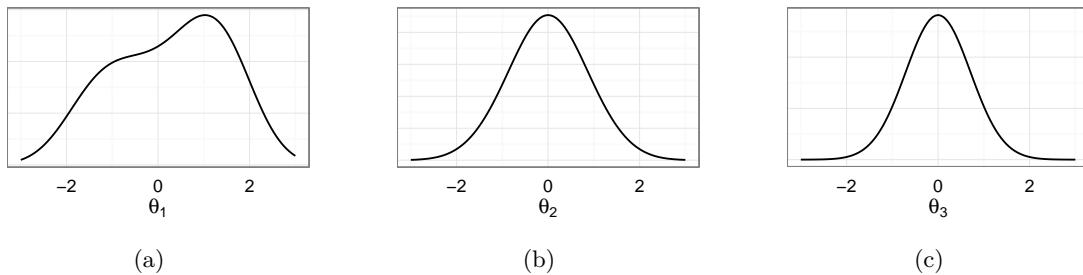


Figure 33.4. The distributions of the three latent traits in the equivalent MIRT model to the Ising model from Figure 33.1

Using the transformations $\tau_i = -\delta_i$ and $-2\sqrt{\lambda_j/2q_{ij}} = \alpha_{ij}$ (arbitrarily using the negative root) defined above we can then form the equivalent MIRT model with discrimination parameters \mathbf{A} and difficulty parameters $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} = \begin{bmatrix} 0.1 & 0.1 & 0.1 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.841 & 0.841 & 0 \\ 1.189 & 0 & 0 \\ 0.841 & -0.841 & 0 \end{bmatrix}.$$

Thus, the model in Figure 33.1 is equivalent to a model with two latent traits: one defining the general coherence between all three nodes and one defining the contrast between the first and the third node. The distributions of all three latent traits can be seen in Figure 33.4. In Table 33.1, we see that the probability is the highest for the two states in which all three nodes take the same value. This is reflected in the distribution of the first latent trait in 33.4a: because all discrimination parameters relating to this trait are positive, the weighted sumscores of $X_1 = X_2 = X_3 = -1$ and $X_1 = X_2 = X_3 = 1$ are dominant and cause a small bimodality in the distribution. For the second trait, 33.4b shows an approximately normal distribution, because this trait acts as a contrast and cancels out the preference for all variables to be in the same state. Finally, the third latent trait is nonexistent, since all of its discrimination parameters equal 0; 33.4c simply shows a Gaussian distribution with standard deviation $\sqrt{\frac{1}{2}}$.

This proof serves to demonstrate that the Ising model is equivalent to a MIRT model with a posterior Gaussian distribution on the latent traits; the discrimination parameter column vector $\boldsymbol{\alpha}_j$ —the item discrimination parameters on the j th dimension—is directly related to the j th eigenvector of the Ising model graph structure $\boldsymbol{\Omega}$, scaled by its j th eigenvector. Thus, the latent dimensions are orthogonal, and the rank of $\boldsymbol{\Omega}$ directly corresponds to the number of latent dimensions. In the case of a Rasch model, the rank of $\boldsymbol{\Omega}$ should be 1 and all ω_{ij} should have exactly the same value, corresponding to the common discrimination parameter; for the uni-dimensional Birnbaum model the rank of $\boldsymbol{\Omega}$ still is 1 but now the the ω_{ij} parameters can vary between items, corresponding to differences in item discrimination.

The use of a posterior Gaussian distribution to obtain a closed form solution for (33.15) is itself not new in the psychometric literature, although it has not previously been linked to the Ising model and the literature related to it. Olkin and Tate (1961)

already proposed to model binary variables jointly with conditional Gaussian distributed continuous variables. Furthermore, Holland (1990) used the ‘‘Dutch identity’’ to show that a representation equivalent to an Ising model could be used to characterize the marginal distribution of an extended Rasch model (Cressie and Holland 1983). Based on these results, Anderson and colleagues proposed an IRT modeling framework using log-multiplicative association models and assuming conditional Gaussian latents (Anderson and Vermunt 2000; Anderson and Yu 2007); this approach has been implemented in the R package ‘‘plRasch’’ (Anderson, Li, and Vermunt 2007; Li and Hong 2014).

With our proof we furthermore show that the clique factorization of the network structure *generated* a latent trait with a functional distribution through a mathematical trick. Thus, the network perspective and common cause perspectives could be interpreted as two different explanations of the same phenomena: cliques of correlated observed variables. In the next section, we show how the Ising model can be estimated.

Estimating the Ising model

We can use (33.8) to obtain the log-likelihood function of a realization \mathbf{x} :

$$\mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x}) = \ln \Pr(\mathbf{X} = \mathbf{x}) = \sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j - \ln Z. \quad (33.16)$$

Note that the constant Z is only constant with regard to \mathbf{x} (as it sums over all possible realizations) and is *not* a constant with regard to the τ and ω parameters; Z is often called the *partition function* because it is a function of the parameters. Thus, while when sampling from the Ising distribution Z does not need to be evaluated, but it *does* need to be evaluated when maximizing the likelihood function. Estimating the Ising model is notoriously hard because the partition function Z is often not tractable to compute (Kolaczyk 2009). As can be seen in (33.8), Z requires a sum over all possible configurations of \mathbf{x} ; computing Z requires summing over 2^k terms, which quickly becomes intractably large as k grows. Thus, maximum likelihood estimation of the Ising model is only possible for trivially small data sets (e.g., $k < 10$). For larger data sets, different techniques are required to estimate the parameters of the Ising model. Markov samplers can be used to estimate the Ising model by either approximating Z (Sebastiani and Sørbye 2002; Green and Richardson 2002; Dryden, Scarr, and Taylor 2003) or circumventing Z entirely via sampling auxiliary variables (Møller et al. 2006; Murray 2007; Murray, Ghahramani, and MacKay 2006). Such sampling algorithms can however still be computationally costly.

Because the Ising model is equivalent to the homogeneous association model in log-linear analysis (Agresti 1990), the methods used in log-linear analysis can also be used to estimate the Ising model. For example, the iterative proportional fitting algorithm (Haberman 1972), which is implemented in the `loglin` function in the statistical programming language *R* (R Core Team 2014), can be used to estimate the parameters of the Ising model. Furthermore, log-linear analysis can be used for model selection in the Ising model by setting certain parameters to zero. Alternatively, while the full likelihood in (33.8) is hard to compute, the conditional likelihood for each node in (33.11) is very easy and does not include an intractable normalizing constant; the conditional likelihood for each node

corresponds to a multiple logistic regression (Agresti 1990):

$$\mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x}) = x_i \left(\tau_i + \sum_j \omega_{ij} x_j \right) - \sum_{x_i} \exp \left(x_i \left(\tau_i + \sum_j \omega_{ij} x_j \right) \right).$$

Here, the subscript i indicates that the likelihood function is based on the conditional probability for node i given the other nodes. Instead of optimizing the full likelihood of (33.8), the pseudolikelihood (PL; Besag 1975) can be optimized instead. The pseudolikelihood approximates the likelihood with the product of univariate conditional likelihoods in (33.11):

$$\ln \text{PL} = \sum_{i=1}^k \mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x})$$

Finally, disjoint pseudolikelihood estimation can be used. In this approach, each conditional likelihood is optimized separately (Liu and Ihler 2012). This routine corresponds to repeatedly performing a multiple logistic regression in which one node is the response variable and all other nodes are the predictors; by predicting x_i from $\mathbf{x}^{(-i)}$ estimates can be obtained for $\boldsymbol{\omega}_i$ and τ_i . After estimating a multiple logistic regression for each node on all remaining nodes, a single estimate is obtained for every τ_i and two estimates are obtained for every ω_{ij} —the latter can be averaged to obtain an estimate of the relevant network parameter. Many statistical programs, such as the *R* function `glm`, can be used to perform logistic regressions. Estimation of the Ising model via log-linear modeling, maximal pseudolikelihood, and repeated multiple logistic regressions and have been implemented in the `EstimateIsing` function in the *R* package *IsingSampler* (Epskamp 2014b).

