Systems with Switching Causal Relations: A Meta-Causal Perspective

Moritz Willig[1], **Tim Nelson Tobiasch**[1], **Florian Peter Busch**[1,2], **Jonas Seng**[1], **Devendra Singh Dhami**[3], **Kristian Kersting**[1,2,4,5]

[1] Department of Computer Science, Technical University of Darmstadt
[2] Hessian Center for AI (hessian.AI)
[3] Dept. of Mathematics and Computer Science, Eindhoven University of Technology
[4] Centre for Cognitive Science, Technical University of Darmstadt
[5] German Research Center for AI (DFKI)
Corresponding author: moritz.willig@cs.tu-darmstadt.de

Abstract

Most work on causality in machine learning assumes that causal relationships are driven by a constant underlying process. However, the flexibility of agents' actions or tipping points in the environmental process can change the qualitative dynamics of the system. As a result, new causal relationships may emerge, while existing ones change or disappear, resulting in an altered causal graph. To analyze these qualitative changes on the causal graph, we propose the concept of *meta-causal states*, which groups classical causal models into clusters based on equivalent qualitative behavior and consolidates specific mechanism parameterizations. We demonstrate how meta-causal states can be inferred from observed agent behavior, and discuss potential methods for disentangling these states from unlabeled data. Finally, we direct our analysis towards the application of a dynamical system, showing that meta-causal states can also emerge from inherent system dynamics, and thus constitute more than a context-dependent framework in which mechanisms emerge only as a result of external factors.

1 Introduction

Causal graphs can be subject to change whenever novel mechanisms emerge or vanish within a system. Prominently, agents can 'break' the natural unfolding of systems dynamics by forecasting system behavior and preemptively intervening in the course of events. As inherent parts of the environment, agents commonly establish or suppress the emergence of causal connections (Zhang & Bareinboim, 2017; Lee & Bareinboim, 2018; Dasgupta et al., 2019).

Consider the scenario shown in Figure 1 (left), where an agent A (with position A_X) follows an agent B (with position B_X) according to its internal policy $A_{\boldsymbol{\pi}}$. We are interested in answering the question 'What is the cause of the current position of agent A?'. In general, the observed system can be formalized as follows: $A_X := f_A(B_X, U_A)$ and $B_X := f_B(U_B)$. Note that, from a classical causal perspective, we observe $B \to A$, since A self-conditions itself to follow B, by instantiating the equation f_A via its policy and thus becomes *dependent* on B. Classical causal considerations, which only consider how relations between variables are constructed, cannot give the correct answer. Only when we take a meta-causal stance and think about *how the equation f_A came to be* in the first place, can we give a sufficient answer to this question.

Contributions. To the best of our knowledge, we are the first to formally introduce typing mechanisms that generalize edges in causal graphs, and the first to provide a formalization of meta-causal

38th Conference on Neural Information Processing Systems (NeurIPS 2024) – Causal Representation Learning Workshop.

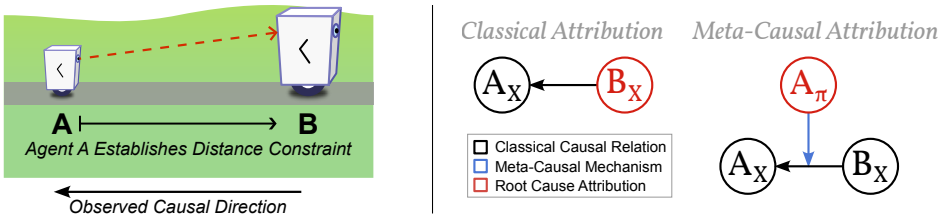

Figure 1: **Meta-Causality Identifies the Policy as a Meta Root Cause.** Agent A intends to maintain its distance from agent B by conditioning its position A_X on the position B_X, which establishes a control mechanism, $A_X := f(B_X)$. In standard causal inference, we would infer $B_X \to A_X$ and, therefore, B to be the root cause. Taking a meta-causal perspective reveals however, that $A_{\boldsymbol{\pi}}$ establishes the edge $B_X \to A_X$ in the first place ($A_{\boldsymbol{\pi}} \to (B_X \to A_X)$) such that $A_{\boldsymbol{\pi}}$ is considered the root cause on the meta-level. *(Best Viewed in Color)*

models (MCM) that are capable of capturing the switching type dynamics of causal mechanisms. Furthermore, we present an approach to discover the number of meta-causal states in the bivariate case. Finally, demonstrate that meta-causal models can be more powerful than classical SCM when it comes to expressing qualitative differences within certain system dynamics.

2 Background

Providing a higher-level perspective on meta-causality touches on a number of existing works that leverage meta-causal ideas, even if not explicitly stated. We will highlight relations of these works in the 'Related Work' of section 5. Here, we continue to provide the necessary concepts on causality, mediation processes and typing, needed for the definitions in our paper:

Causal Models. A common formalization of causality is provided via Structural Causal Models (SCM; (Spirtes et al., 2000; Pearl, 2009)). An SCM is defined as a tuple $\mathcal{M} := (\mathbf{U}, \mathbf{V}, \mathbf{F}, P_{\mathbf{U}})$, where $\mathbf{U}$ is the set of exogenous variables, $\mathbf{V}$ is the set of endogenous variables, $\mathbf{F}$ is the set of structural equations determining endogenous variables, and $P_{\mathbf{U}}$ is the distribution over exogenous variables $\mathbf{U}$. Every endogenous variable $V_i \in \mathbf{V}$ is determined via a structural equation $v_i := f_i(\text{pa}(v_i))$ that takes in a set of parent values $\text{pa}(v_i)$, consisting of endogenous and exogenous variables, and outputs the value of v_i. The set of all variables is denoted by $\mathbf{X} = \mathbf{U} \cup \mathbf{V}$ with values $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ and $N = |\mathbf{X}|$. Every SCM $\mathcal{M}$ entails a directed graph $\mathcal{G}^{\mathcal{M}}$ that can be constructed by adding edges $(i, j) \in \mathbf{X} \times \mathbf{X}$ for all variables $X_j \in \mathbf{X}$ and its parents $X_i \in \text{Pa}(X_j)$. This can be expressed as an adjacency matrix $A \in \mathbb{B}^{N \times N}$ where $A_{ij} := 1$ if $(i, j) \in \mathcal{G}$ and $A_{ij} := 0$ otherwise.

Mediation Processes. Causal effects are embedded in an environment that governs the dynamics and mediates between different causes and effects. We define a *mediation process* $\mathcal{E} = (\mathcal{S}, \sigma)$, adapted from Markov Decision Processes (MDP; Bellman (1957)) for our setting. Here, $\mathcal{S}$ is the state space of the environment and $\sigma : \mathcal{S} \to \mathcal{S}$ is the (possibly non-deterministic) transition function that takes the current state and outputs the next one. If we also have an initial state $\text{s}_0 \in \mathcal{S}$, we call $(\mathcal{E}, \text{s}_0)$ an *initialized mediation process*.

