Group invariance principles for causal generative models

Michel Besserve[1,2,4] **Naji Shajarisales**[1,3] **Bernhard Schölkopf**[1] **Dominik Janzing**[1]
1. Max Planck Institute for Intelligent Systems, Tübingen, Germany
2. Max Planck Institute for Biological Cybernetics, Tübingen, Germany
3. Carnegie Mellon University, Pittsburgh, USA
4. Max Planck ETH Center for Learning Systems

Abstract

The postulate of independence of cause and mechanism (ICM) has recently led to several new causal discovery algorithms. The interpretation of independence and the way it is utilized, however, varies across these methods. Our aim in this paper is to propose a group theoretic framework for ICM to unify and generalize these approaches. In our setting, the cause-mechanism relationship is assessed by perturbing it with random group transformations. We show that the group theoretic view encompasses previous ICM approaches and provides a very general tool to study the structure of data generating mechanisms with direct applications to machine learning.

1 INTRODUCTION

Inferring causal relationships from empirical data is a challenging problem with major applications. While the problem of inferring such relations between arbitrarily many random variables (RVs) has been extensively addressed via conditional statistical independences in graphical models (Spirtes et al., 2000; Pearl, 2000), there are important limitations of this framework, such as the impossibility to infer the direction of causation in a graph with only two observed variables. This has motivated the search for new perspectives on causal inference. A major contribution to this line of research is to exploit a postulate of Independence of Cause and Mechanism (ICM) (Janzing and Schölkopf, 2010; Lemeire and Janzing, 2012; Schölkopf et al., 2012), which assumes that causes and mechanisms are chosen independently by Nature and thus P(cause) and P(effect|cause) do not contain information about each other. This absence of shared information can be formulated rigorously using algorithmic independence (Janzing and Schölkopf, 2010; Lemeire and Janzing, 2012), but also has implications in other contexts such as semi-supervised learning (Schölkopf et al., 2012; Peters et al., 2017). The main use of ICM postulates has been the development of several causal inference algorithms for cause-effect pairs (Janzing et al., 2010; Zscheischler et al., 2011; Daniusis et al., 2010; Janzing et al., 2012; Shajarisales et al., 2015; Sgouritsa et al., 2015); however, results in Schölkopf et al. (2012) also suggest it can be exploited in broader settings, providing guiding principles for the study of learning algorithms. Each of these methods addresses the causal inference problem with specific models, and are thus usable only for a restricted set of applications. Principled ways to generalize them to address new problems are yet unknown. In particular, it is unclear how the notion of "independence" should be defined for a given domain, and how it could impact the results. One conceptual difficulty of the ICM-based approaches is that independence is assessed between two objects of different nature: the input (or cause) and the mechanism; moreover, the appropriate notion is not the usual statistical independence of RVs.

In this paper, we suggest that a group theoretic framework can unify ICM-based approaches and provide useful tools to study generative models in general. This involves defining a group of generic transformations that perturb the relationship between mechanisms and causes, as well as an appropriate contrast function to assess the genericity of the cause-mechanism relationship. We show

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

that this framework encompasses previous ICM approaches (Janzing et al., 2010; Daniusis et al., 2010). In addition, while previous methods based on ICM were focused on cause-effect pairs (addressing which of the variables is the cause and which is the effect), the present paper shows ICM can be exploited for inferring latent variable generative models.

2 EXAMPLE IN VISION

2.1 Occlusion and illusory contours

In this section, we introduce the fundamental properties of our group theoretic perspective by studying a simplified version of a basic inference problem for visual perception: the identification of partially occluded objects. In two dimensional naturalistic visual scenes, an object can partially mask other objects standing behind it in the scene. This phenomenon is usually well addressed by the human visual system, but remains a major challenge for robust object detection in computer vision. Interestingly, even humans can be misled on this task by visual illusions. This is the case of the well know Kanizsa's triangle shown in Fig. 1a. In this figure, an illusory contour emerges from the precise alignment between the edges of the Pac-Man-shaped inducers, instigating the completion of each aligned segment pair into a larger edge and forming the illusion of a white triangle occluding three black disks. One way to describe the specificity of such figure is thus to count the number of lines carrying the straight edges of the three Pac-Man shapes in the figure: there are only three lines, which is atypically (or suspiciously) small for a figure made of three objects totaling six straight edges. The idea that a configuration is "atypical" lies at the heart of our causal inference framework, and we can indeed use the latter to address such scene understanding tasks as follows.

