Supplementary Material for the paper

Variation-based Cause Effect Identification

APPENDIX

A EMPIRICAL DISTRIBUTION

The empirical probability density function (ePDF):

$$p_{x,N}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x) \tag{17}$$

is the derivative of the empirical cumulative distribution (eCDF) defined by

$$F_{x,N}(x) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{x_n \leq x} \tag{18}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function and the inequality is to be understood entry-wise. The eCDF $F_{x,N}(x)$ is the minimum variance unbiased estimator of the true CDF function $F_x(x)$ (Scott, 1992). The ePDF can also be viewed a limit case of kernel-density estimation.

The motivation behind such a modeling choice is that, we normally do not have the output of our *unknown* system/data-generation-process to an arbitrary input x (other than the samples pairs $\{x_n, y_n\}_{n=1}^N$. Hence, in our search for a distinct marginal on e.g. p_x, we are limited to the convex set defined by the mixture distribution. These are the stimuli for which we know the output of our unknown system treating it as a stochastic mapping. This, in turn, allows us to treat the obtained weight vector as a sample weight on the joint distribution p_{xy} and train models to approximate the conditionals $p_{x|y}$ and $p_{y|x}$ accordingly.

One downside is that the search space for a distinct marginal is limited to this convex set, which is itself sensitive to the sampling error. A standard kernel density estimation can alleviate such a problem, but as mentioned, we assume no access to (nor information on) the underlying system allowing us to use this kde-based estimates on the output or joint spaces.

B MAXIMALLY DISTINCT MIXTURE

In this section we detail the derivation of the semidefinite relaxation (SDR) approach to the optimization problem used in our method eq. 6–8.

B.1 FROM THE UNIFORM EMPIRICAL

Problem 1 *Given a set of samples $\mathcal{D}_x = \{x_n\}_{n=1}^N$ from a random variable $x \in \mathbb{X}$, find the weight vector $\boldsymbol{\alpha}$ that renders the mixture distribution $p_{x,N}^{\boldsymbol{\alpha}}$ maximally distinct from $p_{x,N}$ in some discrepancy measure $D(\cdot,\cdot)$.*

With the kernel-based MMD measure $D \equiv \text{MMD}_{k_\mathbb{X}}$, Problem 1 can be formalized as

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \text{MMD}^2_{k_\mathbb{X}}(p_{x,N}^{\boldsymbol{\alpha}},\, p_{x,N}) \tag{19a}$$

$$\text{subject to} \quad \mathbf{1}_N^\top \boldsymbol{\alpha} = 1 \tag{19b}$$

$$\boldsymbol{\alpha} \geqslant 0 \;\; \text{(entry-wise)} \tag{19c}$$

where $\mathbf{1}_N$ refers to a vector of ones with dimensionality N. The quantity being optimized can be reformulated as follows:

$$\text{MMD}^2_{k_\mathbb{X}}(p^{\boldsymbol{\alpha}}_{x,N},\, p_{x,N}) = \|p^{\boldsymbol{\alpha}}_{x,N}(x) - p_{x,N}(x)\|^2_{\mathcal{H}} \tag{20a}$$

$$= \left\| \sum_{n=1}^{N} \alpha \delta_{\boldsymbol{x}_n} - \frac{1}{N} \sum_{n=1}^{N} \delta_{\boldsymbol{x}_n} \right\|^2_{\mathcal{H}} \tag{20b}$$

$$= \sum_{n,n'=1}^{N} \alpha_n \alpha_{n'} \langle \delta_{\boldsymbol{x}_n}, \delta_{\boldsymbol{x}_{n'}} \rangle - \frac{2}{N} \sum_{n,n'=1}^{N} \alpha_n \langle \delta_{\boldsymbol{x}_n}, \delta_{\boldsymbol{x}_{n'}} \rangle + \frac{1}{N^2} \sum_{n,n'=1}^{N} \langle \delta_{\boldsymbol{x}_n}, \delta_{\boldsymbol{x}_{n'}} \rangle \tag{20c}$$

$$= \boldsymbol{\alpha}^\top \mathbf{K}_{xx} \boldsymbol{\alpha} - \frac{2}{N} \boldsymbol{\alpha}^\top \mathbf{K}_{xx} \mathbf{1}_N + \frac{1}{N^2} \mathbf{1}_N^\top \mathbf{K}_{xx} \mathbf{1}_N \tag{20d}$$

where $\mathbf{K}_{xx} = [k(x_i, x_j)]_{i,j=1}^N$ is the Gram matrix of the kernel function $k_\mathbb{X} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^+$ on the sample set $\mathcal{D}_x$. with which the optimization problem becomes:

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \boldsymbol{\alpha}^\top \mathbf{K}_{xx} \boldsymbol{\alpha} - \frac{2}{N} \boldsymbol{\alpha}^\top \mathbf{K}_{xx} \mathbf{1}_N + \frac{1}{N^2} \mathbf{1}_N^\top \mathbf{K}_{xx} \mathbf{1}_N \tag{21a}$$

$$\text{subject to} \quad \mathbf{1}_N^\top \boldsymbol{\alpha} = 1 \tag{21b}$$

$$\boldsymbol{\alpha} \geqslant 0 \;\; \text{(entry-wise)} \tag{21c}$$

The optimization problem is not a convex optimization problem since it is a *maximization* of a convex function. Noting that the closed-form estimator of the squared MMD has a quadratic form in the optimization variable $\boldsymbol{\alpha}$, Park & Boyd (2017) address this problem in a two-step procedure referred to as semidefinite relaxation (SDR). They first *lift* the problem to a higher dimensional space by defining $\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ in which the objective function becomes linear, then apply a convex *relaxation* to the intractable constraints. Without affecting the solution to the problem and using the properties of the **trace** of a matrix, each term of the objective eq. (21a) can be reformulated as:

$$\boldsymbol{\alpha}^\top \mathbf{K}_{xx} \boldsymbol{\alpha} = \textbf{trace}(\boldsymbol{\alpha}^\top \mathbf{K}_{xx} \boldsymbol{\alpha}) \tag{22a}$$

$$= \textbf{trace}(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathbf{K}_{xx}) \tag{22b}$$

$$= \textbf{trace}(\mathbf{A}\mathbf{K}_{xx}) \tag{22c}$$

$$= \mathbf{A} \bullet \mathbf{K}_{xx} \tag{22d}$$

and similarly for the second term:

$$\boldsymbol{\alpha}^\top \mathbf{K}_{xx} \mathbf{1}_N = \textbf{trace}(\boldsymbol{\alpha}^\top \mathbf{K}_{xx} \mathbf{1}_N) \tag{23a}$$

$$= \textbf{trace}(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathbf{K}_{xx} \mathbf{1}_N \mathbf{1}_N^\top) \tag{23b}$$

$$= \mathbf{A} \bullet \mathbf{K}_{xx} \mathbf{1}_N \mathbf{1}_N^\top \tag{23c}$$

where $\bullet$ denotes the dot-product in matrix space defined as $\mathbf{A} \bullet \mathbf{K}_{xx} = \textbf{trace}(\mathbf{A}\mathbf{K}_{xx})$. They then extract all convex constraints from the condition $\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top = [a_{ij}]_{i,j=1}^{N,N}$ The first is the entry-wise non-negativity $a_{ij} = \alpha_i \alpha_j \geqslant 0$ due to the entry-wise non-negativity of $\boldsymbol{\alpha} \in [0, 1]^N$. The second is the consequence of the normalized vector $\mathbf{1}_N^\top \boldsymbol{\alpha} = 1$ which can expressed in $\mathbf{A}$ as $\mathbf{1}_N^\top \mathbf{A} \mathbf{1} = \mathbf{1}_N^\top \boldsymbol{\alpha} (\mathbf{1}_N^\top \boldsymbol{\alpha})^\top = 1$. The last is the similarity of $\mathbf{A} = \mathbf{A}^\top$ by definition. Finally, the equality condition above is relaxed to $\mathbf{A} \succeq \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ and written in its Schur-complement form.