Typing. In the following section, we will make use of an identification function $\mathcal{I}$ to determine the presence or absence of edges between any two variables and in particular identify different *types* of edges. Previous work on typing causality exists (Brouillard et al., 2022), but it primarily concerned with types of variables. Furthermore, several other works in the field of cognitive science consider the perception of different types of mechanistic relations (e.g., 'causing', 'preventing', ...) (Chockler & Halpern, 2004; Wolff, 2007; Sloman et al., 2009; Walsh & Sloman, 2011; Gerstenberg, 2022).

3 Meta-Causality

Meta-Causality is concerned with the general mechanisms that lead to the emergence of causal relations. Usually, the mediating process may be too fine-grained to yield interpretable models. Therefore, we consider a set of variables of interest $\mathbf{X}$ derived via $\varphi : \mathcal{S} \to \boldsymbol{\mathcal{X}}$. Here, φ could be defined as a summarization or abstraction operation over the state space (Rubenstein et al., 2017; Beckers & Halpern, 2019; Anand et al., 2022; Wahl et al., 2023; Kekić et al., 2023; Willig et al.,

2023). In order to identify the type of causal relations from a mediating process, we need to be able to decide on what constitutes such a type.

Definition 1 (Meta-Causal Frame) *For a given mediation process $\mathcal{E} = (\mathcal{S}, \sigma)$ a **meta-causal frame** is a tuple $\mathcal{F} = (\mathcal{E}, \mathbf{X}, \tau, \mathcal{I})$ with:*

- *a type-encoder $\tau : \mathcal{X}_i \times \mathcal{X}_j^{\mathcal{S}} \to \mathcal{T}$ that determines the relationship between the value of a causal variable $\mathcal{X}_i$ and the change of another causal variable $\mathcal{X}_j^{\mathcal{S}} = \{\psi : \mathcal{S} \to \mathcal{X}_j\}$ with respect to a environmental state space $\mathcal{S}$ which we call a **type** $t \in \mathcal{T}$,*
- *an identification function $\mathcal{I} : \mathcal{S} \times \mathbf{X} \times \mathbf{X} \to \mathcal{T}$ with $\mathcal{I}(s, X_i, X_j) \mapsto t := \tau(\varphi(s)_i, \varphi_j \circ \sigma)$ that takes a state of the environment and assigns a type for every pair of causal variables $X_i, X_j \in \mathbf{X}$.*

Types generalize the role of edges in causal graphs. In most classical scenarios, the co-domain $\mathcal{T}$ of the type encoder τ is chosen to be Boolean, representing the *existence* or *absence* of edges. In other cases, different values $t \in \mathcal{T}$ can be understood as particular types of edges, like positive, negative, or the absence of influence. This will help us to distinguish meta-causal states that share the same graph adjacency. The only requirement for $\mathcal{T}$ is that it must contain a special value 0, which indicates the total absence of an edge. *Meta-causal states* now generalize the idea of binary adjacency matrices.

Definition 2 (Meta-Causal State) *In a meta-causal frame $\mathcal{F} = (\mathcal{E}, \mathbf{X}, \tau, \mathcal{I})$, a **meta-causal state** is a matrix $T \in \mathcal{T}^{N \times N}$. For a given environment state $s \in \mathcal{S}$, the **actual meta-causal state** T_s has the entries $T_{s,ij} := \mathcal{I}(s, X_i, X_j) = \tau(\varphi(s)_i, \varphi_j \circ \sigma)$.*

A meta-causal state $T \in \mathcal{T}^{N \times N}$ represents a graph containing edges e_{ij} of a particular type $t \in \mathcal{T}$. In particular T_{ij} indicates the presence ($T_{ij} \neq 0$) or absence ($T_{ij} = 0$) of individual edges. Our goal for meta-causal models (MCM) is to capture the dynamics of the underlying model. In particular, we are interested in modeling how different meta-causal states transition into each other. This behavior allows us to model them in a finite-state machine (Moore et al., 1956):

Definition 3 (Meta-Causal Model) *For a meta-causal frame $\mathcal{F} = (\mathcal{E}, \mathbf{X}, \sigma, \mathcal{I})$, a **meta-causal model** is a finite-state machine defined as a tuple $\mathcal{A} = (\mathcal{T}^{N \times N}, \mathcal{S}, \delta)$, where the set of meta-causal states $\mathcal{T}^{N \times N}$ is the set of machine states, the set of environment states $\mathcal{S}$ is the input alphabet, and $\delta : \mathcal{T}^{N \times N} \times \mathcal{S} \to \mathcal{T}^{N \times N}$ is a transition function.*

Usually, we have the objective to learn the transition function δ for an unknown state transition function σ. The state transition function δ can be approximated as $\delta(T, s) := \mathcal{I}(\sigma'(T, s), X_i, X_j) = \tau(\varphi(s)_i, \varphi_j \circ \sigma'(T, s))$. As standard causal relations emerge from the underlying mediation process, the meta-causal states emerge from different types of causal effects. The transition conditions of the finite-state machine are the configurations of the environment where the quality of some environmental dynamics represented by a type $t \in \mathcal{T}$ changes.

Relation to Contextual Independencies. Changes in the causal graph are often times attributed to changes in the environment, as for example prominently leveraged in invariant causal prediction of Peters et al. (2016); Heinze-Deml et al. (2018). In general, the effects of a changing causal graph can be attributed to the change in some external factor acting on the system. Under certain circumstances, we can indeed represent meta-causal models via standard SCM conditioned on some external factor $Z \in \mathcal{Z}$. Suppose there is a surjective function $\psi : \mathcal{Z} \to \mathcal{T}^{N \times N}$ that can map from the Z to every meta causal state T. As a result, one can define a new set of structural equations $f'_i := \bigcirc_{i \in 1..N} f'_{ij}$ and

$$f_{ij}|Z := \begin{cases} f_{ij}^{\psi(z)} & \textbf{if } \psi(Z)_{ij} \neq 0 \\ (< *) & \textbf{otherwise.} \end{cases}$$

where $f_{ij}^{\psi(z)}$ are the structural equations of the edge e_{ij} that are active under the MCS $T = \psi(z)$ and (in a slight imprecision in the actual definition) $(< *)$ is the function that carries on the previous function of the concatenation and discards X_j. While the 'no-edge' type 0 could be handled like any other type, we have listed it for clarity such that individual variables $\mathbf{X}_i$ become *contextually independent* whenever $\psi(Z)_{ij} = 0$. This mapping can be used under certain circumstances to infer the meta-causal state of a system. We give a brief example in Appendix A

k^* \ k'	$d = 0.0$					$d = 0.1$					$d = 0.2$				
	-	1	2	3	4	-	1	2	3	4	-	1	2	3	4
1	2	81	3	7	7	2	85	3	7	3	1	83	3	8	5
2	41	1	54	4	0	43	1	48	8	0	49	1	47	3	0
3	68	0	4	22	6	63	0	2	30	5	77	2	0	13	8
4	92	0	0	1	7	89	0	0	2	9	87	0	1	5	7

Table 1: **Confusion Matrices for Identifying Meta-Causal Mechanisms.** The table shows identification results for predicting the number of mechanisms for the bivariate case for 100 randomly sampled meta-causal mechanisms. d is the maximum sample deviation from the average mechanism probability. Rows indicate the true number of mechanisms, while columns indicate the algorithms' predictions. '-' indicates setups where the algorithm did not make a decision. In general, the algorithm is rather conservative in its predictions. In all cases where a decision is made, the number of correct predictions along the diagonals dominate. The first and second most frequent predictions are marked in green and orange, respectively. *(Best Viewed in Color)*

4 Applications

In this section, we discuss several applications of the meta-causal formalism. The code used in the experiments is available at `https://anonymous.4open.science/r/metaCausal-CRL/`.

