Discovering causal structures in binary exclusive-or skew acyclic models

Takanori Inazumi¹, Takashi Washio¹, Shohei Shimizu¹, Joe Suzuki², Akihiro Yamamoto³ and Yoshinobu Kawahara¹

¹The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, Japan
²Graduate School of Science, Osaka University, 1-1 Machikaneyama, Toyonaka, Osaka, Japan
³Graduate School of Informatics, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
⁴JST, ERATO, Minato Discrete Structure Manipulation System Project, 1-4-14 Shibata, Kita-ku, Osaka, Japan

Abstract

Discovering causal relations among observed variables in a given data set is a main topic in studies of statistics and artificial intelligence. Recently, some techniques to discover an identifiable causal structure have been explored based on non-Gaussianity of the observed data distribution. However, most of these are limited to continuous data. In this paper, we present a novel causal model for binary data and propose a new approach to derive an identifiable causal structure governing the data based on skew Bernoulli distributions of external noise. Experimental evaluation shows excellent performance for both artificial and real world data sets.

1 INTRODUCTION

Many approaches to causal inference and structure learning of Bayesian networks have been studied in statistics and artificial intelligence (Pearl, 2000), (Spirtes et al., 2000). Most of these derive candidate causal structures from an observed data set by assuming acyclicity of the causal dependencies. These mainly use information up to second-order statistics in the observed variables, and narrow down the candidate directed acyclic graphs (DAGs) by using some constraints and/or scoring functions. However, these approaches often retain multiple candidate causal structures within equivalence classes and local optima.

Recently, some linear non-Gaussianity-based approaches as applied to acyclic models (LiNGAM) have been proposed (Shimizu et al., 2006). Such approaches are known to derive an identifiable causal structure of a linear structural equation model by using information on the non-Gaussian distribution of external noise driving the observed variables. However, these require linearity of the objective system and are only applicable to continuous variables. Recent studies extended this principle to non-linear models with continuous variables (Hoyer et al., 2009) and variables having ordered discrete values (Peters et al., 2011). However, these studies are not primarily applicable to more than two variables, and do not cover any variables having unordered categorical or binary values.

In contrast, many real world domains such as medicine (Pearl, 2000), bioinformatics (Shi et al., 2007) and sociology (Spirtes et al., 2000), maintain massive stochastic binary data sets, and practitioners need to discover those causal structures from the data for various purposes. However, to our best knowledge, all past approaches do not ensure an identifiable solution for any given binary data set. In this paper, our objective is to propose a new approach to discover an identifiable causal structure under some feasible assumptions within a given stochastic binary data set.

In the next section, we briefly review some related work to indicate important technical issues. In the third section, we introduce a novel binary exclusive-or skew acyclic model, termed “BExSAM”, to represent an objective system, and characterize the model in regard to causal ordering and causal structure estimation. In the fourth, we propose novel algorithms and measures for the causal ordering and the causal structure estimation based on the model characterization. In the fifth, we present some experimental evaluations using both artificial and real world data sets.

2 RELATED WORK

Many studies on causal inference and structure learning of Bayesian networks have exerted much effort in developing principles for efficient searches by using information up to second-order statistics (Pearl, 2000), (Spirtes et al., 2000). That has arisen because exhaustive searches are intractable as the number of possible directed acyclic graphs (DAGs) grows exponentially with the number of variables. To address this issue,
constraint-based approaches such as the PC algorithm (Spirtes et al., 2000) and score-based approaches such as the GES algorithm (Chickering, 2002) have been studied for both continuous and discrete variables. However, these admit multiple solutions within equivalence classes and local optima in many cases, and thus often fail to generate an identifiable causal structure.

A recent technique LiNGAM (Shimizu et al., 2006) formulates the DAG structure search in form of an independent component analysis (ICA), that ensures the existence of a unique global optimum under assumptions of linear relations among the observed variables and non-Gaussianity of external noises. However, it may often provide a local optimum solution because of its greedy search. In contrast, more recent DirectLiNGAM (Shimizu et al., 2009) efficiently derives an identifiable solution through its iterative findings of exogenous variables by applying linear regressions and independence measures. (Hoyer et al, 2009) and (Zhang and Hyvarinen, 2009) proposed extensions of this principle to a non-linear additive noise model and a post-nonlinear (PNL) causal model respectively. However, their identifiability has been primarily shown only for two variable cases with the exception of a few special cases. (Mooij et al., 2009) proposed a novel independence-based regression to allow causal inference in a non-linear additive noise model containing multiple variables. However, its identifiability of the causal structure is not confirmed, and its convergence to a unique solution is not ensured because of the non-convexity of the regression problem.

One main advantage of non-Gaussianity-based approaches is the capability to derive an identifiable causal structure under some feasible model assumptions. However, because non-Gaussianity is a property of continuous distributions, these are applicable to data sets containing only continuous variables. In consequence, only a few studies have addressed this issue for discrete variable sets. A study on this topic was recently reported by (Peters et al., 2011). However, its technique is applicable only to two variables having ordered discrete values, because its basic principle is essentially an extension of the aforementioned work (Hoyer et al, 2009). Another study (Sun and Janzing, 2007) proposed a principle to find a causal order of binary variables to explain a given sample distribution by mutually independent Markov kernels. However, its identifiability is not confirmed, and its applicability is limited to very simple Boolean relations because of high computational complexity of the kernel functions for generic cases.

Because binary variables do not constitute a continuous algebra, we should develop a structural model of acyclic relations among the binary variables upon other algebraic systems such as a Boolean algebra. In addition, we need to apply a binary data distribution such as a Bernoulli distribution instead of Gaussian/non-Gaussian distributions in the model. Moreover, we have to invent novel algorithms for both causal ordering and causal structure estimation based upon characteristics of the structural model and the data distribution. In the following sections, we present our ideas concerning these issues.

3 PROPOSED PRINCIPLES

3.1 BExSAM

We first introduce a novel structural model involving generic acyclic causal relations among binary variables. Our model has a DAG structure where every external noise affects its corresponding observed variable via an exclusive-or operation which is different from the continuous additive noise.

