Using Triplet Ordering Preferences for Estimating Causal Effects in the Analysis of Gene Expression Data

Alexander Hartmann [1*] and Grégory Nuel [2]

1 Institut für Physik, Universität Oldenburg, 26111 Oldenburg, Germany
2 LPMA, CNRS 7599, Université Pierre et Marie Curie, Paris, France

* alexander.hartmann@uni-oldenburg.de

Abstract

Triplet ordering preferences are used to perform Monte Carlo sampling of the posterior causal orderings originating from the analysis of gene-expression experiments involving observation as well as, usually few, interventions, like knock-outs. The performance of this sampling approach is compared to a previously used sampling via pairwise ordering preference as well as to the sampling of the full posterior distribution. For a fair comparison, the latter approach is restricted to twice the numerical effort of the triplet-based approach.This is done for artificially generated causal, i.e., directed acyclic networks (DAGs) and for actual experimental data taken from the ROSETTA challenge. The sampling using the triplets ordering turns out to be superior to both other approaches.

Introduction

For the last 10 years, high-throughput omics data have raised many methodological challenges in system biology. Among these challenges, gene-regulation networks have received a great deal of attention. In this context, Gaussian models like the Graphical lasso [1] are very popular for inferring gene regulation networks. Another popular approach, following the work of Pearl [2], focuses on causal Gaussian Bayesian networks and performs intervention calculus [3] proving itself to be able to retrieve bounds on causal effects and thus to partially determine causal relationships using only observational data [4]. In this paper we focus on estimating causal Bayesian networks in the presence of arbitrary mixtures of observational and interventional data [5, 6], i.e., wild-types and knock-out/down experiments with possibly multiple interventions within each experiment.

As explained in [5] estimating the underlying DAG (Directed Acyclic Graph) structure of a causal Bayesian network is equivalent to finding of the so-called *causal ordering* between the genes of interest. In general, this causal ordering is unknown and belongs to a very large ordering space ($p!$ possible orderings for p genes) which cannot be explored exhaustively. The solution suggested by [5] consists in sampling causal orderings in the posterior distribution using Markov chain Monte-Carlo (MCMC) simulations.

At each MCMC step, a new causal ordering is sampled according to a proposal distribution (ex: Mallows distribution) and the maximum likelihood of the model must be computed given the new ordering before to accept/reject the sampled ordering. Thanks to the closed formulas developed in [5], this likelihood maximization can be

done exactly and efficiently but requires an computational effort which still grows with the sixth power of the number p of interacting objects (ex: genes). Thus, each single Monte Carlo step is computationally rather expensive.

Mathematically a proper MCMC is guaranteed to converge to the correct sampling, but only on diverging time scales. Given that for practical applications one only has a finite amount of computational resources available, only small networks can be treated in this way. For this reason, an approximation based solely on pair-wise probabilities of ordering preference has recently been introduced [7]. This resulted in a considerable increase of efficiency, but led in many cases to less reliable parameter estimates.

In this work, we extend this approximation to triplet-wise probabilities. We show that this results in a strongly increased accuracy with respect to the pair-wise approach. Also we show that, when allocating a comparable amount of the numerical resources for the two algorithms, the triplet approach outperforms the sampling based on the full maximum likelihoods. Thus, the triplet algorithm is well balanced: it is sophisticated enough to allow for a rather accurate sampling, while it is computationally cheap enough to be applicable in practice.

The reminder of this work is organized as follows: In Section "Model" we introduce the model we use to analyze causal relationships and state all algorithms we have applied. In Section "Results " we introduce the quantities we have measured to compare the different approaches, and we present the corresponding results. We conclude in Section "Summary and Discussion" with a summary and discussion.

Model and algorithms

Model

We consider directed graphs $G = (V, E)$ with p nodes $i \in V$. Pairs of nodes i, j are connected by directed edges $(i, j) \in E$ and carry a weight $w_{i,j}$. A nonzero weight indicates a causal relationship. We assume that the graph is acyclic, i.e., a directed acyclic graph (DAG). Without loss of generality, we can assume that the nodes are ordered according the causal relationships, i.e., $w_{i,j} > 0 \Rightarrow i < j$. This means within the following random process only nodes i can have causal effects on nodes j if $i < j$:

On each node $j = 1, \ldots, p$ a Gaussian random variable X_j is placed given by

$$X_j = m_j + \sum_{i<j} w_{i,j} X_i + \epsilon_j \quad \text{with} \quad \epsilon_j \sim N(0, \sigma_j^2)\,. \tag{1}$$

The term ϵ_j models fluctuations of the random variables, e.g., for fluctuations of gene expression. Thus, the parameters $\mathbf{m} = (m_1, \ldots, m_p)$ and $\sigma = (\sigma_1, \ldots, \sigma_p)$ represent the mean values and the standard deviations if all interactions were absent. In the following an *experiment* corresponds to one realization of the random process (1).

Within the model is furthermore possible to perform *interventions* on the nodes, i.e., within selected but arbitrary realizations of the process they are fixed to given values instead of generated according to (1). In the DAG these values are uses as inputs to the descendants when generating a realization of the process, i.e., performing an experiment numerically [8].

Estimating model parameters

Given are N experimental data points $\mathbf{x}^k = (x_1^k, \ldots, x_p^k)$ $(1 \le k \le N)$ assumed to be generated according tõ(1). The set of nodes subject to interventions on experiment k is denoted by J_k, respectively ($J_k = \emptyset$ means no intervention for the k'th experiment).

We denote by $K_j = \{k | j \notin J_k\}$ the experiments where there was no intervention on node j and by $N_j = |K_j|$ the number of times node j was not target of an intervention. The log-likelihood of the joint experimental outcome given the parameters can be written (see [5]) as:

$$\ell(\mathbf{m}, \sigma, \mathbf{W}) = \\ -\frac{\log(2\pi)}{2} \sum_j N_j - \sum_j N_j \log(\sigma_i) - \frac{1}{2} \sum_j \frac{1}{\sigma_j^2} \sum_{k \in K_j} (x_j^k - \mathbf{x}^k \mathbf{W} \mathbf{e}_j^T - m_j)^2 \,. \quad (2)$$

Note that we omit the dependence of ℓ on the data here for brevity of notation. For the given N measurements, the parameters $\hat{\mathbf{m}}, \hat{\sigma}, \hat{\mathbf{W}}$ leading to the maximum likelihood estimator (MLE)

$$\ell_{\max} = \ell(\hat{\mathbf{m}}, \hat{\sigma}, \hat{\mathbf{W}}) = \max_{\mathbf{m}, \sigma, \mathbf{W}} \ell(\mathbf{m}, \sigma, \mathbf{W}) \quad (3)$$

can be obtained [5] in a straightforward way by the following procedure: First one obtains for each experiment $k = 1, \ldots, N$ the measurements normalized with respect to the experiments wherethere was no intervention on node j, for each node j:

$$\mathbf{y}^{k,j} = \mathbf{x}^k - \frac{1}{N_j} \sum_{k' \in K_j} \mathbf{x}^{k'} . \quad (4)$$

