Poisson Graphical Granger Causality by Minimum Message Length

Kateřina Hlaváčková-Schindler[1,2] and Claudia Plant[1,3]

[1] Faculty of Computer Science, University of Vienna, Vienna, Austria
[2] Institute of Computer Science of the Czech Academy of Sciences,
Prague, Czech Republic, katerina.schindlerova@univie.ac.at
[3] ds:UniVie, University of Vienna, Vienna, Austria, claudia.plant@univie.ac.at

Abstract. Graphical Granger models are popular models for causal inference among time series. In this paper we focus on the Poisson graphical Granger model where the time series follow Poisson distribution. We use minimum message length principle for determination of causal connections in the model. Based on the dispersion coefficient of each time series and on the initial maximum likelihood estimates of the regression coefficients, we propose a minimum message length criterion to select the subset of causally connected time series with each target time series. We propose a genetic-type algorithm to find this set. To our best knowledge, this is the first work on applying the minimum message length principle to the Poisson graphical Granger model. Common graphical Granger models are usually applied in scenarios when the number of time observations is much greater than the number of time series, normally by several orders of magnitude. In the opposite case of "short" time series, these methods often suffer from overestimation. We demonstrate in the experiments with synthetic Poisson and point process time series that our method is for short time series superior in precision to the compared causal inference methods, i.e. the heterogeneous Granger causality method, the Bayesian causal inference method using structural equation models LINGAM and the point process Granger causality.

Keywords: Granger causality · Poisson graphical Granger model · minimum message length · ridge regression for GLM

1 Introduction

Granger causality is a popular method for causality analysis in time series due to its computational simplicity. Its application to time series with non-Gaussian distribution can be however misleading. Recently, Behzadi et al. in [2] proposed the heterogeneous graphical Granger Model (HGGM) for detecting causal relations among time series having a distribution from the exponential family, which includes a wider class of common distributions. HGGM employs regression in generalized linear models (GLM) with adaptive Lasso as a variable selection method and applies it to time series with a given lag. The approach allows to apply causal inference among time series with discrete values. Poisson graphical

Granger model (PGGM) is a special case of HGGM for detecting Granger-causal relationships among $p \geq 3$ Poisson processes. Each process in the model, represented by time series, is a count. A count process can be e.g. a process of events such as the arrival of a telephone call at a call centre in a time interval. Poisson processes can serve as models of point process data, including neural spike trains, [4]. Poisson graphical Granger model may be appropriate when investigating temporal interactions among processes as e.g. the number of transit passengers of an airport within a time period or in criminology, when temporal relationships among various crimes in some time interval are investigated.

In this paper we approach the inference in the Poisson graphical Granger model by the principle of minimum message length (MML).

- We use minimum message length principle for determination of causal connections in the Poisson graphical Granger model.
- For the highly collinear design matrix of the model we define a corrected form of the Fisher information matrix using the ridge penalty.
- Based on the dispersion coefficient of each time series and on the initial maximum likelihood estimates of the regression coefficients, we propose a minimum message length criterion to select the subset of causally connected time series with each target time series.
- We propose a genetic-type algorithm to find this set.
- To our best knowledge, this is the first work on applying the minimum message length principle to the Poisson graphical Granger model.
- We demonstrate experimentally that our method is superior in precision to the compared causal inference methods, i. e. the heterogeneous Granger causality method, the Bayesian causal inference method using structural equation models LINGAM [23] and the point process Granger causality in the case of short data, i.e. when the number of time observations is approximately of the same order as the number of time series.

The paper is organized as follows. Section 2 presents preliminaries, concretely Granger causality and the Poisson graphical Granger model. Section 3 presents the PGGM as an instance of multiple Poisson regression. The MML code for PGGM is computed and the main theorem is stated in Section 4. The algorithm to compute the MML code for PGGM and the genetic algorithm for variable selection in PGGM are explained in Section 5. Related work is discussed in Section 6. Our experiments are in Section 7. Section 8 is devoted to conclusions and the derivation of the criterion can be found in Appendix.

2 Preliminaries

Relevance of Granger Causality Since its introduction, there has been lead a criticism of Granger causality, since it e.g. does not take into account counterfactuals, [12], [17]. As its name implies, Granger causality is not necessarily true causality. In defense of his method, Granger in [6] wrote: "Possible causation is not considered for any arbitrarily selected group of variables, but only for variables for which the researcher has some prior belief that causation is, in some sense, likely." In other words, drawing conclusions about the existence of a causal relation between time series and about its direction is possible only if theoretical knowledge of mechanisms connecting the time series is accessible. Nevertheless as confirmed by a recent Nature publication [16], if the theoretical background of investigated processes is insufficient, methods to infer causal relations from data rather than knowledge of mechanisms (Granger causality including) are helpful. These methods can also make possible to perform credible analyses with large amount of observational time data, e.g. in social networks [11], since they are less costly than common epidemiological or marketing research approaches.

Graphical Granger model The (Gaussian) graphical Granger model extend the autoregressive concept of Granger causality to $p \geq 2$ time series and time lag $d \geq 1$ [1]. Let $x_1^t, \ldots, x_p^t$ be p time series, $t = 1, \ldots, n$. Consider the vector autoregressive (VAR) models with lag d for $i = 1, \ldots, p$

$$x_i^t = X_{t,d}^{Lag} \beta_i' + \varepsilon_i^t \tag{1}$$

where $X_{t,d}^{Lag} = (x_1^{t-d}, \ldots, x_1^{t-1}, \ldots, x_p^{t-d}, \ldots, x_p^{t-1})$ and β_i be a matrix of the regression coefficients and ε_i^t be white noise. One can easily show that $X_{t,d}^{Lag} \beta_i' = \sum_{j=1}^p \sum_{l=1}^d x_j^{t-l} \beta_j^l$. One says the time series x_j Granger–causes the time series x_i for the given lag d, denote $x_j \rightarrow x_i$ for $i, j = 1, \ldots, p$ if and only if at least one of the d coefficients in $j - th$ row of β_i in (1) is non-zero.

