Sample-efficient Learning of Concepts with Theoretical Guarantees: from Data to Concepts without Interventions

Hidde Fokkema*, Tim van Erven* and Sara Magliacane†.
*: Korteweg-de Vries Institute for Mathematics, †: Informatics Institute

UNIVERSITY OF AMSTERDAM

Summary

We describe a framework that utilizes techniques from **Causal Representation Learning (CRL)** to develop new **Concept Based Models** with theoretical guarantees on the correctness of the learned concepts and number of required concept labels. To achieve this, we propose new estimators that can match learned latent variables from the CRL methods with annotated interpretable concepts from human labellers. In more details:

1. We propose a **linear** estimator with finite-sample high probability consistency results.
2. We propose a **non-parametric** estimator for this mapping with asymptotic high probability consistency results.
3. We conducted experiments that show that multiple versions of our estimator perform well in terms of label accuracy, computational cost and various concept leakage metrics.

Setting

We assume a causal system with latent causal variables $G = (G_1, \dots, G_d) \in \mathcal{G} \subseteq \mathbb{R}^d$ The **observation**, $X \in \mathcal{X} \subseteq \mathbb{R}^D$, is generated through an unobserved invertible function

$$X = f(G) + \varepsilon.$$

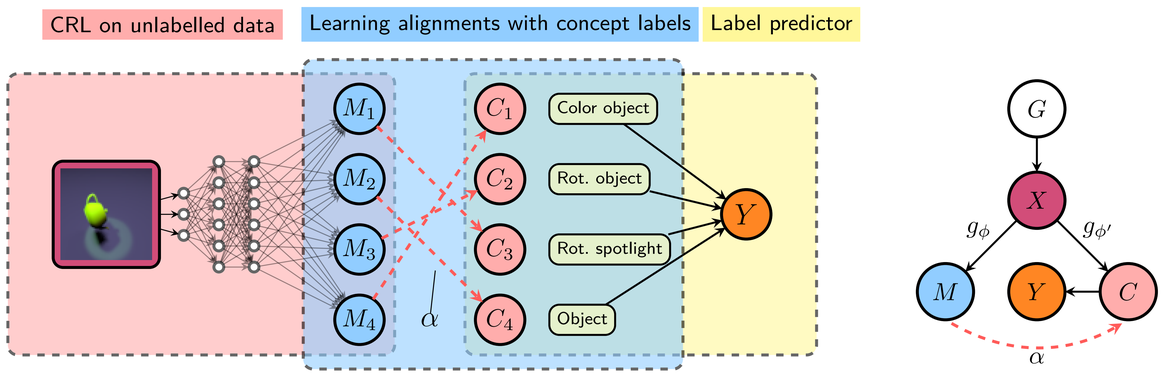
The goal of CRL methods is to recover the causal variables by learning a function $g_\psi \colon \mathcal{X} \to \mathbb{R}^d$ that approximates f^{-1}.

Assumption 3.2. Let $M = g_\psi(X)$ be the causal representations learned by a pretrained CRL method and let $C = g_{\psi'}(X)$ be the interpretable concepts in the observation X. We assume that M and C are **equivalent up to a permutation and diffeomorphism**,

$$PT(M) = \begin{bmatrix} T_{\pi(1)}(M_{\pi(1)}) \\ \vdots \\ T_{\pi(d)}(M_{\pi(d)}) \end{bmatrix} = \begin{bmatrix} C_1 \\ \vdots \\ C_d \end{bmatrix} = C.$$

Matching the learned causal variables with the interpretable concepts now reduces to learning an **alignement map** that consists of a **permutation** and **one-dimensional** regression.

New Framework



Alignment Learning Algorithms

Linear alignment learning

Input: regularization parameter $\lambda > 0$
Data: $\{(C^{(\ell)}, M^{(\ell)})\}_{\ell=1}^n$
for $i = 1, \dots, d$ **do**
  $\widehat{\beta}_i \leftarrow \arg\min_{\beta \in \mathbb{R}^{dp}} \|\mathbf{C}_i - \Phi\beta\|^2 + \lambda\sqrt{p}\|\beta\|_{2,1}$
end for
$\widehat{\pi} \leftarrow \arg\max_{\pi \in \Pi} \sum_{i=1}^d \|\widehat{\beta}_i^{\pi(i)}\|$

Can be solved with a standard Group Lasso solver!

