entropy

MDPI

Article

Heterogeneous 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, 1090 Wien, Austria; claudia.plant@univie.ac.at
2 Institute of Computer Science of the Czech Academy of Sciences, 18207 Prague, Czech Republic
3 ds:UniVie, University of Vienna, 1090 Wien, Austria
* Correspondence: katerina.schindlerova@univie.ac.at

Received: 2 November 2020; Accepted: 7 December 2020; Published: 11 December 2020

Abstract: The heterogeneous graphical Granger model (HGGM) for causal inference among processes with distributions from an exponential family is efficient in scenarios when the number of time observations is much greater than the number of time series, normally by several orders of magnitude. However, in the case of "short" time series, the inference in HGGM often suffers from overestimation. To remedy this, we use the minimum message length principle (MML) to determinate the causal connections in the HGGM. The minimum message length as a Bayesian information-theoretic method for statistical model selection applies Occam's razor in the following way: even when models are equal in their measure of fit-accuracy to the observed data, the one generating the most concise explanation of data is more likely to be correct. Based on the dispersion coefficient of the target 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 and derive its form for various exponential distributions. We propose two algorithms—the genetic-type algorithm (HMMLGA) and exHMML to find the subset. We demonstrated the superiority of both algorithms in synthetic experiments with respect to the comparison methods Lingam, HGGM and statistical framework Granger causality (SFGC). In the real data experiments, we used the methods to discriminate between pregnancy and labor phase using electrohysterogram data of Islandic mothers from Physionet databasis. We further analysed the Austrian climatological time measurements and their temporal interactions in rain and sunny days scenarios. In both experiments, the results of HMMLGA had the most realistic interpretation with respect to the comparison methods. We provide our code in Matlab. To our best knowledge, this is the first work using the MML principle for causal inference in HGGM.

Keywords: Granger causality; graphical Granger model; overestimation; information theory; minimum message length

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 distributions can be, however, misleading. Recently, Behzadi et al. in [1] proposed the heterogeneous graphical Granger Model (HGGM) for detecting causal relations among time series with distributions from the exponential family, which includes a wider class of common distributions. HGGM employs regression in generalized linear models (GLM) with adaptive Lasso penalization [2] as a variable selection method and applies it to time series with a given lag. This approach allows one to apply causal inference among time series, also with discrete values. HGGM, using penalization by adaptive Lasso, showed its efficiency in scenarios when the number of time observations is much greater than the number of time series,

normally by several orders of magnitude—however, on "short" time series, the inference in HGGM suffers often from overestimation.

Overestimation on short time series is a problem which also occurs in general forecasting problems. For example, when forecasting demand for a new product or a new customer, there are usually very few time series observations available. For such short time series, the traditional forecasting methods may be inaccurate. To overcome this problem in forecasting, Ref. [3] proposed to utilize a prior information derived from the data and applied a Bayesian inference approach. Similarly for another data mining problem, a Bayesian approach has shown to be efficient for the clustering of short time series [4].

Motivated by the efficiency of the Bayes approaches in these problems on short time series, we propose to use the Bayesian approach called minimum message principle, as introduced in [5] to causal inference in HGGM. The contributions of our paper are the following:

(1) We used the minimum message length (MML) principle for determination of causal connections in the heterogeneous graphical Granger model.
(2) Based on the dispersion coefficient of the target 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 causally connected time series with each target time series; Furthermore, we derived its form for various exponential distributions.
(3) We found this subset in two ways: by a proposed genetic-type algorithm (HMMLGA), as well as by exhaustive search (exHMML). We evaluated the complexities of these algorithms and provided the code in Matlab.
(4) We demonstrated the superiority of both methods with respect to the comparison methods Lingam [6], HGGM [1] and statistical framework Granger causality (SFGC) [7] in the synthetic experiments with short time series. In the real data experiments without known ground truth, the interpretation of causal connections achieved by HMMLGA was the most realistic with respect to the comparison methods.
(5) To our best knowledge, this is the first work applying the minimum message length principle to the heterogeneous graphical Granger model.

The paper is organized as follows. Section 2 presents definitions of the graphical Granger causal model and of the heterogeneous graphical Granger causal model as well as of the minimum message length principle. Our method including the derived criteria and algorithm are described in Section 3. Related work is discussed in Section 4. Our experiments are summarized in Section 5. Section 6 is devoted to the conclusions and the derivation of the criteria from Section 3 can be found in Appendices A and B.

2. Preliminaries

To make this paper self-contained and to introduce the notation, we briefly summarize the basics about graphical Granger causal model in Section 2.1. The heterogeneous graphical Granger model, as introduced in [1], is presented in Section 2.2. Section 2.3 discusses the strengths and limitations of the Granger causal models. The idea of the minimum message length principle is briefly explained in Section 2.4.

2.1. Graphical Granger Model

The (Gaussian) graphical Granger model extends the autoregressive concept of Granger causality to $p \geq 2$ time series [8]. Let $x_1^t, \ldots, x_p^t$ be the time instances of p time series, $t = 1, \ldots, n$. As it is common, we will use bold font in notation of vectors or matrices. Consider the vector auto-regressive (VAR) models with time lag $d \geq 1$ for $i = 1, \ldots, p$

$$x_i^t = \mathbf{X}_{t,d}^{Lag} \boldsymbol{\beta}_i' + \varepsilon_i^t \tag{1}$$

where $\mathbf{X}_{t,d}^{Lag} = (x_1^{t-d}, \ldots, x_1^{t-1}, \ldots, x_p^{t-d}, \ldots, x_p^{t-1})$ and $\boldsymbol{\beta}_i$ be a matrix of the regression coefficients and ε_i^t be white noise. One can easily show that $\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i' = \sum_{j=1}^{p}\sum_{l=1}^{d} x_j^{t-l}\beta_j^l$.

Definition 1. *One says time series $\mathbf{x}_j$ Granger–causes time series $\mathbf{x}_i$ for a given lag d, denote $\mathbf{x}_j \to \mathbf{x}_i$, for $i, j = 1, \ldots, p$ if and only if at least one of the d coefficients in j-th row of $\boldsymbol{\beta}_i$ in (1) is non-zero.*

The solution of problem (1) has been approached by various forms of penalization methods in the literature, e.g., Lasso in [8], truncated Lasso in [9] or group Lasso [10].

2.2. Heterogeneous Graphical Granger Model

The heterogeneous graphical Granger model (HGGM) [1] considers time series $\mathbf{x}_i$, for which their likelihood function belongs into the exponential family with a canonical parameter θ_i. The generic density form for each $\mathbf{x}_i$ can be written as

$$p(\mathbf{x}_i|\mathbf{X}_{t,d}^{Lag}, \theta_i) = h(\mathbf{x}_i)\exp(\mathbf{x}_i\theta_i - \boldsymbol{\eta}_i(\theta_i)) \tag{2}$$

where $\theta_i = \mathbf{X}_{t,d}^{Lag}(\boldsymbol{\beta}_i^*)'$ ($\boldsymbol{\beta}_i^*$ is the optimum) and $\boldsymbol{\eta}_i$ is a link function corresponding to time series $\mathbf{x}_i$. (The sign $'$ denotes a transpose of a matrix). The heterogeneous graphical Granger model uses the idea of generalized linear models (GLM, see e.g., [11]) and applies them to time series in the following form

$$x_i^t \approx \mu_i^t = \eta_i^t(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i') = \eta_i^t(\sum_{j=1}^{p}\sum_{l=1}^{d} x_j^{t-l}\beta_j^l) \tag{3}$$

for x_i^t, $i = 1, \ldots, p, t = d+1, \ldots, n$ each having a probability density from the exponential family; $\boldsymbol{\mu}_i$ denotes the mean of $\mathbf{x}_i$ and $var(\mathbf{x}_i|\boldsymbol{\mu}_i, \phi_i) = \phi_i v_i(\boldsymbol{\mu}_i)$ where ϕ_i is a dispersion parameter and v_i is a variance function dependent only on $\boldsymbol{\mu}_i$; η_i^t is the t-th coordinate of $\boldsymbol{\eta}_i$.
Causal inference in (3) can be solved as

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \sum_{t=d+1}^{n} (-x_i^t(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i') + \eta_i^t(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i')) + \lambda_i R(\boldsymbol{\beta}_i) \tag{4}$$

for a given lag $d > 0$, $\lambda_i > 0$, and all $t = d+1, \ldots, n$ with $R(\boldsymbol{\beta}_i)$ to be the adaptive Lasso penalty function [1]. (The first two summands in (4) correspond to the maximum likelihood estimates in the GLM).

Definition 2. *One says, time series $\mathbf{x}_j$ Granger–causes time series $\mathbf{x}_i$ for a given lag d, denote $\mathbf{x}_j \to \mathbf{x}_i$, for $i, j = 1, \ldots, p$ if and only if at least one of the d coefficients in j-th row of $\hat{\boldsymbol{\beta}}_i$ of the solution of (4) is non-zero* [1].