Poisson graphical Granger model The Poisson graphical Granger model has the form

$$x_i^t \approx \lambda_i^t = \exp(X_{t,d}^{Lag} \beta_i') = \exp(\sum_{j=1}^{p} \sum_{l=1}^{d} x_j^{t-l} \beta_j^l) \tag{2}$$

for x_i^t, $i = 1, \ldots, p, t = d+1, \ldots, n$ having a Poisson distribution. Applying the HGGM approach to the case, when the link function for each process x_i is function exp, problem (2) can be solved as

$$\hat{\beta}_i = \arg\min_{\beta_i} \sum_{t=d+1}^{n} (x_i^t - \exp(X_{t,d}^{Lag} \beta_i'))^2 + \rho_i R(\beta_i) \tag{3}$$

for a given lag $d > 0$ and all $t = d+1, \ldots, n$ with $R(\beta_i)$ adaptive Lasso penalty function. (The sign $'$ denotes a transpose of a matrix). One says, the time series x_j Granger–causes the time series x_i for the given lag d, denote $x_j \rightarrow x_i$ for $i, j = 1, \ldots, p$ if and only if at least one of the d coefficients in $j - th$ row of $\hat{\beta}_i$ of the solution of (3) is non-zero [2].

3 Poisson Graphical Granger Model as Multiple Poisson Regression

In this section we will derive the Poisson Granger model (2) with a fixed lag d as an instance of a multiple Poisson regression with a fixed design matrix. Consider the full model for p Poisson variables x_i^t and (integer) lag $d \geq 1$ corresponding to the optimization problem (2). To be able to use the maximum likelihood (ML) estimation over the regression parameters, we reformulate the matrix of lagged time series $X_{t,d}^{Lag}$ from (1) into a fixed design matrix form. Assume $n - d > pd$ and denote $x_i = (x_i^{d+1}, x_i^{d+2}, \ldots, x_i^n)$. We construct the $(n-d) \times (d \times p)$ design matrix

$$X = \begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_p^d & \ldots & x_p^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_p^{d+1} & \ldots & x_p^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_p^{n-1} & \ldots & x_p^{n-d+1} \end{bmatrix} \tag{4}$$

and a $1 \times (d \times p)$ vector $\beta_i = (\beta_1^1, \ldots, \beta_1^d, \ldots, \beta_p^1, \ldots, \beta_p^d)$. We can see that problem

$$x_i' \approx \lambda_i = \exp(X\beta_i') \tag{5}$$

is equivalent to problem (2) in the matrix form where we mean by exp a function operating on each coordinate $i = d+1, \ldots, n$ and $\lambda_i = (\lambda_i^{d+1}, \ldots, \lambda_i^{d+1})$.

Denote now by $\gamma_i \subset \mathit{\Gamma} = \{1, \ldots, p\}$ the subset of indices of regressor's variables and $k_i := |\gamma_i|$ its cardinality. Let $\beta_i := \beta_i(\gamma_i) \in \mathbb{R}^{1 \times (d \times k_i)}$ be the vector of unknown regression coefficients with a fixed ordering within the γ_i subset. For illustration purposes and without lack of generality we can assume that the first k_i indices out of p vectors belong into γ_i. Considering only the columns from matrix X in (4) corresponding to γ_i, we define the $(n-d) \times (d \times k_i)$ matrix of lagged vectors with indices from γ_i as

$$X_i := X(\gamma_i) = \begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_{k_i}^d & x_{k_i}^{d-1} & \ldots & x_{k_i}^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_{k_i}^{d+1} & x_{k_i}^d & \ldots & x_{k_i}^2 \\ x_1^{d+2} & \ldots & x_1^3 & \ldots & x_{k_i}^{d+2} & x_{k_i}^{d+1} & \ldots & x_{k_i}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_{k_i}^{n-1} & x_{k_i}^{n-2} & \ldots & x_{k_i}^{n-d+1} \end{bmatrix} \tag{6}$$

The problem (5) for explanatory variables with indices from γ_i is expressed as

$$x_i' \approx \lambda_i = E(x_i'|X_i) = \exp(X_i\beta_i') \tag{7}$$

or alternatively

$$\log(x_i') \approx \log(\lambda_i) = \log(E(x_i'|X_i)) = X_i\beta_i' \tag{8}$$

with $\beta_i := \beta_i(\gamma_i)$ to be a $1 \times (dk_i)$ matrix of unknown coefficients and log operates on each coordinate. Wherever it it clear from context, we will simplify the notation β_i instead of $\beta_i(\gamma_i)$ and X_i instead of $X(\gamma_i)$.

4 Minimum Message Length for Poisson Granger Model

Denote Γ the set of all subsets of covariates $x_i, i = 1, \dots, p$. Assume now a fixed set $\gamma_i \in \Gamma$ of covariates with size $k_i \leq p$ and the corresponding design matrix X_i from (6). It is well known that the Poisson regression model can be still used in over- or underdispersed settings. (However the standard error for Poisson would not be correct for the overdispersed situation.) In the Poisson graphical Granger model, it is the case when for the dispersion of at least one time series holds $\phi_i \neq 1$. So using the Poisson regression model, we assume that the likelihood function for x_i in PGGM does not depend on ϕ_i. It is usual to assume that the targets x_i are independent random variables, conditioned on the features given by X_i, so that the likelihood function can be factorized into the product $p(x_i|\beta_i, X_i, \gamma_i) = \prod_{t=1}^{n-d} p(x_i^t|\beta_i, X_i, \gamma_i)$. The log-likelihood function has then the form

$$L_i := \log p(x_i|\beta_i, X_i, \gamma_i) = \sum_{t=1}^{n-d} \log p(x_i^t|\beta_i, X_i, \gamma_i). \tag{9}$$