**Definition 1** Given a set of $d$ observed binary variables $X = \{x_i | x_i \in \{0,1\}, i = 1, \ldots, d\}$. Let a set of $d$ unobserved binary noises be $E = \{e_i | e_i \in \{0,1\}, i = 1, \ldots, d\}$ where each $e_i$ is a mutually independent stochastic variable. Then, the index $i$ is a permutation $i : \{1, \ldots, d\} \rightarrow \{1, \ldots, d\}$ so that

$$
\begin{align*}
    x_{i(1)} &= f_{i(1)} \oplus e_{i(1)}, \quad \text{and} \\
    x_{i(k)} &= f_{i(k)}(x_{i(1)}, \ldots, x_{i(k-1)}) \oplus e_{i(k)} \\
    & \quad \quad \quad \text{for } k = 2, \ldots, d,
\end{align*}
$$

where $f_{i(1)}$ is a constant in $\{0,1\}$, $f_{i(k)} : \{0,1\}^{k-1} \rightarrow \{0,1\}$ is a deterministic Boolean function for $k \geq 2$, and $\oplus$ is an exclusive-or operator.

$f_{i(k)}$ expresses any deterministic binary relation without loss of generality because such a relation is always represented by Boolean algebraic formulae (Gregg, 1998). In the rest of this paper, subscripts $i(k)$ will be shortened to $i$ for notational simplicity, unless ambiguity requires to use the full notation. In addition, $f_{i(k)}(x_{i(1)}, \ldots, x_{i(k-1)})$ is represented by $f_i$ for brevity.

Furthermore, we introduce a Bernoulli distribution and skewness to every noise as follows.

**Definition 2** Let the probability of $e_i = 1$ be $p(e_i = 1)$. $p(e_i = 1)$ is a constant $p_i$ satisfying $0 < p_i < 1$ and $p_i \neq 1/2$ for all $i = 1, \ldots, d$.

$p(e_i = 0) = 1 - p_i$ satisfies the same constraints $0 < p(e_i = 0) < 1$ and $p(e_i = 0) \neq 1/2$. This noise characteristic is essentially needed in our approach as an analogue to the aforementioned non-Gaussianity in the case of continuous variables. A model based on
the above two definitions is called a binary exclusive-or skew acyclic model, i.e., “BEzSAM” for short.

When a directional edge exists from a variable $x_j$ to another variable $x_i$ ($i \neq j$) in a DAG model, $x_j$ is called a “parent” of $x_i$ and $x_i$ is called a “child” of $x_j$. As widely noted in the causal inference study (Pearl, 2000), (Spirtes et al., 2000), a variable having no parents in a DAG model is called an “exogenous variable”, otherwise, it is called an “endogenous variable.” In this study, we further introduce the following definition of a particular endogenous variable.

**Definition 3** An endogenous variable which does not have any children in a DAG model is called a “sink endogenous variable.”

This sink endogenous variable plays a key role in a BEzSAM in regard to principles and algorithms for causal ordering and causal structure estimation.

**Example 1** The following is an example of a BEzSAM consisting of four binary variables with $i(k) = k$.

\[
\begin{align*}
x_1 & = e_1, \\
x_2 & = x_1 \oplus e_2, \\
x_3 & = x_1x_2 \oplus e_3, \\
x_4 & = (x_1 + x_3) \oplus e_4.
\end{align*}
\]

Here, and or operation of $x_1$ and $x_2$ and an or operation of $x_1$ and $x_3$ are represented by $x_1x_2$ and $x_1 + x_3$ respectively for simplicity. $x_1$ is an exogenous variable, and $x_4$ is a sink endogenous variable.

### 3.2 CHARACTERIZATION

In this subsection, characteristics of BEzSAM associated with a sink endogenous variable are analyzed concerning the discovery of causal structure. First, we define a notion of “control” which specifies the values of some observed variables in $X$.

**Definition 4** Let $X_j = \{0,1\}$ be a domain of $x_j \in X$. Given $X_i = X \setminus \{x_i\}$, i.e., $X_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d\}$, and let $V_i$ be a $(d-1)$-tuple $(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_d)$ where $v_j \in X_j$ for all $x_j \in X_i$. We denote $X_i = V_i$ when we assign the value $v_j$ to its $x_j$ for all $x_j \in X_i$. This assignment is called a “control” of $X_i$ at $V_i$. Similarly, given $X_{ij} = X \setminus \{x_i, x_j\}$. The “control” of $X_{ij}$ at $V_{ij}$ is defined in the same way and denoted by $X_{ij} = V_{ij}$.

We present an important theorem for causal ordering of the variables in $X$ by using the control.

**Theorem 1** The following conditions are equivalent.

1. $x_i \in X$ is a sink endogenous variable.

2. Let $X_i = X \setminus \{x_i\}$. There is a common constant $q_i$ such that $p(x_i = 1|X_i = V_i) = q_i$ or $1 - q_i$ (and therefore, $p(x_i = 0|X_i = V_i) = 1 - q_i$ or $q_i$, equivalently) for all control $X_i = V_i \in \{0,1\}^{d-1}$.

**Proof.** See Appendix 1.

For example, if we are given the two controls $X_4 = \{x_1, x_2, x_3\}$ at $V_4 = (0,0,0)$ and $V_4 = (0,0,1)$ in Example 1, we have the following conditional probabilities of the sink endogenous variable: $x_4 = 1$ as

\[
p(x_4 = 1|X_4 = V_4) = p_{4}, \quad \text{and} \quad p(x_4 = 1|X_4 = V_4') = 1 - p_{4},
\]

respectively. Actually, $p(x_4 = 1|X_4 = V_4) = p_{4}(= q_{4})$ or $1 - p_{4}(= 1 - q_{4})$ holds for any $V_4$ in this case. In contrast, if we are provided with controls $X_3 = \{x_1, x_2, x_4\}$ at $V_3 = (0,0,0)$ and $V_3' = (0,0,1)$,

\[
p(x_3 = 1|X_3 = V_3) = \frac{p_{4}p_{4}}{p_{3}p_{4} + (1 - p_{3})(1 - p_{4})}, \quad \text{and}
\]

\[
p(x_3 = 1|X_3 = V_3') = \frac{p_{3}(1 - p_{4})}{p_{3}(1 - p_{4}) + (1 - p_{3})p_{4}}
\]

hold. Because $p_{3}, p_{4} \neq 1/2$ by Definition 2, these probabilities are not equal, and also their sum is not unity. Accordingly, no constant $q_3$ or $1 - q_3$ can be assigned to both of $p(x_3 = 1|X_3 = V_3)$ and $p(x_3 = 1|X_3 = V_3')$ in this case. These results reflect Theorem 1, that we can find a sink endogenous variable in $X$ by checking the conditional probability of every variable.