Remark 1. *Non-zero values in Definitions 1 and 2 are in practice, distinguished by considering values bigger than a given threshold, which is a positive number "close" to zero.*

For example, Equation (4) for the Poisson graphical Granger model [12] where for each $i = 1, \ldots, p$ $\eta_i^t := \exp$ is considered, can be written as

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \sum_{t=d+1}^{n} (-x_i^t(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i') + \exp(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i')) + \lambda_i R(\boldsymbol{\beta}_i). \tag{5}$$

Equation (4) for the binomial graphical Granger model can be written as

$$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}_i} \sum_{t=d+1}^{n} (-x_i^t(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i') + \log(1+\exp(\mathbf{X}_{t,d}^{Lag}\boldsymbol{\beta}_i'))) + \lambda_i R(\boldsymbol{\beta}_i) \quad (6)$$

and finally Equation (4) for the Gaussian graphical Granger model reduces to the least squares error of (1) with a $R(\boldsymbol{\beta}_i)$ to be adaptive Lasso penalty function. The heterogeneous graphical Granger model can be applied to causal inference among processes, for example in climatology, e.g., Ref. [1] investigated the causal inference among precipitation time series (having gamma distribution) and time series of sunny days (having Poisson distribution).

2.3. Granger Causality and Graphical Granger Models

Since its introduction, Granger causality [13] has faced criticism, since it e.g., does not take into account counterfactuals, [14,15]. In defense of his method, Granger in [16] 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.

Concerning the graphical causal models, including the Granger ones, Lindquist and Sobel in [17] claim that (1) they are not able to discover causal effects; (2) the theory of graphical causal models developed by Spirtes et al. in [18] makes no counterfactual claims; and (3) causal relations cannot be determined non-experimentally from samples that are a combination of systems with different propensities. However, Glymour in [19] argues that each of these claims are false or exaggerated. For arguments against (1) and (3), we refer the reader to [19]. We focus here only to his arguments to (2). Quoting Glymour, claims about what the outcome would be of a hypothetical experiment that has not been done are one form of counterfactual claims. Such claims say that if such and such were to happen then the result would be thus and so—where such and such has not happened or has not yet happened. (Of course, if the experiment is later done, then the proposition becomes factually true or factually false.) Glymour argues that it is not true that the graphical model framework does not represent or entail any counterfactual claims and emphasizes that no counterfactual variables are used or needed in the graphical causal model framework. In the potential outcomes framework, if nothing is known about which of many variables are causes of the others, then for each variable, and for each value of the other variables, a new counterfactual variable is required. In practice that would require an astronomical number of counterfactual variables for even a few actual variables. To summarize, as also confirmed by a recent *Nature* publication [20], if the theoretical background of investigated processes is insufficient, graphical causal methods (Granger causality including), to infer causal relations from data rather than knowledge of mechanisms, are helpful.

2.4. Minimum Message Length Principle

The minimum message length principle of statistical and inductive inference and machine learning was developed by C.S. Wallace and D.M. Boulton in 1968 in the seminal paper [5]. Minimum message length principle is a formal information theory restatement of Occam's razor: even when models are not equal in goodness of fit accuracy to the observed data, the one generating the shortest overall message is more likely to be correct (where the message consists of a statement of the model, followed by a statement of data encoded concisely using that model). 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 [21]. MML is Bayesian (i.e., it incorporates prior beliefs)

and information-theoretic. It has the desirable properties of statistical invariance (i.e., the inference transforms with a re-parametrisation), statistical consistency (i.e., even for very hard problems, MML will converge to any underlying model) and efficiency (i.e., the MML model will converge to any true underlying model about as quickly as is possible). Wallace and Dowe (1999) showed in [22] a formal connection between MML and Kolmogorov complexity, i.e., the length of a shortest computer program that produces the object as output.

3. Method

In this section, we will describe our method in detail. First, in Section 3.1, we will derive a fixed design matrix for HGGM, so that the minimum message length principle can be applied. In Section 3.2, we propose our minimum message length criterion for HGGM. The exact forms of the criterion for various exponential distributions are derived in Section 3.3. Then, we present our two variable selection algorithms and their computational complexity in Sections 3.4 and 3.5.

3.1. Heterogeneous Graphical Granger Model with Fixed Design Matrix

We can see that the models from Section 2 do not have fixed matrices. Since the MML principle proposed for generalized linear models in [23] requires a fixed design matrix, it cannot be directly applied to them. In the following section, we will derive the heterogeneous graphical Granger model (3) with a fixed lag d as an instance of regression in generalized linear models (GLM) with a fixed design matrix.

Consider the full model for p variables x_i^t and lag $d \geq 1$ (be an integer) corresponding to the optimization problem (3). To be able to use the maximum likelihood (ML) estimation over the regression parameters, we reformulate the matrix of lagged time series $\mathbf{X}_{t,d}^{Lag}$ from (1) into a fixed design matrix form. Assume $n - d > pd$ and denote $\mathbf{x}_i = (x_i^{d+1}, x_i^{d+2}, \ldots, x_i^n)$. We construct the $(n-d) \times (d \times p)$ design matrix

$$\mathbf{X} = \begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_p^d & \ldots & x_p^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_p^{d+1} & \ldots & x_p^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_p^{n-1} & \ldots & x_p^{n-d+1} \end{bmatrix} \tag{7}$$

and the $1 \times (d \times p)$ vector $\boldsymbol{\beta}_i = (\beta_1^1, \ldots, \beta_1^d, \ldots, \beta_p^1, \ldots, \beta_p^d)$. We can see that problem

$$\mathbf{x}_i' \approx \boldsymbol{\mu}_i = \boldsymbol{\eta}_i(\mathbf{X}\boldsymbol{\beta}_i') \tag{8}$$

for $i = d+1, \ldots, n$ is equivalent to problem (3) in the matrix form where $\boldsymbol{\mu}_i = (\mu_i^{d+1}, \ldots, \mu_i^{d+1})$ and link function $\boldsymbol{\eta}_i$ operates on each coordinate.

Denote now by $\gamma_i \subset \Gamma = \{1, \ldots, p\}$ the subset of indices of regressor's variables and $k_i := |\gamma_i|$ its cardinality. Let $\boldsymbol{\beta}_i := \boldsymbol{\beta}_i(\gamma_i) \in \mathbb{R}^{1 \times (d \times k_i)}$ 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 to γ_i. Considering only the columns from matrix $\mathbf{X}$ in (7), which correspond to γ_i, we define the $(n-d) \times (d \times k_i)$ matrix of lagged vectors with indices from γ_i as

$$\mathbf{X}_i := \mathbf{X}(\gamma_i) = \begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_{k_i}^d & x_{k_i}^{d-1} & \ldots & x_{k_i}^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_{k_i}^{d+1} & x_{k_i}^d & \ldots & x_{k_i}^2 \\ x_1^{d+2} & \ldots & x_1^3 & \ldots & x_{k_i}^{d+2} & x_{k_i}^{d+1} & \ldots & x_{k_i}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_{k_i}^{n-1} & x_{k_i}^{n-2} & \ldots & x_{k_i}^{n-d+1} \end{bmatrix} \tag{9}$$

The problem (8) for explanatory variables with indices from γ_i is expressed as

$$\mathbf{x}_i' \approx \boldsymbol{\mu}_i = E(\mathbf{x}_i'|\mathbf{X}_i) = \boldsymbol{\eta}_i(\mathbf{X}_i\boldsymbol{\beta}_i'). \tag{10}$$

with $\boldsymbol{\beta}_i := \boldsymbol{\beta}_i(\gamma_i)$ to be a $1 \times (dk_i)$ matrix of unknown coefficients and $\boldsymbol{\eta}_i$ operates on each coordinate. Wherever it is clear from context, we will simplify the notation $\boldsymbol{\beta}_i$ instead of $\boldsymbol{\beta}_i(\gamma_i)$ and $\mathbf{X}_i$ instead of $\mathbf{X}(\gamma_i)$.

