Causality

Stefan Bauer and Bernhard Schölkopf

Roadmap

- Background and motivation
- Structural causal models (SCMs)
- Independent mechanisms and disentangled factorizations
- Do-calculus
- Confounding
- Causal discovery: 2-variable case
- Causal ML: 2-variable case
- Algorithmic SCMs and the arrow of time
- Confounder modelling in exoplanet discovery

- Causal discovery: n-variable case (d-separation, constraint-based methods)
- Time series: Granger causality and other methods
- Causal ML: deep learning (independent mechanisms, disentanglement, …)

Notation

- A, B event
- X, Y, Z random variable
- x value of a random variable
- Pr probability measure
- P_X probability distribution of X
- p density
- p_X or $p(X)$ density of P_X
- $p(x)$ density of P_X evaluated at x
- from the definition of conditional probability, we have the "chain rule"

$$\begin{aligned} p(X_n, \ldots, X_1) &= p(X_n | X_{n-1}, \ldots, X_1) p(X_{n-1}, \ldots, X_1) \\ &= p(X_n | X_{n-1}, \ldots, X_1) p(X_{n-1} | X_{n-2}, \ldots, X_1) \cdot \cdots \cdot p(X_2 | X_1) p(X_1) \end{aligned}$$

- always assume the existence of a joint density, w.r.t. a product measure

Independence

Two events A and B are called *independent* if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

$A_1, \ldots, A_n$ are called *independent* if for every subset $S \subset \{1, \ldots, n\}$ we have

$$\Pr\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \Pr(A_i).$$

Note: for $n \geq 3$, *pairwise independence* $\Pr(A_i \cap A_j) = \Pr(A_i) \cdot \Pr(A_j)$ for all i, j does not imply *independence*.

Independence of random variables

Two real-valued random variables X and Y are called *independent*,

$$X \perp\!\!\!\perp Y,$$

if for every $a, b \in \mathbb{R}$, the events $\{X \leq a\}$ and $\{Y \leq b\}$ are independent.

Equivalently, in terms of densities: for all x, y,

$$p(x, y) = p(x)p(y)$$

Note:
If $X \perp\!\!\!\perp Y$, then $E[XY] = E[X]E[Y]$, and $\mathrm{cov}[X, Y] = E[XY] - E[X]E[Y] = 0$.

The converse is not true: $\mathrm{cov}[X, Y] = 0 \not\Rightarrow X \perp\!\!\!\perp Y$.

However, we have, for large $\mathcal{F}$: $(\forall f, g \in \mathcal{F} : \mathrm{cov}[f(X), g(Y)] = 0) \Rightarrow X \perp\!\!\!\perp Y$

Conditional Independence of random variables

Two real-valued random variables X and Y are called *conditionally independent* given Z,

$$(X \perp\!\!\!\perp Y)\,|Z \quad \text{or} \quad X \perp\!\!\!\perp Y\,|Z \quad \text{or} \quad (X \perp\!\!\!\perp Y\,|Z)_p$$

if

$$p(x, y|z) = p(x|z)p(y|z)$$

for all x, y, and for all z s.t. $p(z) > 0$.

Note: it is possible to find X, Y which are conditionally independent (given Z) but unconditionally dependent, and vice versa.

What is cause and what is effect?

$$p(a,t) = p(a|t)\ p(t) \quad T \to A$$
$$= p(t|a)\ p(a) \quad A \to T$$

MAX-PLANCK-GESELLSCHAFT

- **intervention** on a: raise the city, find that t changes
- hypothetical intervention on a: still expect that t changes, since we can think of a physical mechanism $p(t|a)$ that is **independent** of $p(a)$
- we expect that $p(t|a)$ is **invariant** across, say, different countries in a similar climate zone

MAX-PLANCK-GESELLSCHAFT

Peters, Janzing, Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, 2017
https://www.dropbox.com/s/gkmsow492w3oolt/11283.pdf?dl=1

Principle 2.1 (Independent Mechanisms) *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.

- a "structural" relation not only explains the observed data, it captures a structure connecting the variables; related to autonomy and invariance (Haavelmo 1943, Frisch 1948, ...)
- an equation system becomes **structural** by virtue of invariance to a domain of modifications (Harwich, 1962)
- "Simon's invariance criterion:" the true causal order is the one that is invariant under the right sort of intervention (Simon, 1953; Hoover, 2008)
- each parent-child relationship in the network represents a stable and autonomous physical mechanism (Pearl, 2009)
- formalised using algorithmic information theory (Janzing & Schölkopf, 2010)

Reichenbach's Common Cause Principle

(i) if X and Y are dependent, then there exists Z *causally* influencing both;

(ii) Z screens X and Y from each other (given Z, X und Y become independent)

by permission of the University of Pittsburgh. All rights reserved.

Definition of a Structural Causal Model *(Pearl et al.)*

- directed acyclic graph G with vertices $X_1, \ldots, X_n$
  (following arrows does not lead to loops)
- Semantics: vertices = observables, arrows = direct causation
- $X_i := f_i(\mathrm{PA}_i, U_i)$, with independent RVs $U_1, \ldots, U_n$ that possess a joint density
- U_i stands for "unexplained" (alternatively "noise" or "exogenous variable")
- this is also called a *(nonlinear) structural equation model*

MAX-PLANCK-GESELLSCHAFT

Reichenbach's Principle and causal sufficiency

- Independence of noises is a form of "causal sufficiency:" if the noises were dependent, then Reichenbach's principle would tell us the causal graph is incomplete

- The SCM model satisfies Reichenbach's principle:

  1. functions of independent variables are independent, hence dependence can only arise in two vertices that depend (partly) on the same noise term(s).

  2. if we condition on these noise terms, the variables become independent

MAX-PLANCK-GESELLSCHAFT

Entailed distribution

- $X_i := f_i(\mathrm{PA}_i, U_i)$,
  with independent $U_1, \ldots, U_n$.
- Recursively substitute the parent equations to get $X_i = g_i(U_1, \ldots, U_n)$,
  with independent $U_1, \ldots, U_n$.
- Each X_i is thus a RV and we get a joint distribution of $X_1, \ldots, X_n$,
  called the *observational distribution.*
- The distribution and the DAG form a *directed graphical model* and any directed graphical model can be written as a functional causal model.

MAX-PLANCK-GESELLSCHAFT

Entailed distribution

- A structural causal model entails a joint distribution $p(X_1, \ldots, X_n)$.

  Questions:

  (1) What can we say about it?

  (2) Can we recover G from p?

MAX-PLANCK-GESELLSCHAFT

Markov conditions *(Lauritzen 1996, Pearl 2000)*

Theorem: the following are equivalent:

- Existence of a structural causal model
- Local Causal Markov condition: X_i *statistically independent* of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: "d-separation" (characterizes the set of independences implied by local Markov condition — see below)
- Factorization $p(X_1, \ldots, X_n) = \prod_i p\left(X_i \mid \mathrm{PA}_i\right)$

(subject to technical conditions)

$p\left(X_i \mid \mathrm{PA}_i\right)$ is called a *causal conditional* or *causal Markov kernel*.
It corresponds to the structural "equation" $X_i := f_i(\mathrm{PA}_i, U_i)$.

Not every conditional is causal — only those that condition on the parents in our DAG.

Graphical Causal Inference *(Spirtes, Glymour, Scheines, Pearl, ...)*

Question: How can we recover G from a single p (e.g., from the observational distribution)?
Answer: by conditional independence testing, infer a class containing the correct G
(i.e., track how the noise information spreads).

Problems:

- Markov condition states $(X \perp\!\!\!\perp Y \,|Z)_G \quad \Rightarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$, but
  we need "faithfulness" $(X \perp\!\!\!\perp Y \,|Z)_G \quad \Leftarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$ *(Sprites, Glamour, Scheines 2001)*
  Hard to justify for finite data *(Uhler, Raskutti, Bühlmann, Yu, 2013)*.
- if the f_i are complex, then conditional independence testing based on finite samples becomes arbitrarily hard
- for **two variables** only, there are no conditional independences

Interventions and shifts

- **Definition.** Replacing $X_i := f_i(\mathrm{PA}_i, U_i)$ with another assignment (e.g., $X_i := const.$) is called an *intervention* on X_i.
- The entailed distribution is called the *interventional distribution*.
- This contains as special cases: domain shift distribution and covariate shift distribution (see below).
- A general intervention corresponds to changing some *causal conditionals* $p(X_i|\mathrm{PA}_i)$

Independent mechanisms and the disentangled factorization

- the factorization
$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$
exists since the noise terms are **independent**.
- interventions are feasible if the mechanisms (i.e., causal conditionals) are **independent**: changing one $p\left(X_i \mid \mathrm{PA}_i\right)$ does not change the conditionals $p\left(X_j \mid \mathrm{PA}_j\right)$ for $j \neq i$, they remain **invariant**.
- if both is satisfied, we call this a **disentangled (causal) factorization**.
- **causal shift hypothesis**: a change in a distribution always comes from a sparse change in mechanisms (i.e., structural assignment/function and/or noise variable)
- can be used for causal discovery and causal machine learning (see below)
- **entangled (non-causal) factorizations**: e.g.,
$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p(X_i \mid X_{i+1}, \ldots, X_n).$$
Here, changes will **not** be local.

Cf. *independence of mechanisms* *(Janzing & Schölkopf, 2010)*, *(Schölkopf et al., 2012)*, *(Peters et al., 2017)*, *autonomy*, *(structural) invariance*, *separability*, *exogeneity*, *stability*, *modularity* *(Aldrich, 1989; Pearl, 2009)*

The causal generative process is composed of autonomous modules that do not inform or influence each other.

Counterfactuals

- David Hume (1711–76): *"... we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed."*

- Jerzy Neyman (1923): consider m plots of land and ν varieties of crop.

  Denote U_{ij} the crop yield that *would be* observed if variety $i = 1, \ldots, \nu$ were planted in plot $j = 1, \ldots, m$

  For each plot j, we can only experimentally determine *one* U_{ij} in each growing season.

  The others are called "counterfactuals".

- this leads to the view of causal inference as a missing data problem — the "potential outcomes" framework (Rubin, 1974)

Does it make sense to talk about causality without mentioning time?

Does it make sense to talk about statistics without mentioning time?

MAX-PLANCK-GESELLSCHAFT

Causality in differential equations

Consider the set of differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d,$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$.

Picard–Lindelöf: locally, if f is Lipschitz, there exists a unique solution $\mathbf{x}(t)$

$\Longrightarrow$ the immediate future of $\mathbf{x}$ is implied by its past

Using dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$:

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t)).$$

This tells us which entries of $\mathbf{x}(t)$ cause the future of others $\mathbf{x}(t + dt)$, i.e., the causal structure.

https://arxiv.org/abs/1911.10500

A Modeling Taxonomy

	statistical model	causal model	differential equation model
i.i.d. prediction, pattern recognition, *"generalization"*	y	y	y
Predict under shift & intervention, *"horizontal generalization"*	n	y	y
Provide physical insight, understand predictions	n	(y)	y
Think/Reason, *"act in an imagined space"* (K. Lorenz)	n	?	?
Learn from data	y	(y)	n

MAX-PLANCK-GESELLSCHAFT

From Ordinary Differential Equations to Structural Causal Models: the deterministic case

Joris M. Mooij
Institute for Computing and Information Sciences
Radboud University Nijmegen
The Netherlands

Dominik Janzing
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

[stat.OT] 30 Apr 2013

Abstract

We show how, and under which conditions, the equilibrium states of a first-order Ordinary Differential Equation (ODE) system can be described with a deterministic Structural Causal Model (SCM). Our exposition sheds more light on the concept of causality as expressed within the framework of Structural Causal Models, especially for cyclic models.

algorithms (starting from different assumptions) have been proposed for inferring cyclic causal models from observational data (Richardson, 1996; Lacerda et al., 2008; Schmidt and Murphy, 2009; Itani et al., 2010; Mooij et al., 2011).