Since X_i is highly collinear, to make the ill-posed problem for coefficients β_i a well-posed one, one can use regularization by the ridge regression for GLM (see e.g. [22]). Ridge regression requires an initial estimate of β_i which can be as the maximum likelihood estimator of (7) obtained by the iteratively reweighted least square algorithm (IRLS). For a fixed $\rho_i > 0$, for the ridge estimates of coefficients $\hat{\beta}_{i,\rho_i}$ holds

$$\hat{\beta}_{i,\rho_i} = \underset{\beta_i \in \mathbb{R}^+}{\arg\min} \{-L_i + \rho_i \beta_i' \Sigma_i \beta_i\}. \tag{10}$$

In our paper however, we will not use the GLM ridge regression in form (10). Instead, we will apply the principle of minimum description length. Ridge regression in the minimum description length framework is equivalent to allowing the prior distribution to depend on a hyperparameter (= ridge regularization parameter). To compute the message length using the MML87 approximation proposed in [20], we need the negative log-likelihood function, prior distribution over the parameters and an appropriate Fisher information matrix. [20] proposed the corrected form of Fisher information matrix for a GLM regression with ridge penalty. In our work, we will use this form of ridge regression and apply it to the Poisson graphical Granger model. In the following, we will construct the MML code for every subset of covariates in PGGM. The derivation of the criterion can be found in Appendix.

The MML criterion for PGGM *For each $d \geq 1$ assume $x_i, i = 1, \dots, p, t = 1, \dots, n$ and the estimate of the dispersion parameter $\hat{\phi}_i$ be given. Assume $\hat{\beta}_i$ be an initial solution of (7) achieved as the maximum likelihood estimate.*

(i) The causal graph of the Poisson Granger problem (7) can be inferred from the solutions of p variable selection problems, where for each $i = 1, \dots, p$, the set $\hat{\gamma}_i$ of Granger-causal variables to x_i is found.

(ii) For the estimated set $\hat{\gamma}_i$ holds

$$\hat{\gamma}_i = \arg\min_{\gamma_i \in \Gamma} \{ I(x_i, \hat{\beta}_i, \hat{\phi}_i, \hat{\rho}_i, X_i, \gamma_i) + I(\gamma_i) \} \text{ where} \tag{11}$$

$I(x_i, \hat{\beta}_i, \hat{\phi}_i, \hat{\rho}_i, X_i, \gamma_i) = \min_{\rho_i \in \mathbb{R}^+} \{ MML(x_i, \hat{\beta}_i, \hat{\phi}_i, \rho_i, X_i, \gamma_i) \}$ *where* $MML(x_i, \hat{\beta}_i, \hat{\phi}_i, \rho_i, X_i, \gamma_i)$ *is the minimum message length code of the set* γ_i *and can be expressed as*

$$MML(x_i, \hat{\beta}_i, \hat{\phi}_i, \rho_i, X_i, \gamma_i) = -L_i + \frac{1}{2} \log |X_i' W_i X_i + \rho_i \Sigma_i| - \frac{1}{2} \log |\Sigma_i| \tag{12}$$

$+\frac{k_i}{2} \log(\frac{2\pi}{\rho_i}) + (\frac{\rho_i}{2\hat{\phi}_i}) \hat{\beta}_i' \Sigma_i \hat{\beta}_i + \frac{1}{2} \log(n-d) - \frac{k_i+1}{2} \log(2\pi) + \frac{1}{2} \log((k_i+1)\pi)$ *where* $|\hat{\gamma}_i| = k_i$, Σ_i *is the unity matrix of size* $dk_i \times dk_i$, W_i *is a diagonal matrix with entries* $W_i(t) = \lambda_i^t = \exp(X_i \hat{\beta}_i')^t$, $t = 1, \ldots, n-d$ *for Poisson* x_i *and* $W_i(t) = \lambda_i^t = [x_i^{d+t} - \exp(X_i \hat{\beta}_i')]^2$ *for over- or underdispersed Poisson* x_i, $L_i = \log(p(x_i|\hat{\beta}_i, X_i, \gamma_i)) = \sum_{t=d+1}^{n} x_i^t [X_i \hat{\beta}_i']^t - \exp([X_i \hat{\beta}_i']^t) - \log(x_i^t!)$ *and* $I(\gamma_i) = \log \binom{p}{k_i} + \log(p+1)$.

Remark: Schmidt and Makalic in [21] compared AIC_c criterion with MML code for generalized linear models. We constructed the AIC_c criterion also for PGGM. However this criterion requires pseudoinverse of a matrix multiplication which includes matrices X_i. Since X_is are highly collinear, these matrix multiplications had in our experiments very high condition numbers. This consequently lead the AIC_c criterion for PPGM to spurious results and therefore we do not report them in our paper.

5 Variable Selection in Poisson Graphical Granger Model

For both Poisson and overdispersed Poisson cases we consider the family of models $M(\gamma_i) := \{p(x_i|\beta_i, X_i, \gamma_i), \gamma_i \in \Gamma\}$ defined by Poisson densities $p(x_i|\beta_i, X_i, \gamma_i)$. First, we present the procedure in Algorithm 1 which for each x_i computes the MML code for a set $\gamma_i \subset \Gamma$. Then we present Algorithm 2 for computation of $\hat{\gamma}_i$.