3.2. Minimum Message Length Criterion for Heterogeneous Graphical Granger Model

As before, we assume for each x_i^t, where $i = 1, \ldots, p$, $t = d+1, \ldots, n$ to have a density from the exponential family; furthermore, $\boldsymbol{\mu}_i$ to be the mean of $\mathbf{x}_i$ and $var(\mathbf{x}_i|\boldsymbol{\mu}_i, \phi_i) = \phi_i v_i(\boldsymbol{\mu}_i)$ where ϕ_i is a dispersion parameter and v_i a variance function dependent only on $\boldsymbol{\mu}_i$. In the concrete case, for Poisson regression, it is well known that it can be still used in over- or underdispersed settings. However, the standard error for Poisson regression 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 $\phi_i \neq 1$. In the following, we assume that an estimate of ϕ_i is given. Denote Γ the set of all subsets of covariates $\mathbf{x}_i, i = 1, \ldots, p$. Assume now a fixed set $\gamma_i \in \Gamma$ of covariates with size $k_i \leq p$ and the corresponding design matrix $\mathbf{X}_i$ from (9). Furthermore, we assume that the targets $\mathbf{x}_i$ are independent random variables, conditioned on the features given by $\mathbf{X}_i$, so that the likelihood function can be factorized into the product $p(\mathbf{x}_i|\boldsymbol{\beta}_i, \mathbf{X}_i, \gamma_i) = \prod_{t=1}^{n-d} p(x_i^t|\boldsymbol{\beta}_i, \mathbf{X}_i, \gamma_i)$. The log-likelihood function L_i has then the form $L_i := \log p(\mathbf{x}_i|\boldsymbol{\beta}_i, \mathbf{X}_i, \gamma_i) = \sum_{t=1}^{n-d} \log p(x_i^t|\boldsymbol{\beta}_i, \mathbf{X}_i, \gamma_i)$. Since $\mathbf{X}_i$ is highly collinear, to make the ill-posed problem for coefficients $\boldsymbol{\beta}_i$ (8) a well-posed one, we could use regularization by the ridge regression for GLM (see e.g., [24]). Ridge regression requires an initial estimate of $\boldsymbol{\beta}_i$, which can be set as the maximum likelihood estimator of (10) obtained by the iteratively reweighted least square algorithm (IRLS). For a fixed $\lambda_i > 0$, for the ridge estimates of coefficients $\hat{\boldsymbol{\beta}}_{i,\lambda_i}$ holds

$$\hat{\boldsymbol{\beta}}_{i,\lambda_i} = \arg\min_{\boldsymbol{\beta}_i \in \mathbb{R}^{1 \times dk_i}} \{-L_i + \lambda_i \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_i \boldsymbol{\beta}_i\}. \tag{11}$$

In our paper however, we will not use the GLM ridge regression in form (11), but we 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 (= the ridge regularization parameter). To compute the message length of HGGM using the MML87 approximation, we need the negative log-likelihood function, prior distribution over the parameters and an appropriate Fisher information matrix, similarly as proposed in [23], where it is done for a general GLM regression. Moreover, [23] 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 heterogeneous graphical Granger model. In the following, we will construct the MML code for every subset of covariates in HGGM. The derivation of the criterion can be found in Appendix A.

The MML criterion for inference in HGGM. *Assume $\mathbf{x}_i, i = 1, \ldots, p$ be given time series of length n having distributions from exponential family, and for each of them, the estimate of the dispersion parameter $\hat{\phi}_i$ is given. Consider $\hat{\boldsymbol{\beta}}_i$ be an initial solution of (8) with a fixed $d \geq 1$ achieved as the maximum likelihood estimate. Then*

(i) the causal graph of the heterogeneous graphical Granger problem (8) can be inferred from the solutions of p variable selection problems, where for each $i = 1, \ldots, p$, the set $\hat{\gamma}_i$ of Granger–causal variables to $\mathbf{x}_i$ is found;

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

$$\hat{\gamma}_i = \arg\min_{\gamma_i \in \Gamma} \{HMML(\mathbf{x}_i, \mathbf{X}_i, \gamma_i)\} = \arg\min_{\gamma_i \in \Gamma} \{I(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \hat{\lambda}_i, \mathbf{X}_i, \gamma_i) + I(\gamma_i)\} \tag{12}$$

where $I(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \hat{\lambda}_i, \mathbf{X}_i, \gamma_i) = \min_{\lambda_i \in \mathbb{R}^+}\{MML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \lambda_i, \mathbf{X}_i, \gamma_i)\}$ *and* $MML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \lambda_i, \mathbf{X}_i, \gamma_i)$ *is the minimum message length code of set* γ_i*. It can be expressed as*

$$MML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \lambda_i, \mathbf{X}_i, \gamma_i) = -L_i + \frac{1}{2}\log\det(\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i + \lambda_i\boldsymbol{\Sigma}_i) \tag{13}$$

$+\frac{k_i}{2}\log(\frac{2\pi}{\lambda_i}) + (\frac{\lambda_i}{2\hat{\phi}_i})\hat{\boldsymbol{\beta}}_i'\boldsymbol{\Sigma}_i\hat{\boldsymbol{\beta}}_i + \frac{1}{2}\log(n-d) - \frac{k_i+1}{2}\log(2\pi) + \frac{1}{2}\log((k_i+1)\pi)$ *where* $|\hat{\gamma}_i| = k_i$*,* $\boldsymbol{\Sigma}_i$ *is the unity matrix of size* $dk_i \times dk_i$*,* $I(\gamma_i) = \log\binom{p}{k_i} + \log(p+1)$*,* L_i *is the log-likelihood function depending on the density function of* $\mathbf{x}_i$ *and matrix* $\mathbf{W}_i$ *is a diagonal matrix depending on link function* $\boldsymbol{\eta}_i$*.*

Remark 2. ([23]) *compared* AIC_c *criterion with MML code for generalized linear models. We constructed the* AIC_c *criterion also for HGGM. This criterion however requires the computation of pseudoinverse of a matrix multiplication, which includes matrices* $\mathbf{X}_i$*. Since* $\mathbf{X}_i$*s are highly collinear, these matrix multiplications had, in our experiments, very high condition numbers. This consequently led to the application of* AIC_c *for HGGM, giving spurious results, and therefore we do not report them in our paper.*

3.3. Log-Likelihood L_i*, Matrix* $\mathbf{W}_i$ *and Dispersion* ϕ_i *for* $\mathbf{x}_i$ *with Various Exponential Distributions*

In this section, we will present the form for the log-likelihood function and for matrix $\mathbf{W}_i$ for Gaussian, binomial, Poisson, gamma and inverse-Gaussian distributed time series $\mathbf{x}_i$. The derivation for each case can be found in Appendix B. $\boldsymbol{\mu}_i = \boldsymbol{\eta}_i(\mathbf{X}_i\boldsymbol{\beta}_i')$ holds in each case for the link function as in (10). By $[\mathbf{X}_i\boldsymbol{\beta}_i']^t$, we denote the t-th coordinate of vector $\mathbf{X}_i\boldsymbol{\beta}_i'$.

Case $\mathbf{x}_i$ is Gaussian This is the case when $\mathbf{x}_i$ is an independent Gaussian random variable and link function $\boldsymbol{\eta}_i$ is identity. Assume $\hat{\phi}_i = \sigma_i^2$ to be the variance of the Gaussian random variable. We assume that in model (10) $\mathbf{x}_i$ follows Gaussian distribution with the density function $p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \sigma_i^2, \mathbf{X}_i, \gamma_i) =$

$$\prod_{t=d+1}^{n} p(x_i^t|\hat{\boldsymbol{\beta}}_i, \sigma_i^2, \mathbf{X}_i, \gamma_i) = (\frac{1}{(2\pi\sigma_i^2)})^{(n-d)/2}\exp\,[-\frac{1}{2\sigma_i^2}\sum_{t=d+1}^{n}(x_i^t - [\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^t)^2]. \tag{14}$$

Then

$$L_i = \log p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \sigma_i^2, \mathbf{X}_i, \gamma_i) = -\frac{n-d}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}\sum_{t=d+1}^{n}(x_i^t - [\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^t)^2 \tag{15}$$

and $\mathbf{W}_i := \mathbf{I}_{n-d,n-d}$ is a unit matrix of dimension $(n-d)\times(n-d)$.

Case $\mathbf{x}_i$ is binomial This is the case when $\mathbf{x}_i$ is an independent Bernoulli random variable and it can achieve only two different values. For the link function, it holds $\boldsymbol{\eta}_i = \log(\frac{\boldsymbol{\mu}_i}{1-\boldsymbol{\mu}_j})$. Without lack of generality, we consider $\hat{\phi}_i = 1$ and the density function $p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \sigma_i^2, \mathbf{X}_i, \gamma_i) =$

$$\prod_{t=d+1}^{n} p(x_i^t|\hat{\boldsymbol{\beta}}_i, \sigma_i^2, \mathbf{X}_i, \gamma_i) = \prod_{t=d+1}^{n}([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t)^{x_i^t}(1-([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t))^{(1-x_i^t)}. \tag{16}$$

Then

$$L_i = \log(p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \mathbf{X}_i, \gamma_i)) = \sum_{t=d+1}^{n}\left(x_i^t[\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t - \log(1+\exp[\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t)\right) \tag{17}$$

and

$$\mathbf{W}_i := \mathrm{diag}(\frac{\exp\,([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^1)}{(1+\exp\,([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^1))^2}, \ldots, \frac{\exp\,([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^{n-d})}{(1+\exp\,([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i]^{n-d})^2)}). \tag{18}$$

In the case that we cannot assume accurate fitting to one of the two values, for robust estimation we can consider the sandwich estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_i$ with

$$\mathbf{W}_i = \mathrm{diag}\big([x_i^1 - \frac{\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1)}{(1+\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1))^2}]^2, \ldots, [x_i^{n-d} - \frac{\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d})}{(1+\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d}))^2}]^2\big). \tag{19}$$

