Reliability of decisions based on tests: Fourier analysis of Boolean decision functions

Lourens Waldorp Maarten Marsman Denny Borsboom

University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 NP, the Netherlands
waldorp@uva.nl

Abstract: Items in a test are often used as a basis for making decisions and such tests are therefore required to have good psychometric properties, like unidimensionality. In many cases the sum score is used in combination with a threshold to decide between pass or fail, for instance. Here we consider whether such a decision function is appropriate, without a latent variable model, and which properties of a decision function are desirable. We consider reliability (stability) of the decision function, i.e., does the decision change upon perturbations, or changes in a fraction of the outcomes of the items (measurement error). We are concerned with questions of whether the sum score is the best way to aggregate the items, and if so why. We use ideas from test theory, social choice theory, graphical models, computer science and probability theory to answer these questions. We conclude that a weighted sum score has desirable properties that (i) fit with test theory and is observable (similar to a condition like conditional association), (ii) has the property that a decision is stable (reliable), and (iii) satisfies Rousseau's criterion that the input should match the decision. We use Fourier analysis of Boolean functions to investigate whether a decision function is stable and to figure out which (set of) items has proportionally too large an influence on the decision. To apply these techniques we invoke ideas from graphical models and use a pseudo-likelihood factorisation of the probability distribution.

Keywords and phrases: Fourier analysis, Boolean functions, noise sensitivity, test theory, network psychometrics, graphical models.

1. Introduction

Tests are frequently used to make decisions and their psychometric properties have become an important aspect of any test. Traditionally psychometric properties have been associated with the fit of a particular type of model, one that often relies on measuring constructs, which are defined in terms of latent variables (Lord and Novick, 1968; Holland and Rosenbaum, 1986; Junker, 1993). The normative function of such models determines, for instance, whether an item is considered good or bad. If the model describes the test well, it can be used to make decisions, to pass or fail, for example. Often the decision function is a (weighted) sum score of the items. However, the implications of latent variable models are often unrealistic. For instance, the latent variable is assumed to exist and causally effect the observed variables (Borsboom et al., 2004), or,

equivalently, the items are obtained from infinite events (tail events, Ellis and Junker, 1997). Moreover, it is often unclear what the consequences for the decision based on the items (e.g., a sum score) are in terms of small changes in the items. Using the sum score of the items as a decision function often appears reasonable and its main argument comes from Ellis and Junker (1997) and Junker and Ellis (1997). Ellis and Junker (1997) show that vanishing conditional dependence and conditional association are obtained using infinite events of the items (like weighted sum scores), which makes the possibility of a unidimensional latent variable model "asymptotically empirical". Here we continue to shift the focus away from latent variable models and try to determine what the requirements of an appropriate decision function, like the sum score, should be.

The decision function f is called a Boolean function if it maps the input of n binary items to a single binary outcome that can be coded as 1 for pass and 0 for fail; we also often encode this by 1 and -1, respectively. In order to get at the decision level, we propose a simple experiment where each value is possibly flipped with probability $\frac{1}{2}(1-\rho)$, with $\rho \in [-1,1]$. This simple experiment can be thought of as measurement error, where the original value has been contaminated. We show that this conception of measurement error is in line with the classical definition by Lord and Novick (1968, Section 2.7-2.9). It turns out that ρ is related to the reliability and, for coordinatewise monotone decision functions the decision reliability is proportional to reliability (at the item level). The notion of reliability is closely related to stability. Stability of the decision function refers to the idea that small changes of the items in the test (0 becomes 1 or vice versa), measurement error, results in changes in the decision. We are interested in the question whether such errors of measurement affect the decision f; if a small fraction of flipped items leads to a different decision, then the decision function f is not stable (sensitive to noise). Based on such information we can investigate the influence of the items in the test and even consider the appropriateness of the decision function itself and what form it should have.

One of the main tools we use is Fourier analysis of Boolean functions (see, e.g., De Wolf, 2008; O'Donnell, 2014). In a Fourier analysis the objective is to write the original function as a multilinear (i.e., no squares, cubes, etc.) sum of functions, so that the complexity of the Boolean function can be understood more easily. As in 'regular' Fourier analysis, a function f is written as a sum of (orthogonal) basis functions consisting of products of the original variables and their coefficients, called Fourier coefficients. The Fourier coefficients tell us something about the sensitivity of the decision function f: Stable functions have larger Fourier coefficients at smaller subsets for the basis functions (i.e., using fewer variables in the multilinear products). Additionally, the Fourier coefficients can be used to efficiently obtain information on the influence of each of the items in the test, so that their effect to possibly change the decision may be established.

To apply Fourier analysis of Boolean (decision) functions, we require a factorisation of the probability distribution, similar to independence. Traditionally, in the analysis of Boolean functions the variables are assumed to be identically

and independently distributed. But this is not necessary for the Fourier analysis, only a specific kind of factorisation is required. Here we embed the Fourier analysis in the framework of graphical models and specifically the Ising model, where we trade the joint probability for the product of conditional probabilities (pseudo-likelihood) that is consistent with the joint probability. We use local conditional independence, which refers to the idea that a variable (item) depends on (is connected to) its neighbours (boundary or Markov blanket) only, and conditionally on those neighbours, it is independent of all others. An implication of the use of this definition of independence is that only a small subset of items is involved in determining the properties of the item and not necessarily the entire set of items used in the test.

We first discuss in Section 2 what Boolean (decision) functions are and what properties they should have. Then in Section 3 we discuss the basics of the Fourier analysis for Boolean functions, first in terms of uniform distributions and independent variables, and later in more general terms in Section 4. In Section 4.2 we describe the embedding of tests in graphical models, and in the Ising model in particular, where we need a factorisation to make the Fourier analysis work. After this, we return to the main objective, the investigation into the influence of items on a test in Section 5, and the reliability and stability of the decision function in Section 6. We then use these concepts to determine the usefulness of the sum score in Section 7, where we use the definition of usefulness from Junker (1993). We describe how to apply a Fourier analysis including statistical guarantees in Section 8, which we illustrate with numerical results in Section 9. Proofs can be found in the Appendix.

2. Boolean and decision functions

A Boolean function is a mapping from the n-cube $\{0,1\}^n$ of input configurations to the binary decision $\{0,1\}$. We call such Boolean functions *decision functions* or *decision rules* because of the binary classification corresponding to a two-choice decision. An example of a decision rule is the majority function on three items. For three items $V = \{1,2,3\}$ the following well known $2^3 = 8$ patterns are possible

$$\{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$$

The majority function $\text{maj}_n : \{0,1\}^n \to \{0,1\}$ for n odd is given by

$$\text{maj}_n(x) = \begin{cases} 1 & \text{if } \sum_i x_i > n/2 \\ 0 & \text{if } \sum_i x_i < n/2 \end{cases} \tag{1}$$

Note that because n is odd we do not obtain the equality $\sum_i x_i = n/2$. For convenience we often relabel the 0 and 1 to -1 and 1, respectively, and give an equivalent way to define the majority function (for n odd). The majority function $\text{maj}_n : \{-1,1\}^n \to \{-1,1\}$ is given by

$$\text{maj}_n = \text{sgn}(x_1 + x_2 + \cdots + x_n) \tag{2}$$

for some $x \in \{-1,1\}^n$, where the function $\text{sgn} : \mathbb{R} \to \{-1,1\}$ is the sign function and $\mathbb{R}$ is the set of real numbers. For n odd maj_n is positive if there are more $+1$ than -1 and it is negative if there are more -1 than $+1$. The values for each of the 8 patterns of $\{-1,1\}^3$ are

$$\begin{array}{ll} \text{maj}_3(-1,-1,-1) = -1 & \text{maj}_3(-1,+1,+1) = +1 \\ \text{maj}_3(-1,-1,+1) = -1 & \text{maj}_3(+1,-1,+1) = +1 \\ \text{maj}_3(-1,+1,-1) = -1 & \text{maj}_3(+1,+1,-1) = +1 \\ \text{maj}_3(+1,-1,-1) = -1 & \text{maj}_3(+1,+1,+1) = +1 \end{array} \tag{3}$$

The majority function is easily seen to be a decision rule based on the sum score with a threshold of $n/2$. The majority function is an example of a class of functions called the linear threshold functions (LTF), defined by functions $f : \{-1,1\}^n \to \{-1,1\}$ such that

$$f(x) = \text{sgn}(a_0 + a_1 x_1 + \cdots + a_n x_n) \tag{4}$$

where the a_i are constants in $\mathbb{R}$ and a_0 is referred to as the threshold. LTFs are a more general way of thinking of the sum score. Often the constants $a_1, \ldots, a_n$ are scaled such that $\sum a_i^2 = 1$, which is convenient when considering contrasts or the central limit theorem. Many types of item aggregation belong to the class of LTF. For example, in many cases grading is a form of LTF, where the threshold a_0 is determined according to a percentage that will pass or fail. Or a psychiatric diagnosis may be obtained for some value a_0. But there are examples of items that will make the decision function necessarily nonlinear. For instance, in the $\{0,1\}$ labeling, item i is classified as correct if and only if item j is correct, then we obtain $x_i x_j$, and so f is not a linear function. Another example is a skip-item, where x_{i-1} has the property that when it has value -1, then a set of subsequent items need not be answered, that is,

$$\begin{cases} x_i = x_{i+1} = \cdots = x_{i+m} = \varnothing & \text{if } x_{i-1} = -1 \\ x_i, x_{i+1}, \ldots, x_{i+m} = \pm 1 & \text{if } x_{i-1} = 1 \end{cases}$$

Sometimes it is argued that all values for x_i up to x_{i+m} should be scored as -1, but in general this is a difficult issue with possibly large consequences for the decision. Even in the case where all items following the skip-item are coded as -1, then the influence (impact) of that item is often disproportional. All input functions can be written as *polynomial threshold functions*, where a polynomial is said to have degree k if the highest power in the terms of the polynomial is k. We will not discuss polynomial threshold functions here, see O'Donnell (2014) for more discussion on this.

The majority function is a relatively simple function but turns out to have strong properties. The decision (-1 or 1) based on the majority function is not very sensitive to small changes in the input. But of course, it is in general reasonable to ask what kind of properties we would like our decision function to have so that we can interpret the results in a meaningful way. There are several properties that seem reasonable (Kelly, 1988). A function $f : \{-1,1\}^n \to \{-1,1\}$ is said to be

(a) *monotone* or is *positively responsive* if for $x \leq y$ (i.e., $x_j \leq y_j$ for all j) implies that $f(x) \leq f(y)$;
(b) *odd* or *neutral* if $f(-x) = -f(x)$;
(c) *unanimous* if $f(-1, -1, \ldots, -1) = -1$ and $f(1, 1, \ldots, 1) = 1$;
(d) *symmetric* or *anonymous* if for any permutation $\pi : \{-1, 1\}^n \to \{-1, 1\}^n$ of the coordinates in x we have $f(x^\pi) = f(x)$;
(e) *transitive-symmetric* if for any $i \in V$ there is a permutation $\pi : \{-1, 1\}^n \to \{-1, 1\}^n$ of the coordinates in x that puts x_i in place of x_j, such that $f(x^\pi) = f(x)$.

Monotonicity (a) is the requirement that if the decision is $+1$, then the decision would remain $+1$ whenever there are more $+1$ items at the exact same coordinates. This property is is relevant in many contexts, like voting or testing. We will see in later sections that monotonicity provides computational and theoretical advantages over non-monotonic functions. Monotonicity also stops trivial functions being of interest: If we consider the constant function const_1 that gives a 1 no matter the input, we do not have monotonicity. For an odd function in (b) we have that if all inputs were to be reversed then the decision is reversed. This also excludes trivial functions like const_1. Unanimity in (c) requires that if all inputs are -1 (or $+1$) then the decision has to be in the same direction. The constant function does not satisfy this property. The concept symmetry or anonymity in (d) means that any decision $f(x)$ does not depend on which item was $+1$ or -1, as long as their respective sums remain the same. Its weaker version, transitive-symmetry, in (e) requires that coordinates i and j are exchangeable. There has to be a permutation for each pair i and j in V such that the decision remains the same for the original and permuted version. It is easy to see that if $f(x^\pi) = f(x)$ for any π, then it holds for any subset of permutations that takes i to coordinate j, and so (d) implies (e).

The majority function maj_n in (2) has properties (a)-(d) (see Lemma 5 in the Appendix). May's theorem says that the only function that satisfies (a), (b) and (d) is the majority function maj_n (May, 1952, but see Kelly (1988) for an excellent discussion and proof). So the majority function is an excellent candidate to use for decision making.

Another property that we deem relevant for decision functions based on test items is stability. Stability is related to the idea of measurement error and reliability (which will be made precise later). A Boolean function $f : \{-1, 1\} \to \{-1, 1\}$ is said to be

(f) *stable* if y equals x with probability $\frac{1}{2}(1 + \rho)$ for each coordinate and the probability that $f(x) = f(y)$ is high.

We consider stability in two versions. First, we consider stability upon possibly changing a single item; if y is a copy of x except for a single coordinate (with probability $\frac{1}{2}$), then we want to know the probability that $f(x) \neq f(y)$, so that the decision has changed. We call this influence of an item on the decision. Each item in a test should have approximately the same influence on the decision. We show that this idea is related to the number of connections the item has

in the graph representing conditional dependencies between items. We will also see that the parameter ρ in (f) is proportional to the reliability (correlation) from classical test theory (see, e.g., Lord and Novick, 1968). Using the classical concept of reliability we will show that decision reliability, the correlation between $f(x)$ and $f(y)$ as defined above, is also proportional to the reliability at the level of the items. In order to obtain these results we require to decompose the decision function f so that we can more easily work with the function. We use the Fourier decomposition of Boolean functions for this.

3. Fourier analysis of Boolean functions

We aim to understand how a particular Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ works, or compute its variance or covariance with other functions. Such a function could represent a particular decision on $\{-1, 1\}$ based on n inputs from $\{-1, 1\}^n$. For Fourier analysis it is more convenient to work with the labels $\{-1, 1\}$ instead of $\{0, 1\}$. This is because of the symmetry and the ease with which some factors in computations cancel. Also, we assume in this section only that the variables are independent and have uniform measure. A Fourier analysis of a Boolean function rewrites the function in terms of the reals ($\mathbb{R}$) as a multilinear polynomial. For instance, we could base a decision on the maximum function $\max_n : \{-1, 1\}^n \to \{-1, 1\}$, which for $n = 2$ is

$$\begin{array}{ll} \max_2(-1,-1) = -1 & \max_2(-1,+1) = +1 \\ \max_2(+1,-1) = +1 & \max_2(+1,+1) = +1 \end{array} \tag{5}$$

The Fourier expansion of $\max_2$ is then

$$\max_2(x_1, x_2) = \frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 - \frac{1}{2}x_1x_2 \tag{6}$$

Basically this is an interpolation of the $\max_2$ function on the reals $\mathbb{R}$ where none of the variables is raised to a power (multilinear). The Fourier coefficients $\frac{1}{2}$ and $-\frac{1}{2}$ and the size of the sets of variables tell us something about the complexity of the function. The Fourier expansion of maj_3 given in (3) is

$$\text{maj}_3(x_1, x_2, x_3) = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 - \frac{1}{2}x_1x_2x_3 \tag{7}$$

which can easily be checked. It is clear that the Fourier expansion depends on a product of variables, without squares, cubes etc.

