7-2020

Learning Networks with Categorical Data using Distance Correlation, and A Novel Graph-Based Multivariate Test

Jian Tinker
University of Arkansas, Fayetteville

Learning Networks with Categorical Data using Distance Correlation, and A Novel Graph-Based Multivariate Test

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Mathematics

by

Jian Tinker
China University of Geosciences, Beijing
Bachelor of Science in Mathematics and Applied Mathematics, 2007
University of Arkansas
Master of Science in Mathematics, 2013
University of Arkansas
Master of Science in Statistics, 2014

July 2020
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Qingyang Zhang, Ph.D.
Dissertation Director

Avishek Chakraborty, Ph.D.
Committee Member

Jyotishka Datta, Ph.D.
Committee Member

Abstract

We study the use of distance correlation for statistical inference on categorical data, especially the induction of probability networks. Szekely et al. first defined distance correlation for continuous variables in [42], and Zhang translated the concept into the categorical setting in [57] by defining $\mathrm{dCor}(\boldsymbol{X}, \boldsymbol{Y})$ for categorical variables $X = (x_1, \ldots, x_I)$ and $Y = (y_1, \ldots, y_J)$ where $P(X = x_i) = \pi_i$ and $P(Y = y_j) = \pi_j$ with the formula,

$$\frac{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\pi_{ij}-\pi_{i+}\pi_{+j}\right)^2}}{\left\{\sum_{i=1}^{I}\pi_{i+}^2\left(\sum_{i=1}^{I}\pi_{i+}^2+1\right)-2\sum_{i=1}^{I}\pi_{i+}^3\right\}^{1/4}\left\{\sum_{j=1}^{J}\pi_{+j}^2\left(\sum_{j=1}^{J}\pi_{+j}^2+1\right)-2\sum_{j=1}^{J}\pi_{+j}^3\right\}^{1/4}}.$$

Part I of the dissertation covers the background we need to understand this formula, and prepares us to analyze the properties and performance of its applications.

Part II then presents the main results of the dissertation, applying distance correlation to learn the structure of probability networks with categorical nodes. We cover in detail how the distance correlation measure may be combined with search methods based on graphical models to induce network structure. This leads to our empirical results obtained by enhancing the INeS software library [6]. These results involve experiments using six data sets such as the Danish Jersey cattle blood type determination data and the ALARM network; in terms of accuracy metrics such as edges missed from the true network, induction with distance correlation achieves higher accuracy relative on average than does induction with existing measures such as mutual information and χ^2. We conclude Part II by connecting to earlier joint work with Zhang in [58] on the use of conditional distance covariance for conditional independence and homogeneity tests in large sparse three-way tables. The simulation studies in this work offer another source of intuition for why distance correlation may be able to recover network structure more accurately than traditional measures.

In Part III, we end the dissertation by discussing another application of graphical models, in this case to the derivation of a graph-based multivariate test. The test statistic is computationally cheap, and proven to converge to a χ^2 distribution with favorable asymptotics. We present empirical results in which we use the test to analyze the roles of various oncogenic and suppressor pathways in tumor progression.

Acknowledgements

I have the deepest appreciation for all of the extraordinary support and encouragement that Dr. Qingyang Zhang has provided me during my work on this dissertation.

Dedication

To Michael.

Table of Contents

I BACKGROUND

II LEARNING NETWORKS WITH DISTANCE CORRELATION

III A NOVEL GRAPH-BASED MULTIVARIATE TEST

Part I

Background

Chapter 1

Fundamentals of Distance Correlation

In [42], Szekely et al. introduced the concept of distance correlation in the continuous setting. They wished to overcome some of the weaknesses of traditional correlation measures; for example, detecting nonlinear dependencies or maintaining power in multivariate independence tests. In this chapter we review the continuous theory and then discuss Zhang's translation to the categorical setting from [57].

1.1 The continuous case

Distance correlation is a measure of dependence between random vectors, analogous to product-moment correlation; but unlike the classical notion, it is zero only if the random vectors are independent. Its empirical measures are based on Euclidean distance computed between sample elements rather than sample moments. Nonetheless, these empirical measures retain a compact representation analogous to classical covariance and correlation measures. We will consider their asymptotic properties and usefulness for independence tests.

Notation

In what follows, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are random vectors, where p and q are positive integers. The characteristic functions of X and Y are denoted f_X and f_Y, respectively, and the joint characteristic function of X and Y is denoted $f_{X,Y}$. The scalar product of vectors t and s is denoted by $\langle t, s \rangle$.

For complex-valued functions $f(\cdot)$, the complex conjugate of f is denoted by $\bar{f}$ and $|f|^2 = f\bar{f}$. The Euclidean norm of x in $\mathbb{R}^p$ is $|x|_p$. A sample from the distribution of X in $\mathbb{R}^p$ is denoted by the $n \times p$ matrix X, and the sample vectors (rows) are labeled $X_1, \ldots, X_n$.

A primed variable X' is an independent copy of X; that is, X and X' are independent and identically distributed.

Motivation and introduction

Consider the problem of testing the joint independence of random vectors. For all distributions with finite first moments, we seek a dependency measure $\mathcal{R}$ such that:

1. $\mathcal{R}(X, Y)$ is defined for X and Y in arbitrary dimension;

2. $\mathcal{R}(X, Y) = 0$ characterizes independence of X and Y.

We will see that distance correlation has these properties; in particular, distance correlation satisfies $0 \leq \mathcal{R} \leq 1$ and $\mathcal{R} = 0$ only if X and Y are independent. Concretely, in the bivariate normal case, $\mathcal{R}$ is a function of product-moment correlation ρ, and $\mathcal{R}(X, Y) \leq |\rho(X, Y)|$ with equality when $\rho = \pm 1$.

We also wish a dependency measure $\mathcal{R}$ to reflect the distance $\|f_{X,Y}(t, s) - f_X(t) f_Y(s)\|$ between the joint characteristic function and the product of the marginal characteristic functions; hence allowing a powerful independence test for the null and alternative hypotheses,

$$H_0 : f_{X,Y} = f_X f_Y \quad \text{vs.} \quad H_1 : f_{X,Y} \neq f_X f_Y.$$

The empirical importance of testing independence assumptions is hard to overstate. For example, consider clinical studies on gene interactions which use a *case-only design*; that is, which use only diseased subjects which are assumed independent in the study population. In this design, inferences on multiplicative gene interactions can be highly distorted when there is a departure from independence.

As we will see, distance correlation also performs well here. The power of its associated independence tests is a primary benefit, as Monte Carlo results on distance covariance tests exhibit superior power against non-monotone types of dependence while maintaining good power performance in the multivariate normal case (say, relative to the parametric likelihood ratio test). Distance correlation can also be applied as an index of dependence; for example, meta-analysis suggests distance correlation could be a more generally applicable index than product-moment correlation—without requiring normality for valid inferences.

With these promises in mind, let us now define distance correlation.

Derivation of the distance correlation measure

We need a few preparatory definitions following Szekely et al. in [42].

Definition 1.1. *For complex functions γ defined on $\mathbb{R}^p \times \mathbb{R}^q$ the $\|\cdot\|_w$-norm in the weighted L_2 space of functions on $\mathbb{R}^{p+q}$ is defined by*

$$\|\gamma(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 \, w(t,s) \, dt \, ds,$$

where $w(t,s)$ is an arbitrary positive weight function for which the integral above exists.

Now, for any acceptable choice of weight $w(t,s)$, we may use the $\|\cdot\|_w$-norm to define a measure of dependence.

Definition 1.2. *Given characteristic functions f_X, f_Y, and $f_{X,Y}$ with weight $w(t,s)$ we define the measure $\mathcal{V}^2(X,Y;w)$ by*

$$\mathcal{V}^2(X,Y;w) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 \, w(t,s) \, dt \, ds.$$

In particular, $\mathcal{V}^2(X,Y;w)$ vanishes iff X and Y are independent.

We note that $\mathcal{V}$ is analogous to the absolute value of the classical product-moment covariance. If we divide $\mathcal{V}(X,Y;w)$ by $\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}$ where

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 \, w(t,s) \, dt \, ds,$$

we have a type of unsigned correlation $\mathcal{R}_w$. Of course, not every weight function leads to an "interesting" $\mathcal{R}_w$. The coefficient $\mathcal{R}_w$ should be *scale invariant*; that is, invariant with respect to transformations $(X,Y) \mapsto (\epsilon X, \epsilon Y)$, for $\epsilon > 0$. We also require that $\mathcal{R}_w$ is positive for dependent variables.

One can show that if the weight function $w(t,s)$ is integrable and both X and Y have finite variance, then by Taylor expansions of the underlying characteristic functions,

$$\lim_{\epsilon \to 0} \frac{V^2(\epsilon X, \epsilon Y; w)}{\sqrt{\mathcal{V}^2(\epsilon X; w) V^2(\epsilon Y; w)}} = \rho^2(X,Y),$$

thus for integrable w, if $\rho = 0$, then $\mathcal{R}_w$ can be arbitrarily close to zero even if X and Y are dependent.

However, by applying a nonintegrable weight function, it is possible to obtain an $\mathcal{R}_w$ that is scale invariant and cannot be zero for dependent X and Y. This is the key insight in [42], and it leads to very simple and applicable empirical formulas. The crucial observation is the following lemma.

Lemma 1.3. *If $0 < \alpha < 2$, then for all x in $\mathbb{R}^d$*

$$\int_{\mathbb{R}^d} \frac{1 - \cos(t, x)}{|t|_d^{d+\alpha}} dt = C(d, \alpha)|x|^\alpha,$$

where

$$C(d, \alpha) = \frac{2\pi^{d/2}\Gamma(1 - \alpha/2)}{\alpha 2^\alpha \Gamma((d + \alpha)/2)},$$

and $\Gamma(\cdot)$ is the complete gamma function. The integrals at 0 and ∞ are meant in the principal value sense: $\lim_{\epsilon\to 0} \int_{\mathbb{R}^d\{\epsilon B+\epsilon^{-1}B^c\}}$, where B is the unit ball (centered at 0) in $\mathbb{R}^d$ and B^c is the complement of B.

In the simplest case, $\alpha = 1$, the constant in Lemma 1.3 is

$$c_d = C(d, 1) = \frac{\pi^{(1+d/2\big)}}{\Gamma((1 + d)/2)}.$$

In view of Lemma 1.3, it is natural to choose the weight function corresponding to $\alpha = 1$.

$$w(t, s) = \left(c_p c_q |t|_p^{1+p} |s|_q^{1+q}\right)^{-1}. \tag{1.1}$$

Now, given the weight function 1.1 and the corresponding weighted L_2 norm $\|\cdot\|$, omitting the index w, we write the implied dependence measure from 1.2 as $\mathcal{V}^2(X, Y)$. Also, for conciseness let us write,

$$d\omega = \left(c_p c_q |t|_p^{1+p} |s|_q^{1+q}\right)^{-1} dt\, ds.$$

Then we are studying the integral,

$$\mathcal{V}^2(X, Y) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 \, d\omega. \tag{1.2}$$

For finiteness of 1.2, it is sufficient that $E|X|_p < \infty$ and $E|Y|_q < \infty$. By the Cauchy-Bunyakovsky inequality

$$\begin{aligned} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 &= \left[E\left(e^{i(t,X)} - f_X(t)\right)\left(e^{i(s,Y)} - f_Y(s)\right)\right]^2 \\ &\leq E\left[e^{i(t,X)} - f_X(t)\right]^2 E\left[e^{i(s,Y)} - f_Y(s)\right]^2 \\ &= \left(1 - |f_X(t)|^2\right)\left(1 - |f_Y(s)|^2\right). \end{aligned}$$

If $E\left(|X|_p + |Y|_q\right) < \infty$, then by an application of Fubini's theorem it follows that

$$\begin{aligned} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 \, d\omega &\leq \int_{\mathbb{R}^p} \frac{1 - |f_X(t)|^2}{c_p|t|_p^{1+p}} dt \int_{\mathbb{R}^q} \frac{1 - |f_Y(s)|^2}{c_q|s|_q^{1+q}} \, ds \\ &= E\left[\int_{\mathbb{R}^p} \frac{1 - \cos(t, X - X')}{c_p|t|_p^{1+p}} dt\right] \cdot E\left[\int_{\mathbb{R}^q} \frac{1 - \cos(s, Y - Y')}{c_q|s|_q^{1+q}} ds\right] \\ &= E\left|X - X'\right|_p E\left|Y - Y'\right|_q < \infty. \end{aligned}$$

This leads us to the following definition.

Definition 1.4. *The distance covariance (dCov) between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{V}(X,Y)$ defined by*

$$\begin{aligned} \mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} \, dt \, ds. \end{aligned}$$

Similarly, distance variance (dVar) is defined as the square root of

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2.$$

Note that this definition can be extended to random vectors $E\left(|X|_p + |Y|_q\right) = \infty$ as long as $E\left(|X|_p^\alpha + |Y|_q^\alpha\right) < \infty$ for some $0 < \alpha < 1$, in which case one considers $\mathcal{V}^{(\alpha)}$ and $\mathcal{R}^{(\alpha)}$. In other cases, it may be possible to find a suitable transformation of (X,Y) into bounded random variables $(\tilde{X}, \tilde{Y})$ such that $\tilde{X}$ and $\tilde{Y}$ are independent iff X and Y are independent. These adaptations can allow statistical analysis to proceed using distance covariance even when the first moments are not finite.

Definition 1.5. *The distance correlation (dCor) between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{R}(X,Y)$ defined by*

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases}$$

Clearly the definition of $\mathcal{R}$ suggests an analogy with the product moment correlation coefficient ρ. In fact we can compute an explicit relation between $\mathcal{V}$, $\mathcal{R}$, and ρ in the bivariate normal case. We now move to the empirical dependence measures.

Definition 1.6. *For an observed random sample $(X,Y) = \{(X_k, Y_k) : k = 1,\dots,n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, define,*

$$a_{kl} = |X_k - X_l|_p, \quad \bar{a}_{k.} = \tfrac{1}{n}\textstyle\sum_{l=1}^n a_{kl}, \quad \bar{a}_{.l}, = \tfrac{1}{n}\textstyle\sum_{k=1}^n a_{kl},$$
$$\bar{a}_{..} = \tfrac{1}{n^2}\textstyle\sum_{k,l=1}^n a_{kl}, \quad A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..},$$

where k, $l = 1,\dots,n$. Similarly, define $b_{kl} = |Y_k - Y_l|_q$ and $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$. Then the empirical distance covariance $\mathcal{V}_n(X,Y)$ is the nonnegative number defined by

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{k,\,l=1}^{n} A_{kl}B_{kl}.$$

Similarly, $\mathcal{V}_n(X)$ is the nonnegative number defined by

$$V_n^2(X) = \mathcal{V}_n^2(X,X) = \frac{1}{n^2}\sum_{k,\,l=1}^{n} A_{kl}^2.$$

Although it may not be immediately obvious, it is a fact that $\mathcal{V}_n^2(X,Y) \geq 0$.

Definition 1.7. *The empirical distance correlation $\mathcal{R}_n(X,Y)$ is defined by*

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0 \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0. \end{cases}$$

Note that the statistic $\mathcal{V}_n(X) = 0$ iff every sample observation is identical. Indeed, if it holds that $\mathcal{V}_n(X) = 0$, then $A_{kl} = 0$ for all k, $l = 1,\dots,n$. In particular, $A_{kk} = \bar{a}_{k.} - \bar{a}_{.k} + \bar{a}..$ vanishes, implying that $\bar{a}_{k.} = \bar{a}_{.k} = \bar{a}../2$; and $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..} = a_{kl} = |X_k - X_l|_p$ so $X_1 = \cdots = X_n$. It is immediately clear that ease of computation is a major attraction of the statistic $\mathcal{R}_n$. The harder work is to show that $\mathcal{R}_n$ is *also* a good empirical measure of dependence.

1.2 Properties of distance covariance

It is interesting to note that it would have been natural (though less elementary) to define $\mathcal{V}_n(X,Y)$ as $\left\| f^n_{X,Y}(t,s) - f^n_X(t) f^n_Y(s) \right\|$, where

$$f^n_{X,Y}(t,s) = \frac{1}{n}\sum_{k=1}^{n} \exp\left\{ i\langle t, X_k\rangle + i\langle s, Y_k\rangle \right\},$$

is the empirical characteristic function of the sample, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and

$$f^n_X(t) = \frac{1}{n}\sum_{k=1}^{n} \exp\left\{ i\langle t, X_k\rangle \right\}, \quad f^n_Y(s) = \frac{1}{n}\sum_{k=1}^{n} \exp\left\{ i\left\langle s, Y_k\right\rangle \right\},$$

are the marginal empirical characteristic functions of the X sample and Y sample, respectively. It is a fundamental fact that the two definitions are equivalent, which we now prove to give a sense of the theory in the continuous setting.

Theorem 1.8. *If (X, Y) is a sample from the joint distribution of (X, Y), then*

$$\mathcal{V}^2_n(X,Y) = \left\| f^n_{X,Y}(t,s) - f^n_X(t) f^n_Y(s) \right\|^2.$$

Proof. Lemma 1.3 implies that there exist constants c_p and c_q such that for all X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$,

$$\int_{R^p} \frac{1 - \exp\{i\langle t, X\rangle\}}{|t|_p^{1+p}}\, dt = c_p\, |X|_p, \tag{1.3}$$

$$\int_{R^q} \frac{1 - \exp\{i\langle s, Y\rangle\}}{|s|_q^{1+q}}\, ds = c_q\, |Y|_q, \tag{1.4}$$

$$\int_{R^p}\int_{R^q} \frac{1 - \exp\{i\langle t, X\rangle + i\langle s, Y\rangle\}}{|t|_p^{1+p}|s|_q^{1+q}}\, dt\, ds = c_p\, c_q\, |X|_p\, |Y|_q, \tag{1.5}$$

where the integrals are understood in the principal value sense.

For simplicity, consider the case $p = q = 1$. In this case, the distance between the empirical characteristic functions in the weighted norm $w(t,s) = \pi^{-2}t^{-2}s^{-2}$ involves $\left| f^n_{X,Y}(t,s) \right|^2$, $\left| f^n_X(t) f^n_Y(s) \right|^2$ and $\overline{f^n_{X,Y}(t,s)} f^n_X(t) f^n_Y(s)$. For the first we have

$$f^n_{X,Y}(t,s) \cdot \overline{f^n_{X,Y}(t,s)} = \frac{1}{n^2}\sum_{k,l=1}^{n} \cos\left(X_k - X_l\right) t \cos\left(Y_k - Y_l\right) s + V_1,$$

where V_1 represents terms that vanish when the integral $\left\| f_{X,Y}^n(t,s) - f_X^n(t) f_Y^n(s) \right\|^2$ is evaluated. The second expression is

$$f_X^n(t) f_Y^n(s) \cdot \overline{f_X^n(t) f_Y^n(s)} = \frac{1}{n^2} \sum_{k,l=1}^{n} \cos\left(X_k - X_l\right)\, t \ \cdot \ \frac{1}{n^2} \sum_{k,l=1}^{n} \cos\left(Y_k - Y_l\right)\, s + V_2,$$

and the third is

$$f_{X,Y}^n(t,s) \cdot \overline{f_X^n(t) f_Y^n(s)} = \frac{1}{n^3} \sum_{k,l,m=1}^{n} \cos\left(X_k - X_l\right) t \cos\left(Y_k - Y_m\right) s + V_3,$$

where V_2 and V_3 represent terms that vanish when the integral is evaluated.

Now, to evaluate the integral $\left\| f_{X,Y}^n(t,s) - f_X^n(t) f_Y^n(s) \right\|^2$ in this special case, we can apply Lemma 1.3 and Equations 1.3, 1.4, and 1.5 using the identity

$$\cos u \cos v = 1 - (1 - \cos u) - (1 - \cos v) + (1 - \cos u)(1 - \cos v).$$

After cancellation, the remaining integrals then evaluate as,

$$\begin{aligned} &\textstyle\int_{\mathbb{R}^2} \left(1 - \cos\left(X_k - X_l\right) t\right) \left(1 - \cos\left(Y_k - Y_l\right) s\right) \frac{dt}{t^2} \frac{ds}{s^2} \\ &\quad = \textstyle\int_{\mathbb{R}} \left(1 - \cos\left(X_k - X_l\right) t\right) \frac{dt}{t^2} \times \int_{\mathbb{R}} \left(1 - \cos\left(Y_k - Y_l\right) s\right) \frac{ds}{s^2} \\ &\quad = c_1^2 \left|X_k - X_l\right| \left|Y_k - Y_l\right|. \end{aligned}$$

For random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, we apply the same steps, except now using $w(t,s) = \left\{ c_p c_q |t|_p^{1+p} |s|_q^{1+q} \right\}^{-1}$.

We are left with

$$\left\| f_{X,Y}^n(t,s) - f_X^n(t) f_Y^n(s) \right\|^2 = S_1 + S_2 - 2S_3, \tag{1.6}$$

where

$$S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} \left|X_k - X_l\right|_p \left|Y_k - Y_l\right|_q, \tag{1.7}$$

$$S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} \left|X_k - X_l\right|_p \frac{1}{n^2} \sum_{k,l=1}^{n} \left|Y_k - Y_l\right|_q, \tag{1.8}$$

$$S_3 = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} \left|X_k - X_l\right|_p \left|Y_k - Y_m\right|_q. \tag{1.9}$$

To complete the proof we only need the algebraic identity $V_n^2(X,Y) = S_1 + S_2 - 2S_3$, as worked out in the appendix of [42]. Taken with 1.6, it follows $\mathcal{V}_n^2(X,Y) = \left\| f_{X,Y}^n(t,s) - f_X^n(t) f_Y^n(s) \right\|^2$.

□

With this equivalence in hand, it is relatively straightforward to prove the following key properties of distance covariance and correlation. (See [42] for the proofs.)

Theorem 1.9 (Properties of distance covariance and distance correlation)**.**

(i) *If $E(|X|_p + |Y|_q) < \infty$, then $0 \leq \mathcal{R} \leq 1$, and $\mathcal{R}(X,Y) = 0$ iff X and Y are independent.*

(ii) *If it also holds $E(|X|_p^2 + |Y|_q^2) < \infty$, then given three independent samples we have*

$$\mathcal{V}^2(X,Y) = E\left(|X_1 - X_2|_p |Y_1 - Y_2|_q\right) + E\left(|X_1 - X_2|_p\right) E\left(|Y_1 - Y_2|_q\right) \; - 2E\left(|X_1 - X_2|_p |Y_1 - Y_3|_q\right). \qquad (1.10)$$

1.3 Power comparison with traditional independence tests

Although the theoretical properties of distance correlation and the ease of computing empirical distance correlation are clearly attractive, we of course also want evidence that associated tests perform well in practice. Szekely et al. proposed the following independence test.

Definition 1.10. *With notation as above, let $T(X,Y,\alpha,n)$ be the test that rejects independence of X and Y if*

$$\frac{n\mathcal{V}_n^2(X,Y)}{S_2} > (\phi^{-1}(1-\alpha/2))^2,$$

where $\phi(\cdot)$ is the standard normal cumulative distribution function, and S_2 is from Equation 1.8.

They performed Monte Carlo power comparisons of this test against three classical tests of multivariate independence based on likelihood ratios: the Wilks Lambda statistic, the Puri–Sen rank correlation statistic, and the Puri–Sen sign statistic. The empirical powers of the various tests were comparable for X and Y with different multivariate normal or t distributions.

But when considering X multivariate normal with $p = 5$, and $Y_{kj} = \log(X_{kj}^2)$, the simulations showed the test in Definition 1.10 to be much more powerful. That is to say, given two random vectors with nonlinear relations, traditional measures struggled to detect the dependence structure, while distance correlation remained sensitive.

1.4 Translation to categorical variables

Given these results in the continuous setting, it is of interest to examine applications to categorical random variables as well. Zhang took this step in [57], observing that categorical variables

$X \in 1, 2, \ldots, I$ and $Y \in 1, 2, \ldots, J$ can be represented as random vectors of dimension I and J, respectively, where

$$X = I(X = i)_{1 \leq i \leq I}, \quad Y = I(Y = j)_{1 \leq j \leq J},$$

with $I(\cdot)$ the indicator function. Then under Euclidean distance, $\|X_1 - X_2\|$ vanishes if $X_1 = X_2$ and is $\sqrt{2}$ otherwise. Since the norms of such vectors of course have finite second moments, the second equivalence in Theorem 1.9 applies, and this gives us a path to computing distance covariance and distance correlation in the categorical setting.

Geometric intuition for inter-category distance

To make the intuition here explicit, we mention a geometric interpretation given by Vernizzi and Nakai in [25]. In particular, a categorical variable with K possible outcomes can be represented by K equidistant points $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ with coordinates,

$$\begin{aligned}
\mathbf{v}_1 &= (1, 0, 0, \ldots, 0), \\
\mathbf{v}_2 &= (0, 1, 0, \ldots, 0), \\
&\cdots \\
\mathbf{v}_K &= (0, 0, 0, \ldots, 1).
\end{aligned}$$

All these vertices belong to the K-dimensional space $\mathbb{R}^K$, and they span a $(K-1)$-dimensional convex hull called a regular $(K-1)$-simplex.

For instance, a categorical variable representing binary sex has two possible outcomes that correspond to the two points $\mathbf{v}_1 = (1, 0)$ and $\mathbf{v}_2 = (0, 1)$ on the plane. The segment connecting the two points is a regular one-simplex, and its length is $l = \sqrt{2}$. As another example, a categorical variable associated with employment status might take one of four possible outcomes: full-time, part-time, self-employed, and unemployed. It can be represented by four points $\mathbf{v}_1 = (1, 0, 0, 0), \mathbf{v}_2 = (0, 1, 0, 0), \mathbf{v}_3 = (0, 0, 1, 0)$, and $\mathbf{v}_4 = (0, 0, 0, 1)$ in $\mathbb{R}^4$. The three-dimensional hull spanned by the four vertices is a regular tetrahedron, as in Figure 1.1 below from [25].

Evidently all simplices have edge length $l = \sqrt{2}$, and this geometric interpretation makes it explicit that the Euclidean distance between any pair of vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ of a simplex (and

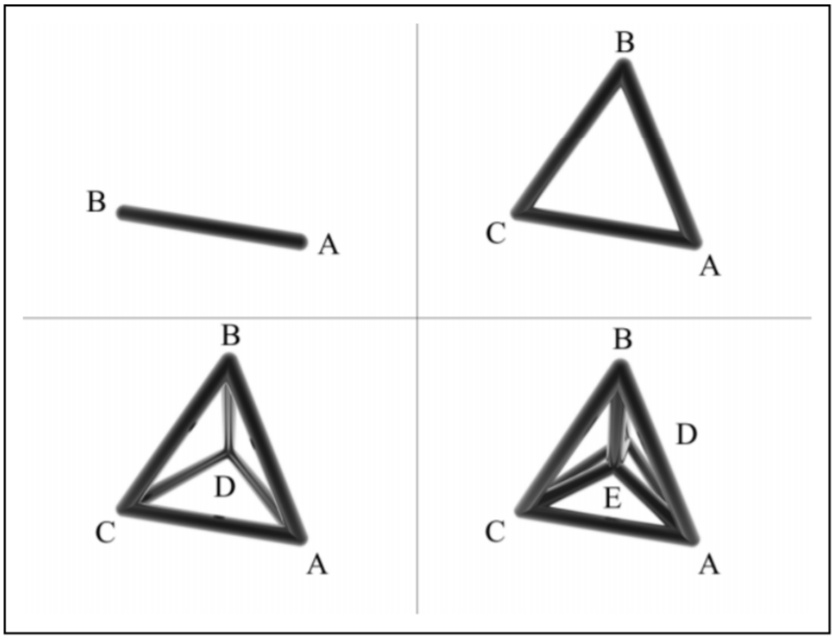


Figure 1.1: Regular simplices representing categorical variables with one to four outcomes

equivalently, between any two outcomes for the underlying categorical variable) is given by $d_{ij} = |\mathbf{v}_i - \mathbf{v}_j| = (1 - \delta_{ij})\sqrt{2}$, where δ_{ij} is the Kronecker delta function.