Next one solves the linear system of size $p(p-1)/2$

$$\sum_{i' | i' < j} \hat{w}_{i',j} \sum_{k \in K_j} y_i^{k,j} y_{i'}^{k,j} = \sum_{k \in K_j} y_i^{k,j} y_j^{k,j} \quad \text{for } i < j,\, 1 \leq i, j, \leq N \quad (5)$$

to obtain estimates $\hat{w}_{i,j}$ of the weights for the MLE. Solving a linear system with $O(p^2)$ variables takes $O(p^6)$ steps.From this solution one obtains, still just following [5], estimates of the mean values

$$\hat{m}_j = \frac{1}{N_j} \sum_{k \in K_j} (x_j^k - \mathbf{x}^k \hat{\mathbf{W}} \mathbf{e}_j^T) \quad (6)$$

and of the variances

$$\hat{\sigma}_j = \frac{1}{N_j} \sum_{k \in K_j} (y_j^{k,j} - \mathbf{y}^{k,j} \hat{\mathbf{W}} \mathbf{e}_j^T)^2 \,. \quad (7)$$

Estimating the posterior distribution

So far, we have assumed that the causal ordering of the model is given by $\mathbf{o}_0 = (1, 2, \ldots, p)$. In experimental situations, if the data was actually generated according the DAG model, the ordering is most of the time unknown, i.e., all estimates will depend on the ordering: $\ell_{\max} = \ell_{\max}(\mathbf{o})$. for the general case, if the data was note generated according to a DAG model, the modeling must involve many orderings. Thus, in experiments and subsequent model estimation, one is actually interested in either the ordering which maximises the MLE, or, alternatively, in obtaining the posterior distribution involving all (or the dominant) orderings weighted by the corresponding ordering-dependent MLEs.

Both can be obtained in principle by iterating over all $p!$ possible causal orderings $\mathbf{o}$, i.e., permutations of the natural numbers $1, \ldots, p$. Each time one has to reorder the

measurement data according this ordering, and obtaining the MLE (3) via solving (4), (5), (6) and (7). Clearly, if p is too large, this enumeration is not possible any more.

One alternative approach is to use a *Markov-chain Monte Carlo* (MCMC) simulation, where orderings $\mathbf{o}(t)$ according the likelihood $\exp(\ell_{\max}(\mathbf{o}))$ are sampled, t denotes the number of steps. A convenient approach to achieve this is the *Metropolis algorithm*. Here, within each step, a *trial order* $\mathbf{o}'$ is generated. For the present study, we use local changes, i.e., an exchange of the order of two nodes with respect to the current ordering $\mathbf{o}(t)$. The trial ordering is *accepted*, i.e., $\mathbf{o}(t+1) = \mathbf{o}'$ with the probability

$$p_{\text{acc}} = \min\{1, \exp[\ell_{\max}(\mathbf{o}') - \ell_{\max}(\mathbf{o}(t))]\}\,. \tag{8}$$

Otherwise, the trial ordering is not accepted and the current ordering kept for the next time step, i.e., $\mathbf{o}(t+1) = \mathbf{o}(t)$.Note that for all simulations we performed (see below for details), the empirical acceptance rate of these locally generated trial orderings was below 0.5. The value of 0.5 is considered by rule of thumb as a good choice, balancing a desired high rate of changing with a desired high acceptance rate. Therefore it would not make sense to consider trial orderings which differ from the current ordering by more than two exchanged positions, since this would increase the fluctuations and therefore decrease the acceptance rate even more.

This type of sampling guarantees, in principle, if the Markov chain is long enough, that the orderings are sampled according the desired posterior distribution. Note that for the computation of the change $\ell_{\max}(\mathbf{o}') - \ell_{\max}(\mathbf{o}(t))$ of the log-likelihood one has to recalculate the log-likelihood for the trial ordering $\mathbf{o}'$ from scratch. Thus, each MCMC Metropolis step takes $O(p^6)$ running time.

By starting with a random ordering $\mathbf{o}(0)$, performing a "long enough" MCMC sampling and by discarding the "initial" part (allowing for *equilibration*), a sample set S of orderings is obtained, which can be used to calculate averaged estimated parameters, see Section "Calculation of averaged estimates".

Calculation of averaged estimates

The aim is to study expectation values in ensembles defined by probabilities or likelihoods $P(\mathbf{o})$. Here we are interested in the true likelihoods $P(\mathbf{o}) \sim e^{\ell_{\max}(\mathbf{o})}$. Thus, for any measured quantity $A(\mathbf{o})$, where the estimate depends on the assumed ordering $\mathbf{o}$, the expectation value is given by

$$\langle A \rangle \equiv \sum_{\mathbf{o}} A(\mathbf{o}) P(\mathbf{o}). \tag{9}$$

Note that the measured quantities of interest are usually estimates which are obtained from the maximum-likelihood calculation, e.g., the estimates of the weights obtained from (5) or estimates of the variances (7), or any other derived values.

If only a finite set S of samples is given, averages can be obtained, approximating the expectation values:

$$\hat{A} \equiv \frac{\sum_{\mathbf{o}\in S} A(\mathbf{o}) P(\mathbf{o})}{\sum_{\mathbf{o}\in S} P(\mathbf{o})} \tag{10}$$

These estimates are most accurate if the process use to generate the sample set follows the desired sampling $P(\mathbf{o}) \sim e^{\ell_{\max}(\mathbf{o})}$ as close as possible. Thus, the sample set S could be generated by a MCMC sampling according to the true probabilities $e^{\ell_{\max}}$, as outlined in the previous section. In this way automatically orderings with high contributions to (10) are preferentially generated. Note that since S is actually a mathematical *set*, there will be no multiple occurrences of orderings in S. If one allowed for multiple occurrence, then one would have to take simple arithmetic averages instead of weighted ones as in (10).

Anyway, here we work with sampling sets. The reason is that, alternatively, these sets can be obtained by sampling according to different probabilities, which only aim at approximating the true probabilities but are computationally much cheaper to calculate. If the size of the set is suitably restricted, we used always $|S| = 100$, the computationally expensive $O(p^6)$ full likelihood calculations have to be performed only for a small number of (here) 100 samples.

The approximate probabilities we have used are introduced in the following section.