To define the Fourier transform we define the parity function $\chi_S : \{-1, 1\}^k \to \mathbb{R}$ for some set $S \subseteq V$ of size $|S| = k$

$$\chi_S(x) = \prod_{i \in S} x_i \tag{8}$$

with the convention $\chi_\varnothing = 1$. The parity function χ_S forms an orthonormal basis for the Fourier expansion for independent variables that have uniform measure,

similar to the sine and cosine functions for ordinary Fourier analysis. To define orthogonality we require an inner product. Here we use the inner product for two functions $f, g : \{-1, 1\}^n \to \mathbb{R}$ defined by

$$\langle f, g\rangle = \frac{1}{2^n} \sum_{x\in\{-1,1\}^n} f(x)g(x) = \mathbb{E}_{\frac{1}{2}}(f(X)g(X)) \tag{9}$$

where we used the $\frac{1}{2}$ in the expectation $\mathbb{E}_{\frac{1}{2}}$ to indicate that we assume that each variable X_j is distributed uniformly on $\{-1, 1\}$. The uniform distribution is not necessary but makes some ideas easier to explain. We extend all ideas to the general case where each item has its own probability in Section 4. With the definition of the inner product we can define orthogonality by $\langle f, g\rangle = 0$. Taking $f = \chi_S$ and $g = \chi_T$ as the functions in (9) we obtain

$$\langle \chi_S, \chi_T\rangle = \begin{cases} 1 & \text{if } S = T \\ 0 & \text{if } S \neq T \end{cases} \tag{10}$$

(see Lemma 8 in the Appendix). We can now define the Fourier expansion as follows. Let $f : \{-1, 1\}^n \to \mathbb{R}$ be a Boolean function and χ_S the parity function for subsets of V. Then the *Fourier expansion* is

$$f(x) = \sum_{S\subseteq V} \hat{f}^{\frac{1}{2}}(S)\chi_S \tag{11}$$

where $\hat{f}^{\frac{1}{2}}(S)$ is the *Fourier coefficient* on S obtained with the uniform measure, defined for subset $S \subseteq V$ by

$$\hat{f}^{\frac{1}{2}}(S) = \langle f, \chi_S\rangle = \mathbb{E}_{\frac{1}{2}}(f(X)\chi_S(X)) \tag{12}$$

Note that for $S = \varnothing$ we have $\hat{f}^{\frac{1}{2}}(\varnothing) = \langle f, 1\rangle = \mathbb{E}_{\frac{1}{2}}(f(X))$.

We return to the example of the function $\max_2$ to compute the Fourier coefficients using (12). For each subset S of V we require the expectation of $\max_2(x)\chi_S$. For $S = \varnothing$ we have that $\mathbb{E}(\max_2(X)\chi_\varnothing(X)) = \mathbb{E}_{\frac{1}{2}}(\max_2(X))$, and so using (5)

$$\widehat{\max}_2(\varnothing) = \mathbb{P}_{\frac{1}{2}}(\max_2(X) = 1) - \mathbb{P}_{\frac{1}{2}}(\max_2(X) = -1) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$$

For the subsets $S = \{1\}$ or $\{2\}$, we have that $\max_2(x)x_1$ equals 1 in 3 out of 4 cases and so

$$\widehat{\max}_2(\{1\}) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$$

for both sets $\{1\}$ and $\{2\}$. Then, for $S = \{1, 2\}$ we obtain $\max_2(x)x_1x_2$ equals 1 in 1 out of 4 cases, and so

$$\widehat{\max}_2(\{1, 2\}) = \frac{1}{4} - \frac{3}{4} = -\frac{1}{2}$$

These are exactly the coefficients we had in the Fourier expansion in (6).

The Fourier coefficients for the function maj_3 function can be obtained similarly, using ther relation $\widehat{\text{maj}}_3(S) = \mathbb{E}_{\frac{1}{2}}(\text{maj}_3(X)\chi_S(X))$ for each S.

$$\begin{array}{llll} \widehat{\text{maj}}_3(\varnothing) = & 0 & \widehat{\text{maj}}_3(\{1,2\}) = & 0 \\ \widehat{\text{maj}}_3(\{1\}) = & \frac{1}{2} & \widehat{\text{maj}}_3(\{1,3\}) = & 0 \\ \widehat{\text{maj}}_3(\{2\}) = & \frac{1}{2} & \widehat{\text{maj}}_3(\{2,3\}) = & 0 \\ \widehat{\text{maj}}_3(\{3\}) = & \frac{1}{2} & \widehat{\text{maj}}_3(\{1,2,3\}) = & -\frac{1}{2} \end{array}$$

It is easily seen that the coefficients correspond to those in (7).

The Fourier expansion of a Boolean function makes it easier to determine the mean and variance of such functions. An often used relation in this context, known as Plancheral's theorem, is that for any Boolean functions $f, g : \{-1,1\}^n \to \mathbb{R}$

$$\langle f, g\rangle = \mathbb{E}_{\frac{1}{2}}(f(X)g(X)) = \sum_{S\subseteq V} \hat{f}^{\frac{1}{2}}(S)\hat{g}^{\frac{1}{2}}(S) \tag{13}$$

From this we can characterise the mean and variance in terms of Fourier coefficients. If we set $g = 1$, we already noted that in a Fourier context we select $S = \varnothing$ for χ_S, and we get

$$\langle f, 1\rangle = \mathbb{E}_{\frac{1}{2}}(f(X)) = \hat{f}^{\frac{1}{2}}(\varnothing) \tag{14}$$

And for the variance we obtain (known as Parseval's theorem)

$$\text{var}^{\frac{1}{2}}(f) = \langle f - \mathbb{E}_{\frac{1}{2}}(f(X)), f - \mathbb{E}_{\frac{1}{2}}(f(X))\rangle = \sum_{S\neq\varnothing} \hat{f}^{\frac{1}{2}}(S)^2 \tag{15}$$

The covariance is then

$$\text{cov}^{\frac{1}{2}}(f, g) = \langle f - \mathbb{E}_{\frac{1}{2}}(f(X)), g - \mathbb{E}_{\frac{1}{2}}(g(X))\rangle = \sum_{S\neq\varnothing} \hat{f}^{\frac{1}{2}}(S)\hat{g}^{\frac{1}{2}}(S) \tag{16}$$

4. Graphical models and Fourier analysis

For the Fourier analysis we have assumed so far that for each item $i \in V$ the probability was uniform over -1 and 1. This provides an unbiased situation since $\mathbb{E}_{\frac{1}{2}}(X_i) = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$ for all $i \in V$. Obviously, this is unrealistic and we need to accommodate the p-biased situation where $p \in (0, 1)$, which also allows dependence. We saw that the Fourier analysis requires orthogonality of the product functions such that $\mathbb{E}_{\frac{1}{2}}(\prod_{i\in S} x_i) = \prod_{i\in S} \mathbb{E}_{\frac{1}{2}}(x_i)$. Such an identity implies independence. Here we use the ideas from graphical models to approximate this situation where we use the product of univariate conditional probabilities (pseudo-likelihood) instead of the joint probability.

4.1. Graphical models

Let $G = (V, E)$ be an undirected graph, where $V = \{1, 2, \ldots, n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges $\{(i, j) : i, j \in V\}$, with size $|E|$. Nodes that are connected are called adjacent or neighbours. Let ∂i be the set of nodes that are adjacent to (neighbours of) node i, $\{j \in V \backslash \{i\} : (i, j) \in E\}$. For a set of nodes B we denote its adjacent nodes $\{j \in V \backslash B : (i, j) \in E \text{ and } i \in B\}$ by ∂B; the set ∂B is also referred to as the *boundary* of B. A subset of nodes W is a cutset or separator set of the graph if removing W results in two (or more) components. For instance, W is a cutset if any path between any two nodes $s \in A$ and $t \in B$ must go through some $q \in W$. A clique is a subset of nodes in $C \subset V$ such that all nodes in C are connected, that is, for any $i, j \in C$ it holds that $(i, j) \in E$. A maximal clique is a clique such that including any other node in $V \backslash C$ will not be a clique.

Consider the example graph in Figure 1. There are 5 nodes and three cliques $C_1 = \{1, 2, 3\}$, $C_2 = \{3, 4\}$ and $C_3 = \{4, 5\}$. The clique C_2 is a cutset because removing C_2 will result in two components. Equivalently, we see that all paths from clique C_1 to C_3 go through C_2.

For an undirected graph G, we associate with each node $i \in V$ a random variable X_i over the set $\{0, 1\}$ or $\{-1, 1\}$. For any subset $A \subset V$ of nodes we define a configuration $x_A = \{x_i : i \in A\}$. Two variables X_i and X_j are independent if $\mathbb{P}(X_i, X_j) = \mathbb{P}(X_i)\mathbb{P}(X_j)$, and we write this as $X_i \perp\!\!\!\perp X_j$. The variables X_i and X_j are conditionally independent on X_k if $\mathbb{P}(X_i, X_j \mid X_k) = \mathbb{P}(X_i \mid X_k)\mathbb{P}(X_j \mid X_k)$. For subsets of nodes A, B, and W, we denote by $X_A \perp\!\!\!\perp X_B \mid X_W$ that X_A is conditionally independent of X_B given X_W.

A random vector X is *Markov with respect to* G if

$$W \text{ is a cutset for disjoint subsets } A \text{ and } B \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_W \tag{17}$$

An equivalent way to define a Markov random field is in terms of the factorisation over cliques of the distribution function, which reveals conditional independence between cliques. For each clique C in the set of all cliques $\mathcal{C}$ of graph G a compatibility function $\psi_C : \{0, 1\}^n \to \mathbb{R}_+$ maps the states of the nodes in clique C to the positive reals. When normalized, the product of the compatibility functions defines the distribution. The distribution of the random vector X *factorises according to graph* G if it can be represented by a product of compatibility functions of the cliques

$$\mathbb{P}(X = x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{18}$$

where Z is the normalisation constant. Here we use the functions $\psi_C(x_C) = \exp(f(x_C))$ for exponential family models. For strictly positive distributions the Hammersly-Clifford theorem says that the Markov and factorisation properties are equivalent (Cowell et al., 1999; Lauritzen, 1996).

Considering Figure 1(a) again, we see that the factorisation over cliques is

$$\psi_{C_1}(x_1, x_2, x_3)\psi_{C_2}(x_3, x_4)\psi_{C_3}(x_4, x_5)$$

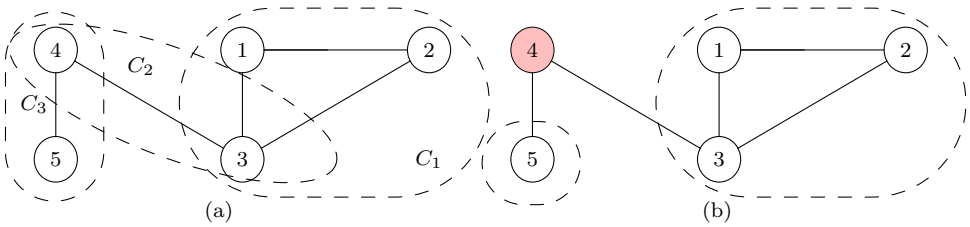


FIG 1. *Graph of 5 nodes with cliques* $C_1 = \{1, 2, 3\}$, $C_2 = \{3, 4\}$ *and* $C_3 = \{4, 5\}$. *In (a) the factorisation property obtained from the three cliques, such that the distribution is defined for potential functions* ψ_{C_1}, ψ_{C_2}, *and* ψ_{C_3}, *as in (18). In (b) conditional independence is shown for* $X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} \mid X_4$.

And, in Figure 1(b) the Markov property implies that $X_5 \perp\!\!\!\perp X_3 \mid X_4$, for instance. Such conditional independence can be checked visually from the paths that go through node 4 to get from 5 to 3.

The Markov and factorisation properties imply that in an undirected graphical model each node is conditionally independent of other nodes given the nodes adjacent to it, except for its adjacent nodes. That is, for any disjoint subsets B and $D \backslash \partial B \subset V$, where the boundary ∂B is excluded, we have the conditional independence $X_B \perp\!\!\!\perp X_{D \backslash \partial B} \mid X_{\partial B}$. So, there is in general conditional independence for nodes when conditioned on their boundary sets.

Here we use the pseudo-likelihood as an approximation to the joint distribution. It has been shown that parameter estimation with the pseudo-likelihood obtains consistent estimators (Hyvärinen, 2006; Nguyen, 2017). A special case that we use here is for pairwise models, specifically the Ising model defined on graph G (Cipra, 1987; Wainwright and Jordan, 2008), or, as it is sometimes called, the auto-logisitc model (Besag, 1974), which has been shown to be consistent (see, e.g., Loh et al., 2013; Yang et al., 2012, 2013; Haslbeck and Waldorp, 2015). Let X be a random vector over $\{0, 1\}^n$ or $\{-1, 1\}^n$. The joint distribution of X is called the *Ising model* if its probability distribution is defined as

$$\mathbb{P}(x) = \frac{1}{Z} \exp\left(\sum_{i \in V} \xi_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right) \tag{19}$$

where

$$Z_V = \sum_{x \in \mathcal{X}^n} \exp\left(\sum_{i \in V} \xi_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right) \tag{20}$$

where $\mathcal{X}$ is either $\{0, 1\}$ or $\{-1, 1\}$. The conditional distribution of X_i, given the nodes adjacent to it in ∂i, is

$$\mathbb{P}(x_i \mid x_{\partial i}) = \frac{\exp\left(x_i \left(\xi_i + \frac{1}{2} \sum_{j \in \partial i} \theta_{ij} x_j \right) \right)}{\exp\left(x_i^c \left(\xi_i + \frac{1}{2} \sum_{j \in \partial i} \theta_{ij} x_j \right) \right) + \exp\left(\xi_i + \frac{1}{2} \sum_{j \in \partial i} \theta_{ij} x_j \right)} \tag{21}$$

where x_i^c is either 0 if $x_i \in \{0,1\}$ or $x_j^c = -1$ if $x_j \in \{-1,1\}$. The parameters $(\xi_j, j \in V)$ are the threshold or external field parameters and $(\theta_{ij}, (i,j) \in E)$ the edge or interaction parameters. We often write simply p_i to denote the conditional probability $\mathbb{P}(X_i = 1 \mid x_{\partial i})$.