Definitions for categorical variables

Now returning to Zhang's work, using a multinomial sampling scheme for the underlying variables X and Y, he derived in particular,

$$\mathrm{dCor}(X,Y) = \frac{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\pi_{ij} - \pi_{i+}\pi_{+j}\right)^2}}{\left\{\sum_{i=1}^{I}\pi_{i+}^2\left(\sum_{i=1}^{I}\pi_{i+}^2 + 1\right) - 2\sum_{i=1}^{I}\pi_{i+}^3\right\}^{1/4}\left\{\sum_{j=1}^{J}\pi_{+j}^2\left(\sum_{j=1}^{J}\pi_{+j}^2 + 1\right) - 2\sum_{j=1}^{J}\pi_{+j}^3\right\}^{1/4}}, \tag{1.11}$$

with a corresponding test statistic for a sample from the joint distribution of

$$T_{\mathrm{dCor}} = \frac{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(p_{ij} - p_{i+}p_{+j}\right)^2}}{\left\{\sum_{i=1}^{I}p_{i+}^2\left(\sum_{i=1}^{I}p_{i+}^2 + 1\right) - 2\sum_{i=1}^{I}p_{i+}^3\right\}^{1/4}\left\{\sum_{j=1}^{J}p_{+j}^2\left(\sum_{j=1}^{J}p_{+j}^2 + 1\right) - 2\sum_{j=1}^{J}p_{+j}^3\right\}^{1/4}}, \tag{1.12}$$

where $p_{ij} = \frac{n_{ij}}{n}$, $p_{i+} = \frac{n_{i+}}{n}$, and $p_{+j} = \frac{n_{+j}}{n}$ are the sampling proportions.

It is clear by inspection of Equation 1.11 that we retain the key property that $\mathrm{dCor}(X,Y) = 0$ iff $X \perp Y$. We will study how an extension of this statistic can be used for independence and homogeneity tests in Chapter 6, in the context of our joint work with Zhang on empirical investigations of the power of T_{dCor} test for conditional independence in sparse three-way contingency tables. The results there will indicate that a test using T_{dCor} has more power to detect dependence relations with sparse data than do traditional tests. This fact was a particular motivation for our empirical study in Chapter 5 of using the $\mathrm{dCor}(X, Y)$ measure to learn probability networks that capture the dependency structure among multiple categorical variables; since even if we start with a large data set, dependence relations that condition on multiple variables will usually result in a sparse test.

The study of network structure learning occupies most of Part II. There is a combinatorial explosion of possible dependency relations to test using the measure; and even given a choice of which tests to conduct, we need a way to induce the overall network structure from their results. This leads us to the idea of graphical models, which are a powerful tool to devise induction algorithms. The usefulness arises because of a deep equivalence between the set of separation statements that hold for a graph, and the set of conditional independence statements that hold among a set of random variables corresponding to the vertices of the graph. Fortunately, the language of graphs is much easier to use and reason about than the language of conditional independence. Hence it is very common to use graphical models when studying the dependence structure of high-dimensional data; and graphical models are, in fact, the common theme that connects Part II and Part III of the dissertation. We now turn to background on this topic.

Chapter 2

Fundamentals of Graphical Models

2.1 Terminology

Although graphs are not our primary objects of investigation, they play a key role in facilitating the algorithms used in this dissertation. In this section we cover the notation and terminology we will need.

Definition 2.1. *A graph is a pair $G = (V, E)$, where V is a (finite) set of vertices or nodes and $E \subseteq X \times X$ is a (finite) set of edges, links, or arcs. It is understood that there are no loops, that is, there are no edges (A, A) for any $A \in V$. G is called undirected iff*

$$\forall A, B \in V : (A, B) \in E \Rightarrow (B, A) \in E.$$

That is, two ordered pairs (A, B) and (B, A) are identified and represent only one (undirected) edge. G is called directed iff

$$\forall A, B \in E : (A, B) \in E \Rightarrow (B, A) \notin E.$$

An edge (A, B) is considered to be directed from A towards B.

The graphs defined above are "simple"; that is, there are no multiple edges between two nodes and no loops. In order to distinguish between directed and undirected graphs, we write $\overrightarrow{G} = (V, \overrightarrow{E})$ for directed graphs.

Definition 2.2. *Let $G = (V, E)$ be an undirected graph. A node $B \in V$ is called adjacent to a node $A \in V$ or a neighbor of A iff there is an edge between them, i.e. iff $(A, B) \in E$. The set of all neighbors of A is*

$$neighbors(A) = \{B \in V | (A, B) \in E\},$$

and deg(A) = |neighbors(A)| is the degree of node A (number of incident edges). The set neighbors(A) is also called the boundary of A. The boundary of A together with A is called the closure of A:

$$closure(A) = neighbors(A) \cup A.$$

Definition 2.3. *Let $G = (V, E)$ be an undirected graph. Two distinct nodes $A, B \in V$ are called connected in G, written $A\widetilde{G}B$, iff there exists a sequence $C_1, C_2, \ldots, C_k, k \geq 2$, of distinct nodes, called path, with $C_1 = A$, $C_k = B$, and $\forall i, 1 \leq i < k : (C_i, C_{i+1}) \in E$.*

Note that in this definition a path is defined as a sequence of nodes, not a sequence of edges. Also note that the nodes on the path must be distinct; that is, the path must not lead back to a node that has been visited. An important special case of an undirected graph is the tree, which restricts the permissible set of paths.

Definition 2.4. *An undirected graph is called singly connected or a tree iff any pair of distinct nodes is connected by exactly one path.*

Definition 2.5. *Let $G = (V, E)$ be an undirected graph. An undirected graph $G_X = (X, E_X)$ is called a subgraph of G (induced by X) iff $X \subseteq V$ and $E_X = (X \times X) \cap E$, that is, iff it contains a subset of the nodes in G and all corresponding edges. An undirected graph $G = (V, E)$ is called complete iff its set of edges is complete, that is, iff all possible edges are present, or formally iff*

$$E = V \times V - \{(A, A) | A \in V\}.$$

A complete subgraph is called a clique. A clique is called maximal iff it is not a subgraph of a larger clique, that is , a clique having more nodes.

Definition 2.6. *Let $\overrightarrow{G} = (V, \overrightarrow{E})$ be a directed graph. A node $B \in V$ is called a parent of a node $A \in V$ and, conversely, A is called the child of B iff there is a directed edge from B to A, that is, iff $(B, A) \in E$. B is called adjacent to A iff it is either a parent or a child of A. The set of all parent of a node A is denoted*

$$parents(A) = \{B \in V | (B, A) \in \overrightarrow{E}\}$$

and the set of its children is denoted

$$children(A) = \{B \in V | (A, B) \in \overrightarrow{E}\}.$$

Definition 2.7. *Let $\overrightarrow{G} = (V, \overrightarrow{E})$ be a directed graph. Two nodes $A, B \in V$ are called d-connected in $\overrightarrow{G}$, written $A \leadsto_{\overrightarrow{G}} B$, iff there is a sequence $C_1, \ldots, C_k, k \geq 2$, of distinct nodes,* *called a directed path, with* $C_1 = A, C_k = B$, *and* $\forall i, 1 \leq i < k : (C_i, C_{i+1} \in \overrightarrow{E})$. $\overrightarrow{G}$ *is called acyclic iff it does not contain a directed cycle, that is, iff for all pairs of nodes* A *and* B *with* $A \rightsquigarrow_{\overrightarrow{G}} B$ *it is* $(B, A) \notin \overrightarrow{E}$.

Note that in a path—in contrast to a directed path—the edge directions are disregarded. (An undirected path is sometimes called a "trail" in order to distinguish it from a directed path.) With directed paths we can now define the notions of ancestor and descendant; as well as the set of all non-descendants, which will be especially important for us.

Definition 2.8. *Let* $\overrightarrow{G} = (V, \overrightarrow{E})$ *be a directed acyclic graph. A node* $A \in V$ *is called an ancestor of another node* $B \in V$ *and, conversely,* B *is called a descendant of* A *iff there is a directed path from* A *to* B. B *is called a non-descendant of* A *iff it is distinct from* A *and not a descendant of* A. *The set of all ancestors of a node* A *is denoted*

$$\mathit{ancestors}(A) = \{B \in V | B \rightsquigarrow_{\overrightarrow{G}} A\},$$

the set of its descendants is denoted

$$\mathit{descendants}(A) = \{B \in V | A \rightsquigarrow_{\overrightarrow{G}} B\},$$

and the set of its non-descendants is denoted

$$\mathit{nondescs}(A) = V - \{A\} - \mathit{descendants}(A).$$

In analogy to undirected graphs there are the special cases of a tree and a polytree, in which the set of paths is severely restricted.

Definition 2.9. *A directed acyclic graph is called singly connected or a polytree iff each pair of distinct nodes is connected by exactly one path. A directed acyclic graph is called a (directed) tree iff it is a polytree and exactly one node (the "root node") has no parents.*

An important concept for directed acyclic graphs is the notion of a topological order of the nodes of the graph. It can be used to test whether a directed graph is acyclic, since it only exists for acyclic graphs, and is often useful to fix the order in which the nodes of the graph are to be processed.

Definition 2.10. *Let $\overrightarrow{G} = (V, \overrightarrow{E})$ be a directed acyclic graph. A numbering of the nodes of $\overrightarrow{G}$, that is, a function $o : V \to \{1, \ldots, |V|\}$ satisfying*

$$\forall A, B \in V : (A, B) \in \overrightarrow{E} \Rightarrow o(A) < o(B),$$

is called a topological order of the nodes of $\overrightarrow{G}$.

For any directed acyclic graph $\overrightarrow{G}$ a topological order can be constructed with the following recursive algorithm: Select an arbitrary childless node A in $\overrightarrow{G}$ and assign to it the number $|V|$. Then remove A and its incident edges from $\overrightarrow{G}$, and recurse to the first step to append a topological order for the reduced graph. It is clear that for graphs with directed cycles there is no topological order, since a directed cycle cannot be reduced by the above algorithm—it will eventually reach a situation in which there is no childless node but the graph is not empty.

2.2 Separation and the graphoid axioms

Graphical models exploit the structural similarity between the sets of conditional independence statements that can hold in high-dimensional distributions; and the sets of node separation statements that can hold in graphs (either directed or undirected). To be more specific, both conditional independence statements and (node) separation statements satisfy the following axioms.

Definition 2.11 (Semi-graphoid and graphoid axioms)**.** *Let V be a set of (mathematical) objects and $(\cdot \perp\!\!\!\perp \cdot|\cdot)$ a three-place relation of subsets of V. Furthermore, let $W, X, Y,$ and Z be four disjoint subsets of V. Then the four statements*

- $(X \perp\!\!\!\perp Y|Z) \Rightarrow (Y \perp\!\!\!\perp X|Z)$, *(symmetry)*
- $(W \cup X \perp\!\!\!\perp Y|Z) \Rightarrow (W \perp Y|Z) \wedge (X \perp\!\!\!\perp Y|Z)$, *(decomposition)*
- $(W \cup X \perp\!\!\!\perp Y|Z) \Rightarrow (X \perp\!\!\!\perp Y|Z \cup W)$, *(weak union)*
- $(X \perp\!\!\!\perp Y|Z \cup W) \wedge (W \perp\!\!\!\perp Y|Z) \Rightarrow (W \cup X \perp\!\!\!\perp Y|Z)$, *(contraction)*

are called the semi-graphoid axioms. A three-place relation $(\cdot \perp\!\!\!\perp \cdot|\cdot)$ *that satisfies the semi-graphoid axioms for all* W, X, Y, *and* Z *is called a "semi-graphoid". The above four statements together with*

$$(W \perp\!\!\!\perp Y|Z \cup X) \wedge (X \perp\!\!\!\perp Y|Z \cup W) \Rightarrow (W \cup X \perp Y|Z), \quad (\textit{intersection})$$

are called the graphoid axioms. A three-place relation $(\cdot \perp\!\!\!\perp \cdot|\cdot)$ *that satisfies the graphoid axioms for all* W, X, Y, *and* Z *is called a "graphoid".*

The rationale underlying graphical models is that the three-place relation named in this definition may either be interpreted as conditional independence or as separation, thus making it possible to use graphs as a "language" for representing sets of conditional independence statements. If the three-place relation is interpreted as conditional independence, $X \perp\!\!\!\perp Y|Z$ means that

$$\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : \forall z \in \mathrm{dom}(Z) :$$
$$P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z),$$

in the probabilistic case. It is easy to show that the semi-graphoid axioms hold for both probabilistic conditional independence, as well as for strictly positive probability measures. When $X \perp\!\!\!\perp Y|Z$ is interpreted as a statement about separation (of nodes) in a graph, it depends on whether the graph is directed or undirected. If it is undirected, the meaning is simply as below.

Definition 2.12. *Let* $G = (V, E)$ *be an undirected graph and* X, Y, *and* Z *be three disjoint subsets of nodes.* Z *u-separates* X *and* Y *in* G, *written* $\langle X|Z|Y\rangle_G$, *iff all paths from a node in* X *to a node in* Y *contain a node in* Z. *A path that contains a node in* Z *is called blocked (by* Z*), otherwise it is called active.*

For directed graphs the conditions are slightly more complicated:

Definition 2.13. *Let* $\vec{G} = (V, \vec{E})$ *be a directed acyclic graph and* X, Y, *and* Z *three disjoint subsets of nodes.* Z *d-separates* X *and* Y *in* $\vec{G}$, *written* $\langle X|Z|Y\rangle_{\vec{G}}$, *iff there is no path from a node in* X *to a node in* Y *along which the following two conditions hold.*

1. *Every node with converging edges either is in* Z *or has a descendant in* Z*; and,*

2. Every other node is not in Z.

A path satisfying these conditions is said to be active, otherwise it is said to be blocked (by Z).

It is easy to show that both u -separation and d -separation satisfy the graphoid axioms. Hence one may use an appropriate graph to capture the set of conditional independence statements that hold in a given distribution.

Definition 2.14. *Let $(\cdot \perp_\delta \cdot|\cdot)$ be a three-place relation representing the set of conditional independence statements that hold in a given distribution δ over a set U of attributes. A (directed or undirected) graph $G = (U, E)$ over U is called a conditional dependence graph or a dependence map with regard to δ, iff for all disjoint subsets $X, Y, Z \subseteq U$ of attributes*

$$X \perp\!\!\!\perp_\delta Y\,|Z \Rightarrow \langle X|Z|Y\rangle_G\ .$$

That is, if G captures by u-separation all (conditional) independences that hold in δ and thus represents only valid (conditional) dependences.

Similarly, G is called a conditional independence graph or an independence map w.r.t. δ, iff for all disjoint subsets $X, Y, Z \subseteq U$ of attributes

$$\langle X|Z|Y\rangle_G \quad \Rightarrow \quad X \perp\!\!\!\perp_\delta Y|Z.$$

That is, if G captures by u-separation only (conditional) independences that are valid in δ. G is said to be a perfect map of the conditional (in)dependences in δ, if it is both a dependence map and an independence map.

Although the correspondence cannot always be made perfect, it is a very convenient tool. Together with the core theorem of graphical models, which connect conditional independence graphs with decompositions of distributions (factorizations in the case of probability distributions), these definitions are the basis of using graphs to capture essential properties of distributions and to derive consistent and efficient methods for drawing inferences in them. When learning graphical models from data, these definitions provide the basis for induction algorithms based on conditional independence tests.

2.3 Markov properties of graphs

Markov properties of graphs allow us to confine ourselves to checking a smaller set of conditional independences when inducing graphs. For undirected graphs the Markov properties are defined as below (see [22], [37], [54], and [38] for references).

Definition 2.15. *Let $(\cdot \perp\!\!\!\perp_\delta \cdot|\cdot)$ be a three-place relation representing the set of conditional independence statements that hold in a given joint distribution δ over a set U of attributes. An undirected graph $G = (U, E)$ is said to have, with regard to the distribution δ:*

- *The pairwise Markov property iff in δ any pair of attributes, which are nonadjacent in the graph, are conditionally independent given all remaining attributes, that is, iff*

$$\forall A, B \in U, A \neq B : \ (A, B) \notin E \Rightarrow A \perp\!\!\!\perp_\delta B|U - \{A, B\}.$$

- *The local Markov property iff in δ any attribute is conditionally independent of all remaining attributes given its neighbors, that is, iff*

$$\forall A \in U : \ A \perp\!\!\!\perp_\delta U - \text{closure}(A)| \ \textit{neighbors}\ (A).$$

- *The global Markov property iff in δ any two sets of attributes which are u-separated by a third are conditionally independent given the attributes in the third set, that is, iff*

$$\forall X, Y, Z \subseteq U : \ \langle X|Z|Y\rangle_G \Rightarrow X \perp\!\!\!\perp_\delta Y|Z.$$

Definition 2.16. *Let $(\cdot \perp\!\!\!\perp_\delta \cdot|\cdot)$ be a three-place relation representing the set of conditional independence statements that hold in a given joint distribution δ over a set U of attributes. A directed acyclic graph $\vec{G} = (U, \vec{E})$ is said to have, with regard to the distribution δ:*

- *The pairwise Markov property iff in δ any attribute is conditionally independent of any non-descendant not among its parents given all remaining non-descendants, that is, iff*

$$\forall A, B \in U : B \in \ \textit{nondescs}\ (A) - \text{parents}(A) \Rightarrow A \perp\!\!\!\perp_\delta B \mid \ \textit{nondescs}\ (A) - \{B\}.$$

- *The local Markov property iff in δ any attribute is conditionally independent of all remaining non-descendants given its parents, that is, iff*

$$\forall A \in U : \quad A \perp\!\!\!\perp_{\delta} \ \mathit{nondescs}\,(A) - \mathrm{parents}(A) \mid \mathrm{parents}(A).$$

- *The global Markov property iff in δ any two sets of attributes which are d-separated by a third are conditionally independent given the attributes in the third set, that is, iff*

$$\forall X, Y, Z \subseteq U : \quad \langle X|Z|Y \rangle_{\vec{G}} \Rightarrow X \perp\!\!\!\perp_{\delta} Y \mid Z.$$

Part II

Learning Networks with Distance Correlation

Chapter 3

Algorithms for Inducing Network Structure

Given the equivalence in Section 2.2 between the separation statements that hold for a graph, and the conditional dependencies of the random variables represented by its nodes, it becomes very natural to look for ways to apply the ideas and algorithms of graph theory to statistical inference. Suppose we a have an i.i.d. sample from the joint distribution of some random variables. In general, the problem of inferring conditional dependencies from the sample is intractable, given the combinatorial explosion of possible dependencies. We must find heuristic methods that can recover (at least some of) the conditional dependencies by testing only a small subset of the possible candidates. It turns out to be much easier to construct these heuristics by drawing on existing ideas from graph theory.

Our discussion will follow Borgelt's exposition in [8]. The main idea is to build up a graph representing the dependence structure by testing marginal and relatively low-order conditional dependencies (that is, with only a few conditioning attributes); where these tests use an arbitrary dependence measure and are selected from the family of possible tests using a search method. It will be clear that we can find examples in which such a heuristic approach fails, where attributes that are not adjacent in the developing graph can ultimately exhibit a strong dependence. But in practice, these approaches can perform reasonably well. Chapter 5 will provide numerous empirical results in which we compare the performance of distance correlation to other well-known dependence measures when using search methods based on graphical models.

It is worth noting that learning graphical models from data is a very broad idea. For example, we can:

1. Test whether a distribution is decomposable with regard to a given graph.
2. Find a suitable graph by measuring the strength of dependences.
3. Find an independence map by performing conditional independence tests.

We are focusing on examples of the third approach. This approach exploits the theorems that connect conditional independence graphs and graphs that represent decompositions. It has the advantage that a single conditional independence test can exclude several candidate graphs. On the other hand, a conditional independence test that yields a false positive can have severe consequences for the accuracy of the final model. To ensure good performance in the general case, it is often necessary to assume that there exists a perfect map for the domain; and that the the result of all conditional independence tests will coincide with true relationship in the underlying distribution.

3.1 Dependence measures

A dependence measure is essentially a scoring function that rates the strength of the (conditional) dependence of two variables, plus a *threshold.* If the value of the scoring function is below the threshold, the variables are considered to be (conditionally) independent; otherwise they are judged to be (conditionally) dependent. As mentioned above, a search method is necessary to decide which conditional dependences to test. There is an abundance of scoring functions, since apart from all the standard dependence measures used in classical statistics, almost any measure used in decision tree induction can be adapted to serve as a scoring function. Even if a measure was first intended for assessing the strength of marginal dependence, it can usually be extended to yield a measure for conditional dependence by simply computing its value for each possible instantiation of the conditioning attributes and then aggregating these values in a suitable manner. For example, one of the most common measures for the strength of marginal probabilistic dependence is the well-known *information gain*,

$$I_{\text{gain}}(A,B) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{i.}p_{.j}} = -\sum_i p_{i.} \log_2 p_{i.} + \sum_j p_{j.} \sum_i p_{i|j} \log_2 p_{i|j},$$

where A and B are two categorical variables; i and j range over their respective values; and p_{ij}, p_i, and $p_{j,j}$ are the probabilities of the joint and individual occurrence of these values, with conditional probabilities defined in the standard way. (This quantity is also known as the mutual information between A and B, and is equivalent to their Kullback-Leibler divergence). We can extend this to conditional information gain as follows:

$$I_{\text{gain}}(A, B|C) = \sum_k p_{..k} \sum_{i,j} p_{ij|k} \log_2 \frac{p_{ij|k}}{p_{i.|k} p_{.j|k}}.$$

That is, by simply taking the expected information gain given the conditions, summing its value over all possible instantiations of C. (Note that C may be an individual variable or a set of variables, so that k will refer to individual values or value vectors, respectively.) It is also worth recalling the standard χ^2 measure,

$$\chi^2(A, B) = N_{..} \sum_{i,j} \frac{(p_{i.}p_{.j} - p_{ij})^2}{p_{i.}p_{.j}}.$$

where $N_{..}$ is the total number of sample cases in the database to learn from. Extended to conditional tests it reads

$$\chi^2(A, B|C) = N_{..} \sum_k p_{..k} \sum_{i,j} \frac{\left(p_{i.|k}p_{.j|k} - p_{ij|k}\right)^2}{p_{i.|k}p_{.j|k}}.$$

We will use the performance of information gain and the χ^2 measure as baselines for evaluating distance correlation in later sections.

A general difficulty with conditional independence tests arises as the number of conditioning attributes (the *order* of the test) increases. In practice, data sets are always of limited size, so high-order conditional independence tests quickly become unreliable. Hence algorithms for learning graphical models from data must be designed to limit the order of the conditional independence tests involved. The Cheng-Bell-Liu algorithm, which we review in the next section, attempts to do so by exploiting an already constructed graphical model to determine suitable condition sets. If the occurring condition sets still become too large, the fallback is enforce to a user-specified limit on the order; all conditional independence tests with a higher order are taken as failures, no matter the score returned by the dependence measure.

3.2 Search methods

We now turn to the question of which conditional dependencies should be tested with our chosen dependence measure. This is the question answered by a *search method.* We will focus on search methods that use graphical models and the results of conditional dependence tests already performed in the search. As mentioned above, the total search space is huge unless the number of attributes is considerably small, so the search cannot be exhaustive. We also only gain a small amount of information from each test—either a score is above the threshold or it is not—and it can be difficult to aggregate this information into measure that provides a useful overall measure of the "correctness" of an intermediate graph being built during the search. The consequence of these obstacles is that, in general, it is very difficult to recover the dependence structure of a large number of attributes from a small sample.

Let us think of a concrete example. Suppose the true network includes a simple undirected chain. Then the two attributes at the ends of the chain must be conditionally independent, given *any* nonempty subset of the attributes between them. (And note these true conditional independence statements are only a small fraction of those we must consider.) For a long chain, there is little hope of actually testing all these conditional independences. A strategy based on only testing pairwise conditional independences will be severely lacking, even if it is able to exploit some of the equivalences of different Markov properties of graphs from Section 2.3. More creative search strategies are necessary; and no known approach is entirely satisfactory.

Baseline algorithm

Most approaches work by evolving from basic ideas developed by Spirtes et al. in [24]. They begin with the assumption that there does exist a perfect map for the domain under consideration. (That is, there exists a graph that represents exactly those conditional independence statements that hold in the joint distribution on the domain.) Then we can infer from a conditional independence $A \perp\!\!\!\perp B|S$, where A and B are attributes and S is a (possibly empty) set of attributes, that there is no edge between A and B. The following exhaustive algorithm hence works to recover the perfect map, with special treatment of the case where $\vec{G}$ is assumed to be directed acyclic.

Algorithm 3.1. *[Recovery of a perfect map from conditional independence tests]*

Given a distribution δ over an attribute set U for which there exists a perfect map $G = (U, E)$, we recover G as follows:

1. *For each pair of attributes A and B, search for a set $S_{AB} \subseteq U - \{A, B\}$, so that $A \perp\!\!\!\perp_{\delta} B | S_{AB}$ holds; that is, so that A and B are conditionally independent given S_{AB}. If there is no such set S_{AB}, connect the two attributes by an undirected edge.*

2. *If $\vec{G}$ is directed acyclic, for each pair of nonadjacent A and B with common neighbor $C \notin S_{AB}$, direct the edges towards C; that is, $A \to C \leftarrow B$.*

3. *If $\vec{G}$ is directed acyclic, repeatedly apply the following rules to direct all remaining undirected edges:*

   - *If for two adjacent attributes A and B there is a strictly directed path from A to B not including the edge from A to B, then set $A \to B$.*
   - *If for two nonadjacent attributes A and B there is an attribute C such that $A \to C$ and $C - B$, then set $C \to B$.*
   - *If neither above rule applies, set an arbitrary direction on a given edge.*

The first step of Algorithm 3.1 simply applies the core idea that attributes are only connected by an edge for if there is no set of attributes that renders them conditionally independent. (But note that if only an independence map exists, but not a perfect map, this step could omit edges incorrectly.)

When the perfect map is directed acyclic, the second step draws on the insight that if a set S_{AB} of attributes renders two attributes A and B conditionally independent, it must block all paths from A to B in the graph. In particular, it must block paths that run via a common neighbor C of A and B. However, if this common neighbor C is not in S_{AB}, the only way in which this path can be blocked (given S_{AB}) is by an edge orientation that converges at C. Moving to the third step for $\vec{G}$, the first rule simply exploits the acyclic nature of the map. If B is reachable from A to B without using $A - B$, the alternative choice $B \to A$ would introduce a cycle, directing the edge from B towards A would introduce a directed cycle. The logic for the second rule is that, under the conditions in this rule, the attribute C must be in a separating set S_{AB}; as otherwise the algorithm would have set $B \to C$ in the second step. But if C is in S_{AB}, the path from A to B via C can only be blocked by S_{AB} if it does *not* have converging edges at

C. Consequently, the edge $C - B$ cannot be directed towards C, but must be directed towards B. The third rule is necessary to avoid "deadlock"; for example, suppose that the underlying perfect map is a simple directed chain. The second step of the algorithm will have no effect; and neither of the two preceding rules will immediately apply. This default rule of randomly assigning a direction will break the deadlock.

