Mutual Information between Categorical and Gaussian data

David Atienza

June 7, 2021

1 Introduction

This documents shows how to calculate the (conditional) mutual information between categorical and Gaussian data. The mutual information is always calculated between two variables X and Y, and can be conditioned on a set of variables $\mathbf{Z}$. Any of this variables can be discrete (categorical) or continuous (Gaussian). When needed, the subscript D and C are used to specify that the variable is discrete or continuous, e.g. X_D is a discrete X variable, $\mathbf{Z}_C$ is a set of conditioning continuous variables. The instantations for a given variable is denoted using lowercase, e.g. x_d, y_c, etc. The number of categories of the variables X and Y are denoted llx and lly, respectively. The total number of categories for all the $\mathbf{Z}_D$ variables is denoted $\text{llz} = \prod_i \text{llz}_i$.

2 Mutual Information

2.1 Mutual Information between Two Discrete Variables

First we will calculate the mutual information between two discrete variables:

$$\begin{aligned}
I(X_D; Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C) &= H(X_D, \mathbf{Z}_D, \mathbf{Z}_C) + H(Y_D, \mathbf{Z}_D, \mathbf{Z}_C) - H(X_D, Y_D, \mathbf{Z}_D, \mathbf{Z}_C) - H(\mathbf{Z}_D, \mathbf{Z}_C) \\
&= H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) + H(X_D, \mathbf{Z}_D) + H(\mathbf{Z}_C \mid Y_D, \mathbf{Z}_D) + H(Y_D, \mathbf{Z}_D) \\
&\quad - H(\mathbf{Z}_C \mid X_D, Y_D, \mathbf{Z}_D) - H(X_D, Y_D, \mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D) - H(\mathbf{Z}_D)
\end{aligned} \tag{1}$$

Note that:

$$I(X_D; Y_D \mid \mathbf{Z}_D) = H(X_D, \mathbf{Z}_D) + H(Y_D, \mathbf{Z}_D) - H(X_D, Y_D, \mathbf{Z}_D) - H(\mathbf{Z}_D) \tag{2}$$

Thus:

$$\begin{aligned}
I(X_D; Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C) &= I(X_D; Y_D \mid \mathbf{Z}_D) \\
&\quad + H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) + H(\mathbf{Z}_C \mid Y_D, \mathbf{Z}_D) - H(\mathbf{Z}_C \mid X_D, Y_D, \mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D)
\end{aligned} \tag{3}$$

where $H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D)$ is the entropy of the Gaussian variables conditioned on $X_D, \mathbf{Z}_D$. This can be easily calculated:

$$
\begin{aligned}
H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) &= -\sum_{x_d \in X_D, \mathbf{z}_d \in \mathbf{Z}_D} \int f(\mathbf{Z}_C, x_d, \mathbf{z}_d) \log \left(\frac{f(\mathbf{Z}_C, x_d, \mathbf{z}_d)}{f(x_d, \mathbf{z}_d)} \right) d\mathbf{Z}_C \\
&= -\sum_{x_d \in X_D, \mathbf{z}_d \in \mathbf{Z}_D} p(x_d, \mathbf{z}_d) \int f(\mathbf{Z}_C \mid x_d, \mathbf{z}_d) \log \left(f(\mathbf{Z}_C \mid x_d, \mathbf{z}_d) \right) d\mathbf{Z}_C \\
&= \sum_{x_d \in X_D, \mathbf{z}_d \in \mathbf{Z}_D} p(x_d, \mathbf{z}_d) H(\mathbf{Z}_C \mid x_d, \mathbf{z}_d)
\end{aligned} \tag{4}
$$

The last integral in (4) is the entropy of the Gaussian distribution trained with the $x_d, \mathbf{z}_d$ configuration.

For a Gaussian distribution, this integral can be solved with a closed-form formula:

$$
H(\mathbf{Z}_C \mid x_d, \mathbf{z}_d) = \frac{k}{2} + \frac{k}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{\Sigma}_{x_d, \mathbf{z}_d}|) \tag{5}
$$

where $k = |\mathbf{Z}_C|$ is the dimensionality of the Gaussian distribution and $\mathbf{\Sigma}_{x_d, \mathbf{z}_d}$ is the covariance of the data with the discrete configuration $x_d, \mathbf{z}_d$.