4.2. General Fourier analysis

What we need in order to work in a general setting with Fourier analysis of Boolean functions, is that the expectation of a product of variables (using the joint distribution) can be decomposed into a product of expectations (using the conditional distributions). We can use the product measure if we have independence (not connected), since then we have $\mathbb{E}(X_1X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2)$, and this holds for any subset S of independent variables. If we have conditional independence, we have a similar construction, that is, $\mathbb{E}(X_1X_2 \mid X_3) = \mathbb{E}(X_1 \mid X_3)\mathbb{E}(X_2 \mid X_3)$, if $X_1 \perp\!\!\!\perp X_2 \mid X_3$. However, whenever there is an edge among the nodes in the subset S, then we do not have (conditional) independence. We therefore propose to work with the pseudo-likelihood, product of conditional distributions (Besag, 1974; Hyvärinen, 2006)

$$\mathbb{P}_\pi(x) := \mathbb{P}(x_1 \mid x_{\partial 1})\mathbb{P}(x_2 \mid x_{\partial 2}) \cdots \mathbb{P}(x_n \mid x_{\partial n}) \tag{22}$$

We write $\mathbb{P}_\pi$ for the measure in (22) with probabilities $\pi = (p_1, p_2, \ldots, p_n)$ for any subset $S \subseteq V$ that is not a singleton set or the empty set. In the Appendix we show that the joint distribution is close to the pseudo likelihood in terms of the Kullbeck-Leibler divergence and give an example.

Using the product of conditionals allows us to obtain the requirements of orthogonality in (10) for the Fourier analysis: The expectation of the basis function needs to be 0 and the expectation of the square needs to be 1. We therefore define the function $\phi : \{-1,1\} \to \mathbb{R}$ by

$$\phi(x_i) = \frac{x_i - \mu_i}{\sigma_i} \tag{23}$$

where we use the conditional expectation

$$\mu_i = \mathbb{E}(X_i \mid x_{\partial i}) = 2p_i - 1 \tag{24}$$

and conditional variance

$$\sigma_i^2 = \mathbb{E}((X_i - \mu_i)^2 \mid x_{\partial i}) = 4p_i(1-p_i) \tag{25}$$

Note that $\phi(1) = \sqrt{(1-p_i)/p_i}$ and $\phi(-1) = -\sqrt{p_i/(1-p_i)}$. We write $\mathbb{E}_{p_i}$ for the conditional expectation $\mathbb{E}(\cdot \mid x_{\partial i})$; we use p_i to refer to conditional statements. For each $i \in V$ we then have that the mean is $\mathbb{E}_{p_i}(\phi(X_i)) = 0$ and the variance is $\mathrm{var}^{p_i}(\phi(X_i)) = 1$. For the product function $\phi_V : \{-1,1\}^n \to \mathbb{R}$ defined by $\phi_V(x) = \prod_{i \in V} \phi(x_i)$, we obtain with $\mathbb{P}_\pi$

$$\mathbb{E}_\pi(\phi_V(X)) := \mathbb{E}_{p_1}(\phi(X_1))\mathbb{E}_{p_2}(\phi(X_2)) \cdots \mathbb{E}_{p_n}(\phi(X_n)) \tag{26}$$

This again results in orthogonality as before in (10). Let S, T be any subsets of V, then

$$\langle \phi_S, \phi_T \rangle = \mathbb{E}_\pi \left(\prod_{i \in S} \phi(X_i) \prod_{i \in T} \phi(X_i) \right) = \begin{cases} 1 & \text{if } S = T \\ 0 & \text{if } S \neq T \end{cases} \tag{27}$$

where we used the mean (24) and variance (25). With this p-biased basis function the p-biased Fourier expansion is

$$f(x) = \sum_{S \subseteq V} \hat{f}^\pi(S) \phi_S(x) \tag{28}$$

where the Fourier coefficient is

$$\hat{f}^\pi(S) = \mathbb{E}_\pi(f(X)\phi_S(X)) \tag{29}$$

and $\pi = (p_1, p_2, \ldots, , p_n)$. To obtain the p-biased Fourier expansion we need only plug-in the transformed variables $x_i = \mu_i + \sigma_i \phi(x_i)$.

As an example, consider again the majority function on three items $\text{maj}_3 : \{-1, 1\}^3 \to \{-1, 1\}$ with Fourier expansion under the uniform measure given in (7). Plugging in the transformed values for x_i gives a p-biased Fourier expansion with Fourier coeffcients for the basis functions ϕ_S with $S \subseteq V$

$$\begin{array}{llll} \widehat{\text{maj}}_3^\pi(\varnothing) = & \frac{1}{2}(\mu_1 + \mu_2 + \mu_3 - \mu_1\mu_2\mu_3) & \widehat{\text{maj}}_3^\pi(\{1,2\}) = & -\frac{1}{2}\mu_3\sigma_1\sigma_2 \\ \widehat{\text{maj}}_3^\pi(\{1\}) = & \frac{1}{2}\sigma_1(1 - \mu_2\mu_3) & \widehat{\text{maj}}_3^\pi(\{1,3\}) = & -\frac{1}{2}\mu_2\sigma_1\sigma_3 \\ \widehat{\text{maj}}_3^\pi(\{2\}) = & \frac{1}{2}\sigma_2(1 - \mu_1\mu_3) & \widehat{\text{maj}}_3^\pi(\{2,3\}) = & -\frac{1}{2}\mu_1\sigma_2\sigma_3 \\ \widehat{\text{maj}}_3^\pi(\{3\}) = & \frac{1}{2}\sigma_3(1 - \mu_1\mu_2) & \widehat{\text{maj}}_3^\pi(\{1,2,3\}) = & -\frac{1}{2}\sigma_1\sigma_2\sigma_3 \end{array} \tag{30}$$

where $\pi = (p_1, p_2, p_3)$ is used to indicate that they are p-biased coefficients. Note that under the uniform measure the coefficients for products with two variables $x_i x_j$ were all 0, whereas in the transformed version there are non-zero coefficients for $\phi(x_i)\phi(x_j)$.

The mean and variance of f can be determined analogously to the unbiased case. Note that

$$\mathbb{E}_\pi(f(X)) = \mathbb{E}_\pi \left(\sum_{S \subseteq V} \hat{f}^\pi(S) \phi_S(x) \right) = \hat{f}^\pi(\varnothing) \tag{31}$$

since $\mathbb{E}_\pi(\phi_S(x)) = 1$ only for the set $S = \varnothing$ and 0 otherwise. So

$$\text{var}^\pi(f) = \mathbb{E}_\pi(f(X)^2) - (\mathbb{E}_\pi f(X))^2 = \sum_{S \neq \varnothing} \hat{f}^\pi(S)^2 \tag{32}$$

which is similar to the unbiased case (15) except that we plug-in the biased Fourier coefficients $\hat{f}^\pi$ (see Lemma 9 in the Appendix). This version of Parseval's theorem implies that the variance of a Boolean function with outcome in $\{-1, 1\}$ and expectation 0 is 1.

5. Influence of items

One of the properties of a decision function is that no single item should dominate the decision. In the social theory context the extreme situation is where a decision is determined by a single item, which is referred to as a dictator function. Formally, a dictator function $f : \{-1, 1\}^n \to \{-1, 1\}$ is such that $f(x) = x_i$ for any x. We would want each of the items to have similar influence on the decision. To determine the influence of the items, we need a particular procedure to flip a value from -1 to 1 or vice versa, so that we can determine whether the decision has changed. That is we need to know for two versions x and y whether $f(x) \neq f(y)$. We start from x, the original, and then for coordinate i we either flip the value or not with probability $\frac{1}{2}$. We denote by x^i the vector x where the ith coordinate has been flipped randomly and independently, so that coordinate i of x^i is

$$x_i^i = \begin{cases} x_i & \text{with probability } \frac{1}{2} \\ -x_i & \text{with probability } \frac{1}{2} \end{cases} \tag{33}$$

and all other coordinates remain the same.

We are interested in determining the coordinate i for which the decision of the Boolean function f is sensitive, that is, when will we have $f(x) \neq f(x^i)$. This is captured by the probability that the decisions are unequal, called the influence. Let x^i be given as in (33) and $f : \{-1, 1\}^n \to \{-1, 1\}$ a Boolean function. Then the *influence of i on f* is

$$\mathbb{I}_i^\pi(f) = \mathbb{P}_\pi(f(X) \neq f(X^i)) \tag{34}$$

The influence is precisely the tool we need to investigate the effect a particular item has on the decision. If the influence is disproportionately large, then this should raise a flag. We can find a threshold value using the Fourier coefficients, which are relatively simple for monotone functions, i.e., where $x \leq y$ coordinatewise implies $f(x) \leq f(y)$. To get the probability for $f(x) \neq f(x^i)$ we obtain an indicator function $\mathbb{1}\{f(x) \neq f(x^i)\}$, and its expectation will give the probability in (34). Such an indicator can be obtained from the discrete derivative function

$$D_i f = \frac{1}{2}(f(x^{(i,1)}) - f(x^{(i,-1)})) \tag{35}$$

where $x^{(i,a)}$ means $(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n)$. Then we see that if the decision changes, then $D_i f$ is -1 or 1, and if the decision remains the same, then $D_i f$ is 0. This means that $(D_i f)^2$ is the indicator function for whenever item i results in $f(x) \neq f(x^i)$. And so

$$\mathbb{E}_\pi(\mathbb{1}\{f(X) \neq f(X^i)\}) = \mathbb{E}_\pi((D_i f(X))^2)$$

The discrete derivative of the function ϕ_S in terms of the Fourier coefficient is

$$D_i \phi_S(x) = \frac{1}{2\sigma_i}(1 - \mu_i - (-1 - \mu_i)) \prod_{j \in S \setminus \{i\}} \phi(x_j) = \frac{1}{\sigma_i} \prod_{j \in S \setminus \{i\}} \phi(x_j)$$

if $i \in S$ and 0 otherwise. And so, since $\mathbb{E}_\pi(\phi_{S\setminus\{i\}}(x)^2) = 1$ for each S,

$$\mathbb{I}_i^\pi(f) = \mathbb{E}_\pi((D_i f(X))^2) = \frac{1}{\sigma_i^2}\sum_{S\ni i}\hat{f}^\pi(S)^2$$

where $S \ni i$ means the sum across all subsets S that contain i. For monotone Boolean functions like maj_n, the influence can be simplified to the Fourier coefficient of the singleton set, that is (see Lemma 11 in the Appendix)

$$\mathbb{I}_i^\pi(f) = \frac{1}{\sigma_i}\hat{f}^\pi(i) \tag{36}$$

where $\hat{f}^\pi(i) = \hat{f}^\pi(\{i\}) = \mathbb{E}_\pi(f(X)\phi(X_i))$. Using this relation, we can easily obtain an upper bound on the influence of an item. We have from (32) that the variance of a Boolean function can be represented as $\text{var}^\pi(f) = \sum_{S\neq\varnothing}\hat{f}^\pi(S)^2$. Clearly this is larger than taking only the subsets $S = \{i\}$, and so, $\text{var}^\pi(f) \geq \sum_{i\in V}\hat{f}^\pi(i)^2$. Assuming transitive-symmetry (e), we know that nodes are exchangeable, and so $\hat{f}^\pi(i) = \hat{f}^\pi(j)$ for any nodes i and j. And so, if we assume that f is monotone, we have that

$$\text{var}^\pi(f) \geq \sum_{i\in V}\hat{f}^\pi(i)^2 = n\hat{f}^\pi(i)^2$$

Hence, we obtain an upper bound for the influence of item i

$$\mathbb{I}_i^\pi(f) \leq \frac{\text{sd}^\pi(f)}{\sigma_i\sqrt{n}} \tag{37}$$

We could use this bound to see whether any of the items come close to this bound, suggesting that its influence maybe too large compared to the other items. Ideally, you would want the items each to have similar influence, so that the test is balanced (see Section 7).

6. Reliability and stability of decisions

The notion of influence can be extended to all items. We set up a procedure to create a new vector Y such that independently we flip the value x_i with probability $\frac{1}{2}(1-\rho)$. We then consider for the contaminated data Y the effect on the decision, i.e., whether $f(x) = f(y)$. We fix the probability of obtaining the original score x_i when considering Y_i, which contains measurement error. To fix the measurement error in the same way in each of the y_i for $i = 1, \ldots, n$, we iterate the experiment independently, so that for any $\rho \in [-1, 1]$

$$Y_i = \begin{cases} x_i & \text{with probability } \frac{1}{2}(1+\rho) \\ -x_i & \text{with probability } \frac{1}{2}(1-\rho) \end{cases} \tag{38}$$

where $x_i \in \{-1, 1\}$. So, in expectation, $n\frac{1}{2}(1+\rho)$ of the Y_i are the same as x_i and $n\frac{1}{2}(1-\rho)$ are the opposite of x_i. This is the same as the experiment where

Y_i is uniform over $\{-1, 1\}$ with probability $1-\rho$, or Y_i stays x_i with probability ρ and ρ is then in the interval $[0, 1]$. This explains why we can consider this form of measurement error an extension of the error defined previously in (33) for the influence of a single item.

The expectation of the random variable with measurement error Y_i conditioned on the value x_i is now

$$\mathbb{E}(Y_i \mid x_i) = x_i \frac{1}{2}(1+\rho) + (-x_i)\frac{1}{2}(1-\rho) = \rho x_i \tag{39}$$

where we take the expectation with respect to the measure $\mathbb{P}_\rho(Y_i = x_i \mid x_i) = \frac{1}{2}(1+\rho)$ from (38) with $\rho \in [-1, 1]$, which we indicate by $\mathbb{E}_\rho$. We use the notation $\mathbb{E}_{p_i,\rho}$ to indicate the expectation first with respect to $\mathbb{P}_\rho$ and then with respect to $\mathbb{P}_{p_i}$. The implication of (39) is that the true score in the operational view of Lord and Novick (1968, Chap. 2) is

$$\mu_\rho(x_i) = \mathbb{E}(Y_i \mid x_i) = \rho x_i = \begin{cases} -\rho & \text{if } x_i = -1 \\ \rho & \text{if } x_i = 1 \end{cases} \tag{40}$$

We are then able to show that the assumptions to obtain the true score are obtained with the process of (38). Let the error be $Y_i - \mu_\rho(x_i)$, then we see that

$$\mathbb{E}_\rho(Y_i - \mu_\rho(x_i)) = x_i \frac{1}{2}(1+\rho) - x_i \frac{1}{2}(1-\rho) - \rho x_i = 0 \tag{41}$$

and so, the correlation between the error and the true score is 0 because

$$\mathrm{cov}^{p_i,\rho}(Y_i - \mu_\rho(x_i), \rho X_i) = \rho \mathbb{E}_{p_i,\rho}(Y_i X_i) - \rho^2 \mathbb{E}_{p_i}(X_i^2) = 0 \tag{42}$$

since $\mathbb{E}_{p_i,\rho}(Y_i X_i) = \rho \mathbb{E}_{p_i}(X_i^2)$. This leads to the criteria to obtain the true score from replications of 'contaminated' data.