4.1 Attributing Responsibility

Consider again the motivational example of Figure 1, where an agent A with position A_X follows an agent B with position B_X as dictated by its policy A_π. Imagine a counterfactual scenario in which we replace the 'following' policy of agent A with, e.g., a 'standing still' policy and find that the $B_X \to A_X$ edge vanishes. As a result, we infer the meta-causal mechanism $\boldsymbol{\pi}_A \to (B_X \to A_X)$ for the system and thus $\boldsymbol{\pi}_A$ as the root cause of for values of A_X. In conclusion, we find that, while A is conditioned on B on a low level, the meta-causal reason for the existence of the edge $B \to A$ is caused by the $\boldsymbol{\pi}_A$. Both attributions, either tracing back causes through the structural equations $A_X := f(B_X)$ or our meta-causal approach, are valid conclusions in their own regard. This scenario can also considered via classical a counterfactual, $A_X^{\boldsymbol{\pi}_A=\text{standing still}} - A_X^{\boldsymbol{\pi}_A=\text{following}}$, which would infer $\boldsymbol{\pi}_A$ on A_X from a purely value-based perspective, missing to explain the change in the underlying mechanistic relation.

4.2 Discovering Meta-Causal States in the Bivariate Case

Our goal in this experiment is to recover the number of meta-causal states $K \in [1..4]$ from data that exists between two variables X, Y that are directly connected by a linear equation with added noise. We assume that each meta-causal state gives rise to a different linear equation $f_k := \alpha_k X + \beta_k + \mathcal{N}, k \in \mathbb{N}$, where α_k, β_k are the slope and intercept of the respective mechanism and $\mathcal{N}$ is a zero-centered, symmetric, and quasiconvex noise distribution[1]. Without loss of generality, we assume Laplacian noise, for which an L1-regression can estimate the true parameters of the linear equations (Hoyer et al., 2008). The causal direction of the mechanism is randomly chosen between different meta-causal states. The exact sampling parameters and plots of the resulting distributions are given in Sec. E and Fig. 3, respectively. Our goal is to recover the number and parameterizations of the causal mechanisms.

Determining the Number of Mechanisms. Given the correct number of parameters $k*$, we apply a LO-RANSAC (Fischler & Bolles, 1981; Chum et al., 2003) algorithm with a local EM (Dempster et al., 1977) optimization to regress the parameters of the mechanisms. The exact approach is presented in the Appendix C. The approach is able to regress the true parameterization of mechanisms when the real number of parameters k^* is given. However, it is still unclear how to determine the correct k^*. Computing the parameters for all $k \in K$ and comparing for the best goodness of fit is generally a bad indicator for choosing the right k, since fitting more mechanisms usually captures

[1]Implying unimodality and monotonic decreasing from zero allows us to distinguish the noise mean and intercept and to recover the parameters from a simple linear regression

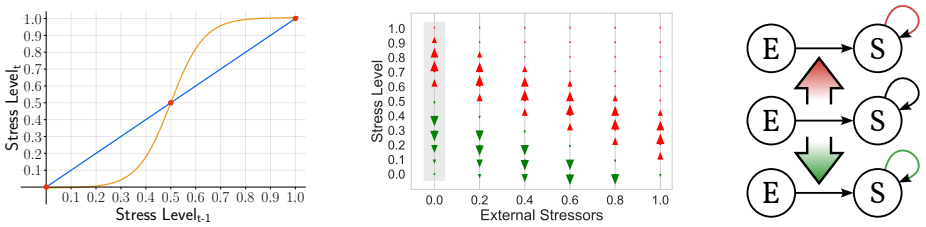

Figure 2: **Mechanistic Decomposition as Meta-Causal States**. **(left)** Effect of the stress level on itself (orange) plotted against the identity (blue). Once a threshold of 0.5 is reached, the function switches behavior from self-suppressing self-reinforcing. **(center)** Contribution of the stress level mechanism for varying external stressors. Red arrows indicate a self-reinforcing effect, while green arrows indicate suppression. The gray area indicates the system without external stressors. **(right)** The standard SCM gets decomposed into different meta-causal states. While the graph adjacency remains the same, the identification function detect the different modes of the causal mechanism.

additional noise and thus reduces the error. We take advantage of the fact that we assumed the noise to be Laplacian distributed. Thus, the residuals of the samples assigned to a particular mechanism can be tested to belong to a Laplacian distribution. Since the assignment of mechanisms to a data point may be ambiguous due to the overlap in the estimated PDFs, we normalize the density values of all mechanisms per data point and consider only those points for which the probability of the dominant mechanism being 0.4 above the second class. Finally, we compute residuals and, make use of the Anderson-Darling test (Anderson & Darling, 1952) to test the residuals against the Laplacian distribution with an $\alpha = 0.95$ (Chen, 2002). If all residual distributions pass the Anderson-Darling test, we choose that number as our predicted number of mechanisms. If the algorithm finds no $k \in [1..4]$ to pass the test, we refrain from making a decision. We provide the pseudo code for our method in Algorithm. 1 in the Appendix.

Evaluation and Results. We evaluate our approach over all $k \in [1..4]$ by generating 100 different datasets for every particular number of mechanisms. For every data set we sample 500 data points from each mechanism $(x_i^k, y_i^k) = f_k(\alpha_k x_i^k + \beta_k + l_i)$ where $l_i \sim \mathcal{L}(0, b_k)$, using the same sampling method described in Appendix E. Finally, we let the algorithm recover the number of mechanisms that generated the data. Table 1 shows the confusion matrices between the actual number of mechanism and the predicted number for different values of maximum class imbalances. In general, we find that our approach is rather conservative when in assigning a number of mechanisms. However, when only considering the cases where the number of mechanisms is assigned, the correct predictions along the main diagonal dominate with over 60% accuracy for $k = 4$ and $d \in \{0.0, 0.1\}$ and rising above 80% for $k \neq 4$ for. In the case of $d = 0.2$, the results indicate higher confusion rates with 41.6% accuracy for the overall worst case of $K = 4$.