The remaining terms $H(\mathbf{Z}_C \mid Y_D, \mathbf{Z}_D)$, $H(\mathbf{Z}_C \mid X_D, Y_D, \mathbf{Z}_D)$ and $H(\mathbf{Z}_C \mid \mathbf{Z}_D)$ can be calculated similarly.

2.2 Mutual Information between a Discrete and Continuous Variable

$$
\begin{aligned}
I(X_D; Y_C \mid \mathbf{Z}_D, \mathbf{Z}_C) &= H(X_D, \mathbf{Z}_D, \mathbf{Z}_C) + H(Y_C, \mathbf{Z}_D, \mathbf{Z}_C) - H(X_D, Y_C, \mathbf{Z}_D, \mathbf{Z}_C) - H(\mathbf{Z}_D, \mathbf{Z}_C) \\
&= H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) + H(X_D, \mathbf{Z}_D) + H(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) + H(\mathbf{Z}_D) \\
&\quad - H(Y_C, \mathbf{Z}_C \mid X_D, \mathbf{Z}_D) - H(X_D, \mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D) - H(\mathbf{Z}_D) \\
&= H(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) + H(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) - H(Y_C, \mathbf{Z}_C \mid X_D, \mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D)
\end{aligned} \tag{6}
$$

where the entropy terms can be calculated as in (4), but in this case the variable Y is added in some of the estimated multivariate Gaussian distributions.

2.3 Mutual Information between Two Continuous Variable

For an unconditional mutual information, the mutual information can be calculated with the correlation coefficient:

$$
I(X_C; Y_C) = -\frac{1}{2} \log \left(1 - \rho^2 \right) \tag{7}
$$

where ρ is the linear correlation coefficient between X and Y.

For the general case:

$$
\begin{aligned}
I(X_C; Y_C \mid \mathbf{Z}_D, \mathbf{Z}_C) &= H(X_C, \mathbf{Z}_D, \mathbf{Z}_C) + H(Y_C, \mathbf{Z}_D, \mathbf{Z}_C) - H(X_C, Y_C, \mathbf{Z}_D, \mathbf{Z}_C) - H(\mathbf{Z}_D, \mathbf{Z}_C) \\
&= H(X_C, \mathbf{Z}_C \mid \mathbf{Z}_D) + H(\mathbf{Z}_D) + H(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) + H(\mathbf{Z}_D) \\
&\quad - H(X_C, Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) - H(\mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D) - H(\mathbf{Z}_D) \\
&= H(X_C, \mathbf{Z}_C \mid \mathbf{Z}_D) + H(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) - H(X_C, Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) - H(\mathbf{Z}_C \mid \mathbf{Z}_D)
\end{aligned} \tag{8}
$$

where the entropy terms can be calculated as in (4), but in this case the variables X and Y are added in some of the estimated multivariate Gaussian distributions.

3 Empirical Degrees of Freedom

This section shows the empirical degrees of freedom by running a simulation over 1000 datasets of 100000 instances that are compatible with the null hypothesis (conditional independence). The empirical degrees of freedom have been rounded to the nearest integer number.

3.1 Empirical Degrees of Freedom between Two Discrete Variables

1 cont. parents				2 cont. parents				3 cont. parents				4 cont. parents			
llx	lly	llz	df	llx	lly	llz	df	llx	lly	llz	df	llx	lly	llz	df
2	4	3	18	2	4	3	36	2	4	3	63	2	4	3	99
2	3	4	16	2	3	4	32	2	3	4	56	2	3	4	88
3	4	2	24	3	4	2	48	3	4	2	84	3	4	2	132

Inducted formula:

$$\text{df} = (\text{llx} - 1) \cdot (\text{lly} - 1) \cdot \text{llz} \cdot \left[1 + \frac{|\mathbf{Z}_C| \cdot (|\mathbf{Z}_C| + 1)}{2}\right] \tag{9}$$

3.2 Empirical Degrees of Freedom between a Discrete and Continuous Variable

1 cont. parents				2 cont. parents				3 cont. parents				4 cont. parents			
llx	conty	llz	df	llx	conty	llz	df	llx	conty	llz	df	llx	conty	llz	df
2		3	6	2		3	9	2		3	12	2		3	15
2		4	8	2		4	12	2		4	16	2		4	20
3		2	8	3		2	12	3		2	16	3		2	20
3		4	16	3		4	24	3		4	32	3		4	40
4		2	12	4		2	18	4		2	24	4		2	30
4		3	18	4		3	27	4		3	36	4		3	45