Pair and triplet probabilities

Instead of sampling the full posterior distribution, in [7] it was proposed to perform an MCMC sampling from a different distribution, the Babington-Smith (BS) ordering distribution [9,10]. It is based on pair preferences $\pi_{i,j}$ $(1 \leq i \neq j \leq p)$ with $\pi_{i,j} \in [0,1]$ and $\pi_{i,j} + \pi_{j,i} = 1$. The meaning is that within the desired ordering distribution in any random ordering element i appears before j with this probability $\pi_{i,j}$. The pair preferences can be estimated from the experimental data with interventions by considering all possible two-node graphs $G_{i,j} \equiv (\{i,j\}, \{(i,j)\})$ with the nodes i and j and with exactly one directed edge (i,j). As above, for brevity of notation, we omit the dependence of the pair preferences and any derived quantities on the data here. Only the data values for the two nodes are considered [1]. For each of the $p(p-1)$ directed two-node graphs the log-likelihood $\ell^{(2)}_{\max}(i,j)$ is obtained. The pair preferences are then given by

$$\pi_{i,j} = \frac{\exp(\ell^{(2)}_{\max}(i,j))}{(\exp(\ell^{(2)}_{\max}(i,j) + \exp(\ell^{(2)}_{\max}(j,i))}. \tag{11}$$

From the pair preferences, the BS probability of a full ordering $\mathbf{o}$ is obtained by

$$P(\mathbf{o}|\pi) \sim \prod_{i<j} \pi_{o_i,o_j} \tag{12}$$

with a suitable normalization. The normalization is not needed here, since, first, we only compare the (relative) values of (12) for different orderings. The corresponding log-likelihoods are denoted as

$$\ell^{\text{pair}} \equiv \ell^{\text{pair}}(\mathbf{o}) = \log \prod_{i<j} \pi_{o_i,o_j}\,. \tag{13}$$

Second, we performed MCMC sampling of orderings using the Metropolis algorithm according (12) where also only relative likelihoods are needed. This was done in an equivalent way as above, only the true MLE is replaced by (13). Thus, starting again from a random ordering $\mathbf{o}(0)$, we generated trial orderings $\mathbf{o}'$ by exchanging the i'th and the j'the entry in the current ordering. The new orderings are accepted with the corresponding Metropolis probability. Note that one does not have to recalculate the BS probability from scratch, since the change in probability is easier to obtain. The Metropolis acceptance probability is given by

$$p^{\text{pair}}_{\text{acc}} = \min\left\{1, \frac{\pi_{o_j,o_i}}{\pi_{o_i,o_j}} \prod_{k|i<k<j} \frac{\pi_{o_j,o_k}\pi_{o_k,o_i}}{\pi_{o_i,o_k}\pi_{o_k,o_j}}\right\}. \tag{14}$$

[1] In case of multiple interventions, we observed in test which are not contributing to the results shown here that the overall performance of the sampling according pair preferences is somehow better if data points with interventions on other nodes than i,j are not considered, respectively.

This takes only $O(p)$ steps compared to the $O(p^2)$ steps needed for the calculation of the full probability. In particular it is much faster than computing the full likelihood which takes $O(p^6)$ steps.

Naturally, when sampling according to (12) the distribution will be different but somehow similar to sampling according the true likelihood. Thus, the final estimates, like the weights, for the posterior distribution are obtained by keeping the n_{incl} samples with the highest Babington-Smith probabilities (12) in the sample set S. For these orderings now the *true MLE* (3) is evaluated and used. This means, (10) is applied for any kind of estimation or averaging, i.e. the Babington-Smith weights are now used in this final averaging step.

In [7] it was found that this sampling approach is in some case similar accurate as a full MCMC sampling as described in Sec. Estimating the posterior distribution, but there are notable differences. Therefore it was proposed to maybe consider triplets instead of pairs.

Thus, it is the purpose of the present work, to study this higher level approximation of the true posterior distribution. Similar to the above defined pair probabilities, we introduce triplet probabilities $\rho_{i,j,k} \in [0,1]$ such that $\rho_{i,j,k} + \rho_{i,k,j} + \rho_{j,i,k} + \rho_{j,k,i} + \rho_{k,i,j} + \rho_{k,j,i} = 1$. These probabilities can be estimated from the experimental data in a similar way as above, by considering all possible sub graphs ($\{i,j,k\}$, $\{(i,j), (i,k), (j,k)\}$) with three nodes and corresponding edges. For these sub graphs the corresponding MLE are obtained and suitably normalized, equivalent to (11) to yield the triplet probabilities $\rho_{i,j,k}$. They can be used to generalize the Babington-Smith probabilities of orderings to

$$P(\mathbf{o}|\rho) \sim \prod_{i<j<k} \rho_{o_i,o_j,o_k} \,. \tag{15}$$

Again, the normalization is not needed here. The corresponding log-likelihood is denoted as

$$\ell^{\text{tripl}} = \log \prod_{i<j<k} \rho_{o_i,o_j,o_k} \,. \tag{16}$$

We perform an MCMC sampling of orderings according these probabilities using the Metropolis algorithm and trial ordering generated via swapping of pairs of elements. For the calculation of the acceptance probabilities only the change in probability of (15) has to be considered, which takes now $O(p^2)$ steps for such a swap.

Again, for all evaluation and estimations, the $n_{\text{incl}} = 100$ highest-probability samples with respect to the triplet probability are kept. For these samples the true likelihood is obtained and used for all averaging processes according to (10).

Data sources

The new approach will be tested and compared to previous approaches using data from biological applications as well for data generated my numerical simulations for DAGs of different sizes.

For the latter one, we consider random DAGs with p nodes. For the edge weights, each edge (i,j) with $i<j$ receives independently a zero weight with probability $1-q$, i.e., these edges are absent. With probability q each edge gets assigned an edge weight which is drawn uniformly from the range $[-1,-0.4] \cup [0.4,1]$. Thus, these edge can be distinguished very well from the absent edges with weight 0. Below, we use $q=1$, i.e. complete graphs, as well as diluted graphs with $q = c/(p-1)$, i.e., these graphs have on average c neighbors. We used $c=6$. Finally, for each DAG instance, for each node i mean values $m_i = 1/2$ are used and the variance values σ_i are drawn randomly uniformly in the interval $[0.01, 0.1]$. We also performed some tests with other values

and verified that our general conclusions do not all all depend on how the means and the standard deviations are chosen. All simulations are performed for 1000 DAG instances generated independently in this way.

Next, for each DAG instance, a certain number of N measurements is performed, where the measurement vectors $\mathbf{x}^k$ $(k = 1, \ldots, N)$ are generated according to (1). Typically, for a DAG of p nodes, we generated $N = 10p$ measurement vectors, other cases are stated when it applies. We used a variable number of interventions to investigate how the different sampling approaches respond to that variation. Note that the scheme exhibited in Section "Estimating model parameters" allows for multiple intervention. Nevertheless, since we are interested in comparing different sampling approaches here, we present for simplicity just single interventions which are systematically done the first r $(r \leq N)$ experiments of each set of experiments. We applied a systematic manner, such that for all nodes at least $\lfloor r/p \rfloor$ interventions are performed while for $r - p\lfloor r/p \rfloor$ nodes one intervention more, i.e., $\lceil r/p \rceil$ interventions are performed. This sums up to r interventions. For each intervention on node j, we set $X_i = 0$, respectively, corresponding to a knock-out.

The advantage of using artificially generated data is that the actual model used to generate the data is available. Therefore all estimated and averaged values, obtained using a sampling via the true likelihoods as well as using a sampling based on pair and triplet probabilities, can be compared to the actual model parameters. This allows for a good comparison of the different sampling approaches. In particular for a varying number of network sizes, even large ones, and for varying number of interventions.