In general, the selection of the best structure γ_i amounts to evaluate values of $MML(\gamma_i)$ for all $\gamma_i \subset \Gamma$, i.e. for all 2^p possible subsets and then to pick the subset with which the minimum of the function was achieved. To avoid the exhaustive search approach, we find γ_i with minimum MML by the proposed genetic algorithm type procedure called MMLGA. The idea of MMLGA is as follows. Consider an arbitrary $\gamma_i \subset \Gamma$ with size k_i for a fixed i and $d \geq 1$. Define a Boolean vector Q_i of length p corresponding to a given γ_i in so that it has ones in the positions of the indices of covariates from γ_i, otherwise zeros. Define $I(Q_i) := I(\gamma_i)$ where $I(\gamma_i)$ is from (11). Genetic algorithm MMLGA executes genetic operations on populations of Q_i. In the first step a population of size m (m be an even integer), is generated randomly in the set of all 2^p binary strings (individuals) of length p. Then we select $m/2$ individuals in the

Algorithm 1 MML Code for γ_i

```
**Input**: $\gamma_i \in \Gamma, d \geq 1$, series is the matrix of $x_i^t$, $\hat{\phi}_i$ dispersion parameter,
$i = 1, \dots, p, t = 1, \dots, n-d$, $\Sigma_i$, $H$ a set of positive numbers;
**Output**: For each $i$ minimum $I(x_i, \hat{\beta}_i, \hat{\rho}_i, X_i, \gamma_i)$ over $H$ is found;
**for all** $x_i$ **do**
    // Construct the d-lagged matrix $X_i$ with time series with indices from $\gamma_i$.
    //Compute matrix $W_i$.
    **for all** $\rho_i \in H$ **do**
        // Compute $L_i$ from (9).
        // Find the initial estimates of $\hat{\beta}_i$.
        //Compute $MML(x_i, \hat{\beta}_i, \rho_i, X_i, \gamma_i)$ from (12).
    **end for**// to $\rho_i$
    // Compute $I(x_i, \hat{\beta}_i, \hat{\rho}_i, X_i, \gamma_i) = \min_{\rho_i \in R^+} MML(x_i, \hat{\beta}_i, \rho_i, X_i, \gamma_i)$.
**end for**// to $x_i$
**return** $I(x_i, \hat{\beta}_i, \hat{\rho}_i, X_i, \gamma_i)$ for each $i$.
```

current population with the lowest value of (11) as the elite subpopulation of parents of the next population. For a predefined number of generated populations n_g, the crossover operation of parents and the mutation operation of a single parent are executed on the elite to create the rest of the new population. A mutation corresponds to a random change in Q_i and a crossover combines the vector entries of a pair of parents. After each run of these two operations on a current population, the current population is replaced with the children with the lowest value of (11) to form the next generation. The algorithm stops after the number of population generations n_g is achieved. The algorithm MMLGA is summarized in Algorithm 2. Our code in Matlab is publicly available at: `https://t1p.de/b3gf`.

5.1 Computational Complexity of MMLGA

For computation of $I(x_i, \hat{\beta}_i, \hat{\rho}_i, X_i, \gamma_i)$ we used Matlab function *fminsearch*. It is well-known that the upper bound of the computational complexity of a genetic algorithm is of order of the product of the size of an individual, of the size of each population, of the number of generated populations and of the complexity of the function to be minimized. Therefore an upper bound of the computational complexity of MMLGA for p time series, size p of an individual, m the population size and n_g the number of population generations is $\mathcal{O}(pmn_g) \times O(\textit{fminsearch}) \times p$ where $O(\textit{fminsearch})$ can be also estimated. The highest complexity in *fminsearch* has the computation of the Hessian matrix, which is the same as for the Fisher information matrix (our matrix W_i) or the computation of the determinant. The computational complexity of Hessian for i fixed for $(n-d) \times (n-d)$ matrix is $\mathcal{O}(\frac{(n-d)(n-d+1)}{2})$. An upper bound on complexity of determinant in (12) is $\mathcal{O}((pd)^3)$. As before we assume $n - d \geq pd$. Denote $M = \max\{pd, (n-d+1)\}$. Then holds also $M^3 \geq \frac{(M-1)M}{2}$. Since we have p optimization functions, our upper bound on the computational complexity of MMLGA is then $\mathcal{O}(p^2 m n_g M^3)$.

Algorithm 2 MMLGA

```
**Input**: $\Gamma$, $d \geq 1, p, n_g, m$ an even integer, $z \leq p$ position for off-spring;
series is the matrix of $x_i^t, i = 1, \ldots, p, t = 1, \ldots, n - d$;
**Output**: $Adj$ := adjacency matrix of the output causal graph;
// For every $x_i$ $Q_i$ with minimum of (11) is found;
**for all** $x_i$ **do**
  Create initial population $\{Q_i^j, j = 1, \ldots, m\}$ at random;
  Compute $I(Q_i^j) := I(x_i, \hat{\beta}_i, \hat{\rho}_i, X_i, Q_i^j) + \binom{p}{k_i^j} + \log(p+1)$ for each $j = 1, \ldots, m$
  where $k_i^j$ is the number of ones in $Q_i^j$; v:=1;
  **while** $v \leq n_g$ **do**
    u:=1;
    **while** $u \leq m$ **do**
      Sort $I(Q_i^j)$ ascendingly and create the elite population; By crossover of $Q_i^j$
      and $Q_i^r$, $r \neq j$ create children and add them to elite; Compute $I(Q_i^j)$ for each
      $j$; Mutate a single parent $Q_i^j$ at a random position; Compute $I(Q_i^j)$ for each
      $j$; Add the children with minimum $I(Q_i^j)$ until the new population not filled;
      u:=u+1;
    **end while**// to $u$
    v:=v+1;
  **end while**// to $v$
**end for**// to $x_i$
The $i - th$ row of Adj: $Adj_i := Q_i$ with min of (11)
**return** ($Adj$)
```

6 Related Work

The minimum message length (MML) is an information theoretic principle based on the statistical inference and data compression. The key idea is, if a statistical model compresses data, then the model has (with a high probability) captured regularities in the data. The MML principle selects the model which most compresses the data (i.e. the one with the "shortest message length") as the most descriptive for the data. To be able to decompress this representation of the data, the details of the statistical model used to encode the data must also be part of the compressed data string. The calculation of the exact message is an NP hard problem, however the most widely used less computationally intensive is the Wallace-Freeman approximation called MML87 [25].