As a result, the following formulation is a relaxation of 19a–19c which is a quadratically constraint quadratic program (QCQP):

$$\underset{\mathbf{A}}{\text{maximize}} \quad \mathbf{A} \bullet \left(\mathbf{K}_{xx} - \frac{2}{N} \mathbf{K}_{xx} \mathbf{1}_N \mathbf{1}_N^\top \right) + \frac{1}{N^2} \mathbf{1}_N^\top \mathbf{K}_{xx} \mathbf{1}_N \tag{24}$$

$$\text{subject to} \quad \begin{bmatrix} \mathbf{A} & \mathbf{A}\mathbf{1}_N \\ \mathbf{1}_N^\top \mathbf{A} & 1 \end{bmatrix} \succeq 0 \quad \text{(positive semidefiniteness)} \tag{25}$$

$$\mathbf{A} \geqslant 0 \qquad \text{(entry-wise)} \tag{26}$$

$$\mathbf{1}_N^\top \mathbf{A} \mathbf{1}_N = 1 \tag{27}$$

$$\mathbf{A} = \mathbf{A}^\top \tag{28}$$

this problem has a convex object (linear) with convex constraints which can be solved using existing packages such as `cvxpy` Diamond & Boyd (2016).

Problem 2: *Given two sets of samples* $\{\boldsymbol{x}_n\}_{n=1}^N$ *and* $\{\tilde{\boldsymbol{x}}_m\}_{m=1}^M$ *from the two distributions* $p_{x,N}$ *and* $p_{\tilde{x},M}$*, respectively, with the corresponding random variables* $x, \tilde{x} \in \mathbb{X}$ *find the weight vector* $\tilde{\boldsymbol{\alpha}} \in [0,1]^M$ *that renders the mixture distribution* $p_{\tilde{x},M}^{\tilde{\boldsymbol{\alpha}}}$ *maximally distinct from* $p_{x,N}$ w.r.t the discrepancy measure $\text{MMD}_{k_{\mathbb{X}}}$.

This problem can be formalized as

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \text{MMD}^2_{k_{\mathbb{X}}}(p_{\tilde{x},M}^{\tilde{\boldsymbol{\alpha}}},\, p_{x,N}) \tag{29a}$$

$$\text{subject to} \quad \mathbf{1}_M^\top \tilde{\boldsymbol{\alpha}} = 1 \tag{29b}$$

$$\tilde{\boldsymbol{\alpha}} \geqslant 0 \;\; \text{(entry-wise)} \tag{29c}$$

Same as in 20 the objective can be reformulated as follows:

$$\text{MMD}^2_{k_{\mathbb{X}}}(p_{\tilde{x},M}^{\tilde{\boldsymbol{\alpha}}},\, p_{x,N}) = \left\| p_{\tilde{x},M}^{\tilde{\boldsymbol{\alpha}}}(\tilde{x}) - p_{x,N}(x) \right\|^2_{\mathcal{H}} \tag{30a}$$

$$= \tilde{\boldsymbol{\alpha}}^\top \mathbf{K}_{\tilde{x}\tilde{x}} \tilde{\boldsymbol{\alpha}} - \frac{2}{N} \tilde{\boldsymbol{\alpha}}^\top \mathbf{K}_{\tilde{x}x} \mathbf{1}_N + \frac{1}{N^2} \mathbf{1}_N^\top \mathbf{K}_{xx} \mathbf{1}_N \tag{30b}$$

Similar to Problem 1, the objective terms can be rewritten as:

$$\tilde{\boldsymbol{\alpha}}^\top \mathbf{K}_{\tilde{x}\tilde{x}} \tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{A}} \bullet \mathbf{K}_{\tilde{x}\tilde{x}} \tag{31a}$$

and similarly for the second term:

$$\tilde{\boldsymbol{\alpha}}^\top \mathbf{K}_{\tilde{x}x} \mathbf{1}_N = \tilde{\mathbf{A}} \bullet \mathbf{K}_{\tilde{x}x} \mathbf{1}_N \mathbf{1}_N^\top \tag{32a}$$

The constraints can be modified as in Problem 1. Hence, a relaxation of 29a–29c is formulated as:

$$\underset{\tilde{\mathbf{A}}}{\text{maximize}} \quad \tilde{\mathbf{A}} \bullet \left(\mathbf{K}_{\tilde{x}\tilde{x}} - \frac{2}{N} \mathbf{K}_{\tilde{x}x} \mathbf{1}_N \mathbf{1}_N^\top \right) + \frac{1}{N^2} \mathbf{1}_N^\top \mathbf{K}_{xx} \mathbf{1}_N \tag{33}$$

$$\text{subject to} \quad \begin{bmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{A}} \mathbf{1}_M \\ \mathbf{1}_M^\top \tilde{\mathbf{A}} & 1 \end{bmatrix} \succeq 0 \quad \text{(positive semidefiniteness)} \tag{34}$$

$$\tilde{\mathbf{A}} \geqslant 0 \qquad \text{(entry-wise)} \tag{35}$$

$$\mathbf{1}_M^\top \tilde{\mathbf{A}} \mathbf{1}_M = 1 \tag{36}$$

$$\tilde{\mathbf{A}} = \tilde{\mathbf{A}}^\top \tag{37}$$

which is a QCQP on the M^2 optimization variables in $\tilde{\mathbf{A}} = [\tilde{a}_{ij}]_{i,j=1}^{M,M}$.