2.2 Formulation of the causal inference problem

We state the following scene understanding problem: two polygonal objects (with different colors) appear in a visual scene, occluding each other, and we want to infer which object partially occludes the other one. An example of such scene is represented on Fig. 1b, for which the most straightforward interpretation is that a red triangle stands in front of a yellow square. However, one could imagine on the contrary that a yellow object is in the foreground and occludes a red one, for instance by picking the objects shown

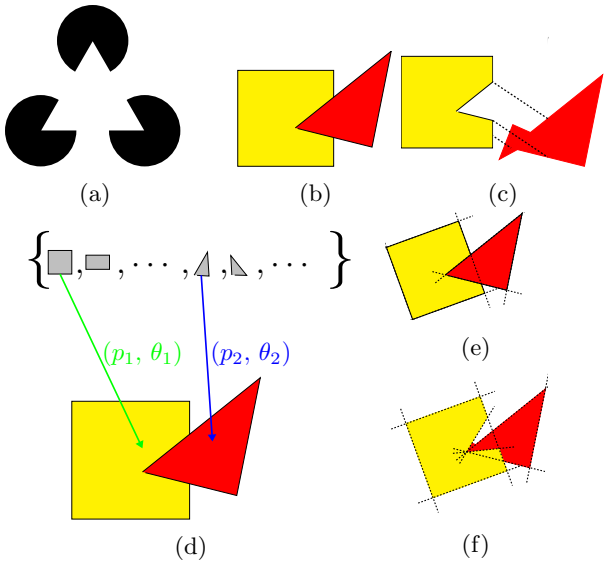


Figure 1: (a) Kanizsa's triangle. (b) Scene of a yellow square occluded by a red triangle. (c) Example of objects leading to the same scene as (b) but where red occludes yellow (dashed segments link superimposed points). (d) Generative model for (b). (e) New scene resulting from an arbitrary rotation of the yellow object in (b). Dashed lines indicate straight lines carrying the edges of the objects. (f) Same as (e) for the case presented in (b). See text for explanations.

in Fig. 1c. Such configuration is however intuitively unlikely if the precise positions and orientations of both objects are not carefully adjusted to lead to the scene shown in Fig. 1b, with the specific purpose to give the "illusion" that a yellow square is occluded. Such considerations have led vision scientists to formulate a *generic viewpoint assumption* in order to perform inference (Freeman, 1994). Such scene understanding problem can be considered as a causal inference problem, as it amounts to inferring a property of the generating mechanism (the objects and their configuration) that leads to an observation (the visual scene).

A generative model of the visual scene (see Fig. 1d for an illustration) may consist in the following mechanism: from a large collection of polygons, a first object O_1 is selected and put in the scene S at position p_1 with orientation θ_1. Then a second object O_2 is selected from the collection and put in front of the first at position p_2 and orientation θ_2. Leading to the Structural Equation Model (SEM)[1]

$$S := m_{p_1,\theta_1,p_2,\theta_2}(O_1, O_2)$$

[1]see next section for a presentation of this concept

where m denotes the object positioning mechanism. Under this generative model, we can either assign the red object R or the yellow one Y to the foreground, leading to $(O_1, O_2) = (Y, R)$ or $(O_1, O_2) = (R, Y)$ respectively. In order to determine which configuration is more likely, we resort to an ICM postulate. The cause being parametrized by the shape of the objects, and the mechanism by a set of positions and orientations, we assume that these last parameters are picked independently from the first ones. As a consequence, if we apply a random rotation to one of the objects, we expect that for most cases, some global properties of the image will be preserved, such that the original scene can be qualified as "typical". In this specific case, the number of lines $C(S)$ in the scene S is a simple global property of the scene that can be exploited.

Indeed, if we apply now a random rotation of angle ϕ to Y, r_ϕ, under the hypothesis that R is in front, we obtain a modified scene $\hat{S}_\phi \coloneqq m_{p_Y,\theta_Y,p_R,\theta_R}(r_\phi Y, R)$ (illustrated in Fig. 1e) that is similar to the original figure in the sense that the total number of lines carrying objects' edges is 7 in both cases (one for each side of each object). We thus can write

$$C(\hat{S}_\phi) = C(S),$$

that we call a genericity equation, stating that the arrangement in S is generic. On the contrary, under the hypothesis that Y is in front, we typically get a configuration $\hat{S}'_\phi \coloneqq m_{p_R,\theta_R,p_Y,\theta_Y}(R, r_\phi Y)$ like the one in Fig. 1f, with a larger number of lines in the scene. Indeed, for almost all choices of ϕ

$$C(\hat{S}'_\phi) \geq C(S) + 2,$$

and therefore the model with $(O_1, O_2) = (R, Y)$ is atypical, as witnessed by the lack of invariance of C values, and thus less likely than the model with $(O_1, O_2) = (Y, R)$ to explain the scene.