While the above mentioned methods of estimating the Ising model are tractable, they all require a considerable amount of data to obtain reliable estimates. For example, in log-linear analysis, cells in the 2^P contingency table that are zero—which will occur often if $N < 2^P$ —can cause parameter estimates to grow to ∞ (Agresti 1990), and in logistic regression predictors with low variance (e.g., a very hard item) can substantively increase standard errors (Whittaker 1990). To estimate the Ising model, P thresholds and $P(P-1)/2$ network parameter have to be estimated, while in standard log linear approaches, rules of thumb suggest that the sample size needs to be three times higher than the number of parameters to obtain reliable estimates. In psychometrics, the number of data points is often far too limited for this requirement to hold. To estimate parameters of graphical models with limited amounts of observations, therefore, regularization methods have been proposed (Meinshausen and Bühlmann 2006; Friedman, Hastie, and Tibshirani 2008).

When regularization is applied, a penalized version of the (pseudo) likelihood is optimized. The most common regularization method is ℓ_1 regularization—commonly known as the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996)—in which the sum of absolute parameter values is penalized to be under some value. Ravikumar, Wainwright, and Lafferty (2010) employed ℓ_1 -regularized logistic regression to estimate the structure of the Ising model via disjoint maximum pseudolikelihood estimation. For each node i the following expression is maximized (Friedman, Hastie, and Tibshirani 2010):

$$\max_{\tau_i, \boldsymbol{\omega}_i} [\mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x}) - \lambda \text{Pen}(\boldsymbol{\omega}_i)] \quad (33.17)$$

Where $\boldsymbol{\omega}_i$ is the i th row (or column due to symmetry) of $\boldsymbol{\Omega}$ and $\text{Pen}(\boldsymbol{\omega}_i)$ denotes the penalty function, which is defined in the LASSO as follows:

$$\text{Pen}_{\ell_1}(\boldsymbol{\omega}_i) = \|\boldsymbol{\omega}_i\|_1 = \sum_{j=1, j \neq i}^k |\omega_{ij}|$$

The λ in (33.17) is the regularization tuning parameter. The problem in above is equivalent to the constrained optimization problem:

$$\max_{\boldsymbol{\tau}_i, \boldsymbol{\omega}_i} [\mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \boldsymbol{x})], \quad \text{subject to } \|\boldsymbol{\omega}_i\|_1 < C$$

in which C is a constant that has a one-to-one monotone decreasing relationship with λ (Lee et al. 2006). If $\lambda = 0$, C will equal the sum of absolute values of the maximum likelihood solution; increasing λ will cause C to be smaller, which forces the estimates of $\boldsymbol{\omega}_i$ to shrink. Because the penalization uses absolute values, this causes parameter estimates to shrink to exactly zero. Thus, in moderately high values for λ a sparse solution to the logistic regression problem is obtained in which many coefficients equal zero; the LASSO results in simple predictive models including only a few predictors.

Ravikumar, Wainwright, and Lafferty (2010) used LASSO to estimate the neighborhood—the connected nodes—of each node, resulting in an unweighted graph structure. In this approach, an edge is selected in the model if either ω_{ij} and ω_{ji} is nonzero (the OR-rule) or if both are nonzero (the AND-rule). To obtain estimates for the weights ω_{ij} and ω_{ji} can again be averaged. The λ parameter is typically specified such that an optimal solution is obtained, which is commonly done through crossvalidation or, more recently, by optimizing the extended Bayesian information criterion (EBIC; Chen and Chen 2008; Foygel and Drton 2010, 2014; Borkulo et al. 2014).

In K -fold crossvalidation, the data is subdivided in K (usually $K = 10$) blocks. For each of these blocks a model is fitted using only the remaining $K - 1$ blocks of data, which are subsequently used to construct a prediction model for the block of interest. For a suitable range of λ values, the predictive accuracy of this model can be computed, and subsequently the λ under which the data were best predicted is chosen. If the sample size is relatively low, the predictive accuracy is typically much better for $\lambda > 0$ than it is at the maximum likelihood solution of $\lambda = 0$; it is preferred to regularize to avoid over-fitting.

Alternatively, an information criterion can be used to directly penalize the likelihood for the number of parameters. The EBIC (Chen and Chen 2008) augments the Bayesian information Criterion (BIC) with a hyperparameter γ to additionally penalize the large space of possible models (networks):

$$\text{EBIC} = -2\mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \boldsymbol{x}) + |\boldsymbol{\omega}_i| \ln(N) + 2\gamma |\boldsymbol{\omega}_i| \ln(k - 1)$$

in which $|\boldsymbol{\omega}_i|$ is the number of nonzero parameters in $\boldsymbol{\omega}_i$. Setting $\gamma = 0.25$ works well for the Ising model (Foygel and Drton 2014). An optimal λ can be chosen either for the entire Ising model, which improves parameter estimation, or for each node separately in disjoint pseudolikelihood estimation, which improves neighborhood selection. While K -fold crossvalidation does not require the computation of the intractable likelihood function, EBIC does. Thus, when using EBIC estimation λ need be chosen per node. We have implemented

ℓ_1 -regularized disjoint pseudolikelihood estimation of the Ising model using EBIC to select a tuning parameter per node in the *R* package *IsingFit* (Borkulo and Epskamp 2014; Borkulo et al. 2014), which uses *glmnet* for optimization (Friedman, Hastie, and Tibshirani 2010).

The LASSO works well in estimating sparse network structures for the Ising model and can be used in combination with crossvalidation or an information criterion to arrive at an interpretable model. However, it does so under the assumption that the true model in the population is sparse. So what if reality is not sparse, and we would not expect many missing edges in the network? As discussed earlier in this chapter, the absence of edges indicate conditional independence between nodes; if all nodes are caused by an unobserved cause we would not expect missing edges in the network but rather a low-rank network structure. In such cases, ℓ_2 regularization—also called ridge regression—can be used which uses a quadratic penalty function:

$$\text{Pen}_{\ell_2}(\boldsymbol{\omega}_i) = \|\boldsymbol{\omega}_i\|_2 = \sum_{j=1, j \neq i}^k \omega_{ij}^2$$

With this penalty parameters will not shrink to exactly zero but more or less smooth out; when two predictors are highly correlated the LASSO might pick only one where ridge regression will average out the effect of both predictors. Zou and Hastie (2005) proposed a compromise between both penalty functions in the *elastic net*, which uses another tuning parameter, α , to mix between ℓ_1 and ℓ_2 regularization:

$$\text{Pen}_{\text{ElasticNet}}(\boldsymbol{\omega}_i) = \sum_{j=1, j \neq i}^k \frac{1}{2}(1 - \alpha)\omega_{ij}^2 + \alpha|\omega_{ij}|$$

If $\alpha = 1$, the elastic net reduces to the LASSO penalty, and if $\alpha = 0$ the elastic net reduces to the ridge penalty. When $\alpha > 0$ exact zeroes can still be obtained in the solution, and sparsity increases both with λ and α . Since moving towards ℓ_2 regularization reduces sparsity, selection of the tuning parameters using EBIC is less suited in the elastic net. Crossvalidation, however, is still capable of sketching the predictive accuracy for different values of both α and λ . Again, the *R* package *glmnet* (Friedman, Hastie, and Tibshirani 2010) can be used for estimating parameters using the elastic net. We have implemented a procedure to compute the Ising model for a range of λ and α values and obtain the predictive accuracy in the *R* package *elasticIsing* (Epskamp 2014a).

One issue that is currently debated is inference of regularized parameters. Since the distribution of LASSO parameters is not well-behaved (Bühlmann and Geer 2011; Bühlmann 2013), Meinshausen, Meier, and Bühlmann (2009) developed the idea of using repeated sample splitting, where in the first sample the sparse set of variables are selected, followed by multiple comparison corrected p -values in the second sample. Another interesting idea is to remove the bias introduced by regularization, upon which ‘standard’ procedures can be used (Geer, Bühlmann, and Ritov 2013). As a result the asymptotic distribution of the so-called de-sparsified LASSO parameters is normal with the true parameter as mean and efficient variance (i.e., achieves the Cramér-Rao bound).. Standard techniques are then applied and even confidence intervals with good coverage are obtained. The limitations here are (i) the sparsity level, which has to be $\leq \sqrt{n/\ln(P)}$, and (ii) the ‘beta-min’

assumption, which imposes a lower bound on the value of the smallest obtainable coefficient (Bühlmann and Geer 2011).