The most straightforward extension to the cyclic case seems to be offered by the structural causal model framework. Indeed, the formalism stays intact when one simply drops the acyclicity constraint. However, the question then arises how to interpret cyclic structural equations. One option is to assume an under-

UAI 2013

See also *Rubenstein, Bongers, Mooij, Schölkopf, 2016*

Recap

- Causal structure formalised by DAG with RVs $X_1, \ldots, X_n$ as vertices
- **Causal Markov condition**: the density $p(x_1, \ldots, x_n)$ factories into

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j | pa_j),$$

  where pa_j denotes the values of the direct causes of X_j.
- Causal conditionals $p(x_j | pa_j)$ represent causal mechanisms

MAX-PLANCK-GESELLSCHAFT

Pearl's do calculus

- Motivation: goal of causality is to infer the effect of interventions

- distribution of Y given that X is set to x: $p(Y|do\,X = x)$ or $p(Y|do\,x)$
- don't confuse it with $p(Y|x)$
- can be computed from p and G

Difference between seeing and doing

$p(y|x)$

Probability an MLSS participant can get a NeurIPS paper accepted

$p(y \mid \text{do } x)$

Probability that anyone can get a NeurIPS paper accepted after being made to participate in MLSS

Computing $p(X_1, \ldots, X_n | do\, x_i)$

from $p(X_1, \ldots, X_n)$ and G

- Start with causal factorization

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j | PA_j)$$

- Replace $p(X_i | PA_i)$ with $\delta_{X_i x_i}$

$$p(X_1, \ldots, X_n | do\, x_i) := \prod_{j \neq i} p(X_j | PA_j) \delta_{X_i x_i}$$

Computing $p(X_k|do\ x_i)$

Sum over x_i to get

$$p(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n|do\ x_i) = \prod_{j \neq I} p(X_j|PA_j(x_i)).$$

- i.e.: for $j \neq i$, drop $p(X_i|PA_i)$ and substitute x_i for X_i
- obtain $p(X_k|do\ x_i)$ by marginalisation

MAX-PLANCK-GESELLSCHAFT

Examples for $p(.|do\,x) = p(.|x)$

Examples for $p(.|do\,x) \neq p(.|x)$

- $p(Y|do\,x) = P(Y) \neq P(Y|x)$

- $p(Y|do\,x) = P(Y) \neq P(Y|x)$

Controlling for confounding / adjustment formula

Y depends on X due to $X \to Y$ *and* the confounder Z

- Causal factorization

$$p(X, Y, Z) = p(Z)\, p(X|Z)\, p(Y|X, Z)$$

- Replace $p(X|Z)$ with δ_{Xx} and integrate out X:

$$\begin{aligned} p(X, Y, Z|do\ x) &= p(Z)\, \delta_{Xx}\, p(Y|X, Z) \\ p(Y, Z|do\ x) &= p(Z)\, p(Y|x, Z) \end{aligned}$$

- marginalize over Z to get the "adjustment formula"

$$p(Y|do\ x) = \sum_z p(z)\, p(Y|x, z)$$

This is different from $p(Y|x)$ (Simpson's paradox).

Simpson's paradox in Covid-19 case fatality rates

(v. Kügelgen, Gresele, https://arxiv.org/abs/2005.07180)

Case fatality rates (CFRs) in Italy are *lower* for each age group, but *higher* overall.

Simpson's paradox: opposite trends in grouped and aggregated data.

Here, it stems from a difference in case demographic:

Thanks to Elias Bareinboim

Coarse-grained causal graph

Data generating process:

- Randomly pick a **country C**
- Given **C**, sample a *positively-tested* patient with **age group A**
- Given **C** and **A**, sample **medical outcome, or mortality, M** (deceased at time of reporting?)

Assumptions & meaning of directed arrows:

- **C** → **A**: general population demographic, inter-generational mixing, age-specific social-distancing, …
- **A** → **M**: age-related health condition & other comorbidities.
- **C** → **M**: number of ventilators & ICU beds, …

Mediation analysis

Only for linear models can **total causal effect (TCE)** be decomposed into direct effect (DE) and indirect effect (IE),

$$\textbf{TCE} = \text{DE} + \text{IE}$$

Due to interactions, DE and IE are *not uniquely defined in general*, but depend on the state of the mediator.

- **Natural Direct Effect (NDE):** case demographic kept as in China while CFRs per age group changed to those in Italy.
- **Natural Indirect Effect (NIE):** CFRs per age group kept as in China, while case demographic changed to that in Italy.

INTERVAL

MAX-PLANCK-GESELLSCHAFT

Recap: Structural Causal Model

- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, with jointly independent $\text{Noise}_1, \ldots, \text{Noise}_n$.

- entails $p(X_1, \ldots, X_n)$ with particular conditional independence structure

  Assuming Markov condition + faithfulness we can recover an equivalence class containing the correct graph using conditional independence testing.

  Problems:

  1. does not work for graphs with only 2 vertices (even with infinite data)

  2. if we don't have infinite data, conditional independence testing can be arbitrarily hard

  Hypothesis:

  Both issues can be resolved by making assumptions on function classes.

Friedrich Nietzsche's

TWILIGHT OF THE IDOLS

or How to Philosophize with a Hammer

Götzen-Dämmerung

oder

Wie man mit dem Hammer philosophirt.

Von

Friedrich Nietzsche.

LEIPZIG.
Verlag von C. G. Naumann.
1889.

Translated, with commentary, by R.J. Hollingdale

[. . .]

The Four Great Errors

1

The error of mistaking cause for consequence. - There is no more dangerous error than that of *mistaking the consequence for the cause*: I call it reason's intrinsic form of corruption. Nonetheless, this error is among the most ancient and most recent

$$X \underset{\leftarrow}{\overset{\rightarrow}{?}} Y$$

Restricting the Structural Causal Model

$X \perp\!\!\!\perp N$

- consider the graph $X \to Y$
- general functional model

$$Y = f(X, N)$$

Note: if N can take d different values, it could switch randomly between mechanisms $f^1(X), \ldots, f^d(X)$

- additive noise model

$$Y = f(X) + N$$

Causal Inference with Additive Noise, 2-Variable Case

additive noise model (ANM):
$Y := f(X) + N_Y$, with $X \perp\!\!\!\perp N_Y$

Identifiability: when is there a backward model of the same form?

answer: generically, there is no model
$X = g(Y) + N_X$ with $Y \perp\!\!\!\perp N_X$

Hoyer et al.: Nonlinear causal discovery with additive noise models. *NIPS* 21, 2009
Peters et al: Causal Discovery with Continuous Additive Noise Models, JMLR 2014
Peters et al.: Detecting the Direction of Causal Time Series. *ICML* 2009

Intuition

- Assume noise of bounded range
- Additive noise model implies range of Y around f is constant
- For nonlinear f, range of X around backward function non-constant

Identifiability Result *(Hoyer, Janzing, Mooij, Peters, Schölkopf, 2008)*

Theorem 1 *Let the joint probability density of x and y be given by*

$$p(x,y) = p_n(y - f(x))p_x(x)\,, \tag{1}$$

where p_n, p_x are positive probability densities on $\mathbb{R}$. If there is a backward model

$$p(x,y) = p_{\tilde{n}}(x - g(y))p_y(y)\,, \tag{2}$$

then, denoting $\nu := \log p_n$ and $\xi := \log p_x$ and assuming sufficient differentiability, the triple (f, p_x, p_n) must satisfy the following differential equation for all x, y with $\nu''(y - f(x))f'(x) \neq 0$:

$$\xi''' = \xi''\left(-\frac{\nu'''f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu''f''f' + \nu'f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'}\,, \tag{3}$$

where we have skipped the arguments $y - f(x)$, x, and x for ν, ξ, and f and their derivatives, respectively. Moreover, if for a fixed pair (f, ν) there exists $y \in \mathbb{R}$ such that $\nu''(y - f(x))f'(x) \neq 0$ for all but a countable set of points $x \in \mathbb{R}$, the set of all p_x for which p has a backward model is contained in a 3-dimensional affine space.

Corollary 1 *Assume that $\nu''' = \xi''' = 0$ everywhere. If a backward model exists, then f is linear.*

Idea of the proof

If $p(x, y)$ admits an additive noise model

$$Y = f(X) + N_Y \quad \text{with } X \perp\!\!\!\perp N_Y$$

we have

$$p(x, y) = q(x) r(y - f(x)) \,.$$

It then satisfies the differential equation

$$\frac{\partial}{\partial x} \left(\frac{\partial^2 \log p(x, y) / \partial x^2}{\partial^2 \log p(x, y) / \partial x \partial y} \right) = 0 \,.$$

If it also holds with exchanging x and y, only specific cases remain.

Alternative View *(cf. Zhang & Hyvärinen, 2009)*

H differential entropy

I mutual information

$N_Y := Y - f(X), \quad N_X := X - g(Y)$ residual noises

Lemma: For arbitrary joint distribution of X, Y and functions f, $g : \mathbf{R} \to \mathbf{R}$, we have:

$$H(X,Y) = H(X) + H(N_Y) - I(N_Y : X) = H(Y) + H(N_X) - I(N_X : Y).$$

Note: $I(N_Y : 0) = 0$ iff there is an additive noise model from X to Y with function f, i.e.,

$$Y = f(X) + N_Y \quad \text{with } N_Y \perp\!\!\!\perp X.$$

Then

$$H(X) + H(N_Y) \leq H(Y) + H(N_X).$$

Hence, we can infer the causal direction by comparing sum of entropies

Causal Inference Method

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

- Compute a function f as non-linear regression of X on Y
- Compute the residual
$$N_Y := Y - f(X)$$
- check whether N_Y and X are statistically independent (uncorrelated is not enough)

Experiments

Relation between altitude (cause) and average temperature (effect) of places in Germany

Our independence tests detect strong dependence.
Hence the method prefers the correct direction

altitude → temperature

Side note on multivariate ANMs

For some DAG G with nodes $X_1, \ldots, X_n$ assume

$$X_j = f_j(PA_j) + N_j$$

where all N_j are independent

- then one can identify the DAG G (except for some rare cases)
- distinguishes even between Markov equivalent DAGs
- avoids conditional independence testing: if all residuals $X_j - f_j(PA_j)$ are independent, $P_{X_1,\ldots,X_n}$ satisfies the Markov condition w.r.t. G

(addresses two problems with the conditional-independence based approach)

Peters et al, UAI 2011

Independence-based Regression *(Mooij et al., 2009)*

- Problem: many regression methods assume a particular noise distribution; if this is incorrect, the residuals may become dependent
- Solution: minimize dependence of residuals rather than maximizing likelihood of data in regression objective
- Use RKHS distance between kernel mean embeddings/Hilbert-Schmidt-norm of cross-covariance operator between two RKHSes as a dependence measure

Mooij, Janzing, Peters, Schölkopf: Regression by dependence minimization and its application to causal inference. *ICML* 2009.
Yamada & Sugiyama: Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise. *AAAI* 2010.

MAX-PLANCK-GESELLSCHAFT

Kernel Independence Testing *(Gretton et al., 2007)*

k bounded p.d. kernel; P Borel probability measure

Define the *kernel mean map*

$$\mu : P \mapsto \mathbf{E}_{x \sim P}[k(x, .)].$$

Theorem: If k is universal, μ is injective.

Discussion: a measure can be represented as an element of the RKHS associated with k without loss of information.

Let's represent $p(X, Y)$ and $p(X)p(Y)$ — they will only map to the same element if they are equal, i.e., if $X \perp\!\!\!\perp Y$.

Kernel Independence Testing: HSIC

Corollary: $x \perp\!\!\!\perp y \iff \Delta := \|\mu(p_{xy}) - \mu(p_x \times p_y)\| = 0.$

- For $k((x,y),(x',y')) = k_x(x,x')k_y(y,y')$:

  Δ^2 = HS-norm of cross-covariance operator between the two RKHSes (HSIC, Gretton et al., 2005)

- empirical estimator $\frac{1}{n^2}\mathrm{tr}[K_x K_y]$ (ignoring centering)

- Why does this characterize independence: $x \perp\!\!\!\perp y$ iff

$$\sup_{f,g \in \text{RHKS unit balls}} \mathrm{cov}(f(x), g(y)) = 0$$

  (cf. Kernel ICA, *Bach & Jordan, 2002)*

Kernel Maps of Structural Causal Models

Assume $X \to Y$ with additive noise, i.e.

$$Y = f(X) + U \text{ with } X \perp\!\!\!\perp U.$$

$$0 = \|\mu(Y) - \mu\left(f(X) + U\right)\|^2 \approx \underbrace{\left\| \frac{1}{m} \sum_{i=1}^{m} \Phi(y_i) - \frac{1}{m^2} \sum_{i,j=1}^{m} \Phi\left(f(x_i) + u_j\right) \right\|^2}_{=:\Delta_{X \to Y}}$$

Can decide causal direction by comparing $\Delta_{X \to Y}$ and $\Delta_{Y \to X}$:

Theorem: if the additive noise model is identifiable and the kernel is characteristic, then

$$\Delta_{X \to Y} \to 0 \text{ and } \Delta_{Y \to X} \not\to 0$$

in probability as $m \to \infty$.