Can be done efficiently and easily!
`scipy.optimize.linear_sum_assignment`

Linear alignment learning

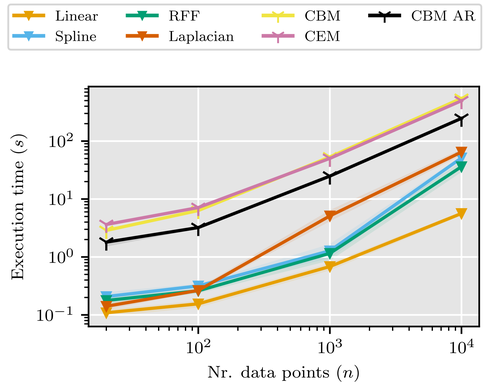
Input: reg. parameter $\lambda > 0$,
  kernels $\kappa_1, \dots, \kappa_d$
Data: $\{(C^{(\ell)}, M^{(\ell)})\}_{\ell=1}^n$
for $j = 1, \dots, d$ **do**
  $(K_j)_{\ell k} \leftarrow \kappa_j(M^{(\ell)}, M^{(k)})$
  $L_j \leftarrow \text{CholeskyDecomposition}(K_j)$
end for
for $i = 1, \dots d$ **do**
  $\widehat{\gamma}_i \leftarrow \arg\min_{\gamma \in \mathbb{R}^{np}} \frac{1}{n}\|\mathbf{C}_i - \sum_{j=1}^d L_j\gamma^j\|^2 + \lambda\|\gamma\|_{2,1}$
end for
$\widehat{\pi} \leftarrow \arg\max_{\pi \in \Pi} \sum_{i=1}^d \|\widehat{\gamma}_i^{\pi(i)}\|$

Accuracy and Concept Leakage Improvement

		Label Acc. ↑ (n)				OIS ↓ (n)			
Model	Method	20	100	1000	10000	20	100	1000	10000
Temporal Causal3DIdent Dataset									
CITRISVAE	Linear	0.74 ± 0.06	0.75 ± 0.06	0.80 ± 0.04	0.81 ± 0.04	0.69 ± 0.02	0.44 ± 0.01	0.19 ± 0.00	0.16 ± 0.00
	Spline	0.74 ± 0.06	**0.80 ± 0.04**	**0.82 ± 0.03**	**0.83 ± 0.03**	0.69 ± 0.02	0.44 ± 0.01	**0.13 ± 0.00**	**0.09 ± 0.00**
	RFF	0.70 ± 0.06	0.76 ± 0.05	0.79 ± 0.04	0.81 ± 0.04	**0.65 ± 0.01**	0.43 ± 0.00	0.16 ± 0.01	0.13 ± 0.01
iVAE	Linear	**0.74 ± 0.06**	0.76 ± 0.05	0.78 ± 0.05	0.80 ± 0.04	0.69 ± 0.02	0.44 ± 0.01	0.17 ± 0.00	0.14 ± 0.00
	Spline	0.73 ± 0.06	0.78 ± 0.04	0.79 ± 0.04	0.81 ± 0.04	0.69 ± 0.02	0.43 ± 0.04	0.17 ± 0.00	0.13 ± 0.00
	RFF	0.69 ± 0.06	0.75 ± 0.05	0.78 ± 0.04	0.80 ± 0.04	0.65 ± 0.01	**0.42 ± 0.02**	0.16 ± 0.00	0.13 ± 0.00
CBM [2]		0.57 ± 0.02	0.53 ± 0.03	0.68 ± 0.04	0.78 ± 0.05	1.00 ± 0.03	0.48 ± 0.02	0.25 ± 0.00	0.18 ± 0.00
CEM [3]		0.64 ± 0.04	0.67 ± 0.03	0.72 ± 0.05	0.78 ± 0.05	0.97 ± 0.05	0.51 ± 0.02	0.36 ± 0.01	0.49 ± 0.01
HardCBM [1]		0.57 ± 0.05	0.53 ± 0.02	0.63 ± 0.03	0.77 ± 0.05	0.96 ± 0.03	0.47 ± 0.02	0.24 ± 0.00	0.16 ± 0.00

Computational efficiency

Training a binary classifier using this framework is more computationally efficient



Theoretical Guarantees

The true model for each concept C_i for $i = 1, \dots, d$ in the **linear setting**, with $\beta_i^\star \in \mathbb{R}^{dp}$ is given by

$$C_i = \varphi(M)\beta_i^\star + \varepsilon_i, \text{ where only } (\beta_i^\star)^{\pi(i)} \neq 0.$$

Corollary 4.3. *Assuming the settings of Theorem 4.2 and that for each $i = 1, \dots, d$, $\|(\beta_i^\star)^j\| > 2c\lambda\sqrt{p}$, then $\widehat{\pi} = \pi$ with probability at least $1 - \delta$.*

The true model for each concept C_i for $i = 1, \dots, d$ in the **non-parametric setting**, with $(\beta_i^\star)^j \in \mathcal{H}_j$ is

$$C_i = \beta_i^\star(M) + \varepsilon_i, \text{ where}$$

$$\beta_i^\star(M) = \sum_{j=1}^d (\beta_i^\star)^j(M_j) = (\beta_i^\star)^{\pi(i)}(M_{\pi(i)}).$$

Theorem 5.2. *Assuming (A-D) in App. B.2. Then, for any sequence of regularization parameters such that $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to +\infty$ when $n \to \infty$, the estimated permutation $\widehat{\pi}$ converges to π in probability.*

References and Link

[1] Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems, NeurIPS*, 2022.
[2] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2020.
[3] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems, NeurIPS*, 2022.

Read our paper online!