Finally, we can use the equivalence between MIRT and the Ising model to estimate a low-rank approximation of the Ising Model. MIRT software, such as the *R* package *mirt* (Chalmers 2012), can be used for this purpose. More recently, Marsman et al. (2015) have used the equivalence also presented in this chapter as a method for estimating low-rank Ising model using Full-data-information estimation. A good approximation of the Ising model can be obtained if the true Ising model is indeed low-rank, which can be checked by looking at the eigenvalue decomposition of the elastic Net approximation or by sequentially estimating the first eigenvectors through adding more latent factors in the MIRT analysis or estimating sequentially higher rank networks using the methodology of Marsman et al. (2015).

Example analysis

To illustrate the methods described in this chapter we simulated two datasets, both with 500 measurements on 10 dichotomous scored items. The first dataset, dataset A, was simulated according to a multidimensional Rasch model, in which the first five items are determined by the first factor and the last five items by the second factor. Factor levels were sampled from a multivariate normal distribution with unit variance and a correlation of 0.5, while item difficulties were sampled from a standard normal distribution. The second dataset, dataset B, was sampled from a sparse network structure according to a Boltzmann Machine. A scale-free network was simulated using the Barabasi game algorithm (Barabási and Albert 1999) in the *R* package *igraph* (Csardi and Nepusz 2006) and a random connection probability of 5%. The edge weights were subsequently sampled from a uniform distribution between 0.75 and 1 (in line with the conception that most items in psychometrics relate positively with each other) and thresholds were sampled from a uniform distribution between -3 and -1 . To simulate the responses the *R* package *IsingSampler* was used. The datasets were analyzed using the *elasticIsing* package in *R* (Epskamp 2014a); 10-fold crossvalidation was used to estimate the predictive accuracy of tuning parameters λ and α on a grid of 100 logarithmically spaced λ values between 0.001 and 1 and 100 α values equally spaced between 0 and 1.

Figure 33.5 shows the results of the analyses. The left panels show the results for dataset A and the right panel show the result for dataset B. The top panels show the negative mean squared prediction error for different values of λ and α . In both datasets, regularized models perform better than unregularized models. The plateaus on the right of the graphs show the performance of the independence graph in which all network parameters are set to zero. Dataset A obtained a maximum accuracy at $\alpha = 0$ and $\lambda = 0.201$, thus in dataset A ℓ_2 -regularization is preferred over ℓ_1 regularization, which is to be expected since the data was simulated under a model in which none of the edge weights should equal zero. In dataset B a maximum was obtained at $\alpha = 0.960$ and $\lambda = 0.017$, indicating that in dataset B regularization close to ℓ_1 is preferred. The middle panels show visualizations of the obtained best performing networks made with the *qgraph* package (Epskamp et al. 2012); green edges represent positive weights, red edges negative weights and the wider and more saturated an edge the stronger the absolute weight. It can be seen that dataset A portrays two clusters while Dataset B portrays a sparse structure. Finally, the bottom

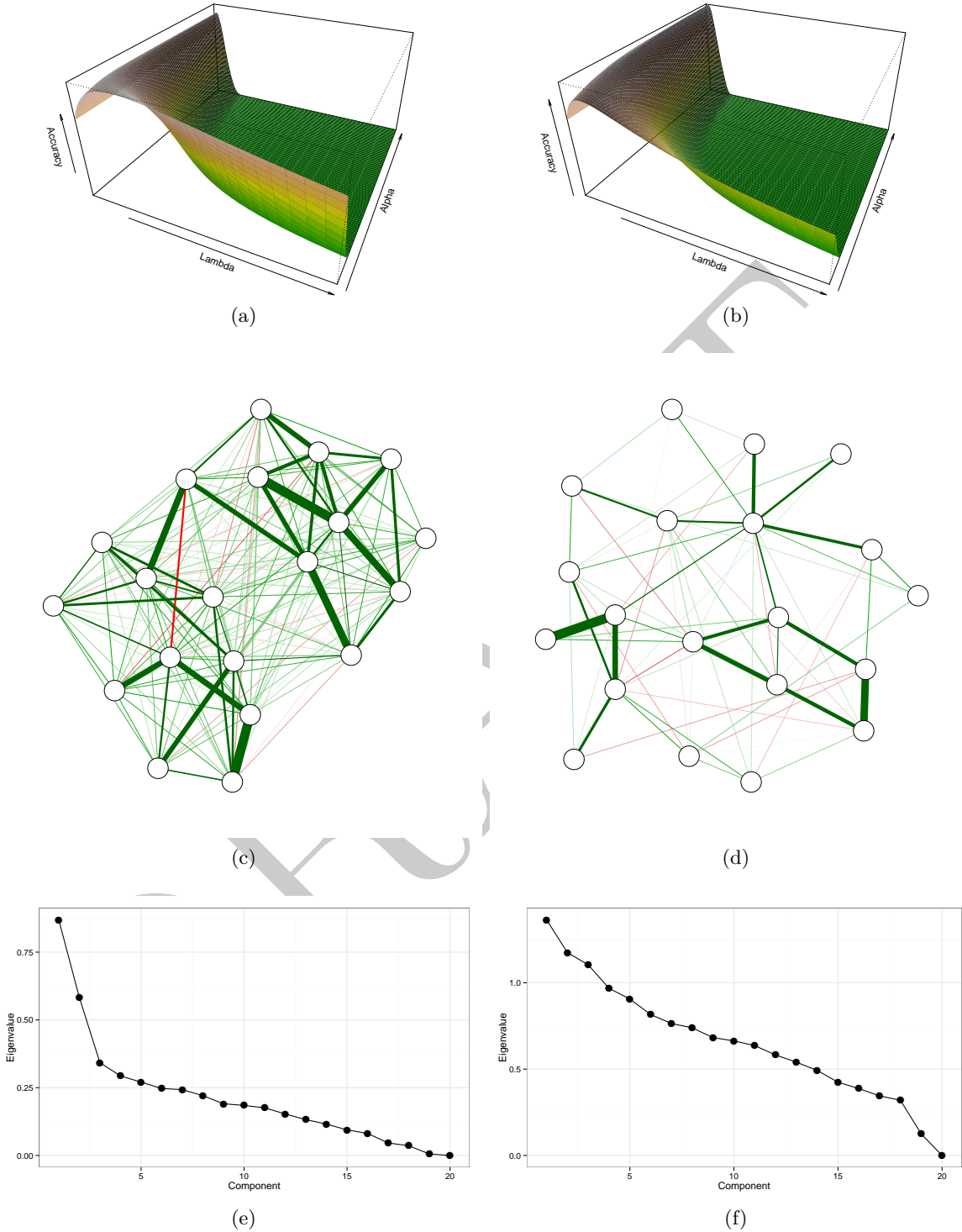


Figure 33.5. Analysis results of two simulated datasets; left panels show results based on a dataset simulated according to a 2-factor MIRT Model, while right panels show results based on a dataset simulated with a sparse scale-free network. Panels (a) and (b) show the predictive accuracy under different elastic net tuning parameters λ and α , panels (c) and (d) the estimated optimal graph structures and panels (e) and (f) the eigenvalues of these graphs.

panels show the eigenvalues of both graphs; Dataset A clearly indicates two dominant components whereas Dataset B does not indicate any dominant component.

These results show that the estimation techniques perform adequately, as expected. As discussed earlier in this chapter, the eigenvalue decomposition directly corresponds to the number of latent variables present if the common cause model is true, as is the case in dataset A. Furthermore, if the common cause model is true the resulting graph should not be sparse but low rank, as is the case in the results on dataset A.