(Schölkopf, Muandet, Fukumizu, Harmeling, Peters, 2015; Scibior et al., 2016))

Detection of Confounders

Given $p(X, Y)$, infer whether

- $X \to Y$
- $Y \to X$
- $X \leftarrow T \to Y$ for some (possibly) unobserved variable T

- Confounded additive noise (CAN) models

$$X = f_X(T) + U_X$$
$$Y = f_Y(T) + U_Y$$

with functions f_X, f_Y and U_X, U_Y, T jointly independent

Note: includes the case

$$Y = f(X) + U$$

by setting $f_X = id$ and $U_X = 0$.

- Estimate $(f_X(T), f_Y(T))$ using dimensionality reduction
- If U_X or U_Y is close to zero, output 'no confounder'
- Identifiability result for small noise

Janzing, Peters, Mooij, Schölkopf: Identifying latent confounders using additive noise models. *UAI 2009*

Independence of input and mechanism

Causal structure:

C cause
E effect
N noise
φ mechanism

Assumption:
$p(C)$ and $p(E|C)$ are "independent"

Janzing & Schölkopf, IEEE Trans. Inf. Theory, 2010; cf. also Lemeire & Dirkx, 2007

Inferring deterministic causal relations

- Does not require noise
- Assumption: $y = f(x)$ with invertible f

Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf:
Inferring deterministic causal relations, *UAI* 2010

Causal independence implies anticausal dependence

Assume that f is a monotonically increasing bijection of $[0,1]$.

View p_x and $\log f'$ as RVs on the prob. space $[0,1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\mathrm{Cov}\,(\log f', p_x) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx,$$

since $\mathrm{Cov}\,(\log f', p_x) = E\,[\log f' \cdot p_x] - E\,[\log f']\,E\,[p_x] = E\,[\log f' \cdot p_x] - E\,[\log f']$.

Proposition: If $f \neq Id$,

$$\mathrm{Cov}\,(\log f^{-1'}, p_y) > 0.$$

u_x, u_y uniform densities for x, y
v_x, v_y densities for x, y induced by transforming u_y, u_x via f^{-1} and f

Equivalent formulations of the postulate:

Additivity of Entropy:

$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

Orthogonality (information geometric):

$$D(p_x \parallel v_x) = D(p_x \parallel u_x) + D(u_x \parallel v_x)$$

which can be rewritten as

$$D(p_y \parallel u_y) = D(p_x \parallel u_x) + D(v_y \parallel u_y)$$

Interpretation:
irregularity of p_y = irregularity of p_x + irregularity introduced by f

Testable implication / inference rule

- If $X \to Y$ then

$$\int \log |f'(x)| p(x) dx \leq \int \log |f^{-1'}(y)| p(y) dy$$

(high density $p(y)$ tends to occur at points with large slope)

- empirical estimator

$$\hat{C}_{X \to Y} := \frac{1}{m} \sum_{j=1}^{m} \log \left| \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \right| \approx \int \log |f'(x)| p(x) dx$$

- infer $X \to Y$ whenever

$$\hat{C}_{X \to Y} < \hat{C}_{Y \to X} \, .$$

"information geometric causal inference"

Benchmark dataset with 106 cause-effect pairs

http://webdav.tuebingen.mpg.de/cause-effect/

discussion of the ground truth and extensive performance studies for bivariate causal inference methods:

Mooij et al: Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks, JMLR 2016

Cause-Effect Pairs − Examples

	var 1	var 2	dataset	ground truth
pair0001	Altitude	Temperature	DWD	→
pair0005	Age (Rings)	Length	Abalone	→
pair0012	Age	Wage per hour	census income	→
pair0025	cement	compressive strength	concrete_data	→
pair0033	daily alcohol consumption	mcv mean corpuscular volume	liver disorders	→
pair0040	Age	diastolic blood pressure	pima indian	→
pair0042	day	temperature	B. Janzing	→
pair0047	#cars/24h	specific days	traffic	←
pair0064	drinking water access	infant mortality rate	UNdata	→
pair0068	bytes sent	open http connections	P. Daniusis	←
pair0069	inside room temperature	outside temperature	J. M. Mooij	←
pair0070	parameter	sex	Bülthoff	→
pair0072	sunspot area	global mean temperature	sunspot data	→
pair0074	GNI per capita	life expectancy at birth	UNdata	→
pair0078	PPFD (Photosynth. Photon Flux)	NEP (Net Ecosystem Productivity)	Moffat A. M.	→

MAX-PLANCK-GESELLSCHAFT

IGCI:
Deterministic
Method

LINGAM:
Shimizu et al.,
2006

AN:
Additive Noise
Model
(nonlinear)

PNL:
AN with post-
nonlinearity

GPI:
Mooij et al.,
2010

(source: Mooij et
al, JMLR 2016)

$X_i := f_i(\mathrm{PA}_i, U_i)$ with independent RVs $U_1, \ldots, U_n$.

Can we recover G from p?

approach	assumptions	method	intuition
graphical approach *(Pearl, Spirtes, Glymour, Scheines)*	noises jointly independent; faithfulness	conditional independence testing ($n \geq 3$)	track how the noises spread
ICM *(Daniušis et al., UAI 2010; Shajarisales et al., ICML 2015)*	noises and f_i independent; f_i learnable	customized tests	noises pick up footprints of the functions
additive noise model *(Peters, Mooij, Janzing, Schölkopf, JMLR 2014)*	$X_i = f_i(\mathrm{PA}_i) + U_i$ with learnable f_i	regression & unconditional independence testing	restriction of function class

Recall different aspects of independence

- **informational:** P_C and $P_{E|C}$ don't contain information about each other
- **modularity:** P_C and $P_{E|C}$ often change independently across datasets

$\Rightarrow$ machine learning should care about the causal direction in prediction tasks

MAX-PLANCK-GESELLSCHAFT

Causal Learning and Anticausal Learning

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, *ICML* 2012

- example 1: predict gene from mRNA sequence

Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png

causal mechanism φ

- example 2: predict class membership from handwritten digit

Covariate Shift and Semi-Supervised Learning

Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

Semi-supervised learning: improve estimate by more data from $p(X)$
Covariate shift: $p(X)$ changes between training and test

Causal assumption: $p(C)$ and mechanism $p(E|C)$ "independent"

Causal learning

$p(X)$ and $p(Y|X)$ independent

1. *semi-supervised learning impossible*
2. $p(Y|X)$ *invariant under change in* $p(X)$

Anticausal learning

$p(Y)$ and $p(X|Y)$ independent

hence $p(X)$ and $p(Y|X)$ dependent

1. *semi-supervised learning possible*
2. $p(Y|X)$ *changes with* $p(X)$

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, 2012, cf. Storkey, 2009; Bareinboim & Pearl, 2012

- Experimental Meta-Analysis confirms prediction

  Schölkopf et al., ICML 2012

- All known SSL assumptions link $p(X)$ to $p(Y|X)$:
  - ***Cluster assumption***: points in same cluster of $p(X)$ have the same Y
  - ***Low density separation assumption***: $p(Y|X)$ should cross 0.5 in an area where $p(X)$ is small
  - ***Semi-supervised smoothness assumption***: $\mathrm{E}(Y|X)$ should be smooth where $p(X)$ is large

MAX-PLANCK-GESELLSCHAFT

SSL Book Benchmark Datasets – Chapelle et al. (2006)

Table 1. Categorization of eight benchmark datasets as Anticausal/Confounded, Causal or Unclear

Category	Dataset
Anticausal/ Confounded	g241c: the class causes the 241 features.
	g241d: the class (binary) and the features are confounded by a variable with 4 states.
	Digit1: the positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS: the class and the features are confounded by the 10-state variable of all digits.
	COIL: the six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr: the amino acid is the cause of the secondary structure.
Unclear	BCI, Text: Unclear which is the cause and which the effect.

MAX-PLANCK-GESELLSCHAFT

UCI Datasets used in SSL benchmark – Guo et al., 2010

Table 2. Categorization of 26 UCI datasets as Anticausal/Confounded, Causal or Unclear

Categ.	Dataset
Anticausal/Confounded	Breast Cancer Wisconsin: the class of the tumor (benign or malignant) causes some of the features of the tumor (e.g., thickness, size, shape etc.).
	Diabetes: whether or not a person has diabetes affects some of the features (e.g., glucose concentration, blood pressure), but also is an effect of some others (e.g. age, number of times pregnant).
	Hepatitis: the class (die or survive) and many of the features (e.g., fatigue, anorexia, liver big) are confounded by the presence or absence of hepatitis. Some of the features, however, may also cause death.
	Iris: the size of the plant is an effect of the category it belongs to.
	Labor: cyclic causal relationships: good or bad labor relations can cause or be caused by many features (e.g., wage increase, number of working hours per week, number of paid vacation days, employer's help during employee 's long term disability). Moreover, the features and the class may be confounded by elements of the character of the employer and the employee (e.g., ability to cooperate).
	Letter: the class (letter) is a cause of the produced image of the letter.
	Mushroom: the attributes of the mushroom (shape, size) and the class (edible or poisonous) are confounded by the taxonomy of the mushroom (23 species).
	Image Segmentation: the class of the image is the cause of the features of the image.
	Sonar, Mines vs. Rocks: the class (Mine or Rock) causes the sonar signals.
	Vehicle: the class of the vehicle causes the features of its silhouette.
	Vote: this dataset may contain causal, anticausal, confounded and cyclic causal relations. E.g., having handicapped infants or being part of religious groups in school can cause one's vote, being democrat or republican can causally influence whether one supports Nicaraguan contras, immigration may have a cyclic causal relation with the class. Crime and the class may be confounded, e.g., by the environment in which one grew up.
	Vowel: the class (vowel) causes the features.
	Wave: the class of the wave causes its attributes.
Causal	Balance Scale: the features (weight and distance) cause the class.
	Chess (King-Rook vs. King-Pawn): the board-description causally influences whether white will win.
	Splice: the DNA sequence causes the splice sites.
Unclear	Breast-C, Colic, Sick, Ionosphere, Heart, Credit Approval were unclear to us. In some of the datasets, it is unclear whether the class label may have been generated or defined based on the features (e.g., Ionoshpere, Credit Approval, Sick).

MAX-PLANCK-GESELLSCHAFT

Datasets, co-regularized LS regression – Brefeld et al., 2006

Table 3. Categorization of 31 datasets (described in the paragraph "Semi-supervised regression") as Anticausal/Confounded, Causal or Unclear

Categ.	Dataset	Target variable	Remark
Anticausal/Confounded	breastTumor	tumor size	causing predictors such as inv-nodes and deg-malig
	cholesterol	cholesterol	causing predictors such as resting blood pressure and fasting blood sugar
	cleveland	presence of heart disease in the patient	causing predictors such as chest pain type, resting blood pressure, and fasting blood sugar
	lowbwt	birth weight	causing the predictor indicating low birth weight
	pbc	histologic stage of disease	causing predictors such as Serum bilirubin, Prothrombin time, and Albumin
	pollution	age-adjusted mortality rate per 100,000	causing the predictor number of 1960 SMSA population aged 65 or older
	wisconsin	time to recur of breast cancer	causing predictors such as perimeter, smoothness, and concavity
Causal	autoMpg	city-cycle fuel consumption in miles per gallon	caused by predictors such as horsepower and weight
	cpu	cpu relative performance	caused by predictors such as machine cycle time, maximum main memory, and cache memory
	fishcatch	fish weight	caused by predictors such as fish length and fish width
	housing	housing values in suburbs of Boston	caused by predictors such as pupil-teacher ratio and nitric oxides concentration
	machine_cpu	cpu relative performance	see remark on "cpu"
	meta	normalized prediction error	caused by predictors such as number of examples, number of attributes, and entropy of classes
	pwLinear	value of piecewise linear function	caused by all 10 involved predictors
	sensory	wine quality	caused by predictors such as trellis
	servo	rise time of a servomechanism	caused by predictors such as gain settings and choices of mechanical linkages
Unclear	auto93 (target: midrange price of cars); bodyfat (target: percentage of body fat); autoHorse (target: price of cars); autoPrice (target: price of cars); baskball (target: points scored per minute); cloud (target: period rainfalls in the east target); echoMonths (target: number of months patient survived); fruitfly (target: longevity of mail fruitflies); pharynx (target: patient survival); pyrim (quantitative structure activity relationships); sleep (target: total sleep in hours per day); stock (target: price of one particular stock); strike (target: strike volume); triazines (target: activity); veteran (survival in days)		

MAX-PLANCK-GESELLSCHAFT

Benchmark Datasets of *Chapelle et al. (2006)*

Asterisk = 1-NN, SVM

Self-training does not help for causal problems (cf. *Guo et al., 2010)*

Relative error decrease = (error(base) –error(self-train)) / error(base)

Co-regularization helps for the **anticausal** problems of *Brefeld et al., 2006*

Co-regularization hardly helps for the **causal** problems of *Brefeld et al., 2006*

Higher-order Semi-Supervised Learning *(with Julius von Kügelgen, UAI 2020)*

Causal factorization

$$P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y)$$

Non-causal factorization

$$P(Y, X_E|X_C) = P(X_E|X_C)P(Y|X_C, X_E)$$

$P(X_E|X_C)$ can be estimated from unlabelled data and contains information about $P(Y|X_C, X_E)$.