On the other hand, the DAG models might not represent all subtleties of biological applications. Thus, to allow for a different viewing angle on the different approaches, we applied also data obtained from biological measurements. Here, we used the Rosetta Compendium data set [11] which contains gene expression data on yeast. It contains data from experiments on mutants with interventions (knock-out or know-down) for single as well as multiple interventions. Also a large amount of data from wild-type experiments (no interventions) is contained. The database can be accessed freely at the location: `http://arep.med.harvard.edu/ExpressDB`. We used in particular a sub network taken from [12] consisting of $p = 17$ genes (STT2, TEC1, NDJ1, KSS1, YLR343W, YLR334C, MFA1, STE6, KAR4, FUS1, PRM1, AGA1, AGA2, TOM6, FIG1, FUS3, YEL059W) and data for $N = 300$ experiments. For this set of genes, no knowledge about any possibly underlying network structure or network parameters is assumed while performing the numerical tests here. Only the actual experimental outcomes taken from the database are used. Thus, the estimated parameters generated using the true likelihood form the set of reference values here to perform a comparison of the different approaches. For this purpose, to allow for an exact enumeration, avoiding sampling errors for the reference values, we selected [7] a sub network consisting of $p = 8$ genes, namely STT2, TEC1, NDJ1, KSS1, YLR343W, YLR334C, MFA1, STE6. Note that these genes form a coherent sub network in the network estimated in [12]: with respect to the full 17 nodes network, only one single interaction involving a node of the sub network STE6 (to FUS1) is missing for this selected subset. For these 8 genes, four of the experiments contained single node interventions, namely knock-downs on nodes KSS1, SST2, and twice on TEC1.

Results

To evaluate and compare the power of the Babington-Smith pair and triple approaches, we applied them to various data obtained from DAG ensembles of different graph sizes as well as to data obtained from biological applications.

First, as shown in Section "Direct comparison", we applied the calculation of the

Babington-Smith pair and triple likelihoods to a single graph, where we enumerated all orderings and compared the result to the full likelihoods. In Section "Application to Rosetta Compendium", the results of the application of the MCMC sampling to the Rosetta data set are shown. Next, in Section "Evaluation for random DAGs", MCMC sampling for all three types of likelihoods, respectively, were applied to data obtained for different random DAGs of size $p = 20$ and $p = 50$ nodes. Finally, in Section "Greedy approach", the results of estimating the most-likely orderings via greedy algorithms based on pair- and triplet-probabilities for random DAGs are shown.

Direct comparison

First, we evaluated the likelihood computation for a single randomly picked realization of a complete ($q = 1$) DAG with $p = 8$ nodes. We performed $N = 100$ experiments, among those $r = 4$ with single-node interventions. For this sample, we enumerated all $p! = 40,320$ orderings, and for each ordering we evaluated the true likelihood (2) together with the pair-wise Babington-Smith log-likelihood (13) and the triplet-wise BS log-likelihood (16).

In the left part of Fig 1, for each ordering the pair-wise Babington-Smith likelihood is shown as a function of the full likelihood. This means a scatter plot of $p!$ orderings of points $(\ell_{\max}(\mathbf{o}), \ell^{\text{pair}}(\mathbf{o}))$ is shown. The ordering $\mathbf{o}^{\max}$ leading to the maximum full likelihood appears to the right of the scatter plot, with $\ell_{\max}(\mathbf{o}^{\max}) \approx 1053$. This ordering will dominate any average according to (10). Obviously, this ordering does not exhibit the maximum pair-wise BS likelihood, which is $\ell^{\text{pair}} \approx -7$, obtained by an ordering which true log-likelihood is about $\ell_{\max} \approx 1016$. The horizontal line in the plot indicates the pair-wise BS log-likelihood ℓ^{pair} of the ordering $\mathbf{o}^{\max}$. A considerable amount of all orderings, actually more than 2700, are located above this line. Thus, they exhibit a value of ℓ^{pair} which is higher than for the ordering $\mathbf{o}^{\max}$. Therefore, when performing an MC sampling according the pair-wise likelihoods (13) plus evaluating the true likelihoods for averaging, one must generate a very large sample if the true maximum-likelihood ordering is to be included.

The corresponding result considering triple-wise BS likelihoods ℓ^{tripl} is shown in the right part of Fig 1 in the same way. Here, the sequence exhibiting the maximum value of $\ell^{\text{tripl}} \approx -62$ has a true log-likelihood of $\ell_{\max} \approx 1040$, which is much closer to the sequence $\mathbf{o}^{\max}$ which has (still) $\ell_{\max}(\mathbf{o}^{\max}) \approx 1053$. Only about 100 sequences exhibit a triple-wise BS likelihood larger than for $\mathbf{o}^{\max}$ (indicated again by the horizontal line). This means an MC sampling using the triplet-wise BS likelihoods allows for much more accurate estimation of model parameters with respect to the true likelihoods. This can be seen also in the next section, where an actual biological application is considered.

Application to Rosetta Compendium

For the experimental data points of the Rosetta Compendium for $p = 8$ nodes and $N = 300$ experiments with four interventions (see Section "Data sources" for details), we obtained the averages of the estimated interaction parameters according to (10). One can either estimate the direct causal effects, i.e., the entries w_{ij} of the weight matrices $\mathbf{W}$. Here, we concentrated on the matrice

$$\mathbf{L} = \mathbf{1} + \mathbf{W} + \mathbf{W}^2 + \ldots + \mathbf{W}^{p-1} = (\mathbf{I} - \mathbf{W})^{-1}, \tag{17}$$

which carry the total (direct and indirect) causal effects [7] mediated through chains of causal effects (note that $\mathbf{W}^p = 0$ because of the DAG structure). Thus, for all cases, we estimated the $8 \times 8 = 64$ entries of the matrix $\mathbf{L}$.

The sampling was performed in four different ways:

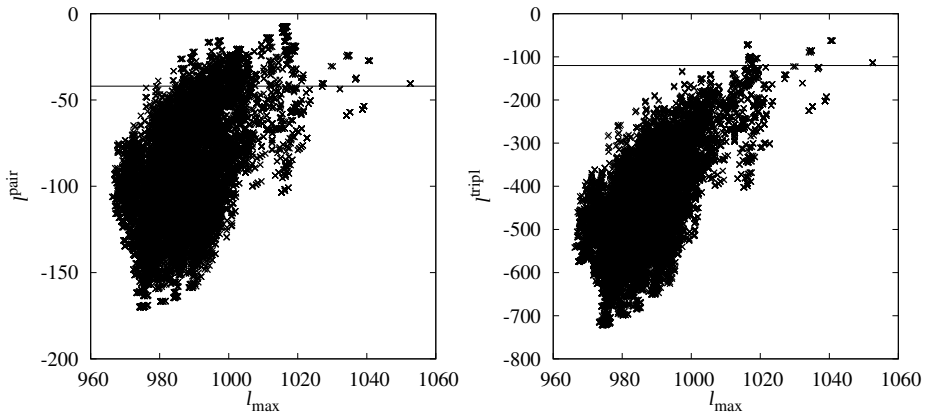

Fig 1. Log-likelihood comparisons. Left: Scatter plot of the true log-likelihood (2) versus the pair-wise log likelihood (13) for data generated for an DAG of size $p = 8$. All 8! orderings were enumerated and the pairs of true likelihood and pair probability plotted. The horizontal line indicates the pair-wise BS log-likelihood for that ordering which exhibits the maximum true likelihood. Right: The same but for true log-likelihood (x-axis) versus triplet log-likelihood (16). Note the different scales of the pair-wise and triple-wise log-likelihoods are only due to the missing normalization of likelihoods.