(i) $\mathbb{E}_\rho(Y_i - \mu_\rho(x_i)) = 0$
(ii) $\mathrm{cor}^{p_i,\rho}(Y_i - \mu_\rho(x_i), \rho X_i \mid x_i) = 0$

As in Lord and Novick (1968), by taking the expectation of the observed score, we obtain the true score $\mu_\rho(x_i)$ because the error is 0 in expectation for each value of x_i separately. Using a law of large numbers argument, this can be achieved empirically with reasonable accuracy. Also, reliability of parallel tests $\mathrm{cor}^{p_i\rho}(Y_i', Y_i)^2$ can be obtained by additionally requiring that errors from the parallel tests are uncorrelated.

Now that we set up the experiment to obtain 'contaminated' data Y_i, we move on to its effect on decisions. We are interested in what the result is of measurement error defined in (38) on the decision, i.e., will the decision remain the same upon changing a proportion $\frac{1}{2}(1-\rho)$ of the original score x_i. We consider the effects of contaminated data on the decisions in two ways: (i) *decision reliability*, defined as the correlation $\mathrm{cor}^{\pi,\rho}(f(X), f(Y))$, similar to reliability, and (ii) *stability*, defined as the difference in probability that the decisions $f(X)$

and $f(Y)$ are the same and that they are not the same. We will see that decision reliability and stability are closely related.

We start with decision reliability. We require the covariance and variances to obtain the correlation. We use Fourier analysis of Boolean functions to obtain these. Using the Fourier representation of $f(y)$ requires orthogonality as in (27), and so we define

$$\phi^\rho(Y_i) = \frac{Y_i - \rho\mu_i}{\sqrt{1-\rho^2\mu_i^2}} \tag{43}$$

such that $\mathbb{E}_{p_i,\rho}(\phi^\rho(Y_i)) = 0$ and $\mathbb{E}_{p_i,\rho}(\phi^\rho(Y_i)^2) = 1$. We denote by $\sigma_i^\rho = \sqrt{1-\rho^2\mu_i^2}$ the standard deviation of Y_i. Then we can easily obtain the reliability. If we use the transformation ϕ in (23) and ϕ^ρ in (43) then we have that $\text{cor}^{p_i,\rho}(X_i, Y_i) = \mathbb{E}_{p_i,\rho}(\phi(X_i)\phi^\rho(Y_i))$, and so

$$\text{cor}^{p_i,\rho}(X_i, Y_i) = \mathbb{E}_{p_i,\rho}\left(\frac{X_i-\mu_i}{\sigma_i}\right)\left(\frac{Y_i-\rho\mu_i}{\sigma_i^\rho}\right) = \rho\frac{\sigma_i}{\sigma_i^\rho} \tag{44}$$

So, the reliability is proportional to the 'raw' correlation $\mathbb{E}_{p_i,\rho}(X_iY_i) = \rho$. We can now determine the reliability (correlation) of the decision function f with input X and Y, i.e., $\text{cor}^{\pi,\rho}(f(X), f(Y))$. In the Appendix (see Lemma 13) we show that the covariance between $f(X)$ and $f(Y)$ is

$$\text{cov}^{\pi,\rho}(f(X), f(Y)) = \sum_{S\neq\varnothing} \omega(S)\rho^{|S|}\hat{f}^\pi(S)\hat{f}^{\pi,\rho}(S) \tag{45}$$

where $\omega(S) = \prod_{i\in S}\frac{\sigma_i}{\sigma_i^\rho}$. The variance is obtained similarly (also in Lemma 13 in the Appendix) and so the correlation is

$$\text{cor}^{\pi,\rho}(f(X), f(Y)) = \frac{\sum_{S\neq\varnothing}\omega(S)\rho^{|S|}\hat{f}^\pi(S)\hat{f}^{\pi,\rho}(S)}{\sqrt{1-\hat{f}^\pi(\varnothing)}\sqrt{1-\hat{f}^{\pi,\rho}(\varnothing)}} \tag{46}$$

If we suppose that the test is unbiased so that $\mathbb{P}_\pi(f(X)=1) = \mathbb{P}_\pi(f(X)=-1)$, then $\hat{f}^\pi(\varnothing) = \hat{f}^{\pi,\rho}(\varnothing) = 0$ then we see from (46) that only the covariance matters. We then see that the reliability of the decision function is mostly determined by subsets of order $|S| = 1$, because of the factor $\rho^{|S|}$. We can therefore use an approximation with only the singleton sets $S = \{i\}$ to use for the correlation.

Proposition 1 (Approximate decision reliability) Let $f : \{-1,1\}^n \to \{-1,1\}$ be unbiased and Y be an n vector defined in (38) such that the reliability is $\text{cor}^{p_i\rho}(X_i, Y_i) = \omega(i)\rho$ for all i. Then the decision reliability is

$$\text{cor}^{\pi,\rho}(f(X), f(Y)) = \rho\sum_{i\in V}\omega(i)\hat{f}^\pi(i)\hat{f}^{\pi,\rho}(i) + O(\rho^2) \tag{47}$$

The proof follows immediately from Lemma 13 in the Appendix and the result in (46). This approximation is useful in practice since we only require estimation of singleton Fourier coefficients.

The second way to consider the impact of measurement error on the decision is to compare the probabilities of $f(X) = f(Y)$ and $f(X) \neq f(Y)$, where Y is the contaminated score. We call this difference stability. Let Y be a vector defined as in (38) and let $f : \{-1, 1\}^n \to \{-1, 1\}$. Then the *stability of* f *at* ρ is

$$\mathbb{S}^{\pi}_{\rho}(f) = \mathbb{P}_{\pi,\rho}(f(X) = f(Y)) - \mathbb{P}_{\pi,\rho}(f(X) \neq f(Y)) \tag{48}$$

Because $f(X)f(Y)$ is 1 whenever $f(X) = f(Y)$, we have that stability equals $\mathbb{E}_{\pi,\rho}(f(X)f(Y))$. We then obtain the Fourier representation of stability immediately from Lemma 13 in the Appendix. Considering the covariance in terms of Fourier coefficients in (45), stability is the covariance of $f(X)$ and $f(Y)$ plus the square of the mean $\mathbb{E}_{\pi}(f(X)) = \hat{f}^{\pi}(\varnothing)$, i.e.,

$$\mathbb{S}^{\pi}_{\rho}(f) = \mathrm{cov}^{\pi,\rho}(f(X), f(Y)) + \mathbb{E}_{\pi}(f(X))\mathbb{E}_{\pi,\rho}(f(Y)) \tag{49}$$

It is often easier to work with the covariance than the stability, so we will use stability in only a few cases.

7. The merit of the sum score

We have used linear threshold functions because they are monotone and are easy to interpret. But, in general there are other arguments to use threshold functions. The first argument is that a linear threshold function satisfies the desirable properties (a)-(d) from Section 2, which are monotonicity, oddness, unanimity and anonymity. Although these are, we believe, necessary to use a decision function, it is by no means clear that satisfying these criteria is a sufficient condition. This is because although the properties (a)-(d) are desirable, they do not show all implications of a decision function. In this section we will elaborate on three types of arguments that seem more convincing to us to use linear threshold functions (LTF).

(1) Using an LTF is connected to the true score in test theory;
(2) an LTF is stable, i.e., insensitive to noise (property (f) in Section 2); and
(3) if a test is balanced, then an LTF matches the input best.

We shall discuss each argument in turn.

7.1. The true score in test theory

In the view of test theory a (weighted) sum score is used to obtain an approximation to the true score. In classical test theory the true score is defined in terms of subpopulations M_i such that all elements in M_i have value x_i. Then the true score is $\mathbb{E}(X_i \mid M_i)$ (Lord and Novick, 1968; Ellis and Junker, 1997). So, for any subpopulation M_i, the (weighted) sum score of the X_i will provide

an approximation to the true score. This idea is generalised in item response theory. The role of the subpopulation is replaced by the latent variable (Junker and Ellis, 1997). In many exponential family distributions, the sum score is a sufficient statistic, implying that to obtain the parameter of interest, no more information is needed than the sum score (Lord and Novick, 1968; Brown, 1986). Under the following assumptions

(i) the latent variable Θ is *unidimensional*
(ii) *local independence*: for any disjoint subsets $S, T \subset V$, $X_S \perp\!\!\!\perp X_T \mid \Theta$
(iii) *latent monotonicity*: $\mathbb{P}_\pi(X_i = 1 \mid \theta) \geq \mathbb{P}_\pi(X_i = 1 \mid \theta')$ if $\theta \geq \theta'$

the true score can be defined as $\mathbb{E}(X_i \mid \theta)$ (e.g., Sijtsma and Molenaar, 1987). These assumptions are not arbitrary, the assumptions (i), (ii) and (iii) together imply conditional association, i.e., for any two-part partition of V, conditional on any function on one part of the test, two monotone functions on the other part of the test are positively related (Holland and Rosenbaum, 1986). Conditional association is therefore an observable quantity, to be verified in applications. Hence, the assumptions above imply restrictions on what we may observe and so can be related to (experimental) data.

Ellis and Junker (1997) show that the latent variable can be defined as a class of events obtained from observed scores that contains all infinite events as elements (tail sigma-algebra of items). Such a conceptualisation is relevant because it implies that the requirements of local independence, unidimensionality and monotonicity can be characterised in terms of observables (the authors call it 'asymptotically empirical'). In that setting the true score is defined as $\mathbb{E}(X_i \mid \tau(X))$, where $\tau(X)$ is the tail sigma algebra. The tail sigma algebra τ is the intersection of sigma algebras $\cap_{i \geq 1} \sigma(X_i, X_{i+1}, \ldots)$; τ contains events like obtaining infinitely often heads on tosses 2, 4, 8, etc (Rosenthal, 2013; Durrett, 2010).

Here we introduce our view on the use of the sum score in defining the true score in terms of the Ising model and show the relations with conditional association and hence with unidimensionality, local independence, and monotonicity. We conceive of an item as being a node in a graph, connected to other nodes. The nodes that are connected to item i contain all relevant information about the item. This can be seen from the Ising model, where the probability of each node is determined by a threshold parameter and the connections to the other nodes. In fact, we have that the conditional probability of obtaining the answer $X_i = 1$ is

$$\mathbb{P}(X_i = 1 \mid x_{\partial i}) = \frac{1}{Z_i(x_{\partial i})} \exp\left(\xi_i + \sum_{j \in \partial i} \theta_{ij} x_j\right)$$

where ∂i is the boundary set of nodes that are connected to node i. The probability of node i is therefore completely determined by the the parameter ξ_i (threshold) and the sum of connected nodes and their connectivity weights $\sum_{j \in \partial i} \theta_{ij} x_j$. It therefore follows naturally that the true score can be defined

by the value of the linear threshold function $\ell_{\beta_i} = \xi_i + \sum_{j\in\partial i}\theta_{ij}x_j$, where β_i is the parameter vector $(\xi_i, \theta_{ij} : j \in \partial i)$, that is,

$$\mathbb{E}(X_i \mid \ell_{\beta_i}) \tag{50}$$

The true score defined in (40) concerns a process on top of the one considered here. In (40) we defined the true score with respect to an experiment where each item could be flipped with probability $\frac{1}{2}(1-\rho)$. It followed that the true score was ρx_i, irrespective of how the probability came about. The true score in (50) is also in line with the ideas of Lord and Novick (1968) and Ellis and Junker (1997), where the true score is defined by the expectation of the variable, conditioning on the infinite sequence of variables (or tail sigma-algebra) here the neighbourhood). We explicitly use the Ising model as a basis for the definition of the true score.

The definition of the true score in (50) leads to several consequences for monotonicity, local independence and unidimensionality, as we discuss next.

We start with monotonicity. The sum $\ell_{\beta_i} = \xi_i + \sum_{j\in\partial i}\theta_{ij}x_j$ is in fact a linear threshold function (LTF)

$$\ell_{\beta_i}(x) = \xi_i + \theta_{i1}x_1 + \cdots + \theta_{ik}x_k = a_0 + a_1x_1 + \cdots + a_kx_k$$

where the sum is over the nodes in the boundary set ∂i with size $|\partial i| = k$, and node $i \notin \partial i$. Because the probability for the Ising model is completely determined by ℓ_{β_i}, we can trade the latent variable or tail sigma-algebra for ℓ_{β_i}. The function ℓ_{β_i} can be characterised as the possible states of 'energy' given a fixed β_i, as is done in the statistical physics literature (see e.g., Emch and Liu, 2013). The conditional probability p_i only changes if the value of ℓ_{β_i} changes. Hence, given β_i, changes in the response pattern x only lead to a different probability if ℓ_{β_i} is different. It is possible that the response pattern is different but that ℓ_{β_i} remains the same. For example, suppose that $\theta_{ij} = 1$ for any $i, j \in V$, then two coordinates x_i and x_j could switch value and ℓ_{β_i} will remain the same (the function ℓ_{β_i} is not injective). So, the function ℓ_{β_i} defines equivalence sets with respect to response patterns. We write $[x]$ for the equivalence class such that $\ell_{\beta_i}(x) = \ell_{\beta_i}(z)$ whenever $z \in [x]$. We must therefore define monotonicity with respect to the equivalence classes $[x]$ determined by ℓ_{β_i}. And so we have monotonicity with respect to ℓ_{β_i} in the sense that

$$\mathbb{P}(X_i = 1 \mid \ell_{\beta_i}) \geq \mathbb{P}(X_i = 1 \mid \ell'_{\beta_i}) \quad \text{if} \quad \ell_{\beta_i} \geq \ell'_{\beta_i} \tag{51}$$

or equivalently

$$\mathbb{P}(X_i = 1 \mid [x]) \geq \mathbb{P}(X_i = 1 \mid [x']) \quad \text{if} \quad x \geq x' \text{ for any } x \in [x] \text{ and } x' \notin [x]$$

Although we obtain monotonicity, with the Ising model there is ambiguity in the 'usefulness' of ℓ_{β_i}. Junker (1993) suggests reasonably that a latent variable is 'useful' when it is identified (injective probability function with respect to θ), is monotone as in (ii) and has at least two distinct values to categorise elements of the population. The problem is that monotonicity can be obtained in multiple ways (is not injective). That is, the order $\ell_{\beta_i} \geq \ell'_{\beta_i}$ can come about because

- $\beta_i = \beta_i'$ but $\sum_j x_{ij} \geq \sum_j x'_{ij}$; or
- $\sum_j x_{ij} = \sum_j x'_{ij}$ but $\beta_i \geq \beta_i'$ (elementwise); or
- a combination of the above.

This makes the ordering in general more difficult. On the other hand, if for a given set of elements of the population the parameters are obtained (the Ising model is identified), then we can 'pinpoint' the origin of the ordering. That is, we can compare subjects on their score and parameters and thereby identify why one has a higher probability of obtaining a correct answer than the other; we are able to determine which parameters are relevant to determine the probability in addition to the set of correct items.

We also obtain a form of local independence because of the Markov property (17). Let $\partial\{i,j\} = \partial i \cup \partial j$, then items i and j are conditionally independent given $\partial\{i,j\}$

$$\mathbb{P}(X_i, X_j \mid \partial\{i,j\}) = \mathbb{P}(X_i \mid \partial\{i,j\})\mathbb{P}(X_j \mid \partial\{i,j\}) \tag{52}$$

for any items i and j such that $i \notin \partial j$ nor $j \notin \partial i$. We call this local conditional-independence.