Compression schemes for Poisson regression have been already studied in the framework of generalized linear models (GLM). Hansen and Yu 2003 in [7] derived objective functions for one-dimensional GLM regression by the minimum description principle. Schmidt and Makalic in [21] used MML87 to derive the MML code of a multivariate GLM ridge regression. The mentioned codes cannot be however directly used for a Granger model due to the lag in the time series and the highly collinear matrix of covariates. To our best knowledge, compression criteria for Poisson graphical Granger model has not been published yet. Other papers inferring Granger causality by MDL are [13], [3], [14]. The inference in this papers is however done for the bivariate Granger causality and the extension to graphical Granger methods is not straightforward.

Kim et al. in [10] proposed the statistical framework Granger causality (SFGC) that can operate on point processes, including neural-spike trains. The proposed framework uses multiple statistical hypothesis testing for each pair of involved neurons. A pair-wise hypothesis test was used for each pair of possible connections among all time series and the false discovery rate (FDR) applied.

For a fair comparison with our method we selected causal inference methods which are designed for $p \geq 3$ non-Gaussian processes. In our experiments, we used SFGC as a comparison method and the publicly available point process time series provided by the authors. As another comparison method we selected the method LINGAM from Shimizu et al. [23] which estimates a causal structure in Bayesian networks among non-Gaussian time series using structural equation models and independent component analysis. The experiments reported in the papers with comparison methods were done only in scenarios when the number of time observations is by several orders of magnitude greater than the number of time series.

7 Experiments

We performed experiments with MMLGA on synthetically generated Poisson processes and on neural spike train data from [10]. We used the method HGGM [2], the method LINGAM [23] and the point process Granger causality SFGC [10] for comparison. To assess similarity between the target and output causal graphs by all methods, we used the commonly applied F-measure, which takes both precision and recall into account.

7.1 Implementation and Parameter Setting

The comparison method HGGM uses Matlab package *penalized* from [18] with adaptive Lasso penalty. The algorithm in this package employs the Fisher scoring algorithm to estimate the coefficients of regressions. As recommended by the author of *penalized* in [18] and employed in [2] we used adaptive Lasso with $\lambda_{max} = 5$, applying cross validation and taking the best result with respect to F measure from the interval $(0, \lambda_{max}]$. We also followed the recommendation of the authors of LINGAM in [23] and used threshold=0.05 and number of boots n/2, where n is the length of the time series. In method SFGC we used the setting recommended by the authors, the significance level 0.05 of FDR. The method SFGC is designed for binomial time series so as expected, using the generated Poisson time series as input of this method gave very low or zero F-measure, so we do not report these values in our results. For a fair comparison to MML, HGGM and LINGAM, we will examine the performance of SFGC with input binomial time series in Section 7.3.

In MMLGA, the initial estimates of β_i were achieved by the iteratively reweighted least square procedure implemented in Matlab function *glmfit*, in the same function we obtained also the estimates of the dispersion parameters of time series. (Considering initial estimates of β_i by the IRLS procedure using function *penalized* with ridge gave poor results in the experiments.) The minimization over ρ_i was done by function *fminsearch* which defined set H from Algorithm 1 as positive numbers greater or equal to 0.1.

7.2 Synthetically Generated Poisson Processes

To be able to evaluate the performance of MMLGA and to compare it to other methods, the ground truth, i.e. the target causal graph in the experiments should be known. In this series of experiments we examined randomly generated Poisson processes together with the correspondingly generated target causal graphs. The performance of MML, HGGM and LINGAM depends on various parameters including the number of time series (features), the number of causal relations in Granger causal graph (dependencies), the length of time series and finally the lag parameter. We examined causal graphs with $p = 5$ and with $p = 9$ time series. The length of generated time series was 'short', varying from 100 to 1000. We generated Poisson time series randomly. Concerning the calculation of an appropriate lag for each time series, theoretically it can be done by AIC or BIC. However, the calculation of AIC and BIC assumes that the degrees of freedom are equal to the number of nonzero parameters, which is only known to be true for the Lasso penalty [29] but not known for adaptive Lasso. In our experiments we followed the recommendation of [2] how to select the lag of time series. They observed that varying the lag parameter from 3 to 50 did not influence either the performance of HGGM nor SFGC significantly. Based on that we considered lags 3 and 4 in our experiments. For the causal graphs with $p = 5$ and with $p = 9$ we tested the performance of algorithms for number of dependencies from 6 to 9. The results of our experiments on causal graphs with 5 features ($p = 5$) are presented in Table 1. Each value in Table 1 represents the mean value of all F-measures over 20 random generations of causal graphs for length n and lag d.

Table 1: $p = 5$, average F-measure for each method, MMLGA, with $n_g = 10$, $m = 50$, HGGM with $\lambda_{max} = 5$, LINGAM with $n/2$ boots.

$d = 3, n =$	50	100	200	300	500	1000
MMLGA	**0.8**	**0.82**	**0.83**	**0.77**	**0.77**	0.73
HGGM	0.67	0.73	0.73	0.73	0.71	**0.8**
LINGAM	0.71	0.71	0.7	0.69	0.65	0.65
$d = 4, n =$	50	100	200	300	500	1000
MMLGA	**0.75**	**0.77**	**0.77**	**0.8**	**0.8**	0.67
HGGM	0.66	0.73	0.71	0.73	0.73	**0.8**
LINGAM	0.64	0.65	0.64	0.63	0.65	0.64

One can see from Table 1 that MMLGA gave significantly higher precision in terms of F-measure than both comparison methods for n up to 500. On the other hand, HGGM gave the highest F-measure for $n = 1000$ which can be for $p = 5$ considered as a scenario of a large data set. The results of our experiments with causal graphs with $p = 9$ are presented in Table 2. Each value in Table 2 represents the mean value of all F-measures over 20 random generations of causal graphs for length n and lag d.