Inducted formula:

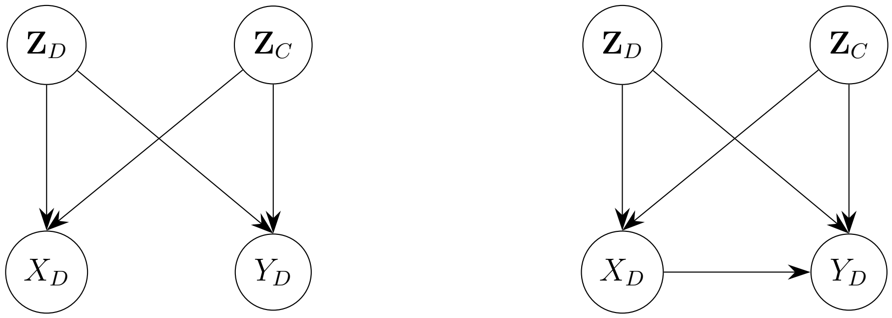
$$\text{df} = (\text{llx} - 1) \cdot \text{llz} \cdot [1 + |\mathbf{Z}_C|] \tag{10}$$

3.3 Empirical Degrees of Freedom between Two Continuous Variables

Inducted formula:

$$\text{df} = \text{llz} \tag{11}$$

1 cont. parents				2 cont. parents				3 cont. parents				4 cont. parents			
contx	conty	llz	df	contx	conty	llz	df	contx	conty	llz	df	contx	conty	llz	df
		2	2			2	2			2	2			2	2
		3	3			3	3			3	3			3	3
		4	4			4	4			4	4			4	4



(a) H_0 model (b) H_1 model

Figure 1: Null hypothesis (left) and alternative (right) models for two discrete variables.

4 Asymptotic Degrees of Freedom

4.1 Asymptotic Degrees of Freedom between Two Discrete Variables

There is a direct relationship between mutual information and a likelihood ratio test (G-test):

$$G = 2 \cdot N \cdot I(X_D; Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C) \tag{12}$$

The G statistic is distributed as a χ^2 if the null hypothesis is true. The degrees of freedom of the χ^2 distribution is the difference in the number of free parameters between a model where there is conditional dependence between X_D and Y_D (Figure 1b), and a model where there is no conditional dependence between X_D and Y_D (Figure 1a).

The only node that contains a different number of parameters is the conditional distribution of Y_D. So we must analyze that distribution to find the degrees of freedom of the χ^2 distribution.

For the H_0 model, the distribution $f(Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C)$ can be defined using the Bayes rule (as in a conditional linear Gaussian networks the discrete nodes do not have continuous parents):

$$f(Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C) = \frac{f(\mathbf{Z}_C \mid Y_D, \mathbf{Z}_D) f(Y_D \mid \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{13}$$

Note that only $\text{lly} - 1$ models are needed to be fitted because the probabilities $f(Y_D \mid \mathbf{Z}_D, \mathbf{Z}_C)$ must sum to 1, so the probability for the last category can be defined as:

$$f(Y_D = \text{lly} \mid \mathbf{Z}_D, \mathbf{Z}_C) = 1 - \sum_{i=1}^{\text{lly}-1} \frac{f(\mathbf{Z}_C \mid Y_D = i, \mathbf{Z}_D) f(Y_D = i \mid \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{14}$$

$f(\mathbf{Z}_C \mid Y_D, \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\text{lly} - 1) \cdot \text{llz} \cdot \frac{|\mathbf{Z}_C| \cdot (|\mathbf{Z}_C| + 3)}{2} \tag{15}$$

Note that $\frac{|\mathbf{Z}_C|\cdot(|\mathbf{Z}_C|+3)}{2}$ is the number of free parameters of a multivariate Gaussian distribution of $|\mathbf{Z}_C|$ dimensions.