The interpretation of latent variables in psychometric models

Since Spearman's (1904) conception of general intelligence as the common determinant of observed differences in cognitive test scores, latent variables have played a central role in psychometric models. The theoretical status of the latent variable in psychometric models has been controversial and the topic of heated debates in various subfields of psychology, like those concerned with the study of intelligence (e.g., Jensen 1998) and personality (McCrae and Costa 2008). The pivotal issue in these debates is whether latent variables posited in statistical models have referents outside of the model; that is, the central question is whether latent variables like g in intelligence or "extraversion" in personality research refer to a property of individuals that exists independently of the model fitting exercise of the researcher (Borsboom, Mellenbergh, and Van Heerden 2003; Van der Maas et al. 2006; Cramer et al. 2010). If they do have such independent existence, then the model formulation appears to dictate a causal relation between latent and observed variables, in which the former cause the latter; after all, the latent variable has all the formal properties of a common cause because it screens off the correlation between the item responses (a property denoted local independence in the psychometric literature; Borsboom 2005; Reichenbach 1991). The condition of *vanishing tetrads*, that Spearman (1904) introduced as a model test for the veracity of the common factor model is currently seen as one of the hallmark conditions of the common cause model (Bollen and Lennox 1991).

This would suggest that the latent variable model is intimately intertwined with a so-called reflective measurement model interpretation (Edwards and Bagozzi 2000; Howell, Breivik, and Wilcox 2007), also known as an effect indicators model (Bollen and Lennox 1991) in which the measured attribute is represented as the cause of the test scores. This conceptualization is in keeping with causal accounts of measurement and validity (Borsboom, Mellenbergh, and Van Heerden 2003; Markus and Borsboom 2013) and indeed seems to fit the intuition of researchers in fields where psychometric models dominate, like personality. For example, McCrae and Costa (2008) note that they assume that extraversion causes party-going behavior, and as such this trait determines the answer to the question "do you often go to parties" in a causal fashion. Jensen (1998) offers similar ideas on the relation between intelligence and the g -factor. Also, in clinical psychology, Reise and Waller (2009, p. 26) note that 'to model item responses to a clinical instrument [with IRT], a researcher must first assume that the item covariation is caused by a continuous latent variable'.

However, not all researchers are convinced that a causal interpretation of the relation between latent and observed variable makes sense. For instance, McDonald (2003) notes that the interpretation is somewhat vacuous as long as no substantive theoretical of empirical identification of the latent variable can be given; a similar point is made by

Borsboom and Cramer (2013). That is, as long as the sole evidence for the existence of a latent variable lies in the structure of the data to which it is fitted, the latent variable appears to have a merely statistical meaning and to grant such a statistical entity substantive meaning appears to be tantamount to overinterpreting the model. Thus, the common cause interpretation of latent variables at best enjoys mixed support.

A second interpretation of latent variables that has been put forward in the literature is one in which latent variables do not figure as common causes of the item responses, but as so-called behavior domains. Behavior domains are sets of behaviors relevant to substantive concepts like intelligence, extraversion, or cognitive ability (Mulaik and McDonald 1978; McDonald 2003). For instance, one can think of the behavior domain of addition as being defined through the set of all test items of the form $x + y = \dots$. The actual items in a test are considered to be a sample from that domain. A latent variable can then be conceptualized as a so-called *tail-measure* defined on the behavior domain (Ellis and Junker 1997). One can intuitively think of this as the total test score of a person on the infinite set of items included in the behavior domain. Ellis and Junker (1997) have shown that, if the item responses included in the domain satisfy the properties of monotonicity, positive association, and vanishing conditional independence, the latent variable can indeed be defined as a tail measure. The relation between the item responses and the latent variable is, in this case, not sensibly construed as causal, because the item responses are a part of the behavior domain; this violates the requirement, made in virtually all theories of causality, that cause and effect should be separate entities (Markus and Borsboom 2013). Rather, the relation between item responses and latent variable is conceptualized as a sampling relation, which means the inference from indicators to latent variable is not a species of causal inference, but of statistical generalization.

Although in some contexts the behavior domain interpretation does seem plausible, it has several theoretical shortcomings of its own. Most importantly, the model interpretation appears to beg the important explanatory question of why we observe statistical associations between item responses. For instance, Ellis and Junker (1997) manifest conditions specify that the items included in a behavior domain should look exactly as if they were generated by a common cause; in essence, the only sets of items that would qualify as behavior domains are infinite sets of items that would fit a unidimensional IRT model perfectly. The question of why such sets would fit a unidimensional model is thus left open in this interpretation. A second problem is that the model specifies infinite behavior domains (measures on finite domains cannot be interpreted as latent variables because the axioms of Ellis and Junker will not be satisfied in this case). In many applications, however, it is quite hard to come up with more than a few dozen of items before one starts repeating oneself (e.g., think of psychopathology symptoms or attitude items), and if one does come up with larger sets of items the unidimensionality requirement is typically violated. Even in applications that would seem to naturally suit the behavior domain interpretation, like the addition ability example given earlier, this is no trivial issue. Thus, the very property that buys the behavior domain interpretation its theoretical force (i.e., the construction of latent variables as tail measures on an infinite set of items that satisfies a unidimensional IRT model) is its substantive Achilles' heel.

Thus, the common cause interpretation of the latent variable model seems too make assumptions about the causal background of test scores that appear overly ambitious given

the current scientific understanding of test scores. The behaviour domain interpretation is much less demanding, but appears to be of limited use in situations where only a limited number of items is of interest and in addition offers no explanatory guidance with respect to answering the question why items hang together as they do. The network model may offer a way out of this theoretical conundrum because it specifies a third way of looking at latent variables, as explained in this chapter. As Van der Maas et al. (2006) showed, data generated under a network model could explain the positive manifold often found in intelligence research which is often described as the g factor or general intelligence; a g factor emerged from a densely connected network even though it was not “real”. This idea suggests the interpretation of latent variables as functions defined as cliques in a network of interacting components (Borsboom et al. 2011; Cramer et al. 2010, 2012). As we have shown in this chapter, this relation between networks and latent variables is quite general: given simple models of the interaction between variables, as encoded in the Ising model, one expects data that conform to psychometric models with latent variables. The theoretical importance of this result is that (a) it allows for a model interpretation that invokes no common cause of the item responses as in the reflective model interpretation, but (b) does not require assumptions about infinite behaviour domains either.

Thus, network approaches can offer a theoretical middle ground between causal and sampling interpretations of psychometric models. In a network, there clearly is nothing that corresponds to a causally effective latent variable, as posited in the reflective measurement model interpretation (Bollen and Lennox 1991; Edwards and Bagozzi 2000). The network model thus evades the problematic assignment of causal force to latent variables like the g -factor and extraversion. These arise out of the network structure as epiphenomena; to treat them as causes of item responses involves an unjustified reification. On the other hand, however, the latent variable model as it arises out of a network structure does not require the antecedent identification of an infinite set of response behaviors as hypothesized to exist in behavior domain theory. Networks are typically finite structures, that involve a limited number of nodes engaged in a limited number of interactions. Each clique in the network structure will generate one latent variable with entirely transparent theoretical properties and an analytically tractable distribution function. Of course, for a full interpretation of the Ising model analogous to that in physics, one has to be prepared to assume that the connections between nodes in the network signify actual interactions (i.e., they are not merely correlations); that is, connections between nodes are explicitly not spurious as they are in the reflective latent variable model, in which the causal effect of the latent variable produces the correlations between item responses. But if this assumption is granted, the theoretical status of the ensuing latent variable is transparent and may in many contexts be less problematic than the current conceptions in terms of reflective measurement models and behavior domains are.

Naturally, even though the Ising and IRT models have statistically equivalent representations, the interpretations of the model in terms of common causes and networks is not equivalent. That is, there is a substantial difference between the causal implications of a reflective latent variable model and of an Ising model. However, because for a given dataset the models are equivalent, distinguishing network models from common cause models requires the addition of (quasi-) experimental designs into the model. For example, suppose that in reality an Ising model holds for a set of variables; say we consider the depression

symptoms “insomnia” and “feelings of worthlessness”. The model implies that, if we were to causally intervene on the system by reducing or increasing insomnia, a change in feelings of worthlessness should ensue. In the latent variable model, in which the association between feelings of worthlessness and insomnia is entirely due to the common influence of a latent variable, an experimental intervention that changes insomnia will not be propagated through the system. In this case, the intervention variable will be associated only with insomnia, which means that the items will turn out to violate measurement invariance with respect to the intervention variable (Mellenbergh 1989; Meredith 1993). Thus, interventions on individual nodes in the system can propagate to other nodes in a network model, but not in a latent variable model. This is a testable implication in cases where one has experimental interventions that plausibly target a single node in the system. Fried et al. (2013) have identified a number of factors in depression that appear to work in this way.