Extension to Meta-Causal State Discovery on Graphs. Our results indicate that identifying meta-causal mechanisms even in the bivariate case comes with an increasing number of uncertainty when it comes to increasing numbers of mechanisms. The meta-causal state of a graph could be identified as all unique combinations of mechanisms that are jointly active at a certain point in time. To recover the meta-causal states, one needs to be able to estimate the set of active parents for every mechanism, the mechanisms parameterizations, and the resulting meta-state assignment of all data points. Since, all of the three components can vary between each meta-causal state, the extension to a unsupervised full-fledged meta-causal state discovery is not obvious to us at the time of writing and we leave it to future work to come up with a feasible algorithm.

4.3 A Meta-State Analysis on Stress-Induced Fatigue

Stress is a major cause of fatigue and can lead to other long-term health problems (Maisel et al., 2021; Franklin et al., 2012; Dimsdale, 2008; Bremner, 2006). We present an idealized system that is radically reduced to the only factors of external stress, modeling everyday environmental factors, and the self-influencing level of internal/perceived stress. We use an identification function to distinguish between two different modes of operation of a causal mechanism. In particular, we are interested whether the inherent dynamics of the internal stress level are self-reinforcing or self-suppressing. For easier analysis of the system we decompose the dynamics of the internal stress variable into

a 'decayed stress' d and 'resulting stress' s computation. The first term are the previous stress levels decayed over time with external factors added. The resulting stress is then the output of a Sigmoidal function (Fig. 2, left) that either reinforces or suppresses the value (Fig. 2,center). The structural equations are defined as follows and we assume that all values lie in the interval of $[0, 1]$: $f_d := 0.95 \, \text{clip}_{[0,1]}(s' + 0.5 \times \text{externalStress})$ and $f_s := 1.01(\frac{1}{1+\exp(-15x+7.5)} - 0.5) + 0.5$. where s and d are the resulting and decayed stress levels, and s' is the previous stress level of s. We define the identification function to be $\mathcal{I} := \text{sign}(\ddot{f}_s)$, where $\ddot{f}_s$ is the second order derivative of the Sigmoidal f_s, with either positive or negative effect on the stress level. Note how the second-order inflection point at 0.5 of the Sigmoid acts as a transition point on the behavior of the mechanism. Stress values below 0.5 get suppressed, while values above 0.5 are amplified, which results in three possible meta-causal states, as governed by the following transition function:

$$\sigma : (t, s) \mapsto \begin{bmatrix} 1 & a \\ 0 & 0 \end{bmatrix} \text{ with } a := \text{sign}(\ddot{f}_s) = \text{sign}(s - 0.5) \in \{-1, 0, 1\}$$

Role of Latent Conditioning. A key takeaway of this example is that the current meta-state persists due to the inherent stress level and dynamics of the system, while the same set of underlying structural equations are employed. As a result, the stressed state of a person would persist even when the initiating external stressors disappear. Intervening on the meta-causal state of the system is now ill-defined, as both positive and negative reinforcing effects are governed by the same equation. Thus, creating a disparity between the intervened meta-causal state and the systems' identified functional behavior.

5 Related Work

Meta-causal models cover cases that reduce to conditionally dependent causal graphs due to changing environments (Peters et al., 2016; Heinze-Deml et al., 2018), but also extend beyond that for dynamical systems. In this sense, the work of Talon et al. (2024) takes a meta-causal view by transporting edges of different causal effects between environments. In general, the transportability of causal relations (Zhang & Bareinboim, 2017; Lee & Bareinboim, 2018; Correa et al., 2022) can be thought of as learning identification functions that have identified the general conditions of the underlying processes to transfer certain types of causal effects between environments Sonar et al. (2021); Yu et al. (2024); Brouillard et al. (2022). This has been studied to some extent under the name of meta-reinforcement learning, which attempts to predict causal relations from the observations of environments Sauter et al. (2023); Dasgupta et al. (2019). Generally, transferability has been considered in reinforcement learning, where the efficient use of data is omnipresent and causality provides guarantees regarding the transferability of the observed mechanisms under changing environments (Sæmundsson et al., 2018; Dasgupta et al., 2019; Zeng et al., 2023). Our MCM framework can also be used to infer and attribute responsibility, as shown in Sec. 4.1, and therefore touches on the topics of fairness, *actual causation* (AC) and work on counterfactual reasoning in cognitive science. (Von Kügelgen et al., 2022; Karimi et al., 2021; Halpern, 2000, 2016; Chockler & Halpern, 2004; Gerstenberg et al., 2014; Gerstenberg, 2024). In this sense, MCM allow for the direct characterization of actual causes of system dynamics types, but a rigorous formalization is open for future work.

6 Conclusion

We formally introduced meta-causal models that are able to capture the dynamics of switching causal types of causal graphs and, in many cases, can better express the qualitative behavior of a causal system under consideration. Within MCM typing mechanisms generalize the notion of specific structural equations and abstract away unnecessary details. We presented a first, motivating example of how a classical causal analysis and a meta-causal perspective can attribute different root causes to the same situation. We extended these claims by considering that MCM are still able to capture and attribute changes in the mechanistic behavior of the system, even though no actual change is apparent. As an application, we presented a first approach to recover metacausal states in the bivariate case. Although our experimental results represent only a first preliminary approach, we find that meta-causal models are a powerful tool for modeling, reasoning, and inferring system dynamics. Finally, we have shown how MCM are able to express more information about a dynamical system than a simple conditioning variable in a standard SCM can do.

Limitations. We have been able to provide examples that illustrate the differences between standard causal, and meta-causal attribution. In particular, the joint of actual and mechanistic view of standard and meta-causal attributions should be explored further. Our approach to recovering MCS from unlabeled data is open to extension. Discovery on the full causal graphs is a desirable goal that, however needs further work to reliably control for the changing sets of parents under changing MCS.

Acknowledgments and Disclosure of Funding

The TU Darmstadt authors acknowledge the support of the German Science Foundation (DFG) project "Causality, Argumentation, and Machine Learning" (CAML2, KE 1686/3-2) of the SPP 1999 "Robust Argumentation Machines" (RATIO). This work is supported by the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK; projects "The Third Wave of AI"). It was funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO. This work was supported from the National High-Performance Computing project for Computational Engineering Sciences (NHR4CES).

The Eindhoven University of Technology authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute.

References

Tara V Anand, Adèle H Ribeiro, Jin Tian, and Elias Bareinboim. Effect identification in cluster causal diagrams. *arXiv preprint arXiv:2202.12263*, 2022.

Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pp. 193–212, 1952.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pp. 679–684, 1957.

J Douglas Bremner. Traumatic stress: effects on the brain. *Dialogues in clinical neuroscience*, 8(4): 445–461, 2006.

Philippe Brouillard, Perouz Taslakian, Alexandre Lacoste, Sébastien Lachapelle, and Alexandre Drouin. Typing assumptions improve identification in causal discovery. In *Conference on Causal Learning and Reasoning*, pp. 162–177. PMLR, 2022.

Colin Chen. Tests for the goodness-of-fit of the laplace distribution. *Communications in Statistics-Simulation and Computation*, 31(1):159–174, 2002.

Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pp. 236–243. Springer, 2003.