To summarize this experiment, if we assume the red object is in front, the number of straight lines is "typical" or generic since an arbitrary rotation of one object will typically lead to the same number of edges. On the contrary, if we assume the yellow object is in front, the number of lines is suspiciously low with respect to what it becomes when modifying the generative model with an arbitrary rotation. We introduced in this example the key elements of our framework: 1/ a generative model of the observed data, 2/ a group of generic transformations (here rotations) that can be applied to the model to simulate "typical" configurations of the generative model and 3/ a contrast (here the number of lines in the scene) that can be evaluated on both the observed data and in typical configurations of the model. Invariance of the contrast to generic transformations then indicates observations are typical. All these elements will be used for defining a general framework for causal inference in section 3.

3 GENERAL FRAMEWORK

3.1 Background and related work

Many machine learning approaches rely on statistical models in order to approximate observed data, ranging from PCA to the recently introduced Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). In order for these models to serve their purpose, they have to represent the observations as faithfully as possible. Such property can be evaluated in a purely statistical sense by testing whether the probability distribution of the model is as close as possible to the empirical distribution of the data (taking into account that such procedure should be properly designed to avoid overfitting). In contrast, enforcing the model to be causal goes beyond this statistical criterion by imposing that the fitted model should to some extent capture the structure of the true data generating process. Concepts pertaining to causality are well formulated using Structural Equation Models (SEMs), which describe the relationship between different variables (observed or hidden) as a set of structural equations, each of them taking the form [2]

$$v_k \coloneqq f_k(v_1, \cdots, v_n).$$

Such equations represent more than algebraic dependencies between the variables, as indicated by the asymmetry of the ":=" symbol, and suggests that the left-hand side variable is generated based on the right-hand side variables. Broadly construed, it means that this relation would still hold if an external agent were to intervene on right-hand side variables (the so called do-operator), and that we can formulate counterfactuals : "what would have happened if one right-hand side variable had been different" (see Pearl (2000) for an overview). As a consequence, a properly inferred causal generative model based on SEMs offers more robustness to changes in the environment than purely statistical models. Such property is exploited in relation to transfer

[2] each right-hand-side variables may refer to either endogenous variables (i.e. a factor influenced by other variables in the model) or exogenous variables (determined by factors outside of the system)

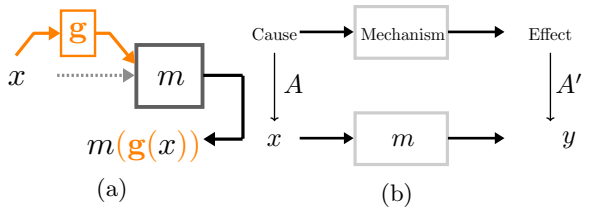


Figure 2: (a) Principle of the group theoretic framework: a generic transformation g is introduced between the cause x and the mechanism m. (b) Introduction of the concept of attribute to describe a structural causal model.

leaning, addressing covariate-shift or changes in the input distribution (Zhang et al., 2013, 2015; Peters et al., 2016; Bareinboim and Pearl, 2016; Rojas-Carulla et al., 2015), and makes causal generative models highly relevant in machine learning.

In section 2, we have shown how to virtually probe SEMs by a counterfactual reasoning which could be stated as: "what would happen if we were to apply a generic transformation to a given variable or mechanism of the SCM". This virtual intervention is represented in Fig. 2a for a SCM with a single cause and mechanism. The applied transformation is generic in the sense that it is sampled at random from a large set of transformations that turns a variable/mechanism into another one that is as likely to occur in a similar scenario. The outcome of this virtual intervention is tested by quantifying whether the counterfactual outcome is qualitatively different from the observed outcome for *most* generic transformations.

In our framework, the set of generic transformations is chosen to be a (compact) group. While readers can refer to appendix A for the relevant definitions regarding group theory, they may just assume the compact group is a set of invertible transformations applied to causes and equipped with a "uniform" probability measure. The choice of this particular structure is motivated by the fact that group actions combine well with a general structural equation framework.

Several points are worth mentioning about the generality of our setting. First causes and effects do not need to be objects of the same kind, and the cause-effect relation may be deterministic (as in 2) or probabilistic. In addition, we do not assume invertibility of this relation. Finally, although confounding can in principle be addressed in ICM-based frameworks (e.g. see Zscheischler et al. (2011)), it will not be investigated in this contribution.