Next, we present an important proposition to estimate causal structures in $X$.

**Proposition 1** Let $x_i \in X$ be a sink endogenous variable. Given a variable $x_i \in X$ ($i \neq j$), let $X_{ij} = V_{ij}$ be a control of $X_{ij} = X \setminus \{x_i, x_j\}$ at $V_{ij} \in \{0,1\}^{d-2}$.

1. $x_j$ is a parent of $x_i$, $\iff$ $p(x_i = v_i|x_j = v_j, X_{ij} = V_{ij}) = p(x_i = v_i|x_j = \bar{v}_j, X_{ij} = V_{ij})$ for some $V_{ij}$.

2. $x_j$ is not a parent of $x_i$, $\iff$ $p(x_i = v_i|x_j = v_j, X_{ij} = V_{ij}) = p(x_i = v_i|x_j = \bar{v}_j, X_{ij} = V_{ij})$ for all $V_{ij}$, where $v_i \in \{0,1\}$ and $\bar{v}_i = v_i \oplus 1$.

If $x_j$ is a parent of $x_i$, $f_i$ depends on $x_j$ under some controls, and its converse statement is also true. If $x_j$ is not a parent of $x_i$, $f_i$ does not depend on $x_j$ under any controls, and vice versa. By combining these facts with Lemma 2 in Appendix 2, we obtain this proposition. In Example 1, the sink endogenous variable $x_4$ satisfies the r.h.s. of the relation 1 of Proposition 1 with $x_1$ under the control $X_{14} = \{x_2, x_3\}$ at $V_{14} = (0,0)$, and $x_1$ is a parent of $x_4$ in fact. In contrast, $x_4$ satisfies the r.h.s. of the relation 2 of Proposition 1 with $x_2$, which is not a parent of $x_4$, under
the control \(X_{24} = \{x_1, x_3\}\) at \(V_{24} = \{0, 0\}\). Proposition 1 indicates a way to identify parent variables of a sink endogenous variable by checking the conditional probability of every other variable.

4 PROPOSED ALGORITHMS AND INDEPENDENCE MEASURES

4.1 OUTLINE\(^1\)

We propose an approach to discover an identifiable causal structure in a BExSAM from a given binary data set \(D = \{V^{(h)} | h = 1, \ldots, n\}\) where each \(V^{(h)}\) is a \(d\)-dimensional vector of all observed values describing \(X\). This approach is based on Theorem 1 and Proposition 1. Since we assume that \(D\) is generated through a process well modeled by a BExSAM, every element \(v_i^{(h)}\) of \(V^{(h)}\) is generated through \(x_i = f_i \oplus e_i\) by Definition 1 where the distributions of \(e_i\) are mutually independent and follow the constraints in Definition 2. Accordingly, if the sample size \(n\) is sufficiently larger than \(2^d\), \(D\) contains varieties of vectors \(V^{(h)}\), and enables to estimate the conditional probabilities under various controls. The required size of \(D\) is tractable when the number of its observed variables is moderate as shown in numerical experiments in a later section.

Figure 1 shows the outline of our proposed algorithm. The first step computes a frequency table \(FT\) of \(D\). Since the algorithm uses \(FT\) only in the latter steps, we can directly apply the algorithm to \(FT\) if \(FT\) is provided as a data set. In the loop from the next step, the algorithm seeks a sink endogenous variable \(x_i^{(k)}\) through a function “find\_sink” at step 3 and the parent variables \(P_i^{(k)}\) of the found \(x_i^{(k)}\) via “find\_parent” at step 4. These functions perform causal ordering and causal structure estimation in the BExSAM. Step 5 reduces the search space in the next loop by removing the estimated \(x_i^{(k)}\) from \(X\) and marginalizing \(x_i^{(k)}\) in \(FT\). The entire list of \(x_i^{(k)}\) and \(P_i^{(k)}\) in the output represents a resultant DAG structure of the BExSAM. This iterative reduction from the bottom in the causal order is similar to the causal ordering of (Mooij et al., 2009) while their causal structure estimation needs a second sweep from the top to the bottom.

4.2 CAUSAL ORDERING

Our causal ordering algorithm is summarized in Fig. 2. In the loop starting from step 1, it computes the conditional probabilities for all controls by following Theorem 1. \(FT(x_i = v_i, X_i = V_i)\) at step 4 expresses the frequency of a value combination; \(x_i = v_i\) and \(X_i = V_i\).  

\(^1\)Our code is available from http://www.ar.sanken.osaka-u.ac.jp/~inazumi/bexsam.html.
input: a frequency table $FT$, its variable list $X$ and a sink endogenous variable $x_i$.

1. $P_1 = \phi$.
2. for $j := 1$ to $d$ and $j \neq i$ do
3. $X_{ij} := X \setminus \{x_i, x_j\}$.
4. for all $V_{ij} \in \{0, 1\}^{d-2}$ do
5. $p(x_i = 1| x_j = v_j, X_{ij} = V_{ij}) := \frac{FT(x_i = 1 | x_j = v_j, X_{ij} = V_{ij}) + FT(x_i = 1 | x_j = v_j, X_{ij} = \neg V_{ij})}{FT(x_j = v_j, X_{ij} = V_{ij})}$ for $v_j \in \{0, 1\}$.
6. compute an independence score $S_p(x_j, V_{ij})$ from $FT$ and $p(x_i = 1| x_j = v_j, X_{ij} = V_{ij})$ for $v_j \in \{0, 1\}$.
7. end
8. apply multiple comparison tests over all $S_p(x_j, V_{ij}) \ (V_{ij} \in \{0, 1\}^{d-2})$ to judge if $x_j$ is a parent of $x_i$. Include the parent $x_j$ into $P_1$.
9. end

output: $P_1$.

Figure 3: Algorithm of find_parent

at step 5 is the frequency of a value combination; $x_i = v_i, x_j = v_j$ and $X_{ij} = V_{ij}$. Furthermore, an independence score $S_p(x_j, V_{ij})$ of each variable $x_j$ is computed at step 6. Finally, variables $x_j$ having low $S_p(x_j, V_{ij})$ are selected as parents of a sink endogenous variable $x_i$ through multiple comparison tests.

