# Goodness of Causal Fit

Robert R. Tucci
tucci@ar-tiste.com

November 2, 2022

## Abstract

We propose a Goodness of Causal Fit (GCF) measure which depends on Judea Pearl's "do" interventions. This is different from Goodness of Fit (GF) measures, which do not use interventions. Given a set $\mathcal{G}$ of DAGs with the same nodes, 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 believe 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 a DAG with a high GCF score.

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.

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 Judea Pearl et al. Without interventions like "do", it might be impossible to distinguish which DAG of a set is the best causal fit. For example, the family of triangular bnets can all represent the same probability distribution because they are fully connected. Hence, from the probability distribution of the triangular bnet alone, it is impossible to decide which bnet in the family is the best causal fit for the physical situation being considered.

When designing a GCF measure, it is important to keep in mind the Data Axiom[1] of CI: A dataset is causal model-free. In the Data Axiom, when we say a "dataset", we are referring to 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 Data Axiom, when we say "causal model", we are referring to a DAG (directed acyclic graph) or a bnet (bnet= DAG + probability table for each node of DAG).

You can try to derive a causal model from a dataset, but you'll soon find out that you can only go so far. The process of finding a *partial* causal 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 causal 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.

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

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

dataset++.

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 $\mathcal{G}_{max}(G_{pd})$ be the DAG set which is generated by giving directions to all undirected edges of $G_{pd}$ in all possible ways. We will refer to the DAG set $\mathcal{G}_{max}(G_{pd})$ as the **maximal generation of** $G_{pd}$ and to any subset $\mathcal{G}(G_{pd})$ of $\mathcal{G}_{max}(G_{pd})$ as a **non-maximal generation of** $G_{pd}$**.** Once we define below our GCF measure, we will evaluate it for the DAGs of non-maximal generation $\mathcal{G}(G_{pg})$.

Henceforth, random variables will be indicated by underlining, as is done in Ref.[3].

# 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** (a.k.a. **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

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

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

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

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

# 3 GCF example 1

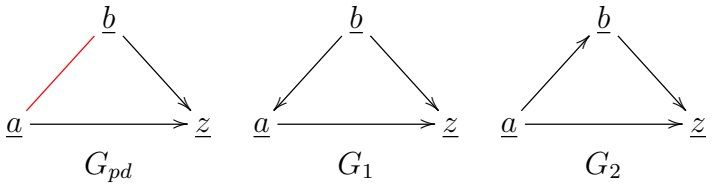


Figure 1: $\mathcal{G}(G_{pd}) = \{G_1, G_2\}$. From the partially directed graph $G_{pd}$, one can generate the DAGs $G_1$ and $G_2$ by giving directions to all undirected edges of $G_{pd}$ in all possible ways. (In this case, there is only one undirected edge in $G_{pd}$.)

For the first example of our GCF measure, 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 calculated 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 calculated from that dataset.

[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.

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

$$\begin{aligned} h_a &= \sum_{z,b} \tilde{P}(z,b|a) \ln \frac{\tilde{P}(z,b|a)}{\tilde{P}(z,b|do(\underline{a})=a)} && (8)\\ &= D_{KL}(\tilde{P}(\underline{z},\underline{b}|a) \parallel \tilde{P}(\underline{z},\underline{b}|do(\underline{a})=a)) && (9)\\ h_{\underline{a}} &= \sum_a \tilde{P}(a) h_a && (10)\\ &= E_a[h_a] && (11) \end{aligned}$$

and

$$\begin{aligned} h_b &= D_{KL}(\tilde{P}(\underline{z},\underline{a}|b) \parallel \tilde{P}(\underline{z},\underline{a}|do(\underline{b})=b)) && (12)\\ h_{\underline{b}} &= \sum_a \tilde{P}(b) h_b && (13)\\ &= E_b[h_b]\ . && (14) \end{aligned}$$

We will refer to $h_{\underline{x}}$ for any node $\underline{x}$ as a **hospitality**. Note that the hospitality for node $\underline{x}$ is zero if node $\underline{x}$ has no incoming arrows (i.e., is "inhospitable"), and becomes positive if node $\underline{x}$ does have some incoming arrows (i.e., is "hospitable"). Note $h_{\underline{b}} = h_{\underline{a}} = 0$ iff there is no arrow between $\underline{a}$ and $\underline{b}$.