Juan D Correa, Sanghack Lee, and Elias Bareinboim. Counterfactual transportability: a formal approach. In *International Conference on Machine Learning*, pp. 4370–4390. PMLR, 2022.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Joel E Dimsdale. Psychological stress and cardiovascular disease. *Journal of the American College of Cardiology*, 51(13):1237–1246, 2008.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Tamara B Franklin, Bechara J Saab, and Isabelle M Mansuy. Neural mechanisms of stress resilience and vulnerability. *Neuron*, 75(5):747–761, 2012.

Tobias Gerstenberg. What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, 377(1866):20210339, 2022.

Tobias Gerstenberg. Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, 2024.

Tobias Gerstenberg, Noah Goodman, David Lagnado, and Josh Tenenbaum. From counterfactual simulation to causal judgment. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12: 317–337, 2000.

Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2023.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

Peter Maisel, Erika Baum, and Norbert Donner-Banzhoff. Fatigue as the chief complaint: epidemiology, causes, diagnosis, and treatment. *Deutsches Ärzteblatt International*, 118(33-34):566, 2021.

Edward F Moore et al. Gedanken-experiments on sequential machines. *Automata studies*, 34: 129–153, 1956.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

Steindór Sæmundsson, Katja Hofmann, and Marc Peter Deisenroth. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*, 2018.

Andreas WM Sauter, Erman Acar, and Vincent Francois-Lavet. A meta-reinforcement learning algorithm for causal discovery. In *Conference on Causal Learning and Reasoning*, pp. 602–619. PMLR, 2023.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Steven A Sloman, Philip M Fernbach, and Scott Ewing. Causal models: The representational infrastructure for moral judgment. *Psychology of learning and motivation*, 50:1–26, 2009.

Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*, pp. 21–33. PMLR, 2021.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Davide Talon, Phillip Lippe, Stuart James, Alessio Del Bue, and Sara Magliacane. Towards the reusability and compositionality of causal representations. *arXiv preprint arXiv:2403.09830*, 2024.

Julius Von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9584–9594, 2022.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *arXiv preprint arXiv:2306.07047*, 2023.

Clare R Walsh and Steven A Sloman. The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, 26(1):21–52, 2011.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Do not marginalize mechanisms, rather consolidate! In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Phillip Wolff. Representing causation. *Journal of experimental psychology: General*, 136(1):82, 2007.

Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. Learning causal dynamics models in object-oriented environments. *arXiv preprint arXiv:2405.12615*, 2024.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *arXiv preprint arXiv:2302.05209*, 2023.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1778–1780, 2017.

Appendix for "Systems with Switching Causal Relations: A Meta-Causal Perspective"

A Inferring Meta-Causal States.

Even if the state transition function is known, it may be unclear from a single observation which exact meta-causal state led to the generation of a particular observation S. This is especially the case when two different meta-causal states can fit similar environmental dynamics. Even in the presence of latent factors (e.g., an agent's internal policy), the current dynamics of a system (e.g., induced by the agent's current policy) can sometimes be inferred from a series of observed environment states. This requires knowledge of the meta-causal dynamics, and is subject to the condition that sequences of observed states uniquely characterize the meta-causal state. Since MCM are defined as finite state machines, the exact condition for identifying the meta-causal state is that observed sequences are homing sequences (Moore et al., 1956). Note that the following example of a game of tag presented below exactly satisfies this condition, where the meta-causal states produce disjunct sets of environment states ('agent A faces agent B', or 'agent B faces agent A'), and thus can be inferred from either a single observation, when movement directions are observed, or two observations, when they need to be inferred from the change in position of two successive observations.

Inferring the Meta-Causal State: 'Game of Tag' Example. Even if the state transition function is known, it may be unclear from a single observation which exact meta-causal state led to the generation of a particular observation S. Since MCM are defined as finite state machines, the exact condition for identifying the meta-causal state is that observed sequences are homing sequences (Moore et al., 1956). Consider an idealized game of tag between two agents, with a simple causal graph and two different meta-states. In general, we expect agent B to make arbitrary moves that increase its distance to A, while A tries to catch up to B, or vice versa. In essence, this is a cyclic causal relationship between the agents, where both states have the same binary adjacency matrix. Note, however, that the two states differ in the *type* of relationship that goes from A to B (and B to A). We can use an identification function that analyzes the current behavior of the agents to identify each edge. Since A can tag B, the behavior of the system changes when the directions of the typed arrows are reversed, so that the type of the edges is either 'escaping' or 'chasing'.

While the underlying policy of an agent may not be apparent from observation as an endogenous variable, it can be inferred by observing the agents' actions over time. Knowing the rules of the game, one can assume that the encircling agent faces the other agent and thus moves towards it, while the fleeing agents show the opposite behavior:

$$(t_{A\to B} = \text{chasing}) \iff (\dot{A}_{pos} \cdot (B_{pos} - A_{pos}) > 0)$$

where $\dot{A}_{pos}$ is the velocity vector of agent A (possibly computed from the position of two consecutive time steps); A_{pos} and B_{pos} are the agent positions, and $\cdot$ is the dot product. Once the edge types are known, the policy can be identified immediately.

Causal and Anti-Causal Meta-Causal States. Assigning meta-causal states to particular system observations can be understood as labeling the individual observations. However, it is generally unclear whether the meta-causal state has an observational or a controlling character on the system under consideration. One could ask the question whether the system dynamics cause the meta-causal state, whether the meta-causal state causes the system dynamics, (or whether they are actually the same concept), similar to the well-known discussion "On Causal and Anticausal Learning" by Schölkopf et al. (2012), but from a meta-causal perspective.

At this point in time, we cannot give definitive conditions on how to answer this question, but we present two examples that support either one of two opposing views. First, in Sec. 4.3 we present a scenario of a dynamical system where the structural equations of both meta-causal states are the same, since the system dynamics are governed by its self-referential system dynamics. Intervening on the meta-causal state will therefore have no effect on the underlying structural equations, and thus can have no effect on the actual system dynamics. In such scenarios, the meta-causal is rather a descriptive label and cannot be considered an external conditioning factor. In the following, we discuss the opposite example, where a meta-causal state can be modeled as an external variable conditioning the structural equations.

B Probabilities for Sample Computation and Upper Bounds

Consider a dataset $\mathcal{D} \in \mathbf{R}^{m\times 2}$ of m samples over two variables where we want to separate n different functions. We assume that the data distribution contains a uniform number of samples from each function, where each class could be under- or overrepresented by an offset of d. Specifically, we assume that each function is represented by $(1 \pm \mathrm{d})\frac{1}{n}|\mathcal{D}|$ samples, i.e., the probability of encountering a particular class X is $\frac{1-\mathrm{d}}{n} \leq P(X) \leq \frac{1+\mathrm{d}}{n}$ and $\mathbb{E}[P(X)] = \frac{1}{n}$. To identify all functions between these two variables, we assume linearity and apply EM with RANSAC on random pairs of samples (see Section 4.2). By selecting n pairs, there is a chance that one pair is chosen from each function (we will refer to such a set of pairs as a "correct" set of samples). In this section, we derive the probability of a correct pair being chosen at random, so that we can estimate how many times pairs need to be sampled to reliably encounter a correct set of samples.