Case $\mathbf{x}_i$ is Poisson If $\mathbf{x}_i$ is an independent Poisson random variable with link function $\eta_i^t = \log(\mu_i^t) = \log([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t)$, the density is

$$p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=d+1}^{n} \frac{\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t)^{x_i^t} \exp(-\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t))}{x_i^t!}. \tag{20}$$

Then

$$L_i = \log(p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \mathbf{X}_i, \gamma_i)) = \sum_{t=d+1}^{n} x_i^t[\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t - \exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t) - \log(x_i^t!) \tag{21}$$

and diagonal matrix

$$\mathbf{W}_i := \mathrm{diag}\big(\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i')^1, \ldots, \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i')^{n-d}\big) \tag{22}$$

for Poisson $\mathbf{x}_i$ with $\hat{\phi}_i = 1$ and

$$\mathbf{W}_i := \mathrm{diag}\big([x_i^{d+1} - \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i')^1]^2, \ldots, [x_i^{d+(n-d)} - \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_i')^{n-d}]^2\big) \tag{23}$$

for over- or underdispersed Poisson $\mathbf{x}_i$, i.e., when $\hat{\phi}_i \neq 1$ and is positive, where $t = 1, \ldots, n-d$.

Case $\mathbf{x}_i$ is gamma If $\mathbf{x}_i$ is an independent gamma random variable, we consider for the inverse of shape parameter κ_i for each t rate parameter $\kappa_i\mu_i^t$ and for the link function it holds $\mu_i^t = \frac{1}{\eta_i^t} = \frac{1}{[\mathbf{X}_i\boldsymbol{\beta}_i]^t}$. For parameters of gamma function a_i, b_i we take $a_i = \frac{1}{\kappa_i}, b_i^t = \kappa_i\hat{\mu}_i^t$ and assume for dispersion $\hat{\phi}_i = \kappa_i$. Then, we have density function

$$p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \frac{1}{\kappa_i}, \kappa_i\hat{\boldsymbol{\mu}}_i, \mathbf{X}_i, \gamma_i) = \prod_{t=d+1}^{n} \frac{(x_i^t)^{(\frac{1}{\kappa_i}-1)} \exp(-\frac{x_i^t}{\kappa_i\mu_i^t})}{(\kappa_i\mu_i^t)^{\frac{1}{\kappa_i}}\Gamma(\frac{1}{\kappa_i})} \tag{24}$$

and log-likelihood $L_i = \log(p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \frac{1}{\kappa_i}, \kappa_i\hat{\boldsymbol{\mu}}_i, \mathbf{X}_i, \gamma_i))$

$$= \sum_{t=d+1}^{n} ((\frac{1}{\kappa_i} - 1)\log x_i^t - \frac{x_i^t}{\kappa_i\hat{\mu}_i^t} - \frac{1}{\kappa_i}\log(\kappa_i\hat{\mu}_i^t) - \log\Gamma(\frac{1}{\kappa_i})) \tag{25}$$

and diagonal matrix

$$\mathbf{W}_i := \mathrm{diag}((\hat{\mu}_i^1)^2, \ldots, (\hat{\mu}_i^{n-d})^2) = \mathrm{diag}(\frac{1}{([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1)^2}, \ldots, \frac{1}{([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d})^2}). \tag{26}$$

Case $\mathbf{x}_i$ is inverse-Gaussian If $\mathbf{x}_i$ is an independent inverse-Gaussian random variable, we consider the inverse of the shape parameter ξ_i and link function $\eta_i^t = \log(\mu_i^t) = \log([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^t)$. Assume dispersion $\hat{\phi}_i = \xi_i$. Then we have density function

$$p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \xi_i, \hat{\boldsymbol{\mu}}_i, \mathbf{X}_i, \gamma_i) = \prod_{t=d+1}^{n} \frac{1}{2\pi\xi_i(x_i^t)^3} \exp[-\frac{1}{2\xi_i}\sum_{t=d+1}^{n} \frac{(x_i^t - \hat{\mu}_i^t)^2}{(\hat{\mu}_i^t)^2 x_i^t}] \tag{27}$$

and log-likelihood $L_i = \log(p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \xi_i\hat{\boldsymbol{\mu}}_i, \mathbf{X}_i, \gamma_i))$

$$= \sum_{t=d+1}^{n} \Big(-\frac{1}{2\xi_i} \sum_{t=d+1}^{n} \frac{(x_i^t - \hat{\mu}_i^t)^2}{(\hat{\mu}_i^t)^2 x_i^t} - \log(2\pi\xi_i) + 3\log(x_i^t) \Big) \tag{28}$$

and diagonal matrix

$$\mathbf{W}_i := \mathrm{diag}\Big(\frac{1}{\hat{\mu}_i^1}, \ldots, \frac{1}{\mu_i^{n-d}}\Big) = \mathrm{diag}\Big(\frac{1}{([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1)}, \ldots, \frac{1}{([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d})}\Big). \tag{29}$$

One could express similarly L_i and $\mathbf{W}_i$ for other common exponential distributions, applied in GLMs.

3.4. Variable Selection by MML in Heterogeneous Graphical Granger Model

For all considered cases of exponential distributions of $\mathbf{x}_i$ we define the family of models $M(\gamma_i) := \{p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \mathbf{X}_i, \gamma_i), \gamma_i \in \Gamma\}$ with the corresponding exponential density $p(\mathbf{x}_i|\hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, X_i, \gamma_i)$. First, we present the procedure which for each $\mathbf{x}_i$ computes the MML code for a set $\gamma_i \subset \Gamma$ in Algorithm 1. Then we present Algorithm 2 for computation of $\hat{\gamma}_i$.

Algorithm 1 MML Code for γ_i

Input: $\gamma_i \in \Gamma, d \geq 1, |\gamma_i| = k_i$, series is the matrix of x_i^t, $\hat{\phi}_i$ dispersion parameter, $i = 1, \ldots, p, t = 1, \ldots, n-d$, $\boldsymbol{\Sigma}_i$ a unity matrix of size $dk_i \times dk_i$, H a set of positive numbers; $I(\gamma_i) = \log\binom{p}{k_i} + \log(p+1)$.
Output: For each i minimum of $HMML(\mathbf{x}_i, \mathbf{X}_i, \gamma_i)$ over H is found;
for all $\mathbf{x}_i$ **do**
  // Construct the d-lagged matrix $\mathbf{X}_i$ with time series with indices from γ_i.
  //Compute matrix $\mathbf{W}_i$.
  for all $\lambda_i \in H$ **do**
    // Compute L_i
    // Find the initial estimates of $\hat{\boldsymbol{\beta}}_i$.
    //Compute $MML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \lambda_i, \mathbf{X}_i, \gamma_i)$ from (13).
  end for// to λ_i
  // Compute $I(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \hat{\lambda}_i, \mathbf{X}_i, \gamma_i) = \min_{\lambda_i \in H} MML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \lambda_i, \mathbf{X}_i, \gamma_i)$.
  // $HMML(\mathbf{x}_i, \mathbf{X}_i, \gamma_i) := I(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \hat{\lambda}_i, \mathbf{X}_i, \gamma_i) + I(\gamma_i)$.
end for// to $\mathbf{x}_i$
return $HMML(\mathbf{x}_i, \mathbf{X}_i, \gamma_i)$ for each i.

In general, the selection of the best structure γ_i amounts to evaluate values of $HMML(\gamma_i)$ 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.

3.5. Search Algorithms

We will find the best structure of γ_i with MML code by two approaches. The first way is by the exhaustive search approach exHMML and the second one is by minimizing the HMML by genetic algorithm type procedure called HMMLGA, which we introduce in the following. Since HMML in (12) is a function having multiple local minima, the achievement of the global minimum by these two approaches is not, in general, guaranteed. In [12], a similar genetic algorithm MMLGA was proposed for the Poisson GGM. In this paper, we propose its modification, which is more appropriate for the objective functions that we have here.

The idea of HMMLGA is as follows. Consider an arbitrary $\gamma_i \subset \Gamma$ with size k_i for a fixed i. Define a Boolean vector Q_i of length p corresponding to a given γ_i, so that it has ones in the positions

of the indices of covariates from γ_i, otherwise zeros. Define $HMML(Q_i) := HMML(\gamma_i)$ where $HMML(\gamma_i)$ is from (12). Genetic algorithm MMLGA executes genetic operations on populations of Q_i. In the first step, a population of size m (m 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 current population with the lowest value of (12) 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. The position of mutation is for each individual selected randomly in contrast to MMLGA, where the position was, for all individuals, the same, and is given as an input parameter. Similarly, the position of crossover in HMMLGA is for each pair of individuals selected randomly. After each run of these two operations on a current population, the current population is replaced with the children with the lowest value of (12) to form the next generation. The algorithm stops after the number of population generations n_g is achieved. Since HMML in (12) has multiple local minima, in contract to MMLGA, we selected in the HMMLGA the following strategy: We do not take the first Q_i with the sorted HMML values ascendently, but based on the parsimony principle, we take that Q_i among all with minimum HMML value, which has the minimum number of ones in Q_i. Concerning the approach by exhaustive search exHMML, similarly we do not take the first Q_i with sorted HMML code ascendently, but also, here, we take that Q_i, among all with a minimum value of HMML, which has the minimum number of ones in Q_i. The algorithm HMMLGA is summarized in Algorithm 2.