Note that a similar argument does not necessarily work with variables that are causal consequences of the observed variables. Both in a latent variable model and in a network model, individual observed variables may have distinct outgoing effects, i.e., affect unique sets of external variables. Thus, insomnia may directly cause bags under the eyes, while feelings of worthlessness does not, without violating assumptions of either model. In the network model, this is because the outgoing effects of nodes do not play a role in the network if they do not feed back into the nodes that form the network. In the reflective model, this is because the model only speaks on the question of where the systematic variance in indicator variables comes from (i.e., this is produced by a latent variable), but not on what that systematic variance causes. As an example, one may measure the temperature of water by either putting a thermometer into the water, or by testing whether one can boil an egg in it. Both the thermometer reading and the boiled egg are plausibly construed as effects of the temperature in the water (the common cause latent variable in the system). However, only the boiled egg has the outgoing effect of satisfying one’s appetite.

In addition to experimental interventions on the elements of the system, a network model rather than a latent variable model allows one to deduce what would happen upon changing the connectivity of the system. In a reflective latent variable model, the associations between variables are a function of the effect of the latent variable and the amount of noise present in the individual variables. Thus, the only ways to change the correlation between items is by changing the effect of the latent variable (e.g., by restricting the variance in the latent variable so as to produce restriction of range effects in the observables) or by increasing noise in the observed variables (e.g., by increasing variability in the conditions under which the measurements are taken). Thus, in a standard reflective latent variable model, the connection between observed variables is purely a correlation, and one can only change it indirectly through the variable that have proper causal roles in the system (i.e., latent variables and error variables).

However, in a network model, the associations between observed variables are not spurious; they are real, causally potent pathways, and thus externally forced changes in connection strengths can be envisioned. Such changes will affect the behavior of the system in a way that can be predicted from the model structure. For example, it is well known that increasing the connectivity of an Ising model can change its behavior from being linear (in which the total number of active nodes grows proportionally to the strength of external perturbations of the system) to being highly nonlinear. Under a situation of high connec-

tivity, an Ising network features tipping points: in this situation, very small perturbations can have catastrophic effects. To give an example, a weakly connected network of depression symptoms could only be made depressed by strong external effects (e.g., the death of a spouse), whereas a strongly connected network could tumble into a depression through small perturbations (e.g., an annoying phone call from one's mother in law). Such a vulnerable network will also feature very specific behavior; for instance, when the network is approaching a transition, it will send out early warning signals like increased autocorrelation in a time series (Scheffer et al. 2009). Recent investigations suggest that such signals are indeed present in time series of individuals close to a transition (Leemput et al. 2014). Latent variable models have no such consequences.

Thus, there are at least three ways in which network models and reflective latent variable models can be distinguished: through experimental manipulations of individual nodes, through experimental manipulations of connections in the network, and through investigation of the behavior of systems under highly frequent measurements that allow one to study the dynamics of the system in time series. Of course, a final and direct refutation of the network model would occur if one could empirically identify a latent variable (e.g., if one could show that the latent variable in a model for depression items was in fact identical with a property of the system that could be independently identified; say, serotonin shortage in the brain). However, such identifications of abstract psychometric latent variables with empirically identifiable common causes do not appear forthcoming. Arguably, then, psychometrics may do better to bet on network explanations of association patterns between psychometric variables than to hope for the empirical identification of latent common causes.

Discussion

The correspondence between the Ising model and the MIRT model offers novel interpretations of long standing psychometric models, but also opens a gateway through which the psychometric can be connected to the physics literature. Although we have only begun to explore the possibilities that this connection may offer, the results are surprising and, in our view, offer a fresh look on the problems and challenges of psychometrics. In the current chapter, we have illustrated how network models could be useful in the and conceptualisation of psychometric data. The bridge between network models and latent variables offers research opportunities that range from model estimation to the philosophical analysis of measurement in psychology, and may very well alter our view of the foundations on which psychometric models should be built.

As we have shown, network models may yield probability distributions that are exactly equivalent to this of IRT models. This means that latent variables can receive a novel interpretation: in addition to an interpretation of latent variables as common causes of the item responses (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000), or as behaviour domains from which the responses are a sample (Ellis and Junker 1997; McDonald 2003), we can now also conceive of latent variables as mathematical abstractions that are defined on cliques of variables in a network. The extension of psychometric work to network modelling fits current developments in substantive psychology, in which network models have often been motivated by critiques of the latent variable paradigm. This has for instance happened in the context of intelligence research (Van der Maas et al. 2006), clinical psychology (Cramer

et al. 2010; Borsboom and Cramer 2013), and personality (Cramer et al. 2012; Costantini et al. 2014). It should be noted that, in view of the equivalence between latent variable models and network models proven here, even though these critiques may impinge on the common cause interpretation of latent variable models, they do not directly apply to latent variable models themselves. Latent variable models may in fact fit psychometric data well *because* these data result from a network of interacting components. In such a case, the latent variable should be thought of as a convenient fiction, but the latent variable model may nevertheless be useful; for instance, as we have argued in the current chapter, the MIRT model can be profitably used to estimate the parameters of a (low rank) network. Of course, the reverse holds as well: certain network structures may fit the data because cliques of connected network components result from unobserved common causes in the data. An important question is under which circumstances the equivalence between the MIRT model and the Ising model breaks down, i.e., which experimental manipulations or extended datasets could be used to decide between a common cause versus a network interpretation of the data. In the current paper, we have offered some suggestions for further work in this direction, which we think offers considerable opportunities for psychometric progress.

As psychometrics starts to deal with network models, we think the Ising model offers a canonical form for network psychometrics, because it deals with binary data and is equivalent to well-known models from IRT. The Ising model has several intuitive interpretations: as a model for interacting components, as an association model with at most pairwise interactions, and as the joint distribution of response and predictor variables in a logistic regression. Especially the analogy between networks of psychometric variables (e.g., psychopathology symptoms such as depressed mood, fatigue, and concentration loss) and networks of interacting particles (e.g., as in the magnetisation examples) offers suggestive possibilities for the construction of novel theoretical accounts of the relation between constructs (e.g., depression) and observables as modelled in psychometrics (e.g., symptomatology). In the current chapter, we only focused on the Ising model for binary data, but of course the work we have ignited here invites extensions in various other directions. For example, for polymotous data, the generalized Potts model could be used, although it should be noted that this model does require the response options to be discrete values that are shared over all variables, which may not suit typical psychometric applications. Another popular type of PMRF is the Gaussian Random Field (GRF; Lauritzen 1996), which has exactly the same form as the model in (33.18) except that now \boldsymbol{x} is continuous and assumed to follow a multivariate Gaussian density. This model is considerably appealing as it has a tractable normalizing constant rather than the intractable partition function of the Ising model. The inverse of the covariance matrix—the precision matrix—can be standardized as a partial correlation matrix and directly corresponds to the $\boldsymbol{\Omega}$ matrix of the Ising model. Furthermore, where the Ising model reduces to a series of logistic regressions for each node, the GRF reduces to a multiple linear regression for each node. It can easily be proven that also in the GRF the rank of the (partial) correlation matrix—cliques in the network—correspond to the latent dimensionality if the common cause model is true (Chandrasekaran, Parrilo, and Willsky 2012). A great body of literature exists on estimating and fitting GRFs even when the amount of observations is limited versus the amount of nodes (Meinshausen and Bühlmann 2006; Friedman, Hastie, and Tibshirani 2008; Foygel

and Drton 2010). Furthermore, promising methods are now available for the estimation of a GRF even in nongaussian data, provided the data are continuous (Liu, Lafferty, and Wasserman 2009; Liu et al. 2012).