Practical difficulties

Note that the actual learning takes place in the first two steps of Algorithm 3.1; where the first step learns the skeleton of the graph in both the directed and undirected cases, and the second step identifies the v-structures. (In the third step, we simply choose the remaining directions in a fashion that does not introduce directed cycles or additional v-structures. Of course these directions are not deterministic, given the third rule; but it is a fact that all directed acyclic graphs with the same skeleton and the same v-structures are Markov-equivalent. Hence we could also skip the third rule and accept a final graph with some possibly undirected edges as a compact representation of the class of all Markov equivalent graphs that are perfect maps for the domain being studied.) Although this learning in the first two steps is guaranteed to recover a perfect map if it exists, there is no known efficient implementation. Indeed, to ensure that there is no set S_{AB} that renders two attributes A and B conditionally independent, in principle we must check all subsets of $U - \{A, B\}$, of which there are

$$s = \sum_{i=1}^{|U|-2} \left(\begin{array}{c} |U| - 2 \\ i \end{array} \right).$$

Even worse, unless $|U|$ is quite small, some of these sets contain a large number of attributes, so we must perform high-order conditional independence tests. With a finite database of samples, it is likely we will not be able to get an accurate estimate of the joint distribution of A and B for each instantiation of the conditioning attributes in U.

One response to this problem is to assume the existence of a sparse perfect map; that is, a perfect map with only a limited number of edges. In this case, any pair of attributes can be separated by an attribute set of limited size. Thus we only have to test for conditional independence up the order that is fixed by the chosen size limit. If for two attributes A and B, all conditional independence tests with an order up to this upper bound failed, we infer from the sparseness assumption there is no set S_{AB} that renders them conditionally independent. The sparsest connected graph is, of course, a tree (or its directed counterpart, a polytree). In this extreme case, there are special versions of Algorithm 3.1. That is, when the underlying perfect map is a polytree, the conditional independence tests can be restricted to at most order one, since for any pair of attributes there is only one path connecting them; and this path can be blocked with at most one attribute. A useful overview of specializations that consider other, less restricted, classes of graphs is the contribution of Campos et al. in [16]. When a search algorithm uses a sparseness assumption that does not hold in the true network, the result will still be at least an independence map. This follows since assuming a sparse graph reduces the set of tests for conditional independence, and may only lead to additional edges relative to the true network. (In practice sparseness assumptions frequently do not hold.)

The second problem with Algorithm 3.1 arises when even the weaker assumption of existence of a perfect map does not hold. The consequences of violating this assumption are more severe than just finding an independence map; if the domain does not admit a perfect map, the network induced by can be very distorted. To avoid this outcome, we must insert additional edges. This requires much more than just a specialization of Algorithm 3.1, and Chapter 4 below will be devoted to a reviewing more robust search algorithm due to Cheng et al., as well as an adaptation due to Borgelt. We will then move in Chapter 5 to empirical studies of how well a dependence measure using distance correlation will perform with this search method when learning networks. For this we will build on a software library also developed by Prof. Borgelt.

3.3 The INeS Software Package

The Induction of Network Structure (INeS) software package [6] is available for download from `https://www.borgelt.net/ines.html`. The package defines file formats to represent attribute domains, probability networks, and databases of sample observations. It then features several programs written in the C language, as below.

- `gendb` – Generates a database of a given number of random samples with attributes from a given domain, where those attributes have a given joint distribution.

- `ines` – Given a database of samples, induces a probability network that estimates the joint distribution of the attributes in the sample database. The induction can use any combination of dependence measure and search method supported by INeS.
- `neval` – Given a true network, a database of generated samples, and an induced network, evaluates the log-likelihood of the database given the induced network; and compares the induced network to the true network. This comparison consists of edges added and missed relative to the DAG representing the true network. The `neval` program also reports the number of parameters required to specify the conditions in the induced network.

Although INeS supports a wide variety of search methods (ranging from optimum weight spanning tree construction to hypertree simulated annealing); and dependence measures (ranging from information gain to stochastic complexity), it does not support any form of distance correlation "out-of-the-box". Hence a necessary step in our investigation must be to enhance the `ines` program with an implementation of distance correlation for categorical variables as defined by Zhang in [57].

Chapter 4

Chow-Liu Search Methods

There are many proposed search methods which try to overcome the weaknesses of Algorithm 3.1—that is, its extremely high computational cost and need to assume existence of a perfect map. We will focus on just few search methods with which to evaluate the performance of distance correlation as a dependence measure. This will primarily be two methods which both begin by constructing a skeleton graph called a Chow-Liu tree. The first method, know as the Cheng-Bell-Liu algorithm, was introduced in [4], and builds the final model using a directed graph. The second, from Borgelt in [7], uses an undirected graphical model for its intermediate steps.

4.1 Induction with a directed graphical model

Drafting

The Cheng-Bell-Liu algorithm begins by drawing on ideas of Chow et al. in [14]. A so-called *Chow-Liu tree* is formed from the attributes by evaluating all possible edges—that is, pairs of attributes—using a provided independence test in which higher scores reject the null hypothesis of independence. The evaluation compares each edge score to a minimum threshold; any edge below the threshold is ignored. (The classical example uses information gain as the independence test and a threshold of 0.1 bits.) Edges that survive this filter are then weighted by their score; and we use a fast algorithm from graph theory to construct the optimum weight spanning tree based on these edge weights. This is the algorithm's “first draft” of the graphical model representing the induced network.

Thickening

Next the Cheng-Bell-Liu algorithm “thickens” the developing graph with edges for which we cannot find evidence of a conditional independence between the incident attributes. The evidence for a candidate edge $A - B$ is evaluated in a series of conditional independence tests where the conditioning set is iteratively refined. (Note that at this stage, edges are still undirected.)

To form the initial conditioning set, we pick one of the nodes, say A, and identify all adjacent nodes which lie on a path from A to B. After computing the score of the independence test using this set, we discard the node which leads to the *largest decrease* in the score of the independence test and repeat. This continues until the score falls below the independence threshold, in which case the edge is not added to the developing graph; or no further decrease is possible, in which case the edge is added.

Since it is not yet known whether A or B will be the parent in the final directed graph, the above process must be repeated with the roles reversed. The rationale for the iterative reduction in this step is the local Markov property of directed graphs. In particular, the conditioning sets may have adjacent nodes that are children in the directed graph. This is a problem because in the underlying graphical model there may be a v-structure with the child; and including the child in the conditioning set could activate a path and break the conditional independence test.

Thinning

It might happen during the thickening phase that the developing graph is too sparse to reveal the conditional independence of a pair of attributes. That is, some of the true paths between the attributes may not be present at the time of the independence test, resulting in a Type II error. The algorithm attempts to fix such mistakes in the fourth step by retesting every edge in the graph constructed in the first three steps, and "thinning" out any edges whose nodes are now judged conditionally independent. The test for thinning an edge is conducted in two ways. First, the algorithm repeats the approach used during thickening; that is, the score of the conditional independence test given all neighbors of one incident attribute is compared to the threshold. Second, the algorithm attempts to address certain degenerate cases which the first test may mishandle. This requires a second test which conditions on not just the neighbors of one attribute; but also the neighbors of these neighbors. (See [7] for a complete discussion of the degenerate cases for when this is necessary.) If either test judges an edge to connect conditionally independent attributes, the algorithm removes that edge.

Orienting

The final step of the algorithm assigns directions to the edges that survived thinning. The approach is actually identical to steps two and three of Algorithm 3.1. For each edge $A - B$, we obtain an appropriate $S_{AB} \subseteq U \setminus \{A, B\}$ by another sequence of conditional independence tests, starting with the "strict" test whose conditioning set is both the neighbors of A; and *their* neighbors. Nodes are removed from conditioning set in a greedy fashion as long as the result of the test still falls below the independence threshold.

4.2 Induction with an undirected graphical model

We now turn to Borgelt's adaption of the Cheng-Bell-Liu algorithm. The *result* can be roughly understood as what would be obtained by executing the Cheng-Bell-Liu algorithm, then removing the edge directions and adapting the result by adding edges between non-adjacent parents of a node to obtain a roughly equivalent set of separation statements. Of course, Borgelt's algorithm does not actually form a directed graph and then remove the directions. Instead it drafts a Chow-Liu tree as above, then uses a modified version of the thickening step and replaces the final orienting step with a so-called "moralizing" step which comes directly after thickening. We review these differences next.

Thickening

Borgelt's algorithm continues after drafting by ordering the edges with a score above the threshold, but not included in the Chow-Liu tree, by decreasing weight. It traverses them in this order, and performs tests for conditional independence of the incident attributes, *given* the neighbors of the incident attributes which lie on any of their connecting paths in the developing graph. If this conditional independence test scores above the threshold, the edge is added to the candidate model. The rationale derives from the local Markov property of undirected graphs as discussed in Section 2.3, which guarantees an attribute is conditionally independent of any other attribute given its neighbors. It is possible to conduct additional tests in this step, conditioning on just the neighbors of one attribute or the other. This can improve the robustness of the overall algorithm, since the initial test may give a false positive for independence if the current graph is still too sparse and not all neighbors necessary to render the attributes conditionally independent are already present.

However many conditioning sets are used, if any of the conditional independence tests returns a score above the threshold, the developing graph is "thickened" by adding the edge. A key difference with the thickening step in Subsection 4.1 above is that since here we do not require a directed graph, the work performed by the Chow-Bell-Liu algorithm in iterative reduction of condition sets is superfluous. This does result in some increased computational efficiency in Borgelt's approach—see [7] for details and simulation studies using INeS that compare the total number of tests required by each algorithm in various settings.

Moralizing

The next step of the algorithm would not be necessary given an assumption of the existence of an undirected perfect map of the domain, as made in Algorithm 3.1. However, this would cripple the algorithm in practice, as dependence structures that contain directed v-structures arise quite frequently. To address such structures in the undirected setting, the algorithm now "moralizes" the graph by attempting to connect the parents of these underlying v-structures. The reasoning is that even though the parents are independent given their common ancestors (which may be the empty set), they become dependent in the presence of a common child. Note the effect on the set of conditions is non-monotone—enlarging the set of conditions destroys a conditional independence. Thus the algorithm may leave some conditional independences unrepresented. In practice this is not a major problem, at least for reasoning applications, since an independence map will often be adequate. Precisely, moralizing considers the edges in the developing graph which share a common neighbor; but whose non-identical incident attributes were judged conditionally independent during drafting. These attributes are now tested for conditional independence given all of one of their neighbors. If the independence test scores above the threshold, the algorithm adds edge between the non-identical incident attributes. Note that such edges are the only candidates for connecting the parents of a v-structure in a directed independence map of the domain.

Additional thinning

The efficiency of Borgelt's algorithm may improve in some cases by inserting an additional thinning phase between thickening and moralizing above. The idea is that any edges removed by this thinning would have been removed by the final thinning phase, but the order of the conditional independence tests required may be lower, since a graph with fewer edges may of course contain fewer neighbors to be included in the conditioning sets used. An added benefit is that when edges are thinned before moralizing, this also reduces the number of tests that may need to be carried out during moralizing. The final thinning phase will then target only edges between attributes that received a new incident edge during moralizing. Careful caching of test results can also reduce the overhead of inserting the "additional" thinning.

Chapter 5

Experimental Evaluation of Distance Correlation

5.1 Implementation with INeS

It can be a challenging and time-consuming task to implement network induction algorithms correctly and efficiently. For example, the INeS software package has been under development as part of Prof. Borgelt's research for over twenty years (see again [6]). We hence looked to build on this foundation by extending the package with support for categorical distance correlation as a dependence measure. As part of our research, we extended the INeS codebase with an implementation of categorical distance correlation following Zhang's formulation from [57]. This implementation was done in C using the data structures built up during the development of INeS. We then needed only to add the various "hooks" for this new dependence measure into the `ines` runtime framework, and we were able to take full advantage of the various search algorithms already implemented in `ines`, as well as the complementary tools such as `gendb`, `neval`, and other scripts for conducting numerical experiments. This enhanced version of INeS is available from the author on request.

5.2 Description of numerical studies

The focus of this chapter is a series of numerical studies using the INeS package enhanced with distance correlation as just described. Each study has the following general pattern.

1. Let X be a source domain with known joint distribution, where this distribution is represented by a probability network in the INeS *.net* file format.

2. For various sample sizes N (or independence thresholds t), perform K experiments at each sample size (or independence threshold) by repeating the following nested steps.

    a) Use `gendb` to generate a training set of size N and a test set of fixed size for the experiment as INeS *.tab* files.

    b) For the three dependence measures based on distance correlation, information gain, and the χ^2 statistic, use `ines` to induce a network from the training data with one or both of the search methods described in Chapter 4 and an independence threshold t.

   c) Evaluate the relative performance of the dependence measures using `neval` to compute the log likelihood of the test data, number of edges missed in the induced networks, and number of edges added to the induced networks.

3. Summarize the average relative performance of the dependence measures across the K experiments for each sample size (or independence threshold); visualize examples of induced networks if desired for additional intuition.

When visualizing an example of an induced network, we will add styling cues and edge data to indicate which edges appeared in the true network, which edges were added, and which edges from the true network were missed. Figure 5.1 clarifies this visualization strategy. Note in particular that *only* the thick edges are present in the induced network (shaded darker if also in the true network, and lighter if not). The thinnest and lightest edges were missed in the induced network, and are added purely for visualization.

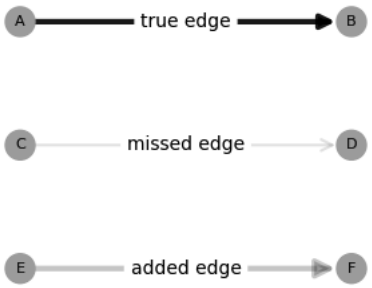


Figure 5.1: Interpretation of edge styles

The concluding sections of this chapter are each devoted to experimental results on a particular dataset. We will use varying choices of sample size, and conduct $K = 10$ experiments in the fashion described above, so that we learn something about how the various dependence measures are able to take advantage of more data to induce more accurate networks. It will then be useful to plot curves displaying sample size against average performance in terms of one of our main performance metrics—that is, true edges missed in the induced network, edges incorrectly added in the induced network, and the log-likelihood of the test data. It is useful to look more concretely at this last metric.

5.3 Intuition of test log-likelihood

Suppose we have a small Bayesian network $\mathcal{N}$ over eight attributes labeled from A to H as in Figure 5.2 on the following page. For simplicity, let each attribute has two levels, so that for example $\mathrm{dom}(A) = \{0, 1\}$. Now suppose we want to evaluate the performance of the network on a test set $\mathbf{X} = \{X_1, \ldots, X_n\}$ where each vector $X_i = (a_i, b_i, c_i, d_i, e_i, f_i, g_i, h_i) \in \{0, 1\}^8$ is an independent sample drawn from the true distribution. We can then compute the log-likelihood of the observed data given the network by factorizing the joint distribution in terms of just the conditional dependences that are present in $\mathcal{N}$. In Equation 5.1 below we do this, abbreviating for example $p_B(B = b_i \mid A = a_i)$ as $p_B(b_i|a_i)$ for readability. This equation makes it clear that to specify the entire joint distribution $p(A, \ldots, H)$ we really only need to specify certain marginal and conditional distributions for the individual attributes. Throughout this chapter, we will always take such distributions to be multinomial.

$$\begin{aligned}
l(\mathbf{X}|\mathcal{N}) &= \sum_{i=1}^{n} \log p(X_i|\mathcal{N}) \\
&= \sum \log \Big[p_A(a_i) * p_B(b_i|a_i) * p_C(c_i|a_i, b_i) * p_D(d_i|c_i, b_i, a_i) * p_E(e_i|c_i, d_i, b_i, a_i) \\
&\qquad\qquad * p_F(f_i|e_i, d_i, c_i, b_i, a_i) * p_G(g_i|f_i, e_i, d_i, c_i, b_i, a_i) \\
&\qquad\qquad * p_H(h_i|e_i, f_i, e_i, d_i, c_i, b_i, a_i) \Big] \\
&\underset{\text{indep.}}{=} \sum \log \Big[p_A(a_i) * p_B(b_i|a_i) * p_C(c_i) * p_D(d_i|c_i) * p_E(e_i|c_i) \\
&\qquad\qquad * p_F(f_i|d_i, b_i) * p_G(g_i|f_i) * p_H(h_i|e_i, f_i) \Big].
\end{aligned} \tag{5.1}$$

Clearly this simplification will apply to any Bayesian network—the more conditional independences that are identified during learning, the simpler the final model will be. The more conditional dependences that are learned, the more complex the final model will be.

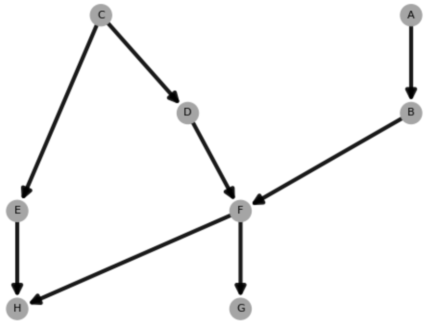
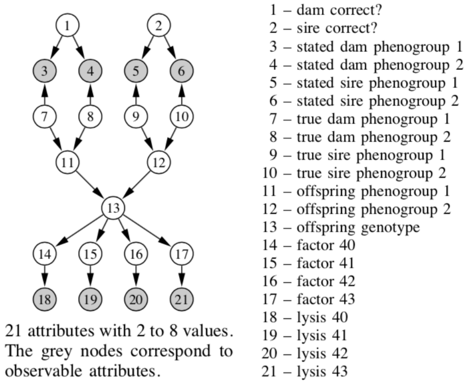


Figure 5.2: Example network

5.4 The Danish Jersey cattle dataset

Description

A standard benchmark for testing the performance of a network induction algorithm is the dataset from [40] that includes genotypical and phenotypical data associated to several lineages of cattle. The dataset was originally used to guide the development of an expert domain model for blood group determination of Danish Jersey cattle in the F-group blood system. This expert model took the form of a Bayesian network, visualized below in Figure 5.3. The example is particularly interesting because of its rich empirical derivation, and the INeS package is distributed with a *djc.net* network file that expresses the “true” hand-crafted network in the INeS format.



1 – dam correct?
2 – sire correct?
3 – stated dam phenogroup 1
4 – stated dam phenogroup 2
5 – stated sire phenogroup 1
6 – stated sire phenogroup 2
7 – true dam phenogroup 1
8 – true dam phenogroup 2
9 – true sire phenogroup 1
10 – true sire phenogroup 2
11 – offspring phenogroup 1
12 – offspring phenogroup 2
13 – offspring genotype
14 – factor 40
15 – factor 41
16 – factor 42
17 – factor 43
18 – lysis 40
19 – lysis 41
20 – lysis 42
21 – lysis 43

21 attributes with 2 to 8 values. The grey nodes correspond to observable attributes.

Figure 5.3: Manually constructed Bayesian network for cattle blood group determination

Experimental results with the Cheng-Bell-Liu search algorithm

Given a sequence of sample sizes $N = 1000, 1500, 2000, 2500, 3000$ we proceeded as in Section 5.2 above. The results using the Cheng-Bell-Liu search algorithm from Chapter 4 with an independence threshold of 0.1 are displayed in Figure 5.4 below. Alongside the performance of the networks induced with the χ^2, information gain, and distance correlation measures, we plot the performance of the true generating network. This provides a sense of scale for the log-likelihood of the test data. The three measures are quite similar on this metric; that is, they learn networks which on average assign about the same likelihood to the test data drawn from the original generating network. With more training data, the all induce networks which give slightly higher average log-likelihood to the test data. The interesting differences appear with the other performance metrics.

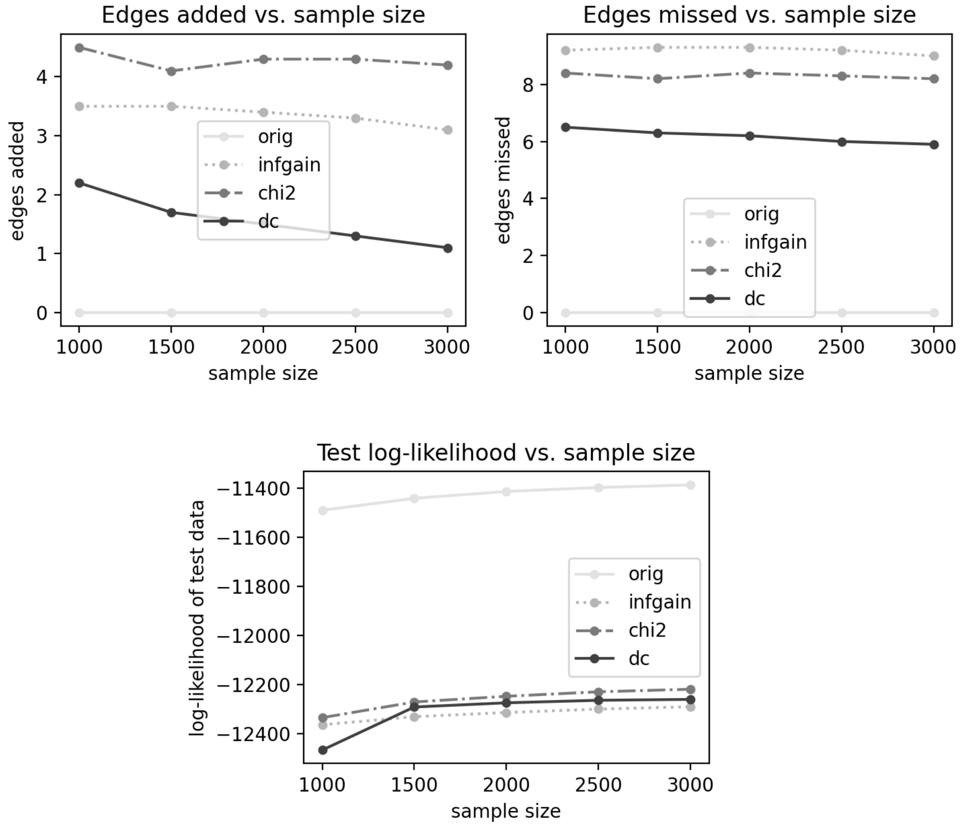
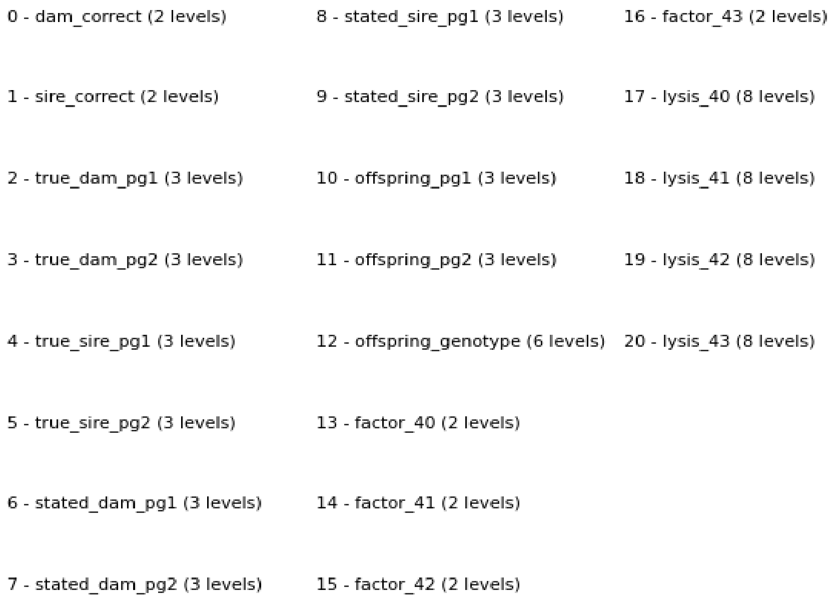


Figure 5.4: Average performance on Danish Jersey cattle data (Cheng-Bell-Liu search)

Here we see that networks with induced using distance correlation do have some noticeable benefits. Networks induced with distance correlation—on average—miss fewer edges from the true network; and add fewer edges that were absent from the true network. We can get a sense of what this difference looks like by comparing some particular networks that were induced with each dependence measure using the same training data set. (Recall the meanings of the three edge styles from Figure 5.1.) In Figures 5.6, 5.7, and 5.8 on the following pages we visualize networks induced using, respectively, distance correlation, the χ^2 measure, and information gain. All were induced from a particular training database of size $N = 3000$ generated in one of the experiments conducted above. The attributes represented by each vertex are in Figure 5.5 below.

In Figure 5.6 we have the network induced with distance correlation. We see that distance correlation was able to discover several true dependences that were missed when using χ^2 and information gain; for example, the dependence of `lysis_40` on `factor_40`, the dependence of `lysis_42` on `factor_42`, and the dependence of `lysis_43` on `factor_43`. This provides more empirical evidence for one of the main themes of this dissertation—namely, that distance correlation is more sensitive to some types of dependences than traditional measures. We can also see that distance correlation is able to avoid most of the tendency of χ^2 and information gain to add spurious edges to the induced network. Both of the latter add three false dependences on `offspring_genotype`. Since this attribute has six levels, this increases the model complexity in terms of the number of parameters needed to specify the factorized distribution for these networks relative to the network induced with distance correlation, as discussed in Section 5.3. For example, the edge added from `offspring_genotype` (attribute 12) to `lysis_40` (attribute 17) below in Figure 5.7 uses 28 more parameters than are required to represent the true conditional dependence of `lysis_40` on `factor_40`.

Note also the monotonic nature of the various performance curves in Figure 5.4. As the sample size increases, the induced networks tend to miss fewer edges, add fewer edges, and assign a higher log-likelihood to the test data. This shows that the learning process has the capacity to extract more information about the generating distribution when it can perform more accurate independence and conditional independence tests.



0 - dam_correct (2 levels)

1 - sire_correct (2 levels)

2 - true_dam_pg1 (3 levels)

3 - true_dam_pg2 (3 levels)

4 - true_sire_pg1 (3 levels)

5 - true_sire_pg2 (3 levels)

6 - stated_dam_pg1 (3 levels)

7 - stated_dam_pg2 (3 levels)

8 - stated_sire_pg1 (3 levels)

9 - stated_sire_pg2 (3 levels)

10 - offspring_pg1 (3 levels)

11 - offspring_pg2 (3 levels)

12 - offspring_genotype (6 levels)

13 - factor_40 (2 levels)

14 - factor_41 (2 levels)

15 - factor_42 (2 levels)

16 - factor_43 (2 levels)

17 - lysis_40 (8 levels)

18 - lysis_41 (8 levels)

19 - lysis_42 (8 levels)

20 - lysis_43 (8 levels)

Figure 5.5: Danish Jersey cattle attributes

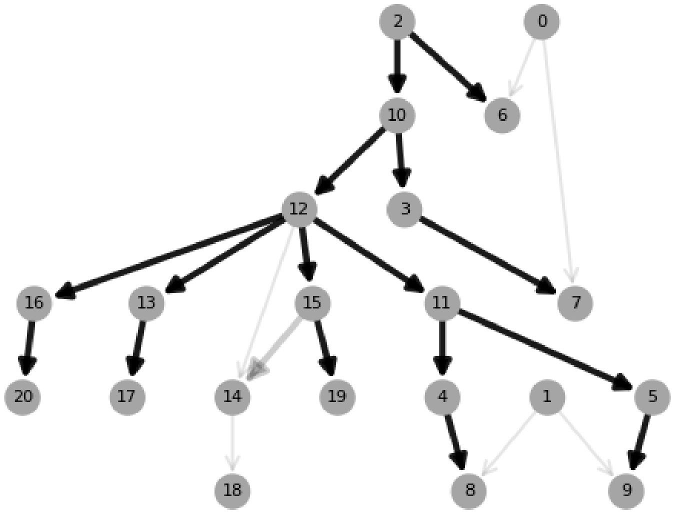


Figure 5.6: Danish Jersey cattle network induced with distance correlation

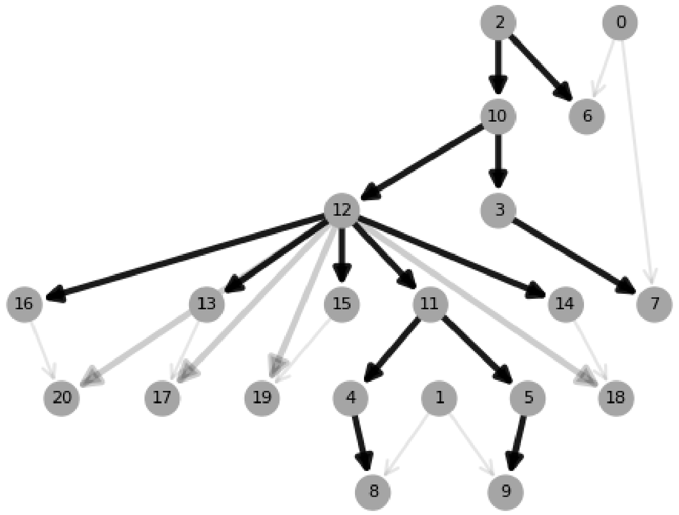


Figure 5.7: Network induced with the χ^2 measure

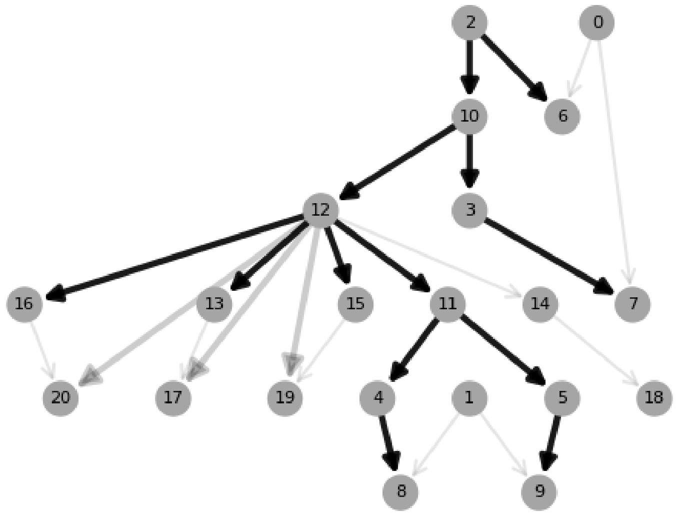


Figure 5.8: Network induced with information gain

Experimental results with Borgelt's algorithm

We now consider the results of the above experiments when using Borgelt's adaptation of the Cheng-Bell-Liu algorithm. These are in Figure 5.9 below. The overall conclusions are very similar to the conclusions with the Cheng-Bell-Liu algorithm. That is, across each set of $K = 10$ experiments at each sample size, the average log-likelihood of the test data for networks learned with each of the three measures is quite similar. However the average number of edges added and edges missed in networks learned using distance correlation is lower than for networks learned with the χ^2 measure or information gain.

