Goodness of Causal Fit

Robert R. Tucci
tucci@ar-tiste.com

May 5, 2021

Abstract

We propose a Goodness of Causal Fit (GCF) measure which depends on Pearl "do" interventions. This is different from a measure of Goodness of Fit (GF), which does not use interventions. Given a DAG set $\mathcal{G}$, to find a good $G \in \mathcal{G}$, we propose plotting $GCF(G)$ versus $GF(G)$ for all $G \in \mathcal{G}$, and finding a graph $G \in \mathcal{G}$ with a large amount of both types of goodness.

1 Introduction

Frequently, when students first encounter Bayesian Networks (bnets) and Causal Inference (CI) (Refs.[1], [3]), they experience serious doubts about the usefulness of this theory, because they believe finding the underlying model (i.e., DAG) for most realistic physical situations is too difficult or impossible. I think that part of the problem is that these students are assuming, perhaps unconsciously, that there exists a unique DAG that fits Nature perfectly, and a mind-boggling number of possibilities to sift through to find that DAG. Rather than looking for a unique DAG, I think a better strategy is to write down a set $\mathcal{G}$ of likely DAGs, and to calculate for each DAG in $\mathcal{G}$, a measure called Goodness of Causal Fit (GCF). Then use the DAGs with the highest GCF scores.

The goal of this paper is to propose a GCF measure. Such a measure is of course not unique, and someone may propose in the future a measure that is better than ours.

When designing a GCF measure, it is important to keep in mind the First Dictum[1] of CI: The data is model-less. In the First Dictum, when we say "data", we are referring to what is commonly called a dataset. A dataset is a table of data, where all the entries of each column have the same units, and measure a single feature, and each row refers to one particular sample or individual. Datasets are particularly useful for estimating probability distributions and for training neural nets. In the First Dictum, when we say "model", we are referring to a DAG (directed acyclic graph) or a bnet (Bayesian Network= DAG + probability table for each node of DAG).

You can try to derive a model from a dataset, but you'll soon find out that you can only go so far. The process of finding a *partial* model from a dataset is called structure learning (SL). SL can be done quite nicely with Marco Scutari's open source program `bnlearn` (Ref[2]). The problem is that SL often cannot narrow down the model to a single one. It finds an undirected graph (UG), and it can determine the direction of some of the arrows in the UG, but it is often incapable, for well understood fundamental —not just technical— reasons, of finding the direction of *all* the arrows of the UG. So it often fails to fully specify a DAG model.

Let's call the ordered pair (dataset, model) a **data SetMo**. Then what the First Dictum is saying is that a dataset is model-free or model-less (although sometimes one can find a partial model hidden in there). A dataset is not a data SetMo.

Graphs which contain both directed and undirected edges are called **partially directed (PD) graphs**. `bnlearn` takes a dataset as input and returns a PD graph G_{pd}. Given a PD graph G_{pd}, let $\mathcal{G}(G_{pd})$ be the DAG set $\mathcal{G}$ which is generated by giving directions to all undirected edges of G_{pd} in all possible ways. We will refer to $\mathcal{G}(G_{pd})$ as the **DAG set generated by** G_{pd}. and to any $\mathcal{G}' \subset \mathcal{G}(G_{pd})$ as a **DAG set partially generated by** G_{pd}**.** Once we define below our GCF measure, we will

[1] This is just my whimsical name for it.

apply it to sets of the type $\mathcal{G}(G_{pd})$ as an example.

It's clear that any measure of GCF will have to involve interventions such as the "do" intervention (see Refs. [1] and [3]) invented by Pearl et al for CI. Without interventions like "do", it is impossible to distinguish causally the DAGs in a set $\mathcal{G}(G_{pd})$.

Henceforth, random variables will be indicated by underlining.

2 Goodness of Fit

Before trying to define a GCF measure, it is instructive to review the closely related, well established, measures of Goodness of Fit (GF).

Consider two probability distributions $PO(x)$ and $PE(x)$, where $x \in S_{\underline{x}}$. By a GF measure, we mean a measure of the difference between PO and PE. Usually PO is the observed probability distribution and PE is the expected, theoretical one.

