Causal Discovery in Hawkes Processes by MDL Technical Appendix

Anonymous authors

1 Proofs

1.1 Proof of Theorem 2

Theorem 2. *If in MDL-based model selection for exp-MHP,*

$$\pi(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \pi_i(\boldsymbol{\gamma}_i), \quad v(\boldsymbol{\theta}) = \prod_{i=1}^{p} v_i(\boldsymbol{\theta}_i), \tag{1}$$

then the MDL function can be rewritten as p independent terms

$$L_v(\boldsymbol{\gamma}; \boldsymbol{x}) = \sum_{i=1}^{p} L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}), \tag{2}$$

such that each $L_v^i(\boldsymbol{\gamma}_i, \boldsymbol{x})$ can be computed by Algorithm 2.

Proof. Define

$$\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}} - \log p(\boldsymbol{x}|\boldsymbol{\theta}_i) - \log v_i(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}$ is the space of all possible values for $\boldsymbol{\theta}_i$. We have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}) &= \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} - \log p(\boldsymbol{x}|\boldsymbol{\theta}) - \log v(\boldsymbol{\theta}) \\ &= [\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_1}(\boldsymbol{x})^T, \hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_2}(\boldsymbol{x})^T, \ldots, \hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_p}(\boldsymbol{x})^T]^T.\end{aligned}$$

Define

$$COMP(M_{\boldsymbol{\gamma}_i}; v) = \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) d\boldsymbol{s}.$$

Let C denote $COMP(M_{\boldsymbol{\gamma}}; v)$.

We have

$$\begin{aligned}C &= \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s})) v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s})) d\boldsymbol{s} \\ &= \log \int_{\mathcal{X}} [\prod_{i=1}^{p} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))] d\boldsymbol{s} \\ &= \log \prod_{i=1}^{p} \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) d\boldsymbol{s} \\ &= \sum_{i=1}^{p} \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s})) d\boldsymbol{s} \\ &= \sum_{i=1}^{p} COMP(M_{\boldsymbol{\gamma}_i}; v).\end{aligned}$$

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

As in Eq. 22 in the paper, the negative log-likelihood can be written in independent terms for each dimension. Therefore, for each dimension $1 \leq i \leq p$ we may define

$$\begin{aligned}L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}) := &- \log \pi_i(\boldsymbol{\gamma}_i) - \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \\ &- \log v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) + COMP(M_{\boldsymbol{\gamma}_i}; v).\end{aligned} \tag{3}$$

Hence, we have

$$\begin{aligned}\sum_{i=1}^{p} L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}) = &- \sum_{i=1}^{p} \log \pi_i(\boldsymbol{\gamma}_i) - \sum_{i=1}^{p} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \\ &- \sum_{i=1}^{p} \log v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \\ &+ \sum_{i=1}^{p} COMP(M_{\boldsymbol{\gamma}_i}; v) \qquad (4) \\ = &- \log \pi(\boldsymbol{\gamma}) - \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})) \\ &- v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})) + COMP(M_{\boldsymbol{\gamma}}; v) \qquad (5) \\ = &L_v(\boldsymbol{\gamma}; \boldsymbol{x}). \qquad (6)\end{aligned}$$

Thus, L_v^i as defined above satisfies Eq. 2 as required. Algorithm 2 summarizes the procedure for computing $L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$. First, we compute the MDL estimator $\hat{\boldsymbol{\theta}}_i$ for the i-th dimension by optimizing goodness-of-fit, which is a convex optimization problem by an appropriate choice of luckiness function v, as discussed in subsection 4.3. Next, we estimate the model complexity by using Algorithm 1. Finally, we compute MDL objective as in Eq. 2. □

2 Experimental Setup

Here we provide more details to our experiments.

2.1 Synthetic Experiments

The comparison methods are estimation methods which search for MHP kernel functions and baseline vector based on the data. To extract a causal graph from such output, based on Theorem 1, we put a threshold on the kernel norm to distinguish zero and non-zero kernels. This threshold is set to 0.01.

Each of the comparison methods has a set of hyperparameters like penalty and the level of regularization. For

ML, LS, and NPHC we have penalties: L1 (lasso), L2, elastic net, and none. For ADM4 has lasso-nuclear. We evaluated each of the baseline methods with all possible penalties, i.e. for the set of possible values for

$$C \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, ...\}.$$

For ADM4, we use two hyperparameters, the penalization parameter and the nuclear lasso ratio. Together we considered 5 times 10 cases, i.e. 50 different cases where the lasso-nuclear-ratio was 0, 0.1, 05, 0.9, 1.

The reported numbers for "Random" in Table 1 are the result of a random adjacency matrix with the same number of non-zero entries as the average test case. We report the highest F1 score achieved by each method based on different hyper-parameters. As we do not perform train/test validation and instead we take the highest F1, the validated F1 scores for baseline methods would be presumably lower.

Information criteria (AIC, BIC, and HQ) generally do not perform well for small data and it was the case also in our experiments. These methods rely on reducing model loss (i.e., negative log-likelihood) for the price of adding new parameters to the model. The least price that these models suggest for increasing the size of parameter set is about 1 unit of log-likelihood, hence, these model selection methods allow for adding any edge to the graph only if the log-likelihood could be increased by at least 1 unit compared to the empty graph model. This is not the case for a small data set ("short" data), as the amount of log-likelihood that we have for the naive model and also for the maximum-likelihood model are both very small (of order 0.01 or 0.1 in all of our experimental settings), and therefore, the 1 unit improvement is not possible, and this prevents the IC methods from discovering any edge in the causal graph.

Our method is a MDL-based model selection with no hyper-parameters, however, we can choose the number of MC simulations N for integral estimation, and the higher N the better estimate. Limited by our computational resources, we used 1000 iterations for default case (for dimension $p = 7$), and 500 iterations for the sparse graph scenario (for dimension $p = 20$). As discussed in Section 4 in Subsection Amortization, we first do the MC simulations and compute model complexity values, which takes about one hour in each experimental setting (i.e., fixed p and T), and then we perform 100 test runs, each taking a about ten seconds.

2.2 Real-world Data

In our synthetic experiments we observed that for $p = 7$ and $T = 400$ method ML outperformed the other baseline methods. Elastic net regularization was the best regularization for ML in our synthetic experiments. So we ran this method on the data for the set of levels for regularization as listed above, and in all cases the bi-directed edge between US and Japan as an edge of the causal graph was returned. This is not plausible based on expert knowledge from the domain.