3.2 Formal definition

Each SEM variable is characterized by a quantity that we call *attribute*. The covariance matrix is an example of an attribute for a multivariate random variable[3]. Given an effect generated by a cause through a mechanism as described in Fig. 2b, we measure attributes of cause and effect using functions A and A' with codomain $\mathcal{A}$ and $\mathcal{A}'$ respectively. To allow a less formal presentation, we will abusively consider the mechanism m as acting directly on the attribute space $\mathcal{A}$, and x and y will indicate indistinctly the cause and effect or their attribute.

Applying the ICM framework requires assessing genericity of the relationship between input and mechanism quantitatively. For that we define two objects: (1) the *generic group* $\mathcal{G}$ is a compact topological group that acts on $\mathcal{A}$, thus equipped with a unique Haar probability measure $\mu_{\mathcal{G}}$ (see appendix A), (2) the *contrast*[4] C is a real valued function with domain $\mathcal{A}'$. The contrast and generic group introduced in such a way allow to compute the expected value when randomly "breaking" the cause-mechanism relationship using generic transformations according to the following definition.

Definition 1 *Given a contrast C, the Expected Generic Contrast (EGC) of a cause mechanism pair (x, m) is defined as:*

$$\langle C \rangle_{m,x} = \mathbb{E}_{g\sim\mu_{\mathcal{G}}} C(mgx)\,. \tag{1}$$

We say that the relation between m and x is $\mathcal{G}$-generic under C, whenever

$$C(mx) = \langle C \rangle_{m,x} \tag{2}$$

holds approximately.

We call eq. (2) the *genericity equation*. Note that this equation is used to express an idealized ICM postulate (hence the term "holds approximately") that is not meant to be satisfied exactly in practice but justified by assuming, in appropriate contexts, that the value of $C(mx)$ concentrates around its mean (see (Janzing et al., 2010) for an example). Genericity can be formulated more rigorously as a statistical test, assessing the null hypothesis: "*is $C(mx)$ likely to be sampled from the generic distribution $D_{m,x}$ generated by $C(mgx)$, $g \sim \mu_{\mathcal{G}}$?*" (Zscheischler et al., 2011).

[3] typically the attribute of a random variable will be a function of its probability distribution

[4] the term *contrast* refers to the field of Independent Component Analysis, where such function is used as a proxy to quantify independence

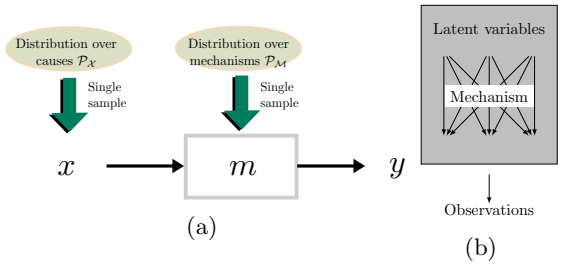


Figure 3: (a) Generative model including distributions over causes and mechanisms. (b) Causal structure of a latent generative model.

3.3 Invariant generative models

There is an interesting probabilistic interpretation of the concept of genericity. If we are given a generative model such that the cause x is a single sample drawn from a meta-distribution[5] $\mathcal{P}_\mathcal{X}$ (see Fig. 3a). To estimate genericity irrespective of the possible values of x, we consider the *genericity ratio* $C(mx)/\langle C\rangle_{m,x}$: this quantity should be close to one with high probability in order to satisfy ICM assumptions. Assume $\mathcal{P}_\mathcal{X}$ is a $\mathcal{G}$-invariant distribution, under mild assumptions (Wijsman, 1967) x can be parametrized as $x = g\tilde{x}$ where g is a sample from a $\mu_\mathcal{G}$-distributed variable G, and $\tilde{x}$ is a sample from anther RV independent of G. Then (see appendix B for details)

$$\mathbb{E}_x\left[\frac{C(mx)}{\langle C\rangle_{m,x}}\right] = \mathbb{E}_{\tilde{x}}\mathbb{E}_g\left[\frac{C(mg\tilde{x})}{\langle C\rangle_{m,g\tilde{x}}}\right] = 1 \qquad (3)$$

This tells us that the postulate of genericity is valid at least "on average" for the generative model. On the contrary, if this average would be different from 1 as it may happen for a non-invariant $\mathcal{P}_\mathcal{X}$, the postulate is unlikely be valid for individual examples. As represented on Fig. 3a, the same reasoning can be applied when sampling the mechanism from an invariant distribution.

It is noteworthy that, although meta-distributions could be considered as priors, we do not need to explicitly formulate this prior. This is important as most parameters live in unbounded domains, for which non-informative proper priors do not exist, and Bayesian techniques would be sensitive to the choice of such prior, in contrast to our approach.