Conditional cluster assumption: points in the same cluster of $p(X_E|X_C)$ have the same Y

Algorithmic structural causal model

- for every node x_j there exists a program u_j that computes x_j from its parents pa_j

- all u_j are jointly independent
- the program u_j represents the causal mechanism that generates the effect from its causes
- u_j are the analog of the unobserved noise terms in the statistical functional model

Theorem: this model implies the causal Markov condition (replacing Shannon entropy with Kolmorogov complexity).

(Janzing & Schölkopf, IEEE Trans. Information Theory 2010)

Gedankenexperiment

Particles scattered at an object

- incoming beam: ‘cause’
- scattering at object: ‘mechanism’
- outgoing beam: ‘effect’, contains information about the object

Independence assumption

- s initial state of a physical system
- M the system dynamics applied for some fixed time

Independence Principle: s and M are algorithmically independent

$$I(s : M) \stackrel{+}{=} 0,$$

i.e., knowing s does not enable a shorter description of M and vice versa.

Thermodynamic Arrow of Time

Theorem [non-decrease of entropy]. Let M be a bijective map on the set of states of a system then $I(s : M) \stackrel{+}{=} 0$ implies

$$K(M(s)) \stackrel{+}{\geq} K(s)$$

Proof idea: If $M(s)$ admits a shorter description than s, knowing M admits a shorter description of s: just describe $M(s)$ and then apply M^{-1}.

Janzing, Chaves, Schölkopf: Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. New J. of Physics, 2016

Milky Way Galaxy
Kepler Search Space
3,000 light years
Sagittarius Arm
Sun
Orion Spur
Perseus Arm

Exoplanet Transits

MAX-PLANCK-GESELLSCHAFT

with Hogg, Wang, Foreman-Mackey, Janzing, Simon-Gabriel, Peters, Montet, and Morton.
ICML 2015
Astrophysical Journal 2015
PNAS 2016

half-siblings

Q-E[Q] = Y – E[Y|X]

Kepler 5088536 Quarter 5
CCD channel 25 Row 875 Column 322

| 3 months |

Kepler 5949551 Quarter 5
CCD channel 25 Row 57 Column 756

MAX-PLANCK-GESELLSCHAFT

The Milkman Problem

Half-Sibling Regression

Idea: remove $E[Y|X]$ from Y to reconstruct Q.

$X \perp\!\!\!\perp Q$

X and Y share information (only) through N

If we try to predict Y from X, we only pick up the part due to N

with David Hogg, Dan Foreman-Mackey, Dun Wang, Dominik Janzing
Jonas Peters, Carl-Johann Simon-Gabriel *(ICML 2015)*

MAX-PLANCK-GESELLSCHAFT

Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable. Suppose

- $Y = Q + f(N)$ *(additive noise model)*

- $f(N) = \psi(X)$ for some ψ *(complete information).*

Then $\hat{Q} := Y - \mathbb{E}[Y|X] = Q - \mathbb{E}[Q]$.

Q can be reconstructed, up to a constant offset, from Y and $\mathbb{E}[Y|X]$.

Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable. Suppose

- $Y = Q + f(N)$ *(additive noise model)*

Then $E[(\hat{Q} - (Q - E[Q]))^2] = E[\mathrm{Var}[f(N)|X]]$.

If $f(N)$ can (in principle) be predicted well from X, then Q can be reconstructed well.

Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

Planet-Hunting Kepler Spacecraft Suffers Major Failure, NASA Says

By Mike Wall May 15, 2013 Science & Astronomy

An artist's interpretation of the Kepler observatory in space. (Image: © NASA.)

This story was updated at 5:20 p.m. EDT.

The planet-hunting days of NASA's prolific Kepler space telescope, which has discovered more than 2,700 potential alien worlds to date, may be over.

The second of Kepler's four reaction wheels — devices that allow the observatory to maintain its position in space — has failed, NASA officials announced Wednesday (May 15).

Credits: NASA Ames/W Stenzel

The Astrophysical Journal, 806:215 (13pp), 2015 June 20 doi:10.1088/0004-637X/806/2/215
© 2015. The American Astronomical Society. All rights reserved.

A SYSTEMATIC SEARCH FOR TRANSITING PLANETS IN THE *K2* DATA

Daniel Foreman-Mackey[1], Benjamin T. Montet[2,3], David W. Hogg[1,4,5],
Timothy D. Morton[6], Dun Wang[1], and Bernhard Schölkopf[7]
[1] Center for Cosmology and Particle Physics, Department of Physics, New York University,
4 Washington Place, New York, NY 10003, USA; danfm@nyu.edu
[2] Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA
[3] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA
[4] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117, Heidelberg, Germany
[5] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[6] Department of Astrophysics, Princeton University, Princeton, NJ 08544, USA
[7] Max Planck Institute for Intelligent Systems, Spemannstrasse 38, D-72076, Tübingen, Germany
Received 2015 February 16; accepted 2015 May 12; published 2015 June 18

ABSTRACT

Photometry of stars from the *K2* extension of NASA's *Kepler* mission is afflicted by systematic effects caused by small (few-pixel) drifts in the telescope pointing and other spacecraft issues. We present a method for searching *K2* light curves for evidence of exoplanets by simultaneously fitting for these systematics and the transit signals of interest. This method is more computationally expensive than standard search algorithms but we demonstrate that it can be efficiently implemented and used to discover transit signals. We apply this method to the full Campaign 1 data set and report a list of 36 planet candidates transiting 31 stars, along with an analysis of the pipeline performance and detection efficiency based on artificial signal injections and recoveries. For all planet candidates, we present posterior distributions on the properties of each system based strictly on the transit observables.

Key words: catalogs – methods: data analysis – methods: statistical – planetary systems – stars: statistics

1. INTRODUCTION

The *Kepler* Mission was incredibly successful at finding transiting exoplanets in the light curves of stars. The Mission

a few percent of the data are actually stored and downloaded to Earth, there is not enough information in the data to derive or infer a complete or accurate flat-field map. Therefore, work on

Table 2
The Catalog of Planet Candidates and their Observable Properties

EPIC	*Kepler* mag	R.A. (J2000)	Decl. (J2000)	P (days)	t_0 [BJD-2456808]	$R_P/R_\star$
201208431	14.41	174.745640	−3.905585	$10.0040^{+0.0018}_{-0.0016}$	$7.5216^{+0.0098}_{-0.0090}$	$0.0349^{+0.0034}_{-0.0026}$
201257461	11.51	178.161109	−3.094936	$50.2677^{+0.0083}_{-0.0074}$	$20.3735^{+0.0147}_{-0.0098}$	$0.0334^{+0.0054}_{-0.0017}$
201295312	12.13	174.011630	−2.520881	$5.6562^{+0.0007}_{-0.0007}$	$3.7228^{+0.0086}_{-0.0091}$	$0.0175^{+0.0020}_{-0.0009}$
201338508	14.36	169.303502	−1.877976	$10.9328^{+0.0022}_{-0.0021}$	$6.5967^{+0.0088}_{-0.0081}$	$0.0339^{+0.0025}_{-0.0030}$
201338508	14.36	169.303502	−1.877976	$5.7350^{+0.0006}_{-0.0006}$	$0.8626^{+0.0054}_{-0.0055}$	$0.0331^{+0.0025}_{-0.0023}$
201367065	11.57	172.334949	−1.454787	$10.0542^{+0.0004}_{-0.0004}$	$5.4186^{+0.0018}_{-0.0018}$	$0.0354^{+0.0022}_{-0.0011}$
201367065	11.57	172.334949	−1.454787	$24.6470^{+0.0014}_{-0.0016}$	$4.2769^{+0.0030}_{-0.0029}$	$0.0272^{+0.0016}_{-0.0013}$
201384232	12.51	178.192260	−1.198477	$30.9375^{+0.0029}_{-0.0052}$	$19.5035^{+0.0053}_{-0.0039}$	$0.0260^{+0.0011}_{-0.0011}$
201393098	13.05	167.093771	−1.065755	$28.6793^{+0.0105}_{-0.0116}$	$16.6212^{+0.0305}_{-0.0177}$	$0.0231^{+0.0028}_{-0.0020}$
201403446	11.99	174.266344	−0.907261	$19.1535^{+0.0050}_{-0.0050}$	$7.3437^{+0.0116}_{-0.0143}$	$0.0154^{+0.0014}_{-0.0013}$
201445392	14.38	169.793665	−0.284375	$10.3527^{+0.0011}_{-0.0011}$	$5.6110^{+0.0047}_{-0.0051}$	$0.0349^{+0.0045}_{-0.0025}$
201445392	14.38	169.793665	−0.284375	$5.0644^{+0.0006}_{-0.0006}$	$5.0690^{+0.0059}_{-0.0064}$	$0.0274^{+0.0025}_{-0.0020}$
201465501	14.96	176.264468	0.005301	$18.4488^{+0.0015}_{-0.0015}$	$14.6719^{+0.0035}_{-0.0032}$	$0.0531^{+0.0061}_{-0.0039}$
201505350	12.81	174.960319	0.603575	$11.9069^{+0.0005}_{-0.0004}$	$9.2764^{+0.0013}_{-0.0015}$	$0.0446^{+0.0009}_{-0.0006}$
201505350	12.81	174.960319	0.603575	$7.9193^{+0.0001}_{-0.0001}$	$5.3840^{+0.0006}_{-0.0008}$	$0.0747^{+0.0016}_{-0.0013}$
201546283	12.43	171.515165	1.230738	$6.7713^{+0.0001}_{-0.0001}$	$4.8453^{+0.0012}_{-0.0011}$	$0.0481^{+0.0020}_{-0.0012}$
201549860	13.92	170.103081	1.285956	$5.6083^{+0.0005}_{-0.0006}$	$4.1195^{+0.0045}_{-0.0047}$	$0.0283^{+0.0041}_{-0.0023}$
201555883	15.06	176.075940	1.375947	$5.7966^{+0.0002}_{-0.0002}$	$5.3173^{+0.0027}_{-0.0050}$	$0.0604^{+0.0068}_{-0.0032}$
201565013	16.91	176.992193	1.510249	$8.6381^{+0.0003}_{-0.0002}$	$3.4283^{+0.0016}_{-0.0015}$	$0.1538^{+0.0355}_{-0.0243}$
201569483	11.77	167.171299	1.577513	$5.7969^{+0.0000}_{-0.0000}$	$5.3130^{+0.0002}_{-0.0003}$	$0.3587^{+0.0379}_{-0.0334}$
201577035	12.30	172.121957	1.690636	$19.3062^{+0.0013}_{-0.0013}$	$11.5790^{+0.0025}_{-0.0027}$	$0.0380^{+0.0023}_{-0.0012}$
201596316	13.15	169.042002	1.986840	$39.8415^{+0.0136}_{-0.0155}$	$21.8572^{+0.0120}_{-0.0101}$	$0.0267^{+0.0034}_{-0.0022}$
201613023	12.14	173.192036	2.244884	$8.2818^{+0.0006}_{-0.0007}$	$7.3752^{+0.0055}_{-0.0052}$	$0.0205^{+0.0012}_{-0.0008}$
201617985	14.11	179.491659	2.321476	$7.2823^{+0.0007}_{-0.0008}$	$4.6337^{+0.0050}_{-0.0050}$	$0.0333^{+0.0072}_{-0.0032}$
201629650	12.73	170.155528	2.502696	$40.0492^{+0.0186}_{-0.0259}$	$4.5363^{+0.0202}_{-0.0172}$	$0.0241^{+0.0025}_{-0.0020}$
201635569	15.55	178.057026	2.594245	$8.3681^{+0.0002}_{-0.0002}$	$3.4514^{+0.0015}_{-0.0014}$	$0.0991^{+0.0120}_{-0.0078}$
201649426	13.22	177.234262	2.807619	$27.7704^{+0.0001}_{-0.0001}$	$13.3476^{+0.0001}_{-0.0002}$	$0.4365^{+0.0777}_{-0.0583}$
201702477	14.43	175.240794	3.681584	$40.7365^{+0.0026}_{-0.0025}$	$3.5451^{+0.0026}_{-0.0025}$	$0.0808^{+0.0043}_{-0.0114}$
201736247	14.40	178.110797	4.254747	$11.8106^{+0.0016}_{-0.0019}$	$3.8483^{+0.0093}_{-0.0071}$	$0.0347^{+0.0030}_{-0.0024}$
201754305	14.30	175.097258	4.557340	$19.0726^{+0.0048}_{-0.0049}$	$1.4893^{+0.0128}_{-0.0133}$	$0.0297^{+0.0042}_{-0.0030}$
201754305	14.30	175.097258	4.557340	$7.6202^{+0.0012}_{-0.0011}$	$3.6813^{+0.0061}_{-0.0057}$	$0.0281^{+0.0034}_{-0.0026}$
201779067	11.12	168.542699	4.988131	$27.2429^{+0.0001}_{-0.0001}$	$12.2599^{+0.0002}_{-0.0003}$	$0.2535^{+0.0369}_{-0.0259}$
201828749	11.56	175.654342	5.894323	$33.5093^{+0.0023}_{-0.0018}$	$5.1554^{+0.0037}_{-0.0032}$	$0.0267^{+0.0021}_{-0.0020}$
201855371	13.00	178.329775	6.412261	$17.9715^{+0.0015}_{-0.0017}$	$9.9412^{+0.0033}_{-0.0038}$	$0.0311^{+0.0030}_{-0.0017}$
201912552	12.47	172.560460	7.588391	$32.9410^{+0.0039}_{-0.0032}$	$28.1834^{+0.0057}_{-0.0105}$	$0.0513^{+0.0035}_{-0.0056}$
201929294	12.97	174.656969	7.959611	$5.0084^{+0.0001}_{-0.0001}$	$4.5703^{+0.0022}_{-0.0012}$	$0.1163^{+0.0011}_{-0.0014}$