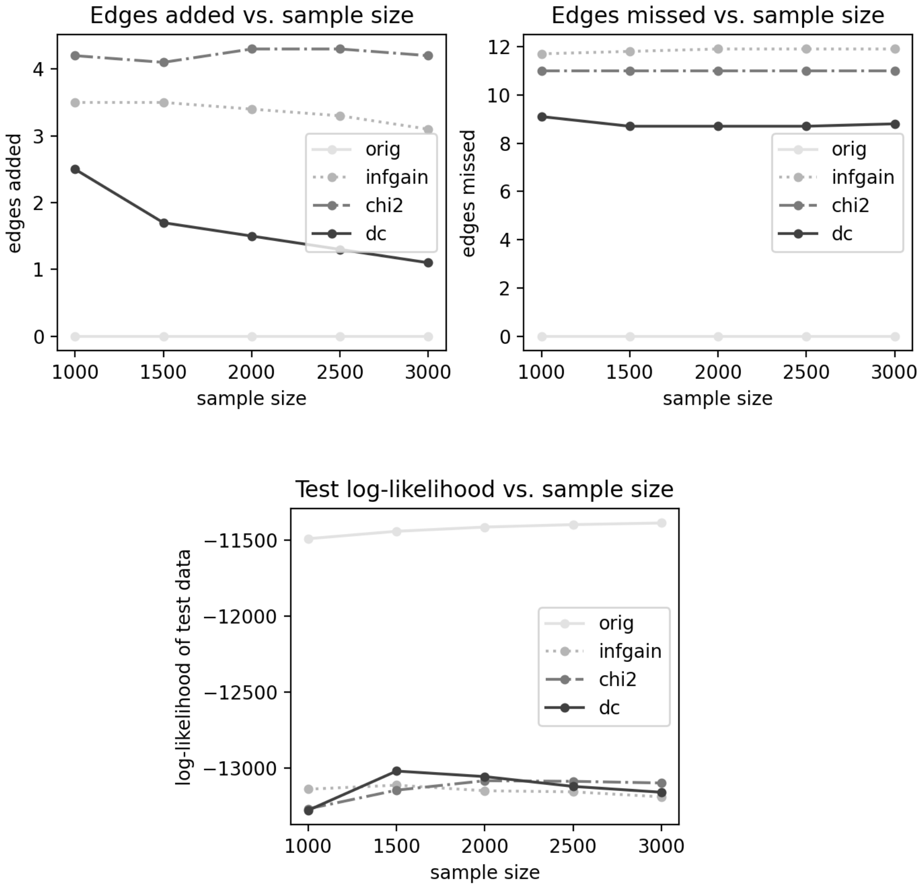


Figure 5.9: Average performance on Danish Jersey cattle data (Borgelt search)

It is also possible to compare the performance of learning networks with distance correlation, using both the Cheng-Bell-Liu algorithm and Borgelt's adaptation. These results are in Figure 5.10. We see that the additional conditional independence tests conducted by the Cheng-Bell-Liu result in generally better performance. For subsequent data sets we will focus on results using Cheng-Bell-Liu only.

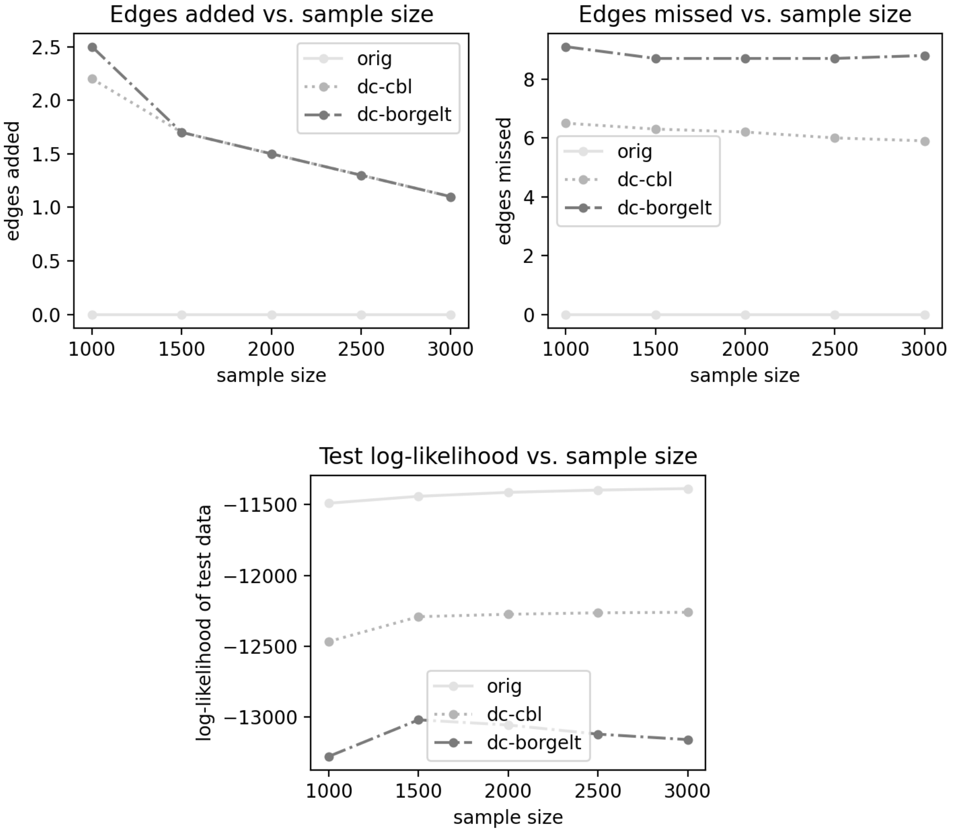


Figure 5.10: Average performance on Danish Jersey cattle data (search comparison)

5.5 The ALARM data set

Description

In developing the Cheng-Bell-Liu algorithm in [4], Cheng et al. tested their approach using experimental results based on the ALARM network database from [10]. This database was deployed in combination with a medical diagnostic alarm message system for patient monitoring in intensive care units. The conditions for an alarm are modeled using a belief network of 37 attributes that represent various observable facts about a patient, as well as eight candidate diagnoses. For example, the `hr` attribute has three levels `low`, `normal`, and `high` that correspond to the patient's heart rate; and the `intubation` attribute has three levels `normal`, `esophageal`, and `onesided` that correspond to the method of intubation. The "true" joint distribution of these attributes is given by a sample database of $10,000$ observations.

Experimental results

We translated the ALARM belief network into the INeS *.net* format and followed the same methodology as in Section 5.4, performing $K = 10$ network induction experiments with distance correlation, the χ^2 measure, and information gain at the same sample sizes as above ($N = 1000, 1500, 2000, 2500, 3000$). Figure 5.12 contains the attribute ids and number of levels for the ALARM dataset; and Figure 5.13 includes an example network induced using distance correlation. The most obvious result is once again that networks learned with distance correlation miss fewer edges than networks induced with either information gain or the χ^2 measure.

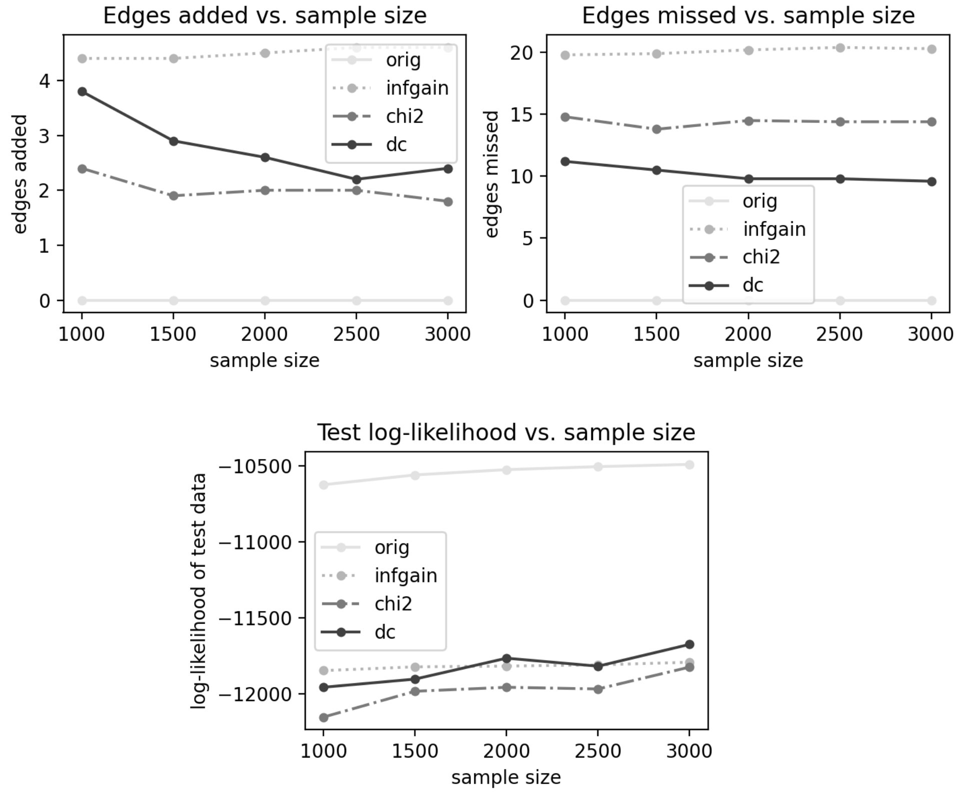
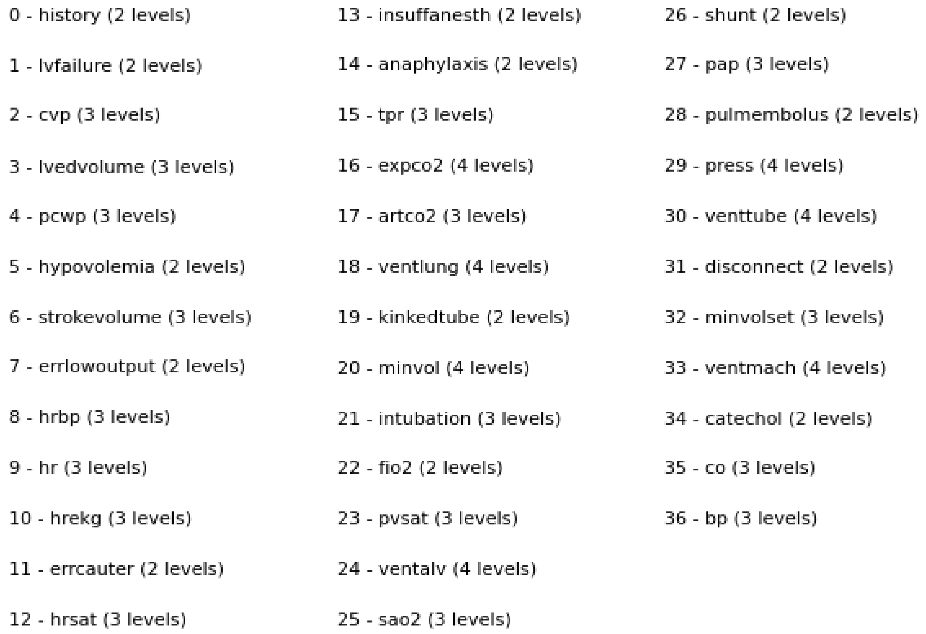


Figure 5.11: Average performance on ALARM data (Cheng-Bell-Liu search)



0 - history (2 levels)

1 - lvfailure (2 levels)

2 - cvp (3 levels)

3 - lvedvolume (3 levels)

4 - pcwp (3 levels)

5 - hypovolemia (2 levels)

6 - strokevolume (3 levels)

7 - errlowoutput (2 levels)

8 - hrbp (3 levels)

9 - hr (3 levels)

10 - hrekg (3 levels)

11 - errcauter (2 levels)

12 - hrsat (3 levels)

13 - insuffanesth (2 levels)

14 - anaphylaxis (2 levels)

15 - tpr (3 levels)

16 - expco2 (4 levels)

17 - artco2 (3 levels)

18 - ventlung (4 levels)

19 - kinkedtube (2 levels)

20 - minvol (4 levels)

21 - intubation (3 levels)

22 - fio2 (2 levels)

23 - pvsat (3 levels)

24 - ventalv (4 levels)

25 - sao2 (3 levels)

26 - shunt (2 levels)

27 - pap (3 levels)

28 - pulmembolus (2 levels)

29 - press (4 levels)

30 - venttube (4 levels)

31 - disconnect (2 levels)

32 - minvolset (3 levels)

33 - ventmach (4 levels)

34 - catechol (2 levels)

35 - co (3 levels)

36 - bp (3 levels)

Figure 5.12: ALARM attributes

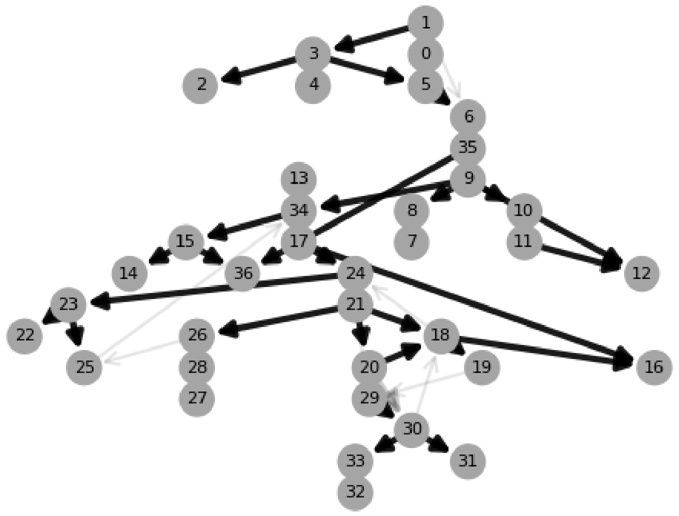


Figure 5.13: ALARM network induced with distance correlation

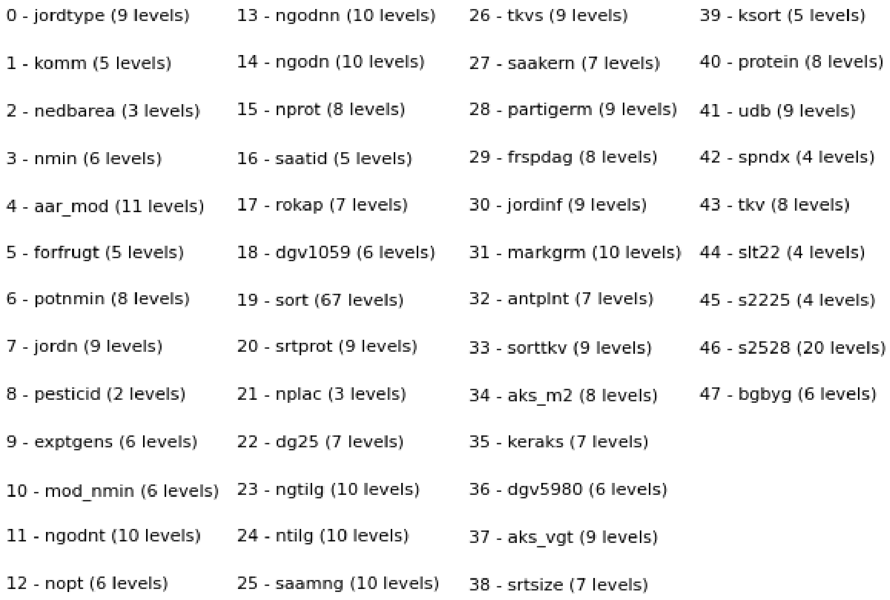
5.6 The barley data set

Description

Kristensen et al. studied pesticide use when producing beer from Danish malting barley in [34]; the associated data are available for download at [35]. Their goal was to create a decision support system that would enable growers to predict yield and quality when growing barley without pesticides in the presence of observables such as fungal diseases, weed infestations, the techniques used for weed control and cultivation. Figure 5.14 reports the attribute metadata and Figure 5.15 shows the complexity of the true network. Compared to either prior data sets, there are significantly more attributes, edges, and numbers of levels per edge.

Experimental results

We can see from Figure 5.16 that, as with the Danish Jersey cattle and ALARM data, in every case networks induced using distance correlation missed fewer edges than networks induced with either the χ^2 measure or information gain. This was especially true at the larger sample sizes, though the greater accuracy did not lead to a dramatic difference in log-likelihood.



0 - jordtype (9 levels)
1 - komm (5 levels)
2 - nedbarea (3 levels)
3 - nmin (6 levels)
4 - aar_mod (11 levels)
5 - forfrugt (5 levels)
6 - potnmin (8 levels)
7 - jordn (9 levels)
8 - pesticid (2 levels)
9 - exptgens (6 levels)
10 - mod_nmin (6 levels)
11 - ngodnt (10 levels)
12 - nopt (6 levels)
13 - ngodnn (10 levels)
14 - ngodn (10 levels)
15 - nprot (8 levels)
16 - saatid (5 levels)
17 - rokap (7 levels)
18 - dgv1059 (6 levels)
19 - sort (67 levels)
20 - srtprot (9 levels)
21 - nplac (3 levels)
22 - dg25 (7 levels)
23 - ngtilg (10 levels)
24 - ntilg (10 levels)
25 - saamng (10 levels)
26 - tkvs (9 levels)
27 - saakern (7 levels)
28 - partigerm (9 levels)
29 - frspdag (8 levels)
30 - jordinf (9 levels)
31 - markgrm (10 levels)
32 - antplnt (7 levels)
33 - sorttkv (9 levels)
34 - aks_m2 (8 levels)
35 - keraks (7 levels)
36 - dgv5980 (6 levels)
37 - aks_vgt (9 levels)
38 - srtsize (7 levels)
39 - ksort (5 levels)
40 - protein (8 levels)
41 - udb (9 levels)
42 - spndx (4 levels)
43 - tkv (8 levels)
44 - slt22 (4 levels)
45 - s2225 (4 levels)
46 - s2528 (20 levels)
47 - bgbyg (6 levels)

Figure 5.14: Barley attributes

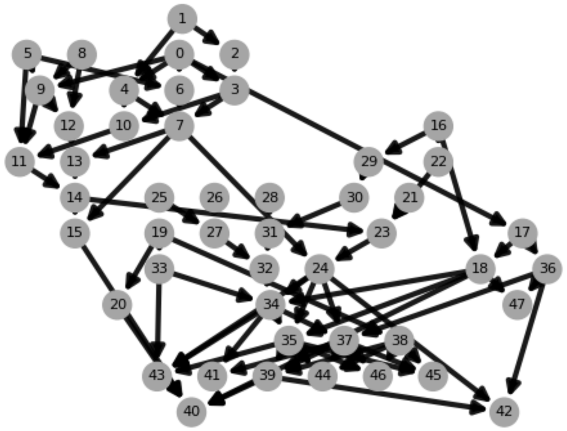


Figure 5.15: True network for barley data

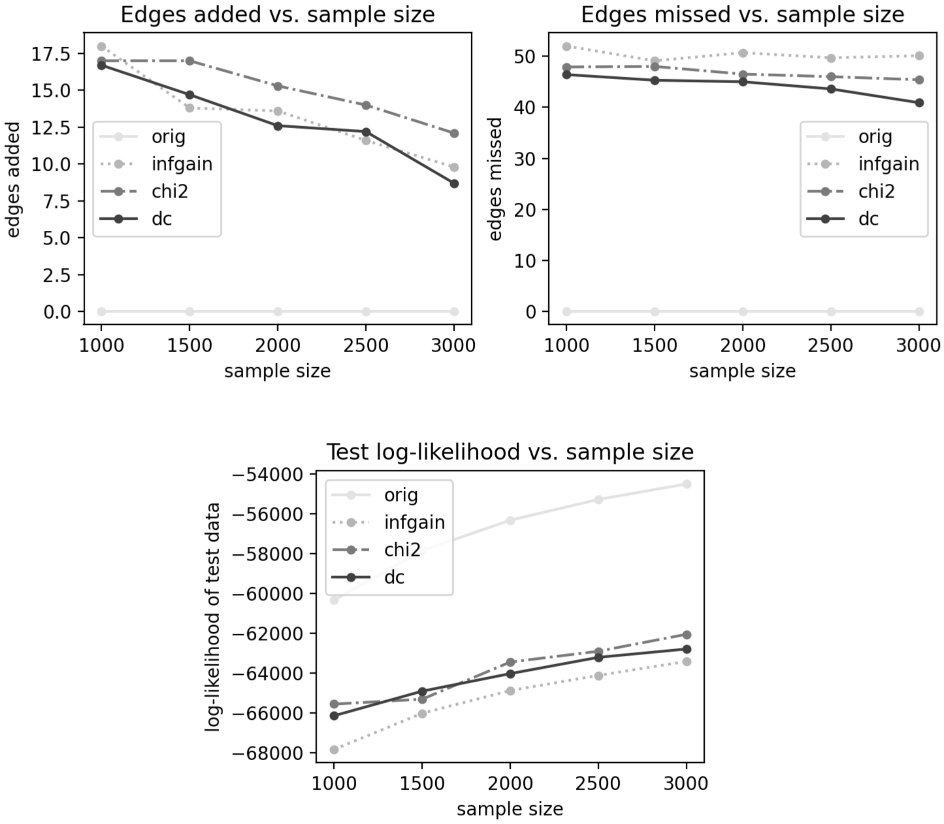


Figure 5.16: Average performance on barley data (Cheng-Bell-Liu search)

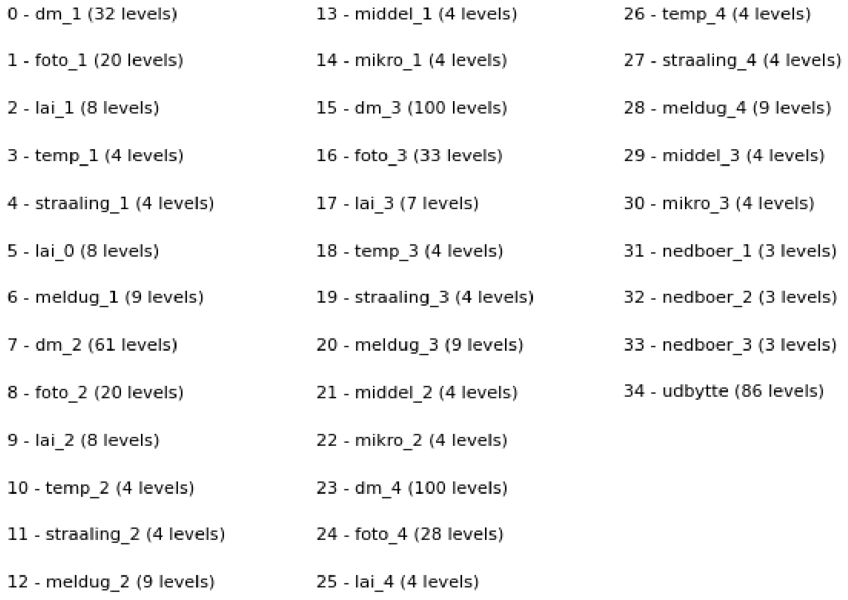
5.7 The mildew data set

Description

In [30], Jensen et al. used a Bayesian network to model the interactions of various environmental factors with the timing and dosing of mildew treatments in wheat. Figure 5.17 displays the 35 attributes they identified; note that some attributes have significantly more levels than in the datasets above. For example, the dm_4 attribute has 100 levels reflecting a wide range of choices of kilograms used per square meter of a particular mildew treatment. Figure 5.18 visualizes the structure of this network.

Experimental results

On this data set, the performance of networks induced with distance correlation is mostly equivalent to that of networks induced with the χ^2 measure and information gain. The only clear difference is in the average number of edges added, where at smaller sample sizes, distance correlation does noticeably better. This may be an artifact of the particularly high risk of sparseness when conditioning on attributes with many levels.