4 REINTERPRETING PAIRWISE CAUSAL INFERENCE

We show in this section that the group invariance framework encompasses previous causal inference methods that have been proposed in the literature to solve the pairwise case: given two observables X and Y, can we decide between the alternatives "X causes Y" or "Y causes X"?

4.1 The Trace Method

We consider the case of X and Y n- and l-dimensional RVs, respectively, causally related by the noisy linear structural equation

$$Y := MX + E\,, \qquad (4)$$

where M is an $l \times n$ structure matrix and E is a multivariate additive noise term independent of X.

If we take the output covariance Σ_Y as attribute, the normalized trace $\tau_l(\Sigma_Y) = \mathrm{tr}(\Sigma_Y)/l$ as a contrast, and use generic matrices U distributed according to the Haar measure over the group $SO(n)$: $\mu_{SO(n)}$, the EGC writes

$$\langle C\rangle_{M,X} = \mathbb{E}_{U\sim\mu_{SO(n)}}\tau_l(MU\Sigma_X U^T M^T + \Sigma_E)\,.$$

This quantity can be evaluated using the following result (see appendix B for a proof).

Proposition 2 *Let U be a random matrix drawn from $SO(n)$ according to $\mu_{SO(n)}$ and let A and B be two symmetric matrices in $M_{n,n}(\mathbb{R}) = \mathbb{R}^{n\times n}$. Then*

$$\mathbb{E}_{U\sim\mu_{SO(n)}}\,\mathrm{tr}\left(U^T AUB\right) = \tfrac{1}{n}\,\mathrm{tr}(A)\,\mathrm{tr}(B). \qquad (5)$$

This leads to $\langle C\rangle_{M,X} = \tau_n(\Sigma_X)\tau_l(MM^T) + \tau_l(\Sigma_E)$, where Σ_E denotes the noise covariance. Then the genericity equation $\langle C\rangle_{M,X} = C(\Sigma_Y)$ writes

$$\tau_l(M\Sigma_X M^T) = \tau_n(\Sigma_X)\tau_l(MM^T)\,, \qquad (6)$$

which is exactly the *Trace Condition* postulate used in the Trace Method (Janzing et al., 2010).

4.2 Automorphisms on the unit interval

Another example of ICM based causal inference addresses the case where the cause, a random variable X on the unit interval, is mapped to the effect by an invertible C^1 function m. Using the density of Y (p_Y) as its attribute, the differential entropy $H(Y)$ as a contrast, and modulo 1 translations of the unit interval $(g_\tau)_\tau$ –parametrized by a shift variable $\tau \in [0, 1]$– as the generic group (the associated Haar measure being the Lebesgue measure), the EGC writes

$$\langle C\rangle_{m,X} \quad = \quad \int_0^1 H(m\circ g_\tau(X))d\tau$$

[5] meta-distributions have some similarities with the approach of Lopez-Paz et al. (2015)

This easily leads to

$$\langle C \rangle_{m,X} = H(X) + \int_0^1 \log \left| \frac{dm}{dx}(x) \right| dx \qquad (7)$$

and the genericity equation $\langle C \rangle_{m,X} = H(Y)$ can be rewritten

$$\int_0^1 \log \left| \frac{dm}{dx}(x) \right| p_X(x) dx = \int_0^1 \log \left| \frac{dm}{dx}(x) \right| dx \,, \qquad (8)$$

which corresponds exactly to the orthogonality postulate exploited in Information Geometric Causal Inference (IGCI) (Daniusis et al., 2010; Janzing et al., 2012).

4.3 Linear Non-Gaussian Additive Noise Models (LiNGAM)

While we showed in previous sections that several ICM-based approaches fit in our group theoretic framework, it can also be related to other causal inference methods such as LiNGAM (Shimizu et al., 2006), which relies on the assumption that the additive noise of a causal mechanism is independent from its input. Assume the cause X is a real random variable with zero mean, and Y is generated from X through a linear additive noise model, i.e.,

$$X \mapsto Y := \alpha X + \epsilon \,, \qquad (9)$$

where $\alpha \in \mathbb{R}^*$ and ϵ is a zero mean i.i.d noise random variable. Let $\mathbf{x}$, $\boldsymbol{\epsilon}$ and $\mathbf{y}$ be the corresponding N-tuples of i.i.d. samples drawn from the joint distribution $P(X, \epsilon, Y)$, they follow the structural equation $\mathbf{y} := m_{\alpha,\boldsymbol{\epsilon}}\mathbf{x}$ with deterministic mechanism

$$m_{\alpha,\boldsymbol{\epsilon}} : \mathbf{x} \mapsto \alpha\mathbf{x} + \boldsymbol{\epsilon} \,.$$