The frequency $FT(x_i = 1, x_j = v_j, X_{ij} = V_{ij})$ which is required in the numerator at step 5 follows a binomial distribution $B(FT(x_j = v_j, X_{ij} = V_{ij}), p(x_i = 1| x_j = v_j, X_{ij} = V_{ij}))$. Thus, we easily derive the variance $\sigma^2(X_{ij} = V_{ij})$ of $\delta p(X_{ij} = V_{ij}) := p(x_i = 1| x_j = v_j, X_{ij} = V_{ij}) - p(x_i = 1| x_j = v_j, X_{ij} = \neg V_{ij})$ and compute $S_p(x_j, V_{ij})$ which is a P-value of null-hypothesis: $p(x_i = 1| x_j = v_j, X_{ij} = V_{ij}) = p(x_i = 1| x_j = \neg v_j, X_{ij} = V_{ij})$ by using Gaussian approximation. We test this null-hypothesis over all controls $V_{ij} \in \{0, 1\}^{d-2}$ by multiple comparison tests (Benjamini, 1995). If accepted, we cannot say that condition 2 in Proposition 1 does not hold, and we presume that $x_j$ is not a parent of $x_i$. Otherwise, condition 1 in Proposition 1 is likely to hold, and $x_j$ is judged to be a parent. These multiple comparison tests do not cause severe failures in the causal structure estimation, because their errors are localized into the parent estimation of each sink endogenous variable. This is different from the errors of multiple conditioning used in PC algorithms which can induce global failures. We should note that a significance level $\alpha$ used in these tests is a unique parameter in our algorithms.

4.4 COMPUTATIONAL COMPLEXITY

The largest table used in the above algorithms is the frequency table $FT$ of size $2^d$. According to the requirement size, $n > 2^d$, the memory complexity is given by the size of $FT$, $O(n)$.

The loop involved in the “find_sink” function compute the conditional probabilities $2^{d-1}$ times, and compute a score $S_p(x_i)$ by aggregating these $2^{d-1}$ probabilities and the frequency table $FT$. Thus, these are $O(n)$. Since the loop repeats $d$ times, the time complexity of “find_sink” is $O(dn)$. Similarly, the loop involved in the “find_parent” function computes the conditional probabilities and scores $2^{d-2}$ times. The multiple comparison tests requires sorting $2^{d-2}$ scores. Therefore, the time complexity is $O(n)$. Because these are repeated $d$ times in the outer loop, “find_parent” which are $O(dn)$ are repeated $d$ times in the main algorithm. Accordingly, the total time complexity of the proposed algorithms is $O(d^2n)$. This complexity is favorable compared with past work. For example, DirectLiNGAM which also has an iterative algorithm structure and is considered to be one of the most efficient algorithms has $O(d^3n)$ complexity.

5 EXPERIMENTAL EVALUATION

5.1 BASIC PERFORMANCE FOR ARTIFICIAL DATA

For numerical experiments, we generated artificial data by using BEZSAMs where every $f_i(k)$ is represented by the following algebraic normal form (ANF).

$$f_i(k) = a_{0}^k \oplus a_{1}^k x_{i(1)} \oplus a_{2}^k x_{i(2)} \oplus \ldots \oplus a_{d}^k x_{i(d)} \oplus a_{1,2}^k x_{i(1)}x_{i(2)} \oplus \ldots \oplus a_{k-1,k}^k x_{i(k-1)}x_{i(k)} \oplus \ldots \oplus a_{1,2,\ldots,k}^k x_{i(1)}x_{i(2)} \ldots x_{i(k)} \oplus e_i(k).$$

Each coefficient $a_{k}^i$ is a constant in $\{0, 1\}$. Under given parameters of $d$, $n$ and $p(a_{k}^i = 1)$, we randomly determine the permutation $i : \{1, \ldots, d\} \rightarrow \{1, \ldots, d\}$ in Definition 1, and randomly generate all $a_{k}^i$ in this expression for all $k = 1, \ldots, d$. Because any Boolean formula $f_i(k)$ is known to be represented by an ANF with arbitrary coefficients $a_{k}^i$ (Carlet and Guilhot, 1999), a set of the randomized $a_{k}^i$ generates a generic BEZSAM.

Moreover, we randomly generate $e_{i(1)}^{(h)}, \ldots, e_{i(d)}^{(h)}$ $(h = 1, \ldots, n)$ under given $p(e_i(k) = 1)$ $(k = 1, \ldots, d)$, and obtain $D = \{V(h) | h = 1, \ldots, n\}$ through the generated BEZSAM. Similarly to the other constraint based approaches (Spirtes et al., 2000), our algorithm basically requires a complete data set $D$ where all value patterns of $X$ are included as mentioned in subsection 4.1. Accordingly, any incomplete $D$ is excluded if it is generated. A series of the model generation,
the data generation and our approach application is repeated 1000 times.

Three performance indices are used for the evaluation. Given an adjacency matrix \( B \) representing the true DAG structure in a \textit{BE}x\textit{SAM} which we generated. According to the estimated order of the variables in the output of Main Algorithm in Fig. 1: \([x_{i(k)}], k = 1, \ldots, d\), the rows and the columns of \( B \) are simultaneously permuted. Then, we compute an error rate \( 0 \leq ER_a \leq 1 \) defined by a rate of nonzero elements in the strict upper triangular part of the permuted \( B \), because the true \( B \) is strictly lower triangular. This \( ER_a \) was also used in (Shimizu et al., 2006) to evaluate the accuracy of the causal ordering. Another index is an error rate \( 0 \leq ER_s \leq 1 \) defined by a rate of mismatched elements in the estimation \( B \) against its true \( B \). \( B \) is derived from the output of Main Algorithm in Fig. 1: \([x_{i(k)}, P_{i(k)}]\)|\( k = 1, \ldots, d\). \( ER_s \) reflects the accuracy of the causal structure estimation together with the given causal order. The third is simply a total computation time \( CT \) (msec) of our algorithms explained in section 4. These indices are averaged over the 1000 trials.