0 - dm_1 (32 levels)
1 - foto_1 (20 levels)
2 - lai_1 (8 levels)
3 - temp_1 (4 levels)
4 - straaling_1 (4 levels)
5 - lai_0 (8 levels)
6 - meldug_1 (9 levels)
7 - dm_2 (61 levels)
8 - foto_2 (20 levels)
9 - lai_2 (8 levels)
10 - temp_2 (4 levels)
11 - straaling_2 (4 levels)
12 - meldug_2 (9 levels)
13 - middel_1 (4 levels)
14 - mikro_1 (4 levels)
15 - dm_3 (100 levels)
16 - foto_3 (33 levels)
17 - lai_3 (7 levels)
18 - temp_3 (4 levels)
19 - straaling_3 (4 levels)
20 - meldug_3 (9 levels)
21 - middel_2 (4 levels)
22 - mikro_2 (4 levels)
23 - dm_4 (100 levels)
24 - foto_4 (28 levels)
25 - lai_4 (4 levels)
26 - temp_4 (4 levels)
27 - straaling_4 (4 levels)
28 - meldug_4 (9 levels)
29 - middel_3 (4 levels)
30 - mikro_3 (4 levels)
31 - nedboer_1 (3 levels)
32 - nedboer_2 (3 levels)
33 - nedboer_3 (3 levels)
34 - udbytte (86 levels)

Figure 5.17: Mildew attributes

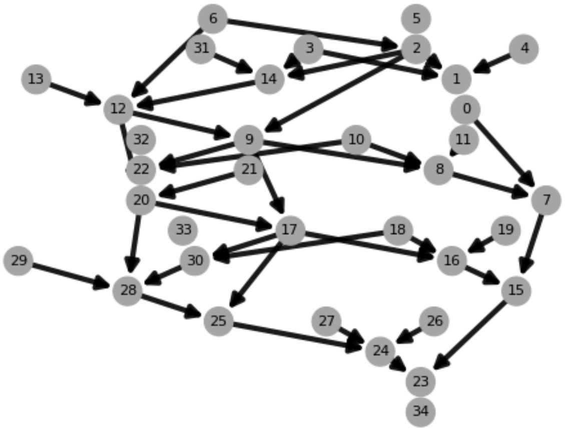


Figure 5.18: True network for mildew data

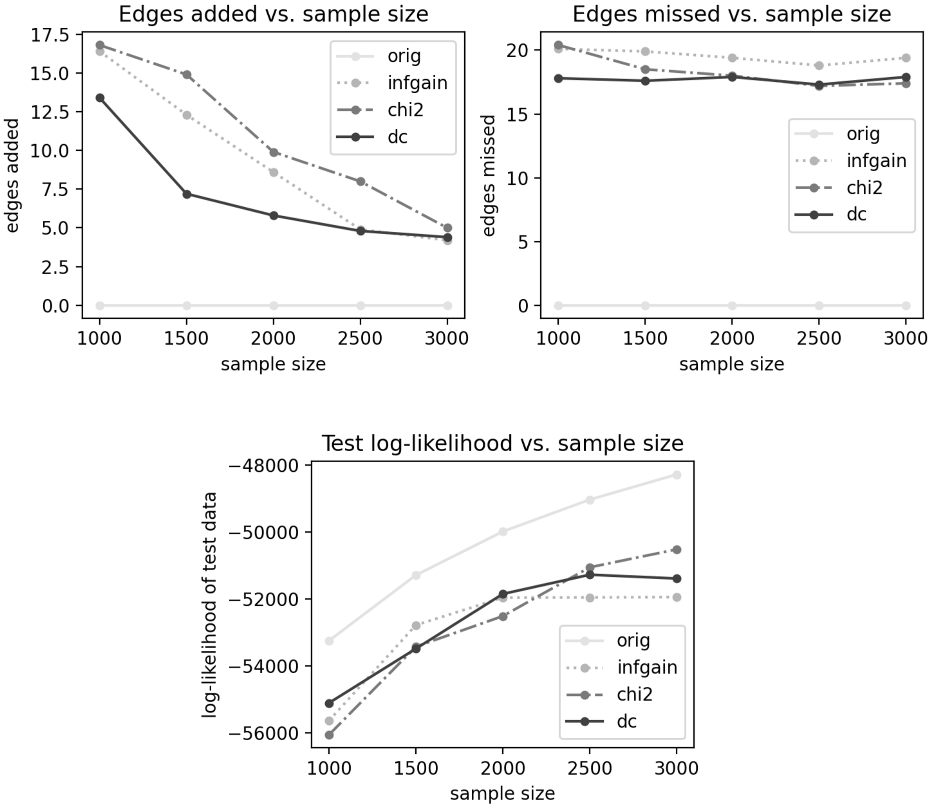


Figure 5.19: Average performance on mildew data (Cheng-Bell-Liu search)

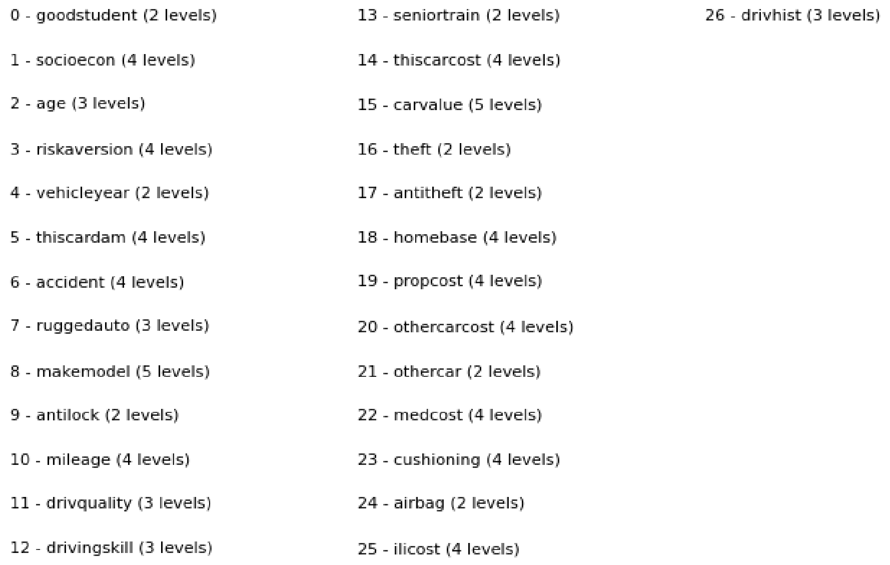
5.8 The insurance dataset

Description

This dataset comes from a study by Binder et al. in [44] on the use of prior knowledge about structure to improve the learning rate of a probabilistic network. It features a Bayesian network of 27 attributes from the auto insurance domain, where the goal is to support decision-making in insurance pricing. Example attributes include a (hidden) node `riskaversion` with levels such as `cautious` and `adventurous`. The attribute list is in Figure 5.20, and the true network is in Figure 5.21, where for example we see the dependence of `riskaversion` on `age` and `socioecon`.

Experimental results

The results here in Figure 5.22 continue to give empirical evidence of the superior power of distance correlation in testing conditional independence, as networks learned with distance correlation continue to miss the fewest edges on average. Perhaps because some of the edges missed by the χ^2 measure and information gain represent particularly strong dependences, on this data set we see that average log-likelihood of the test data is substantially higher for networks induced with distance correlation.



0 - goodstudent (2 levels)

1 - socioecon (4 levels)

2 - age (3 levels)

3 - riskaversion (4 levels)

4 - vehicleyear (2 levels)

5 - thiscardam (4 levels)

6 - accident (4 levels)

7 - ruggedauto (3 levels)

8 - makemodel (5 levels)

9 - antilock (2 levels)

10 - mileage (4 levels)

11 - drivquality (3 levels)

12 - drivingskill (3 levels)

13 - seniortrain (2 levels)

14 - thiscarcost (4 levels)

15 - carvalue (5 levels)

16 - theft (2 levels)

17 - antitheft (2 levels)

18 - homebase (4 levels)

19 - propcost (4 levels)

20 - othercarcost (4 levels)

21 - othercar (2 levels)

22 - medcost (4 levels)

23 - cushioning (4 levels)

24 - airbag (2 levels)

25 - ilicost (4 levels)

26 - drivhist (3 levels)

Figure 5.20: Insurance attributes

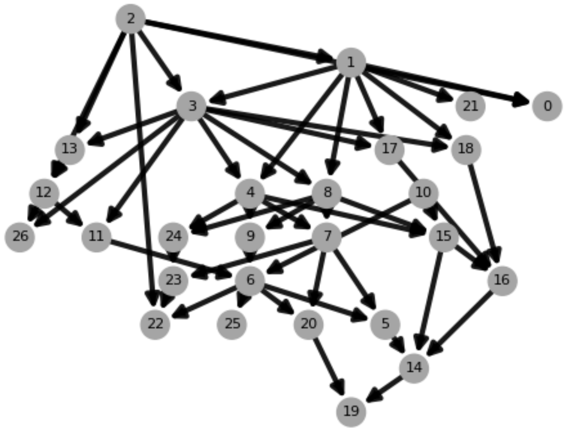


Figure 5.21: True network for insurance data

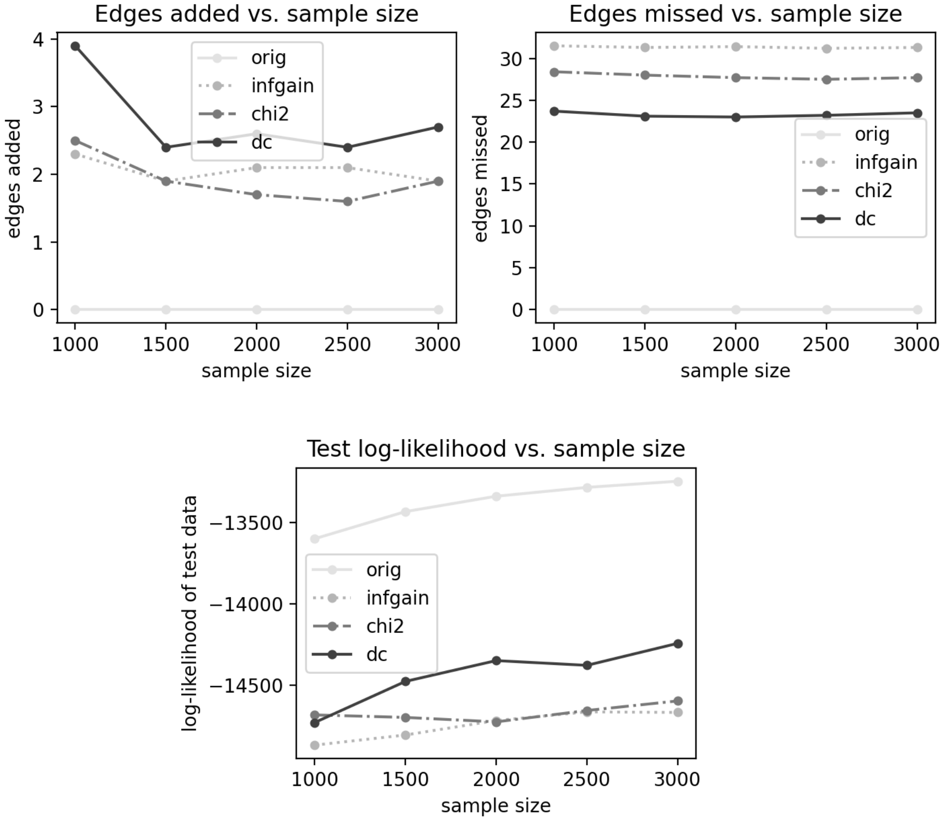


Figure 5.22: Average performance on insurance data (Cheng-Bell-Liu search)

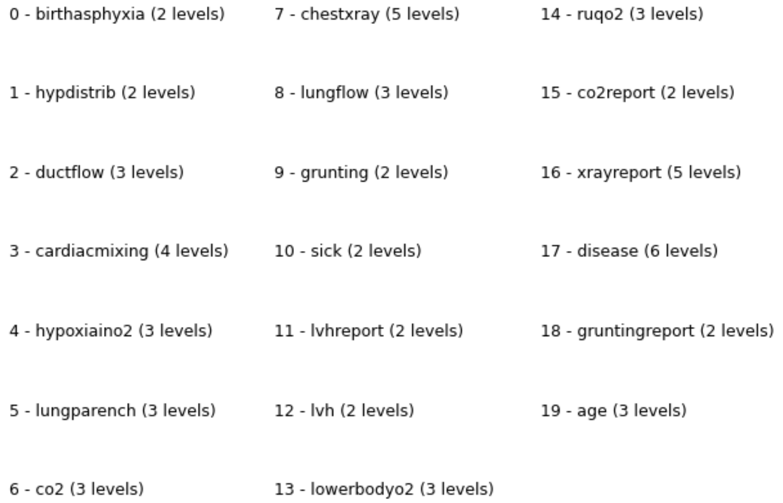
5.9 The child data set

Description

In [48], Spiegelhalter et al. describe a Bayesian network used to aid diagnosis of congenital heart disease in children born with certain symptoms. The key presenting symptom is `birthasphyxia` with levels `yes` and `no`, and its child node `disease` then has six levels reflecting various possible diagnoses. Other observables are present in Figure 5.23. We can see in Figure 5.24 that the overall network is somewhat more structured than in most examples above.

Experimental results

Much as with the insurance data set above, this is a case in which networks learned with distance correlation enjoy substantially better performance in terms of test log-likelihood; at the larger sample sizes, the test log-likelihood using distance correlation is quite close to the log-likelihood of the test data relative to the true network. Once again the simplest explanation lies in the consistently better performance on edges missed. Compared to information gain, networks learned with distance correlation miss approximately 50% fewer edges on average; and compared to networks learned with the χ^2 measure, approximately 33% fewer edges.



0 - birthasphyxia (2 levels)

1 - hypdistrib (2 levels)

2 - ductflow (3 levels)

3 - cardiacmixing (4 levels)

4 - hypoxiaino2 (3 levels)

5 - lungparench (3 levels)

6 - co2 (3 levels)

7 - chestxray (5 levels)

8 - lungflow (3 levels)

9 - grunting (2 levels)

10 - sick (2 levels)

11 - lvhreport (2 levels)

12 - lvh (2 levels)

13 - lowerbodyo2 (3 levels)

14 - ruqo2 (3 levels)

15 - co2report (2 levels)

16 - xrayreport (5 levels)

17 - disease (6 levels)

18 - gruntingreport (2 levels)

19 - age (3 levels)

Figure 5.23: Child attributes

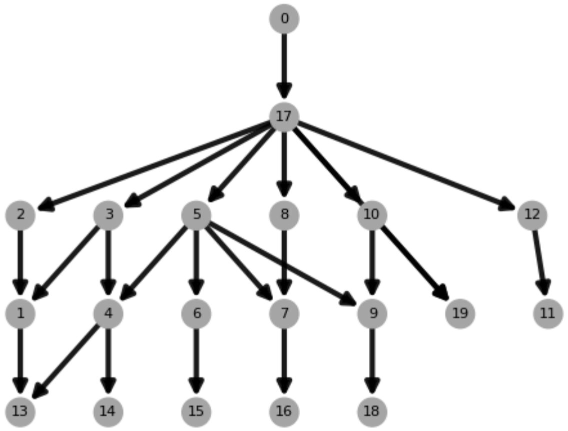


Figure 5.24: True network for child data

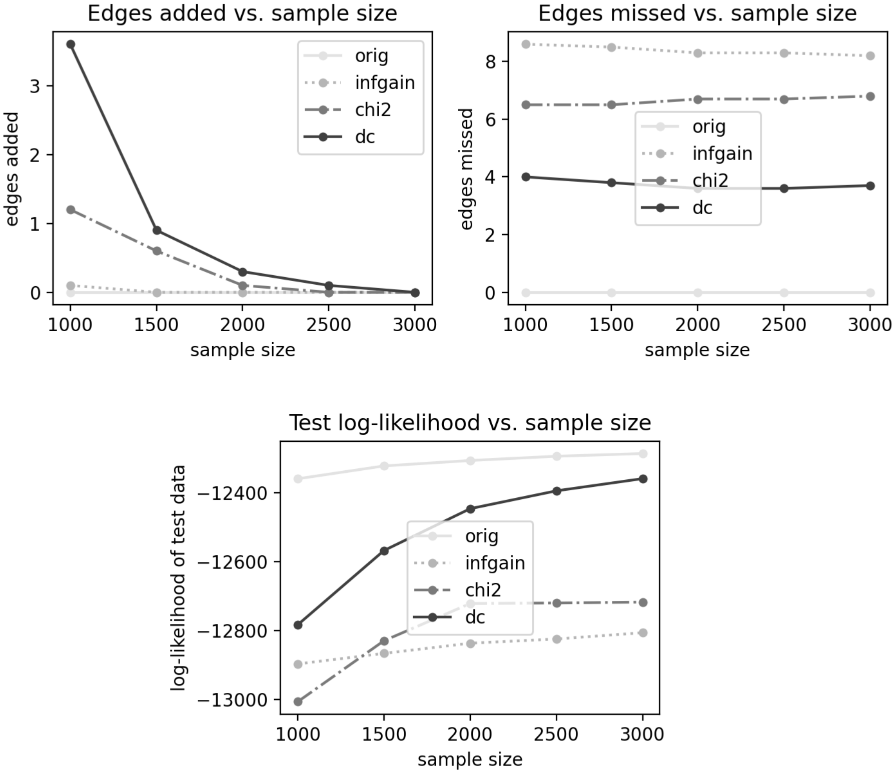


Figure 5.25: Average performance on child data (Cheng-Bell-Liu search)

5.10 Results with varying independence thresholds

All the above experiments were conducted with the standard independence threshold of $t = 0.1$. It is interesting to also explore how performance on various metrics changes with different thresholds. Intuitively, with a higher threshold there is less risk of adding edges, as it is harder to reject the null hypothesis of independence when performing tests. But of course this comes with an increased risk of missing edges, since we may not recognize some weaker dependences or conditional dependences based on the test with a higher threshold. In the figures following below, we display the results of experiments for each of the six datasets above using a sample size of $N = 3000$ and a sequence of independence thresholds $t = 0.05, 0.10, 0.15, 0.20, 0.25$. The overall conclusion is quite clear—the conventional choice of $t = 0.10$ represents a good average-case tradeoff for edges added and edges missed. There are special cases, as with the Danish Jersey cattle data and distance correlation, we can see some overall improvement with higher thresholds. But in general $t = 0.10$ appears a good choice.

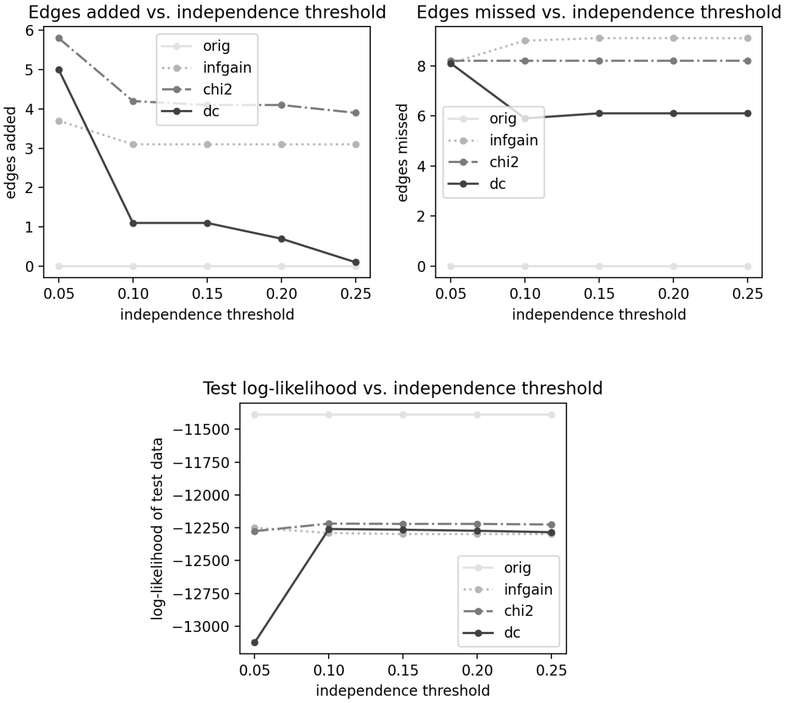


Figure 5.26: Performance of various thresholds with the Danish Jersey cattle data

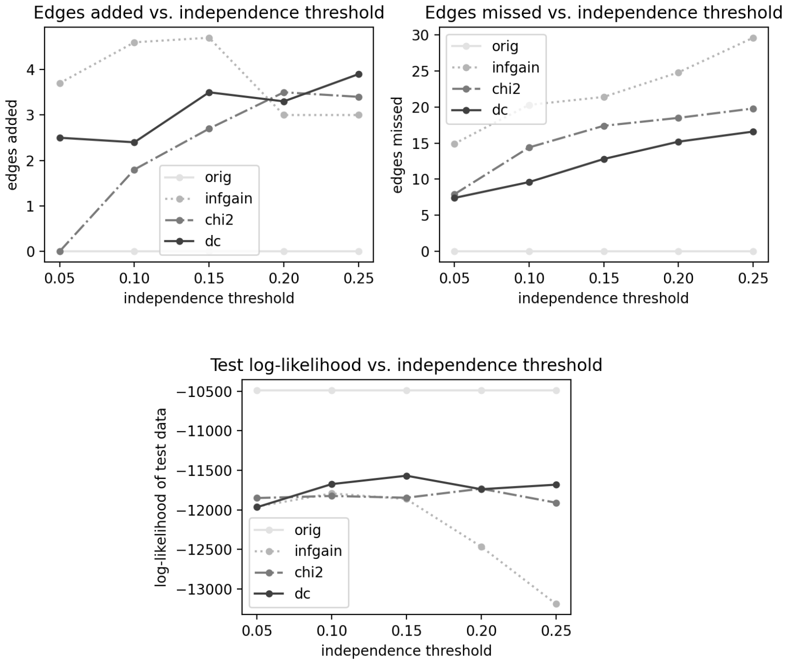


Figure 5.27: Performance of various thresholds with the ALARM data

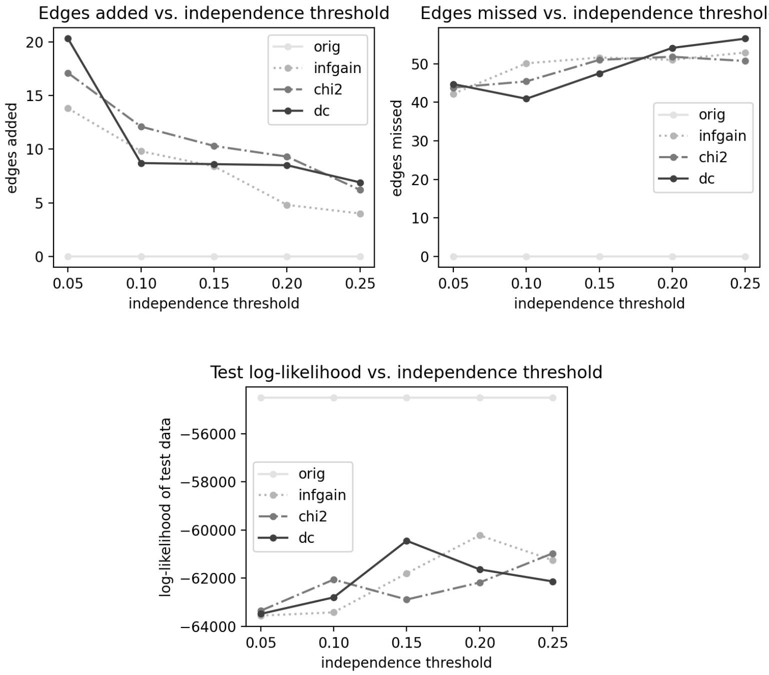


Figure 5.28: Performance of various thresholds with the barley data

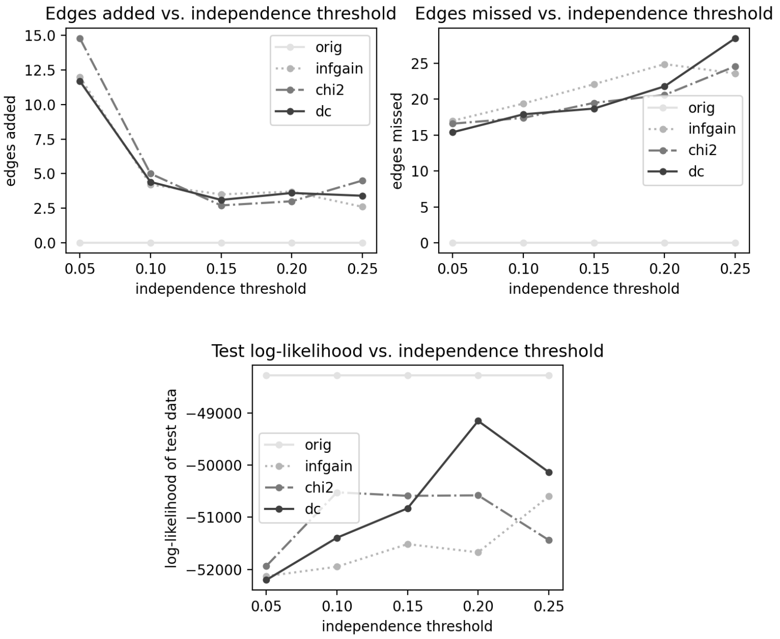


Figure 5.29: Performance of various thresholds with the mildew data

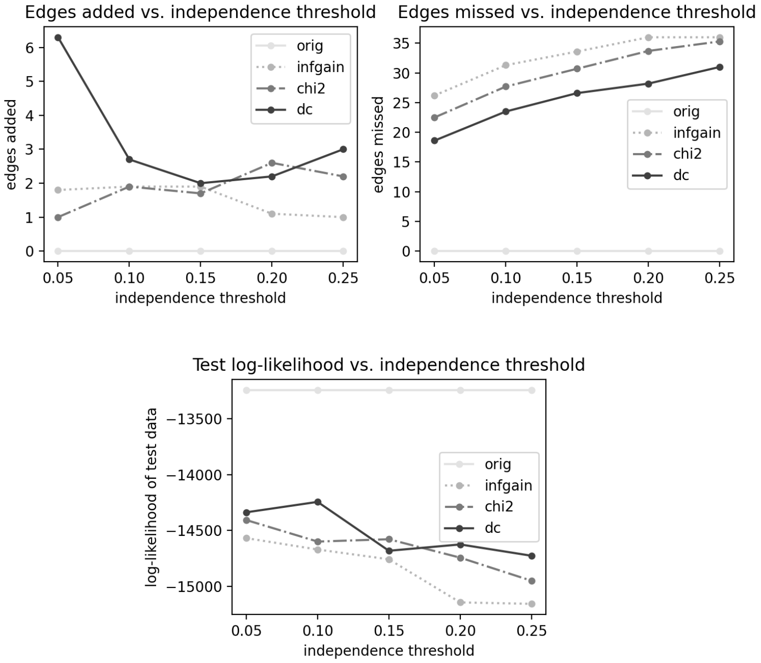


Figure 5.30: Performance of various thresholds with the insurance data

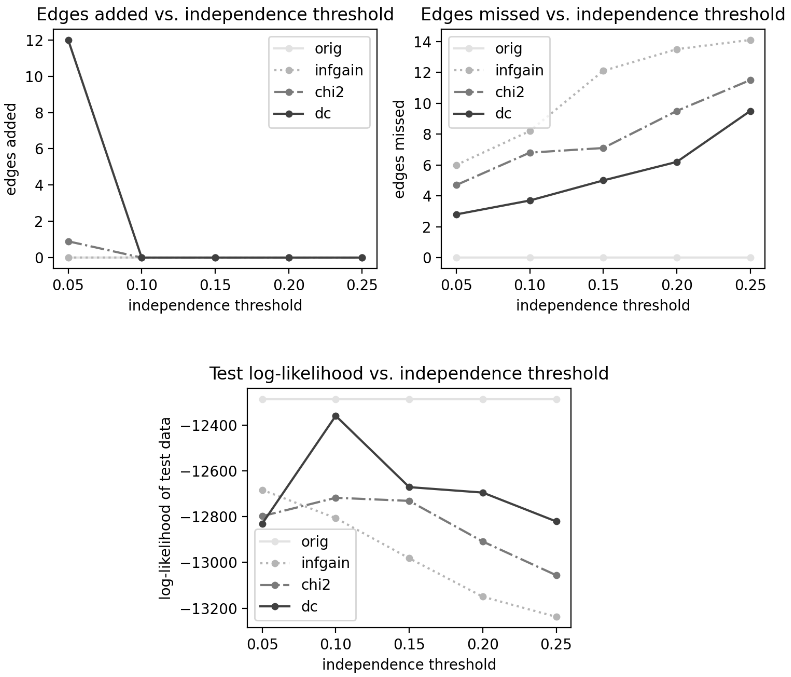


Figure 5.31: Performance of various thresholds with the child data

5.11 Computational cost of the distance correlation measure

There are many factors that influence the time needed to learn a network. Certainly the size of the training data is important. Assuming the behavior of the learning algorithm is reasonably correlated to the nature of the true network, the factors will also include number of attributes, edges, and parameters in the true underlying network. The number of independence tests and conditional independence tests performed by the search algorithm will also have a large impact. Even given the same data and the same algorithm, it is also possible to implement an algorithm more or less efficiently; and of course the underlying hardware being used will affect the running time.