Note that if the truth is $G_2$ with $\underline{a} \to \underline{b}$, then

$$h_a = 0 \text{ for all } a \text{ so } \underbrace{h_{\underline{a}}}_{0} \leq h_{\underline{b}} \tag{15}$$

and if the truth is $G_1$ with $\underline{b} \to \underline{a}$, then

$$h_b = 0 \text{ for all } b \text{ so } h_{\underline{a}} \geq \underbrace{h_{\underline{b}}}_{0}\ . \tag{16}$$

Hence, no matter what the truth is, the arrow connecting nodes $\underline{a}$ and $\underline{b}$ always points towards the larger of the 2 hospitalities (i.e., the arrow "seeks the most hospitable node")

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

# 4 GCF example 2

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

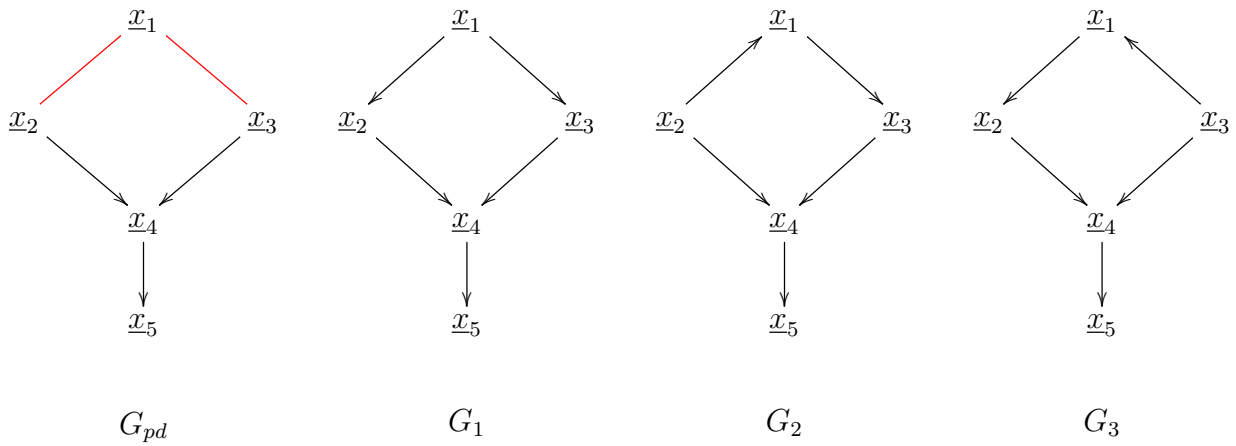


Figure 2: $\mathcal{G} = \{G_1, G_2, G_3\}$. $\mathcal{G}$ is a set of observationally equivalent (OE) graphs. These are graphs that have the same value for GF, and are therefore indistinguishable by means of GF alone. For more info about OE graphs, see Chapter entitled "Observationally Equivalent DAGs"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

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

$$h_{\underline{x}_2} \leq h_{\underline{x}_1} \leq h_{\underline{x}_3} \ . \tag{17}$$

For any two hospitalities $h_{\underline{a}}$ and $h_{\underline{b}}$, let

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

If we abbreviate $\underline{x}_j$ by $j$ when used as a subscript, we can define the GCF for each of the graphs in $\mathcal{G}$ by:

$$GCF(G_1) = \frac{-d_{2,1} + d_{1,3}}{d_{2,1} + d_{1,3}} \tag{19a}$$

$$GCF(G_2) = \frac{d_{2,1} + d_{1,3}}{d_{2,1} + d_{1,3}} = 1 \tag{19b}$$

$$GCF(G_3) = \frac{-d_{2,1} - d_{1,3}}{d_{2,1} + d_{1,3}} = -1 \tag{19c}$$

# 5 GCF in general

Suppose $G_i \in \mathcal{G}$, where $\mathcal{G}$ is a non-maximal generation of $G_{pd}$. In that case, we define a GFC measure as follows. Note that the following definition generalizes the definition of GFC measure that was used in the 2 special cases that we have considered so far.

For any bnet $G_i \in \mathcal{G}$ with a node $\underline{a}$, and such that $\underline{a}^c$ are all nodes of $G_i$ other than $\underline{a}$, define the **hospitality** $h_{\underline{a}}$ by

$$\begin{aligned} h_a &= \sum_{a^c} \tilde{P}(a^c|a) \ln \frac{\tilde{P}(a^c|a)}{\tilde{P}(a^c|do(\underline{a}) = a)} && (20) \\ &= D_{KL}(\tilde{P}(\underline{a}^c|a) \parallel \tilde{P}(\underline{a}^c|do(\underline{a}) = a)) && (21) \\ h_{\underline{a}} &= \sum_a \tilde{P}(a) h_a && (22) \\ &= E_a[h_a] && (23) \end{aligned}$$

For any two hospitalities $h_{\underline{a}}$ and $h_{\underline{b}}$, define the **hospitality distance** by

$$d_{\underline{b},\underline{a}} = |h_{\underline{b}} - h_{\underline{a}}| \tag{24}$$

Note that $d_{\underline{b},\underline{a}} = 0$ iff $h_{\underline{a}} = h_{\underline{b}}$. See Appendix A for a proof that if $h_{\underline{a}} = h_{\underline{b}}$ and there is no fine tuning, then there is no arrow between $\underline{a}$ and $\underline{b}$.