In the first experiment, every combination of \( d = 2, 4, 6, 8 \) and \( n = 100, 500, 1000, 5000, 10000 \) was evaluated with the probabilities \( 0 \leq p(a_i^k = 1) \leq 1 \) and \( p(e_i(k)) = 1 \) \( \in \{0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9\} \) chosen at uniformly random for all \( k = 1, \ldots, d \). This choice of \( p(e_i(k)) = 1 \) ensures the skew and nondeterministic distribution of each \( e_i(k) \) requested by Definition 2. \( p(a_i^k = 1) \) reflects density of the variable couplings in the generated \textit{BE}x\textit{SAM}. Table 1 summarizes the performance of our approach. The experiments for \( n = 100, 500 \) and \( d = 8 \) were skipped, since \( n \) were insufficient for the completeness of \( D \). Both error rates \( ER_a \) and \( ER_s \) are less than or around 0.05 if \( n \geq 1000 \) and \( d \leq 4 \). This indicates that our approach works for a moderate data size around 1000 if we can limit the number of variables less than or equal to 4. For the case of \( d \geq 6 \), we need to collect a large number of data more than 5000. \( CT \) particularly increases along with \( d \). This is consistent with the aforementioned complexity analysis, but does not cause practical problems since the required \( CT \) is very small even for \( d = 8 \) and \( n = 10000 \).

In the second experiment, we gave an identical value to every \( p(e_i(k) = 1) \) of all \( k = 1, \ldots, d \), and evaluated \( ER_a \) and \( ER_s \) under \( d = 4 \), \( n = 1000 \) and the aforementioned random \( p(a_i^k = 1) \). Figure 4 depicts the resultant relations of \( ER_a \) and \( ER_s \) with \( p(e_i(k) = 1) \). If \( p(e_i(k) = 1) \) is close to 0 or 1, our approach loses the chance to accurately estimate the conditional probabilities and thus its accuracy is degraded. If \( p(e_i(k) = 1) \) is closed to 0.5, again the accuracy of our approach is lost, because our approach totally relies on Theorem 1 and Proposition 1 which require \( p(e_i(k) = 1) \neq 0.5 \).

Through some extra experiments, we confirmed that \( ER_a \) and \( ER_s \) do not show strong dependency on \( p(a_i^k = 1) \), \( i.e. \), the density of the \textit{BE}x\textit{SAM}. We also confirmed that a false negative rate to miss true parent variables is dominant rather than a false positive rate to find spurious parent variables. This is reasonable since the multiple comparison tests check a null hypothesis that a variable is not a parent. The experiments also showed that the computation time \( CT \) is almost independent of both \( p(a_i^k = 1) \) and \( p(e_i(k) = 1) \).

### 5.2 COMPARISON WITH OTHER ALGORITHM

Our algorithm, PC algorithm (Spirtes et al., 2000), CPC algorithm (Ramsey et al., 2006) and GES algorithm (Chickering, 2002) were compared through their applications to data generated by a following artificial \textit{BE}x\textit{SAM} forming a Y structure (Mani et al., 2006).

| \( n \) \( \backslash d \) | 2   | 4   | 6   | 8   |
|-----------------|-----|-----|-----|-----|
| 100             | 0.035 | 0.20 | 0.29 | -   |
| 500             | 0.043 | 0.20 | 0.34 | -   |
| 1000            | 0.96  | 2.6  | 7.1  | -   |
| 5000            | 0.004 | 0.033 | 0.23 | -   |
| 10000           | 0.00955 | 0.068 | 0.24 | -   |
| 100000          | 0.90  | 2.7  | 7.9  | -   |
| 0.00            | 0.017 | 0.16 | 0.15 | -   |
| 0.00            | 0.0043 | 0.041 | 0.18 | 0.25 |
| 0.00            | 0.89  | 2.8  | 8.3  | 32.0 |
| 0.00            | 0.0063 | 0.011 | 0.055 | 0.16 |
| 0.00            | 1.1   | 3.4  | 9.9  | 35.0 |
| 1.2            | 4.2   | 11.0 | 0.27 | 0.14 |
| 1.2            | 4.2   | 11.0 | 0.27 | 0.14 |

Table 1: Performance under various \( n \) and \( d \) (top: \( ER_a \), middle: \( ER_s \), and bottom: \( CT \) (msec) in each cell)
Table 2: Performance of two algorithms

|       | true \ est. | directed | no edge | undirected |
|-------|------------|----------|---------|------------|
| PC Algorithm | directed | 53       | 7       | 0          |
|        | no edge   | 7        | 173     | 0          |
| CPC Algorithm | directed | 31       | 20      | 9          |
|        | no edge   | 9        | 162     | 9          |
| GES Algorithm  | directed | 32       | 13      | 15         |
|        | no edge   | 1        | 162     | 17         |

Each $p(c_i(k) = 1)$ was given similarly to the first experiment in the previous subsection. Table 2 shows frequencies of estimated relations between variables corresponding to its columns over their true relations corresponding to its row for 20 trials. Because of the Y structure among four variables, the number of the true directed edges is $3 \times 20$ trials = 60 in total while the number of the rest true no edges are $(4 \times 3 - 3) \times 20$ trials = 180 by double counting two missed directed edges between two variables for a no edge. This counting method gives double penalties to a wrong estimation of an edge direction which often comes from causal ordering failures affecting the global structure estimation. The italics show numbers of the correct estimations. Note that this Y structure is a typical example which enables a valid estimation of the PC algorithm. However, our approach based on the skewness of the binary data distribution provides better accuracy. Similar advantageous results of our approach was obtained for the case where $x_3 = (x_1 + x_2) \oplus e_3$. The results of CPC and GES are similar with PC, since CPC and GES do not have any significant advantage to identify Y-structures comparing with PC.

5.3 EXAMPLE APPLICATIONS TO REAL-WORLD DATA

Our approach has been applied to two real-world data sets. One is on leukemia deaths/survivals ($LE = 1/0$) in children in southern Utah with high/low exposure to radiation ($EX = 1/0$) from the fallout of nuclear tests in Nevada (Finkelstein and Levin, 1990), (Pearl, 2000). As this contains only two binary variables, conventional constraints/score-based approaches cannot estimate any unique causal structure. In contrast, our approach found a causal order $EX \rightarrow LE$ consistent with our intuition, though their parent-child relation was not significant under $\alpha = 0.05$ in the multiple comparison tests.