The consequence of this complexity is that it is hard to draw any clear conclusion about the relative computational cost of learning networks with distance correlation rather than the χ^2 measure or information gain. Since our experiments indicate that, on average, distance correlation misses fewer edges from the true network than the χ^2 measure or information gain, we might expect it to do more conditional independence tests during execution of the Cheng-Bell-

Liu algorithm and require somewhat more time. As a basic comparison of computational cost, we selected five datasets with varied characteristics from the BNLEARN repository of Bayesian networks, available for download from `https://www.bnlearn.com/bnrepository`. The runtime of using INeS to learn networks from training data of size $N = 3000$ for each network appears in the table below. (The author's system is a MacBook Pro with a 2.3 GHz 8-Core Intel Core i9 processor and 16GB DDR4 memory.)

Network	Attributes	Edges	Parameters	`infgain` secs	χ^2 secs	`dc` secs
`ALARM`	37	46	509	0.232	0.247	0.398
`HAILFINDER`	56	66	2656	0.388	0.419	0.769
`HEPAR2`	70	123	1453	0.484	0.484	0.482
`BARLEY`	48	84	114005	1.836	3.203	2.885
`MILDEW`	35	46	540150	7.702	10.262	7.840

5.12 Conclusions

In summary, there are several main conclusions we may draw from these empirical studies.

1. The most consistent and obvious result is the superior performance of distance correlation on the "edges missed" metric. Of the 30 experiments we performed (five different sample sizes for each of six datasets), in 29 of them the networks induced with distance correlation missed fewer edges from the true network, on average, than networks induced with either information gain or the χ^2 measure.

2. There are no clear trends in the relative performance of distance correlation versus information gain or the χ^2 measure when comparing the log-likelihood on the test data or the number of edges added to the true network.

3. Combined with the Cheng-Bell-Liu search algorithm, distance correlation is able to reliably take advantage of increased sample data and learn more accurate networks, as evidenced by the plotted curves across our six datasets.

Chapter 6

Distance Correlation and Sparse Three-Way Contingency Tables

In this chapter we return to our joint work with Zhang in [58] on the use of distance covariance for conditional independence tests and marginal distance covariance for homogeneity tests. It may help give some intuition for why the networks induced in Chapter 5 using distance correlation could achieve better performance to networks induced with other measures. All the results here build on Zhang's work in [57] translating distance correlation to the categorical setting, which we outlined in Section 1.4. We also draw on the work of Wang et al. in [50] for mathematical context. The primary goal is to move from two categorical variables X and Y in isolation, to considering their dependence structure conditioned on a third categorical variable Z. Many of the calculations and derivations are analogous to the unconditioned case, which is why we omitted some details in the background material, and present them here in a more general context.

In [57], Zhang also carried out a number of numerical studies comparing the empirical power of an independence test based on T_{dCor} from Equation 1.7 with other previously developed tests. For example, given nominal X and Y with varying degrees of dependence and sample sizes of varying sparsity, he compared with Pearson's χ^2 test at a significance level of $\alpha = 0.05$. Note that the p-value we compare to α when using T_{dCor} is computed differently than in Pearson's χ^2 test. In Pearson's test, the statistic has an explicit asymptotic null distribution. However, the asymptotic null distribution of T_{dCor} depends on the true distributions of X and Y; so if it was available, there would be no need for an independence test! (See again [42] for more details.) So instead we use a bootstrap procedure. That is, we permute the sample $Y_1, \ldots, Y_n$ m times; compute a T'_{dCor} for each permutation; and let the p-value for T_{dCor} be the proportion of permutations with a smaller T'_{dCor}.

6.1 Tests in three-way contingency tables

There is an extensive literature analyzing contingency tables with three-way classifications, particularly in the special case of a $2 \times 2 \times K$ table since this is a useful framework for testing association between treatment (for example, drug vs. placebo) and response (for example, success vs. failure) while controlling for the effects of some covariates (for example, clinic, gender, and age group). Similarly in genetic association studies, it is crucial to quantify the association between genotype and phenotype while adjusting for environmental factors. In the most general case, the analysis concerns the association between two categorical variables $X \in \{1, 2, \dots, R\}$ and $Y \in \{1, 2, \dots, C\}$ while controlling for a possibly confounding variable $Z \in \{1, 2, \dots, K\}$, and regularly appears in scientific contexts.

Two of the most popular traditional tests are the Cochran-Mantel-Haenszel test and the conditional mutual information test. They are relatively simple to use, and perform quite well given a large sample.

Cochran-Mantel-Haenszel (CMH) test

The CMH test was first proposed in 1959 for the special case of $2 \times 2 \times K$ tables and later generalized to $R \times C \times K$ tables (see [1]). It is essentially a score test that attempts to summarize the dependence information from the K partial tables of size $R \times C$. For the case $R = C = 2$, the test statistic is simply,

$$\mathrm{CMH} = \frac{\left[\sum_{k=1}^{K} \left(n_{11k} - E\left(n_{11k}\right)\right)\right]^2}{\sum_{k=1}^{K} \mathrm{Var}\left(n_{11k}\right)},$$

where $E\left(n_{11k}\right)$ and $\mathrm{Var}\left(n_{11k}\right)$ can be estimated as follows:

$$\begin{aligned}\widehat{E\left(n_{11k}\right)} &= n_{1+k} n_{+1k} / n_{++k} \\ \widehat{\mathrm{Var}(n_{11k})} &= n_{1+k} n_{2+k} n_{+1k} n_{+2k} / \left(n_{++k}^2 (n_{++k} - 1)\right).\end{aligned}$$

Under the null hypothesis of independence and for large sample sizes, this statistic approximately follows a χ^2 distribution with one degree of freedom, and the approximation improves as the total sample size n increases, regardless of whether K is small or large.

To generalize the statistic, we let $\mathbf{n}_k$ be the vector of $(R-1)(C-1)$ free counts in the kth partial table, that is,

$$\mathbf{n}_k = \left(n_{11k}, n_{12k}, \dots, n_{R-1,C-1,k}\right)^T,$$

with μ_k the expectation of $\mathbf{n}_k$ given conditional independence. The sample estimate of μ_k can be obtained as

$$\widehat{\mu}_k = \left(n_{1+k}n_{+1k}, n_{1+k}n_{+2k}, \ldots, n_{R-1,+,k}n_{+,C-1,k}\right)^T / n_{++k}.$$

Again assuming conditional independence, the covariance matrix $\mathbf{V}_k$ of $\mathbf{n}_k$ is made up of elements estimable by the formula,

$$\widehat{\text{Cov}}\left(n_{rck}, n_{r'c'k}\right) = \frac{n_{r+k}n_{+ck}\left(\delta_{rr'}n_{++k} - n_{r'+k}\right)\left(\delta_{cc'}n_{++k} - n_{+c'k}\right)}{n_{++k}^2\left(n_{++k} - 1\right)},$$

where $\delta_{ab} = 1$ if $a = b$ and $\delta_{ab} = 0$ otherwise. Then the generalized CMH statistic is,

$$CMH = (\mathbf{n} - \widehat{\mu})^T\widehat{\mathbf{V}}^{-1}(\mathbf{n} - \widehat{\mu}),$$

where $\mathbf{n} = \sum_{k=1}^K \mathbf{n}_k$, $\widehat{\mu} = \sum_{k=1}^K \widehat{\mu}_k$, and $\hat{\mathbf{V}} = \sum_{k=1}^K \widehat{\mathbf{V}}_k$. The general CMH statistic approximately follows a χ^2 distribution with $(R-1)(C-1)$ degrees of freedom.

The conditional mutual information (CMI) test

CMI is an information-theoretic measure of conditional dependence. It has been applied to many statistical problems, including the structural recovery of discrete networks, as discussed in Section 3.1. The mutual information between X and Y given Z is defined as

$$I(X;Y|Z) = \sum_{1\leq r\leq R}\sum_{1\leq c\leq C}\sum_{1\leq k\leq K} \pi_{rck}\log\frac{\pi_{++k}\pi_{rck}}{\pi_{r+k}\pi_{+ck}},$$

and it is known that $I(X;Y|Z) = 0$ iff X and Y are independent conditioning on Z (see [15]). The empirical estimate for $I(X;Y|Z)$ is naturally,

$$\hat{\mathrm{I}}(\mathrm{X};\mathrm{Y}|\mathrm{Z}) = \frac{1}{n}\sum_{1\leq r\leq R}\sum_{1\leq c\leq C}\sum_{1\leq k\leq K} n_{rck}\log\frac{n_{++k}n_{rck}}{n_{r+k}n_{+ck}},$$

where we define $n_{rck}\log\left\{n_{++k}n_{rck}/(n_{r+k}n_{+ck})\right\} = 0$ if $n_{\mathrm{rck}} = 0$. Note that when $n_{\mathrm{rck}} > 0$, the term $n_{++k}n_{\mathrm{rck}}/\left(n_{r+k}n_{+ck}\right) > 0$ as n_{r+k}, n_{+ck}, and n_{++k} are all positive.

Given a large sample, one can normalize $I(X;Y|Z)$ by the marginal entropies $H(X,Z) + H(Y,Z)$ and then apply Fisher's Z transformation to calculate the p value; where the asymptotic distribution under H_0 is $N(0, 1/\sqrt{n-K-3})$. For smaller samples, it is preferable to use a permutation test to evaluate the significance (see [27]).

Motivation for new tests

As mentioned several times above in discussing CMH and CMI, these traditional tests of conditional independence need large samples to perform well. For three-way tables with many sparse counts, the tests become overly conservative and lose power (as will be clearly visible in the simulation studies we present later in this chapter). It is thus desirable to find test statistics based on measures which remain sensitive to dependence relations even with less data available. This is our goal in what follows, as we will derive the explicit formula for conditional distance covariance between nominal X and Y given some Z, and propose two types of statistics based on this measure for testing the hypothesis of conditional independence for X and Y. We will also propose a new statistic for testing homogeneity of XY at different levels of Z.

6.2 Definition and properties of conditional distance correlation

Wang et al. first extended distance covariance to conditional distance covariance in [50]. The result is a nonparametric measure of conditional dependence for multivariate random variables with arbitrary dimensions, retaining the property that measured correlation is zero almost surely iff two multivariate random variables are conditionally independent given a third random variable.

Definition 6.1. *The conditional distance covariance (CDCov), notated $\mathcal{D}(X,Y|Z)$, between random vectors X and Y with finite moments given Z is defined as the square root of,*

$$\begin{aligned}\mathcal{D}^2(X,Y|Z) &= \left\|\phi_{X,Y|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)\right\|^2, \\ &= \tfrac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \tfrac{\left|\phi_{X,Y|Z}(t,s) - \phi_{X|Z}(t)\phi_{Y|Z}(s)\right|^2}{|t_p|_p^{p+1}|s|_q^{q+1}}\, dt\, ds,\end{aligned} \tag{6.1}$$

where $c_p = \frac{\pi(p+1)/2}{\Gamma((p+1)/2)}$ and $c_q = \frac{\pi(q+1)/2}{\Gamma((q+1)/2)}$. Similarly, the conditional distance variance (CDVar) of a random vector X with finite moments is is defined as the square root of,

$$\mathcal{D}^2(X|Z) = \mathcal{D}^2(X,X|Z). \tag{6.2}$$

The definition of conditional distance correlation is also completely analogous to the unconditioned case.

Definition 6.2. *The conditional distance correlation (CDCor) between random vectors X and Y with finite moments given Z is defined as the square root of*

$$\rho^2(X,Y|Z) = \frac{\mathcal{D}^2(X,Y|Z)}{\sqrt{\mathcal{D}^2(X|Z)\mathcal{D}^2(Y|Z)}}, \tag{6.3}$$

if the denominator in Equation 6.3 is not zero; or 0 *otherwise.*

Wang et al. demonstrated that the CDCov measure is nonnegative and equals zero iff X and Y are independent conditioning on Z, as desired. It is important to note that CDCov is not the same as the notion of partial distance correlation developed by Szekel and Rizzo in [49]. Partial distance correlation (PDC) is defined as

$$R^*(X,Y|Z) = \frac{\mathcal{R}^2(X,Y) - \mathcal{R}^2(X,Z)R^2(Y,Z)}{\sqrt{1-\mathcal{R}^4(X,Z)}\sqrt{1-\mathcal{R}^4(Y,Z)}},$$

where $\mathcal{R}^2(X,Y) = \mathcal{V}^2(X,Y)/(\mathcal{V}(X,X)\mathcal{V}(Y,Y))$ represents the marginal distance correlation between X and Y. In particular, PDC is defined in a similar manner to Pearson's correlation coefficient; and in general $\mathcal{R}^*(X,Y|Z) = 0$ does not imply conditional independence. We will compare CDCor and PDC using numerical studies in what follows.

6.3 Translation to three-way contingency tables

Our main task is now to translate the above concepts into a form meaningful for categorical variables whose joint outcomes form a $R \times C \times K$ contingency table. We will wish to test a hypothesis of conditional independence, as follows. Let $\pi_{\mathrm{rc|k}}, \pi_{\mathrm{r+|k}}$ and $\pi_{\mathrm{+c|k}}$ be the joint and marginal probabilities in the kth partial table, that is, the two-way XY table obtained by fixing Z. The null hypothesis of conditional independence of X and Y given Z, versus the alternative hypothesis of conditional dependence, is then formulated as

$$H_0 : \pi_{\mathrm{rc|k}} = \pi_{\mathrm{r+|k}} * \pi_{\mathrm{+c|k}} \text{ for } 1 \leq r \leq R, 1 \leq c \leq C, 1 \leq k \leq K,$$

vs.

$$H_\alpha : \pi_{rck} \neq \pi_{r+|k} * \pi_{+c|k} \text{ for some } (r,c,k).$$

In the special case of a $2 \times 2 \times K$ table, the null hypothesis is equivalent to $\theta_1 = \theta_2 = \ldots = \theta_k = 1$, where $\theta_k = \pi_{11|k}\pi_{22|k}/(\pi_{12|k}\pi_{21|k})$ is the odds ratio in the k-th partial table; because independence is equivalent to an odds ratio of one.

Again using the geometric intuition from Section 1.4, we rewrite X, Y, and Z as random vectors, $X = \{X_r\}_{1 \le r \le R}, Y = \{Y_c\}_{1 \le c \le C}$, and $Z = \{Z_k\}_{1 \le k \le K}$, where X_r, Y_c, and Z_k are indicator variables for categories r, c and k, respectively. The characteristic functions are as follows:

$$\phi_{X,Y|Z_k=1}(t,s) = E\left(e^{i<X,t>+i\langle Y,s\rangle}|Z_k = 1\right) = \sum_{r=1}^{R}\sum_{c=1}^{C} e^{i(t_r+s_c)}\pi_{rc|k},$$

$$\phi_{X|Z_k=1}(t) = E\left(e^{i<X,t>}\right) = \sum_{r=1}^{R} e^{it_r}\pi_{r+|k},$$

$$\phi_{Y|Z_k=1}(s) = E\left(e^{i\langle Y,s\rangle}\right) = \sum_{c=1}^{C} e^{is_c}\pi_{+c|k}.$$

(Here $< v_1, v_2 >$ represents the inner product of vectors v_1 and v_2.) So starting from Equation 6.1, we consider

$$\phi_{X,Y|Z_k=1}(t,s) - \phi_{X|Z_k=1}(t)\phi_{Y|Z_k=1}(s) = \sum_{r=1}^{R}\sum_{c=1}^{C} e^{i(t_r+s_c)}\left(\pi_{rc|k} - \pi_{r+|k}\pi_{+c|k}\right),$$

And continue,

$$|\phi_{X,Y|Z_k=1}(t,s) - \phi_{X|Z_k=1}(t)\phi_{Y|Z_k=1}(s)|^2 = \left[\sum_{r=1}^{R}\sum_{c=1}^{C} \cos\left(t_r + s_c\right)\left(\pi_{rc|k} - \pi_{r+|k}\pi_{+c|k}\right)\right]^2 + \left[\sum_{r=1}^{R}\sum_{c=1}^{C} \sin\left(t_r + s_c\right)\left(\pi_{rc|k} - \pi_{r+|k}\pi_{+c|k}\right)\right]^2.$$

Now, by the sum and differences formula from Lemma 1 in [42], we can further simply to

$$\mathcal{V}^2\left(X, Y|Z_k = 1\right) = 2\sum_{r=1}^{R}\sum_{c=1}^{C}\left(\pi_{rc|k} - \pi_{r+|k}\pi_{+c|k}\right)^2. \tag{6.4}$$

It follows that $\mathcal{V}^2\left(X, Y|Z_k = 1\right) = 0$ is equivalent to $\pi_{rc|k} = \pi_{r+|k}\pi_{+c|k}$ for all choices of (r, c); that is, equivalent to independence between X and Y in the kth partial table.

The expression in Equation 6.4 is in fact a natural extension of the marginal distance covariance between two categorical variables. Now to obtain an estimator for conditional distance covariance, we let n_{rck} be the count for cell (r, c, k), and write,

$$n_{rc+} = \sum_{k=1}^{K} n_{rck},$$
$$n_{r+k} = \sum_{c=1}^{C} n_{rck},$$
$$n_{+ck} = \sum_{r=1}^{R} n_{rck}.$$

That is, let the appearance of + in an index position denote the corresponding marginal sum, so that for example $n_{++k} = \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rck}$. We then reach the following definition by replacing the unknown parameters $(\pi_{rc|k}, \pi_{r+|k}, \pi_{+c|k})$ for all (r, c) with the maximum likelihood estimates.

Definition 6.3. *Given an i.i.d. sample of size n for vectors X, Y, and Z as above, the sample conditional distance covariance is*

$$\mathcal{V}_n^2\left(X, Y|Z_k = 1\right) = 2\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k} - p_{r+|k}p_{+c|k}\right)^2,$$

where $p_{rc|k} = n_{rck}/n_{++k}, p_{r+k} = n_{r+k}/n_{++k}$ and $p_{+c|k} = n_{+ck}/n_{++k}$.

6.4 Conditional independence tests with CDCov

We are now in a position to define tests for the hypothesis of conditional independence

$$H_0 : \mathcal{V}^2\left(X, Y|Z_k = 1\right) = 0 \quad \forall k = 1, 2, \ldots, K. \tag{6.5}$$

We propose two such tests.

Definition 6.4 (Average-type test statistic)**.** *Given X, Y, and Z as above, the average-type test statistic for H_0 in Equation 6.5 is,*

$$\sum_{k=1}^{K} p_k \mathcal{V}_n\left(X, Y|Z_k = 1\right) = \sum_{k=1}^{K} p_k \sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k} - p_{r+|k}p_{+c|k}\right)^2}.$$

Definition 6.5 (Maximum-type test statistic)**.** *Given X, Y, and Z as above, the average-type test statistic for H_0 in Equation 6.5 is,*

$$\max_k \mathcal{V}_n\left(X, Y | Z_k = 1\right) = \max_k \sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k} - p_{r+|k}p_{+c|k}\right)^2}.$$

Before these statistics can be used to test H_0, we need an understanding of their asymptotic distributions given H_0. The following lemma due to Biau and Gyorfi (see [5]) will be very useful.

Lemma 6.6. *Given n i.i.d. observations from a categorical variable $X \in \{1, 2, \ldots, I\}$, where $\pi_i = P(X = i)$ and $p_i = n_i/n$ are the true and estimated probabilities of X; for any $\epsilon > 0$, it holds that,*

$$P\left(\sum_{i=1}^{I} |\pi_i - p_i| > \epsilon\right) < 2^I e^{-n\epsilon^2/2}.$$

We can now prove the strong consistency of the proposed maximum-type test statistic.

Theorem 6.7. *With the notation from Lemma 6.6, under the multinomial model and null hypothesis of conditional independence, for any $\epsilon > 0$, it holds,*

$$P\left(\max_k \sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k} - p_{r+|k}p_{+c|k}\right)^2} > \epsilon\right) < \left(2^{RC} + 2^R + 2^C\right)\sum_{k=1}^{K} e^{\frac{-n_{++k}\epsilon^2}{18}}.$$

Proof. We begin with the triangle inequality,

$$\sum_{r=1}^{R}\sum_{c=1}^{C}|p_{rck} - p_{r+|k}p_{+c|k}| \leq$$
$$\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k} - \pi_{rc|k}\right| + \sum_{r=1}^{R}\sum_{c=1}^{C}\left|\pi_{rc|k} - \pi_{r+|k}\pi_{+c|k}\right| + \sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{r+|k}p_{+c|k} - \pi_{r+|k}\pi_{+c|k}\right|.$$

Under independence, the second term is zero; therefore, we only need bound the first and third terms. By Lemma 6.6, the first term can be bounded as follows:

$$P\left(\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k} - \pi_{rc|k}\right| > \frac{\epsilon}{3}\right) < 2^{RC} e^{\frac{-n_{++k}\epsilon^2}{18}}.$$

For the third term, we have,

$$\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{r+|k}p_{+c|k}-\pi_{r+|k}\pi_{+c|k}\right| \leq \sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{r+|k}p_{+c|k}-\pi_{r+|k}p_{+c|k}\right|+\sum_{r=1}^{R}\sum_{c=1}^{C}\left|\pi_{r+|k}p_{+c|k}-\pi_{r+|k}\pi_{+c|k}\right|.$$

Which is equivalent to,

$$\sum_{r=1}^{R}\left|p_{r+|k}-\pi_{r+|k}\right|+\sum_{c=1}^{C}\left|p_{+c|k}-\pi_{+c|k}\right|.$$

And again by Lemma 6.6 we have for any $\epsilon > 0$

$$P\left(\textstyle\sum_{r=1}^{R}\left|p_{r+|k}-\pi_{r+|k}\right| > \frac{\epsilon}{3}\right) < 2^{R}e^{\frac{-n_{++k}\epsilon^2}{18}},$$
$$P\left(\textstyle\sum_{c=1}^{C}\left|p_{+c|k}-\pi_{+c|k}\right| > \frac{\epsilon}{3}\right) < 2^{C}e^{\frac{-n_{++k}\epsilon^2}{18}}.$$

Summarizing the results above, we have

$$P\left(\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k}-p_{r+|k}p_{+c|k}\right| > \epsilon\right) < \left(2^{RC}+2^{R}+2^{C}\right)e^{\frac{-n_{++k}\epsilon^2}{18}},$$

Now, by the Cauchy-Schwarz inequality, it also holds

$$\sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k}-p_{r+|k}p_{+c|k}\right)^2} \leq \sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k}-p_{r+|k}p_{+c|k}\right|$$

and we can complete the proof,

$$\begin{aligned}
P\left(\max_k\sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(p_{rc|k}-p_{r+|k}p_{+c|k}\right)^2} > \epsilon\right) &\leq P\left(\max_k\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k}-p_{r+|k}p_{+c|k}\right| > \epsilon\right)\\
&< \sum_{k=1}^{K}P\left(\sum_{r=1}^{R}\sum_{c=1}^{C}\left|p_{rc|k}-p_{r+|k}p_{+c|k}\right| > \epsilon\right)\\
&< \left(2^{RC}+2^{R}+2^{C}\right)\sum_{k=1}^{K}e^{\frac{-n_{++k}\epsilon^2}{18}}.
\end{aligned}$$

□

An immediate corollary is that the average-type statistic is also strongly consistent, as it is less than or equal to the maximum-type statistic. Even given strong consistency, the distributions of the average-type and maximum-type statistics are generally difficult to derive. So as in Zhang's approach in [57], it is preferable to use a permutation procedure to approximate the p-value. For instance, one can randomly permute the X values of all samples and compute the average-type test statistic $\sum_{k=1}^{K} p_k \mathcal{V}_n (X, Y|Z_k = 1)$ for each permuted table. This yields a p-value based on the proportion of permuted test statistics smaller than the observed average-type test statistic. We will use this approach in the numerical studies below; but first introduce a third statistic meant to test the related but different hypothesis of homogeneity in a three-way table.

6.5 Homogeneity test via marginal distance covariance

The homogeneity of XY on Z is equivalent to the joint independence between (X, Y) and Z. Hence there are two possible formulations. First,

$$\begin{aligned} &H_0 : \pi_{rc1} = \pi_{rc2} = \ldots = \pi_{rcK} \quad \text{for all } 1 \le r \le R, 1 \le c \le C, \\ &\mathrm{H}_\alpha : \pi_{rck} = \pi_{rck'} \quad \text{for some } \ (r, c, k, k'). \end{aligned} \tag{6.6}$$

Or equivalently,

$$\begin{aligned} &\mathrm{H}_0 : \pi_{rck} = \pi_{rc+} * \pi_{++k} \quad \text{for all } \ 1 \le r \le R, 1 \le c \le C, 1 \le k \le K, \\ &H_\alpha : \pi_{rck} \ne \pi_{rc+} * \pi_{++k} \quad \text{for some } \ (r, c, k). \end{aligned}$$

The most popular homogeneity test for three-way tables is the Pearson χ^2 test, for which the test statistic is,

$$X^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{k=1}^{K} \frac{\left(n_{rck} - \overline{n}_{rc}\right)^2}{\overline{n}_{rc}},$$

where $\overline{n}_{rc} = \sum_{k=1}^{K} n_{nck}/K$. It is well-known that with large samples, X^2 tends to a χ^2 distribution with $(RC-1)(K-1)$ degrees of freedom. Once again, the difficulty (which will appear in our studies in Section 6.6), is that Pearson's test suffers from low statistical power for without the benefit of a large sample size.

To this end, we propose a new test of homogeneity of XY on Z that is based on marginal distance covariance. It will be convenient to use the notation

$$\mathcal{R}((X, Y), Z) \equiv \mathcal{R}(W, Z) \quad \text{where } W_{rc} = 1 \iff X_r = Y_c = 1.$$

Let $\{X, Y, W, Z\}$, $\{X^{'}, Y^{'}, W^{'}, Z^{'}\}$, and $\{X^{''}, Y^{''}, W^{''}, Z^{''}\}$ be three independent copies of $\{X, Y, W, Z\}$. Now we can exploit another alternative formulation of $\mathcal{V}^2(\cdot, \cdot)$, again due to Szekely et al. [42], that is stated in terms of inter-point distance $\|\cdot\| = |\cdot|_2$,

$$\begin{aligned}\mathcal{V}^2((X,Y),Z) &= \mathcal{R}(W,Z)\\ &= E(\|W-W'\|\,\|Z-Z'\|) + E(\|W-W'\|)E\left(\|Z-Z'\|\right) - 2E(\|W-W'\|\,\|Z-Z''\|).\end{aligned}$$

Now using the same geometric intuition as from Section 1.4, we see

$$E\left(\|Z-Z'\|\right) = \sqrt{2}P\left(Z \neq Z'\right) = \sqrt{2}\left(1 - \sum_{k=1}^{K} \pi_{++k}^2\right).$$

And similarly,

$$E\left(\|W-W'\|\right) = \sqrt{2}\left(1 - P\left(X = X', Y = Y'\right)\right) = \sqrt{2}\left(1 - \sum_{r=1}^{R}\sum_{c=1}^{C} \pi_{rc+}^2\right).$$

For the covariance terms, we can also compute,

$$\begin{aligned}E\left(\|W-W'\|\,\|Z-Z'\|\right) &= 2P\left(W \neq W', Z \neq Z'\right)\\ &= 2\left(P\left(Z \neq Z'\right) - P\left(X = X', Y = Y', Z \neq Z'\right)\right)\\ &= 2\left(1 - \sum_{k=1}^{K}\pi_{++k}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}\left(\pi_{rc+} - \pi_{rck}\right)\right).\end{aligned}$$

$$\begin{aligned}E\left(\|W-W'\|\,\|Z-Z''\|\right) &= 2P\left(W \neq W', Z \neq Z''\right)\\ &= 2\left(P\left(Z \neq Z''\right) - P\left(X = X', Y = Y', Z \neq Z''\right)\right)\\ &= 2\left(1 - \sum_{k=1}^{K}\pi_{++k}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}\pi_{rc+}\left(1 - \pi_{++k}\right)\right).\end{aligned}$$

Combining summations over the same indices,

$$\begin{aligned}\mathcal{V}^2((X,Y),Z) =& -2\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}\left(\pi_{rc+} - \pi_{rck}\right) - 2\left(1 - \sum_{k=1}^{K}\pi_{++k}^2\right)\sum_{r=1}^{R}\sum_{c=1}^{C}\pi_{rc+}^2\\ &+ 4\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}\pi_{rc+}\left(1 - \pi_{++k}\right)\\ =&\ 2\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}^2 - 4\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rck}\pi_{rc+}\pi_{++k} + 2\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi_{rc+}^2\pi_{++k}^2\\ =&\ 2\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\pi_{rck} - \pi_{rc+}\pi_{++k}\right)^2.\end{aligned}$$

Replacing the parameters with their maximum likelihoods, we have the following test statistic $\mathcal{V}_n^2((X,Y),Z)$ for H_0 in Equation 6.6,

$$\mathcal{V}_n^2((X,Y),Z) = 2\sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(p_{rck} - p_{rc+}p_{++k}\right)^2. \tag{6.7}$$

Statistical significance of the test may be evaluated by a permutation approach, ranking the observed statistic within a population of statistics computed from permuted data. The homogeneity test statistic is also strongly consistent.