In latent variable modelling, unidimensionality, together with local independence and monotonicity of the latent variable, restricts the marginal distribution $\mathbb{P}(x)$ to be conditionally associative (Holland and Rosenbaum, 1986). Conditional association is an important property because it turns the latent variable assumptions (i)-(iii) above into observable quantities (see the equivalence in Ellis and Junker, 1997, Theorem 5; it is actually almost observable because vanishing conditional dependence is also required). Conditional association is defined as follows: For any partition K and L of the nodes V of the graph G, the covariance between monotone functions f and g is ≥ 0, conditional on $h(x_L) = c$ (Junker, 1993). By using the Fourier expansion in (28) we obtain conditional association if the sum of the product of Fourier coefficients of f and g and variances of all items is positive. More generally, we could assume that the Fourier coefficients of f and g agree for the most part.

Proposition 2 Let X be the random variables induced by the Ising probability in (19) with respect to G. Then X is conditionally associated with respect to $\mathbb{P}_\pi$ if the Fourier coefficients of the monotone functions f and g agree in sign for most of the variables. In particular, if f and g are LTF, then the coefficients a_i of f and b_i of g imply conditional association if $\sum_{i\in K} a_i b_i \sigma_i^2 \geq 0$ for all partitions K and L of V.

For example, if the weights are all $a_i = \frac{1}{n}$ $(i \in V)$, we obtain an equally weighted sum score and obtain conditional association. Or we could choose the weights according to the underlying graph, which maximises the agreement between items and the decision. We discuss this last idea further in Section 7.3.

In summary, when defining the true score in terms of the 'energy' function ℓ_{β_i} we obtain an observable quantity, in contrast to the definition using the latent

variable or tail events. However, there is non-uniqueness in that the value ℓ_{β_i} can be obtained for different response patterns x, and so we are forced to work with equivalence classes $[x]$. Given this limitation, we can use the underlying graph to obtain the weights a_i used in the LTF.

7.2. An LTF is stable

From a practical point of view, the items in a test should have more or less equal influence. There are exceptions, of course, where certain questions are necessary (but not sufficient) to answer correctly in order to pass the test. In view of the graphical perspective of a test, equal degree nodes in the graph (regular graph) implies similar influence of the items.

Lemma 3 Let the graph G be induced by a regular (equal degree nodes) Ising model (19) with equal interaction and threshold parameters. Then the decision function $f : \{-1, 1\}^n \to \{-1, 1\}$ will have equal influences $\mathbb{I}_i^\pi(f) = \mathbb{I}_j^\pi(f)$ for any $i \neq j \in V$.

A consequence of Lemma 3 is that low degree nodes in the graph, nodes with no or few connections to other nodes, have low influence. This is in line with item-test correlations (or regressions) as described in Lord and Novick (1968, Section 3.7), which would yield 0 coefficients if the item were not correlated to other items.

In practice the interaction parameters of the Ising model need not be the same, of course. But suppose you think a well-designed test is one where a latent variable explains all dependency among the items in an equal way. Then, from the argument that the marginal distribution is in that case an Ising model (Marsman et al., 2018), this Ising model will have equal interaction parameters, and so, satisfies the above assumptions. In general, without the assumption of a latent variable, it is reasonable to suppose that in a test no one item has extremely many connections (i.e., high degree), and that the interaction parameters are approximately equal. This is similar to requiring that the item-rest correlations are similar and none of them stand out. We call a test with underlying graph that has approximately equal interaction parameters and similar degrees, a balanced test.

This idea of a balanced test leads to the influences of the items being approximately equal, as we saw in Lemma 3. In such cases it has been shown that for large tests an LTF is the stablest function out of all functions (Mossel, O'Donnell and Oleszkiewicz, 2010). In particular, the stability of the majority function maj_n is

$$\mathbb{S}_\rho^\pi(\text{maj}_n) = \frac{2}{\pi} \arcsin \rho \quad \text{as } n \to \infty$$

The assumptions in Mossel, O'Donnell and Oleszkiewicz (2010) are that the influence of each item is no higher than some small value, and that the expectation

of the Boolean function (here maj_n) is 0. Then it is shown that for any other function, the stability of the majority function is higher, i.e. $\mathbb{S}^\pi_\rho(\text{maj}_n) \geq \mathbb{S}^\pi_\rho(f)$ for any f. The implication is that, with respect to measurement error as defined in (38), an LTF with non-dominant coefficients (for all i, $a_i \leq \tau$, for some $\tau > 0$) is the stablest function among all functions that have small influences. This means that using an LTF decision function to determine a final score, will, when a small percentage of the items have been flipped (measurement error), not immediately lead to a different decision, and there is no other function that will be better in this respect. In view of the relation between stability and the covariance function (49) and Proposition 1, which shows that relation with reliability at the item level, it is interesting to see that this theorem by Mossel, O'Donnell and Oleszkiewicz (2010) shows that the reliability of the sum score as a decision function is highest among all other decision function (given the conditions of the theorem).

7.3. Rousseau's criterion

Another way to see that the LTF is appropriate, is to consider Rousseau's viewpoint in social choice theory. Rousseau's criterion (see, e.g., Schwartzberg, 2008) states that the ideal decision function is one which has the decision in the same direction as most of the items (voters originally). In the p-biased case where each item has probability of success p_i, we can amend Rousseau's original idea and demand that the decision agrees with most of the items made above average $\mu_i = 2p_i - 1$ (and so have positive sign). The agreement between the items being made above average and the decision for a function $f : \{-1, 1\}^n \to \{-1, 1\}$ can be measured by the covariance between the decision $f(X)$ and the value $\phi(X_i)$. For all covariances together, we obtain

$$\mathbb{E}_\pi(f(X)[\phi(X_1) + \phi(X_2) + \cdots + \phi(X_n)]) \leq \mathbb{E}_\pi[|\phi(X_1) + \phi(X_2) + \cdots + \phi(X_n)|]$$

where the inequality arises because $f(x)$ is -1 or 1. We have equality only if $f(x) = \text{sgn}(\phi(x_1) + \cdots + \phi(x_n))$, and so the maximum is achieved for a linear threshold function (LTF). Rewriting $\phi(x_1) + \cdots + \phi(x_n)$ gives the LTF

$$\ell_\phi(x) = -\sum_{i \in V} \frac{\mu_i}{\sigma_i} + \frac{1}{\sigma_1} x_1 + \cdots + \frac{1}{\sigma_n} x_n = a_0 + a_1 x_1 + \cdots + a_n x_n \tag{53}$$

where $a_0 = \sum_{i \in V} \mu_i / \sigma_i$ and $a_i = 1/\sigma_i$ for $i \geq 1$. Now if f is monotone, we see that

$$\mathbb{E}_\pi[f(X)(\phi(X_1) + \phi(X_2) + \cdots + \phi(X_n))] = \sum_{i \in V} \mathbb{E}_\pi(f(X)\phi(X_i)) = \sum_{i \in V} \hat{f}^\pi(i)$$

The last term is the *total influence* for monotone functions

$$\mathbb{I}^\pi(f) = \sum_{i \in V} \mathbb{I}^\pi_i(f) = \sum_{i \in V} \frac{1}{\sigma_i} \hat{f}^\pi(i)$$

And so, we proved that the LTF in (53) maximises the agreement between the decision $f(x)$ and the sum of values above or below the average, expressed in terms of the total influence.

Theorem 4 (Rousseau's criterion) Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be monotone. Then ℓ_ϕ in (53) maximises the unscaled total influence $\sum_{i \in V} \sigma_i \mathbb{I}_i^\pi(f)$. In particular

$$\sum_{i \in V} \sigma_i \mathbb{I}_i^\pi(f) \leq \sum_{i \in V} \sigma_i \mathbb{I}_i^\pi(\ell_\phi)$$

for any monotone function f.

According to Rousseau's criterion, then, the sum of covariances of all items with the decision (influence for monotone functions) is maximised by an LTF. In the case of uniform probability over $\{-1, 1\}$, the expected number of items to agree with the decision $f(x)$ is $\frac{n}{2} + \frac{1}{2}\mathbb{I}(f)$ for monotone functions (see Lemma 12), indicating that one would expect at least half of the items to point in the same direction as the decision, which is a reduced version of the unanimous property (c). From this expected number of items that agree with the decision, we see that we can interpret Rousseau's criterion as having at least half the items in line with the decision and more items will agree with the decision corresponding to the total influence of the items on the decision. Originally, this a formalisation of an argument in favour of democracy.

Rousseau's criterion suggests choosing weights that incorporate the values $a_i = \sigma_i^{-1}$ for $i \geq 1$, where $\sigma_i = 1 - \mu_i^2$ for each item and $a_0 = \sum_{i \in V} \mu_i / \sigma_i$. The criterion does not specify that the weights are exactly those given in the theorem. As an alternative, a version with $b_i \phi(X_i)$, for some $b_i \in \mathbb{R}$ would also work. But any alternative must have the property that its expectation is 0.

8. Applying Fourier analysis

In applications of Fourier analysis of Boolean functions where we have biased variables X_i, we need to obtain estimates of the functions $\phi(x_i)$ for all $i \in V$ to obtain the Fourier coefficients. This requires the probabilities $p_i = \mathbb{P}(X_i = 1)$ for all variables. To obtain these probabilities we use the Ising conditional probabilities (21). This in turn requires knowledge of the parameters of the conditional distribution for each variable $i \in V$. Here, we suggest to first estimate the parameters of the conditional distribution using the lasso and then using these estimates to obtain the Fourier coefficients. The lasso for the conditional distribution in binary data has been shown to lead to consistent estimates (Van de Geer, 2008; Ravikumar, Wainwright and Lafferty, 2010; Bühlmann and van de Geer, 2011). Although violation of the sparsity or multicollinearity assumptions deteriorate accuracy of threshold and edge parameters, predictions using the probabilities should still be accurate (Waldorp, Marsman and Maris, 2019).

8.1. The algorithm

The z-transformation $\phi(x_i) = (x_i - \mu_i)/\sigma_i$ for all $i \in V$ requires the mean (24) and variance (25), which are determined by the probabilities $p_i = \mathbb{P}(X_i = 1)$. Because we assume a graph for the items x_i we can model the probability p_i by the nodes in the neighbourhood ∂i, the nodes that are directly connected to node i. For each node $i \in V$ we estimate the probability p_i that $X_i = 1$ using the Ising model by

$$\hat{p}_i = \frac{\exp(\hat{\xi}_i + \sum_{j\in\partial i} \hat{\theta}_{ij} x_{ij})}{1 + \exp(\hat{\xi}_i + \sum_{j\in\partial i} \hat{\theta}_{ij} x_{ij})} \tag{54}$$

where $\hat{\xi}_i$ and $\hat{\theta}_{ij}$ are estimates of the ξ_i and θ_{ij}, respectively. Note that we only use local information with respect to the graph in that the nodes j in the neighbourhood ∂i are responsible for determining the probability of node i. Obtaining the estimate $\hat{p}_i$ for each node $i \in V$ we can determine $\hat{\mu}_i = 2\hat{p}_i - 1$ and $\hat{\sigma}_i = 1 - \hat{\mu}_i^2$. With these estimates we obtain

$$\hat{\phi}(x_i) = \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} \tag{55}$$

which we can plug in the p-biased Fourier coefficient $\hat{f}^{\hat{\pi}}(S)$ in (29) with $\hat{\pi} = (\hat{p}_1, \ldots, \hat{p}_n)$.

To obtain the estimates $\hat{\xi}_i$ and $\hat{\theta}_{ij}$ for all i and j in V we use the lasso penalty on the conditional distributions (Van de Geer, 2008; Ravikumar, Wainwright and Lafferty, 2010; van Borkulo et al., 2014). The lasso version for logistic regression optimises the pseudo-likelihood (Besag, 1974). Choose node $i \in V$ and let the logit function for $p_i = \mathbb{P}(X_i = 1)$ be

$$g_i(x_t) = \log\left(\frac{p_i}{1 - p_i}\right) = \xi_i + \sum_{j\in\partial i} \theta_{ij} x_{ij,t} \tag{56}$$

for observation x_t with $t \in U$ the observation units. Then the pseudo-likelihood is then

$$\min_{\beta\in\mathbb{R}^n} \frac{1}{|U|} \sum_{t\in U} \Big(-x_i g_i(x_t) + \log(1 + \exp(g_i(x_t))) \Big) + \lambda ||\beta_i||_1 \tag{57}$$

where $\lambda > 0$ is the penalty parameter and $\beta_i = (\xi_i, \theta_{ij}; j \in V \backslash \{i\})$ is the parameter vector for node i and $||\beta_i||_1 = \sum_{j=1}^n |\beta_{i,j}|$ is the ℓ_1 norm. This function is convex and so can be optimised using for instance the coordinate descent algorithm, where each parameter $\beta_{i,j}$ (coordinate) is optimised in turn (Hastie, Tibshirani and Wainwright, 2015; Waldorp, Marsman and Maris, 2019). We estimate the parameters β_i for each node i in turn.

Once we have estimates $\hat{\beta}_i$ for all nodes $i \in V$ we have estimates $\hat{p}_i$ of the probabilities for all nodes and so the transformation $\phi(x_i)$, all based on the

conditional distributions. We can then compute for each subset $S \subseteq V$ the functions

$$\hat{\phi}_S = \prod_{i \in S} \hat{\phi}(x_i) \tag{58}$$

where we plugged in the estimates $\hat{\phi}$ in (27). This leads to the estimates of the Fourier coefficients

$$\hat{f}_U^{\hat{\pi}}(S) = \frac{1}{|U|} \sum_{t \in U} f(x_t) \hat{\phi}_S(x_t) \tag{59}$$

Note that by Proposition 1 we require only the first few coefficients, singleton sets $S = \{i\}$ and duo sets $S = \{i, j\}$ because the effect of the higher order sets is small.

The Fourier coefficients are required to compute the stability and noise sensitivity. The definition of stability in (48) uses both the original values in x and the ones with measurement error y; and we of course only have those with measurement error.

8.2. Statistical guarantees

Here we investigate the rate of convergence of the estimates of the Fourier coefficients we can expect based on estimation of the Ising parameters.

Our algorithm in Section 8.1 involves estimation of the Ising parameters $\beta_i = (\xi_i, \theta_{ij}, j \in V \backslash \{i\})$ with the lasso. The lasso is known to have a difficult distribution (Pötscher and Schneider, 2009) and hence we cannot directly use this to obtain bounds. In order to obtain convergence rates on the estimate, we use results from the so-called desparsified lasso (van de Geer, Bühlmann and Ritov, 2013; Javanmard and Montanari, 2014), where a projection of the residuals is added to 'desparsify' the lasso (make the 0s non-zero again based on the residuals).