We denote S as the event of "correctly" sampling all n pairs from all n different functions. If all classes have the same number of samples, the chance of randomly selecting a pair from a new class is $\frac{n}{n} \cdot \frac{1}{n}$ for the first pair of samples, $\frac{n-1}{n} \cdot \frac{1}{n}$ for the second, ..., and $\frac{1}{n} \cdot \frac{1}{n}$ for the last; in short:

$$\mathbb{E}[P_n(S)] = \frac{n!}{n^n} \cdot \left(\frac{1}{n}\right)^n = \frac{n!}{n^{2n}}. \tag{1}$$

If the data distribution is not perfectly uniform, i.e., $\mathrm{d} \neq 0$, we can also calculate a lower bound for the same probability. Consider two probabilities per sample: the **probability of selecting a new class** $P_n(S^{\text{new}})$ and the **probability of selecting a second sample of the same class** $P_n(S^{\text{same}})$ afterwards. Across all samples, these correspond to $\frac{n!}{n^n}$ and $(\frac{1}{n})^n$ in $\mathbb{E}[P_n(S)]$, respectively.

Let us first consider the **probability of selecting a new class**. When the first sample is taken, only one new class can be selected (probability of 1). If this sample was taken from the largest class first, the probability that subsequent samples will be taken from new classes decreases, since the space of "unsampled" classes is smaller. For this lower bound, we therefore assume that maximally large classes are sampled from as much and as early as possible. According to our assumptions, the largest classes each take up a fraction of $\frac{1+\mathrm{d}}{n}$ of the data. Therefore, the probability of selecting a new class for successive samples has the following probabilities $\frac{n}{n}, \frac{n-(1+\mathrm{d})}{n}, \frac{n-(1+\mathrm{d})2}{n}, \ldots$. First, consider the case where n is even. Here, after all the $\frac{n}{2}$ largest classes have been selected, only the small classes remain. For the last, second to last, ... classes, this probability is represented by $\frac{(1-\mathrm{d})}{n}, \frac{2(1-\mathrm{d})}{n}, \ldots$. Overall, we get the probability

$$P_n(S^{\text{new}}_{\text{even}}) \geq \left(\prod_{i=\frac{n}{2}}^{n} \frac{n-(1+\mathrm{d})(n-i)}{n}\right)\left(\prod_{i=1}^{\frac{n}{2}-1} \frac{i(1-\mathrm{d})}{n}\right).$$

If n is uneven, an average size class between the largest and smallest classes must be included:

$$P_n(S^{\text{new}}_{\text{odd}}) \geq \left(\prod_{i=\frac{n}{2}+0.5}^{n} \frac{n-(1+\mathrm{d})(n-i)}{n}\right)\left(\frac{n-(1+\mathrm{d})(\frac{n}{2}-0.5)-1}{n}\right)\left(\prod_{i=1}^{\frac{n}{2}-1.5} \frac{i(1-\mathrm{d})}{n}\right).$$

The **probability of selecting a second sample of the same class** is easier to calculate. Instead of constant probabilities as in the expectation with $\frac{1}{n}$, we now have two different probabilities in the even case and three in the odd case. In the even case, we have $\frac{n}{2}$ large batches and the same number of small batches, so the probability of choosing the right batch each time is

$$P_n(S^{\text{same}}_{\text{even}}) \geq \left(\frac{1+\mathrm{d}}{n}\right)^{\frac{n}{2}}\left(\frac{1-\mathrm{d}}{n}\right)^{\frac{n}{2}}.$$

In the uneven case, we also have to also consider the batch that has an average size

$$P_n(S^{\text{same}}_{\text{odd}}) \geq \left(\frac{1+\mathrm{d}}{n}\right)^{\frac{n}{2}}\frac{1}{n}\left(\frac{1-\mathrm{d}}{n}\right)^{\frac{n}{2}}.$$

Note that for both $P_n(S_{\text{new}})$ and $P_n(S_{\text{same}})$, the distribution of the data into the largest and smallest possible batches (according to our assumptions) results in the smallest possible probabilities; hence,

the computed probability is a lower bound. If the batches were more evenly sized, the probability would be larger. We can also see that a deviation of up to 1 results in a probability of 0, since it is impossible to sample from a class that is not represented in the data.

In total, the probability of selecting of a correct sample set is the product of the two probabilities above, i.e,

$$P_n(S) = \begin{cases} P_n(S^{\text{new}}_{\text{even}})P_n(S^{\text{same}}_{\text{even}}) & \text{if } n \text{ is even} \\ P_n(S^{\text{new}}_{\text{odd}})P_n(S^{\text{same}}_{\text{odd}}) & \text{if } n \text{ is odd.} \end{cases} \tag{2}$$

$P_n(S)$ is the (lower bound on the) probability of selecting a correct set of samples. We can calculate the number of trials k needed to find such a set of samples with at least 95% probability. The opposite probability, of never finding it with less than 5% probability, is easier to calculate:

$$\begin{aligned} (1 - P(S))^k &\leq 1 - 0.95 = 0.05 \\ k \ln(1 - P(S)) &\leq \ln(0.05) \\ k &\leq \frac{\ln(0.05)}{\ln(1 - P(S))} \end{aligned}$$

This allows us to determine how many attempts might be necessary. Note that while this would leave a 5% chance of not picking the right samples, there are various practical reasons why the actual probability of finding a working set of samples will be higher, e.g., if the number of samples from each class is not as uneven as assumed, or if some samples are distributed in such a way that even picking a sample from the "wrong" class might still lead to the identification of the correct mechanisms.

For example, if $n = 2$ and $\mathrm{d} = 0.2$, we have an expected probability of $\mathbb{E}[P(S)] = \frac{2!}{2^{2 \cdot 2}} = 0.125$ and a lower bound of $P_2(S) = P(S^{\text{new}}_{\text{even}})P(S^{\text{same}}_{\text{even}}) \geq \frac{0.8}{2} \cdot \frac{2}{2} \cdot \frac{1.2}{2} \cdot \frac{0.8}{2} = 0.096$. Larger deviations decrease the probability while a deviation of $\mathrm{d} = 0$ results in the same probability as with $\mathbb{E}[P(S)]$. For the lower bound, this results in $k = 30$ samples for a confidence of 95% using the above calculation steps. All resulting sample counts can be found in Table 2.

C LO-RANSAC Bivariate Discovery

The problem we are trying to solve is twofold : first, we are initially unaware of the underlying meta-causal state t that generated a particular data point (x_i, y_i), which prevents us from estimating the parameterization (α_k, β_k) of the mechanism. Conversely, our lack of knowledge about the mechanism parameterizations $(\alpha_k, \beta_k)_{k \in [1..K]}$ prevents us from assigning class probabilities to the individual data points. Since neither the state assignment nor the mechanism parameterizations are initially known, we perform an Expectation-Maximization (EM; Dempster et al. (1977)) procedure to iteratively estimate and assign the observed data points to the discovered MCS parameterizations. Due to the local convergence properties of the EM algorithm, we further embed it into a locally optimized local random sample consensus approach (LO-RANSAC; Fischler & Bolles (1981); Chum et al. (2003)). RANSAC approaches repeatedly sample initial parameter configurations to avoid local minima, and successively perform several steps of local optimization - here the EM algorithm - to regress the true parameters of the mechanism.