References

- Aggen, Steven H, Michael C Neale, and Kenneth S Kendler. 2005. "DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models." *Psychological Medicine* 35 (04): 475–487.
- Agresti, Alan. 1990. *Categorical data analysis*. New York, NY: John Wiley & Sons Inc.
- Anderson, Carolyn J, Zhushan Li, and J Vermunt. 2007. "Estimation of models in the Rasch family for polytomous items and multiple latent variables." *Journal of Statistical Software* 20 (6): 1–36.
- Anderson, Carolyn J, and Jeroen K Vermunt. 2000. "Log-Multiplicative Association Models As Latent Variable Models For Nominal and/or Ordinal Data." *Sociological Methodology* 30 (1): 81–121.
- Anderson, Carolyn J, and Hsiu-Ting Yu. 2007. "Log-multiplicative association models as item response models." *Psychometrika* 72 (1): 5–23.
- Barabási, Albert-László, and Réka Albert. 1999. "Emergence of scaling in random networks." *science* 286 (5439): 509–512.
- Besag, Julian. 1975. "Statistical analysis of non-lattice data." *The statistician* 24:179–195.
- Birnbaum, Allan. 1968. "Some latent trait models and their use in inferring an examinee's ability." In *Statistical theories of mental test scores*, edited by F. Lord and M. Novick. Reading, MA: Addison-Wesley.
- Bollen, Kenneth, and Richard Lennox. 1991. "Conventional wisdom on measurement: A structural equation perspective." *Psychological bulletin* 110 (2): 305–314.
- Borkulo, C. D. van, Denny Borsboom, Sacha Epskamp, Tessa F Blanken, Lynn Boschloo, Robert A Schoevers, and Lourens J Waldorp. 2014. "A new method for constructing networks from binary data." *Scientific reports* 4 (5918): 1–10.
- Borkulo, C. D. van, and Sacha Epskamp. 2014. *IsingFit: Fitting Ising models using the eLasso method*. R package version 0.2.0. <http://CRAN.R-project.org/package=IsingFit>.
- Borsboom, Denny. 2005. *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.
- Borsboom, Denny, and Angélique O. J. Cramer. 2013. "Network analysis: an integrative approach to the structure of psychopathology." *Annual review of clinical psychology* 9:91–121.

- Borsboom, Denny, Angélique O. J. Cramer, Verena D Schmittmann, Sacha Epskamp, and Lourens J Waldorp. 2011. “The small world of psychopathology.” *PloS one* 6 (11): e27407.
- Borsboom, Denny, Gideon J Mellenbergh, and Jaap Van Heerden. 2003. “The theoretical status of latent variables.” *Psychological review* 110 (2): 203–219.
- Bühlmann, Peter. 2013. “Statistical significance in high-dimensional linear models.” *Bernoulli* 19 (4): 1212–1242.
- Bühlmann, Peter, and S. van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York, NY: Springer.
- Chalmers, R. Philip. 2012. “mirt: A Multidimensional Item Response Theory Package for the R Environment.” *Journal of Statistical Software* 48 (6): 1–29. <http://www.jstatsoft.org/v48/i06/>.
- Chandrasekaran, Venkat, Pablo A Parrilo, and Alan S Willsky. 2012. “Latent variable graphical model selection via convex optimization (with discussion).” *The Annals of Statistics* 40 (4): 1935–1967.
- Chen, Jiahua, and Zehua Chen. 2008. “Extended Bayesian information criteria for model selection with large model spaces.” *Biometrika* 95 (3): 759–771.
- Costantini, Giulio, Sacha Epskamp, Denny Borsboom, Marco Perugini, René Mõttus, Lourens J. Waldorp, and Angélique O. J. Cramer. 2014. “State of the aRt personality research: A tutorial on network analysis of personality data in R.” *Journal of Research in Personality*. ISSN: 0092-6566. doi:10.1016/j.jrp.2014.07.003. <http://www.sciencedirect.com/science/article/pii/S0092656614000701>.
- Cox, David R. 1972. “The analysis of multivariate binary data.” *Applied statistics* 21:113–120.
- Cox, David R, and Nanny Wermuth. 1994. “A note on the quadratic exponential binary distribution.” *Biometrika* 81 (2): 403–408.
- Cramer, Angélique O. J., Sophie Sluis, Arjen Noordhof, Marieke Wichers, Nicole Geschwind, Steven H Aggen, Kenneth S Kendler, and Denny Borsboom. 2012. “Dimensions of normal personality as networks in search of equilibrium: You can’t like parties if you don’t like people.” *European Journal of Personality* 26 (4): 414–431.
- Cramer, Angélique O. J., Lourens J Waldorp, Han L. J. van der Maas, and Denny Borsboom. 2010. “Comorbidity: a network perspective.” *Behavioral and Brain Sciences* 33 (2-3): 137–150.
- Cressie, Noel, and Paul W Holland. 1983. “Characterizing the manifest probabilities of latent trait models.” *Psychometrika* 48 (1): 129–141.
- Csardi, Gabor, and Tamas Nepusz. 2006. “The igraph software package for complex network research.” *InterJournal Complex Systems*:1695. <http://igraph.org>.

- Dryden, Ian L, Mark R Scarr, and Charles C Taylor. 2003. "Bayesian texture segmentation of weed and crop images using reversible jump Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52 (1): 31–50.
- Edwards, Jeffrey R, and Richard P Bagozzi. 2000. "On the nature and direction of relationships between constructs and measures." *Psychological methods* 5 (2): 155–174.
- Ellis, Jules L, and Brian W Junker. 1997. "Tail-measurability in monotone latent variable models." *Psychometrika* 62 (4): 495–523.
- Epskamp, Sacha. 2014a. *elasticIsing: Ising network estimation using Elastic net and k-fold cross-validation*. R package version 0.1. github.com/SachaEpskamp/elasticIsing.
- . 2014b. *IsingSampler: Sampling methods and distribution functions for the Ising model*. R package version 0.1.1. github.com/SachaEpskamp/IsingSampler.
- Epskamp, Sacha, Angélique O. J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, and Denny Borsboom. 2012. "qgraph: Network Visualizations of Relationships in Psychometric Data." *Journal of Statistical Software* 48 (4): 1–18. <http://www.jstatsoft.org/v48/i04/>.
- Fischer, Gerhard H. 1974. *Einführung in die Theorie psychologischer tests: Grundlagen und Anwendungen*. Bern, CH: Huber.
- Fitzmaurice, Garrett M, Nan M Laird, and Andrea G Rotnitzky. 1993. "Regression models for discrete longitudinal responses." *Statistical Science* 8:284–299.
- Foygel, Rina, and Mathias Drton. 2010. "Extended Bayesian information criteria for Gaussian graphical models," 23:2020–2028.
- . 2014. "High-dimensional Ising model selection with Bayesian information criteria." *arXiv preprint arXiv:1403.3374*. doi:10.1214/154957804100000000.
- Fried, EI, RM Nesse, K Zivin, C Guille, and S Sen. 2013. "Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors." *Psychological medicine*:1–10.
- Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9 (3): 432–441.
- . 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. ISSN: 1548-7660. <http://www.jstatsoft.org/v33/i01>.
- Geer, S. van de, Peter Bühlmann, and Ya'acov Ritov. 2013. "On asymptotically optimal confidence regions and tests for high-dimensional models." *arXiv preprint arXiv:1303.0518*.
- Green, Peter J, and Sylvia Richardson. 2002. "Hidden Markov models and disease mapping." *Journal of the American statistical association* 97 (460): 1055–1070.
- Haberman, Shelby J. 1972. "Log-linear fit for contingency tables—Algorithm AS51." *Applied Statistics* 21:218–225.