C Experimental Setup and Further Analysis

In this section we detail the experimental setup that was used in estimating the results presented in fig. 4. We first standardize the dataset using the `RobustScaler`from the `sklearn` library [B1]. As a second step we extract randomly M samples to use further in the optimization problem from 3.1. The next steps are then to be followed as stated in the Algorithm 1 where the hyperparameters were defined as follows:

1. We use a squared exponential kernel (SEK), with its maximum likelihood estimate of its lengthscale parameter using a KDE on a 5-fold cross validation scheme.
2. We use the Exact-GP as our predictive model class $\mathcal{M}$ (SEK as a kernel).
3. We use $b_\alpha = 0.2$
4. We use the mean value for the predition of the GP model
5. All experiments took place on an 8-core processor from a single PC (without GPU compute power).

Note that in case of a large dataset (such as the pair-07 in Tübingen benchmark) we extract a subset that represents the distribution of the original set, referred to as a coreset $\mathcal{D}_{\mathbf{C}}$ which is estimated as follows. From a KDE estimate [B2] on either of the marginals (on x and y), include the k *rare* samples of with probability lower than 0.05 in either of the marginal KDEs. This is then further complemented with $M - k$ samples drawn randomly. This last step (the random draw of $M - k$ samples) is repeated a number of times, and the case with the minimal MMD to the original set is selected. In case of a small dataset, the coreset is automatically identical to the main set.

C.1 BENCHMARK DATASETS

Simulated data:

Another benchmark of synthetic data, inspired by Peters et al. (2014), is a diversified dataset of causal pairs with additive, location-scale, and multiplicative noise. For some deterministic non-linear function f, samples from a Gaussian Process, non-linear additive noise (`AN`) models of the form $y = f(x) + \epsilon$ were considered with $x \sim \mathcal{N}(0, \sqrt{2})$ and $\epsilon \sim \mathcal{N}(0, \sigma)$ with $\sigma \sim \mathcal{U}[1/5, \sqrt{2/5}]$. In the same manner of additive noise, f was replaced by a sigmoid function to generate the `AN-s` scenario. The third scenario consists of location-scale (`LS`) data-generation processes in which the effect's mean and variance are functions of the cause, i.e. $y = f(x) + g(x)\,\epsilon$, where ϵ and x are the same as described in additive noise models. `LS` and `LS-s` correspond, therefore, to the GP-sampled and sigmoid functions described for `AN` and `AN-s` scenarios, respectively. The last scenario has multiplicative models (`MN`) that are defined as $y = f(x)\,\epsilon$, where f is a sigmoid function and $\epsilon \sim \mathcal{U}[0, 1]$.

Each of the above mentioned datasets contains 100 pairs of size $N = 1000$ samples. All pairs have the same weights, and variable ordering is determined by a coin flip, resulting in balanced datasets.

Real-world data: [9] [10].

C.2 BASELINES

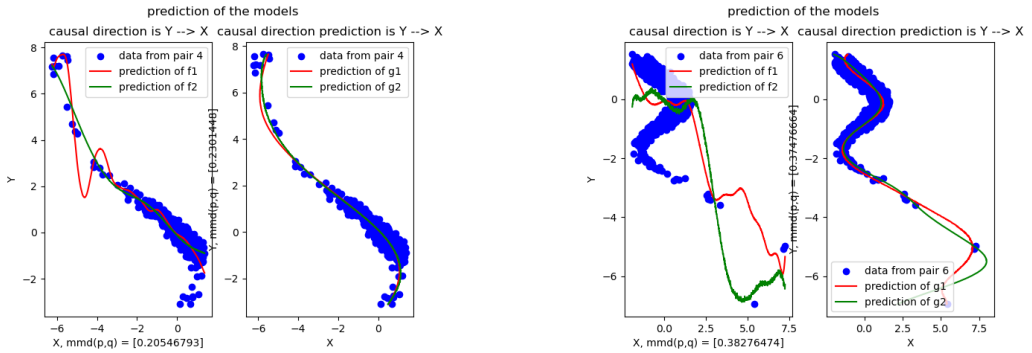
For relevance, we compared our method only to high performing methods shown in Figure 4. All these baselines distinguish between cause and effect solely in the bivariate case. Other methods like `CAM`, `RESIT` and `LinGaM` allow for higher dimensional causal discovery. For more details about the baselines, their implementation and computational complexity we refer to Tagasovska et al. (2018).