Another data set is on college plans of 10318 Wisconsin high school seniors (Sewell and Shah, 1968), (Spirtes et al., 2000). While the original study aimed to find a feasible causal structure among five variables constituting a Y structure, we focus on the causality between three variables; yes/no college plans ($CP = 0/1$), low/high parental encouragement ($PE = 0/1$) and lowest to highest socioeconomic status ($SES = 0, \ldots, 3$). Conventional constraints-based approaches are known to give multiple candidate causal structures in an equivalence class for the three variables. We selected 3756 male seniors ($SEX = male$) having a higher intelligence quotient by excluding the lowest group ($IQ \neq lowest$) to control their individual characters while maintaining the sample size. We further transformed $SES$ to a binary variable according to the observation that the samples having its lowest value have a particular low probability of yes college plans; $p(CP = 0) = 0.21$ while the other samples have $p(CP = 0) = 0.51$. To preserve this strong dependency, it is discretized as $SES = 0$ if it is originally the lowest, and $SES = 1$ otherwise. The application of our approach with a significance level $\alpha = 0.05$ to this data set provided a unique causal structure depicted in Fig. 5. It states that the socioeconomic status of the parents affects both the parental encouragement and the college plan of a senior, and the parental encouragement further influences the college plan. This is consistent with our intuition and the structure estimated from the original five variables constituting a Y structure by the PC algorithm.

6 DISCUSSION

When we apply our approach to a data set, we presume that the data generation process approximately follows a $BExSAM$. A crucial property requested in a $BExSAM$ is the skew distribution of every external noise. Because the noise is not observable, a measure to indirectly check this property by using the given
observed data set is desirable. The following lemma can be used for this purpose.

**Lemma 1** Assuming that a data set $D$ is generated by a $BExSAM$, if a variable $x_i \in X$ has a distribution $p(x_i = 1) \neq 0.5$, then the following conditions hold.

$$p(e_i = 1) \neq 0.5 \text{ and } p(f_i = 1) \neq 0.5,$$

where $x_i = f_i \oplus e_i$.

*Proof. See Appendix 3.*

We simply check if the frequency of $x_i = 1$ is apart from 0.5 for every observed variable in $X$. If it is, the skewness of their noise distribution is ensured. Another strong assumption of a $BExSAM$ is the interventions of external binary noises via exclusive-or operations. However, this Lemma 1 also suggests the applicability of the $BExSAM$ to generic Boolean interventions of the noises. For example, if a given data is generated by $x_i = f_i + e_i$ (+ is an or operator), we have $p(x_i = 1) = p(e_i = 1) + p(f_i = 1) - p(e_i = 1)p(f_i = 1)$. On the other hand, a $BExSAM$ $x_i = f_i \oplus e_i$ provides $p(x_i = 1) = p(e_i = 1) + p(f_i = 1) - 2p(e_i = 1)p(f_i = 1)$. Accordingly, these two models show very similar distributions of the observed variables, if $p(e_i = 1), p(f_i = 1) \ll 0.5$. These conditions can be checked by the insights of Lemma 1.

As mentioned in subsection 5.1, our algorithm basically requires a complete data set $D$ similarly to the other constraint based approaches (Spirtes et al., 2000). Therefore, to analyze an incomplete data set $D$, we need to estimate the missing data in $D$ by introducing some data completion techniques such as (Bernaards et al., 2007). Another associated issue is its applicability to many variables, since the low error rates are ensured only for less than 10 variables with thousands of samples. A promising way to overcome this issue may be to combine our approach with the other constraint based approaches such as the PC algorithm as discussed in (Zhang and Hyvarinen, 2009). The extensions of our approach toward these issues are remained for future studies.

### 7 CONCLUSION

In this paper, we presented a novel binary structural model involving exclusive-or noise and proposed a new efficient approach to derive an identifiable causal structure governing a given binary data set based on the skewness of the distributions of external noises. The approach has low computational complexity and requires only one tunable parameters. The experimental evaluation shows promising performance for both artificial and real world data sets.

This study provides an extension of the non-Gaussianity-based causal inference for continuous variables to the causal inference for discrete variables, and suggests a new aspect on more generic causal inference.

### Acknowledgements

This work was partially supported by JST, ERATO, Minato Discrete Structure Manipulation System Project and JSPS Grant-in-Aid for Scientific Research(B) #22300054. The authors would like to thank Dr. Tsuyoshi Ueno, a research fellow of the JST ERATO project, for his valuable technical comments.

### References

Y. Benjamini and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, 57(1):289-300.

C.A. Bernaards, T.R. Belin and J.L. Schafer (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6):1368-1382.

C. Carlet and P. Guilhot (1999). A new representation of Boolean functions. *Proc. of 13th AAECC, Lecture Notes in Computer Science*, 1719:94-103.

D.M. Chickering (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507-554.

M.O. Finkelstein and B. Levin (1990). *Statistics for lawyers*. Springer-Verlag, New York.

J.R. Gregg (1998). *Ones and Zeros: Understanding Boolean Algebra, Digital Circuits, and the Logic of Sets*, Chap.5:101-123. Hoboken, NJ. John Wiley & Sons.

P.O. Hoyer, D. Janzing, J. Mooij, J. Peters and B. Schölkopf (2009). Nonlinear causal discovery with additive noise models. *Proc. of NIPS2008: In Advances in Neural Information Processing Systems*, 21:689-696.

S. Mani, G. Cooper and P. Spirtes (2006). A theoretical study of $Y$ structures for causal discovery. *Proc. of the 22nd Conference in Uncertainty in Artificial Intelligence*, 314-323.

J. Mooij, D. Janzing, J. Peters and B. Schölkopf (2009). Regression by dependence minimization and its application to causal inference in additive noise models. *Proc. of ICML2009: the 26th Annual International Conference on Machine Learning*, 94-101.

J. Pearl (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, UK. Cambridge University Press.
J. Peters, D. Janzing and B. Schölkopf (2011). Causal inference on discrete data using additive noise models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 61(2):282–293.

J. Ramsey, J. Zhang and P. Spirtes (2006). Adjacency-Faithfulness and Conservative Causal Inference. Proc. of UAI2006: the 22nd Conference Annual Conference on Uncertainty in Artificial Intelligence, 401-408.