1. All $p! = 40320$ orderings were enumerated and the true expectation value for all 64 matrix entries was obtained via (9).

2. To estimate the influence of a finite sample size, a subset S of 1000 orderings with the highest true likelihoods $e^{\ell_{\max}(\sigma)}$ was taken. For this subset the averages of estimates of the 64 matrix entries were obtained via (10).

3. An MCMC sampling according the pair BS probabilities (12) was performed.1000 independent MCMC chains where performed, each starting with an independently chosen random ordering. The length of each MCMC chain consisted of 100 pair-exchange trial steps according to (14). From these orderings, the set S of the 1000 orderings exhibiting the highest pair BS probabilities was taken and the average estimates of the matrix entries were obtained via (10).

4. An MCMC sampling according the triplet BS probabilities (12) is performed, in an equivalent way as for the pair BS probabilities. All parameters were the same and the analysis was performed in the same way. Thus, everything was the same, except that the pair BS probabilities were replaced by the more demanding triplet BS probabilities.

In Fig 2 the averages obtained from the approaches 2-4 are compared to the exact expectation values obtained from the first approach. For a perfect estimation of the averages, all data points would lie on the diagonal. Clearly deviations are visible, which is to be expected since the averages are only approximations of the expectation values. The main result is that the deviations are much stronger for the sampling using the pair probabilities. On the other hand, for the triplet probabilities, the scatter of the data points is comparable to the scatter of the exact sampling of restricted size. This shows that the sampling of a finite size set of ordering samples is already close to perfect when using the triplet probabilities.

In the inset of Fig 2 we also show the mean-squared errors $e^2 = (\hat{A}^{\text{a}} - \langle A \rangle)^2$, where A are the different matrix entries L_{ij} and 'a' denotes the algorithm (a= pairs, triplets,

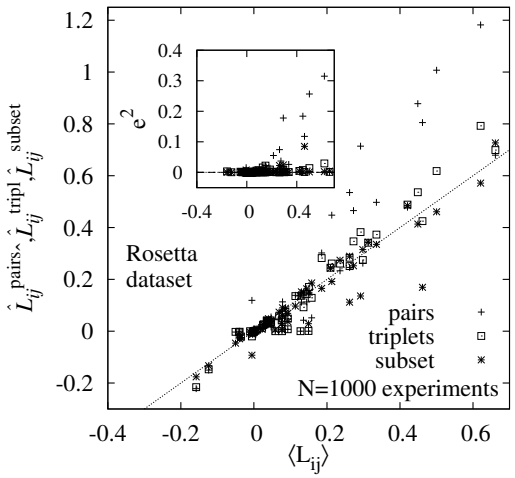

Fig 2. Comparison on the Rosetta Dataset. Comparison of the estimations of the 64 entries of the total causal effects matrix **L** using the exact expectation values $\langle L_{ij} \rangle$ (from a complete enumeration) and estimates $\hat{L}_{ij}$ obtained from the three approaches: pair-wise sampling, triplet-wise sampling, and a subset of the exact sample. For each matrix entry, the average value obtained via one of the three approaches is shown, respectively, as a function of the exact expectation value. The data is taken from the module Rosetta data set (8 genes). The inset shows the mean-squared error e^2 between averaged entry and exact expectation value, as a function again of the exact expectation values.

subset). The above findings are supported by MSE values, which are comparable for triplets probability and subset (exact probability) sampling, but much larger for the pair probability sampling.

Evaluation for random DAGs

Next, we show results for numerically generated data for an ensemble of DAGs. This has the advantage, that due to the average the influence of fluctuations is negligible when comparing the efficiencies of the different sampling approaches. Furthermore, we were able to perform the simulations for different DAG sizes, here we studied DAGs with $p = 20$ and with $p = 50$ nodes. Also, we could vary the number r of interventions over a wide range to get a grip on how this influences the performance of the different algorithms. Finally, we could compare the estimated parameters with the original values used to generate the data. Thus, to measure the efficiency, we consider all edge weights $w_{i,j}$, where $w_{i,j}$ might be zero because it does not match the causal ordering, or because the causal interaction is just absent (in the case of edge probability $q < 1$). This is done in the following way: From each sampling, we obtain averaged estimated edge weights $\hat{w}_{i,j}$ $(i, j = 1, \ldots, p)$ according to (10). Now, we count the "bad" estimates of the edge weights as follows:

$$\delta_{\mathrm{bad}}(i,j) = \begin{cases} \Theta(|\hat{w}_{i,j}| - w_0) & \text{if } w_{i,j} = 0 \\ \Theta\left(\left|\frac{w_{i,j} - \hat{w}_{i,j}}{\hat{w}_{i,j}}\right| - w_1\right) & \text{if } w_{i,j} \neq 0 \end{cases} . \tag{18}$$

$\Theta(x)$ denotes the threshold function which is $\Theta(x) = 0$ for $x \leq 0$ and $\Theta(x) = 1$ for $x > 0$. Thus, for a weight which is zero in the original DAG used to generated the data, the averaged estimate is counted as bad if its absolute value exceeds a threshold

value w_0. For an edge with nonzero weight of the original DAG the average estimate is counted as bad, if the relative deviation of the average estimated weight and the original weight exceeds threshold value w_1. We used $w_0 = 0.2$ and $w_1 = 0.5$. In general, details of the results might depend on the actual values of w_0 and w_1, but we verified that the principal trends, with respect to which sampling approach performs better, remain the same. To exclude the influence of the actual threshold values, we also performed an *Receiver Operator Characteristics* (ROC) analysis, see below. We iterated over all edges, i.e. measured

$$n_{\mathrm{bad}} = \frac{1}{p(p-1)} \sum_{i \neq j} \delta_{\mathrm{bad}}(i,j)\,. \tag{19}$$

The results we show are an average over all 1000 random DAGs.