The lasso has a soft threshold such that parameter values within the range of the penalty λ of 0 will be set to exactly 0. This implies that the sampling distribution of the lasso estimate has unit mass at these points, which destroys the nice property of the sampling distribution (Pötscher and Schneider, 2009), usually obtained with the central limit theorem. The desparsified lasso projects the residuals from the lasso based on an approximation of the inverse of the second order derivative of the optimisation function in (56). For each β_i in the list of nodewise optimisations, let X_t denote the tth observation of the $n-1$ items (without item i), and let $p_i(X_t)$ denote the conditional probability in (21) with the tth observation for the $n-1$ remaining items plugged in. Furthermore, let the function for the nodewise regression with respect to item i, without the lasso penalty, be

$$\psi(\beta_i) = -x_{it} m_i(x_t) + \log(1 + \exp(m_i(x_t))) \tag{60}$$

where x_{it} is the value from observation t of item i. Then we have the $n \times n$ second order derivative matrix for the n parameters in $\beta = (\xi_i, \theta_{ik}, k \in V \backslash \{i\})$

$$\nabla^2 \psi(\beta_i) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_\pi(p_i(X_j) p_i(-X_j) X_j X_j^\mathsf{T}) \tag{61}$$

We denote this theoretical second order derivative by $\Sigma_i = \nabla^2 \psi(\beta_i)$ and assume that its eigenvalues are > 0 (van de Geer, Bühlmann and Ritov, 2013; Waldorp, Marsman and Maris, 2019). We obtain an estimate $\hat{\Sigma}_i$ by removing the expectation operator in (61). We often have that $\hat{\Sigma}$ is singular, certainly so when $n > m$. Therefore, in general we construct an approximate inverse $\hat{\Theta}_i$ such that the difference $||\hat{\Sigma}_i \hat{\Theta}_i - I_n||_\infty = O_p(\sqrt{\log(n)/m})$, where $||\cdot||_\infty$ is the max norm and $O_p(v)$ means that the random variable V_m is for all ε, $\mathbb{P}(|V_m| \leq K_\varepsilon b_m) > 1 - \varepsilon$ for $m \to \infty$ and some $K_\varepsilon > 0$, i.e., V_m / b_m is bounded in probability by K_ε (see, e.g., van de Geer, Bühlmann and Ritov, 2013; Javanmard and Montanari, 2014). Then we can construct the desparsified lasso by

$$\hat{\beta}_i^{dL} = \hat{\beta}_i + \hat{\Theta}_i \nabla \psi(\beta_i) \tag{62}$$

where $\nabla \psi(\beta_i)$ is the n vector of first order derivatives of ψ with respect to β_i with jth element

$$\nabla_j \psi(\beta_i) = \frac{1}{m} \sum_{t=1}^{m} (-x_{jt} + p_i(x_t)) x_{jt} \tag{63}$$

The matrix $\hat{\Theta}_i$ can be obtained, for instance, by performing nodewise regressions on the remaining (i.e., predictor) items (see, e.g., van de Geer et al., 2014). Then we obtain the approximation for the desparsified lasso

$$\hat{\beta}_i^{dL} = \beta_i + \frac{1}{\sqrt{m}} Z_i + o_p(1) \tag{64}$$

where the n vector Z_i is a normal random variable with mean 0 and variance matrix $\text{var}(Z_i) = \hat{\Theta}_i \Sigma_i \hat{\Theta}_i$ and $o_p(1)$ means a random variable that converges in probability to 0, i.e., for every $\varepsilon > 0$, $\mathbb{P}(|V_m| \leq \varepsilon) > 1 - \varepsilon$ as $m \to \infty$. We assume here that the lasso estimate $\hat{\beta}_i$ is obtained with penalty $\lambda \geq \sqrt{\log(p)/n}$ (van de Geer et al., 2014).

With the representation in (64) we can obtain bounds on the estimation error of the Fourier coefficients. We first plug in the representation (64) in the conditional distribution p_i in (21). For the conditional probabilities we can plug in the representation and obtain

$$p_i(\hat{\beta}_i^{dL}) = \text{logit}(\tilde{x}_t^\mathsf{T} \beta_i + Z_i / \sqrt{m} + o_p(1))$$

where $\tilde{x}_t$ is the tth observation $(1, x_t)$, where the 1 is included for the threshold parameter ξ_i, and the logit(z) function is $1/(1 + \exp(-z))$. We then see that we have in the denominator

$$\exp(-\tilde{x}^\mathsf{T} \beta_i) \exp(Z_i / \sqrt{m} + o_p(1))$$

The second term $\exp(Z_i/\sqrt{m}+o_p(1))$ converges with m to 1 at the rate of $1/\sqrt{m}$. Hence, the conditional probability $p_i(\hat{\beta}_i^{dL})$ converges to the true conditional probability p_i at rate $1/\sqrt{m}$. The transformation $\hat{\phi}$ using $\hat{\beta}_i^{dL}$ with the mean $\mu_i = 2p_i - 1$ and standard deviation $\sigma_i = \sqrt{1-\mu_i^2}$, do not alter the convergence in probability, and hence, we obtain $\hat{\phi}(x_i) = \phi(x_i) + o_p(1)$. And finally, the computation of the Fourier coefficient (of order 1) in (59) gives

$$\hat{f}_m^{\pi}(i) = \frac{1}{m}\sum_{t=1}^{m} f(X_t)\hat{\phi}(X_{it}) = \frac{1}{m}\sum_{t=1}^{m} f(X_t)\phi(X_{it}) + o_p(1)$$

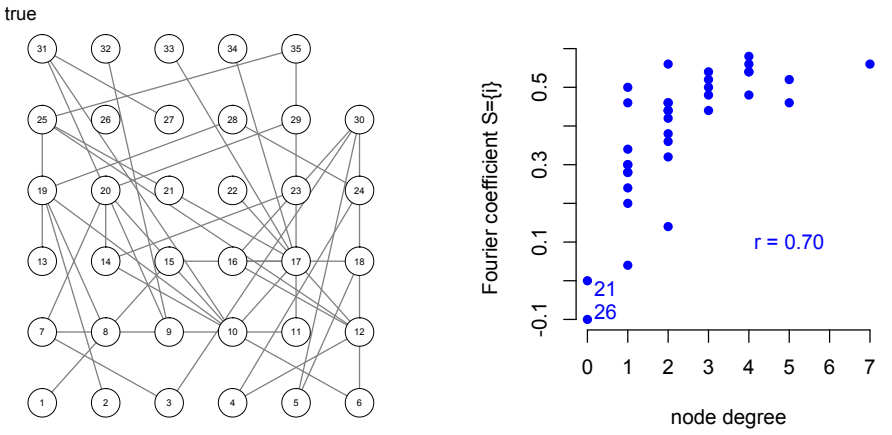
Because the convergence of the average of the Fourier coefficient with the correct transformation with rate $1/\sqrt{m}$, we find that the Fourier coefficient converges at the reasonable rate $1/\sqrt{m}$. This convergence rate is for the coefficients of order 1, for higher order coefficients with ϕ_S where $|S| = k$, say, we obtain slower convergence due to the product in ϕ_S.

9. Numerical illustrations

To show the usefulness and accuracy of a Fourier analysis on test data we consider some numerical illustrations and simulations. In a single case we show what the usefulness is of Fourier analysis of Boolean functions applied in the context of tests. Then we show with simulations what could happen with Fourier analysis in practice when the graph structure is incorrectly recovered, a realistic situation.

We start with a single case. Here we generate 0-1 data according to an Ising model with $n = 35$ items and $m = 100$ observations according to a random graph with probability of an edge 0.05. Each nonzero edge has weight $\theta_{ij} = 3$ and the thresholds are $-\frac{1}{2}\sum_{i=1}^{n}\theta_{ij}$ for each $j = 1, 2, \ldots, 35$. The data are generated with *IsingSampler* (described in van Borkulo et al., 2014) in R (R Development Core Team, 2012). The parameters of the conditional Ising model were estimated by *IsingFIt* (van Borkulo et al., 2014), which estimates the parameters in a nodewise fashion. We used $\gamma = 0.25$ for the extended BIC model selection to obtain the appropriate lasso penalty (only on edge parameters). An example of the graph that generated current data is shown in Figure 2(a). Estimates of the graph parameters are accurate when signal to noise ratios are large as they are here (see Monte Carlo simulations below).

To compute the Fourier coefficients we determined for each $i \in V$ the transformed score $\phi(y_i)$, where $y_i = 2x_i - 1$ to transform them from the domain $\{0, 1\}$ to $\{-1, 1\}$. The conditional Ising probabilities are adjusted accordingly (plugging in the transformed values $x_i = \frac{1}{2}(y_i + 1)$). The decisions for each observation $1, 2, \ldots, m$ were determined with an LTF with $a_0 = -0.6$ (pass at 60% correct) and $a_i = \frac{1}{n}$ (equally weighted) for all i. The Fourier coefficients were then determined using (29) only for singleton sets $S = \{i\}$. The Fourier coefficients can be seen in the scatterplot in Figure 2(b). It is clear that the coefficients correlate highly with the degree of the nodes in the graph. The correlation between the Fourier coefficients and degree of this particular graph is



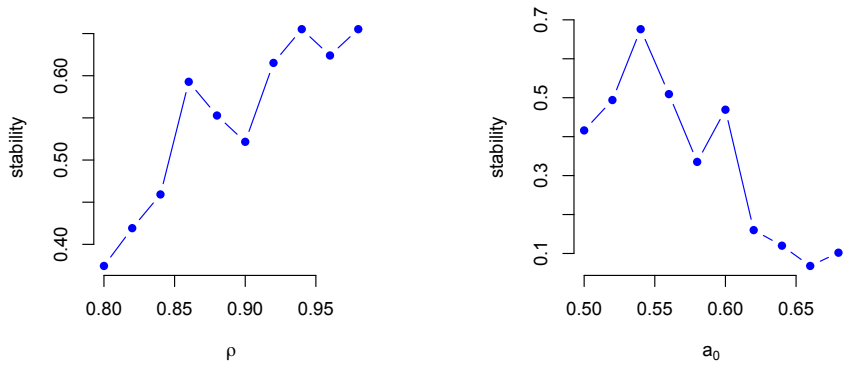
(a) (b)

FIG 2. *Visualisation of the example random graph in (a) and the scatterplot for the degree of the nodes of the graph in (a) and the Fourier coefficients in (b). The correlation is 0.70. Nodes 21 and 26 have the lowest influence and can be seen to be isolated in the network, literally having no influence on the network.*

0.70. It can also be seen that the isolated nodes, nodes 21 and 26, have the lowest influence (which is the Fourier coefficient for singleton sets) of 0.0 and -0.1, respectively. This is exactly what can be expected from theory, since the values of the isolated nodes are independent from the rest, and so will have very little influence on the decision based on the nodes that are connected to each other.

Turning to stability, we are considering whether the rule we chose (LTF with threshold $a_0 = -0.6$, so that 60% is required to be 1) is stable with respect to measurement error. We apply (38) to the generated data and then compute the stability using the Fourier coefficients in (29). Figure 3(a) shows the stability as a function of the correlation ρ between the original values x and the ones with measurement error y. It clearly shows, as expected, that the stability increases as the correlation ρ increases. We also considered the threshold a_0 for several values, shown in Figure 3(b). This indicates that stability decreases with increasing threshold. An explanation is that at high thresholds the values in y never reach it and so the decisions $f(y)$ are nearly constant, resulting in near 0 correlation.

To study the accuracy of the estimates of the Fourier coefficients we use Monte Carlo simulations. The Fourier coefficients depend on the estimates $\hat{p}_i$ of the conditional probabilities that are used for the p-biased parity functions ϕ. For all simulations we used $n = 35$ and $m = 100$ and we varied the signal to noise ratio by increasing the edge weights (β) from 1 to 3. To determine the accuracy of the graph we determined whether we obtained neighbourhoods accurately



(a) (b)

FIG 3. *Stability of the small example data as a function of the correlation ρ between the original and corrupted data (a), and as a function of the threshold a_0 (b).*

and considered the adjacency matrices of the estimated graphs. As dependent measures we used precision (ratio of correctly identified edges to the number of identified edges) and recall (ratio of correctly identified edges to the number of correct edges). We see in Figure 4(a) that the edge parameters are accurately estimated when the edge weights are large ($\beta \geq 2$), but are estimated poorly when the edge weights are lower. Although precision remains high, the number of correctly recovered edges is low for low signal to noise ratios. However, in Figure 4(b) we see that the mean absolute deviations of the Fourier coefficients remain small, regardless of the effect size β. This is because the conditional probabilities are still reasonable at low signal to noise ratios. So, even though the graph itself is not accurately recovered, the predictions for the probabilities are reasonable and hence the Fourier coefficients are accurately estimated. The fact that the graph need not be correctly recovered for the conditional probabilities is that the edge weights and threshold parameters are exchangeable in the Ising model. In Waldorp, Marsman and Maris (2019) we explain in more detail the relation between prediction and graph recovery.

10. Conclusions and discussion

Here we presented another way to think about decisions based on (binary) tests. We focussed on the representation of a test in terms of a graph and showed that the Fourier coefficients can be interpreted as the influence of an item on the decision based on the test. This makes it possible to determine if there are any items that seem inappropriate. We showed that small influence (relative to the other items) is an indication of an 'isolated' item, an item that has few or no connections to other items and so should not be considered as part of the test.

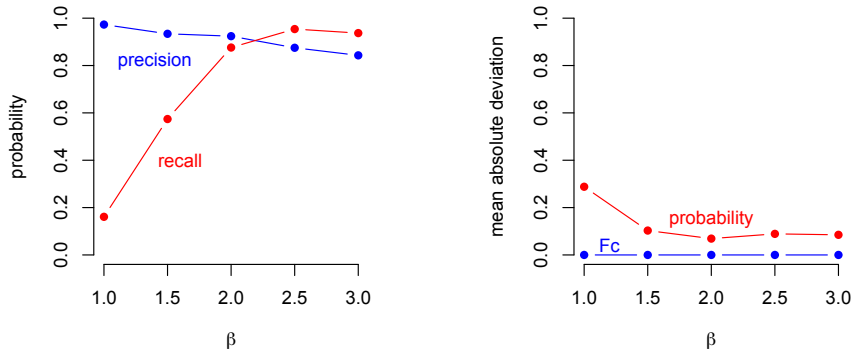


FIG 4. *Recovery of edge parameters in terms of precision (blue) and recall (red) as a function of the edge weights (β) shown in (a). In (b) is the mean absolute deviation of the conditional probabilities (red) and Fourier coefficients (Fc, blue).*

Additionally, it is possible within this framework to consider different decision functions (functions f to decide pass or fail, say) in order to obtain a reliable or stable decision. Decision reliability can be interpreted in terms of the stability of the decision when a proportion $\frac{1}{2}(1-\rho)$ has been flipped. Hence, decision reliability provides information on the stability of the decision with respect to possible changes in the answers to items.