The Astrophysical Journal, 809:25 (15pp), 2015 August 10 doi:10.1088/0004-637X/809/1/25
© 2015. The American Astronomical Society. All rights reserved.

STELLAR AND PLANETARY PROPERTIES OF *K2* CAMPAIGN 1 CANDIDATES AND VALIDATION OF 17 PLANETS, INCLUDING A PLANET RECEIVING EARTH-LIKE INSOLATION

Benjamin T. Montet[1,2], Timothy D. Morton[3], Daniel Foreman-Mackey[4,5], John Asher Johnson[2], David W. Hogg[4,5,6], Brendan P. Bowler[1,8], David W. Latham[2], Allyson Bieryla[2], and Andrew W. Mann[7,9]

[1] Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA; btm@astro.caltech.edu
[2] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA
[3] Department of Astrophysics, Princeton University, Princeton, NJ 08544, USA
[4] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[5] Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, NY 10003, USA
[6] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[7] Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA
Received 2015 March 27; accepted 2015 July 2; published 2015 August 5

ABSTRACT

The extended *Kepler* mission, *K2*, is now providing photometry of new fields every three months in a search for transiting planets. In a recent study, Foreman-Mackey and collaborators presented a list of 36 planet candidates orbiting 31 stars in *K2* Campaign 1. In this contribution, we present stellar and planetary properties for all systems. We combine ground-based seeing-limited survey data and adaptive optics imaging with an automated transit analysis scheme to validate 21 candidates as planets, 17 for the first time, and identify 6 candidates as likely false positives. Of particular interest is K2-18 (EPIC 201912552), a bright ($K = 8.9$) M2.8 dwarf hosting a 2.23 ± 0.25 $R_\oplus$ planet with $T_{\rm eq} = 272 \pm 15$ K and an orbital period of 33 days. We also present two new open-source software packages which enable this analysis. The first, `isochrones`, is a flexible tool for fitting theoretical stellar models to observational data to determine stellar properties using a nested sampling scheme to capture the multimodal nature of the posterior distributions of the physical parameters of stars that may plausibly be evolved. The second is `vespa`, a new general-purpose procedure to calculate false positive probabilities and statistically validate transiting exoplanets.

Key words: catalogs – planetary systems – planets and satellites: detection – stars: fundamental parameters

Table 3
Planet Properties for All Objects of Interest

Candidate	Period (days)	Epoch (BJD-2456808)	Radius ($R_\oplus$)	$a/R_\star$	a (AU)	T_{eq} (K)	Disposition
201208431.01/K2-4b	10.00329 ± 0.00159	7.5212 ± 0.0080	2.37 ± 0.40	27.79 ± 0.72	0.0777 ± 0.0012	563 ± 11	Planet
201257461.01	50.27762 ± 0.00785	20.3735 ± 0.0397	209.52 ± 99.23	6.19 ± 0.52	0.3049 ± 0.0030	1466 ± 52	FP
201295312.01	5.65706 ± 0.00079	3.7187 ± 0.0082	2.16 ± 0.57	12.94 ± 4.07	0.0633 ± 0.0019	1211 ± 154	Candidate
201338508.01/K2-5c	10.93406 ± 0.00205	6.5947 ± 0.0080	1.92 ± 0.20	32.27 ± 0.71	0.0783 ± 0.0007	511 ± 9	Planet
201338508.02/K2-5b	5.73491 ± 0.00061	0.8640 ± 0.0063	1.92 ± 0.23	20.99 ± 0.46	0.0509 ± 0.0004	634 ± 12	Planet
201367065.01/K2-3b	10.05448 ± 0.00033	5.4177 ± 0.0015	1.98 ± 0.10	30.72 ± 0.75	0.0740 ± 0.0009	504 ± 9	Planet
201367065.02/K2-3c	24.64745 ± 0.00152	4.2759 ± 0.0030	1.56 ± 0.10	55.85 ± 1.36	0.1345 ± 0.0016	374 ± 7	Planet
201384232.01/K2-6b	30.94191 ± 0.00467	19.5014 ± 0.0090	2.50 ± 0.88	50.27 ± 24.56	0.1898 ± 0.0056	615 ± 105	Planet
201393098.01/K2-7b	28.67992 ± 0.00947	16.6155 ± 0.0149	2.67 ± 0.56	40.29 ± 8.19	0.1814 ± 0.0043	651 ± 61	Planet
201403446.01	19.15344 ± 0.00607	7.3412 ± 0.0152	2.04 ± 0.46	27.05 ± 5.87	0.1408 ± 0.0040	889 ± 88	Candidate
201445392.01/K2-8b	10.35176 ± 0.00133	5.6119 ± 0.0053	2.97 ± 0.51	24.94 ± 0.79	0.0856 ± 0.0012	691 ± 14	Planet
201445392.02	5.06468 ± 0.00063	5.0663 ± 0.0071	2.31 ± 0.33	15.49 ± 0.49	0.0531 ± 0.0008	877 ± 17	Candidate
201465501.01/K2-9b	18.44883 ± 0.00137	14.6723 ± 0.0030	1.60 ± 0.42	74.76 ± 6.66	0.0848 ± 0.0050	284 ± 14	Planet
201505350.01/K2-19c	11.90691 ± 0.00037	9.2764 ± 0.0018	4.31 ± 0.49	24.09 ± 2.48	0.0965 ± 0.0017	797 ± 42	Planet
201505350.02/K2-19b	7.91943 ± 0.00007	5.3836 ± 0.0005	7.11 ± 0.81	18.35 ± 1.89	0.0735 ± 0.0013	913 ± 48	Planet
201546283.01	6.77131 ± 0.00012	4.8440 ± 0.0022	5.77 ± 3.24	17.56 ± 9.24	0.0668 ± 0.0029	991 ± 239	Candidate
201549860.01	5.60840 ± 0.00055	4.1181 ± 0.0047	2.20 ± 0.40	17.42 ± 0.46	0.0555 ± 0.0008	766 ± 14	Candidate
201555883.01	…	…	…	…	…	…	FP[b]
201565013.01	8.63810 ± 0.00024	3.4284 ± 0.0016	15.99 ± 9.19	28.07 ± 2.68	0.0669 ± 0.0031	536 ± 37	Candidate
201569483.01	5.79687 ± 0.00000	5.3135 ± 0.0004	27.81 ± 3.56	15.68 ± 1.91	0.0589 ± 0.0015	930 ± 51	FP
201577035.01/K2-10b	19.30691 ± 0.00127	11.5768 ± 0.0033	3.92 ± 0.69	32.74 ± 5.15	0.1374 ± 0.0025	703 ± 55	Planet
201596316.01/K2-11b	39.93767 ± 0.23229	21.8290 ± 0.1156	7.55 ± 9.33	45.08 ± 58.53	0.2257 ± 0.0143	734 ± 253	Planet
201613023.01/K2-12b	8.28212 ± 0.00060	7.3734 ± 0.0054	2.33 ± 0.58	17.47 ± 5.05	0.0802 ± 0.0021	1003 ± 121	Planet
201617985.01	7.28161 ± 0.00078	4.6366 ± 0.0047	1.78 ± 0.43	26.04 ± 1.16	0.0586 ± 0.0012	518 ± 16	Candidate
201629650.01/K2-13b	39.91488 ± 0.32477	4.5250 ± 0.0146	1.89 ± 0.95	79.69 ± 63.37	0.2114 ± 0.0061	511 ± 126	Planet
201635569.01/K2-14b	8.36802 ± 0.00019	3.4513 ± 0.0013	4.81 ± 0.42	30.16 ± 0.69	0.0627 ± 0.0006	488 ± 8	Planet
201649426.01	27.77045 ± 0.00008	13.3482 ± 0.0012	32.79 ± 9.01	59.26 ± 13.58	0.1517 ± 0.0097	441 ± 42	FP
201702477.01	40.73620 ± 0.00266	3.5455 ± 0.0025	7.28 ± 1.10	56.98 ± 7.61	0.2205 ± 0.0053	529 ± 36	Candidate
201736247.01/K2-15b	11.81040 ± 0.00204	3.8509 ± 0.0076	2.48 ± 0.30	28.84 ± 1.98	0.0910 ± 0.0018	676 ± 26	Planet
201754305.01/K2-16c	19.07536 ± 0.00490	1.4854 ± 0.0119	2.14 ± 0.41	41.43 ± 1.34	0.1220 ± 0.0021	523 ± 12	Planet
201754305.02/K2-16b	7.62067 ± 0.00095	3.6802 ± 0.0054	2.13 ± 0.37	22.47 ± 0.73	0.0662 ± 0.0011	710 ± 16	Planet
201779067.01	27.24273 ± 0.00012	12.2601 ± 0.0003	31.73 ± 5.25	38.25 ± 3.72	0.1718 ± 0.0022	707 ± 34	FP
201828749.01	33.51569 ± 0.00232	5.1504 ± 0.0034	3.83 ± 3.25	67.09 ± 67.64	0.1875 ± 0.0090	613 ± 239	Candidate
201855371.01/K2-17b	17.96753 ± 0.00152	9.9462 ± 0.0035	2.23 ± 0.20	39.38 ± 0.85	0.1190 ± 0.0020	487 ± 10	Planet
201912552.01/K2-18b[a]	32.94488 ± 0.00281	28.1849 ± 0.0027	2.24 ± 0.23	83.83 ± 9.03	0.1491 ± 0.0055	272 ± 15	Planet
201929294.01	…	…	…	…	…	…	FP[b]

Habitable Zone Gallery

Home Plots Table Gallery Movies About Links

This site is dedicated to tracking the orbits of exoplanets in relation to their Habitable Zones.