- Holland, Paul W. 1990. "The Dutch Identity: A new tool for the study of item response models." *Psychometrika* 55 (1): 5–18.
- Howell, Roy D, Einar Breivik, and James B Wilcox. 2007. "Reconsidering formative measurement." *Psychological methods* 12 (2): 205–218.
- Ising, Ernst. 1925. "Beitrag zur theorie des ferromagnetismus." *Zeitschrift für Physik A Hadrons and Nuclei* 31 (1): 253–258.
- Jensen, Arthur Robert. 1998. *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Kac, Mark. 1966. *Mathematical mechanism of phase transition*. New York, NY: Gordon & Breach.
- Kindermann, Ross, James Laurie Snell, et al. 1980. *Markov random fields and their applications*. Vol. 1. Providence, RI: American Mathematical Society.
- Kolaczyk, Eric D. 2009. *Statistical analysis of network data*. New York, NY: Springer.
- Lauritzen, Steffen L. 1996. *Graphical models*. Oxford, UK: Clarendon Press.
- Lee, Su-In, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. 2006. "Efficient ℓ_1 regularized logistic regression." In *Proceedings of the National Conference on Artificial Intelligence*, 21:401. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- Leemput, Ingrid A. van de, Marieke Wichers, Angélique O. J. Cramer, Denny Borsboom, Francis Tuerlinckx, Peter Kuppens, Egbert H. van Nes, et al. 2014. "Critical slowing down as early warning for the onset and termination of depression." 111 (1): 87–92.
- Li, Zhushan, and Feng Hong. 2014. *plRasch: Log Linear by Linear Association models and Rasch family models by pseudolikelihood estimation*. R package version 1.0. <http://CRAN.R-project.org/package=plRasch>.
- Lin, Keh-Ying. 1992. "Spontaneous Magnetization of the Ising Model." *Chinese Journal of Physics* 30 (3): 287–319.
- Liu, Han, Fang Han, Ming Yuan, John D Lafferty, and Larry Wasserman. 2012. "High-dimensional semiparametric Gaussian copula graphical models." *The Annals of Statistics* 40 (4): 2293–2326.
- Liu, Han, John D Lafferty, and Larry Wasserman. 2009. "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *The Journal of Machine Learning Research* 10:2295–2328.
- Liu, Qiang, and Alexander Ihler. 2012. "Distributed parameter estimation via pseudolikelihood." *Proceedings of the International Conference on Machine Learning (ICML)*.
- Markus, Keith A, and Denny Borsboom. 2013. "Reflective measurement models, behavior domains, and common causes." *New Ideas in Psychology* 31 (1): 54–64.
- Marsman, Maarten, Gunter Maris, Timo Bechger, and Cees Glas. 2015. "Bayesian inference for low-rank Ising networks." *Scientific reports* 5 (9050): 1–7.

- McCrae, Robert R, and Paul T Costa. 2008. "Empirical and theoretical status of the five-factor model of personality traits." *Sage handbook of personality theory and assessment* 1:273–294.
- McDonald, Roderick P. 2003. "Behavior Domains in Theory and in Practice." *Alberta Journal of Educational Research* 49:212–230.
- Meinshausen, Nicolai, and Peter Bühlmann. 2006. "High-dimensional graphs and variable selection with the lasso." *The Annals of Statistics* 34 (3): 1436–1462.
- Meinshausen, Nicolai, Lukas Meier, and Peter Bühlmann. 2009. "P-values for high-dimensional regression." *Journal of the American Statistical Association* 104 (488).
- Mellenbergh, Gideon J. 1989. "Item bias and item response theory." *International Journal of Educational Research* 13 (2): 127–143.
- Meredith, William. 1993. "Measurement invariance, factor analysis and factorial invariance." *Psychometrika* 58 (4): 525–543.
- Molenaar, Peter CM. 2003. "State space techniques in structural equation modeling." <http://statistics.ucla.edu/preprints/uclastat-preprint-2011:14>.
- Møller, Jesper, Anthony N Pettitt, R Reeves, and Kasper K Berthelsen. 2006. "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants." *Biometrika* 93 (2): 451–458.
- Mulaik, Stanley A, and Roderick P McDonald. 1978. "The effect of additional variables on factor indeterminacy in models with a single common factor." *Psychometrika* 43 (2): 177–192.
- Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT press.
- Murray, Iain. 2007. "Advances in Markov chain Monte Carlo methods." PhD diss., Gatsby Computational neuroscience unit, University College London.
- Murray, Iain, Zoubin Ghahramani, and David J. C. MacKay. 2006. "MCMC for doubly-intractable distributions." In *Uncertainty in Artificial Intelligence (UAI)*, 359–366. AUAI Press.
- Olkin, Ingram, and Robert Fleming Tate. 1961. "Multivariate correlation models with mixed discrete and continuous variables." *The Annals of Mathematical Statistics* 32:448–465.
- Pearl, Judea. 2000. *Causality: models, reasoning and inference*. Vol. 29. Cambridge, UK: Cambridge University Press.
- R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Rasch, G. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, DM: Danish Institute for Educational Research.

- Ravikumar, Pradeep, Martin J Wainwright, and John D Lafferty. 2010. "High-dimensional Ising model selection using ℓ_1 -regularized logistic regression." *The Annals of Statistics* 38 (3): 1287–1319.
- Reckase, Mark D. 2009. *Multidimensional item response theory*. New York, NY: Springer.
- Reichenbach, Hans. 1991. *The direction of time*. Vol. 65. Berkeley, CA: University of California Press.
- Reise, Steven P, and Niels G Waller. 2009. "Item response theory and clinical measurement." *Annual review of clinical psychology* 5:27–48.
- Scheffer, Marten, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H van Nes, Max Rietkerk, and George Sugihara. 2009. "Early-warning signals for critical transitions." *Nature* 461 (7260): 53–59.
- Sebastiani, Giovanni, and Sigrunn Holbek Sørbye. 2002. "A Bayesian method for multi-spectral image data classification." *Journal of Nonparametric Statistics* 14 (1-2): 169–180.
- Spearman, Charles. 1904. "General Intelligence," Objectively Determined and Measured." *The American Journal of Psychology* 15 (2): 201–292.
- Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58:267–288.
- Van der Maas, Han L. J., Conor V Dolan, Raoul P. P. P. Grasman, Jelte M Wicherts, Hilde M Huizenga, and Maartje E. J. Raijmakers. 2006. "A dynamical model of general intelligence: the positive manifold of intelligence by mutualism." *Psychological review* 113 (4): 842.
- Wainwright, Martin J, and Michael I Jordan. 2008. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* 1 (1-2): 1–305.
- Whittaker, Joe. 1990. *Graphical models in applied multivariate statistics*. Chichester, UK: John Wiley & Sons.
- Wickens, Thomas D. 1989. *Multway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Zhao, Lue Ping, and Ross L Prentice. 1990. "Correlated binary regression using a quadratic exponential model." *Biometrika* 77 (3): 642–648.
- Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320.

Appendix A

Proof of equivalence between the Ising model and MIRT

To prove the equivalence between the Ising model and MIRT, we first need to rewrite the Ising Model in matrix form:

$$p(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\boldsymbol{\tau}^\top \mathbf{x} + \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x}\right), \quad (33.18)$$

in which $\boldsymbol{\Omega}$ is an $P \times P$ matrix containing network parameters ω_{ij} as its elements, which corresponds in graph theory to the adjacency or weights matrix. Note that, in this representation, the diagonal values of $\boldsymbol{\Omega}$ are used. However, since x_i can be only -1 or 1 , $x_i^2 = 1$ for any combination, and the diagonal values are cancelled out in the normalizing constant Z . Thus, arbitrary values can be used in the diagonal of $\boldsymbol{\Omega}$. Since $\boldsymbol{\Omega}$ is a real and symmetrical matrix, we can take the usual eigenvalue decomposition:

$$\boldsymbol{\Omega} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^\top,$$

in which $\boldsymbol{\Lambda}$ is a diagonal matrix containing eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_P$ on its diagonal, and \mathbf{Q} is an orthonormal matrix containing eigenvectors $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_P$ as its columns. Inserting the eigenvalue decomposition into (33.18) gives:

$$p(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_i \tau_i x_i\right) \prod_j \exp\left(\frac{\lambda_j}{2} \left(\sum_i q_{ij} x_i\right)^2\right). \quad (33.19)$$

Due to the unidentified and arbitrary diagonal of $\boldsymbol{\Omega}$ we can force $\boldsymbol{\Omega}$ to be positive semi-definite—requiring all eigenvalues to be nonnegative—by shifting the eigenvalues with some constant c :

$$\boldsymbol{\Omega} + c\mathbf{I} = \mathbf{Q} (\boldsymbol{\Lambda} + c\mathbf{I}) \mathbf{Q}^\top.$$