Three popular measures of the difference between PO and PE are:

1. The **Kullback-Liebler divergence**:

$$D_{KL}(PO \parallel PE) = \sum_{x \in S_{\underline{x}}} PO(x) \ln \frac{PO(x)}{PE(x)} \; . \tag{1}$$

2. The **Pearson divergence** (aka **Pearson Chi-squared test statistic**):

$$D_{\chi^2}(PO \parallel PE) = \sum_{x \in S_{\underline{x}}} \frac{[PO(x) - PE(x)]^2}{PE(x)} = \sum_{x \in S_{\underline{x}}} \frac{PO^2(x)}{PE(x)} - 1 \; . \tag{2}$$

   It's easy to show using $\ln(1+\delta) = \delta + \mathcal{O}(\delta^2)$ that if $\left| \frac{PO(x)}{PE(x)} - 1 \right| << 1$ for all $x \in S_{\underline{x}}$, then

$$D_{KL}(PO \parallel PE) \approx D_{\chi^2}(PO \parallel PE) \tag{3}$$

3. The **Euclidean distance squared**:

$$D_E(PO, PE) = \sum_{x \in S_{\underline{x}}} [PO(x) - PE(x)]^2 \tag{4}$$

Note that of these 3 measures, only $D_E(PO, PE)$ is symmetric in PO and PE.

Given any bnet G with full probability distribution [2] $P_G(x.)$ and a probability distribution[3] $\tilde{P}(x.)$ derived empirically from a dataset, let

[2] We define $x.$ to be a vector with components x_i

[3] Empirical distributions will be denoted by P with a tilde over it.

$$D(G) = \sum_{x.} \tilde{P}(x.) \ln \frac{\tilde{P}(x.)}{P_G(x.)} \quad (5)$$

$$= D_{KL}(\tilde{P}(\underline{x}.) \parallel P_G(\underline{x}.)) \quad (6)$$

We define **Goodness of Fit (GF)** of the bnet G by

$$GF(G) = \ln \frac{1}{D(G)} \quad (7)$$

3 GCF example 1

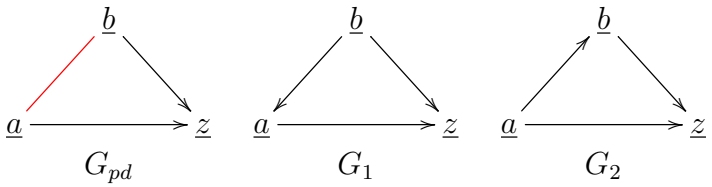

Figure 1: $\mathcal{G}(G_{pd}) = \{G_1, G_2\}$. The partially directed graph G_{pd} generates the DAGs G_1 and G_2 by giving directions to all undirected edges of G_{pd} in all possible ways. (In this case, there is only one undirected edge in G_{pd}.)

For the first example of our measure of GCF, we consider $\mathcal{G}(G_{pd}) = \{G_1, G_2\}$ given by Fig.1. We will assume the following:

- First, we assume that we have collected a dataset from which we have extracted a full empirical distribution $\tilde{P}(z, a, b)$. From $\tilde{P}(z, a, b)$, we assume that the following have been calculated. $\tilde{P}(z, b|a)$ $\tilde{P}(z, a|b)$ $\tilde{P}(a)$, $\tilde{P}(b)$.

- Second, we assume that a dataset has been collected for which $\underline{a}$ was held fixed to each of the possible values $a \in S_{\underline{a}}$ of $\underline{a}$. Furthemore, we assume that the distribution $\tilde{P}(z, b|do(\underline{a}) = a)$ has been extracted from that dataset.

- Third, we assume that a dataset has been collected for which $\underline{b}$ was held fixed to each of the possible values $b \in S_{\underline{b}}$ of $\underline{b}$. Furthemore, we assume that the distribution $\tilde{P}(z, a|do(\underline{b}) = b)$ has been extracted from that dataset.

We will refer to $\tilde{P}(z, b|do(\underline{a}) = a)$ and $\tilde{P}(z, a|do(\underline{b}) = b)$ as **empirical do-probabiity distributions**.

Now define

$$\begin{aligned} D_a &= \sum_{z,b} \tilde{P}(z,b|a) \ln \frac{\tilde{P}(z,b|a)}{\tilde{P}(z,b|do(\underline{a})=a)} && (8) \\ &= D_{KL}(\tilde{P}(\underline{z},\underline{b}|a) \parallel \tilde{P}(\underline{z},\underline{b}|do(\underline{a})=a)) && (9) \\ D_{\underline{a}} &= \sum_a \tilde{P}(a) D_a = E_a[D_a] && (10) \end{aligned}$$

and

$$\begin{aligned} D_b &= D_{KL}(\tilde{P}(\underline{z},\underline{a}|b) \parallel \tilde{P}(\underline{z},\underline{a}|do(\underline{b})=b)) && (11) \\ D_{\underline{b}} &= \sum_a \tilde{P}(b) D_b = E_b[D_b] \, . && (12) \end{aligned}$$

We will refer to $D_{\underline{a}}$ and $D_{\underline{b}}$ as **do-divergences**.