Table 2: $p = 9$, average F-measure for each method, MMLGA, with $n_g = 10$, $m = 50$, HGGM with $\lambda_{max} = 5$, LINGAM with $n/2$ boots.

$d = 3, n =$	50	100	200	300	500	1000
MMLGA	**0.67**	**0.62**	**0.69**	**0.69**	**0.69**	0.67
HGGM	0.6	0.48	0.5	0.62	0.65	**0.82**
LINGAM	0.4	0.28	0.29	0.33	0.35	0.36
$d = 4, n =$	50	100	200	300	500	1000
MMLGA	**0.62**	**0.61**	**0.67**	**0.64**	**0.64**	0.63
HGGM	0.5	0.51	0.54	0.54	0.55	**0.8**
LINGAM	0.4	0.39	0.39	0.4	0.37	0.38

Similarly as in the experiments with $p = 5$, one can see in Table 2 for $p = 9$ that MMLGA gave significantly higher F-measure than for both comparison methods for n up to 500. HGGM gave higher F-measure than MMLGA for $n = 1000$ which can be for $p = 5$ considered as a scenario of a large data set. In both networks with $p = 5$ and $p = 9$ time series, method LINGAM had the lowest F-measures for all investigated n.

7.3 Neural Spike Train Data

In this section we examine the performance of MMLGA, SFGC, HGGM and LINGAM on time series representing the spike train data. We used the nine-neuron network with the spike train data from [10] and the corresponding target network in Figure 1-B of the paper. Based on the experimental settings described in [10], the authors generated 100,000 samples for each neuron, and the total number of spikes for each neuron ranged from 2176 through 2911, i.e. three orders of magnitude more than the number of time series. The target network corresponds thus to the long time series. For fair comparison of all methods, we compared their precision on short time series in terms of F-measure to the target network from Figure 1-B of the paper.

Spike train data is a special case of a temporal point process. A temporal point process is a stochastic time series of binary events that occur in continuous time. It can only take on two values at each point in time, indicating whether or not an event has actually occurred. When considering the data set of size

n, the point process model of a spike train for neuron i can be defined as a counting process, which can be denoted as $\{N(t), 1 \leq t \leq n\}$. A counting process represents the total number of occurrences or events that have happened up to and including time t. It is a Poisson process. We used the point process time series from [10] as input of SFGC and their Poisson representation as described above as input of methods MMLGA, HGGM and LINGAM. We experimented with short time series with n from 100 up to 1000. Table 3 gives the F-measures of the methods. Rephrasing the F-measures into percent, one can see that for $n = 500$ method MMLGA is able to reconstruct 59 % of the target causal network, while the best reconstruction results are for SFGC 12 % (achieved for $n = 900$ and $n = 1000$), for HGGM 45 % (achieved for $n = 300$) and for LINGAM 19 % (achieved for $n = 900$). One can see that MMLGA outperformed significantly the other three methods in precision measured by F-measure for almost all investigated n.

Table 3: $p = 9$, F-measure for each method, MMLGA with $d = 3$,$n_g = 30$, $m = 50$, HGGM with $\lambda_{max} = 5$, LINGAM with $n/2$ boots.

$d = 3, n =$	100	200	300	400	500	600	700	800	900	1000
MMLGA	0.27	0.43	0.45	0.51	**0.59**	0.55	0.47	0.46	0.47	0.46
SFGC	0	0.06	0	0	0	0	0.06	0	**0.12**	**0.12**
HGGM	0.34	0.44	**0.45**	0.41	0.44	0.38	0.39	0.38	0.41	0.39
LINGAM	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.33	**0.42**	0.39

7.4 Analysis of Chicago Violence Crime Dataset

Chicago's violent crime rate is substantially higher than the US average. Although national crime rates in the US have stayed near historic lows, Chicago had nearly half of 2016's increase in crimes in the US [15]. Thus any research on possible causes of the increased number of crimes is valuable for the law enforce agencies. We used the data set of 5 most frequent crimes in Chicago from [15], i.e. battery, narcotics consumption, criminal damage (= violation of property rights), theft and other offense (e.g. harassment by telephone, a weapon violation). These are yearly measurements from 2001 to 2017. Due to the small data size, this is rather a toy example. We investigated temporal interactions of these time series. No target graph was given. Our goal was to find out whether the resulting causal graphs for each test method support empirical evidence. Statistical distribution fitting test confirmed Poisson distribution of all time series. We investigated causal graphs for lags 1 to 3 (to keep the condition $n - d \geq pd$ as discussed above). Method HGGM gave for each lag a different causal graph, for $d = 1$ it gave the complete graph. So we excluded it from further analysis. Methods LINGAM and MMLGA gave only one causal graph as output for all considered lags and the causal graphs can be found (for $n_g = 10$ and $m = 30$)

in Figure 1. Focusing on crime narcotic consumptions, MMLGA outputs other offenses as causal to narcotics and narcotics causal to crime battery. Both claims support the empirical evidence, in the second case it is known that drug consumption increases the effects generating violence, as stated in the reports of the Bureau of Justice Statistics, US Department of Justice, e.g. [24]. On the other hand, LINGAM outputs battery as a cause of narcotics, which seems unrealistic. So the output of MMLGA gave a more realistic causal graph than LINGAM.

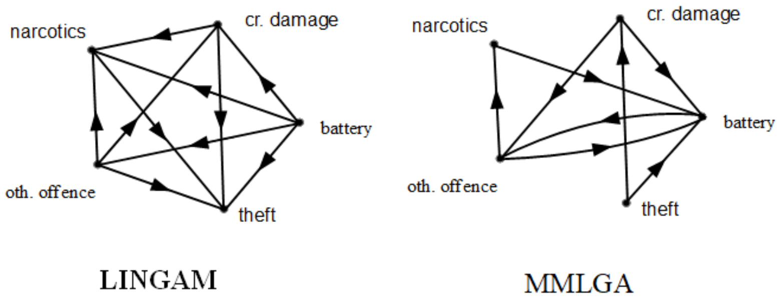

Fig. 1: Results of causal interactions among 5 most frequent crimes in Chicago for LINGAM and MMLGA.