The measurement data was obtained for $N = 10p$ experiments, i.e., $N = 200$ experiments for $p = 20$ nodes and $N = 500$ experiments for $p = 50$. We performed interventions for a varying number $0 \leq r \leq N$ of experiments as explained in Section "Data sources". The different sets S of sampled orderings, for which the averages $\hat{w}_{i,j}$ were calculated using (10), were obtained via four different sampling approaches, respectively:

pairs An MCMC sampling according the pair BS probabilities (12) is performed. 100 independent MCMC chains where performed, each starting with an independently chosen random ordering. The length of each MCMC chain consisted of 10100 pair-exchange trial steps according to (14). During the last 100 steps of each MCMC chain, configurations were stored, i.e., the initial 10000 steps are for equilibration. From these 10000 stored orderings, the set S of the 100 orderings exhibiting the highest pair BS probabilities was taken and the average entries, now using the true maximum likelihoods of these configurations, were obtained via (10).

triplets An MCMC sampling according the triplet BS probabilities (12) is performed, in an equivalent way as for the pair BS probabilities. All parameters were the same and the analysis was performed in the same way. Thus, everything was the same, except that the pair BS probabilities were replaced by the more demanding triplet BS probabilities.

full In a similar way an MCMC sampling with the full maximum likelihoods was performed. Here only 10 independent runs starting with random orderings were done. Note hat in the limit of infinite long simulation time, each of such an MCMC chain should yield the true expectation values (9). Nevertheless, for a fair comparison, the length of the MCMC chains was chosen such that the full simulation CPU time was slightly above two times the running time of the MCMC simulation using the triplet BS probabilities. Since each MCMC step involves a full $O(p^6)$ calculation of the maximum likelihoods, this means per MCMC chains only 50 steps could be performed.

exact The set S consisted only of the original ordering of nodes which was used generate the data. Thus, only one single $O(p^6)$ maximum likelihood computation has to be performed. This usually yielded the best estimates of the parameters. Clearly, in true experiments, this ordering is not available.

In Fig 3 the resulting average values for the fraction n_{bad} of incorrectly estimated edge weights is shown as a function of the relative number r/p of single-node interventions. One can observe that with increasing number of interventions, the

quality of the averaged weight estimate increases. This is especially true for the range $r < p$ where the number of interventions is smaller than the number of nodes in the DAG. For $r > p$ the quality of the averaged estimates increases only slightly.

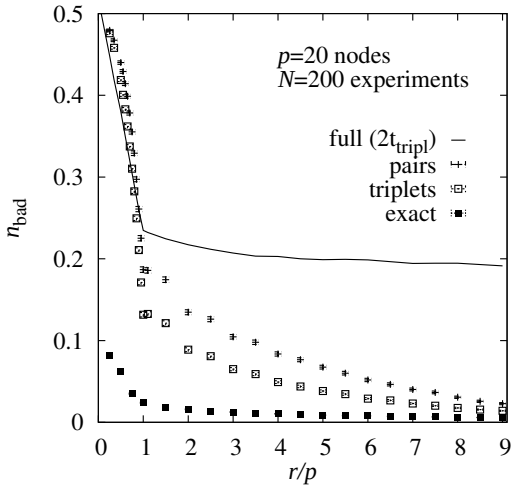

Fig 3. Topological errors. Average fraction n_{bad} of incorrectly estimated edge weights as a function of the number of interventions r per node. The data was generated for 1000 randomly generated DAGs of size $p = 20$ nodes. The results are obtained using four different sampling approaches using the true maximum likelihoods (**full**), the pair BS probabilities (**pair**), the triplet BS probabilities (**triplet**) and using just the exact ordering of nodes of the DAGs. The running time for the sampling using the true maximum likelihoods was restricted to two times the CPU time of the triplet sampling.

Also one can observe that the full sampling, due to the limited number of MCMC steps performed, is the worst approach, except for a very small number of interventions, where the estimates are bad anyway. Furthermore, the quality of the estimates is much better when using the triplet probabilities as compared to the pair probabilities. Still, one cannot reach the quality of the estimate which we obtained when using the single true ordering. Thus, the result from the true ordering constitutes a lower limit for what is possible using sampling.

As mentioned already, the details of the results for n_{bad} depend on the choice of the threshold values w_0 and w_1. For this reason we determined the ROC for whether a weight is considered non-zero or not. For this purpose we used a simple thresholding, i.e., a weight for edge i, j is considered non-zero if is estimate exceeds a threshold $\hat{w}_{i,j} \geq w_2$. Thus, for a large threshold value, only few weights will be considered as nonzero, while for a small value of w_2 many weights will be considered as non-zero. Since we know the weights used to generate the data, we know those edges which are correctly identified as being non-zero, i.e., the number of *true positives* N_{pos}, as well as the number of incorrectly as being non-zero identified edges, the *false positives* N_{false}. For the corresponding normalized rates $n_{\mathrm{pos}} = N_{\mathrm{pos}}/(p(p-1))$ and $n_{\mathrm{false}} = N_{\mathrm{false}}/(p(p-1))$, the function $n_{\mathrm{pos}}(n_{\mathrm{false}})$ can be obtained by varying w_2. This is the actual ROC curve. The steeper it grows for small values of n_{false}, i.e., the more true positives are found at the cost of accepting false negative estimates, the better is the determination of the non-zero edge weights. Thus, the *area* A_{ROC} under the ROC (AUROC) is a measure for the quality of the estimate. Since the AUROC is a number obtained via the variation of the threshold w_2 it has the advantage of being

parameter-free. Due to the normalization, the AUROC is bounded by one, which is the optimum case of finding all true-positive non-zero weights without false-positive ones.

In Fig 4 the AUROC is shown for the same data of the $p = 20$ complete DAGs. Clearly, with increasing number r of interventions, the AUROC grows. The increase is strongest for values $r < p$, beyond this point the increase in the quality of the estimates is much smaller. One can also observe that again the triplet-based sampling outperforms the pair-based sampling. Also, the sampling using the true maximum likelihoods, restricted to about two times the numerical effort of the triplet-based approach, is better for about $r < 0.8p$ but worse for $r > 0.8p$, confirming the previous results.

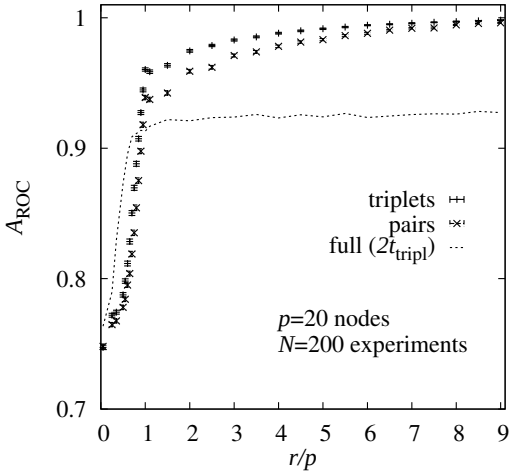

Fig 4. AUROC. Area A_{ROC} under ROC curve (AUROC) for estimating non-zero edge weights as a function of the number of interventions r per node. The results are obtained using three different sampling approaches using the true maximum likelihoods (**full**), the pair BS probabilities (**pair**), and the triplet BS probabilities (**triplet**). The running time for the sampling using the true maximum likelihoods was restricted to two times the CPU time of the triplet sampling.