Note that

$$D_a(G_2) = 0 \text{ for all } a \text{ so } \underbrace{D_{\underline{a}}(G_2)}_{0} \leq D_{\underline{b}}(G_2) \tag{13}$$

and

$$D_b(G_1) = 0 \text{ for all } b \text{ so } D_{\underline{a}}(G_1) \geq \underbrace{D_{\underline{b}}(G_1)}_{0} \, . \tag{14}$$

If $D_{\underline{a}} \leq D_{\underline{b}}$, then $\underline{a} \to \underline{b}$, and if $D_{\underline{a}} \geq D_{\underline{b}}$ then $\underline{a} \leftarrow \underline{b}$. Thus, the arrow and the inequality sign point in opposite directions. Alternatively, just remember that the arrow points to the larger of the two D's.

If $D_{\underline{a}} \leq D_{\underline{b}}$, then define $GCF(G_1) = -1$ and $GCF(G_2) = 1$.

If $D_{\underline{b}} \leq D_{\underline{a}}$, then define $GCF(G_1) = 1$, $GCF(G_2) = -1$.

4 GCF example 2

For the second example of our measure of GCF, consider $\mathcal{G} = \{G_1, G_2, G_3\}$ given by Fig.2.

Which of the do-divergences $D_{\underline{x}_2}$, $D_{\underline{x}_1}$ and $D_{\underline{x}_3}$, is smallest, middle and highest, depends on the empirical do-probability distributions. For definiteness, suppose the sizes of these do-divergences are related as follows:

$$D_{\underline{x}_2} \leq D_{\underline{x}_1} \leq D_{\underline{x}_3} \, . \tag{15}$$

For any two do-divergences $D_{\underline{a}}$ and $D_{\underline{b}}$, define the distance

$$d_{\underline{b},\underline{a}} = |D_{\underline{b}} - D_{\underline{a}}| \tag{16}$$

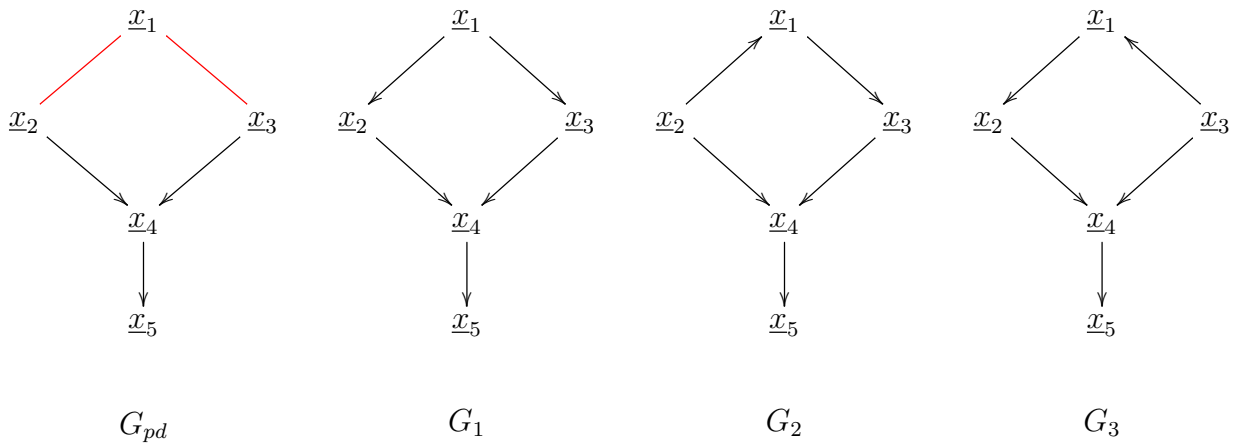

Figure 2: $\mathcal{G} = \{G_1, G_2, G_3\}$. $\mathcal{G}$ is a set of observationally equivalent (OE) graphs. These are graphs that have the same value for GF, and are therefore indistinguishable from GF alone. For more info about OE graphs, see Chapter entitled "Observationally Equivalent DAGs" in Ref.[3]. Note that $\mathcal{G}(G_{pd})$ includes one more DAG, the one in which node $\underline{x}_1$ is a collider.