C.3 DETAILED ANALYSIS

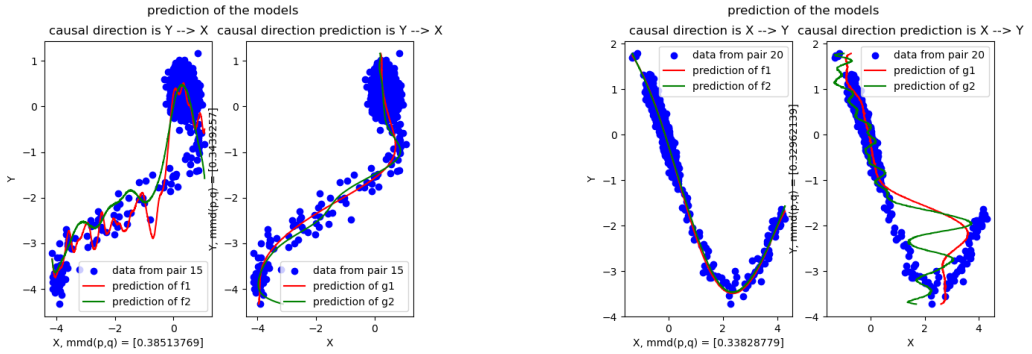
Functional asymmetry illustrative figures: In a first step and similar to the toy-example [1] presented in the beginning, Figure 5 contains some illustrative examples of our method tested on real data, which again shows stability of the causal predictive models compared to acausal ones.

[9] Version Dec 2017.

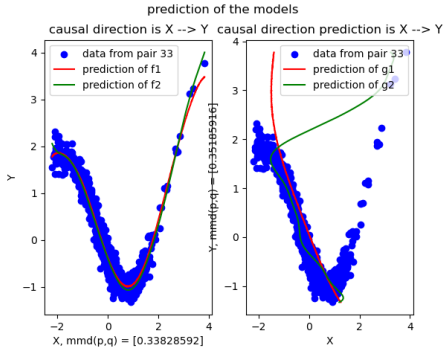
[10] Accompanied weights were used to estimate the final identification accuracies to avoid bias introduced by multiple copies of the same data (see: https://webdav.tuebingen.mpg.de/cause-effect/ for details).



(a)



(b)



(c)

Figure 5: Sample results showing the trained models in each (causal) direction over some files of the SIM dataset.

computational and parameter analysis: In the following, we tackle our method from different angles to analyse its performance.

Figure 6 shows how the maximized objective function (i.e. solution to eq. 9–13 and 15) changes per subset for the SIM dataset (which includes 100 pairs) for a certain b_α. Figure 7 illustrates that the maximum MMD is a monotonic function of the controllable hyperparameter b_α, and that its variations increase as the optimization problem gets less constrained (i.e. larger b_α).

One should rather note that (as discussed in Section 3.3 paragraph: Dirac Distributions), while the maximum MMD increases for an increasing b_α, the resultant distribution p^α actually becomes more and more degenerate (i.e. towards a Dirac delta). Effectively, the models f^α and g^α trained on such an empirical distribution are less stable. In other words, there should be a peak point in the range of

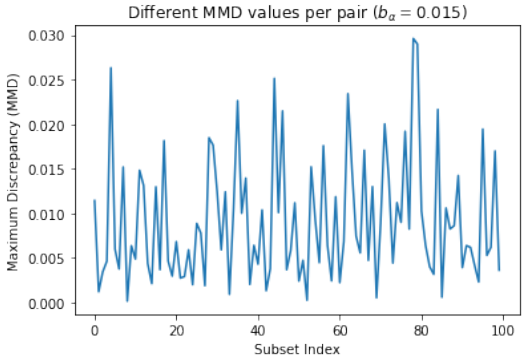


Figure 6: The Maximum Mean Discrepancy (MMD) for each pair of the SIM benchmark dataset with a fixed bound $b_\alpha = 0.015$ and sample size $M = 70$ samples. As observed, the obtained max. MMD value differs across subsets, and is not directly controllable.