Using the empirical estimate of the third order cumulant of a centered RV, $C(\mathbf{y}) = \overline{\mathbf{y}^3}$, as a contrast (where the bar indicates the empirical average over the samples), and the symmetric group $S(N)$ of all permutations of N-tuples as generic group, then the EGC writes

$$\langle C \rangle_{\mathbf{x}, m_{\alpha,\boldsymbol{\epsilon}}} = \mathbb{E}_{g \sim \mu_{S(N)}} \overline{(m_{\alpha,\boldsymbol{\epsilon}} \circ g(\mathbf{x}))^3} . \qquad (10)$$

Considered as a function of both g and X, $g(\mathbf{x})$ behaves like a N-tuple $\mathbf{x}'$ sampled from X', an independent copy of X. Using the law of large numbers, for large N we thus get approximately

$$\langle C \rangle_{\mathbf{x}, m_{\alpha,\epsilon}} \approx \mathbb{E}_{X'} \mathbb{E}_{\epsilon} (\alpha X' + \boldsymbol{\epsilon})^3 \,, \qquad (11)$$

where X' and ϵ are independent. Because the cumulant of a sum of independent variables is the sum of individual cumulants, we get

$$\langle C \rangle_{\mathbf{x}, m_{\alpha,\epsilon}} \approx \alpha^3 C(\mathbf{x}) + C(\boldsymbol{\epsilon}) \,, \qquad (12)$$

and the corresponding genericity equation

$$C(\mathbf{y}) \approx \alpha^3 C(\mathbf{x}) + C(\boldsymbol{\epsilon}) \,. \qquad (13)$$

This relates to the cumulant-based approach of Hyvärinen and Smith (2013) to infer the causal direction in LiNGAM models. As shown in appendix B, the use of the cumulant relies on the fact that the genericity equation will be valid for the forward model (because the additive noise is independent from X), while it is violated in the backward direction (additive noise becomes dependent), provided X is skewed ($\mathbb{E}X^3 \neq 0$). Overall, this example illustrates the idea that LiNGAM models can be framed as a group theoretic approach in the limit of large samples, because stating statistical independence between the additive noise and the cause amounts to have the model insensitive to random permutations of the samples of the cause.

5 UNSUPERVISED LEARNING

5.1 Causal generative models

Classically generative models aim at modeling the probability distribution of observations. However, we often expect such model to capture information about the true generative process, in order to better understand its underlying mechanisms. Take for example the case of clustering using Gaussian mixture models: when experimental scientists cluster a dataset, they expect that the resulting clusters reflect a reliable structure that will be robust to moderate changes of experimental parameters, such that the results can be replicated. Such required property, although not explicitly stated, puts the clustering task in a causal inference perspective. Like for any causal inference problem, finding plausible causal generative models will require assumptions on the data generating mechanism. We can thus try to exploit the ICM postulate to learn the structure of generative models from a causal perspective. As suggested in (Schölkopf et al., 2012), many real world datasets have an intuitive underlying causal structure that we may exploit to improve learning algorithms. For instance, in a character recognition datasets such as MNIST, the character that a human intents to write is a cause for the observed handwritten character image.

In this section, we assume the setting of Fig. 3b in which observations are generated by composing latent variables and partially unknown mechanisms, and we formulate ICM postulates for such systems.

5.2 Clustering

Consider the following classical Gaussian Mixture Model of the observed multivariate random vector $\boldsymbol{X}$ using latent variable Z.

$$Z \sim \text{Mult}(\pi_1, \pi_2, \cdots, \pi_K), \quad (14)$$
$$\boldsymbol{X}|\{z=k\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (15)$$

where z indicates the cluster membership of one observation, and $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are means and covariances of the p-dimensional Gaussian distribution of each cluster.