Planets: 3706 **Systems:** 2806
Planets with orbits entirely within the Habitable Zone: 129 [?]
Updated: 2019 08 29 14:39:34 PDT

"The Earth is the only world known so far to harbor life. There is nowhere else, at least in the near future, to which our species could migrate. Visit, yes. Settle, not yet. Like it or not, for the moment the Earth is where we make our stand." - Carl Sagan

NATIONAL GEOGRAPHIC

Water found on a potentially life-friendly alien planet

A super-Earth about 111 light-years away is "the best candidate for habitability that we know right now," astronomers say.

Search · The Guardian · International edition

3 MINUTE READ

Water found on most habitable known world beyond solar system

But humans would not fare well on planet K2-18b despite wispy clouds and huge red sun

tagesschau.de

Startseite | Videos & Audios | Inland | Ausland | Investigati

Startseite › Ausland › Erstmals Wasserdampf in Planetenatm

Astronomische Sensation

Wasserdampf auf Planet K2-18l

SCIENTIFIC AMERICAN. Subscribe

Observations

No, the Exoplanet K2-18b Is *Not* Habitable

News outlets that said otherwise are just crying wolf—but they're not the only ones at fault

By Laura Kreidberg on September 23, 2019

Draft version September 11, 2019
Typeset using LaTeX **twocolumn** style in AASTeX62

arXiv:1909.04642v1 [astro-ph.EP] 10 Sep 2019

Water Vapor on the Habitable-Zone Exoplanet K2-18b

Björn Benneke,[1] Ian Wong,[2,3] Caroline Piaulet,[1] Heather A. Knutson,[4] Ian J.M. Crossfield,[5] Joshua Lothringer,[6] Caroline V. Morley,[7] Peter Gao,[8,3] Thomas P. Greene,[9] Courtney Dressing,[8] Diana Dragomir,[5,10] Andrew W. Howard,[11] Peter R. McCullough,[6] Eliza M.-R. Kempton,[12,13] Jonathan J. Fortney,[14] and Jonathan Fraine[15]

[1] *Institute for Research on Exoplanets and Department of Physics, Université de Montréal, Montreal, QC, Canada*
[2] *Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA*
[3] *51 Pegasi b Fellow*
[4] *Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA*
[5] *Department of Physics and Kavli Institute of Astronomy, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA*
[6] *Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD 21218, USA*
[7] *Department of Astronomy, University of Texas, Austin, TX 78712, USA*
[8] *Department of Astronomy, University of California - Berkeley , Berkeley, CA, 94720, USA*
[9] *NASA Ames Research Center, Moffett Field, CA, 94035, USA*
[10] *NASA Hubble Fellow*
[11] *Department of Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[12] *Department of Astronomy, University of Maryland, College Park, MD 20742, USA*
[13] *Department of Physics, Grinnell College, 1116 8th Avenue, Grinnell, IA 50112, USA*
[14] *Department of Astronomy, University of California, Santa Cruz, CA 95064, USA*
[15] *Center for Extrasolar Planetary Systems, Space Science Institute, Boulder, CO 80301, USA*

ABSTRACT

Ever since the discovery of the first exoplanet, astronomers have made steady progress towards finding and probing planets in the habitable zone of their host stars, where the conditions could be right for liquid water to form and life to sprawl. Results from the Kepler mission indicate that the occurrence rate of habitable-zone Earths and super-Earths may be as high as 5–20%. Despite this abundance, probing the conditions and atmospheric properties on any of these habitable-zone planets is extremely difficult and has remained elusive to date. Here, we report the detection of water vapor and the likely presence of liquid water clouds in the atmosphere of the 8.6 $M_\oplus$ habitable-zone planet K2-18b. With a 33 day orbit around a cool M3 dwarf, K2-18b receives virtually the same amount of total radiation from its host star (1441 ± 80 W/m^2) as the Earth receives from the Sun (1370 W/m^2), making it a good candidate to host liquid water clouds. In this study we observed eight transits using HST/WFC3 in order to achieve the necessary sensitivity to detect water vapor. While the thick gaseous envelope of K2-18b means that it is not a true Earth analogue, our observations demonstrate that low-mass habitable-zone planets with the right conditions for liquid water are accessible with state-of-the-art telescopes.

Keywords: planets and satellites: individual (K2-18b) – planets and satellites: atmospheres

1. INTRODUCTION

The recent discovery of the transiting 8.63 ± 1.35 $M_\oplus$ exoplanet K2-18b in the habitable zone of a bright, nearby M3-dwarf provides us with an opportunity to carry out the spectroscopic study of the atmosphere of a habitable-zone planet outside our solar system (Montet et al. 2015, Benneke et al. 2017, Cloutier et al. 2019). K2-18b is an intriguing planet because its equilibrium temperature (265 ± 5K at an albedo of $A = 0.3$) is potentially very close to that of the Earth (257 K). The planet's predicted temperature provides the right con-

Corresponding author: Björn Benneke
bbenneke@astro.umontreal.ca

nature astronomy
LETTERS
https://doi.org/10.1038/s41550-019-0878-9

Water vapour in the atmosphere of the habitable-zone eight-Earth-mass planet K2-18 b

Angelos Tsiaras ★, Ingo P. Waldmann ★, Giovanna Tinetti, Jonathan Tennyson and Sergey N. Yurchenko

In the past decade, observations from space and the ground have found water to be the most abundant molecular species, after hydrogen, in the atmospheres of hot, gaseous extrasolar planets[1–4]. Being the main molecular carrier of oxygen, water is a tracer of the origin and the evolution mechanisms of planets. For temperate, terrestrial planets, the presence of water is of great importance as an indicator of habitable conditions. Being small and relatively cold, these planets and their atmospheres are the most challenging to observe, and therefore no atmospheric spectral signatures have so far been detected[6]. Super-Earths—planets lighter than ten Earth masses—around later-type stars may provide our first opportunity to study spectroscopically the characteristics of such planets, as they are best suited for transit observations. Here, we report the detection of a spectroscopic signature of water in the atmosphere of K2-18 b—a planet of eight Earth masses in the habitable zone of an M dwarf[7]—with high statistical confidence (Atmospheric Detectability Index[5] = 5.0, ~3.6σ (refs. [8,9])). In addition, the derived mean molecular weight suggests an atmosphere still containing some hydrogen. The observations were recorded with the Hubble Space Telescope/Wide Field Camera 3 and analysed with our dedicated, publicly available, algorithms[5,9]. Although the suitability of M dwarfs to host habitable worlds is still under discussion[10–13], K2-18 b offers an unprecedented opportunity to gain insight into the composition and climate of habitable-zone planets.

Atmospheric characterization of super-Earths is currently within reach of the Wide Field Camera 3 (WFC3) on board the Hubble Space Telescope (HST), combined with the recently implemented spatial scanning observational strategy[14]. The spectra of three hot transiting planets with radii less than 3.0 Earth radii ($R_\oplus$) have been published so far: Gliese 1214 b[15], HD 97658 b[16] and 55 Cancri e[17]. The first two do not show any evident transit depth modulation with wavelength, suggesting an atmosphere covered by thick clouds or made of molecular species heavier than hydrogen, while only the spectrum of 55 Cancri e has revealed a light-weight atmosphere, suggesting hydrogen–helium (H_2–He) still being present. In addition, transit observations of six temperate Earth-size planets around the ultra-cool dwarf TRAPPIST-1—planets b, c, d, e, f[5] and g[18]—have not shown any molecular signatures and have excluded the presence of cloud-free H_2–He atmospheres around them.

K2-18 b was discovered in 2015 by the Kepler spacecraft[7] and is orbiting around an M2.5 (metallicity [Fe/H] = 0.123 ± 0.157 dex (units of decimal exponent), effective temperature $T_{\rm eff} = 3{,}457 \pm 39$ K, stellar mass $M_\star = 0.359 \pm 0.047$ solar masses ($M_\odot$), stellar radius $R_\star = 0.411 \pm 0.038$ solar radii ($R_\odot$))[19] dwarf star, 34 pc away from the Earth. The star–planet distance of 0.1429 AU (ref. [19]) suggests a planet within the star's habitable zone (~0.12–0.25 AU) (ref. [20]), with effective temperature between 200 K and 320 K, depending on the albedo and the emissivity of its surface and/or its atmosphere. This crude estimate accounts for neither possible tidal energy sources[21] nor atmospheric heat redistribution[11,13], which might be relevant for this planet. Measurements of the mass and the radius of K2-18 b (planetary mass $M_p = 7.96 \pm 1.91$ Earth masses ($M_\oplus$) (ref. [22]), planetary radius $R_p = 2.279 \pm 0.0026\, R_\oplus$ (ref. [19])) yield a bulk density of 3.3 ± 1.2 g cm^{-1} (ref. [22]), suggesting either a silicate planet with an extended atmosphere or an interior composition with a water (H_2O) mass fraction lower than 50% (refs. [22–24]).

We analyse here eight transits of K2-18 b, obtained with the WFC3 camera on board the HST. We used our publicly available tools, specialized for HST/WFC3 data[5,9], to perform the end-to-end analysis from the raw data to the atmospheric parameters. The techniques used here have been validated by the analysis of the largest catalogue of exoplanetary spectra from WFC3[5]. Details can be found in Methods, and links to the data and the codes used can be found in 'Data availability' and 'Code availability', respectively. Along with the data, we provide descriptions of the data structures and instructions on how to reproduce the results presented here. Our analysis resulted in the detection of an atmosphere around K2-18 b with an Atmospheric Detectability Index[5] (ADI; a positively defined logarithmic Bayes factor) of 5.0, or approximately 3.6σ confidence[8,9], making K2-18 b the first habitable-zone planet in the super-Earth mass regime (1–10 $M_\oplus$) with an observed atmosphere around it.

More specifically, nine transits of K2-18 b were observed as part of the HST proposals 13665 and 14682 (principal investigator: Björn Benneke), and the data are available through the Mikulski Archive for Space Telescopes (MAST; see 'Data availability'). Each transit was observed during five HST orbits, with the G141 infrared grism (1.1–1.7 μm), and each exposure was the result of 16 up-the-ramp samples in the spatial scanning mode. The ninth transit observation suffered from pointing instabilities, and we therefore decided not to include it in this analysis. We extracted the white and the spectral light curves from the reduced images, following our dedicated methodology[5,17,25], which has been integrated into an automated, self-consistent and user-friendly Python package named Iraclis (see 'Code availability'). No systematic variations of the white light curve, $R_p/R_\star$, appeared between the eight different observations. This level of stability among the extracted broadband transit depths is not always guaranteed, as consistency problems among different observations emerged in previous analyses[5,16].

In our analysis, we found that the measured mid-transit times were not consistent with the expected ephemeris[19]. We used these results to refine the ephemeris of K2-18 b to be $P = 32.94007 \pm 0.00003$ days and $T_0 = 2457363.2109 \pm 0.0004$ BJD$_{\rm TDB}$ (ref. [26]), where P

Department of Physics & Astronomy, University College London, London, UK. *e-mail: atsiaras@star.ucl.ac.uk; ingo@star.ucl.ac.uk

http://people.tue.mpg.de/bs/K2-18b.html

Outlook: Causal discovery (Spirtes, Glymour, Scheines, Pearl,…)

- **Question**: given $p(X_1,\ldots,X_n)$, can we infer G?
- **Answer**:
  - Impossible without additional information
  - Possible with *interventions* *(subject to philosophical debates re. counterfactuals)*
  - Assuming *faithfulness* (and suitable assumptions to make conditional independence testing feasible), can estimate a *Markov equivalence class* containing G by running conditional independence tests (PC, FCI, …).

Outlook: Causal representation learning

Problem: SCMs are usually at the 'symbolic' level; they assume the causal variables are given.

Goal: embed an SCM into a deep learning model whose inputs and outputs are high-dimensional and unstructured.

Observation: natural connection between SCMs and deep generative models: both use the *reparametrization trick* (Kingma & Welling, 2013), i.e., they make randomness an (exogenous) input to the model rather than an intrinsic component.

Idea: realize the U_i as (latent) noise variables in a generative model.

Causal mechanisms in machine learning

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$

- causal conditionals are robust
  (Schölkopf et al., ICML 2012)
- use them for causal inference in multi-task settings *(Zhang et al., ICML 2013 & AAAI 2015, Rojas-Carulla et al., JMLR 2018)*
- related: transportability *(Bareinboim & Pearl, 2012)*
- semi-supervised learning *(Schölkopf et al., ICML 2012)*
- generative models *(Parascandolo et al., Besserve et al., Locatello et al., Leeb et al., Träuble et al., Goyal et al., ...)*

End of Part 1

Inferring the DAG

- Key postulate: Causal Markov condition

- Essential mathematical concept: d-separation
  (describes the conditional independences required by a causal DAG)

d-separation (Pearl 1988)

Path = sequence of pairwise distinct nodes where consecutive ones are adjacent

A path q is said to be **blocked** by the set Z if

- q contains a *chain* $i \rightarrow m \rightarrow j$ or a *fork* $i \leftarrow m \rightarrow j$ such that the middle node is in Z, or
- q contains a *collider* $i \rightarrow m \leftarrow j$ such that the middle node is not in Z and such that no descendant of m is in Z.