The functions that interest us most are monotone functions, and specifically those that are also unanimous, odd, anonymous, and stable. A function that satisfies these properties is a linear threshold function (LTF), and according to May's theorem, an LTF used for a binary decision is the only function that satisfies these properties. To determine the appropriateness of LTFs we also provided several other arguments. (1) We provided a graphical modeling framework to define the true score based on a weighted sum score of the items directly connected to the item, relating it to monotonicity, local independence and conditional association. (2) A weighted sum score is the most stable decision function in the sense that flipping a small percentage of the items will not immediately change the decision. And (3), the decision function that is most in line with the items is an LTF. These arguments suggest that a weighted sum score is an appropriate basis for binary decisions based on binary items.

Our view on defining the true score with the neighbours in the Ising model suggests interesting extensions. For instance, the probability of making an item correct can be considered a property of an individual, and does not necessarily have to be described with respect to a population of other individuals. This suggests that we can obtain the parameters of the Ising model for each individual separately. In theory this is possible, using time series for instance, assuming

that the process is time Markov. Such an individualistic definition of the true score could lead to mechanistic investigations of what underlies the increase (or decrease) of the probability of making an item correctly.

References

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36** 192-236.

Borsboom, D., Mellenbergh, G. J. and van Heerden, J. (2004). The concept of validity. *Psychological Review* **111** 1061-1071.

Brown, L. D. (1986). *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cipra, B. A. (1987). An introduction to the Ising model. *The American Mathematical Monthly* **94** 937-959.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*, 2nd ed. Wiley and Sons.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer.

De Wolf, R. (2008). A Brief Introduction to Fourier Analysis on the Boolean Cube. *Theory of Computing, Graduate Surveys* **1** 15.

Durrett, R. (2010). *Probability: Theory and examples*. Duxbury Press.

Ellis, J. L. and Junker, B. W. (1997). Tail-measurablility in monotone latent varible models. *Psychometrika* **62** 495-523.

Emch, G. G. and Liu, C. (2013). *The logic of thermostatistical physics*. Springer Science & Business Media.

Goldberg, D., Bridges, K., Duncan-Jones, P. and Grayson, D. (1988). Detecting anxiety and depression in general medical settings. *British Medical Journal* **297** 897–899.

Haslbeck, J. and Waldorp, L. J. (2015). Structure estimation for mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.05677*.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Holland, P. W. and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics* 1523–1543.

Hyvärinen, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation* **18** 2283–2292.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression Technical Report, arXiv:1306.317.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics* **21** 1359–1378.

Junker, B. W. and Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics* **25** 1327-1343.

Kelly, J. S. (1988). *Social choice theory: An introduction*. Springer Science & Business Media.

Kindermann, R., Snell, J. L. et al. (1980). *Markov random fields and their applications* **1**. American Mathematical Society Providence, RI.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Loh, P.-L., Wainwright, M. J. et al. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics* **41** 3022–3049.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Maris, G. and Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika* **77** 615–633.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., Maas, H. v. d. and Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research* **53** 15–35.

May, K. O. (1952). A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica: Journal of the Econometric Society* 680–684.

Mossel, E., O'Donnell, R. and Oleszkiewicz, K. (2010). Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics* **17** 295–341.

Nguyen, H. D. (2017). Near Universal Consistency of the Maximum Pseudolikelihood Estimator for Discrete Models. *Annals of Statistics* **2** 22-23.

O'Donnell, R. (2014). *Analysis of Boolean functions*. Cambridge University Press.

Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* **139** 2775–2790.

Ravikumar, P., Wainwright, M. and Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics* **38** 1287-1319.

Rosnthal, J. S. (2013). *A first look at rigorous probability theory*, 2nd ed. World Scientific Publishing.

Schwartzberg, M. (2008). Voting the general will: Rousseau on decision rules. *Political Theory* **36** 403–423.

Sijtsma, K. and Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika* **52** 79–97.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A. and Waldorp, L. J. (2014). A new

method for constructing networks from binary data. *Scientific reports* **4**.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 614–645.

van de Geer, S., Bühlmann, P. and Ritov, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166-1202.

van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement* **4** 469–492.

van der Linden, W. J. (1987). The use of test scores for classification decisions with threshold utility. *Journal of educational statistics* **12** 62–75.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* **1** 1-305.

Waldorp, L., Marsman, M. and Maris, G. (2019). Logistic regression and Ising networks: prediction and estimation when violating lasso assumptions. *Behaviormetrika* **46** 49.

Yang, E., Allen, G., Liu, Z. and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* 1358–1366.

Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2013). On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*.

Appendix

Proofs are given in their numerical order, not in order of appearance in the main text.

Proof of Proposition 2 We can use the Fourier expansion $f(x) = \sum_{S \subseteq V} \hat{f}^{\pi}(S)\phi_S(x)$ to obtain

$$\text{cov}^{\pi}(f(X_K), g(X_K) \mid h(x_L)) = \sum_{\varnothing \neq S,T \subseteq K} \hat{f}^{\pi}(S)\hat{g}^{\pi}(T)\mathbb{E}_{\pi}(\phi_S(X)\phi_T(X) \mid h(x_L))$$

If $T = S \subseteq K$, then

$$\mathbb{E}_{\pi}(\phi(X_S)^2 \mid h(x_L)) = \frac{1}{\mathbb{P}_{\pi}(h(x_L))} \int \mathbb{1}\{h(x_L)\} d\mathbb{P}_{\pi}(x_L) \prod_{i \in K} \int \phi(x_i)^2 d\mathbb{P}_{\pi}(x_i)$$

We clearly have that $\mathbb{P}_{\pi}(h(x_L)) = \int \mathbb{1}\{h(x_L)\} d\mathbb{P}_{\pi}(x_L)$, and so the first part equals 1. Furthermore, since $\int \phi(x_i)^2 d\mathbb{P}_{\pi}(x_i) = \mathbb{E}_{p_i}(\phi(x_i)^2) = 1$, we obtain that

$\mathbb{E}_\pi(\phi(X_S)^2 \mid h(x_L)) = 1$, for any two-way partition K and L of V. If $S \neq T$ this equals 0 by orthogonality (see (27)). And so

$$\text{cov}^\pi(f(X_K), g(X_K) \mid h(x_L)) = \sum_{\varnothing \neq S \subseteq K} \hat{f}^\pi(S)\hat{g}^\pi(S)$$

which is the covariance between f and g limited to the set $K \subseteq V$. Then the question is when will the Fourier coefficients $\hat{f}^\pi$ and $\hat{g}^\pi$ on K agree. From here we can assume sufficient conditions on f and g such that the covariance is positive. The first order Fourier coefficients are $\hat{f}^\pi(i) = \mathbb{E}_\pi(f(X)\phi(X_i))$. We could assume that the function f is an LTF such that

$$\hat{f}^\pi(i) = \mathbb{E}_\pi((a_0 + a_1X_1 + \cdots + a_nX_n)\phi(X_i)) = a_i\mathbb{E}_\pi(X_i\phi(X_i)) = a_i\sigma_i$$

for $i \in K$, by orthogonality. And so, for the first order coefficients we obtain

$$\text{cov}^\pi(f(X_K), g(X_K) \mid h(X_L)) = \sum_{i \in K} \hat{f}^\pi(i)\hat{g}^\pi(i) = \sum_{i \in K} a_i b_i \sigma_i^2$$

which is what was required. □

Proof of Lemma 3 Suppose that the probabilities for each node were the same, $\mathbb{P}(x_i \mid x_{\partial i}) = p$ for all $i \in V$. We then have that $\mu_i = \mu$ and $\sigma_i = \sigma$ for all $i \in V$. And so the functions $\phi(x_i)$ are all the same, given a value x_i. Then we find for the influence

$$\mathbb{E}_\pi(f(X)\phi(X_i)) = \mathbb{E}_{p^{n-1}}\left(f(x^{(i,1)})\phi(1)p - f(x^{(i,-1)})\phi(-1)(1-p)\right)$$

where p^{n-1} refers to the sequence of $n-1$ equal probabilities $(p, p, \ldots, p)$. Recall that $\phi(1) = \sqrt{(1-p)/p}$ and $\phi(-1) = \sqrt{p/(1-p)}$, so that $p\phi(1) = (1-p)\phi(-1) = \sqrt{p(1-p)}$ and $\sigma = 2\sqrt{p(1-p)}$. Then we obtain

$$\mathbb{E}_\pi(f(X)\phi(X_i)) = \mathbb{E}_{p^{n-1}}\left(f(x^{(i,1)}) - f(x^{(i,-1)})\right)\frac{1}{2}\sigma$$

Recall that the part of the right hand side in the expectation operator equals the discrete differential operator $D_i f$, and its expectation $\mathbb{E}_\pi(D_i f)$ is by definition the influence. Because all probabilities $\mathbb{P}(x_i \mid x_{\partial i}) = p$ are equal, we obtain each time the same expectation.

We next consider the equality of the probabilities p_i for $i \in V$. Because we assumed that the graph is regular, we see that the size of the boundary sets $|\partial i| = r$ is equal for all $i \in V$. By assuming additionally that the threshold and interaction parameters are all the same, i.e., $\xi = \xi_i$ for all nodes and $\beta = \beta_{ij}$ for all edges, we obtain that the probabilities, defined by the function $\xi + \beta\sum_{k \in \partial i} X_{ik}$, with the same size sets ∂i for all i, are equal up to differences in $S_r = \sum_{k \in \partial i} X_{ik}$. However, S_r contains for each node the same number of X_{ik} that have the same conditional distribution. Hence, we have a spatially stationary (shift invariant) process in which the probabilities are the same across the graph. □

Lemma 5 The majority function $\text{maj}_n : \{-1,1\}^n \to \{-1,1\}$ defined by $\text{maj}_n(x) = \text{sgn}\sum_{\imath=1}^n x_i$has the properties (a)-(e) below. A function f is

(a) *monotone* or is *positively responsive* if for $x \leq y$ ($x_j \leq y_j \forall j$) implies that $f(x) \leq f(y)$, and
(b) *odd* or *neutral* if $f(-x) = -f(x)$;
(c) *unanimous* if $f(-1,-1,\ldots,-1) = -1$ and $f(1,1,\ldots,1) = 1$;
(d) *symmetric* or *anonymous* if for any permutation $\pi : \{-1,1\}^n \to \{-1,1\}^n$ of the coordinates in x we have $f(x^\pi) = f(x)$;
(e) *transitive-symmetric* if for any $i \in V$ there is a permutation $\pi : \{-1,1\}^n \to \{-1,1\}^n$ of the coordinates in x that puts x_i in place of x_j, such that $f(x^\pi) = f(x)$.

Proof (a) The majority function is monotone because if $x_i \leq y_i$ for all i, then $\sum_i x_i \leq \sum_i y_i$, and by consequence $\text{sgn}\sum_i x_i \leq \text{sgn}\sum_i y_i$. (b) Since for x_i in the $\{-1,1\}$ domain multiplying by -1 is the negation of x_i, we have that $\text{sgn}(-x_1 - x_1 - \cdots - x_n) = -\text{sgn}(x_1 + x_2 + \cdots + x_n)$, where we obtain the negation of the sign function. (c) Follows from (a). (d) Because the majority function only considers the sum, any permutation x^π of x will have $\sum_i x_i = \sum_i x^\pi$. Finally, (e) is implied by (d) because (e) refers to particular permutations while (d) is about any permutation. □

Pseudo-likelihood We are consistent with the univariate conditional probabilities to the joint distribution in the sense that we only require a rescaling (see Lemma 6 below). We consider this with an example. Suppose we have $S = \{1,2\}$, containing two nodes, and consider a product of the two variables X_1 and X_2 over the space $\{-1,1\}^2$, where we take the expectation and hence use the joint probability

$$\begin{aligned}\mathbb{E}(X_1X_2) = \mathbb{P}(X_1 = 1, X_2 = 1) - \mathbb{P}(X_1 = -1, X_2 = 1) \\ - \mathbb{P}(X_1 = 1, X_2 = -1) + \mathbb{P}(X_1 = -1, X_2 = -1)\end{aligned}$$

We need each of the joint probabilities to factorise into a product. Let $Z_{\{1,2\}}$ be the normalising constant of the joint probability as in (20) and $Z_i(x_{\partial i})$ is the normalising constant of the conditional distribution based on the nodes in the boundary set ∂i; here $Z_1(x_2)$ and $Z_2(x_1)$. Then with the interaction parameter $\theta_{12}/2$ in the conditional probability we see that

$$\mathbb{P}(x_1, x_2) = \frac{Z_1(x_2)}{Z_{\{1,2\}}^{\frac{1}{2}}}\mathbb{P}(x_1 \mid x_2)\frac{Z_2(x_1)}{Z_{\{1,2\}}^{\frac{1}{2}}}\mathbb{P}(x_2 \mid x_1)$$

This is because $Z_1(x_2)\mathbb{P}(x_1 \mid x_2)Z_2(x_1)\mathbb{P}(x_2 \mid x_1)$ is

$$\exp\left(\xi_1 x_1 + x_1\frac{1}{2}\theta_{12}x_2\right)\exp\left(\xi_2 x_2 + x_2\frac{1}{2}\theta_{12}x_2\right) = \exp\left(\xi_1 x_1 + \xi_2 x_2 + \theta_{12}x_1x_2\right)$$

and we have the normalising constant $Z_{\{1,2\}} = Z_{\{1,2\}}^{\frac{1}{2}} Z_{\{1,2\}}^{\frac{1}{2}}$. So, the difference in the pseudo-likelihood probability and the joint probability is in the normalisation, $Z_1(x_2)Z_2(x_1)/Z_{\{1,2\}}$.

We can express the difference between the joint $\mathbb{P}$ and product of conditionals $\mathbb{P}_\pi$ in terms of the Kullback-Leibler divergence (Cover and Thomas, 2006). We see that the distributions are similar up to scaling.