5.2.1 Invariance hypothesis

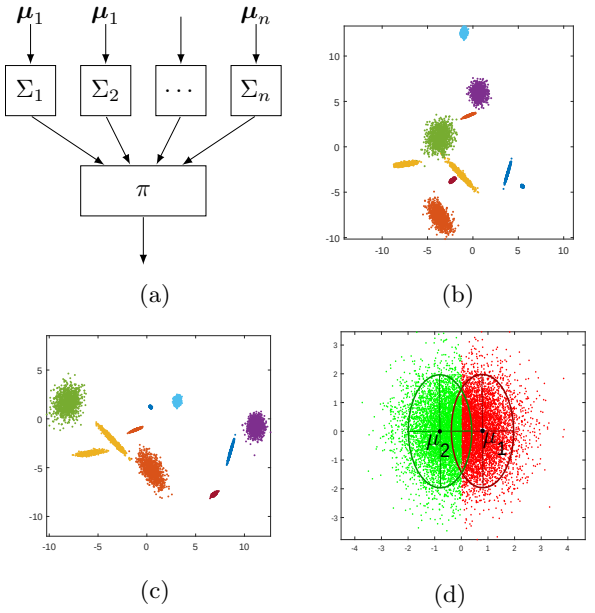


Figure 4: (a) Generative model for a mixture of Gaussians. (b) Cluster data generated with random parameters (projected on 2 components). (c) Data in (b) after a generic transformation. (d) Suspicious dependency between cluster means and covariances in a case of a mispecified number of clusters.

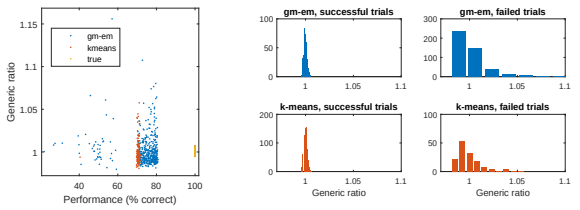
To get an insight of what form of genericity is relevant for such generative model, imagine the collected data reflects the phenotype of different subspecies of plants (similarly to the popular Iris dataset). Each cluster mean $\boldsymbol{\mu}_k$ reflects the average characteristics of the subspecies k, while the covariance matrices $\boldsymbol{\Sigma}_k$ express the variations of these characteristics across the subpopulation. If we assume that each subspecies has emerged independently (say on different continents) and that they never interacted with each other (no competition for resources), we suggest that the variability within each subspecies should be unrelated to the variations across species. As a consequence, randomizing the properties of $\boldsymbol{\mu}_k$'s while keeping $\boldsymbol{\Sigma}_k$'s constant may lead to a model as likely to have been generated by Nature as the observed dataset sampled from $\boldsymbol{X}$.

This intuition leads to formulate a causal generative model as described in Fig. 4a, where a noise vector of covariance Σ_k is added to the mean vector of each cluster k (constituting the causes), before clusters are mixed according to π in order to generate the observed effect $\boldsymbol{X}$. In order to randomize the configuration of $\boldsymbol{\mu}_k$'s with respect to the noise distributions, we apply a random orthogonal transformation to them by left multiplication.

To ease analysis, we represent the model by the sum of an (intra-cluster) variability component $\boldsymbol{V}$ and a cluster mean component $\boldsymbol{M}$. Then

$$\boldsymbol{X} = \boldsymbol{V} + \boldsymbol{M}, \begin{cases} \boldsymbol{V}|\{z=k\} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_k), \\ \boldsymbol{M}|\{z=k\} = \boldsymbol{\mu}_k. \end{cases}$$

We then choose $O(p)$ as generic group and p-dimensional orthogonal matrices from this group act on the mean vectors by left multiplication to the variable $\boldsymbol{M}$, before $\boldsymbol{V}$ is added. Application of one generic transformation results in clusters with the same intra-cluster variability as the original data, but whose locations in the feature space have been randomized, as illustrated on Figs. 4b and 4c with an 5-dimensional feature space and 10 clusters. In this illustration, the structure of the observations does not seem to be affected by the transformation, suggesting the original data is "typical" in some sense. This makes sense as mean and covariance parameters have been drawn independently at random. However, there are simple pathological examples where a clustering algorithm can fail to capture the underlying structure of the data and generate an atypical dependency between means and covariances. Assume for example that, focusing on one single Gaussian cluster, a clustering algorithm fails to identify a single cluster and instead cuts it in two clusters. This situation illustrated on Fig. 4d shows an interesting dependency between the centroids of the two clusters and their within cluster empirical covariance matrices: the difference between centroids is oriented in the direction (eigenspace) of smallest variance. We postulate that such suspicious dependencies may appear when the clustering algorithm fails to capture the causal structure of the data.



(a) (b)

Figure 5: (a) Clustering performance of 'gm-em' and 'kmeans' algorithms against genericity ratio for 800 simulations; yellow ground truth points on the very right hand-side indicate how the ground truth genericity ratio concentrates (note they hide points of successful trials for both algorithms). (b) Distribution of the genericity ratios for both algorithms in case of successful trials (performance exceeds 99%) and failed trials (performance below 99%).

5.2.2 Contrast

To detect such suspicious dependencies in the inferred generative model using the group theoretic approach, we propose the following 4th order tensor contrast

$$C(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{X}} \operatorname{tr}\left[\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{X}^\top\right] .$$

Using this contrast is justified by the following.