Z is said to **d-separate** X and Y in the DAG G, formally

$$(X \perp\!\!\!\perp Y \,|Z)_G$$

if Z blocks every path from a node in X to a node in Y.

Example (blocking of paths)

path from X to Y is blocked by conditioning on U or Z or both

Example (unblocking of paths)

- path from X to Y is blocked by $\emptyset$
- unblocked by conditioning on Z or W or both

Correlation by conditioning on common effects

Berkson's paradox (1946)
Example: X, Y, Z binary

$$X \perp\!\!\!\perp Y \quad \text{but} \quad X \not\perp\!\!\!\perp Y \,|\, Z$$

- assume: there is no correlation between getting a vaccination (X) and being naturally immune (Y)
- a person survives if he/she is vaccinated or immune
- among the survivors, there is a *negative* correlation between being naturally immune and having received the vaccination

Asymmetry under inverting arrows

(Reichenbach 1956)

$X \perp\!\!\!\perp Y$ $\qquad$ $X \not\perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \,|Z$ $\qquad$ $X \perp\!\!\!\perp Y \,|Z$

Examples (d-separation)

$$(X \perp\!\!\!\perp Y \,|\, ZW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, ZUW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, VZUW)_G$$

$$(X \not\perp\!\!\!\perp Y \,|\, VZU)_G$$

Causal inference for time-ordered variables

assume $X \not\perp\!\!\!\perp Y$ and X earlier. Then $X \leftarrow Y$ excluded, but still two options:

Example (Fukumizu 2007): barometer falls before it rains, but it does not cause the rain

Conclusion: time order makes causal problem (slightly?) easier but does not solve it

Time series and Granger causality

Does X cause Y and/or Y cause X?

exclude instantaeous effects and common causes

- if

$$Y_{present} \not\perp X_{past} \,|Y_{past}$$

  there must be arrows from X to Y (otherwise d-separation)

- Granger (1969): the past of X helps when predicting Y_t from its past

- strength of causal influence often measured by transfer entropy

$$I(Y_{present}; X_{past} \,|Y_{past})$$

Confounded Granger

Hidden common cause Z relates X and Y

due to different time delays we have

$$Y_{present} \not\perp\!\!\!\perp X_{past} \,|Y_{past}$$

but

$$X_{present} \perp\!\!\!\perp Y_{past} \,|X_{past}$$

Granger infers $X \rightarrow Y$

Recent work on finding causation in confounded time series: Mastakouri et al., 2020 https://arxiv.org/abs/2005.08543

Why transfer entropy does not quantify causal strength (Ay & Polani, 2008)

deterministic mutual influence between X and Y

- although the influence is strong

$$I(Y_{present}; X_{past} \,|Y_{past}) = 0\,,$$

  because the past of Y already determines its present

- quantitatively still wrong for non-deterministic relation

- see paper on definitions of causal strength: Janzing, Balduzzi, Grosse-Wentrup, Schölkopf, *Annals of Statistics* 2013

Quantifying causal influence for general DAGs

Given:
causally sufficient set of variables $X_1, \ldots, X_n$ with

- known causal DAG G
- known joint distribution $P(X_1, \ldots, X_n)$

Goal:
construct a measure that quantifies the strength of $X_i \rightarrow X_j$
with the following properties:

Postulate 1: (mutual information)

For this simple DAG we postulate

$$c_{X \to Y} = I(X;Y)$$

(no other path from X to Y, hence the dependence is caused by the arrow $X \to Y$)

Postulate 2: (locality)

causes of causes and effects of effects don't matter

here we also postulate $c_{X \to Y} = I(X;Y)$

Postulate 3: (strength majorizes conditional dependence, given the other parents)

$$c_{X \to Y} \geq I(X; Y \,|Z)$$

(without $X \to Y$ the Markov condition would imply $I(X; Y \,|Z) = 0$)

Why $c_{X \to Y} = I(X; Y \,|Z)$ is a bad idea

contains

as a limiting case
(weak influence $Z \to Y$),

where we postulated $c_{X \to Y} = I(X; Y)$ instead of $I(X; Y \,|Z)$

Our approach: "edge deletion"

- define a new distribution

$$P_{X\to Y}(x,y,z) = P(z)P(x|z)\sum_{x'} P(y|x',z)P(x')$$

- define causal strength by the 'impact of edge deletion'

$$c_{X\to Y} := D(P\|P_{X\to Y})$$

- intuition of edge deletion:

  cut the wire between devices and feed the open end with an iid copy of the original signal

related work:
Ay & Krakauer (2007)

Properties of our measure

- strength also defined for set of edges
- satisfies all our postulates
- also applicable to time series
- conceptually more reasonable than Granger causality and transfer entropy

Inferring the causal DAG without time information

- Setting: given observed n-tuples drawn from $p(X_1, \ldots, X_n)$, infer G
- Key postulates: Causal Markov condition and causal faithfulness

Causal faithfulness

Spirtes, Glymour, Scheines

p is called faithful relative to G if only those independences hold true that are implied by the Markov condition, i.e.,

$$(X \perp\!\!\!\perp Y \,|Z)_G \quad \Leftarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$$

Recall: Markov condition reads

$$(X \perp\!\!\!\perp Y \,|Z)_G \quad \Rightarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p$$

Examples of unfaithful distributions (1)

Cancellation of direct and indirect influence in linear models

$$
\begin{aligned}
X &= U_X \\
Y &= \alpha X + U_Y \\
Z &= \beta X + \gamma Y + U_Z
\end{aligned}
$$

with independent noise terms U_X, U_Y, U_Z

$$\beta + \alpha\gamma = 0 \quad \Rightarrow \quad X \perp\!\!\!\perp Z$$

Examples of unfaithful distributions (2)

binary causes with XOR as effect

- for $p(X), p(Y)$ uniform: $X \perp\!\!\!\perp Z$, $Y \perp\!\!\!\perp Z$.
  i.e., unfaithful (since X, Z and Y, Z are connected in the graph).
- for $p(X), p(Y)$ non-uniform: $X \not\perp\!\!\!\perp Z$, $Y \not\perp\!\!\!\perp Z$.
  i.e., faithful

unfaithfulness considered unlikely because it only occurs for non-generic parameter values

Conditional-independence based causal inference

Spirtes, Glymour, Scheines and Pearl

Causal Markov condition + Causal faithfulness:

- accept only those DAGs G as causal hypotheses for which

$$(X \perp\!\!\!\perp Y \,|Z)_G \quad \Leftrightarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p \,.$$

- identifies causal DAG up to Markov equivalence class
  (DAGs that imply the same conditional independences)

Markov equivalence class

Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same v-structures.

skeleton: corresponding undirected graph
v-structure: substructure $X \rightarrow Y \leftarrow Z$ with no edge between X and Z

Markov equivalent DAGs

same skeleton, no v-structure

$$X \perp\!\!\!\perp Z \,|\, Y$$

Markov equivalent DAGs

same skeleton, same v-structure at W

Algorithmic construction of causal hypotheses

IC algorithm by Verma & Pearl (1990) to reconstruct DAG from p

idea:

1. Construct skeleton
2. Find v-structures
3. direct further edges that follow from
    - graph is acyclic
    - all v-structures have been found in 2)

Construct skeleton

Theorem: X and Y are linked by an edge iff there is no set S_{XY} such that

$$(X \perp\!\!\!\perp Y \,|S_{XY}\,.$$

(assuming Markov condition and Faithfulness)

Explanation: dependence mediated by other variables can be screened off by conditioning on an appropriate set

$$X \perp\!\!\!\perp Y \,|\{Z, W\}$$

...but not by conditioning on all other variables!

S_{XY} is called a Sepset for (X, Y)

Efficient construction of skeleton

PC algorithm by Spirtes & Glymour (1991)

iteration over size of Sepset

1. remove all edges $X - Y$ with $X \perp\!\!\!\perp Y$

2. remove all edges $X - Y$ for which there is a neighbor $Z \neq Y$ of X with $X \perp\!\!\!\perp Y \,|Z$

3. remove all edges $X - Y$ for which there are two neighbors $Z_1, Z_2 \neq Y$ of X with $X \perp\!\!\!\perp Y \,|Z_1, Z_2$

4. ...

Advantages

- many edges can be removed already for small sets
- testing all sets S_{XY} containing the adjacencies of X is sufficient
- depending on sparseness, algorithm only requires independence tests with small conditioning tests
- polynomial for graphs of bounded degree

Find v-structures

- given $X - Y - Z$ with X and Y non-adjacent
- given S_{XY} with $X \perp\!\!\!\perp Y \,|S_{XY}$

a priori, there are 4 possible orientations:

$$\left.\begin{array}{l} X \to Z \to Y \\ X \leftarrow Z \to Y \\ X \leftarrow Z \leftarrow Y \end{array}\right\} \qquad Z \in S_{XY}$$

$$X \to Z \leftarrow Y \qquad Z \notin S_{XY}$$

Orientation rule: create v-structure if $Z \notin S_{XY}$

Direct further edges (Rule 1)

(otherwise we get a new v-structure)

Direct further edges (Rule 2)

(otherwise one gets a cycle)

Direct further edges (Rule 3)

could not be completed
without creating a cycle
or a new v-structure

Direct further edges (Rule 4)

could not be completed
without creating a cycle
or a new v-structure

Examples

(taken from Spirtes et al, 2010)
true DAG

start with fully connected undirected graph

remove all edges $X - Y$ with $X \perp\!\!\!\perp Y \,|\emptyset$

$$X \perp\!\!\!\perp W \quad Y \perp\!\!\!\perp W$$

remove all edges having Sepset of size 1

$$X \perp\!\!\!\perp Z \,|Y \quad X \perp\!\!\!\perp U \,|Y \quad Y \perp\!\!\!\perp U \,|Z \quad W \perp\!\!\!\perp U \,|Z$$

find v-structure

$Z \notin S_{YW}$

orient further edges (no further v-structure)

edge $X - Y$ remains undirected

Improvements

- CPC (conservative PC) by Ramsey, Zhang, Spirtes (1995) uses weaker form of faithfulness

- FCI (fast causal inference) by Spirtes, Glymour, Scheines (1993) and Spirtes, Meek, Richardson (1999) infers causal links in the presence of latent common causes

- for implementations of the algorithms see homepage of the TETRAD project at Carnegie Mellon University Pittsburgh

Equivalence of Markov conditions

Theorem: the following are equivalent:

- Existence of a structural causal model
- Local Causal Markov condition: X_j *statistically independent* of non-descendants, given parents
- Global Causal Markov condition: d-separation
- Factorization $p(X_1, \ldots, X_n) = \prod_j p\left(X_j \mid PA_j\right)$

(subject to technical conditions)

Local Markov $\Rightarrow$ factorization *(Lauritzen 1996)*

- Assume X_n is a terminal node, i.e., it has no descendants, then $ND_n = \{X_1, \ldots, X_{n-1}\}$. Thus the local Markov condition implies

$$X_n \perp\!\!\!\perp \{X_1, \ldots, X_{n-1}\} \,|PA_n \,.$$

- Hence the general decomposition

$$p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_1, \ldots, x_{n-1})$$

  becomes

$$p(x_1, \ldots, x_n) = p(x_n|pa_n)p(x_1, \ldots, x_{n-1}) \,.$$

- Induction over n yields

$$p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j|pa_j) \,.$$

Factorization $\Rightarrow$ global Markov

(Lauritzen 1996)

Need to prove $(X \perp\!\!\!\perp Y \,|Z)_G \Rightarrow (X \perp\!\!\!\perp Y \,|Z)_p$.

Assume $(X \perp\!\!\!\perp Y \,|Z)_G$

- define the smallest subgraph G' containing X, Y, Z and all their ancestors
- consider moral graph G'^m (undirected graph containing the edges of G' and links between all parents)
- use results that relate factorization of probabilities with separation in undirected graphs

Global Markov $\Rightarrow$ local Markov

Know that if Z d-separates X, Y, then $X \perp\!\!\!\perp Y \,|Z$.
Need to show that $X_j \perp\!\!\!\perp ND_j \,|PA_j$.