Theorem 6.8. *Under multinomial model and the null hypothesis of homogeneity, we have*

$$P\left(\mathcal{V}_n^2((X,Y),Z) > \epsilon\right) < \left(2^{RCK} + 2^{RC} + 2^{K}\right) e^{\frac{-n\epsilon^4}{36}},$$

for any $\epsilon > 0$.

Proof. The proof proceeds in close analogy to the proof of Theorem 6.7. □

6.6 Simulation studies and improved power

We now compare the empirical power of our new tests of conditional independence to the CMH, CMI, and PDC tests discussed above. The comparison is based on simulated large sparse tables, generated in four settings as below. In all cases $K = 2$; and $X \perp Y|Z_1 = 1$ while $X \not\perp Y|Z_2 = 1$.

(Weak dependence, $R = C = 10$**)** With $\pi_{rc|1} = \frac{1}{100}$; and $\pi_{rc|2} = \frac{9}{270}$ for 10 randomly selected cells; and $\pi_{rc|2} = \frac{2}{270}$ for the remaining 90 cells.

(Strong dependence, $R = C = 10$**)** With $\pi_{rc|1} = \frac{1}{100}$; and $\pi_{rc|2} = \frac{18}{270}$ for 10 randomly selected cells; and $\pi_{rc|2} = \frac{1}{270}$ for the remaining 90 cells.

(Weak dependence, $R = C = 20$**)** With $\pi_{rc|1} = \frac{1}{400}$; and $\pi_{rc|2} = \frac{38}{3800}$ for 20 randomly selected cells; and $\pi_{rc|2} = \frac{8}{3800}$ for the remaining 380 cells.

(Strong dependence, $R = C = 20$**)** With $\pi_{rc|1} = \frac{1}{400}$; and $\pi_{rc|2} = \frac{114}{3800}$ for 20 randomly selected cells; and $\pi_{rc|2} = \frac{4}{3800}$ for the remaining 380 cells.

The results are below in Figure 6.1. In settings with only a weak dependence test, the CDC test is significantly better than all tests across all sample sizes. When the dependence relation is relatively strong, all tests achieve satisfactory power with large samples. However, the CDC test is much more powerful when data is sparse. For these results we used the maximum-type test statistic from Definition 6.5. The justification for this choice over the average-type statistic

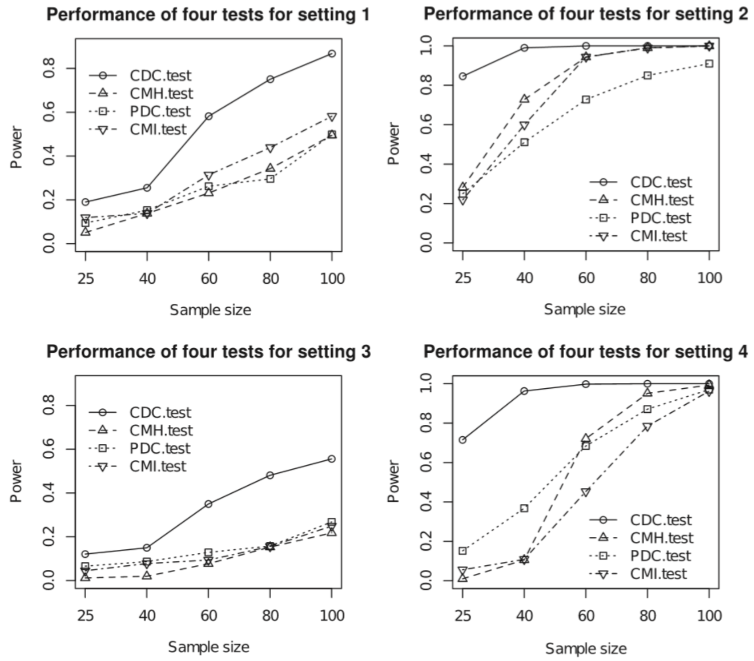


Figure 6.1: Empirical results of the four independence tests

from Definition 6.4 is in Figure 6.2; especially when dependence is weak, our results indicate the maximum-type statistic is strictly preferable.

Finally, we performed the same experiments for the homogeneity statistic in Equation 6.7, comparing its power to Pearson's test for homogeneity (whose statistic is asymptotic χ^2 with $df = RC-1$ since $K = 1$). Figure 6.3 shows that in all settings, the marginal distance covariance test greatly outperforms Pearson's test, especially for relatively small sample sizes, where the Pearson's test failed to detect any true positives.

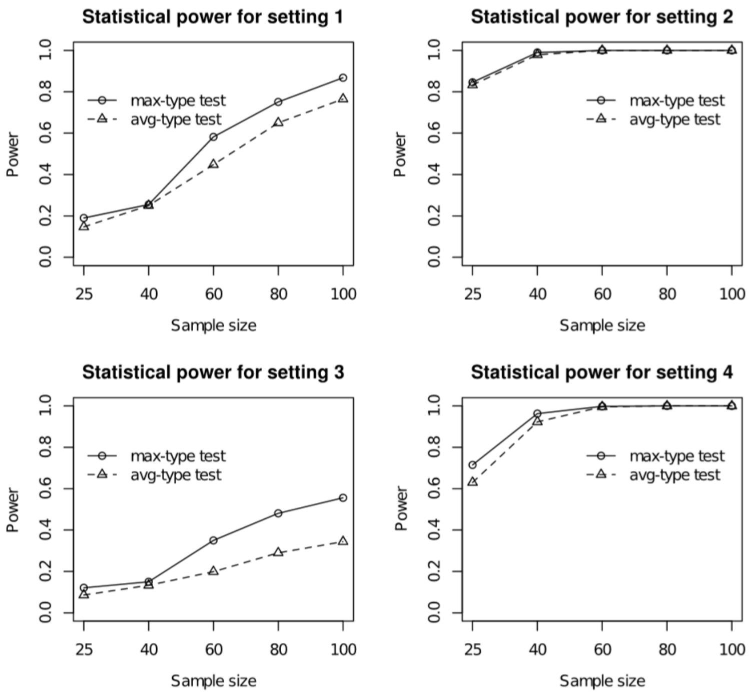


Figure 6.2: Empirical results of the test statistic types

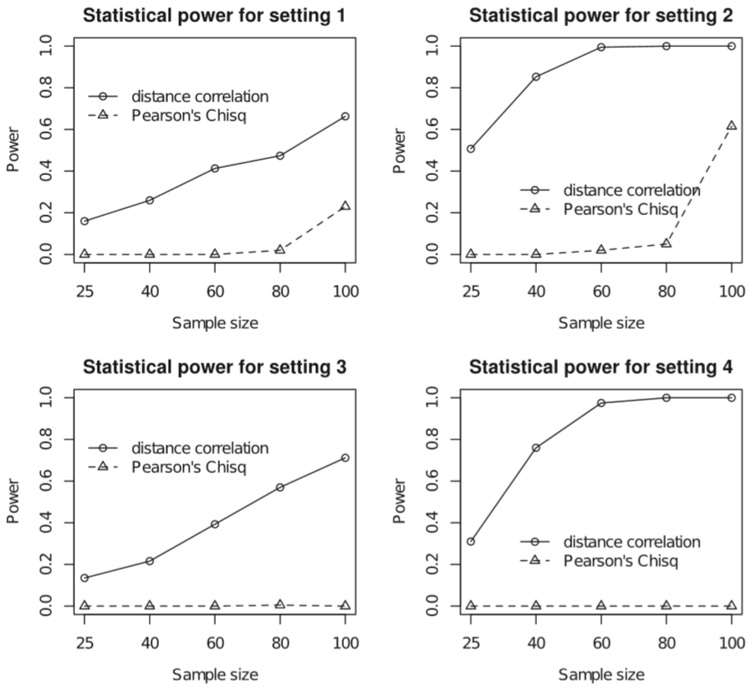


Figure 6.3: Empirical results of the two homogeneity tests

Part III

A Novel Graph-Based Multivariate Test

Chapter 7

Motivation and Definition of the Test

Biological processes are often highly heterogeneous. This can make it very complicated to understand their evolution. Unfortunately, this includes cancer. There are many genetic and epigenetic factors which, if they become abnormal, can cause cancer. Examples include gene expression level, DNA methylation level, somatic mutation, and copy number variation. It is possible to study these genes individually by performing univariate two-sample comparisons between normal and tumor groups; and over the past decades, many genes have been discovered to affect specific kinds of cancer. For instance, genes *BRCA1*, *BRCA2*, *PIK3C*, and *GATA3* for breast cancer (see [18]); and genes *MYC*, *RIT1*, *ECFR* and *ERBB2* for prostate cancer (see [17]). But this kind of analysis provides limited insight into the molecular mechanisms of tumorigenesis, as it ignores regulatory relations between genes (there is much literature establishing the importance of these relations; see for example [46], [23], [45], [41], [29], [36], [2], and [17]).

Thus we need robust multivariate tests when studying the genetic alterations involved in pathways theorized to affect tumorigenesis. In this part of the dissertation, we present joint work from [51] with Zhang, Mahdi, and Chen in which we propose a nonparametric and data-driven approach to the problem. It begins by refining and expanding the pathways to be studied; and then applies a novel graph-based multivariate test to analyze how genetic alterations in these pathways affect the progression of ovarian cancer. Drawing on the KEGG pathway database and a dataset from the Cancer Genome Atlas in [9], we find evidence that the cell cycle and ERBB pathways play key roles in early-stage transitions; and evidence that the ECM receptor and apoptosis pathways contribute to the progression from Stage III to Stage IV.

7.1 Typical challenges for parametric and non-parametric approaches

Researchers have in fact developed many computational models of genetic alterations caused by metabolic pathways. (Examples include EnrichNet [52], GAGE [26], PAGE [33], MEGO [56], GeneTrail [32], and Catmap [19]; to name a few.) But these models have tended to come from knowledge-based enrichment analyses and lack the flexibility to analyze user-defined pathways or gene sets. The work of Edelman et al. in [20] is an interesting exception, as it takes a more data-driven approach focused on pathway dependencies. This makes it flexible enough to handle user-defined pathways as well. The analysis is hierarchical, modeling the transitions from normal to primary tumor and from primary tumor to metastasis. However, this desirable property come at a high computational price. Edelman et al. require steps such as regularized multi-task learning, inverse regression, gradient learning, and leave-one-out cross-validation. This limits the usefulness of their approach with large-scale data sets. We would like a data-driven approach that retains the flexibility to analyze, for example, all 186 KEGG pathways from `http://www.genome.jp/kegg` relative to a large clinical data set, at a reasonable computational cost.

7.2 Statistical formulation of the problem

We first need a precise formulation of the problem, which is to detect differentially acted pathways between cancer stages. In statistical terms, we must test the equality of two or more joint distributions, where each random variable represents the expression level of one gene. So given a pathway to study, let $i \in \{1, 2, \ldots, p\}$ be the index for cancer stages, and $(X_1^{(i)}, \ldots, X_d^{(i)})$ be the expression levels of d genes in the pathway with a joint distribution $\boldsymbol{F}^{(i)}$.

In this setting, given n_i i.i.d. observations in stage i—that is, $(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)})$ where $\mathbf{x}_k^{(i)} = (x_{k1}^{(i)}, x_{k2}^{(i)}, \ldots, x_{kd}^{(i)})$—we wish to test the hypothesis,

$$\begin{aligned} &H_0 : \boldsymbol{F}^{(1)} = \boldsymbol{F}^{(2)} = \ldots = \boldsymbol{F}^{(p)} \\ &H_\alpha : \boldsymbol{F}^{(j)} \neq \boldsymbol{F}^{(j')} \quad \text{for some } j, j', 1 \leq j, j' \leq p. \end{aligned} \tag{7.1}$$

For a particular transition step, from stage i to stage $i+1$, we can conduct the following pairwise test (corresponding to the special case $p = 2$):

$$\begin{aligned} &H_0 : \boldsymbol{F}^{(i)} = \boldsymbol{F}^{(i+1)} \\ &H_\alpha : \boldsymbol{F}^{(i)} \neq \boldsymbol{F}^{(i+1)}. \end{aligned} \tag{7.2}$$

For background, let us recall that the two-sample multivariate tests in the statistics literature are broadly categorized as parametric or non-parametric.

The classic example of a parametric test is Hotelling's T^2 test, which compares the mean vectors of two multivariate Gaussian distributed populations, and generally works well in low-dimensional cases. High-dimensional tests for mean vectors have received special attention recently (see for example [55], [3], and [31]). The famous Kolmogorov-Smirnov test (see [28] for discussion of computational costs) is a nonparametric test for equality of arbitrary multivariate distributions. Another very interesting family of nonparametric tests are "edge-count" tests that draw from the intuition that if two groups have different distributions, samples will tend to be "closer" (in an appropriate sense) to samples in the same group than to those from the other group. If we then form a graph whose vertices represent the samples, and add edges only between "sufficiently close" vertices, then counting the number of inter-group edges can yield statistics for testing the hypothesis in Equation 7.2. We reject H_0 if the number of between-group edges is significantly less than expected. (See work of Friedman and Rafksy in [21]; and Rosenbaum in [43] for examples of existing work in this family.)

7.3 Derivation of a novel graph-based test

All such tests have notable limitations in practical applications. The Kolmogorov-Smirnov test is known to be overly conservative; it too often fails to reject H_0. Moreover, sophisticated implementations are needed to make the test usable in high-dimensional settings, as a brute force approach is very computationally intensive. Existing edge-count tests are easier to implement efficiently, but are problematic relative to choice of location and scale, as Chen and Friedman detail in [12]. Their work on a modified edge-count test in the same work overcomes these problems using a statistic with a computable asymptotic distribution. Not only does the test work properly under different location and scale alternatives, it exhibits substantial power gains over existing edge-count tests. The core ideas are still as above; for instance, the similarity graph can be a minimum spanning tree (MST) as described in [13], where edge assignments are made based on Euclidean distance.

In order to test the hypotheses in Equation 7.1, we will extend Chen and Friedmans test to our multi-sample setting, and derive an asymptotic distribution for easy p-value approximation. The first step is to pool the samples from all p groups (with n_i observations in group i), writing

$N = \sum_{i=1}^{p} n_i$. The next step is to construct a similarity graph G with a vertex for each pooled observation, letting R_i denote the number of edges in the graph that connect observations within sample i. Now consider the permutation null distribution which assigns probability $1/\begin{pmatrix} N \\ n_1, n_2, \ldots, n_p \end{pmatrix}$ to each of the $\begin{pmatrix} N \\ n_1, n_2, \ldots, n_p \end{pmatrix}$ choices of n_i out of the total N observations—where each observation is chosen once.

We can then denote by $\mathrm{P_P}$, $\mathrm{E_P}$, $\mathrm{Var_p}$, $\mathrm{Cov_p}$ probability expectation, variance, and covariance, respectively, under the permutation null distribution. Direct computation then yields,

$$\begin{aligned}
\mathrm{E_P}(R_i) \quad & = |G|\tfrac{n_i(n_i-1)}{N(N-1)} \triangleq \mu_i, \\
\mathrm{Var_P}(R_i) \quad & = \mu_i(1-\mu_i) \\
& \quad + 2C\tfrac{n_i(n_i-1)(n_i-2)}{N(N-1)(N-2)} \\
& \quad + (|G|(|G|-1)-2C)\tfrac{n_i(n_i-1)(n_i-2)(n_i-3)}{N(N-1)(N-2)(N-3)} \triangleq \sigma_i^2, \\
\mathrm{Cov_P}(R_i, R_j) \quad & = (|G|(|G|-1)-2C)\tfrac{n_i n_j (n_i-1)(n_j-1)}{N(N-1)(N-2)(N-3)} - \mu_i\mu_j \quad \text{for } i \neq j.
\end{aligned} \tag{7.3}$$

where $|G|$ is the number of edges in graph G and C is the constant $\frac{1}{2}\sum_{k=1}^{N} |G_k|^2 - |G|$ with G_k being the subgraph in G that includes all edge(s) that connect to node k. (That is, C is the number of edge pairs that share a common node in G).

We are now in a position to define our test statistic, as follows:

$$S := (R_1-\mu_1, R_2-\mu_2, \ldots, R_p-\mu_p)\,\Sigma^{-1}\begin{pmatrix} R_1-\mu_1 \\ R_2-\mu_2 \\ \cdots \\ R_p-\mu_p \end{pmatrix}. \tag{7.4}$$

where Σ is the covariance matrix of vector $(R_1, R_2, \ldots, R_p)^T$, whose explicit analytic expression follows immediately from $\mathrm{Var_p}$, $\mathrm{Cov_p}$ in Equation 7.3. Having derived the test statistic, we now consider some of key properties.

7.4 Properties of the test statistic

The fundamental property of the test statistic is its asymptotic convergence to a χ^2 distribution under certain conditions on the underlying graph. The proof, which is quite technical, is included in Appendix A for completeness.

Theorem 7.1. *For an edge $e \in G$, let*

$$A_e = \{e\} \cup \{e' \in G : e' \text{ and } e \text{ share a node } \}\,.$$

$$B_e = A_e \cup \{e'' \in G : \exists e' \in A_e, \text{ such that } e'' \text{ and } e' \text{ share a node}\}.$$

Now suppose that,

- $|G| = O(N)$*; and,*
- $\sum_{k=1}^{N} |G_k|^2 - 4|G|^2/N = O(N)$*; and,*
- $\sum_{e \in G} |A_e|\,|B_e| = o\left(N^{1.5}\right)$*; and,*
- $\lim_{N\to\infty} n_i/N = \lambda_i \in (0,1)$.

Then it follows that $S \to \chi^2_p$ under the permutation null.

7.5 Simulation studies

We now turn to two empirical studies of our statistic. First, we consider how accurately we can approximate the p-value of the test statistic using a permutation approach. Second, we compare the power of our test to three similar tests.

Accuracy of p-value approximation

The approximate p-value of the test statistic is obtained by a permutation approach in which the sample is permuted m times, a S' is computed for each permutation, and the p-value is taken to be the proportion of S' smaller than the observed S. To study the finite performance of this approach under moderate sample sizes, we compared the approximate p-value from a known χ^2 distribution with the permutation p-value where $m = 10000$. This was done using varying dimensions ($d = 10$ and $d = 100$) and sample sizes ($N = 80$, $n_1 = n_2 = n_3 = n_4 = 20$ and $N = 140$, $n_1 = n_2 = 20, n_3 = n_4 = 50$ and $N = 200$, $n_1 = n_2 = n_3 = n_4 = 50$). The data were generated from a p-dimension Gaussian distribution with zero mean vector and identity covariance matrix. The similarity graph was chosen to be a k-MST ($k = 1, 3, 5$) over the pooled observations. (Recall that the k-MST is defined as the union of the first k disjoint minimum spanning trees; it can be obtained cheaply using Chazelle's algorithm from [11].)

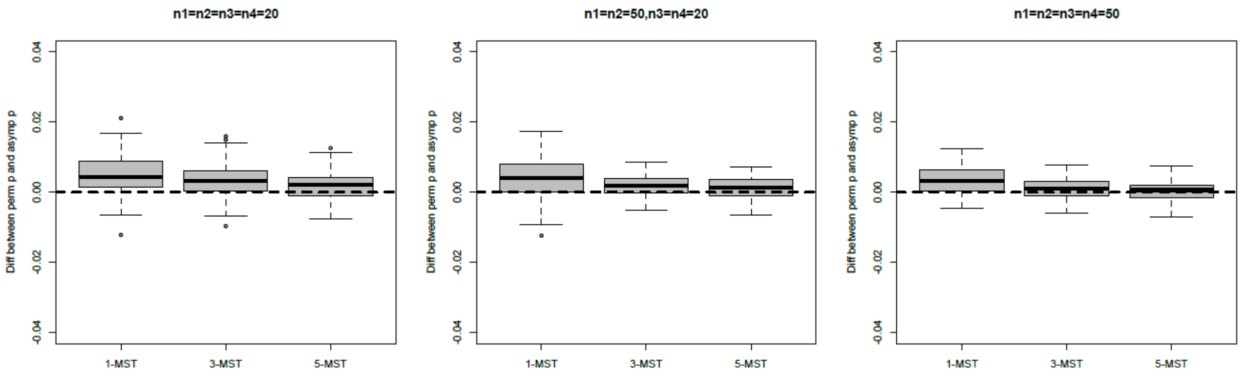


Figure 7.1: Boxplots for approximation accuracy with dimension $d = 10$

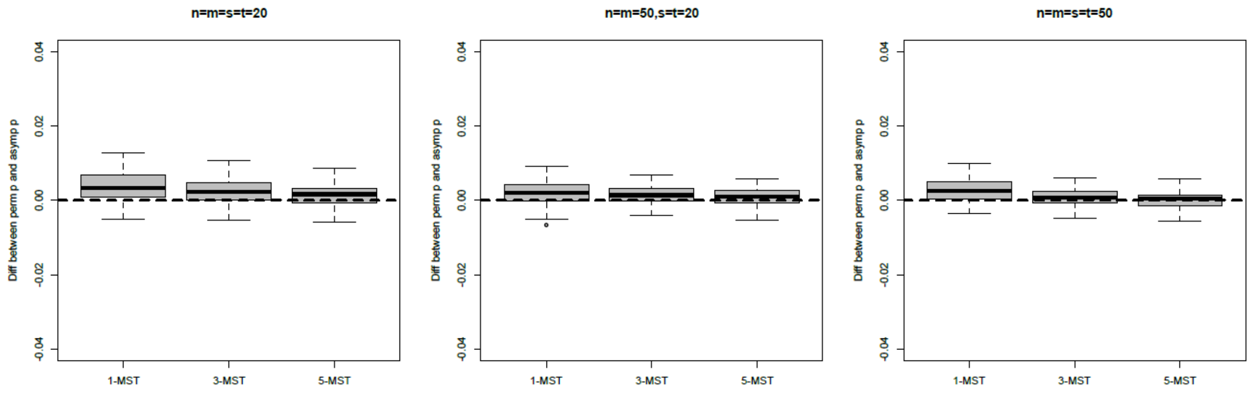


Figure 7.2: Boxplots for approximation accuracy with dimension $d = 100$

Since we know the true asymptotic distribution is, for example, $\chi^2_{df=4}$, the approximate p-value is simply $\Pr\left(\chi^2_{df=4} > S\right)$. Figure 7.1 and Figure 7.2 above summarize the accuracy of the approximate p-values by plotting the approximate p-value minus the permutation p-value under the various dimensions, sample sizes, and similarity graphs mentioned above. It can be seen that under all conditions, the approximate p-values tend to be slightly conservative and increasing dimension leads to a slightly decreasing accuracy. We also note that using a denser similarity graph such as a 3-MST or 5-MST can slightly improve the p-value approximation. As for sample size, $\min_i n_i > 20$ appears to suffice in practice in order to reasonably approximate the p-value from the χ^2 distribution.

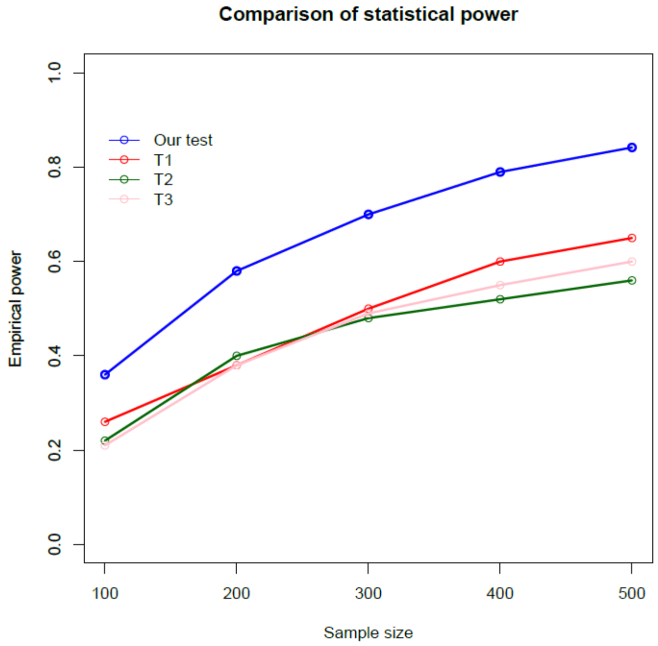


Figure 7.3: Empirical power of various tests

Power comparison with other tests

This second study compares the empirical statistical power of our proposed multi-sample edge-count test with three similar tests examined by Chen and Friedman in [12]. The test statistics are as follows:

$$T_1 = \sum_{i=1}^{4} \left| R_i - \mu_i \right|,$$

$$T_2 = \sum_{i=1}^{4} \left| R_i - \mu_i \right| / \sqrt{\Sigma_{ii}},$$

$$T_3 = \sum_{i=1}^{4} \left(R_i - \mu_i \right)^2,$$

where R_i and μ_i are as before; that is, the number of edges connecting samples within group i and its expectation, respectively. Using $p = 4$ and dimension $d = 100$, we generated the study data from a multivariate Gaussian distribution for each group, where the mean vector of group i is $\mathbf{c}_i$ and the covariance matrix is the identity. The choices of mean vector were $\boldsymbol{c}_1 = 0$, $\boldsymbol{c}_2 = 0.1$, $\boldsymbol{c}_3 = 0.2$, and $\mathbf{c}_4 = -0.3$; the sample size varied from $N = 100$ to $N = 500$. Figure 7.3 shows the empirical statistical power; our proposed test outperforms the three reference test statistics.

Chapter 8

Application to KEGG Pathways in Ovarian Cancer

8.1 The Cancer Genome Atlas dataset

In [18], Hsu et al. published a rich data set as part of the Cancer Genome Atlas (TCGA) on 565 subjects diagnosed with ovarian cancer. Beyond clinical record data such as age, race, outcome of debulking surgery, and chemotherapy, the data set has information on 17,813 genes; the gene profiles include gene expression, exon expression, genotype, and copy number variation (CNV). The subjects fell into four natural groups based on the clinical classification of their cancer (Stage I, Stage II, Stage III, and Stage IV). In this chapter, we present an application of the multivariate test from Chapter 7 to an enrichment of this data set. The goal is to discover any metabolic pathways from the KEGG database (`http://www.genome.jp/kegg`) whose gene profiles change in a significant way among the groups of subjects with different stages of cancer. Our results recover the known importance of the ERBB, cell cycle, prostrate cancer, TGF β signaling, pancreatic cancer, and p53 signaling pathways in the transition from Stage I to Stage II cancer. They also suggest the ECM receptor and apoptosis pathways play a significant role in the transition from Stage III to Stage IV.