Proposition 3 *Let $\boldsymbol{X}$ be a centered p-dimensional random variable, and $\langle C\rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ the generic contrast obtained by random orthogonal transformation applied to cluster means, then*

$$C(\boldsymbol{X}) - \langle C\rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}} = 4\sum_k \pi_k \Big(\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k \boldsymbol{\mu}_k - \|\boldsymbol{\mu}_k\|^2 \tfrac{\operatorname{tr}[\boldsymbol{\Sigma}_k]}{p}\Big) .$$

Indeed, this result shows that differences between the contrast of the data and the EGC quantify the alignment between the cluster means $\boldsymbol{\mu}_k$ and the principal axes (eigenvectors) of the covariance matrices $\boldsymbol{\Sigma}_k$. We will use the resulting genericity ratio to detect suspicious dependencies in the solution of clustering algorithms, such that of Fig. 4d.

5.2.3 Experiments

We test this approach to detect bad clustering of a simulated dataset. We generate 5 random clusters in a 20 dimensional space. The cluster means are drawn at random from an isotropic Gaussian distribution with standard deviation 2. Cluster covariances are generated with random axes (with isotropic distribution) and eigenvalues. We test the performance of two clustering algorithm: K-means ('kmeans') and the Expectation Minimization algorithm based on the simulated Gaussian Mixture model ('gm-em'). The scatter plot shown on Fig. 5a suggests that the genericity ratios are broadly distributed on the interval $[0.98, 1.1]$ when the algorithms do not reach a good estimation of the original clusters. Comparison of the distributions of the genericity ratio in case of success and failure of the clustering shown on Fig. 5b shows a much more concentrated distribution when the clusters are correctly retrieved. This suggests that a genericity ratio far from one indeed witnesses the failure of the algorithm to cluster the data properly and could be exploited to improve the performance of clustering algorithms.

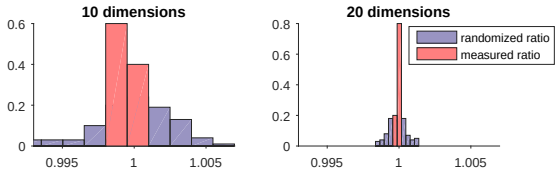


Figure 6: Normalized histograms of the genericity ratio for the CIFAR-10 dataset. Left panel: measured ratio (M= 0.9994, SD= 0.0002), randomized ratio (M= 1.0001, SD= 0.0027). Right panel: measured ratio (M= 0.99991, SD= 6.3 e-5), randomized ratio (M= .99989, SD= 6.4 e-4).

We further addressed the question whether real datasets satisfy the ICM principle exploited in our framework. For that we use the CIFAR-10 dataset containing 64x64 color pictures of 10 different types of objects (Krizhevsky, 2009). In order to reduce the dimension and eliminate correlation between neighboring pixels, we preprocessed the images by extracting the 10 or 20 first principal components of the data, using singular value decomposition. We computed the generic ratio for 5 non-overlaping batches of 10000 images preprocessed separately and compared it with the generic distribution of this ratio (denoted randomized ratio) estimated by applying random orthogonal transformations the cluster mean vector. The results shown in Fig. 6 suggest that the CIFAR datasets satisfies the ICM assumption. Result for the same approach applied to the MNIST dataset lead to the same conclusion (Fig. 7 in appendix).

Acknowledgements

MB would like to thank Marek Kaluba for helpful discussions and comments. MB acknowledges funding from the Max Planck ETH Center for Learning Systems.

References

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

P. Daniusis, D. Janzing, K. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Uncertainty in Artificial Intelligence*, 2010.

M. L. Eaton. Group invariance applications in statistics. volume 1. Institute of Mathematical Statistics, 1989.

W. T. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368 (6471):542, 1994.

K. Fukumizu, A. Gretton, B. Schölkopf, and B.K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, pages 473–480, 2009.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

A. Hyvärinen and S.M. Smith. Pairwise likelihood rations for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152, 2013.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *Information Theory, IEEE Transactions on*, 56(10):5168–5194, 2010.

D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning*, 2010.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182–183:1–31, 2012.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, pages 1–23, 2012.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *arXiv preprint arXiv:1507.05333*, 2015.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial Intelligence and Statistics*, 2015.

N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning*, 2015.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

W. K. Tung. Group theory in physics. 1985.

R A Wijsman. Cross-sections of orbits and their application to densities of maximal invariants. 1967.

R.A. Wijsman. Invariant measures on groups and their use in statistics. IMS, 1990.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. 2013.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. 2015.

J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *Uncertainty in Artificial Intelligence*, 2011.