$f(Y_D \mid \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\mathrm{lly}-1)\cdot \mathrm{llz} \tag{16}$$

$f(\mathbf{Z}_C \mid \mathbf{Z}_D)$ does not contain free parameters because it can be represented using the previous functions:

$$f(\mathbf{Z}_C \mid \mathbf{Z}_D) = \sum_{y_d \in Y_D} f(\mathbf{Z}_C \mid Y_D = y_d, \mathbf{Z}_D) f(Y_D = y_d \mid \mathbf{Z}_D) \tag{17}$$

For the H_1 model, the distribution $f(Y_D \mid X_D, \mathbf{Z}_D, \mathbf{Z}_C)$ can be defined using the Bayes rule (as in a conditional linear Gaussian networks the discrete nodes do not have continuous parents):

$$f(Y_D \mid X_D, \mathbf{Z}_D, \mathbf{Z}_C) = \frac{f(\mathbf{Z}_C \mid X_D, Y_D, \mathbf{Z}_D) f(Y_D \mid X_D, \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D)} \tag{18}$$

Note that only lly-1 models are needed to be fitted because the probabilities $f(Y_D \mid X_D, \mathbf{Z}_D, \mathbf{Z}_C)$ must sum to 1, so the probability for the last category can be defined as:

$$f(Y_D = \mathrm{lly} \mid X_D, \mathbf{Z}_D, \mathbf{Z}_C) = 1 - \sum_{i=1}^{\mathrm{lly}-1} \frac{f(\mathbf{Z}_C \mid X_D, Y_D = i, \mathbf{Z}_D) f(Y_D = i \mid X_D, \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D)} \tag{19}$$

$f(\mathbf{Z}_C \mid X_D, Y_D, \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\mathrm{lly}-1)\cdot \mathrm{llx}\cdot \mathrm{llz}\cdot \frac{|\mathbf{Z}_C|\cdot(|\mathbf{Z}_C|+3)}{2} \tag{20}$$

$f(Y_D \mid X_D, \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\mathrm{lly}-1)\cdot \mathrm{llx}\cdot \mathrm{llz} \tag{21}$$

$f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D)$ does not contain free parameters because it can be represented using the previous functions:

$$f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) = \sum_{y_d \in Y_D} f(\mathbf{Z}_C \mid X_D, Y_D = y_d, \mathbf{Z}_D) f(Y_D = y_d \mid X_D, \mathbf{Z}_D) \tag{22}$$

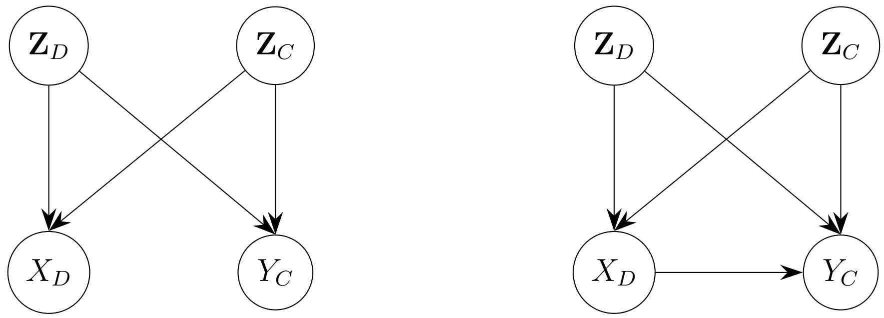
The difference in parameters (and the degrees of freedom of the χ^2) is equal to:

$$\begin{aligned} \mathrm{df} &= \mathrm{lly}-1)\cdot(\mathrm{llx}-1)\cdot \mathrm{llz}\cdot \frac{|\mathbf{Z}_C|\cdot(|\mathbf{Z}_C|+3)}{2} + (\mathrm{lly}-1)\cdot(\mathrm{llx}-1)\cdot \mathrm{llz} \\ &= \boxed{(\mathrm{lly}-1)\cdot(\mathrm{llx}-1)\cdot \mathrm{llz}\left[1 + \frac{|\mathbf{Z}_C|\cdot(|\mathbf{Z}_C|+3)}{2}\right]} \end{aligned} \tag{23}$$