Algorithm 2 HMMLGA

Input: Γ, $d \geq 1, p, n_g, m$ an even integer;
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 $\mathbf{x}_i$ Q_i with minimum of (12) is found;
for all $\mathbf{x}_i$ **do**
  Create initial population $\{Q_i^j, j = 1, \ldots, m\}$ at random; Compute
  $HMML(Q_i^j) := I(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\phi}_i, \hat{\lambda}_i, \mathbf{X}_i, Q_i^j) + \binom{p}{k_i^j} + \log(p+1)$ 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 $HMML(Q_i^j)$ ascendently and create the elite population; By crossover of Q_i^j and $Q_i^r, r \neq j$ at a random crossing position create children and add them to elite; Compute $HMML(Q_i^j)$ for each j; Mutate a single parent Q_i^j at a random position; Compute $HMML(Q_i^j)$ for each j; Add the children with minimum $HMML(Q_i^j)$ until the new population is 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 (12) such that $|Q_i|$ is minimum.
return (Adj)

Our code in Matlab is publicly available at: https://t1p.de/26f3.

Computational Complexity of HMMLGA and of exHMML

We used Matlab function *fminsearch* for computation of $HMML(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_i, \hat{\lambda}_i, \mathbf{X}_i, \gamma_i)$. 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 HMMLGA for p time series, size p of an individual, m the population size and n_g the number of population generations is $\mathcal{O}(pmn_g) \times O(fminsearch) \times p$, where $O(fminsearch)$ can also be 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 $\mathbf{W}_i$) 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 the complexity of determinant in (13) is $\mathcal{O}((pd)^3)$ (for proof see e.g., [25]). Denote $M = \max\{(pd)^3, \frac{(n-d)(n-d+1)}{2}\}$. Since we have p optimization functions, our upper bound on the computational complexity of HMMLGA is then $\mathcal{O}(p^2mn_gM)$. The computational complexity of $exHMML$ is $p2^pO(fminsearch) = p2^pM$.

4. Related Work

In this section, we discuss the related work on the application of two description length based compression schemes for generalized linear models, further the related work on these compression principles applied to causal inference in graphical models, and finally, other papers on causal inference in graphical models for non-Gaussian time series.

Minimum description length (MDL) is another principle based on compression. Similarly as for MML, by viewing statistical modeling as a means of generating descriptions of observed data, the MDL framework (Rissanen [26], Barron et al. [27], and Hansen and Yu [28]) discriminates between competing model classes based on the complexity of each description. The minimum description length principle is based on the idea that one chooses the model that gives the shortest description of data. The methods based on MML and MDL appear mostly equivalent, but there are some differences, especially in interpretation. MML is a Bayesian approach: it assumes that the data-generating process has a given prior distribution. MDL avoids assumptions about the data-generating process. Both methods make use of two-part codes: the first part always represents the information that one is trying to learn, such as the index of a model class (model selection) or parameter values (parameter estimation); the second part is an encoding of the data, given the information in the first part.

Hansen and Yu 2003 in [29] derived objective functions for one-dimensional GLM regression by the minimum description principle. The extension to the multi-dimensional case is however not straighforward. Schmidt and Makalic in [23] used MML87 to derive the MML code of a multivariate GLM ridge regression. Since these works were not designed for time series and do not consider any lag, the mentioned codes cannot be directly used for Granger models.

Marx and Vreeken in [30,31] and Budhathoki and Vreeken [32] applied the MDL principle to the Granger causal inference. The inference in these papers is however done for the bivariate Granger causality and the extension to graphical Granger methods is not straightforward. Hlaváčková-Schindler and Plant in [33] applied both MML and MDL principle to the inference in the graphical Granger models for Gaussian time series. Inference in graphical Granger models for Poisson distributed data using the MML principle was done by the same authors in [12]. To our best knowledge, papers on compression criteria for heterogeneous graphical Granger model have not been published yet.

Among the causal inference on time series, Kim et al. in [7] 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. The method can also be used for time series from exponential family.

For a fair comparison with our method, we selected the causal inference methods, which are designed for $p \geq 3$ non-Gaussian processes. In our experiments, we used SFGC as a comparison method, and as another comparison method, we selected the method LINGAM from Shimizu et al. [6], which estimates causal structure in Bayesian networks among non-Gaussian time series using structural equation models and independent component analysis. Finally, as a comparison method, we used the

HGGM with the adaptive Lasso penalisation method, as introduced in [1] and described in Section 2.2. The experiments reported in the papers with comparison methods were done only in scenarios when the number of time observations is several orders of magnitude greater than the number of time series.

5. Experiments

We performed experiments with HMMLGA and with exHMML on processes, which have an exponential distribution of types given in Section 3.3. We used the methods HGGM [1], LINGAM [6] and SFGC [7] for comparison. To assess similarity between the target and output causal graphs in synthetic experiments by all methods, we used the commonly applied F-measure, which takes both precision and recall into account.

5.1. Implementation and Parameter Setting

The comparison method HGGM uses Matlab package *penalized* from [34] with adaptive Lasso penalty. The algorithm in this package employs the Fisher scoring algorithm to estimate the regression coefficients. As recommended by the author of *penalized* in [34] and employed in [1], 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 [6] and used threshold = 0.05 and the 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.

In HMMLGA and exHMML, the initial estimates of $\boldsymbol{\beta}_i$ were achieved by the iteratively re-weighted 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 $\boldsymbol{\beta}_i$ by the IRLS procedure using function *penalized* with ridge penalty gave poor results in the experiments.) In case of gamma distribution, we achieved the estimates of parameters κ_i by statistical fitting, concretely by Matlab function *gamfit*. The minimization over λ_i was done by function *fminsearch*, which defined set H from Algorithm 1 as positive numbers from interval $[0.1, 1000]$.

5.2. Synthetically Generated Processes

To be able to evaluate the performance of HMML 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 processes, having an exponential distribution of Gaussian and gamma types from Section 3.3, together with the correspondingly generated target causal graphs. The performance of all tested algorithms 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, on the lag parameter. 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 [35], but not known for adaptive Lasso. In our experiments, we followed the recommendation of [1] on how to select the lag of time series in HGGM. It was 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.

We examined causal graphs with mixed types of time series for $p = 5$ and $p = 8$ number of features. For each case, we considered causal graphs with higher edge density (dense case) and lower edge density (sparse case), which corresponds to the parameter "dependency" in the code, where the full graph has for p time series $p(p-1)$ possible directed edges. Since we concentrate on a short time series in the paper; the length of generated time series varied from 100 to 1000.

5.2.1. Causal Networks with 5 and 8 Time Series

We considered 5 time series with 2 gamma, 2 Gaussian and 1 Poisson distributions, which we generated randomly together with the corresponding network. For the denser case with 5 time series,

we generated randomly graphs with 18 edges, and for the sparser case, random graphs with 8 edges. 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 10 random generations of causal graphs for length n and lag d. For dependency 8, we took strength = 0.9; for dependency 18, we took strength = 0.5 of causal connections.

Table 1. $p = 5$, average F-measure for each method, HMML, $n_g = 10$, $m = 20$, HGGM with $\lambda_{max} = 5$, LINGAM with $n/2$ boots. The first one subtable is for $d = 3$, the second one for $d = 4$.

dense g. 18, $n =$	100	300	500	1000;	sparse g. 8, $n =$	100	300	500	1000
exHMML	0.69	0.83	0.82	0.88	**exHMML**	0.70	0.72	0.72	**0.67**
HMMLGA	**0.73**	**0.90**	**0.89**	**0.90**	**HMMLGA**	**0.73**	**0.76**	**0.74**	**0.67**
HGGM	0.5	0.48	0.54	0.52	HGGM	0.52	0.36	0.66	0.36
LINGAM	0.57	0.58	0.62	0.58	LINGAM	0.58	0.54	0.69	0.45
SFGC	0.33	0.26	0.26	0.33	SFGC	0.14	0.35	0.44	0.31
dense g. 18, $n =$	100	300	500	1000;	sparse g. 8, $n =$	100	300	500	1000
exHMML	0.71	0.73	0.83	0.83	**exHMML**	**0.67**	**0.80**	**0.80**	0.68
HMMLGA	**0.82**	**0.79**	**0.87**	**0.92**	**HMMLGA**	**0.67**	0.73	0.77	**0.70**
HGGM	0.44	0.37	0.40	0.39	HGGM	0.53	0.47	0.65	0.36
LINGAM	0.71	0.58	0.58	0.65	LINGAM	0.33	0.52	0.74	0.46
SFGC	0.43	0.55	0.42	0.63	SFGC	0.35	0.59	0.42	0.38

One can see from Table 1 that HMMLGA and exHMML gave considerably higher precision in terms of F-measure than three other comparison methods, for all considered n up to 1000.