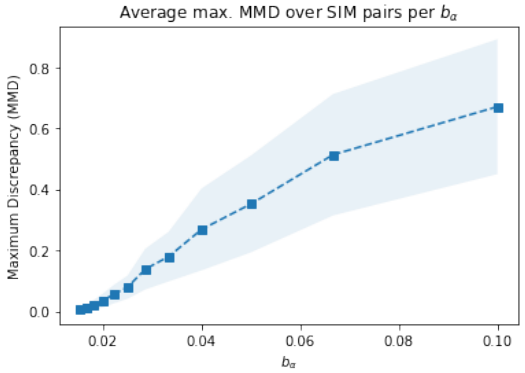


Figure 7: The Maximum Mean Discrepancy (MMD) for different values of b_α with a sample size of 70 samples over all SIM pairs. As $b_\alpha \to 1.0$ the search space for the maximally distinct subset increases, and thus the max. MMD is a monotonic function of b_α. Shaded are represent the variation in the max. MMD for the subsets of the SIM benchmark dataset for a given value of the hyper-parameters b_α and M.

the parameter b_α that reaches an optimal accuracy (i.e. giving non-negligible MMD value, but also non-degenerate empirical distributions). This is depicted in Figure 8 for the SIM dataset (10 trial runs of VCEI with $M = 70$ samples). Note that, the results in this figure shouldn't be compared to the accuracies presented in the experiments section of the paper (Figure 4) since the number of samples has been drastically reduced for these analysis.

Another hyperparameter in our framework is the size of the subset we use. The expected behavior here is that as more samples are included, the optimization is able to search across a larger space for distinct subset. In other words, the maximum MMD obtained should be a monotonic function of the number of samples M for a fixed optimization hyper-parameters (e.g. b_α). This is shown empirically for a single trial on the SIM dataset in Figure 9.

For a fixed b_α, the effect of increasing M has multiple (potentially conflicting) effects on the scores $S_{x \to y}$ and $S_{y \to x}$. On the one hand, increasing M gives more distinct subsets as a solution to the optimization problem 9–13. Yet, it also gives more training samples for the models f and g as compared to the corresponding f^α and g^α which are normally fixed to b_α^{-1} training samples. The identification accuracy vs M for a fixed value of b_α is depicted in Figure 10. An approach to mitigate this issue is to make b_α a function of the subset size (e.g. $b_\alpha = 2/M$). the identification accuracy vs M for a linked value of b_α (for 4 trials on the SIM dataset) is shown in Figure 11

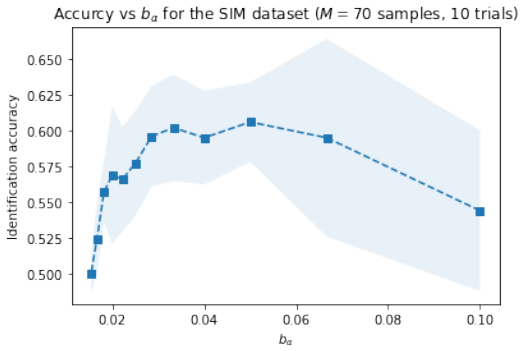


Figure 8: The identification accuracy for different values of b_α with a sample size of $M = 70$. As expected, small values of b_α renders the marginals p and p^α almost similar (since $\alpha \to \frac{1}{M}\mathbf{1}$), leading to inferior performance. As $b_\alpha \to 1.0$, the marginal p^α degenerate to a Dirac delta measure, which leads to an unstable training of the models f^α or g^α. Variations (shaded region) are across 10 trials on the same dataset.