8.2 Preprocessing of TCGA attributes

The CGA data set in [18] does require some preprocessing and refinement before we can apply our test. From the profiling data, we decided to focus on gene expression level, CNV, and DNA methylation from the genetic profiling data, which were downloaded from the Genomic Data Commons portal in January 2017. Out of all genes in the data set, 12,831 had methylation level measured for each CpG island located in their promoter regions. For genes containing more than one CpG island, we took the average methylation level, as is standard practice. The copy numbers were measured on each chromosome segment by circular binary segmentation. (If a gene spans two chromosomal segments, its overall copy number is the average of the numbers that would be assigned for each segment.) The expression level of each gene was quantified by the count of reads mapped to the gene, where these quantifications were performed using the

HTSeq software package, version 0.9.1 (see [39]); and the count data were log-transformed for further processing.

We normalized each data type for each gene by subtracting the median and dividing by the standard deviation to avoid any data type dominating. In addition, we removed the effects due to different age groups and batches using a median-matching and variance-matching strategy (see [9]). For example, the batch effect can be removed in the following way:

$$g^*_{ijk} = M_i + (g_{ijk} - M_{ij}) \frac{\hat{\sigma}_{g_i}}{\hat{\sigma}_{g_{ij}}}, \tag{8.1}$$

where g_{ijk} refers to the expression value for gene i from sample k in batch j, M_{ij} represents the median of $g_{ij} = (g_{ij1}, \ldots, g_{ijn})$, M_i refers to the median of $g_i = (g_{i1}, \ldots, g_{iJ})$; and $\hat{\sigma}_{g_i}$, $\hat{\sigma}_{g_{ij}}$ are the standard deviations of g_i and g_{ij}, respectively.

8.3 Pathway expansion and refinement

Now we describe how to integrate these gene profile data with the 186 metabolic pathways in the KEGG data set. Consider a pathway, say the ERBB pathway, which contains 86 genes. For each gene, define three variables corresponding to each data type above; that is, expression level, CNV, and methylation level. (Since it is well known that the expression level of a gene can be greatly affected by genetic or epigenetic changes such as copy number variation and DNA methylation, including these factors together may provide insight about the upstream cause of abnormal expression.) Now, to better adapt the KEGG pathways, we further refine the derived variables by removing genes that are irrelevant to phenotypic changes. Following the example of Edelman et al. in [20], we use an F-test to calculate the p-value of each single gene's derived variables. We exclude the variable if its p-value exceeds a predefined threshold of 0.1. This refinement, for example, reduces the ERBB pathway expression level variables from 87 genes to 29 genes; it also excludes 165 out of the 174 methylation and CNV variables. Thus in our refined data set, an observation of the KEGG pathway is a vector $x_i \in \mathbb{R}^{38}$.

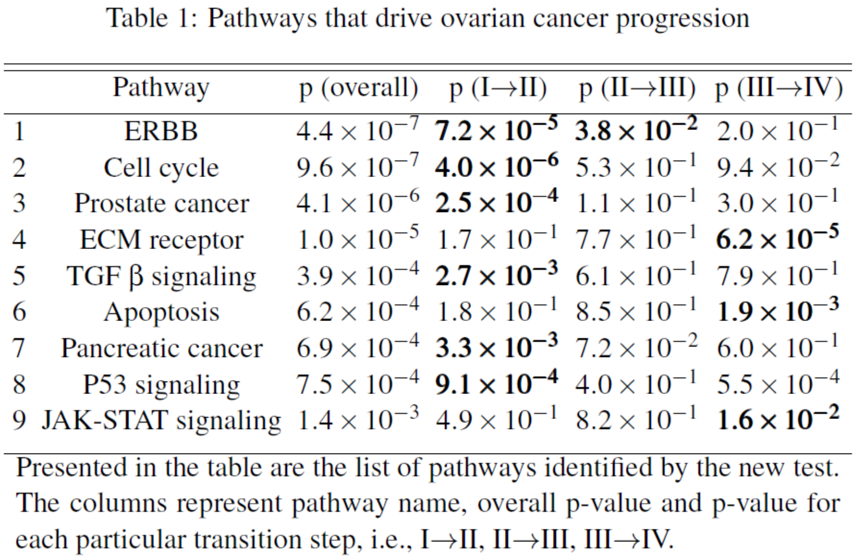


Table 1: Pathways that drive ovarian cancer progression

	Pathway	p (overall)	p (I→II)	p (II→III)	p (III→IV)
1	ERBB	4.4×10^{-7}	$\mathbf{7.2 \times 10^{-5}}$	$\mathbf{3.8 \times 10^{-2}}$	2.0×10^{-1}
2	Cell cycle	9.6×10^{-7}	$\mathbf{4.0 \times 10^{-6}}$	5.3×10^{-1}	9.4×10^{-2}
3	Prostate cancer	4.1×10^{-6}	$\mathbf{2.5 \times 10^{-4}}$	1.1×10^{-1}	3.0×10^{-1}
4	ECM receptor	1.0×10^{-5}	1.7×10^{-1}	7.7×10^{-1}	$\mathbf{6.2 \times 10^{-5}}$
5	TGF β signaling	3.9×10^{-4}	$\mathbf{2.7 \times 10^{-3}}$	6.1×10^{-1}	7.9×10^{-1}
6	Apoptosis	6.2×10^{-4}	1.8×10^{-1}	8.5×10^{-1}	$\mathbf{1.9 \times 10^{-3}}$
7	Pancreatic cancer	6.9×10^{-4}	$\mathbf{3.3 \times 10^{-3}}$	7.2×10^{-2}	6.0×10^{-1}
8	P53 signaling	7.5×10^{-4}	$\mathbf{9.1 \times 10^{-4}}$	4.0×10^{-1}	5.5×10^{-4}
9	JAK-STAT signaling	1.4×10^{-3}	4.9×10^{-1}	8.2×10^{-1}	$\mathbf{1.6 \times 10^{-2}}$

Presented in the table are the list of pathways identified by the new test. The columns represent pathway name, overall p-value and p-value for each particular transition step, i.e., I→II, II→III, III→IV.

Figure 8.1: Pathway p-values for overall and individual transitions

8.4 Analysis of cancer-driving pathways

With these preprocessing steps completed, we are in a position to test H_0 from Equation 7.1 for each of the refined 186 KEGG pathways. By a Benjamin-Hochberg procedure with level 0.05 using our graph-based test from Chapter 7, we identified a set of nine pathways. These included some well-studied cancer-related pathways such as the cell cycle, ERBB, p53 signaling, and JAK-STAT signaling pathways. For each identified pathway, we continued with a two-sample test in order to investigate the role of the pathways in each particular transition step. Figure 8.1 summarizes the results.

It is quite interesting to note that most of the pathways contributed only to a particular step; the ERBB pathway is a notable exception, with significant p-values in both the transition from Stage I to Stage II, as well as from Stage II to Stage III. Overall, five pathways were found to contribute only to the first stage transition; and three pathways to contribute only to the final transition. These results echo some prior knowledge; for instance, the ERBB pathway contains important proto-oncogenes and tumor suppressors such as PIK3C, KRAS and STAT5. It is also known that the ERBB pathway can be involved in the excessive signaling of growth factor receptors ERBB1 and ERBB2, which are critical factors in the malignancy of solid tumors. For example, several studies have shown a significant role for ERBB in the early progression of ovarian cancer and breast cancer. The cell cycle pathway also contains many genes that co-regulate cell proliferation, including ATM, RB1, CCNE1 and MYC. When this regulation becomes abnormal, it can cause cells to over-proliferate and tumors to accumulate—see again [18]. The refined gene sets of ERBB pathway and cell cycle pathway clearly differentiate the Stage I and Stage II groups, indicating their substantial involvement in this transition. One novel finding from our analysis is the critical role of the extracellular matrix (ECM) receptor pathway in the late-stage transition from Stage III to Stage IV. The ECM is a major component of the local microenvironment in a cancer cell, and plays an important roles in cancer development (see [53]).

Appendix A

Proof of Theorem 7.1

We need a technical device of Chen and Friedmans (see [47] for details) to establish the asymptotic distribution of S in Theorem 7.1.

Theorem A.1 (Chen and Shao). *Let $\mathcal{J}$ be an index set and let ξ_i denote a random variable with $\mathrm{E}\xi_i = 0$ and $\mathrm{E}\left(\xi_i^2\right) = 1$. Consider sums of the form $V = \sum_{i\in\mathcal{J}} \xi_i$.*

Now assume a restriction on dependence among the $\{\xi_i \colon i \in \mathcal{J}\}$. In particular, assume that, for each $i \in \mathcal{J}$, there exists $K_i \subset L_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{K_i^c}$ and ξ_{K_i} is independent of $\xi_{L_i^c}$. Under this assumption it holds that,

$$\sup_{h\in\mathrm{Lip}(1)} |Eh(V) - Eh(Z)| \leq \delta,$$

where $\mathrm{Lip}(1) = \{h : \mathbb{R} \to \mathbb{R}; \|h'\| \leq 1\}$, $Z \sim \mathcal{N}(0,1)$, and

$$\delta = 2\sum_{i\in\mathcal{J}} \left(E\left|\xi_i\eta_i\theta_i\right| + \left|E\left(\xi_i\eta_i\right)\right| E\left|\theta_i\right|\right) + \sum_{i\in\mathcal{J}} E\left|\xi_i\eta_i^2\right|,$$

with $\eta_i = \sum_{j\in K_i} \xi_j$ and $\theta_i = \sum_{j\in L_i} \xi_j$, for some K_i and L_i whose existence is guaranteed by the above assumption.

The proof for Theorem 7.1 is then as below.

Proof. We begin by studying our statistic under the bootstrap null distribution, which is defined as follows. For each observation, assign it to be from sample i with probability n_i/N, independent of other observations. Let V_i be the number of observations assigned to sample i. Then, conditioning on $\{V_i = n_i\}_{i=1,\ldots,p}$, the bootstrap null distribution is in fact the permutation null distribution. We use $\mathrm{P_B}, \mathrm{E_B}$, $\mathrm{Var}\ _{\mathrm{B}}$ to denote the probability, expectation, and variance under the bootstrap null distribution, respectively. Again direct computation shows,

$$\mathrm{E_B}\left(R_i\right) = \frac{n_i^2}{N^2}|G| \triangleq \mu_i^{\mathrm{B}},$$

$$\mathrm{Var_B}\left(R_i\right) = \frac{n_i^2\left(N-n_i\right)^2}{N^4}|G| + \frac{n_i^3\left(N-n_i\right)}{N^4}\sum_{k=1}^{N}\left|G_k\right|^2 \triangleq \left(\sigma_i^{\mathrm{B}}\right)^2.$$

So let,

$$W_i^{\mathrm{B}} = \frac{R_i - \mu_i^{\mathrm{B}}}{\sigma_i^{\mathrm{B}}}, \quad W_i = \frac{R_i - \mu_i}{\sigma_i},$$
$$U_i = \frac{V_i - n_i}{\sqrt{N\lambda_{i,N}\left(1-\lambda_{i,N}\right)}},$$

where $\lambda_{i,N} = n_i/N$. We can now prove the following as $N \to \infty$,

1. Under the bootstrap null, $\left(W_1^{\mathrm{B}}, W_2^{\mathrm{B}}, \ldots, W_p^{\mathrm{B}}, U_1, \ldots, U_{p-1}\right)$ goes to a multivariate normal distribution with positive definite covariance matrix for $(U_1, \ldots, U_{p-1})$.

2. $\frac{\sigma_i^{\mathrm{B}}}{\sigma_i} \to c_i$, $\frac{\mu_i^{\mathrm{B}} - \mu_i}{\sigma_i^{\mathrm{B}}} \to 0$ for constants c_i.

3. $\mathrm{rank}(\Sigma) = p$.

Taken together, these will imply the conclusion of the theorem. First, (1) implies the conditional distribution of $\left(W_1^{\mathrm{B}}, W_2^{\mathrm{B}}, \ldots, W_p^{\mathrm{B}}\right)'$ given $(U_1, \ldots, U_{p-1})$ becomes a multivariate Gaussian distribution under the bootstrap null distribution as $N \to \infty$. Since the permutation null distribution is equivalent to the bootstrap null distribution given $U_i = 0 \quad \forall i$, in particular $\left(W_1^{\mathrm{B}}, W_2^{\mathrm{B}}, \ldots, W_p^{\mathrm{B}}\right)'$ becomes a multivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Furthermore, since

$$W_i = \frac{\sigma_i^{\mathrm{B}}}{\sigma_i}\left(W_i^{\mathrm{B}} + \frac{\mu_i^{\mathrm{B}} - \mu_i}{\sigma_i^{\mathrm{B}}}\right),$$

given (2), it follows that $(W_1, W_2, \ldots, W_p)'$ *also* becomes a multivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Together with (3) we have the conclusion in the theorem.

Let us now prove facts (1), (2), and (3). For the first part of (1), by the Cramér-Wold device, we only need to show that $W = \sum_{i=1}^p a_i W_i^{\mathrm{B}} + \sum_{i=1}^{p-1} b_i U_i$ is asymptotically Gaussian distribution for any combination of a_i's and b_i's such that $\mathrm{Var}_{\mathrm{B}}(W) > 0$. To show this, we can use Stein's method. For $e \in G$, let

$$\xi_e = \sum_{i=1}^{p} a_i \frac{I_{J_e = i} - \lambda_{i,N}}{\sigma_i^{\mathrm{B}}},$$

where $\{J_e = i\}$ means the edge connects two observations from sample i.

For $k \in \{1, \ldots, N\}$, let

$$\xi_k = \sum_{i=1}^{p-1} b_i \frac{I_{g_k=i} - \lambda_{i,N}}{\sqrt{N\lambda_{i,N}\left(1-\lambda_{i,N}\right)}},$$

where $\{g_k = i\}$ means node k is from sample i. Then $W = \sum_{e \in G} \xi_e + \sum_{k=1}^{N} \xi_k$. We have $\mathrm{E}_{\mathrm{B}}\left(\xi_e\right) = 0, \forall e \in G$ and $\mathrm{E}_{\mathrm{B}}\left(\xi_k\right) = 0, \forall k \in \{1, \ldots, N\}$. Now we continue,

$$a = \max\left(\max_{i \in \{1,\ldots,p\}} \left|a_i\right|, \max_{i \in \{1,\ldots,p-1\}} \left|b_i\right|\right),$$

and

$$\sigma = \min\left(\min_{i \in \{1,\ldots,p\}} \sigma_i^B, \min_{i \in \{1,\ldots,p-1\}} \sqrt{N\lambda_{i,N}\left(1-\lambda_{i,N}\right)}\right).$$

Then $\left|\xi_e\right|, \left|\xi_k\right| \leq pa/\sigma$ for all $e \in G$ and $i \in \{1, \ldots, N\}$. Since also $|G| = O(N)$, we have $\sigma = O(\sqrt{N})$. For $e = (e_-, e_+) \in G$, let

$$\begin{aligned} K_e &= A_e \cup \left\{e_-, e_+\right\}, \\ L_e &= B_e \cup \left\{\text{nodes in} A_e\right\}. \end{aligned}$$

Note K_e and L_e satisfy the assumption in Theorem A.1. Then for $k \in \{1, \ldots, N\}$, let

$$\begin{aligned} K_k &= \{e \in G_k\} \cup \{i\}, \\ L_k &= \{e \in G_{k,2}\} \cup \{\text{ nodes in } G_i\}. \end{aligned}$$

Here K_k and L_k also meet the dependence assumption in Theorem A.1. Now let $\mathcal{J} = \{e : e \in G\} \cup \{1, \ldots, N\}$. For $j \in \mathcal{J}$, let $\eta_j = \sum_{k \in K_j} \xi_k$ and $\theta_j = \sum_{k \in L_j} \xi_k$. By Theorem A.1, we then have $\sup_{h \in \mathrm{Lip}(1)} |\mathrm{E}h(W) - \mathrm{E}h(Z)| \leq \delta$ for $Z \sim \mathcal{N}(0, 1)$ where

$$\begin{aligned} \delta &= \tfrac{1}{\sqrt{\mathrm{Var}_B(W)}} \left(2 \textstyle\sum_{j \in \mathcal{J}} \left(\mathrm{E}_B \left|\xi_j \eta_j \theta_j\right| + \left|E_B\left(\xi_j \eta_j\right)\right| E_B \left|\theta_j\right|\right) + \textstyle\sum_{j \in \mathcal{J}} E_B \left|\xi_j \eta_j^2\right|\right) \\ &\leq \tfrac{1}{\sqrt{\mathrm{Var}_B(W)}} \left(5 \textstyle\sum_{e \in G} \tfrac{p^3 a^3}{\sigma^3} \left(\left|A_e\right| + 2\right)\left(\left|B_e\right| + \left|A_e\right| + 1\right) + 5 \textstyle\sum_{k=1}^{N} \left(\left|G_i\right| + 1\right)\left(\left|G_{i,2}\right| + 1\right)\right) \\ &\leq \tfrac{1}{\sqrt{\mathrm{Var}_B(W)}} \tfrac{90 p^3 a^3}{\sigma^3} \textstyle\sum_{e \in G} \left|A_e\right| \left|B_e\right|. \end{aligned}$$

Since $\sigma = O(\sqrt{N})$, when $\left|A_e\right| \left|B_e\right| = o\left(N^{1.5}\right)$, we have $\delta \to 0$ as $N \to \infty$.

Finally we must check the covariance matrix of $(U_1, \ldots, U_{p-1})$. The diagonal elements are all unity and the off-diagonal element (i, j) with $i \neq j$ is,

$$-\sqrt{\frac{\lambda_{i,N}\lambda_{j,N}}{\left(1-\lambda_{i,N}\right)\left(1-\lambda_{j,N}\right)}}.$$

So the covariance matrix can be written as $D - vv^T$ where D is a diagonal matrix with the i-th diagonal element $1+\lambda_{i,N}/\left(1-\lambda_{i,N}\right)$, and $v = (\sqrt{\lambda_{1,N}/\left(1-\lambda_{1,N}\right)}, \ldots, \sqrt{\lambda_{p-1,N}/\left(1-\lambda_{p-1,N}\right)})^T$. Since $1 - v^T D^{-1} v = 1 - \sum_{i=1}^{p-1} \lambda_{i,N} = \lambda_{p,N} \neq 0$, $D - vv^T$ is invertible. Now the covariance matrix is inherently non-negative definite; and when it is full rank, it is positive definite.

We next prove the second fact from above. Notice that σ_i^2 can be re-written as

$$\sigma_i^2 = \frac{n_i\left(n_i-1\right)\left(N-n_i\right)\left(N-n_i-1\right)}{N(N-1)(N-2)(N-3)}\left(|G| + \frac{n_i-2}{N-n_i-1}\left(\sum_{k=1}^{N}|G_i|^2 - \frac{4|G|^2}{N}\right) - \frac{2}{N(N-1)}|G|^2\right),$$

since both $|G| = O(N)$ and $\sum_{k=1}^{N}|G_k|^2 - 4|G|^2/N = O(N)$. Now write $a_0 = \lim_{N\to\infty}|G|/N$ and $b_0 = \lim_{N\to\infty}\left(\sum_{k=1}^{N}|G_k|^2 - 4|G|^2/N\right)/N$. It follows,

$$\lim_{N\to\infty}\sigma_i^2/N = \lambda_i^2\left(1-\lambda_i\right)^2\left(a_0 + b_0\lambda_i/\left(1-\lambda_i\right)\right),$$

$$\lim_{N\to\infty}\left(\sigma_i^B\right)^2/N = \lambda_i^2\left(1-\lambda_i\right)^2 a_0 + \lambda_i^3\left(1-\lambda_i\right)\left(b_0 + 4a_0^2\right).$$

Hence,

$$\lim_{N\to\infty}\frac{\sigma_i^{\mathrm{B}}}{\sigma_i} = \sqrt{1 + \frac{4a_0^2\lambda_i}{a_0\left(1-\lambda_i\right) + b_0\lambda_i}}.$$

And since $\mu_i^{\mathrm{B}} - \mu_i = |G|\frac{n_i(N-n_i)}{N^2(N-1)}$, we conclude

$$\lim_{N\to\infty}\frac{\mu_i^{\mathrm{B}} - \mu_i}{\sigma_i^{\mathrm{B}}} = \lim_{N\to\infty}\frac{a_0\lambda_i\left(1-\lambda_i\right)}{\sigma_i^{\mathrm{B}}} = 0.$$

It remains to prove the third fact from above. The diagonal elements of Σ are σ_i^2 's. The off-diagonal elements are, for $i \neq j$,

$$\Sigma[i,j] = \frac{n_i n_j\left(n_i-1\right)\left(n_j-1\right)}{N(N-1)(N-2)(N-3)}\left(|G| - \left(\sum_{k=1}^{N}|G_k|^2 - \frac{4|G|^2}{N}\right) - \frac{2}{N(N-1)}|G|^2\right).$$

Then the leading terms of Σ can be decomposed as $\tilde{D} + \left(a_0 - b_0\right)uu^T$, where $\tilde{D}$ is a diagonal matrix with

$$\tilde{D}[i,i] = \lambda_i^2\left(\left(1-2\lambda_i\right)a_0 + \lambda_i b_0\right) \text{ and } u = \left(\lambda_1^2, \ldots, \lambda_p^2\right)^T.$$

Since $1 + \left(a_0 - b_0\right)u^T\tilde{D}^{-1}u = 1 + \sum_{i=1}^{p}\frac{\lambda_i^2(a_0-b_0)}{(1-2\lambda_i)a_0+\lambda_i b_0}$ is strictly larger than 0 (this quantity is strictly increasing in a_0 and equals 0 when $a_0 = 0$), Σ is of full rank. □

Bibliography

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley-Interscience, Hoboken, NJ, second edition, 2007.

[2] D. Talantov, A. Mazumder, J.X. Yu, T. Briggs, Y. Jiang, J. Backus, D. Atkins, and Y. Wang. Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clinical Cancer Research*, 11(20):7234–7242, 2005.

[3] K.B. Gregory, R.J. Carroll, V. Baladandayuthapani, and S.N Lahiri. A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association*, 110(510):837–849, 2015.

[4] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning Bayesian networks from data: An information-theory-based approach. *Artificial Intelligence - AI*, 137:43–90, 05 2002.

[5] G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1 -test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

[6] C. Borgelt. INeS - Induction of Network Structures.

[7] C. Borgelt. A conditional independence algorithm for learning undirected graphical models. *J. Comput. Syst. Sci.*, 76:21–33, 02 2010.

[8] C. Borgelt and R. Kruse. *Graphical Models - Methods for Data Analysis and Mining*. Wiley, 2002.

[9] Q. Zhang, J. Burdette, and J. Wang. Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*, 8:1338, 12 2014.

[10] I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, pages 246–256, 1989.

[11] B. Chazelle. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *Journal of the ACM (JACM)*, 47(6):1028–1047, 2000.

[12] Hao Chen and Jerome H. Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409, 2017.

[13] D. Cheriton and R.E. Tarjan. Finding minimum spanning trees. *SIAM Journal on Computing*, 5(4):724–742, 1976.

[14] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462 – 467, 06 1968.

[15] L. de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7:2149–2187, 10 2006.

[16] L. de Campos, J. Huete, and S. Moral. Independence in uncertainty theories and its applications to learning belief networks. In D. Gabbay and R. Kruse, editors, *DRUMS Handbook on Abduction and Learning*. Kluwer, Dordrecht, Netherlands, 2000.

[17] X. Cao, S.M. Dhanasekaran, and D.R. Rhodes. Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, 39(1):41–51, 2007.

[18] H. Hsu, E. Serpedin, T. Hsiao, A.J.R. Bishop, E.R. Dougherty, and Y. Chen. Reducing confounding and suppression effects in TCGA data: An integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics*, 13(6), 2012.

[19] T. Breslin, P. Edn, and M. Krogh. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, 5(1):193–193, 2004.

[20] E.J. Edelman, J. Guinney, J. Chi, P.G. Febbo, and S. Mukherjee. Modeling cancer progression via pathway dependencies. *PLoS computational biology*, 4(2), 2008.

[21] J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

[22] M. Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17(4):333–353, 1990.

[23] A.V. Ivshina, J. George, and O. Senko. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, 66(21):10292–10301, 2006.

[24] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* 01 2001.

[25] V. Graziano and M. Nakai. A geometrical framework for covariance matrices of continuous and categorical variables. *Sociological Methods & Research*, 44(1):48–79, 2015.

[26] W. Luo, M.S. Friedman, K. Shedden, K.D. Hankenson, and P.J. Woolf. Gage: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161–161, 2009.

[27] X. Zhang, X. Zhao, K. He, and L. Chen. Inferring gene regulatory networks from gene expression data by pc-algorithm based on conditional mutual information. *Bioinformatics (Oxford, England)*, 28:98–104, 11 2011.

[28] R.H.C. Lopes, P.R. Hobson, and D. Reid. Computationally efficient algorithms for the two-dimensional Kolmogorov–Smirnov test. *Journal of Physics: Conference Series*, 119(4):042019, jul 2008.

[29] A.P. Smith, K. Hoek, and D. Becker. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. *Cancer Biology and Therapy*, 4(9):1018, 2005.

[30] A. Jensen and F. Jensen. MIDAS — An influence diagram for management of mildew in winter wheat, 2013.

[31] M.S. Srivastava, S. Katayama, and Y. Kano. A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358, 2013.

[32] C. Backes, A. Kelley, and J. Kuentzer. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Research*, 35:186–192, 2007.

[33] S. Kim and D.J. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144–144, 2005.

[34] K. Kristensen and I. Rasmussen. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33:197–217, 03 2002.

[35] K. Kristensen and I. Rasmussen. A preliminary model for the production of beer from Danish malting barley grown without pesticides, 2013.

[36] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S. Pomeroy, T.R. Golu, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[37] S. Lauritzen, A. Dawid, B. Larsen, and H. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–505, 08 1990.

[38] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

[39] S. Anders, P. Pyl, and W. Huber. HTSeq: A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 09 2014.

[40] L. K. Rasmussen. Blood group determination of Danish Jersey cattle in the F-blood group system. *Dina Research Report*, 1992.

[41] M. Pancione, A. Remo, and V. Colantuoni. Genetic and epigenetic events generate multiple pathways in colorectal cancer progression. *Pathology Research International*, 2012:509348, 07 2012.

[42] G. Szekely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[43] P.R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[44] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, 1997.

[45] X. Ma, R. Salunga, and J.T. Tuggle. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5974–5979, 2003.

[46] K.S. Hoek, N.C. Schlegel, P. Brafford, A. Sucker, S. Ugurel, R. Kumar, B.L. Weber, K.L. Nathanson, D.J. Phillips, M. Herlyn, D. Schadendorf, and R. Dummer. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Research*, 19(4):290–302, 2006.

[47] H.Y. Chen, Q. Shao, and L. Goldstein. *Normal Approximation by Stein's Method*. Springer, 2011.

[48] D. J. Spiegelhalter and R. G. Coewll. *Bayesian Statistics 4: Learning in probabilistic expert systems.* Clarendon Press, Oxford, England, 1992.

[49] J. Szkely and L. Maria. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

[50] X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110:0–0, 01 2015.

[51] Q. Zhang, G. Mahdi, J. Tinker, and H. Chen. A graph-based multi-sample test for identifying pathways associated with cancer progression. *Computational Biology and Chemistry*, 87:107285, 2020.

[52] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: Network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)*, 28(18):i451, 2012.

[53] P. Lu, V.M. Weaver, and Z. Werb. The extracellular matrix: A dynamic niche in cancer progression. *The Journal of Cell Biology*, 196(4):395–406, 2012.

[54] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester, 1990.

[55] T.T. Cai, W. Liu, Y. Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(2):349–372, 2014.

[56] K. Tu, H. Yu, and M. Zhu. MEGO: Gene functional module expression based on gene ontology. *BioTechniques*, 38(2):277–283, 2005.

[57] Q. Zhang. Independence test for large sparse contingency tables based on distance correlation. *Statistics and Probability Letters*, 148:17–22, 2019.

[58] Q. Zhang and J. Tinker. Testing conditional independence and homogeneity in large sparse three-way tables using conditional distance covariance. *Stat*, 8, 2019.