4.2 Asymptotic Degrees of Freedom between a Discrete and Continuous Variable

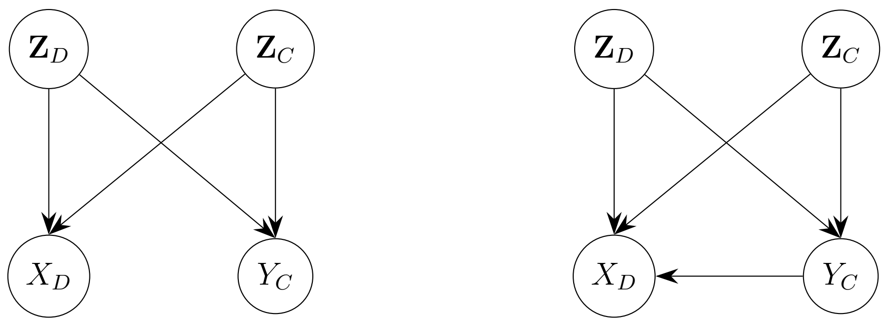
The H_0 model is shown in Figure 2a and the H_1 model is shown in Figure 2b.

The distribution $f(Y_C \mid \mathbf{Z}_D, \mathbf{Z}_C)$ has a number of free parameters equal to:



(a) H_0 model (b) H_1 model

Figure 2: Null hypothesis (left) and alternative (right) models for a continuous and discrete variable.



(a) H_0 model (b) H_1 model

Figure 3: Variation of the null hypothesis (left) and alternative (right) models for a continuous and discrete variable.

$$\mathrm{llz} \cdot (|\mathbf{Z}_C| + 2) \tag{24}$$

The distribution $f(Y_C \mid X_D, \mathbf{Z}_D, \mathbf{Z}_C)$ has a number of free parameters equal to:

$$\mathrm{llx} \cdot \mathrm{llz} \cdot (|\mathbf{Z}_C| + 2) \tag{25}$$

The difference in parameters (and the degrees of freedom of the χ^2) is equal to:

$$\boxed{\mathrm{df} = (\mathrm{llx} - 1) \cdot \mathrm{llz} \cdot (|\mathbf{Z}_C| + 2)} \tag{26}$$

The same result can be derived using a different H_0 model (Figure 3a) and H_1 model (Figure 3b). In this case, the difference in the number of parameters happens to be in the conditional distribution of X_D.

The $f(X_D \mid \mathbf{Z}_D, \mathbf{Z}_C)$ can be defined using the Bayes rule:

$$f(X_D \mid \mathbf{Z}_D, \mathbf{Z}_C) = \frac{f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D) f(X_D \mid \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{27}$$

Note that only $\mathrm{llx} - 1$ models are needed to be fitted because the probabilities $f(X_D \mid \mathbf{Z}_D, \mathbf{Z}_C)$ must sum to 1, so the probability for the last category can be defined as:

$$f(X_D = \text{llx} \mid \mathbf{Z}_D, \mathbf{Z}_C) = 1 - \sum_{i=1}^{\text{llx}-1} \frac{f(\mathbf{Z}_C \mid X_D = i, \mathbf{Z}_D) f(X_D = i \mid \mathbf{Z}_D)}{f(\mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{28}$$

The distribution $f(\mathbf{Z}_C \mid X_D, \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\text{llx} - 1) \cdot \text{llz} \cdot \frac{|\mathbf{Z}_C| \cdot (|\mathbf{Z}_C| + 3)}{2} \tag{29}$$

The distribution $f(X_D \mid \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\text{llx} - 1) \cdot \text{llz} \tag{30}$$

$f(\mathbf{Z}_C \mid \mathbf{Z}_D)$ does not contain free parameters because it can be represented using the previous functions:

$$f(\mathbf{Z}_C \mid \mathbf{Z}_D) = \sum_{x_d \in X_D} f(\mathbf{Z}_C \mid X_D = x_d, \mathbf{Z}_D) f(X_D = x_d \mid \mathbf{Z}_D) \tag{31}$$

The $f(X_D \mid \mathbf{Z}_D, \mathbf{Z}_C)$ can be defined using the Bayes rule:

$$f(X_D \mid Y_C, \mathbf{Z}_D, \mathbf{Z}_C) = \frac{f(Y_C, \mathbf{Z}_C \mid X_D, \mathbf{Z}_D) f(X_D \mid \mathbf{Z}_D)}{f(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{32}$$