Simply need to show that the parents PA_j d-separate X_j from its non-descendants ND_j:

All paths connecting X_j and ND_j include a $P \in PA_j$, but never as a collider

$$\cdot \rightarrow P \leftarrow X_j$$

Hence all paths are chains

$$\cdot \rightarrow P \rightarrow X_j$$

or forks

$$\cdot \leftarrow P \rightarrow X_j$$

Therefore, the parents block every path between X_j and ND_j.

structural causal model $\Rightarrow$ local Markov condition

(Pearl 2000)

- augmented DAG G' contains unobserved noise
- local Markov-condition holds for G':

  (i): the unexplained noise terms U_j are jointly independent, and thus (unconditionally) independent of their non-descendants

  (ii): for the X_j, we have
  $$X_j \perp\!\!\!\perp ND'_j \,|PA'_j$$
  because X_j is a (deterministic) function of PA'_j.
- local Markov in G' implies global Markov in G'
- global Markov in G' implies local Markov in G (proof as previous slide)

factorization $\Rightarrow$ structural causal model

generate each $p(X_j|PA_j)$ in

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j|PA_j)$$

by a deterministic function:

- define a vector valued noise variable U_j
- each component $U_j[pa_j]$ corresponds to a possible value pa_j of PA_j
- define structural equation

$$x_j = f_j(pa_j, u_j) := u_j[pa_j]\,.$$

- let component $U_j[pa_j]$ be distributed according to $p(X_j|pa_j)$.

Note: joint distribution of all $U_j[pa_j]$ is irrelevant, only marginals matter

different point of view

=g(X), U chooses g ∈ G

- G denotes set of deterministic mechanisms
- U randomly chooses a mechanism

Example: X, Y binary

$= g(X), \quad U$ chooses $g \in \{ID, NOT, 1, 0\}$

the same $p(X, Y)$ can be induced by different distributions on G:

- model 1 (no causal link from X to Y)

$$P(g = 0) = 1/2, \quad P(g = 1) = 1/2$$

- model 2 (random switching between ID and NOT)

$$P(g = ID) = 1/2, \quad P(g = NOT) = 1/2$$

both induce the uniform distribution for Y, *independent* of X

- **Disentangled (causal) factorization**

$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$

  according to the causal graph: **independent noises and conditionals**

- a change in a distribution always comes from a sparse change in causal conditionals / mechanisms (i.e., structural assignment/function and/or noise variable)

- changing one $p\left(X_i \mid \mathrm{PA}_i\right)$ does not change the other $p\left(X_j \mid \mathrm{PA}_j\right)$ ($j \neq i$); they remain **invariant**

- **Entangled (non-causal) factorizations**: e.g.,

$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p(X_i \mid X_{i+1}, \ldots, X_n).$$

  Here, changes will **not** be local.

Cf. *independence of mechanisms (Janzing & Schölkopf, IEEE Trans. Inf. Th. 2010; Schölkopf et al., ICML 2012)*, *independence of cause and mechanism (Janzing et al., 2012)*, *autonomy*, *(structural) invariance*, *separability*, *exogeneity*, *stability*, *modularity (Aldrich, 1989; Pearl, 2009)*

Disentanglement methods: VAE + Regularizer

Beta-VAE, (Higgins et al., 2017);
Fix capacity of VAE bottleneck.
Annealed-VAE, (Burgess et al., 2017);
Progressively increase capacity of VAE bottleneck.
Factor-VAE, (Kim & Mnih, 2018);
Penalize Total Correlation with adversarial training.
Beta-TCVAE, (Chen et al., 2018);
Penalize Total Correlation with Monte Carlo estimate.
DIP-VAE I and **II**, (Kumar et al., 2018)
Match moments with a disentangled prior.

Regularization encourages factorizing representations

- in general **not possible in an unsupervised way** (Locatello, Bauer et al. - ICML 2019)
- mostly studied within highly controlled and unrealistic settings (**synthetic** datasets and **perfectly independent** factors of variation)

Weakly-Supervised Disentanglement

F. Locatello, B. Poole, G. Rätsch, B.Schölkopf, O. Bachem, M. Tschannen

arXiv:2002.02886 (submitted to ICML'20)

Proposed algorithms:

- Ada-ML-VAE
- Ada-GVAE

Theoretical result

Under some assumptions, disentangled representations can be learned and identified from weak supervision.

Practical aspects

Model selection based on the weakly-supervised loss, without relying on supervised disentanglement metrics

Disentanglement Challenge: From Simulation to Real World

(S. Bauer, M. Wüthrich, …)

http://www.disentanglement-challenge.com/

- 2 stages
  - Stage 1: Simulated-to-real image transfer learning
  - Stage 2: Transfer to unseen objects
- Low-to-high image resolution (each over 1m images)
- The competition spanned across approx. 3 months
- Approximately 200 participants

Low-quality rendering

High-quality rendering

Real-life

Disentangling correlated factors is nontrivial *(F. Träuble,…)*

Model with highest DCI score
and line width 20%

Mutual information

object size and azimuth has been correlated
in shapes3d

Causal representation learning

Problem: SCMs are usually at the 'symbolic' level; they assume the causal variables are given.

Goal: embed an SCM into a deep learning model whose inputs and outputs are high-dimensional and unstructured.

Observation: natural connection between SCMs and deep generative models: both use the *reparametrization trick* (Kingma & Welling, 2013), i.e., they make randomness an (exogenous) input to the model rather than an intrinsic component.

Idea: realize the U_i as (latent) noise variables in a generative model.

Given (high-dimensional) $X = (X_1, \ldots, X_d)$ (think of X as an image with pixels $X_1, \ldots, X_d$), construct causal variables $S_1, \ldots, S_n$ ($n \ll d$) and mechanisms

$$S_i := f_i(\mathbf{PA}_i, U_i), \quad (i = 1, \ldots, n)$$

such that we get a **disentangled representation**

$$p(S_1, \ldots, S_n) = \prod_{i=1}^{n} p(S_i \mid \mathbf{PA}_i)$$

with $p(S_i \mid \mathbf{PA}_i)$ independently manipulable and largely invariant across related problems.

Step 1: (anticausal) **encoder** $q : \mathbb{R}^d \to \mathbb{R}^n$ taking X to a latent representation $U = (U_1, \ldots, U_n)$

Step 2: **structural causal model** $f(U)$ determined by the assignments/mechanisms $f_1, \ldots, f_n$ (possibly recurrent)

Step 3: (causal) **decoder** $p : \mathbb{R}^n \to \mathbb{R}^d$.

Embedding training: $p \circ f \circ q \approx id$ on the observed images (reconstruction error)

Causal training: make the U_i statistically independent, and make the mechanisms f_i independently manipulable **(ICM Principle)** by requiring that they produce valid images after interventions.

Branched Decoders

Felix Leeb

➔ ~200-700k parameters

Quantitative Results

Disentanglement by Architecture

12-Branched
AE

Rec

Gen (hyb)

Key Insights

The SCM architecture of the decoder seems to facilitate discovering the structure of the underlying mechanisms can be discovered

The Branched Decoder learns to **distinguish** and **order mechanisms** from more abstract higher level features (orientation, shape, size) to the low level ones (object color, background colors)

Learning independent mechanisms

(with Parascandolo, Kilbertus, Rojas-Carulla, ICML 2018)

- Data drawn from $p(x)$, transformed by M mechanisms $f_1, \ldots, f_M$
- Goal: learn the independent mechanisms / factors of variation
- Method: generative model with competing mechanisms

Original data

Transformed data

Method

- Mechanisms initialized ≈ identity
- The highest scoring mechanism against the discriminator D wins the example and is updated to increase the score
- D is trained on the original data and against the winning outputs

$$\max_{\theta_D}\left(\mathbb{E}_{x\sim P}\log(D_{\theta_D}(x)) + \frac{1}{N'}\sum_{j=1}^{N'}\mathbb{E}_{x'\sim Q}\left(\log(1-D_{\theta_D}(E_{\theta_j}(x')))\right)\right)$$

Training progress

randomly sampled inputs

Accuracy of a CNN trained on MNIST for different test sets

Generalizing to Omniglot characters

Recurrent Independent Mechanisms

with **Anirudh Goyal**,
Alex Lamb,
Jordan Hoffmann,
Shagun Sodhani,
Sergey Levine,
Yoshua Bengio

A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, 2019. Recurrent independent mechanisms. arXiv:1909.10893.

Interventional Representations *(Besserve et al., ICLR 2020)*

MAX-PLANCK-GESELLSCHAFT

Interventional Representations *(Besserve et al., ICLR 2020)*

Interventional Representations *(Besserve et al., ICLR 2020)*

MAX-PLANCK-GESELLSCHAFT

Causality and fairness

gender-fair mapping

{ , ✗ }

suitable for the job?

"unawareness"

other features ↓ gender

"demographic parity"

$\boxed{\text{decision} \perp\!\!\!\perp \text{gender}}$

Too weak & too strong...

"Equalized odds/equal opportunity"

$\boxed{\text{decision} \perp\!\!\!\perp \text{gender} \mid \text{suitability}}$

Is the *observed suitability* biased?

Dwork et al. 2011, Hardt et al. 2016, Berk et al. 2017, Kleinberg et al. 2016, Kamiran, Calders, …

Removing proxy discrimination

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, D. Janzing, M. Hardt, B. Schölkopf, NIPS 2017

control this to cancel "unfair" paths
(if the influence of proxy on skills is additive & linear)

Criterion:

$$\mathbb{P}(\text{decision} \,|\, do(\text{proxy} = p_1)) = \mathbb{P}(\text{decision} \,|\, do(\text{proxy} = p_2))$$

Pearl 2009, VanderWeele & Robinson 2014, Bonchi et al. 2015, Qureshi et al. 2016, Kusner et al. 2017, Zhang & Wu 2017, Nabi & Shpitser 2017

Towards causal world models

$$p(X_1, \ldots, X_n) = \prod_{I=1}^{n} p\left(X_i \mid \mathrm{PA}_i\right)$$

- learn a multi-task/environment model
  - data from multiple tasks in multiple environments
  - re-usable components that are robust across tasks, i.e., causal (independent) mechanisms
  - representation learning should move towards representations of causal world models: *"thinking is acting is an imagined space"* (Konrad Lorenz)

arXiv:1911.10500v1 [cs.LG] 24 Nov 2019

Causality for Machine Learning

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany
bs@tuebingen.mpg.de

Abstract

Graphical causal inference as pioneered by Judea Pearl arose from research on artificial intelligence (AI), and for a long time had little connection to the field of machine learning. This article discusses where links have been and should be established, introducing key concepts along the way. It argues that the hard open problems of machine learning and AI are intrinsically related to causality, and explains how the field is beginning to understand them.

1 Introduction

The machine learning community's interest in causality has significantly increased in recent years. My understanding of causality has been shaped by Judea Pearl and a number of collaborators and colleagues, and much of it went into a book written with Dominik Janzing and Jonas Peters (Peters et al., 2017). I have spoken about this topic on various occasions,[1] and some of it is in the process of entering the machine learning mainstream, in particular the view that causal modeling can lead to more invariant or robust models. There is excitement about developments at the interface of causality and machine learning, and the present article tries to put my thoughts into writing and draw a bigger picture. I hope it may not only be useful by discussing the importance of causal thinking for AI, but it can also serve as an introduction to some relevant concepts of graphical or structural causal models for a machine learning audience.

In spite of all recent successes, if we compare what machine learning can do to what animals accomplish, we observe that the former is rather bad at some crucial feats where animals excel. This includes transfer to new problems, and any form of generalization that is not from one data point to the next one (sampled from the same distribution), but rather from one problem to the next one —— both have been termed *generalization*, but the latter is a much harder form thereof. This shortcoming is not too surprizing, since machine learning often disregards information that animals use heavily: interventions in the world, domain shifts, temporal structure —- by and large, we consider these factors a nuisance and try to engineer them away. Finally, machine learning is also bad at *thinking* in the sense of Konrad Lorenz, i.e., acting in an imagined space. I will argue that causality, with its focus on modeling and reasoning about interventions, can make a substantial contribution towards understanding and resolving these issues and thus take the field to the next level. I will do so mostly in non-technical language, for many of the difficulties of this field are of a conceptual nature.

Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

Thank You