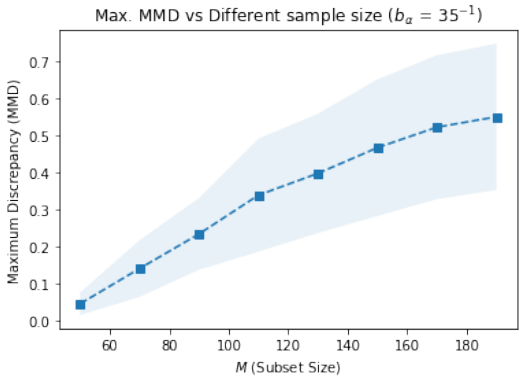


Figure 9: The Maximum Mean Discrepancy (MMD) for different sample size M with a fixed value for the hyper-parameter $b_\alpha = 1/35$. Similar to the effect of b_α, increasing the sample size M widens the search space for the maximally distinct marginal. Effectively, the max. MMD is also a monotonic function of the sample size M. Variations (shaded area) are across different subsets of the SIM bechmark dataset.

Figure 12 provides an overview about the computational complexity of our method in sort of histograms of processing time per dataset pair for the benchmark datasets SIM and Tübingen with $M = 70$ samples. For the Tübingen benchmark, for instance, most of the dataset pairs took 25 seconds of a processing time, with some outliers reaching up to 250 seconds. The total processing time of this dataset on a 28-core workstation is around 25-30 minutes. In the supplementary material, we also show histogram plots of the processing time vs sub-set size M.

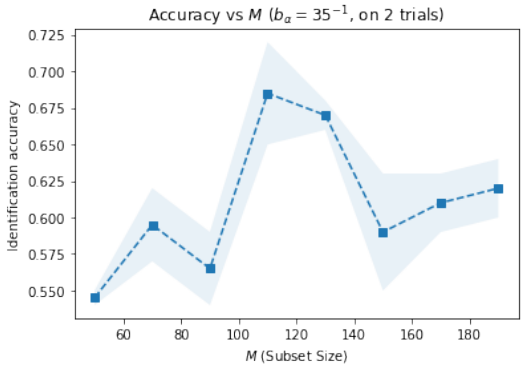


Figure 10: The identification accuracy for different subset sample size M with a fixed bound $b_\alpha = 1/35$ on the SIM benchmark dataset. Variations are across 2 trials with random seeds.

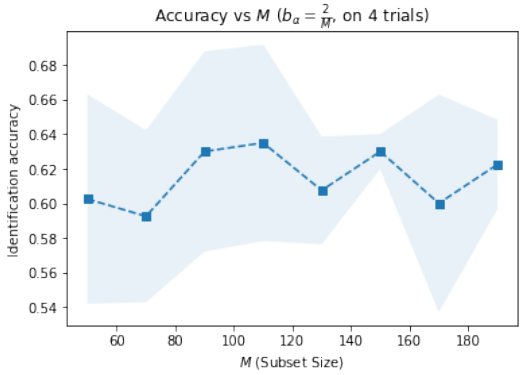


Figure 11: The identification accuracy for different subset sample size M with a linked hyper-parameter $b_\alpha = \frac{2}{M}$ on the SIM benchmark dataset. Variations are across 4 trials.

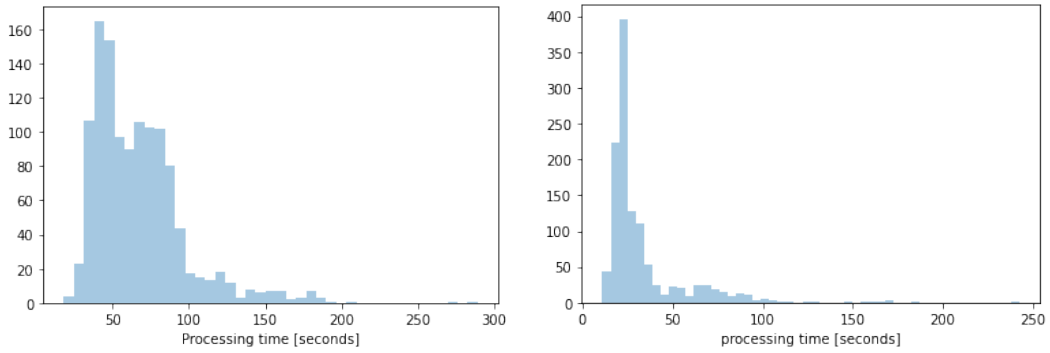


Figure 12: A histogram of the processing time of each pair (aka subset) for the (left) SIM and (right) Tübingen benchmark datasets for a sample size of $M = 70$ and a $b_\alpha = 65^{-1}$. The total processing time is around 25min. (for either of the datasets) on a 28 core workstation.