W. Sewell and V. Shah (1968). Social class, parental encouragement, and educational aspirations. American Journal of Sociology, 73:559-572.

Y.Y. Shi, G.A. Miller, O. Denisenko, H. Qian and K. Bomsztyk (2007). Quantitative model for binary measurements of protein-protein interactions. Journal of Computational Biology, 14(7):1011-1023.

S. Shimizu, P.O. Hoyer, A. Hyvarinen and A. Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003-2030.

S. Shimizu, A. Hyvarinen, Y. Kawahara and T. Washio (2009). A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. Proc. of UAI2009: the 25th Conference on Uncertainty in Artificial Intelligence, Causality II & Graphical Models.

P. Spirtes, C. Glymour, and R. Scheines (2000). Causation, Prediction, and Search (Adaptive Computation and Machine Learning). Cambridge, MA. The MIT Press.

X. Sun and D. Janzing (2007). Exploring the causal order of binary variables via exponential hierarchies of Markov kernels. Proc. of ESANN2007: European Symposium on Artificial Neural Networks, 25-27.

K. Zhang and A. Hyvarinen (2009). On the identifiability of the post-nonlinear causal model. Proc. of UAI2009: the 25th Conference on Uncertainty in Artificial Intelligence, Causality II & Graphical Models, 647-655.

Appendix 1

Proof of Theorem 1

(1 \Rightarrow 2)

Let the value of $x_i$ be $v_i \in \{0, 1\}$. Under a control $X_i = V_i$, $f_i$ is a constant in $\{0, 1\}$. Accordingly, the following relation holds by Definition 1.

$$x_i = v_i \Leftrightarrow v_i = f_i \oplus e_i \Leftrightarrow e_i = v_i \oplus f_i.$$  

Because $x_i$ is a sink endogenous variable, $e_i$ and $X_i$ are mutually independent by Definition 1. Under this fact and Definition 2,

$$p(x_i = v_i | X_i = V_i) = p(v_i = e_i \oplus f_i) = \begin{cases} p_i, & \text{for } v_i \oplus f_i = 1 \\ 1 - p_i, & \text{for } v_i \oplus f_i = 0 \end{cases}$$  

By letting $q_i = p_i$ or $q_i = 1 - p_i$, $1 \Rightarrow 2$ holds.

(1 \Rightarrow 2)

Assume that $x_i$ is not a sink endogenous variable. Let $X_i$ be partitioned into $X_i^u$ and $X_i^vu$ where $X_i^u$ is a set of all descendants of $x_i$ in a BExSAM and $X_i^vu$ is the complement of $X_i^u$ in $X_i$. Then, the following holds.

$$p(x_i = v_i | X_i = V_i) = p(x_i = v_i | X_i^u = V_i^u, X_i^vu = V_i^vu) = \sum_{v' \in \{v_i, e_i\}} p(x_i = v_i | V_i^u, X_i^vu = V_i^vu)p(x_i = v_i | V_i^u = V_i^u, X_i^vu = V_i^vu) = (1).$$  

where $v_i = v_i \oplus 1$. Furthermore, let $X_j = X \setminus \{x_j\}$ of $x_j \in X_i^u$ be partitioned into $X_j^u$ and $X_j^vu$ similarly to $X_i^u$ and $X_i^vu$ of $x_i$. By Definition 1, each $f_j$ of $x_j \in X_j^u$ is given by $X_j^u = V_j^u$, and thus $x_j = f_j(X_j^u = V_j^u) \oplus e_j$ for all $x_j \in X_j$. Accordingly, $X_i^vu$ is equivalent to $e_j = v_j \oplus f_j(X_j^u = V_j^u)$ for all $x_j \in X_j$. We rewrite the r.h.s.: $v_j \oplus f_j(X_j^u = V_j^u)$ as $l_j(v_j)$, since $x_j$ is an ancestor of $x_j$, i.e., $x_j \in X_j^u$ and the values of all variables except $x_i = v_i$ are constant under a control $X_i = V_i$. Because every $e_j$ is independent of its upper variables,

$$p(X_i^u = V_i^u | x_i = v_i, X_i^vu = V_i^vu) = \prod_{x_j \in X_j^u} p(e_j = l_j(v_j)).$$  

$x_i$ has at least one child $x_h \in X_i^u$ where $l_h(1) \neq l_h(0)$ for some control $X_i = V_i$ from the assumption that $x_i$ is not a sink endogenous variable. Accordingly,

$$p(X_i^u = V_i^u | x_i = v_i, X_i^vu = V_i^vu) = (1) \Pi_{x_j \in X_j^u} p(e_j = l_j(v_j)), \quad \text{for } l_h(v_i) = 1; \quad \text{for } l_h(v_i) = 0. \quad \text{(2)}$$  

On the other hand, since $e_i$ is independent of $X_i^u$ and $x_i = v_i \Leftrightarrow e_i = v_i \oplus f_i,$

$$p(x_i = v_i | X_i^u = V_i^u) = p(e_i = v_i \oplus f_i) = \begin{cases} p_i, & \text{for } v_i \oplus f_i = 1 \\ 1 - p_i, & \text{for } v_i \oplus f_i = 0 \end{cases} \quad \text{(3)}$$  

By substituting Eq.(2) and (3) to Eq.(1), we obtain the following four cases.

$$p(x_i = v_i | X_i = V_i) = \frac{\alpha_{p_i}}{\alpha_{p_i} + \beta(1-p_i)(1-p_h)} \quad \text{(4)}$$  

$$\frac{\alpha_{p_i}}{\alpha_{p_i} + \beta(1-p_i)(1-p_h)} \quad \text{(5)}$$  

$$\frac{\alpha_{p_i}(1-p_h)}{\alpha_{1-p_i} + \beta p_i(1-p_h)} \quad \text{(6)}$$  