Following the work of Kac (1966), we can use the the following identity:

$$e^{y^2} = \int_{-\infty}^{\infty} \frac{e^{-2ct-t^2}}{\sqrt{\pi}} dt,$$

with $y = \sqrt{\frac{\lambda_j}{2}} (\sum_i q_{ij} x_i)$ and $t = \theta_j$ to rewrite (33.19) as follows:

$$p(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \int_{-\infty}^{\infty} \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \exp\left(x_i \left(\tau_i + \sum_j -2\sqrt{\frac{\lambda_j}{2}} q_{ij} \theta_j\right)\right) d\boldsymbol{\theta}.$$

Reparameterizing $\tau_i = -\delta_i$ and $-2\sqrt{\frac{\lambda_j}{2}} q_{ij} = \alpha_{ij}$ we obtain:

$$p(\mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} \frac{1}{Z} \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right) d\boldsymbol{\theta}. \quad (33.20)$$

The same transformations can be used to obtain a different expression for Z :

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \sum_{\mathbf{x}} \prod_i \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right) d\boldsymbol{\theta} \\ &= \int_{-\infty}^{\infty} \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right) d\boldsymbol{\theta}. \end{aligned} \quad (33.21)$$

Finally, inserting (33.21) into (33.20), multiplying by $\frac{\prod_i \sum_{x_i} \exp(x_i(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i))}{\prod_i \sum_{x_i} \exp(x_i(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i))}$, and rearranging gives:

$$\begin{aligned} p(\mathbf{X} = \mathbf{x}) &= \int_{-\infty}^{\infty} \frac{\frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}{\int_{-\infty}^{\infty} \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right) d\boldsymbol{\theta}} \\ &\quad \cdot \prod_i \frac{\exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}{\sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)} d\boldsymbol{\theta}. \end{aligned} \quad (33.22)$$

The first part of the integral on the right hand side of (33.22) corresponds to a distribution that sums to 1 for a P -dimensional random vector $\boldsymbol{\Theta}$:

$$f(\boldsymbol{\theta}) \propto \frac{\exp\left(\sum_j -\theta_j^2\right)}{\sqrt{\pi^P}} \prod_i \sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right),$$

and the second part corresponds to the 2-parameter logistic MIRT probability of the response vector as in (33.13):

$$P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_i \frac{\exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}{\sum_{x_i} \exp\left(x_i \left(\boldsymbol{\alpha}_i^\top \boldsymbol{\theta} - \delta_i\right)\right)}.$$

We can look further at this distribution by using Bayes' rule to examine the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$:

$$\begin{aligned} f(\boldsymbol{\theta} \mid \mathbf{X} = \mathbf{x}) &\propto \Pr(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\Theta} = \boldsymbol{\theta}) f(\boldsymbol{\theta}) \\ &\propto \exp\left(\mathbf{x}^\top \mathbf{A} \boldsymbol{\theta} - \boldsymbol{\theta}^\top \boldsymbol{\theta}\right) \\ &\propto \exp\left(-\frac{1}{2} \left(\boldsymbol{\theta} - \frac{1}{2} \mathbf{A}^\top \mathbf{x}\right)^\top 2\mathbf{I} \left(\boldsymbol{\theta} - \frac{1}{2} \mathbf{A}^\top \mathbf{x}\right)\right) \end{aligned}$$

and see that the posterior distribution of $\boldsymbol{\Theta}$ is a multivariate Gaussian distribution:

$$\boldsymbol{\Theta} \mid \mathbf{X} = \mathbf{x} \sim N_P \left(\pm \frac{1}{2} \mathbf{A}^\top \mathbf{x}, \sqrt{\frac{1}{2}} \mathbf{I} \right), \quad (33.23)$$

in which \mathbf{A} is a matrix containing the discrimination parameters \mathbf{a}_i as its rows and \pm indicates that columns \mathbf{a}_j could be multiplied with -1 due to that both the positive and negative root can be used in $\sqrt{\frac{\lambda_j}{2}}$, simply indicating whether the items overall are positively or negatively influenced by the latent trait θ . Additionally, Since the variance-covariance matrix of θ equals zero in all nondiagonal elements, θ is orthogonal. Thus, the multivariate density can be decomposed as the product of univariate densities:

$$\Theta_j | \mathbf{X} = \mathbf{x} \sim N \left(\pm \frac{1}{2} \sum_i a_{ij} x_i, \sqrt{\frac{1}{2}} \right).$$

DRAFT

Appendix B
Glossary of Notation

Symbol	Dimension	Description
$\{\dots\}$		Set of distinct values.
(a, b)		Interval between a and b .
P	\mathbb{N}	Number of variables.
N	\mathbb{N}	Number of observations.
\mathbf{X}	$\{-1, 1\}^P$	Random vector of binary variables.
\mathbf{x}	$\{-1, 1\}^P$	A possible realization of \mathbf{X} .
$n(\mathbf{x})$	\mathbb{N}	Number of observations with response pattern \mathbf{x} .
i, j, k and l	$\{1, 2, \dots, P\}, j \neq i$	Subscripts of random variables.
$\mathbf{X}^{-(i)}$	$\{-1, 1\}^{P-1}$	Random vector of binary variables without X_i .
$\mathbf{x}^{-(i)}$	$\{-1, 1\}^{P-1}$	A possible realization of $\mathbf{X}^{-(i)}$.
$\mathbf{X}^{-(i,j)}$	$\{-1, 1\}^{P-2}$	Random vector of binary variables without X_i and X_j .
$\mathbf{x}^{-(i,j)}$	$\{-1, 1\}^{P-2}$	A possible realization of $\mathbf{X}^{-(i,j)}$.
$\Pr(\dots)$	$\rightarrow (0, 1)$	Probability function.
$\phi_i(x_i)$	$\{-1, 1\} \rightarrow \mathbb{R}_{>0}$	Node potential function.
$\phi_i(x_i, x_j)$	$\{-1, 1\}^2 \rightarrow \mathbb{R}_{>0}$	Pairwise potential function.
τ_i	\mathbb{R}	Threshold parameter for node X_i in the Ising model. Defined as $\tau_i = \ln \phi_i(1)$.
$\boldsymbol{\tau}$	\mathbb{R}^P	Vector of threshold parameters, containing τ_i as its i th element.
ω_{ij}	\mathbb{R}	Network parameter between nodes X_i and X_j in the Ising model. Defined as $\omega_{ij} = \ln \phi_{ij}(1, 1)$.
$\boldsymbol{\Omega}$	$\mathbb{R}^{P \times P}$ and symmetrical	Matrix of network parameters, containing ω_{ij} as its ij th element.
$\boldsymbol{\omega}_i$	\mathbb{R}^P	The i th row or column of $\boldsymbol{\Omega}$.
$\text{Pen}(\boldsymbol{\omega}_i)$	$\mathbb{R}^P \rightarrow \mathbb{R}$	Penalization function of $\boldsymbol{\omega}_i$.
β	$\mathbb{R}_{>0}$	Inverse temperature in the Ising model.
$H(\mathbf{x})$	$\{-1, 1\}^P \rightarrow \mathbb{R}$	Hamiltonian function denoting the energy of state \mathbf{x} in the Ising model.
$\nu_{\dots}(\dots)$	$\rightarrow \mathbb{R}$	The log potential functions, used in loglinear analysis.
M	\mathbb{N}	The number of latent factors.
$\boldsymbol{\Theta}$	\mathbb{R}^M	Random vector of continuous latent variables.
$\boldsymbol{\theta}$	\mathbb{R}^M	Realization of $\boldsymbol{\Theta}$.
$\mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x})$	$\rightarrow \mathbb{R}$	Likelihood function based on $\Pr(\mathbf{X} = \mathbf{x})$.
$\mathcal{L}_i(\boldsymbol{\tau}, \boldsymbol{\Omega}; \mathbf{x})$	$\rightarrow \mathbb{R}$	Likelihood function based on $\Pr(X_i = x_i \mid \mathbf{X}^{-(i)} = \mathbf{x}^{-(i)})$.
λ	$\mathbb{R}_{>0}$	LASSO tuning parameter
α	$(0, 1)$	Elastic net tuning parameter