If we abbreviate $D_{\underline{x}_j}$ by D_j, we can define the GCF for each of the graphs in $\mathcal{G}$ by:

$$GCF(G_1) = \frac{-d_{2,1} + d_{1,3}}{d_{2,1} + d_{1,3}} \tag{17a}$$

$$GCF(G_2) = \frac{d_{2,1} + d_{1,3}}{d_{2,1} + d_{1,3}} = 1 \tag{17b}$$

$$GCF(G_3) = \frac{-d_{2,1} - d_{1,3}}{d_{2,1} + d_{1,3}} = -1 \tag{17c}$$

5 GCF in general

Eqs.(17) are a special case of the following formulas.

For any two do-divergences $D_{\underline{a}}$ and $D_{\underline{b}}$, define the **do-divergence distance** by

$$d_{\underline{b},\underline{a}} = |D_{\underline{b}} - D_{\underline{a}}| \ . \tag{18}$$

For any $G_i \in \mathcal{G}$, define the **edge-sign function** by

$$\sigma_{G_i}(\underline{a}—\underline{b}) = \begin{cases} +1 & \text{if edge } \underline{a}—\underline{b} \text{ in } G_i \text{ points towards larger of } D_{\underline{a}} \text{ and } D_{\underline{b}}. \\ -1 & \text{otherwise} \end{cases} \tag{19}$$

Finally, suppose that $\mathcal{G}$ is either partially or fully generated by a PD graph G_{pd} with undirected edges $\{\underline{a}_k\text{—}\underline{b}_k\}_{k=0,1,\ldots,nk-1}$. Then define the GCF of graph $G_i \in \mathcal{G}$ by

$$GCF(G_i) = \frac{\sum_{k=0}^{nk-1} \sigma_{G_i}(\underline{a}_k\text{—}\underline{b}_k) d_{\underline{a}_k,\underline{b}_k}}{\sum_{k=0}^{nk-1} d_{\underline{a}_k,\underline{b}_k}} \ . \tag{20}$$

Note that $-1 \leq GCF(G_i) \leq 1$.

If the DAG set $\mathcal{G}$ contains only one DAG G, define $GCF(G) = 1$, because all arrows in G are in the correct direction, as far as we know.

Call the **skeleton** of a DAG, the undirected graph that one obtains by turning all its edges from directed to undirected ones.

So far, we have applied our measure of GCF to a DAG set $\mathcal{G}$ which is either fully or partially generated by a PD graph G_{pd}, or is a singleton set. But what if we want a GCF that can score every DAG in a DAG set $\mathcal{G}$ that contains DAGs with different skeletons? In that case, let $\{\underline{a}_k\text{—}\underline{b}_k\}_{k=0,1,\ldots,nk-1}$ be *all* the edges of graph $G_i \in \mathcal{G}$, and use the relative GCF given by Eq.(20), or use an absolute GCF defined by

$$GCF_a(G_i) = \sum_{k=0}^{nk-1} \sigma_{G_i}(\underline{a}_k\text{—}\underline{b}_k) d_{\underline{a}_k,\underline{b}_k} \ . \tag{21}$$

So let $\mathcal{G}$ be an arbitrary DAG set. Our GCF (or GCF_a) measure is not enough to decide the best possible G in $\mathcal{G}$, because there might be several graphs with $GCF \approx 1$. For this reason, we recommend plotting $GCF(G)$ (or GCF_a) versus $GF(G)$ for all $G \in \mathcal{G}$. Then choose a G with a large amount of both types of goodness. It might even be advantageous to average over a small subset of DAGs in $\mathcal{G}$ that have large amounts of both types of goodness. This would be similar to averaging over an ensemble of decision trees to get a random forest.

A plot of $GCF(G)$ versus $GF(G)$ agrees with the spirit of the First Dictum and data setmos, because also in those, we acknowledge a separation between the dataset (GF) and model (GCF) degrees of freedom.

References

[1] Judea Pearl. *Causality: Models, Reasoning, and Inference, Second Edition.* Cambridge University Press, 2013.

[2] Marco Scutari. bnlearn. `https://www.bnlearn.com/`.

[3] Robert R. Tucci. Bayesuvius (book). `https://github.com/rrtucci/Bayesuvius/raw/master/main.pdf`.