8 Conclusions

Common graphical Granger models are usually applied in scenarios when the number of time observations is by several orders of magnitude greater than the number of time series. In the opposite case of short time series, these methods often suffer from overestimation. In this paper we used minimum message length principle for determination of causal connections in the Poisson graphical Granger model. Based on the dispersion coefficient of each time series and on the initial maximum likelihood estimates of the regression coefficients, we proposed a minimum message length criterion to select the subset of time series causal to each target time series. We used a genetic-type algorithm MMLGA to find this set. We demonstrated in the experiments on synthetic Poisson time series and on point process time series that our method is on short time series superior in precision to the compared causal inference methods, i. e. the heterogeneous Granger causality method, the Bayesian causal inference method using structural equation models LINGAM and the point process Granger causality. Both MMLGA and HGGM use penalization, MMLGA uses ridge, HGGM adaptive Lasso. The superiority of MMLGA with respect to HGGM for short time series can be explained by using the dispersion of the time series in the criterion as additional information with respect to HGGM. To our best knowledge, this is the first work on applying the minimum message length principle to the Poisson graphical Granger model. In our future work we would like to investigate other utilization of the minimum message principle for PGGM, for example for the dispersion parameter of the involved time series.

Acknowledgement: This work was supported by the Czech Science Foundation, project $GA19-16066S$.

References

1. Arnold, A., Liu, Y., Abe, N.: Temporal causal modeling with graphical Granger methods. ACM SIGKDD, 66-75 (2007).
2. Behzadi, S., Hlaváčková-Schindler, K., Plant, C.: Granger Causality for Heterogeneous Processes, PAKDD 2019.
3. Budhathoki, K., Vreeken, J.: Origo: causal inference by compression, Knowledge and Information Systems, 56, 2, 285–307 (2018).
4. Brown E.N.: Theory of point processes for neural systems. In: Chow C, et al. Methods and models in neurophysics. Paris: Elsevier, 691–726 (2005).
5. Granger, C.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 424–438 (1969).
6. Granger, C.W.: Some recent development in a concept of causality. Journal of econometrics 39(1-2):199–211 (1988).
7. Hansen, M.H., and Yu, B.: Minimum description length model selection criteria for generalized linear models. Lecture Notes-Monograph Series 145–163 (2003).
8. Hansen, P.C.: Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. SIAM Journal on Scien. and Stat. Computing 11(3):503–518 (1990).
9. Huber, P. J.: The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, 221–233. University of California Press (1967).
10. Kim, S. Putrino, D., Ghosh, S., Brown, E.N.: A Granger causality measure for point process models of ensemble neural spiking activity, PLOS Computational Biology, 1–13 (2011).
11. Kwak, H., Lee, C. Park, H., Moon, S.: What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, 591–600. ACM (2010).
12. Mannino, M., Bressler, S.L.: Foundational perspectives on causality in large-scale brain networks. Physics of life reviews 15:107–123 (2015).
13. Marx, A. and Vreeken, J.: Telling cause from effect using MDL-based local and global regression, IEEE ICDM,307–316 (2017).
14. Marx, A. and Vreeken, J.: Causal Inference on Multivariate and Mixed-Type Data, ECML PKDD 2018, 655–671 (2018).
15. Mangipudi, V.: Analysis of crimes in Chicago 2001-2017. `https://rstudio-pubs-static.s3.amazonaws.com/294927b602318d06b74e4cb2e6be336522e94e.html` Accessed Feb 21,2020.
16. Marinescu, I.E., Lawlor, P.N., Kording, K.P.: Quasi-experimental causality in neuroscience and behavioural research. Nature Human Behaviour 1 (2018).
17. Maziarz, M.: A review of the granger-causality fallacy. The journal of philosophical economics: Reflections on economic and social issues 8(2):86–105 (2015).

18. McIlhagga, W. H.: penalized: A MATLAB toolbox for fitting generalized linear models with penalties. Journal of Statistical Software. 72(6) (2016).
19. Peterson, L.E.: PIRLS: Poisson iteratively reweighted least squares computer program for additive, multiplicative, power, and non-linear models. J Stat Software, 2, 1-28 (1997).
20. Schmidt, D.F., Makalic, E.: MML invariant linear regression. In Advances in Artificial Intelligence, 312–321 (2009).
21. Schmidt, D.F., Makalic, E.: Minimum Message Length Ridge Regression for Generalized Linear Models. In Australasian Joint Conference on Artificial Intelligence, pp. 408-420, Springer, Cham (2013).
22. Segerstedt, B.: On ordinary ridge regression in generalized linear models. Communications in Statistics-Theory and Methods, 21(8), 2227-2246 (1992).
23. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12: 1225–1248 (2011).
24. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. `https://www.bjs.gov/content/pub/pdf/DRRC.PDF`
25. Wallace, C.S., Freeman, P.R: Estimation and inference by compact coding, Journal of the Royal Statistical Society: Series B, **49**(3), 240–252, Wiley Online Library (1987).
26. Wong, C.K., Makalic, E., Schmidt, D.F.: Minimum message length inference of the Poisson and geometric models using heavy-tailed prior distributions. Journal of Mathematical Psychology 83:1–11 (2018).
27. Zhou, D., Xiao, Y., Zhang, Y., Xu, Z., Cai, D.: Granger causality network reconstruction of conductance-based integrate-and-fire neuronal systems. PloS one, 9(2) (2004).
28. Zou, H.: The adaptive lasso and its oracle property. Journal of the American Statistical Association 1418–1429 (2008).
29. Zou, H., Hastie, T., Tibshirani, R.: On the "degrees of freedom" of the lasso. The Annals of Statistics, 35(5), 2173-2192 (2007).