We also considered diluted DAGs. In Fig 5, the number n_{bad} of strongly incorrectly-estimated edge weights is shown as a function of relative number r/p of interventions for the case of diluted DAGs which exhibit one average $c = 6$ neighbours, which is less than one third compared to the case of the complete graphs. Here, the results of the pair and triplet-based sampling approaches are much closer to each other, but the general trend remains, showing that the triplet-based sampling outperforms the pair-based sampling, and the full likelihood-based sampling for a comparable numerical effort.

The facts that the results of the pair and triplet-based approaches are closer to each other can be some expected, because the effective number of parameters to be estimated is smaller, thus the corresponding likelihoods or probabilities will be closer to each other. Thus, we also studied larger DAGs with $p = 50$ nodes. Here we generated $N = 500$ experimental outcomes per node for each DAG. For the MCMC sampling we used again 100 independent runs for the pair-based and the triplet-based sampling, 10 independent runs for the sampling based on the true maximum likelihoods ℓ_{max}. For the former two, we used 15100 MC steps for equilibration and 100 steps for measurement, for each of the independent runs. For the sampling based on ℓ_{max}, due to its expensive $O(p^6)$ computation, we could perform only 25 MCMC steps in order to consume about two times the CPU time needed for the triplet-based

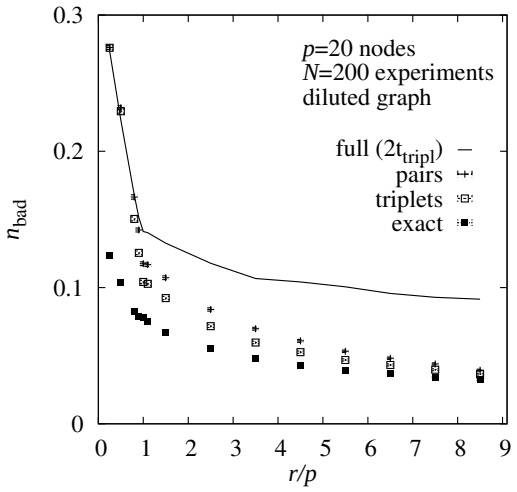

Fig 5. Topological errors for diluted DAGs. For a diluted graph with $p = 20$ nodes: Average fraction n_{bad} of incorrectly estimated edge weights as a function of the number of interventions r per node. The results are obtained using four different sampling approaches using the true maximum likelihoods (**full**), the pair BS probabilities (**pair**), the triplet BS probabilities (**triplet**) and using just the exact ordering of nodes of the DAGs. The running time for the sampling using the true maximum likelihoods was restricted to two times the CPU time of the triplet sampling.

sampling.

The corresponding results of $n_{\mathrm{bad}}(r/p)$ for complete graphs are shown in Fig 6. Here the differences between the approaches are indeed larger compared to the $p = 20$ case, but the general trend is confirmed that the triplet-based approach outperforms the pair-based approach, which in turn outperforms the exact sampling. The results when just using the original causal ordering form again a lower bound on what can be achieved for n_{bad}.

Greedy approach

Finally, to allow for a comparison of the approaches from a different perspective, we consider the case where we do not aim at estimating parameters of the model, e.g., the weights of the causal interactions. Instead we focus on the estimation of the causal ordering itself which was used to numerically generate the data. This is a much harder task. One approach could be to enumerate all orderings and take that one exhibiting the largest maximum likelihood $\ell_{\max}$ as an estimate of the correct ordering. This represents a double-nested optimization: For each given ordering, the exact maximum likelihood is obtained in a straightforward way as explained in Section "Estimating model parameters". This has to be repeated for all possible orderings. Thus it would require an numerical effort $O(p!)$ for system consisting of p nodes, i.e. more than exponentially.

This is not feasible for systems beyond exhibiting few nodes. Therefore, we follow a different approach here. We apply a *greedy* construction of an estimate for the true ordering.

For this purpose, we again use the pair-wise and the triplet-wise probabilities, respectively. This works as follows: We initialize the ordering with the a single pair (i, j) of nodes, for the pair-based approach, or the triplet (i, j, k) of nodes, for the

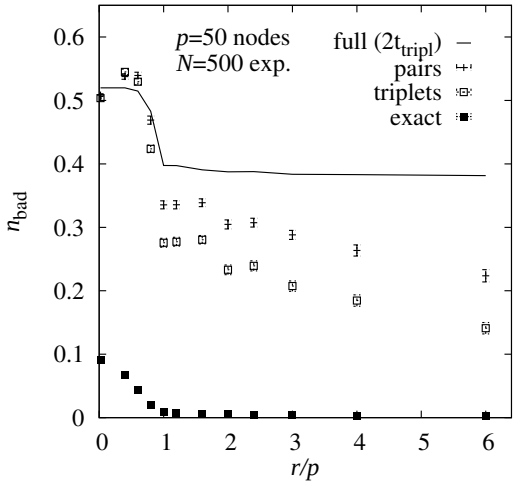

Fig 6. Topological errors for larger complete DAGs. Average fraction n_{bad} of incorrectly estimated edge weights as a function of the number of interventions r per node for complete graph of $p = 50$ nodes. The results are obtained using four different sampling approaches using the true maximum likelihoods (**full**), the pair BS probabilities (**pair**), the triplet BS probabilities (**triplet**) and using just the exact ordering of nodes of the DAGs. The running time for the sampling using the true maximum likelihoods was restricted to two times the CPU time of the triplet sampling.

triplet-based approach, which exhibits the largest value of pair preference $\pi_{i,j}$ or the largest triplet preference $\rho_{i,j,k}$, respectively. Next, iteratively nodes are included in the ordering, one-by-one, such that the resulting combined BS probability, evaluated according to (12) or (15), respectively, is largest. The construction is finished when a full ordering of length p is obtained. This means in each step, one chooses among $O(p)$ nodes and $O(p)$ insertion positions, i.e., one considers $O(p^2)$ choices. Also, like in the MCMC steps, one has to consider $O(p)$ terms when evaluation the influence of on the pair-wise likelihood for each extension of the ordering. Similarly, for the triplet-based greed approach, each insertion choice requires the calculation of $O(p^2)$ factors. This leads to an overall running-time of $O(p^3)$ for the pair-based and $O(p^4)$ for the triplet-based greedy approaches.

To evaluate the resulting ordering, we compared it to the original ordering which was used to generate the data, while again varying the number of interventions in the same way as before. For the comparison, we used *Kendal's tau-distance* K, which is defined for two orderings $\mathbf{o}$, $\mathbf{o}'$ as the number of pairs of nodes which appear in different relative orders in the two orderings.

$$K(\mathbf{o}, \mathbf{o}') = |\left\{\{i,j\} | o_i < o_j \wedge o'_i > o'_j\right\}| \tag{20}$$

Note that Kendal's *tau* distance is also called *bubble-sort distance* because it states the number of elementary sorting swaps to arrange one ordering in the order of the other given ordering. The maximum possible value is $p(p-1)/2$ for p elements.