Note that only $\text{llx} - 1$ models are needed to be fitted because the probabilities $f(X_D \mid Y_C, \mathbf{Z}_D, \mathbf{Z}_C)$ must sum to 1, so the probability for the last category can be defined as:

$$f(X_D = \text{llx} \mid Y_C, \mathbf{Z}_D, \mathbf{Z}_C) = 1 - \sum_{i=1}^{\text{llx}-1} \frac{f(Y_C, \mathbf{Z}_C \mid X_D = i, \mathbf{Z}_D) f(X_D = i \mid \mathbf{Z}_D)}{f(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D)} \tag{33}$$

The distribution $f(Y_C, \mathbf{Z}_C \mid X_D, \mathbf{Z}_D)$ has a number of free parameters equal to:

$$(\text{llx} - 1) \cdot \text{llz} \cdot \frac{(|\mathbf{Z}_C| + 1) \cdot (|\mathbf{Z}_C| + 4)}{2} \tag{34}$$

The distribution $f(X_D \mid \mathbf{Z}_D)$ has a number of free parameters equal to:

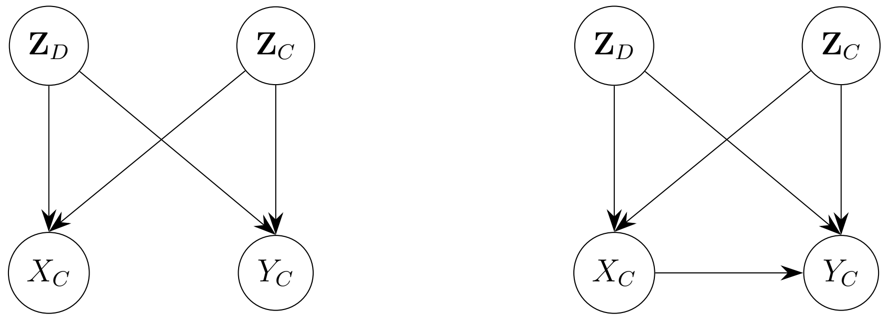
$$(\text{llx} - 1) \cdot \text{llz} \tag{35}$$

$f(\mathbf{Z}_C \mid \mathbf{Z}_D)$ does not contain free parameters because it can be represented using the previous functions:

$$f(Y_C, \mathbf{Z}_C \mid \mathbf{Z}_D) = \sum_{x_d \in X_D} f(Y_C, \mathbf{Z}_C \mid X_D = x_d, \mathbf{Z}_D) f(X_D = x_d \mid \mathbf{Z}_D) \tag{36}$$

The difference in parameters (and the degrees of freedom of the χ^2) is equal to:

$$\begin{aligned}
\text{df} &= (\text{llx} - 1) \cdot \text{llz} \cdot \left(\frac{|\mathbf{Z}_C|^2 + 5|\mathbf{Z}_C| + 4 - |\mathbf{Z}_C|^2 - 3|\mathbf{Z}_C|}{2} \right) \\
&= (\text{llx} - 1) \cdot \text{llz} \cdot \left(\frac{2|\mathbf{Z}_C| + 4}{2} \right) \\
&= \boxed{(\text{llx} - 1) \cdot \text{llz} \cdot (|\mathbf{Z}_C| + 2)}
\end{aligned} \tag{37}$$



(a) H_0 model (b) H_1 model

Figure 4: Null hypothesis (left) and alternative (right) models for two continuous variables.

4.3 Asymptotic Degrees of Freedom between Two Continuous Variables

The H_0 model is shown in Figure 4a and the H_1 model is shown in Figure 4b.

For the H_0 model, the distribution $f(Y_C \mid \mathbf{Z}_D, \mathbf{Z}_C)$ has a number of free parameters equal to:

$$\text{llz} \cdot (|\mathbf{Z}_C| + 2) \tag{38}$$

For the H_1 model, the distribution $f(Y_C \mid X_C, \mathbf{Z}_D, \mathbf{Z}_C)$ has a number of free parameters equal to:

$$\text{llz} \cdot (|\mathbf{Z}_C| + 3) \tag{39}$$

The difference in parameters (and the degrees of freedom of the χ^2) is equal to:

$$\boxed{\text{llz}} \tag{40}$$