9 Appendix

Derivation of the MML Criterion for PGGM

Assume p independent Poisson random variables expressed by time series with lag $d > 0$ $x_i^t, t = d+1, \dots, n$, and the problem (7). Assume the estimate $\hat{\phi}_i$ is given. We consider now γ_i fixed, so for simplicity of writing we omit it from the list of variables of the functions. First we need to express the log-likelihood function in terms of parameters β_i. Since we use Poisson model for x_i having the Poisson distribution or overdispersed Poisson, we omit ϕ_i from the list of parameters which condition function p. For a given set of parameters β_i, the probability of attaining $x_i^{d+1}, \dots, x^n$ is given by $p(x_i^{d+1}, \dots, x_i^n | X_i, \beta_i)$ $= \prod_{t=d+1}^n \frac{(\lambda_i^t)^{x_i^t} \exp(-\lambda_i^t)}{(x_i^t)!} = \prod_{t=d+1}^n \frac{\exp([X_i \beta_i']^t)^{x_i^t} \exp(-\exp([X_i \beta_i']^t))}{x_i^t!}$ where $[X_i \beta_i']^t$ denotes the t-th coordinate of the vector $X_i \beta_i'$. The log-likelihood in terms of β_i is $L_i = ll(\beta_i | x_i, X_i) = \log p(\beta_i | x_i, X_i) = \sum_{t=d+1}^n x_i^t [X_i \beta_i']^t - \exp([X_i \beta_i']^t) - \log(x_i^t!)$.

Having function L_i, we can now compute an initial estimate of $\hat{\beta}_i$ from (7) which is the solution to the system of score equations. Since $-ll(\beta_i|x_i, X_i)$ is a convex function, one can use standard convex optimization techniques (e.g. Newton-Raphson method) to solve these equations numerically. (In our code, we use the Matlab implementation of an iteratively reweighted least squares (IRLS) algorithm of the Newton-Raphson method). Assume now we have an initial solution $\hat{\beta}_i$ from (7).

1. Now we derive matrix W_i for x_i with Poisson distribution:
The Fisher information matrix $J_i = J(\beta_i) = -\mathbb{E}_{\beta_i}(\nabla^2 ll(\beta_i|x_i, X_i))$ may be obtained by computing the second order partial derivatives of ll for $r, s = 1, \dots, k_i$. This gives
$\frac{\delta^2 ll(\beta_i|x_i,X_i)}{\delta^2 \beta_i^r \beta_i^s} = \frac{\delta ll}{\delta \beta_i^s} \sum_{t=d+1}^{n} [x_i^t \sum_{l=1}^{d} x_r^{t-l} - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{t-l} \beta_j^l) \sum_{l=1}^{d} x_r^{t-l}] =$
$= -\sum_{t=d+1}^{n} \exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{t-l} \beta_j^l)(\sum_{l=1}^{d} x_s^{t-l})(\sum_{l=1}^{d} x_r^{t-l})$. If we denote
$W_i := diag(\exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{d+1-l} \beta_j^l), \dots, \exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{n-l} \beta_j^l))$ then we have Fisher information matrix $J(\beta_i) = (X_i)' W_i X_i$.

2. Derivation of matrix W_i for x_i with overdispersed Poisson distribution:
Assume now the dispersion parameter $\phi_i > 0, \neq 1$. The variance of the overdispersed Poisson distribution is $\phi_i \lambda_i$. We know that the Poisson regression model can be still used in overdispersed settings and the function ll is the same as $ll(\beta_i)$ derived above. We use the robust sandwich estimate of covariance of $\hat{\beta}_i$, proposed in [9] for a general Poisson regression. The Fisher information matrix of overdispersed problem is $J_i = J(\beta_i) = (X_i)' W_i X_i$ where W_i is constructed for PGGM based on [9] and has the form $W_i =$
$diag([x_i^{d+1} - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{d+1-l} \beta_j^l)]^2, \dots, [x_i^n - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^{d} x_j^{n-l} \beta_j^l)]^2)$.
Having parameters $\hat{\beta}_i$, $\hat{\phi}_i$, Σ_i W_i and ρ_i, we still need to construct the function $MML(\gamma_i)$.

Construction of function $MML(\gamma_i)$:
Having these parameters, we use for each $i = 1, \dots, p$ and regression (7) formula (18) from [21] i.e. for the case when in $\alpha := 0$ and $\beta := \beta_i$ and $X := X_i$, $y := x_i$, $n := n - d$, $k := k_i$, $\theta := \hat{\beta}_i$, $\lambda := \hat{\rho}_i$, $\phi := \phi_i$, $S = \Sigma_i$ is the unity matrix of dimension dk_i, the corrected Fisher information matrix for the parameters β_i is then $J(\beta_i|\phi_i, \rho_i) = (\frac{1}{\phi_i}) X_i' W_i X_i + \rho_i \Sigma_i$ where $\lambda_i = \exp(X_i \beta_i)$. Function $c(m)$ for $m := k_i + 1$ is then $c(k_i + 1) = -\frac{k_i+1}{2} \log(2\pi) + \frac{1}{2} \log((k_i + 1)\pi) - 0.5772$ and the constants independent of k_i we omitted from MML code, since the optimization over γ_i is of them independent. Among all subsets $\gamma_i \in \mathit{\Gamma}$, there are $\binom{p}{k_i}$ subsets of size k_i. If nothing is known a priori about the likelihood of any covariate x_i being included in the final model, a prior that treats all subset sizes equally likely $\pi(|\gamma_i|) = 1/(p + 1)$ is appropriate [21]. This gives the code length $I(\gamma_i) = \log \binom{p}{k_i} + \log(p + 1)$ as in (11).