Assuming for the moment that the correct number of mechanisms k^* has been chosen, we assume that the EM algorithm is able to regress the parameters of the mechanisms, α^*_k, β^*_k, whenever there exists a pair of points for each of the mechanisms, where both points of the pair are samples generated by that particular mechanism. The chances of sampling such an initial configuration decrease rapidly with an increasing number of mechanisms (e.g. 0.036% probability for $k = 4$ and equal class probabilities). Furthermore, we assume that the sampling probabilities of the individual mechanisms in the data can deviate from the mean by up to a certain factor d. In our experiments, we consider setups with $d \in \{0.0, 0.1, 0.2\}$. Given the number of classes and the maximum sample deviation of the mechanisms, one can compute an upper bound on the number of resamples required to have a 95% chance of drawing at least one valid initialization. The bound is maximized by assigning the first half of the classes the maximum deviation probability $P_{\text{k-max}} = (1 + d)/k$ and the other half the minimum deviation probability $P_{\text{k-min}} = (1 - d)/k$. We provide the formulas for the upper bound estimation and in Sec. B in the Appendix and provide the calculated required resample counts in Table 2. In the

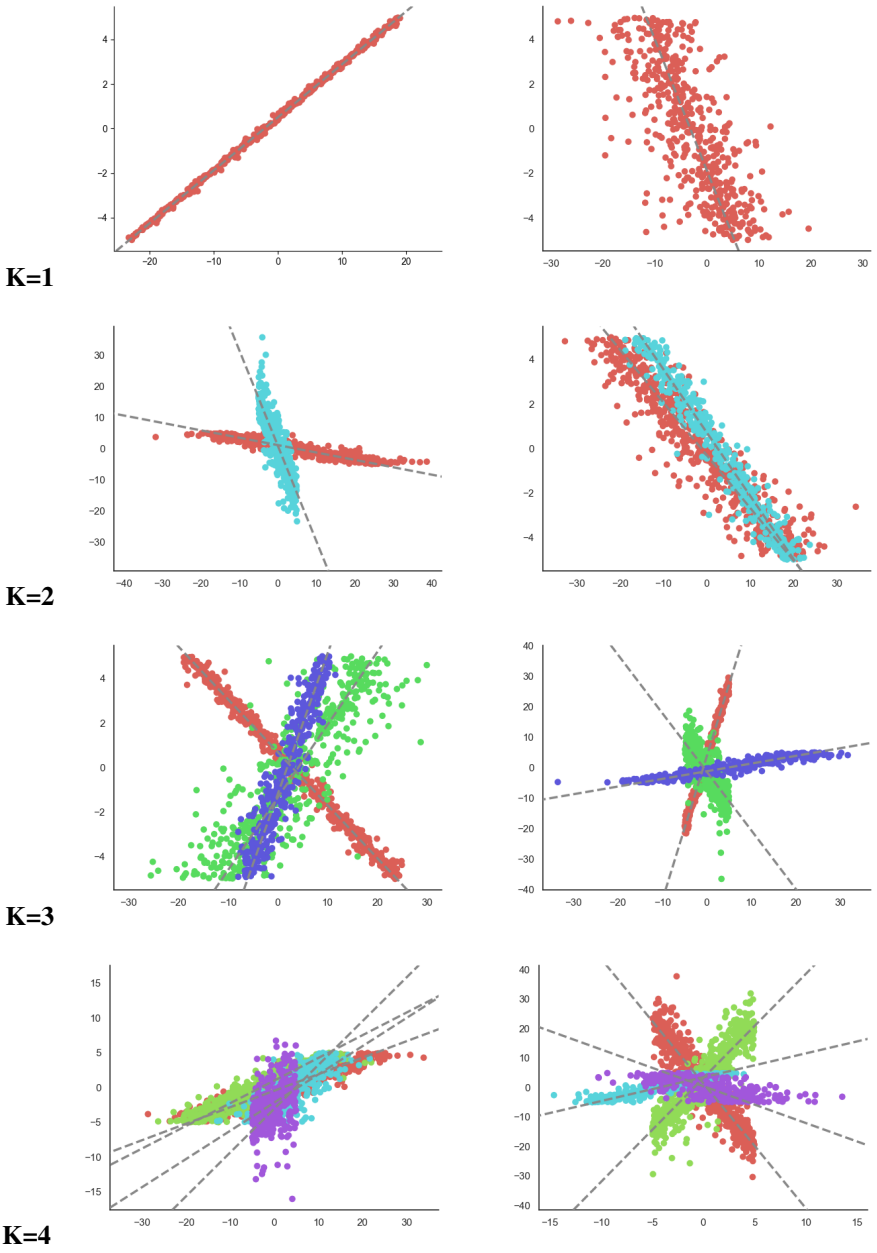

Figure 3: **Sampled Mechanisms.** The figure shows a selection of different randomly sampled mechanism distributions, ranging from one to up to four simultaneously present mechanisms. The gray dotted lines represent the generating ground truth mechanisms. *(Best Viewed In Color)*

worst case, for a scenario with $k = 4$ mechanisms and $d = 0.2$ maximum class probability deviation, nearly $14,900$ restarts of the EM algorithm are required, drastically increasing the potential runtime.

In our experiments, we find that our assumptions about EM convergence are rather conservative. Our evaluations show that the EM algorithm is still likely to be able to regress the true parameters, given that some of the initial points are sampled from incorrect mechanisms. We measure the empirical convergence rate by measuring the convergence rate of the EM algorithm over 5,000 different setups (500 randomly generated setups with 10 parameter resamples each). We perform 5 EM steps for setups with $k = 1$ and $k = 2$ mechanisms, and increase to 10 EM iterations for 3 and 4 mechanisms.

	1 Mechanism	2 Mechanisms	3 Mechanisms	4 Mechanisms
Max. Class Deviation = 0.0				
Theoretical	1	23	363	8,179
Empirical	2	8	24	173
Max. Class Deviation = 0.1				
Theoretical	1	26	429	10,659
Empirical	2	8	25	177
Max. Class Deviation = 0.2				
Theoretical	1	30	526	14,859
Empirical	2	8	26	177

Table 2: **Estimated Number of Required Resamples for Obtaining a 95% Convergence Rate with the LO-RANSAC Algorithm per Number of Mechanisms and Maximum Class Deviation.** The empirical observed convergence rates of the EM algorithm drastically reduce the required number of samples of the theoretical derived bound. This reveals that our made assumptions where chosen to be quite conservative, and attests a good fit of the EM algorithm for regressing the mechanisms' parameters.