Lemma 6 Let $\mathbb{P}(x)$ be the joint distribution of the Ising model with probability of x in $\{0,1\}^n$ or $\{-1,1\}^n$ as in (19). Furthermore, in the conditional probability $\mathbb{P}_i$ we use the parameterisation $\theta_{ij}/2$, and let $Z_\pi(x) = \prod_{i\in V} Z_i(x_{\partial i})$. Then we have the factorisation of graph G

$$\mathbb{P}(x) = \frac{Z_1(x_{\partial 1})}{Z_V^{\frac{1}{n}}}\mathbb{P}(x_1 \mid x_{\partial 1}) \times \cdots \times \frac{Z_n(x_{\partial n})}{Z_V^{\frac{1}{n}}}\mathbb{P}(x_n \mid x_{\partial n}) \tag{65}$$

Proof The statement about the probability for each configuration x is easy to see. The product

$$\mathbb{P}(x) = \frac{Z_1(x_{\partial 1})}{Z_V^{\frac{1}{n}}}\mathbb{P}(x_1 \mid x_{\partial 1}) \times \cdots \times \frac{Z_n(x_{\partial n})}{Z_V^{\frac{1}{n}}}\mathbb{P}(x_n \mid x_{\partial n})$$

gives the normalising constant Z_V and the product of conditionals with the normalising term removed by $Z_{p_i}(x_{\partial i})$, leads to

$$\prod_{i\in V} \exp\left(\xi_i x_i + \frac{1}{2} x_i \sum_{j\in\partial i} x_j\right) = \exp\left(\sum_i \xi_i x_i + 2\frac{1}{2}\sum_{(i,j)\in E} x_i x_j\right)$$

because each node is visited twice and all neighbourhoods $\partial 1, \partial 2, \ldots, \partial n$ together give the edge set E. □

Lemma 7 Let $\mathbb{P}$ be the joint Ising probability (19) and $\mathbb{P}_\pi$ be the product of conditionals (22). Then the Kullback-Leibler divergence is

$$D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{P}_\pi) = \log \frac{Z_V}{\prod_{i\in V} Z_i}$$

Proof The Kullback-Leibler divergence is

$$D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{P}_\pi) = \sum_{x\in\{-1,1\}^n} \mathbb{P}(x)\log\mathbb{P}(x) - \sum_{x\in\{-1,1\}^n}\sum_{i\in V} \mathbb{P}(x)\log\mathbb{P}_i(x)$$

By Lemma 6, where we use the rescaled univariate conditionals with interaction parameters $\theta_{ij}/2$, we have

$$\mathbb{P}(x) = \frac{\prod_{i\in V} Z_i(x_{\partial i})}{Z_V} \prod_{i\in V} \frac{1}{Z_i(x_{\partial i})} \exp\left(\xi_i x_i + \frac{1}{2}\sum_{j\in\partial i}\theta_{ij}x_i x_j\right)$$

And so we see that we get from the rescaled product of conditionals to the joint distribution of the Ising probability. Plugging in the KL divergence for each i the univariate conditional

$$\mathbb{P}_i(x) = \frac{1}{Z_i(x_{\partial i})} \exp\left(\xi_i x_i + \frac{1}{2}\sum_{j\in\partial i}\theta_{ij}x_i x_j\right)$$

gives the result. □

Lemma 8 The function $\chi_S : \{-1,1\} \to \{-1,1\}$ forms an orthonormal basis for the space $L^2(\{-1,1\})$.

Proof The parity functions form an orthonormal basis for the space $L^2(\{-1,1\})$. First note that

$$\chi_S\chi_T = \prod_{i\in S} x_i \prod_{i\in T} x_i = \prod_{i\in S\Delta T} x_i \prod_{i\in S\cap T} x_i^2 = \prod_{i\in S\Delta T} x_i$$

where $S\Delta T$ is the symmetric difference $S\cap T^c \cup T\cap S^c$. Furthermore,

$$\mathbb{E}\chi_S(X) = \mathbb{E}\prod_{i\in S} x_i = \prod_{i\in S}\mathbb{E}(x_i)$$

which equals 1 if $S = \varnothing$ since then $\chi_\varnothing(x) = 1$ and 0 otherwise. Hence, if $S = T$ then we obtain $S\Delta T = \varnothing$ and so obtain 1, and 0 if $S \neq T$. □

Lemma 9 Let f be a Boolean function and let the x_i be biased in that $\mathbb{E}_{p_i}(x_i) = 2p_i - 1$ for each i (possibly based on the conditional distributions obtained with the Ising model). Then the mean, variance and covariance are, respectively,

$$\mathbb{E}_\pi(f(X)) = \mathbb{E}_\pi\left(\sum_{S\subseteq V}\hat{f}^\pi(S)\phi_S(x)\right) = \hat{f}^\pi(\varnothing)$$

and

$$\mathrm{var}^\pi(f) = \mathbb{E}_\pi f(X)^2 - (\mathbb{E}_\pi f(X))^2 = \sum_{S\neq\varnothing}\hat{f}^\pi(S)^2$$

and

$$\text{cov}^\pi(f) = \mathbb{E}_\pi f(X)g(X) - \mathbb{E}_\pi f(X)\mathbb{E}_\pi g(X) = \sum_{S\neq\varnothing} \hat{f}^\pi(S)\hat{g}^\pi(S)$$

Proof The mean is obvious since the only term of $\mathbb{E}_\pi(\phi_S(x))$ that is non-zero is when $S = \varnothing$. For the variance, we have

$$\mathbb{E}_\pi f(X)^2 = \sum_{S\subseteq V}\sum_{T\subseteq V} \hat{f}^\pi(S)\hat{f}^\pi(T)\mathbb{E}_\pi(\phi_S(x)\phi_T(x))$$

and

$$\mathbb{E}_\pi(\phi_S(x)\phi_T(x)) = \mathbb{E}_\pi\left(\prod_{i\in S}\phi(x_i)\prod_{i\in T}\phi(x_i)\right)$$

If $S = T$ then all terms are $\mathbb{E}_\pi\phi(x_i)^2$, which for all $i \in S$ equal 1. If $S \neq T$ then there is at least one $i \notin S \cap T$ (if $i \notin T$, say) and so we get

$$\mathbb{E}_{p_i}\phi(x_i)\mathbb{E}_{\pi\setminus p_i}\left(\prod_{j\in S\setminus\{i\}}\phi(x_j)\prod_{j\in T}\phi(x_j)\right)$$

where $\mathbb{E}_{p_i}\phi(x_i)$ equals 0. The covariance cov^π is analogous to the variance, since both f and g are decomposed in terms of the functions ϕ_S. □

Lemma 10 For a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ with the variables x_i having mean $\mu_i = 2p_i - 1$ and variance $\sigma_i^2 = 1 - \mu_i^2$, the influence is for item i is

$$\mathbb{I}_i^\pi(f) = \sum_{S\ni i} \hat{f}^\pi(S)^2$$

where $S \ni i$ denotes the subsets $S \subseteq V$ such that $i \in S$. Furthermore, for monotone Boolean functions f $\mathbb{I}_i^{p_i}(f) = \sigma_i\mathbb{I}_i(f)$.

Proof Following the definition of influence, we obtain

$$\mathbb{I}_i^\pi(f) = \mathbb{E}_\pi(D_{\phi,i}f(x)^2) = \mathbb{E}_\pi\left(\sum_{S\ni i}\hat{f}^\pi(S)\phi_{S\setminus\{i\}}(x)\sum_{S\ni i}\hat{f}^\pi(T)\phi_{T\setminus\{i\}}(x)\right)$$

$$\mathbb{E}_\pi\phi_{S\setminus\{i\}}(x)\phi_{T\setminus\{i\}}(x) = \begin{cases}1 & \text{if } S = T\\ 0 & \text{if } S \neq T\end{cases}$$

because $\mathbb{E}_{p_i}\phi(x_i)^2 = 1$ and 0 for $i \neq j$. Hence,

$$\mathbb{I}_i^{\pi}(f) = \sum_{S \ni i} \hat{f}^{\pi}(S)^2$$

Furthermore, in terms of the unbiased version, $D_{\phi,i}f = \sigma_i D_i f$. The result now follows for monotone functions f. □

Lemma 11 The influence $\mathbb{I}_i^{\pi}(f)$ of $i \in V$ on function monotone $f : \{-1,1\}^n \to \{-1,1\}$ can be described in terms of the Fourier coefficients $\hat{f}^{\pi}(S)$ by singleton sets $S = \{i\}$

$$\mathbb{I}_i^{\pi}(f) = \frac{1}{\sigma_i}\hat{f}^{\pi}(i)$$

where $\hat{f}^{\pi}(i) = \hat{f}^{\pi}(\{i\})$.

Proof Because f is monotone by assumption, we have that $D_i f \geq 0$, and so $D_i f(x) = D_i f(x)^2$ and is either 0 or 1, hence $\mathbb{I}_i(f) = \mathbb{E}_{\pi}(D_i f(x))$. And

$$\mathbb{E}_{\pi}(D_i f(x)) = \mathbb{E}_{\pi}\left(\sum_{\substack{S \subseteq V \\ i \in S}} \hat{f}^{\pi}(S)\phi_{S\setminus\{i\}}(x)\right) = \sum_{i \in S} \hat{f}^{\pi}(S)\mathbb{E}_{\pi}(\phi_{S\setminus\{i\}}(x))$$

And

$$\mathbb{E}_{\pi}(\phi_{S\setminus\{i\}}(x)) = \begin{cases} 1 & \text{if } i \in S = \{i\} \\ 0 & \text{if } i \notin S \end{cases}$$

because if $i \in S$ then $S\setminus\{i\} = \varnothing$ only if $S = \{i\}$. Hence, we have nonzero coefficients with $\hat{f}^{\pi}(S) = \hat{f}^{\pi}(\{i\})$.

□

Lemma 12 Let $f : \{-1,1\}^n \to \{-1,1\}$ be any Boolean function for unbiased variables X_i and W is the number of items that agree with f. Then

$$\mathbb{E}(W) = \frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n} \hat{f}(i)$$

Proof Because

$$\sum_{i=1}^{n} \hat{f}(i) = \sum_{i=1}^{n} \mathbb{E}(f(x)x_i) = \mathbb{E}(f(x)(x_1 + \cdots + x_n))$$

and $f(x)(x_1 + \cdots + x_n)$ is the the number of items that agree with $f(x)$ or with $-f(x)$, and so is $w - (n - w) = 2w - n$. Hence,

$$\sum_{i=1}^{n} \hat{f}(i) = \sum_{i=1}^{n} \mathbb{E}(f(x)x_i) = \mathbb{E}(f(x)(x_1 + \cdots + x_n)) = \mathbb{E}(2W - n) = 2\mathbb{E}(W) - n$$

leading to the result. □

Lemma 13 Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a Boolean function and for all i, Y_i are obtained from the experiment where with probability $\frac{1}{2}(1 + \rho)$, $y_i = x_i$ and with probability $\frac{1}{2}(1 - \rho)$, $y_i = -x_i$. Then the variance of $f(Y)$ is

$$\text{var}^{\pi,\rho}(f(Y)) = \sum_{S \neq \varnothing} \hat{f}^{\pi,\rho}(S)^2$$

and the covariance between $f(X)$ and $f(Y)$ is

$$\text{cov}^{\pi,\rho}(f(X), f(Y)) = \sum_{S \neq \varnothing} \omega(S)\rho^{|S|} \hat{f}^{\pi}(S)\hat{f}^{\pi,\rho}(S)$$

where $\omega(S) = \prod_{i \in S} \frac{\sigma_i}{\sigma_i^p}$.

Proof We require the covariance between $f(X)$ and $f(Y)$ and their variances. We start with the variance and then determine the covariance. The variance of $f(X)$ is determined in Lemma 9. To determine the variance of $f(Y)$ we use the Fourier expansion $f(y) = \sum_{S \subseteq V} \hat{f}^{\pi}(S)\phi_S(y)$, where we need to redefine the Fourier expansion for y. We first note that $\mathbb{E}_{p_i,\rho}(Y_i) = \rho\mu_i$ and $\mathbb{E}_{p_i,\rho}(Y_i^2) = 1 - \rho^2\mu_i^2$. Then

$$\phi^{\rho}(Y) = \frac{Y_i - \rho\mu_i}{\sqrt{1 - \rho^2\mu_i^2}}$$

so that $\mathbb{E}_{p_i,\rho}(\phi^{\rho}(Y_i)) = 0$ and $\mathbb{E}_{p_i,\rho}(\phi^{\rho}(Y_i)^2) = 1$. We obtain the Fourier expansion

$$\hat{f}^{\pi,\rho}(S) = \mathbb{E}_{\pi,\rho}(f(Y)\phi_S^{\rho}(Y))$$

and we have again orthonormality for the Fourier expansion as before (27). Then, for the variance, we have

$$\mathbb{E}_{\pi,\rho} f(Y)^2 = \sum_{S \subseteq V} \sum_{T \subseteq V} \hat{f}^{\pi}(S)\hat{f}^{\pi}(T)\mathbb{E}_{\pi,\rho}(\phi_S^{\rho}(Y)\phi_T^{\rho}(Y))$$

which is nonzero only if $S = T$. Now we can apply Lemma 9 for the variance and obtain the result.

For the covariance we have

$$\text{cov}^{\pi,\rho}(f(X), f(Y)) = \sum_{S,T\subseteq V} \hat{f}^{\pi}(S)\hat{f}^{\pi,\rho}(T)\mathbb{E}_{\pi,\rho}(\phi_S(X)\phi^{\rho}_T(Y))$$

and by orthogonality

$$\mathbb{E}_{\pi,\rho}(\phi_S(X)\phi^{\rho}_S(Y)) = \prod_{i\in S} \mathbb{E}_{\pi,\rho}(\phi(X_i)\phi^{\rho}(Y_i))$$

We have that the expectation of $\phi(X_i)\phi^{\rho}(Y_i)$, with respect to the measure $\mathbb{P}_{\rho}$ and then with respect to $\mathbb{P}_{p_i}$, is

$$\mathbb{E}_{p_i,\rho}(\phi(X_i)\phi^{\rho}(Y_i)) = \frac{1}{2}(1+\rho)\mathbb{E}_{p_i}(\phi(X_i)\phi^{\rho}(Y_i)) + \frac{1}{2}(1-\rho)\mathbb{E}_{p_i}(\phi(X_i)\phi^{\rho}(-X_i))$$

These expectations are

$$\mathbb{E}_{p_i}(\phi(X_i)\phi^{\rho}(Y_i)) = \frac{\sqrt{1-\mu_i^2}}{\sqrt{1-\rho^2\mu_i^2}} \quad \text{and} \quad \mathbb{E}_{p_i}(\phi(X_i)\phi^{\rho}(-X_i)) = -\frac{\sqrt{1-\mu_i^2}}{\sqrt{1-\rho^2\mu_i^2}}$$

which is the ratio of standard deviations, σ_i in the nominator and σ_i^{ρ} in the denominator. And so, considering the function $\prod_{i\in S}\mathbb{E}_{\pi,\rho}(\phi(X_i)\phi^{\rho}(Y_i))$, and letting $\omega(S) = \prod_{i\in S}\frac{\sigma_i}{\sigma_i^{\rho}}$, we obtain for a given $S \subseteq V$

$$\mathbb{E}_{\pi,\rho}(\phi_S(X)\phi^{\rho}_S(Y)) = \prod_{i\in S}\text{cor}^{\pi,\rho}(X_i, Y_i) = \prod_{i\in S}\frac{\sigma_i}{\sigma_i^{\rho}}\rho = \omega(S)\rho^{|S|}$$

Using the Fourier representation we obtain

$$\mathbb{E}_{\pi,\rho}(f(X)f(Y)) = \sum_{S\subseteq V}\omega(S)\rho^{|S|}\hat{f}^{\pi}(S)\hat{f}^{\pi,\rho}(S)$$

The mean of $f(X)$ is $\hat{f}^{\pi}(\varnothing)$ and for $f(Y)$ the mean is $\hat{f}^{\pi,\rho}(\varnothing)$, and so we obtain the covariance. □