In the second network, we considered 8 time series with 7 gamma and 1 Gaussian distributions, which we generated randomly together with a corresponding network. For the denser case, we randomly generated graphs with 52 edges and for the sparser case random graphs with 15 edges. The results are presented in Table 2. Each value in Table 2 represents the mean value of all F-measures over 10 random generations of causal graphs for length n and lag d. For graph with 52 dependencies, we had strength = 0.3; for graph with 15 dependencies, strength = 0.9. Similarly as in the experiments with $p = 5$, one can see in Table 2 for $p = 8$ that both exHMML and HMMLGA gave considerably higher F-measure than the comparison methods for considered n up to 1000. The pair-wise hypothesis test used in SFGC for each pair of possible connections among all time series showed its efficiency for long time series in [1,7], however, it was in all experiments in our short-time series scenarios outperformed by LINGAM. The performance of method HGGM, efficient in long-term scenarios [1], was for 5 times series comparable to Lingam; for 8 times, this was the performance of HGGM the weakest from all the methods.

5.2.2. Performance of exHMML and MMLGA

The strategy to select the set γ_i with minimum HMML and with minimum number of regressors is applied in both methods. In exHMML, all 2^p possible values of HMML were sorted ascendently. Among those having the same minimum value, that one in the list is selected so that it has minimum number of ones (regressors) and is the last in the list. Similarly, this strategy is applied iteratively in HMMLGA on populations of individuals which have size $m < 2^p$. This strategy is an improvement with respect to MMLGA [12], where the first γ_i in the list with minimum MML function was selected. However, since the function HMML has multiple local minima, the convergence to the global minimum by both exHMML and HMMLGA cannot be guaranteed. The different performance of exHMML and HMMLGA for various p and various causal graph density is given by the nature of the objective function in (12) to be minimized. This function has multiple local minima. The above described implementation of both procedures for the exhaustive search and for the genetic algorithm, therefore, without any prior knowledge of the ground truth causal graph, can give different performance of

HMMLGA and exHMML. However as shown in the experiments, the achieved local minima are for both methods much closer to the global one than in case of the three rival methods.

Table 2. $p = 8$, average F-measure for each method, HMML, with $d = 3$, $n_g = 10$, $m = 20$, HGGM with $\lambda_{max} = 5$, LINGAM with $n/2$ boots. The first subtable is for $d = 3$, the second one for $d = 4$.

dense g. 52, $n =$	100	300	500	1000;	sparse g. 15, $n =$	100	300	500	1000
exHMML	0.68	**0.78**	**0.79**	0.82	**exHMML**	**0.69**	**0.73**	**0.77**	**0.64**
HMMLGA	**0.84**	0.67	0.66	**0.87**	**HMMLGA**	0.57	0.69	0.7	0.56
HGGM	0.16	0.17	0.17	0.17	HGGM	0.2	0.09	0.18	0.17
LINGAM	0.62	0.54	0.51	0.55	LINGAM	0.28	0.33	0.4	0.19
SFGC	0.32	0.21	0.35	0.20	SFGC	0.3	0.24	0.22	0.19
dense g. 52, $n =$	100	300	500	1000;	sparse g. 15, $n =$	100	300	500	1000
exHMML	0.59	0.64	0.56	0.75	**exHMML**	**0.58**	**0.84**	**0.80**	**0.69**
HGGMGA	**0.77**	**0.72**	**0.63**	**0.79**	**HMMLGA**	0.42	0.69	0.70	0.56
HGGM	0.16	0.16	0.18	0.17	HGGM	0.17	0.10	0.18	0.19
LINGAM	0.62	0.54	0.51	0.55	LINGAM	0.27	0.33	0.40	0.18
SFGC	0.36	0.45	0.82	0.83	SFGC	0.29	0.29	0.24	0.20

5.3. Climatological Data

We studied dynamics among seven climatic time series in a time interval. All time series were measured in the station of the Institute for Meteorology of the University of Life Science in Vienna 265 m above sea level [36]. Since weather is a very changeable process, it makes sense to focus on shorter time interval. We considered time series of dewpoint temperature (degree C, dew p), air temperature (degree C, air tmp), relative humidity (%, rel hum), global radiation (W m^{-2}, gl rad), wind speed (km/h, w speed), wind direction (degree, w dir), and air pressure (hPa, air pr). All processes were measured every ten minutes, which corresponds to $n = 432$ time observations for each time series. We concentrated on the temporal interactions of these processes during two scenarios. The first one corresponded to 7 to 9 July 2020 which were days with no rain. The second one corresponded to 16 to 18 July 2020 which were rainy days.

Before we applied the methods, we tested the distributions of each time series. In the first scenario (rainy days), air temperature (2) and global radiation (4) followed a gamma distribution and the remaining processes, the dew point temperature (1), relative humidity (3), wind speed (5), wind direction (6), and air pressure (7), following a Gaussian distribution. In the second scenario (dry days), wind direction (6) and air pressure (7) followed a Gaussian distribution, the dew point temperature (1), air temperature (2), relative humidity (3), global radiation (4) and wind speed (5), following a gamma distribution. To decide which of the algorithms exHMML or HMMLGA would be preferable to apply in this real valued experiment, we executed synthetic experiments for constellations of 5 gamma and 2 Gaussian (dry days), as well as of 2 gamma and 5 Gaussian (rainy days), with $n = 432$ for sparse and dense graphs with $d = 4$ and 5, each for 10 random graphs. Higher F-measure was obtained for HMMLGA, therefore we applied the HMMLGA procedure in the climatological experiments.

The resulting output graphs for methods HMMLGA, Lingam and HGGM for rainy and dry days gave the same graphs each for both lags; for dry days, we obtained, in HGGM, different graphs for each lag. We were interested in (a) how realistic were the detected temporal interactions of the processes by each method and in each scenario and (b) how realistic were the detected temporal interactions by each method, coming from the difference of graphs for dry and rainy days. In this case, we focused here only on the connections which differed in both graphs for each method. The figures of the output graphs for methods HMMLGA, Lingam, SFGC and HGGM for rainy and dry days can be for lag $d = 4$, seen in Figures 1 and 2.

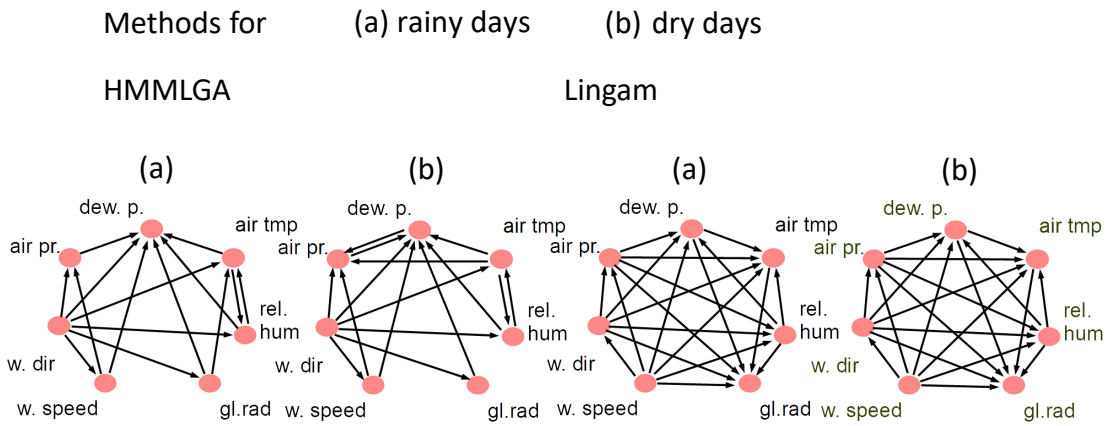

Figure 1. Output causal graphs for method HMMLGA and Lingam for rainy days and dry day scenarios.