Class Deviation	1 Mechanism	2 Mechanisms	3 Mechanisms	4 Mechanisms
0.0	4219 (84.38%)	1740 (34.80%)	592 (11.84%)	86 (1.72%)
0.1	"	1702 (34.04%)	577 (11.54%)	84 (1.68%)
0.2	"	1567 (31.34%)	555 (11.10%)	84 (1.68%)

Table 3: **Empirical estimated convergence percentage of the EM algorithm for different class imbalances and number of mechanisms.** The table shows the number of samples converged and the convergence rates (in parentheses) for a single random initialization and for estimating the parameterization of the underlying system for a given number of mechanisms. All results are reported per 5000 samples.

For each initialization, we count the EM algorithm as converged if the slope and intercept of the true and predicted values do not differ by more than an absolute value of 0.2.

We compare theoretically computed and empirical in convergence results in Table 2. In practice, we observe that the convergence of the EM algorithm is more favorable than estimated, reducing from 23 to 8 required examples for the simple case of $k = 2, d = 0.0$, and requiring up to 83-times fewer samples for the most challenging setup of $k = 4, d = 0.2$, reducing from a theoretical of 14,859 to an empirical estimate of 177 samples. The actual convergence probabilities and the formula for deriving to sample counts are given in Table 3 and the next Appendix Sec. D.

D EM Convergence Results

The required number of resamples for a 95% success rate of the RANSAC algorithm is calculated by $\log(0.05)/\log(1 - C^1)$, where $S1$ are the convergence rates for the individual samples computed in Sec. B.

Empirical convergence probabilities and resulting resampling counts for the LO-RANSAC algorithm are shown in Tables 2 and 3. Table 4 lists the goodness of fit for all converged samples. In general, we find that in cases where the approach is able to converge, it undercuts the required parameter convergence boundary of 0.2 by factors of 4.8 and 3.5 for the slope and intercept, respectively.

E Mechanism Sampling

For our experiments in Sec. 4.2 we uniformly sample the number of mechanisms to be in $K \in \{1..4\}$. The slopes of the linear equations are uniformly sampled between $\alpha \in \pm[0.2..5]$ and the intercepts are in the range $\beta \in [-5, 5]$. We add Laplacian noise $\mathcal{L}(x|\mu, b) = \frac{1}{2b} * \exp(-\frac{|x-\mu|}{b})$ with $\mu = 0$ and $b \in [0.1, 4.0]$. X values are uniformly sampled in the range $[-5, 5]$ and $y_i = \alpha x_i + \beta + \mathcal{L}(x|0, b)$. The average number of samples per class is set to 500. Throughout the experiments, we vary the class probabilities by a class deviation factor $D \in \{0.0, 0.1, 0.2\}$. Specifically, we maximize the class

Class Deviation	1 Mechanism	2 Mechanisms	3 Mechanisms	4 Mechanisms
0.0	0.0349|0.0590	0.0370|0.0543	0.0380|0.0592	0.0389|0.0516
0.1	"	0.0381|0.0555	0.0381|0.0566	0.0346|0.0528
0.2	"	0.0414|0.0548	0.0412|0.0567	0.0402|0.0570

Table 4: **Mean average error for the slope and intercept of the correctly predicted mechanism for different class imbalances and number of mechanisms.** Mechanisms are accepted if their parameters do not differ by more than 0.2 from the ground truth parameterization. The results show that converged samples are typically estimated with an error well below the threshold.

deviation by assigning $K/2$ classes the maximum probability $1/K * (1 + D)$ and $K/2$ classes the minimum probability $1/K * (1 - D)$. If K is odd, a class is assigned the average probability $1/K$. We show a selection of the resulting sample distributions in Fig. 3.

Algorithm 1 Recovering Mechanisms for the Bivariate Case

procedure RECOVERMECHANISMS($\mathbf{x}, \mathbf{y}$, maxClassDev $K_{\max}$, EMSteps)
  for all $k \in [1..K_{\max}]$ **do**
    bestModelLogL $\leftarrow -\infty$
    N $\leftarrow$ requiredSamples$(k, \text{maxClassDev}, 0.95)$ ▷ Compute # of samples (c.f. Sec. B)
    for all $n \in [1..\text{N}]$ **do** ▷ RANSAC iteration.
      $\mathbf{px}, \mathbf{py} \leftarrow \texttt{sample}(x_i, y_i, 2 * k)$ ▷ Initialize parameters with $2 \times k$ points.
      for all $k' \in [0..k)$ **do**
        $\alpha_k \leftarrow (py_{2k+1} - py_{2k})/(px_{2k+1} - px_{2k})$
        $\beta_k \leftarrow py_{2k} - \alpha_k * x_{2k}$
        $b_k \leftarrow 1.0$ ▷ Assume initial avg. deviation of the Laplacion to be 1.
        $d_k \leftarrow$ 'XY' ▷ Assume $X \rightarrow Y$ direction first.
        $\mathbf{c}_k \leftarrow P_{\texttt{Laplacian}}(\mathbf{x}, \mathbf{y}; \alpha, \beta, b, d)$ ▷ Initial class probabilities for all samples.
      end for
      for all $l \in [1..\text{EMSteps}]$ **do** ▷ EM Iteration.
        $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{d} \leftarrow \texttt{regressLines}(\mathbf{x}, \mathbf{y}; \mathbf{c})$ ▷ (Weighted) median regression.
        $\mathbf{c} \leftarrow P_{\texttt{Laplacian}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{d})$
      end for
      modelLogL $\leftarrow \sum_i \text{LogP}_{\texttt{Laplacian}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{d})_i$ ▷ Obtain the joint log probs.
      if modelLogL $>$ bestModelLogL **then**
        bestModelLogL $\leftarrow$ modelLogL
        bestParameters $\leftarrow (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{d})$
      end if
    end for
    allLaplacian $\leftarrow$ `true` ▷ Check if residuals are Laplacian distributed.
    for all $l \in [1..k]$ **do**
      $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{d}) \leftarrow$ bestParameters
      $\mathbf{x}^l, \mathbf{y}^l, \mathbf{c}^l \leftarrow \texttt{selectClass}(\mathbf{x}, \mathbf{y}, \mathbf{c}, l)$ ▷ Select points where $\text{argmax}_{\mathbf{c}} = l$.
      $\mathbf{x}^l, \mathbf{y}^l \leftarrow \texttt{filter}(\mathbf{x}^l, \mathbf{y}^l, \mathbf{c}^l, 0.4)$ ▷ Keep points s.t. $\max^{\#2}\mathbf{c}^l < 0.4(1 - \max^{\#1}(\mathbf{c}^l)$.
      $r = \texttt{LinEq}(\mathbf{x}^l, \mathbf{y}^l; \alpha_l, \beta_l, d_l) - \mathbf{y}^l$
      if not $\texttt{AndersonDarling}(r, 0.95)$ **then**
        allLaplacian $\leftarrow$ `false`
        break
      end if
    end for
    if allLaplacian **then**
      return k
    end if
  end for
  return 0
end procedure