In Fig 7 the average of K is shown for complete DAGs with $p = 20$ nodes as a function of the number r of interventions. Here a larger (quite unrealistic) number of $N = 1000$ experiments is numerically performed. This allowed us to change the number r of interventions in a very large range such that we could also access the region were the greedy approach actually determines the true ordering with high

probability. One observes that indeed when increasing the number of interventions, the greedy orderings resemble the original DAG ordering more and more. Compared to the maximum value $p(p-1)/2 = 190$, the orderings found by the greed approaches are quite similar to the true ordering. Interestingly, as seen in the inset of Fig 7, for about $O(50)$ interventions, the greedy approaches finds the true ordering, among $20! \approx 2 \times 10^{18}$ ones, in more than half of all cases! This is in particular striking, because apparently the numerical effort ($O(p^3)$ or $O(p^4)$) as well as the number of interventions (linear) appears to grow only polynomially with the number of nodes. Nevertheless, for any value of r, the triplet-based greedy approach clearly outperforms the pair-based approach significantly. This confirms the result found above using the MCMC sampling.

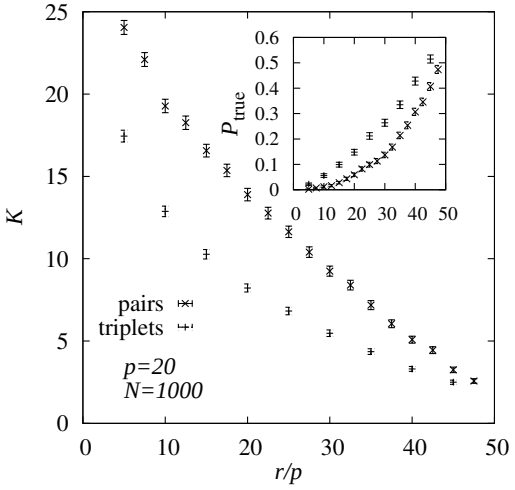

Fig 7. Kendall's tau. Average Kendall's τ distance K to the original ordering for ordering obtained via applying the greedy approach for pairs and triplet Babington-Smith probabilities. The average is obtained over 1000 DAGs of size $p = 20$, while varying the number r of interventions performed within the numerically generated measurement data. The inset shows the frequency P_{true} that the original DAG ordering is found, i.e., the frequency that $K = 0$.

Summary and Discussion

To summarize, we studied the estimation of causal orderings and corresponding parameters in sampled data using interventions. In particular we compared pair-wise Babington-Smith sampling, which was discussed before [7] with triplet-wise sampling which we introduced in this work. All results show a much better performance for the triplet sampling approach. When limiting the numerical effort to about two times the running time of the triplet sampling, a sampling using the full maximum likelihood turned out to be much worse than both pair- and triplet-wise sampling.

These results were confirmed for various cases: for data from actual biological measurements as well as for artificial data generated in a controlled way for a DAG-based Gaussian causal model. We studied small and larger DAGs, as well as completely connected and diluted ones. The general result stays also the same independently of whether one compares the estimated weight parameters directly, uses thresholding to find correct estimates, or performs an ROC analysis of the estimated nonzero weights. Also when restricting the analysis to just the prediction of the

orderings, the triplet approach turns out to be much more efficient than the pair approach.

Therefore, the triplet-based approach appears to be well balanced: It is computationally efficient enough such that long MCMC chains can be easily generated, for systems large enough for practical applications. This would be impossible when using a sampling based on the full likelihood, except for small systems. On the other hand, in combination with the final computation of the true maximum-likelihood estimators for a comparable small subset of "best" configurations, the triplet approach allows for accurate results, much better than the pair-based approach.

In principle, one could also try a similar approach based on quadruplets of nodes. Nevertheless, in contrast to when moving from pairs to triplets, we believe that this will not result in a considerable increase of accuracy. One reason, e.g., is that for the study of the Rosetta data set, the accuracy using the triplet sampling was comparable to the exact evaluation for a *finite* subset of orderings with the hightest exact likelihoods (see Fig 2). One the other hand, the numerical effort for evaluating the Metropolis criterion in each MCMC step would increase to $O(p^3)$ for a quadruplet-based algorithm.Thus, the triplet approach seems to be multi-criterion (accuracy, numerical demand) efficient within the hierarchy of approaches based on n-nodes sub graphs.

On the other hand, for further applications, it might be fruitful to perform a MCMC chains which consist of mixture of triplet-wise (first part of chain) and full maximum-likelihood sampling (last part). But this is beyond of the scope of the current study.

Furthermore, it could be interesting to study more thoroughly the point $r = p$ where most results exhibit a notable change of characteristics. It could be interesting whether this change corresponds to a kind of information-driven phase transition, similar to neural networks where the memory of a network changes if the amount of data to be learned is increased beyond a threshold. We have already started research in this direction.

Acknowledgments

A visit of GN to the University of Oldenburg, during which the project was set up, was supported by the Graduiertenkollen 1885 "Molecular Basis of Sensory Biology" funded by the Deutsche Forschungsgemeinschaft (DFG). AKH is grateful to the Foundation Sciences Mathḿatiques de Paris (FSMP) for providing financial support for a three month (February to April 2016) visit at the LPTMA of the Université Pierre et Marie Curie during which this project mainly conducted.

The simulations were performed on the HERO cluster of the University of Oldenburg jointly funded by the DFG (INST 184/108-1 FUGG) and the ministry of Science and Culture (MWK) of the Lower Saxony State.

References

1. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–441.

2. Pearl J. Causality. Cambridge university press; 2009.

3. Maathuis MH, Kalisch M, Bühlmann P, et al. Estimating high-dimensional intervention effects from observational data. The Annals of Statistics. 2009;37(6A):3133–3164.

4. Maathuis MH, Colombo D, Kalisch M, Bühlmann P. Predicting causal effects in large-scale systems from observational data. Nature Methods. 2010;7(4):247–248.

5. Rau A, Jaffrézic F, Nuel G. Joint estimation of causal effects from observational and intervention gene expression data. BMC systems biology. 2013;7(1):1.

6. Hauser A, Bühlmann P. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015;77(1):291–318.

7. Nuel G, Rau A, and FJ. Using Pairwise Ordering Preferences to Estimate Causal Effects in Gene Expression from a Mixture of Observational and INtervention Experiments. Qulaity Techn Quant Manag. 2014;11:23–37.

8. Hartmann AK. Big Practical Guide to Computer Simulations. Singapore: World Scientific; 2015.

9. Critchlow DE, Fligner MA, Verducci JS. Probability models on rankings. J Math Psych. 1991;35:294–318.

10. Joe H, Verducci JS. On the Babington-Smith class of models for rankings. Cell. 1993;102:109–126.

11. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell. 2000;102:109–126.

12. Pe´er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics. 2001;17 (suppl 1):S215–S224.