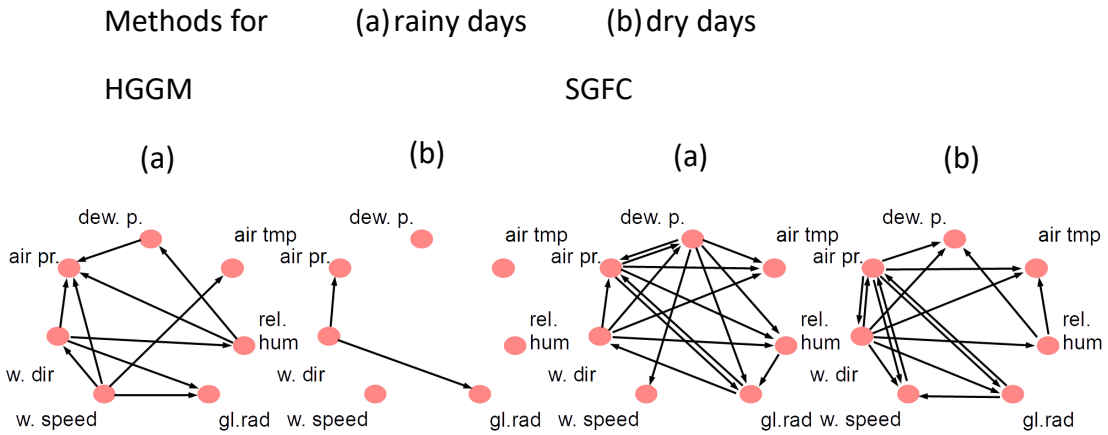

Figure 2. Output causal graphs for method HGGM and SFGC for rainy days and dry day scenarios.

For Lingam, the ouput graphs for rainy and dry days were identical and complete, so we omitted this method from further analysis.

Based on the expert knowledge [37], the temporal interactions in HMMLGA output graphs in both the rainy and dry scenarios correspond to the reality. In $HMMLGA_{D-R}$, which is the subgraph of HMMLGA of connections of the complement for dry days and of rainy days, the following directed edges in the form (cause, effect) were detected: (air tmp, air pr) and (dew p, air pr). The (direct) influence of dew point on air pressure is more strongly observable during sunny days, since the dew point is not possible to determine during rainy days. Similarly, the causal influence of air temperature on the air pressure is stronger during sunny days than during rainy days. So, both detected edges in HMMLGA were realistic. $HMMLGA_{R-D}$ was empty. Output graph $HGGM_{D-R}$ gave no edges. For $HGGM_{R-D}$, we obtained these directed edges: (dew p, air pr), which is, during rain, not observable, but the achieved influence (rel hum, dew p) is also during rain observable. Moreover, (rel hum, air pr) are observable (as humidity increases, pressure decreases). The edge (w speed, w dir) is not observable in reality, (w speed, air pr) is observable (higher wind speeds will show lower air pressure); also (w speed, air tmp) and (w speed, gl rad) are observable, however direct effect (w dir, rel hum) is not observable in reality. So, $HGGM_{R-D}$ had 2 falsely detected directions out of 8. Graph $SFGC_{R-D}$ gave this edge (dew p, air pr). Similarly, as in the case of HGGM, this edge is, during rain, not observable; (dew p, air tmp)—is during rain not observable; (dew p, w speed)—is during rain not observable; (dew p, rel hum)—is during rain not observable; (dew p, gl rad)—is during rain not observable; (rel hum, gl rad)—is during rain observable; (gl rad, w speed)—is during rain not observable; (gl rad, w dir)—is during rain not observable. So, $SFGC_{R-D}$ had 7 falsely detected directions out of 8. The output of $SFGC_{D-R}$ gave these edges: (rel hum, dew p)—this is during a dry

period observable; (rel hum, air tmp)—this is during a dry period observable; (gl rad, w speed)—this is during a dry period observable; (dev p, air tmp)—this is during a dry period observable; (air press, w dir)—this is during a dry period observable; (w speed, air pr)—this is during a dry period observable; (air pr, w speed) is during dry period in reality observable. So, $SFGC_{D-R}$ had 7 correctly detected directions out of 7.

We conclude that, in this climatological experiment, method HMMLGA, followed by SFGC, gave the most realistic causal connections with respect to the comparison methods.

5.4. Electrohysterogram Time Series

In the current obstetrics, there is no effective way of preventing preterm birth. The main reason is that no good objective method is known to evaluate the stepwise progression of pregnancy through to labor [38]. Any better understanding of the underlying labor dynamics can contribute to prevent preterm birth, which is the main cause of mortality in newborns. External recordings of the electrohysterogram (EHG) can provide new knowledge on uterine electrical activity associated with contractions.

We considered a database of 4-by-4 electrode EHG recordings performed on pregnant women, which were recorded in Iceland between 2008 and 2010 and are available via PhysioNet (PhysioBank ATM) [39]. This EHC grid (in the matrix form) was placed on the abdomen of the pregnant women. The electrode numbering, as considered in [38], can be found in Figure 3.

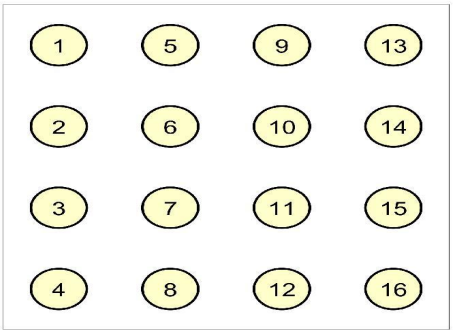

Figure 3. The ordering of the electrodes as mounted on the abdomen of women.

We applied the recordings, concretely for EHG signal for women in the third phase of pregnancy and during labor, to all the methods. We selected all (five) mothers for which the recordings were performed, both in the third trimester and during labor. Since there is no ground truth known on how the dynamics among the electrodes should look like for both modalities, we set a modest objective for us, whether HMMLGA and the comparison methods are able to distinguish labor from pregnancy from the EHG recordings. During labor, a higher density of interactions among electrodes is expected than during pregnancy, due to the higher occurrence of contractions of the uterine smooth muscles, which is also supported by some recent research in obstetrics, e.g., [40].

The 16 electromyographic time series (channels) were taken for all women (woman 11, 27, 30, 31 and 34), for each in the third trimester (P) and during labor (L). The observations in time series correspond to the time resolution every 5th microsecond. The time series in the databasis are commented by information about contraction, possible contraction, participant movement, participant change of position, fetal movement and equipment manipulation. By statistical fitting, we found out that all 16 time series followed Poisson distribution (setting raw ADC units in the Physionet database). We analysed the causal connections of each method for labor and pregnancy for all five women.

Since HMMLGA had higher F-measure than exHMML in the synthetic experiments with 16 Poisson time series, we considered further only HMMLGA in this real data experiment. In the synthetic experiments in [12], Poisson time series showed the highest F-measure on short time series, i.e., the case when the number of time observations is smaller than approximately two

orders times the number of time series. Based on this, we took the last 1200 observations for labor, since in the last phase, it was sure the labor had already started and the contractions had increased in time. Labor still continued for another few hours after the EHG recording finished for each of five women. For pregnancy time series, we took also 1200 observations, starting the moment where all electrodes had been fixed. The hypothesis, that during labor all electrodes were activated was confirmed by HMMLGA, HGGM and Lingam at all mothers. The hypothesis, that the causal graph during labor had higher density of causal connections than in the pregnancy case, was confirmed at all mothers by HMMLGA, for HGGM for mothers 30 and 31, but for SFGC and Lingam, we could not confirm it. In fact, Lingam gave identical complete causal graphs for both labor and pregnancy cases. The real computational time for Lingam (with 100 boots, as recommended by the authors) was for 16 time series and both labor and pregnancy modalities cca 12 h (in HP Elite Notebook); on the other side, for other methods, the time was in order of minutes. We present the causal graphs of all methods for labor and pregnant phase of mother 31 in Figure 4.

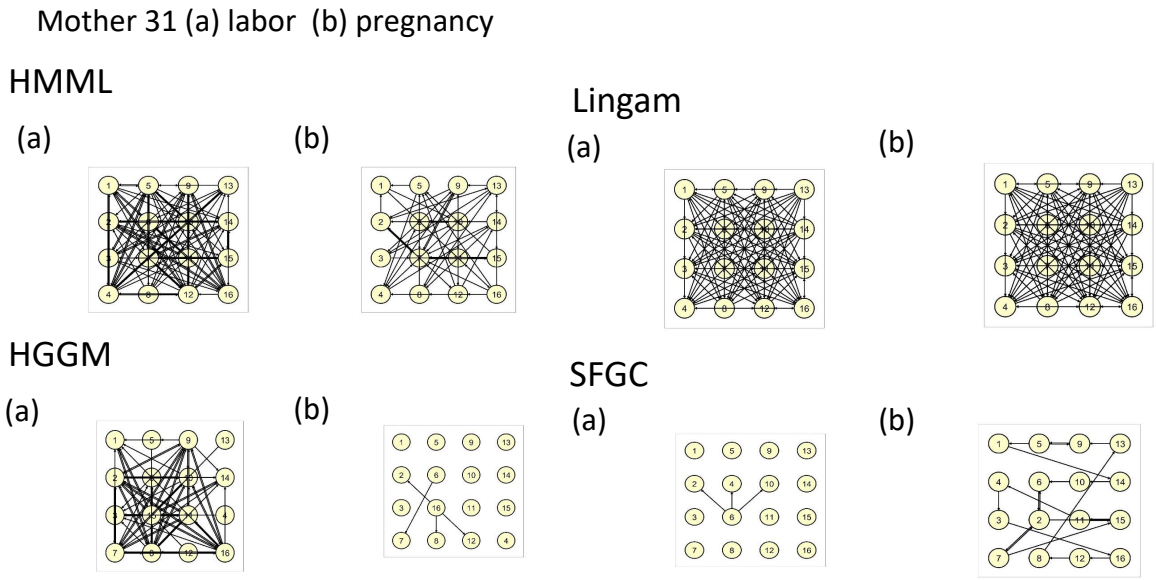

Figure 4. Output causal graphs for mother 31 during (**a**) labor and (**b**) pregnancy for all methods.