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

$$\rho_{G_i}(\underline{a}—\underline{b}) = \begin{cases} +1 & \text{if edge } \underline{a}—\underline{b} \text{ in } G_i \text{ points towards the larger of } h_{\underline{a}} \text{ and } h_{\underline{b}}. \\ -1 & \text{otherwise} \end{cases} \tag{25}$$

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

$$GCF(G_i) = \frac{\sum_{k=0}^{nk-1} \rho_{G_i}(\underline{a}_k—\underline{b}_k) d_{\underline{a}_k,\underline{b}_k}}{\sum_{k=0}^{nk-1} d_{\underline{a}_k,\underline{b}_k}} \ . \tag{26}$$

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 directions of arrows in $G$ are known.

Call an undirected graph a **frame** and define the **frame of a DAG** to be the frame that one obtains by turning all the edges of the DAG from directed to undirected ones.

So far, we have applied our GFC measure to a DAG set $\mathcal{G}$ which is either a maximal or non-maximal generation of 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 frames but the same nodes? In that case, let $F$ be the frame which is the union of all edges in all $G \in \mathcal{G}$. For each edge $\underline{a}$—$\underline{b}$ of $F$, if all the $G \in \mathcal{G}$ give the same direction to that edge, then give that direction to that edge in $F$. After doing this for all edges of $F$, call $G_{pd}$ the resulting PD graph. Modify each $G \in \mathcal{G}$ by adding to it the undirected edges that occur in $G_{pd}$ but not in $G$. The new $G$, call it $[G]_{mod}$, is PD. Remove $G$ from $\mathcal{G}$ and add to $\mathcal{G}$ the elements of the maximal generation $\mathcal{G}_{max}([G]_{mod})$. At this point, we have reduced our seemingly more complicated situation where $\mathcal{G}$ contains different frames with the same nodes to the original situation in which $\mathcal{G}$ is a non-maximal generation of $G_{pd}$.

So let $\mathcal{G}$ be an arbitrary set of DAGs with the same nodes. Our GCF 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 $GCF(G)$ versus $GF(G)$ for all $G \in \mathcal{G}$. Then choose a $G$ with a large amount of both types of goodness. A plot of $GCF(G)$ versus $GF(G)$ agrees with the spirit of the Data Axiom, because in that axiom we also acknowledge a separation between the degrees of freedom of the dataset and those of the causal model.

# A Appendix

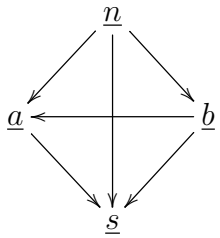


Figure 3: Bnet used to prove Claim 1. The proof is also valid if the direction of arrow $\underline{n} \to \underline{s}$ is reversed.

**Claim 1** *Suppose $\underline{a}, \underline{b}$ are any two nodes of a bnet $G_i$. Then*

*(a) $h_{\underline{b}} = h_{\underline{a}}$ implies $h_{\underline{b}} = h_{\underline{a}} = 0$ or bnet $G_i$ obeys the following* **fine tuning** *equations.*

$$\begin{cases} H(\underline{a}|\underline{b}, \underline{n}) = H(\underline{b}|\underline{n}) \\ h_{\underline{b}} = h_{\underline{a}} = 0 \text{ is false} \end{cases} \tag{27}$$

*where the $H(\cdot|\cdot)$ are conditional entropies and node $\underline{n}$ is defined in Fig.3.*

*(b) $h_{\underline{b}} = h_{\underline{a}} = 0$ implies $\underline{a}$ and $\underline{b}$ are both root nodes and there is no arrow between $\underline{a}$ and $\underline{b}$.*

**proof:** $(b)$ is obvious.

To prove $(a)$, consider Fig.3. In that figure, $\underline{n}$ and $\underline{s}$ might each represent multiple nodes of $G_i$. Let $\underline{x}$ denote all the nodes of $G_i$. Note that

$$\begin{aligned} h_{\underline{a}} &= \sum_x \tilde{P}(x) \ln \frac{\tilde{P}(a^c|a)}{\tilde{P}(a^c|do(\underline{a}) = a)} && (28) \\ &= \sum_x \tilde{P}(x) \ln \tilde{P}(a|b,n) && (29) \\ &= -H(\underline{a}|\underline{b},\underline{n}) && (30) \end{aligned}$$

$$\begin{aligned} h_{\underline{b}} &= \sum_x \tilde{P}(x) \ln \frac{\tilde{P}(b^c|b)}{\tilde{P}(b^c|do(\underline{b}) = b)} && (31) \\ &= \sum_x \tilde{P}(x) \ln \tilde{P}(b|n) && (32) \\ &= -H(\underline{b}|\underline{n}) && (33) \end{aligned}$$

**QED**

# 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`.