$$\frac{\alpha_{1-p_i}}{\alpha_{p_i} + \beta(1-p_i)(1-p_h)} \quad \text{(7)}$$  

where $\alpha = \prod_{x_j \in X_j^u, x_j \neq x_h} p(e_j = l_j(v_j))$ and $\beta = \prod_{x_j \in X_j^u, x_j \neq x_h} p(e_j = l_j(v_j))$. $\alpha$ and $\beta$ are nonzero.
from Definition 2, and each of (t4) = (t5), (t4) = (t6), (t5) = (t7) and (t6) = (t7) for any $\alpha$ and $\beta$, i.e., any $X_i^t = V_i^t$ excluding $x_h$, implies $p_i = 1/2$ or $p_h = 1/2$ respectively. Because neither of $p_i \neq 1/2$ nor $p_h \neq 1/2$ are allowed by Definition 2, this implies that all conditions (t4) $\neq$ (t5), (t4) $\neq$ (t6), (t5) $\neq$ (t7) and (t6) $\neq$ (t7) hold simultaneously under some $X_i^t = V_i^t$ excluding $x_h$. If we assume (t4) = (t7) and (t5) = (t6) simultaneously under such $X_i^t = V_i^t$ excluding $x_h$, they imply $p_i^2 p_h^2 = (1 - p_i)^2 (1 - p_h)^2$ and $p_i^2 (1 - p_h)^2 = (1 - p_i)^2 p_h^2$, i.e., $p_i = 1/2$ and $p_h = 1/2$. Accordingly, $p_i \neq 1/2$ and $p_h \neq 1/2$ from Definition 2 imply that one of (t4) = (t7) and (t5) = (t6) do not hold under the $X_i^t = V_i^t$ excluding $x_h$. This result shows that $p(x_i = v_i | X_i = V_i)$ takes more than two values for some given control $X_i = V_i$ if $x_i$ is not a sink endogenous variable. By taking its contrapositive, we obtain $1 \Leftarrow 2$. \hfill \Box

Appendix 2

Lemma 2 Let $x_i \in X$ be a sink endogenous variable. For a variable $x_j \in X$ ($j \neq i$), the followings hold under a control of $X_{ij} = X \setminus \{x_i, x_j\}$ at $V_{ij} \in \{0, 1\}^{d-2}$.

1. $f_i$ depends on $x_j$ under $X_{ij} = V_{ij}$, $\Leftrightarrow$
   \[
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) = p(x_i = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}),
   \]

2. $f_i$ does not depend on $x_j$ under $X_{ij} = V_{ij}$, $\Leftrightarrow$
   \[
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) = p(x_i = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}),
   \]

where $v_\ast \in \{0, 1\}$ and $\bar{v}_\ast = v_\ast \oplus 1$.

Proof. $x_i = f_i \oplus e_i$ from Definition 1, in addition, $e_i$ and $X_{ij} = V_{ij}$ are mutually independent since $x_i$ is a sink endogenous variable. Accordingly,

\[
\begin{align*}
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) &= (1 - p_i) p(f_i = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(f_i = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= (1 - p_i) p(f_i = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(f_i = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= (1 - p_i) p(f_i = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(f_i = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}).
\end{align*}
\]

Since $f_i$ is a deterministic Boolean function, it is always represented by a Boolean algebraic formula. If $f_i$ depends on $x_j$ under $X_{ij} = V_{ij}$, $f_i = c \oplus x_j$ where $c$ is a constant in $\{0, 1\}$. From Eq. (11),

\[
\begin{align*}
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) &= (1 - p_i) p(c \oplus v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(c \oplus \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= (1 - p_i) p(c \oplus v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(c \oplus \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= \{ 1 - p_i, \text{ for } c \oplus v_i = v_i \Leftrightarrow c \oplus \bar{v}_i = \bar{v}_i \\
   &\quad \quad \text{ and } c \oplus \bar{v}_i = \bar{v}_i \Leftrightarrow c \oplus v_i = v_i,
\end{align*}
\]

Therefore,

\[
\begin{align*}
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) &= (1 - p_i) p(c = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(c = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= (1 - p_i) p(c = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(c = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}).
\end{align*}
\]

\[
\begin{align*}
   p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij}) &= (1 - p_i) p(c = v_i | x_j = v_j, X_{ij} = V_{ij}) + p_i p(c = \bar{v}_i | x_j = v_j, X_{ij} = V_{ij}) \\
   &= \{ 1 - p_i, \text{ for } c = v_i \\
   &\quad \quad \text{ and } c = \bar{v}_i \Leftrightarrow c \oplus v_i = v_i.
\end{align*}
\]

Therefore, $p(x_i = v_i | x_j = v_j, X_{ij} = V_{ij})$ does not depend on $v_j$.

Proof of Lemma 1

Without loss of generality, $p(e_i = 1)$, $p(e_i = 0)$, $p(f_i = 1)$, $p(f_i = 0)$ are represented as

\[
\begin{align*}
   p(e_i = 1) &= 1 + \epsilon_i, \quad p(e_i = 0) = 1 - \epsilon_i \quad (-1 \leq \epsilon_i \leq 1), \\
   p(f_i = 1) &= 1 + \xi_i, \quad p(f_i = 0) = 1 - \xi_i \quad (-1 \leq \xi_i \leq 1).
\end{align*}
\]

Because $e_i$ and $f_i$ are mutually independent, and have a relation $x_i = f_i \oplus e_i$, $p(x_i = 1) = p(e_i = 1) p(f_i = 1) + p(e_i = 0) p(f_i = 1) = (1 + \epsilon_i) (1 - \xi_i) + (1 - \epsilon_i) (1 + \xi_i) = 1 - \epsilon_i \xi_i$. Therefore,

\[
\begin{align*}
   p(x_i = 0) &= p(e_i = 1) p(f_i = 1) + p(e_i = 0) p(f_i = 0) \\
   &= (1 + \epsilon_i) (1 - \xi_i) + (1 - \epsilon_i) (1 + \xi_i) \\
   &= 1 + \epsilon_i \xi_i.
\end{align*}
\]

Thus, $p(x_i = 0) - p(x_i = 1) = \epsilon_i \xi_i$. Accordingly, if $p(x_i = 1) \neq p(x_i = 0)$ then $\epsilon_i \xi_i \neq 0$. This implies $\epsilon_i \neq 0$ and $\xi_i \neq 0$. Therefore, if $p(x_i = 1) \neq 0.5$ then $p(e_i = 1) \neq 0.5$ and $p(f_i = 1) \neq 0.5$ because of $p(x_i = 0) = 1 - p(x_i = 1)$, $p(e_i = 0) = 1 - p(e_i = 1)$ and $p(f_i = 0) = 1 - p(f_i = 1)$. \hfill \Box