One can see that the density of connections by HMMLGA for labor is higher than for pregnancy. Causal graphs of HMMLGA for all mothers were for labor also denser than the pregnancy one. To make some more concrete hypotheses about the temporal interactions among the electrodes based on contractions, we would probably have to consider only intervals about which we know that they are without or with a limited number of artifacts in terms of participant movement, participant change of position, etc.

6. Conclusions

Common graphical Granger models in scenarios with short time series suffer often from overestimation, including the heterogeneous graphical Granger model. To remedy this, in this paper, we proposed to use the minimum message length principle for determination of causal connections in the heterogeneous graphical Granger model. Based on the dispersion coefficient of the target 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 causally connected time series with each target time series, and we derived its concrete form for various exponential distributions. We found this subset by a genetic-type algorithm (HMMLGA), which we have proposed as well as by exhaustive search (exHMML). We evaluated the complexity of these algorithms. The code in Matlab is provided. We demonstrated superiority of both methods with respect to the comparison methods in synthetic

experiments in short data scenarios. In two real data experiments, the interpretation of the causal connections as the result of HMMLGA was the most realistic with respect to the comparison methods. The superiority of HMMLGA with respect to the comparison methods for short time series can be explained by utilizing the dispersion of time series in the criterion as an additional (prior) information, as well as the fact that this criterion is optimized in the finite search space.

Author Contributions: Conceptualization, K.H.-S.; Data curation, K.H.-S.; Formal analysis, K.H.-S.; Investigation, K.H.-S.; Methodology, K.H.-S.; Resources, C.P.; Software, K.H.-S.; Supervision, C.P.; Validation, K.H.-S.; Visualization, K.H.-S.; Writing—original draft, K.H.-S.; Writing—review & editing, C.P., K.H.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Czech Science Foundation grant number GA19-16066S.

Acknowledgments: This work was supported by the Czech Science Foundation, project GA19-16066S. The authors thank to Dr. Irene Schicker and Dipl.-Ing. Petrina Papazek from [37] for their help with analysing the results of the climatological experiments. Open Access Funding by the University of Vienna.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of the MML Criterion for HGGM

Assume p independent random variables from the exponential family which are represented by time series $x_i^t, t = d+1, \ldots, n$ and for each i be $\hat{\phi}_i$ the given estimate of its dispersion. Consider the problem (10) for a given lag $d > 0$.

We consider now γ_i fixed, so for simplicity of writing we omit it from the list of variables of the functions. Having function L_i, we can now compute an initial estimate of $\hat{\boldsymbol{\beta}}_i$ from (10) which is the solution to the system of score equations. Since L_i forms 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 $\hat{\boldsymbol{\beta}}_i$ from (10).

Having parameters $\hat{\boldsymbol{\beta}}_i$, $\hat{\phi}_i$, $\boldsymbol{\Sigma}_i$ $\mathbf{W}_i$ and λ_i, we need to construct the function $HMML(\gamma_i)$: We use for each $i = 1, \ldots, p$ and for regression (10) Formula (18) from [23] i.e., the case when we plug in variables $\alpha := 0$ and $\boldsymbol{\beta} := \hat{\boldsymbol{\beta}}_i$ and $\mathbf{X} := \mathbf{X}_i$, $y := \mathbf{x}_i$, $n := n - d$, $k := k_i$, $\theta := \hat{\boldsymbol{\beta}}_i$, $\lambda := \lambda_i$, $\phi := \hat{\phi}_i$, $\mathbf{S} = \boldsymbol{\Sigma}_i$ be the unity matrix of dimension dk_i. The corrected Fisher information matrix for the parameters $\boldsymbol{\beta}_i$ is then $J(\hat{\boldsymbol{\beta}}_i|\hat{\phi}_i, \lambda_i) = (\frac{1}{\phi_i})\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i + \lambda_i\boldsymbol{\Sigma}_i$. 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 which are independent of k_i we omitted from the HMML code, since the optimization w.r.t. γ_i is independent of them. Among all subsets $\gamma_i \in \Gamma$, there are $\binom{p}{k_i}$ subsets of size k_i. If nothing is known a priori about the likelihood of any covariate $\mathbf{x}_i$ being included in the final model, a prior that treats all subset sizes equally likely $\pi(|\gamma_i|) = 1/(p+1)$ is appropriate [23]. This gives the code length $I(\gamma_i) = \log\binom{p}{k_i} + \log(p+1)$ as in (12).

Appendix B. Derivation of L_i, $\mathbf{W}_i$, ϕ_i for Various Exponential Distributions of $\mathbf{x}_i$

Case $\mathbf{x}_i$ is Gaussian Since in this case is $\phi_i = \sigma_i^2$ its variance, we will omit ϕ_i from the list of parameters which condition function p. L_i in (15) is obtained directly from (14) by applying logarithm on it. By plugging values for identity link corresponding to the Gaussian case as $\eta_i^t = \mu_i^t = [\mathbf{X}_i\boldsymbol{\beta}_i]^t$ and $\frac{\delta\eta_i^t}{\delta\mu_i^t} = 1$ into Formula (13) from [23], matrix $\mathbf{W}_i = \mathbf{I}_{k_i d \times k_i d}$ is directly obtained.

Case $\mathbf{x}_i$ is binomial Assuming ϕ_i be a constant, we can omit ϕ_i from the list of parameters which condition function p. L_i in (17) is obtained directly from (16) by applying logarithm on it. As in the previous case, it is obtained by plugging values into formula (13) from [23]. Value of $\mathbf{W}_i$ from (19) is obtained by plugging values for logit link corresponding to the binomial case as $\eta_i^t = [\mathbf{X}_i\boldsymbol{\beta}_i]^t = \log(\frac{\mu_i^t}{1-\mu_i^t})$ and $\frac{\delta\eta_i^t}{\delta\mu_i^t} = \frac{1}{\mu_i^t(1-\mu_i^t)}$ into Formula (13) from [23]. In case we cannot assume $\phi_i = 1$, we apply the sandwich estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_i$ for robust estimation which for a general logistic

regression can be found in e.g., [41]) and in our case it gives matrix $\mathbf{W}_i$ in the form $\mathbf{W}_i = \text{diag}\big((x_i^1 - \frac{\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1)}{(1+\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^1))^2})^2, \ldots, (x_i^{n-d} - \frac{\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d})}{(1+\exp([\mathbf{X}_i\hat{\boldsymbol{\beta}}_i']^{n-d}))^2})^2\big)$.

Case $\mathbf{x}_i$ is Poisson First we will express the log-likelihood function L_i in terms of parameters $\boldsymbol{\beta}_i$. Since we use Poisson model for $\mathbf{x}_i$ 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 $\boldsymbol{\beta}_i$, the probability of attaining $x_i^{d+1}, \ldots, x_i^n$ is given by $p(x_i^{d+1}, \ldots, x_i^n | \mathbf{X}_i, \boldsymbol{\beta}_i) = \prod_{t=d+1}^n \frac{(\mu_i^t)^{x_i^t} \exp(-\mu_i^t)}{(x_i^t)!} = \prod_{t=d+1}^n \frac{\exp([\mathbf{X}_i\boldsymbol{\beta}_i']^t)^{x_i^t} \exp(-\exp([\mathbf{X}_i\boldsymbol{\beta}_i']^t))}{x_i^t!}$ and $\eta_i^t = \exp([\mathbf{X}_i\boldsymbol{\beta}_i']^t)$, (recalling the notation from Section 3.2, $[\mathbf{X}_i\boldsymbol{\beta}_i']^t$ denotes the t-th coordinate of the vector $\mathbf{X}_i\boldsymbol{\beta}_i'$). The log-likelihood in terms of $\boldsymbol{\beta}_i$ is $L_i = \log p(\boldsymbol{\beta}_i | \mathbf{x}_i, \mathbf{X}_i) = \sum_{t=d+1}^n x_i^t [\mathbf{X}_i\boldsymbol{\beta}_i']^t - \exp([\mathbf{X}_i\boldsymbol{\beta}_i']^t) - \log(x_i^t!)$. Now we derive matrix $\mathbf{W}_i$ for $\mathbf{x}_i$ with (exact) Poisson distribution: The Fisher information matrix $J_i = J(\boldsymbol{\beta}_i) = -\mathbb{E}_{\boldsymbol{\beta}_i}(\nabla^2 L_i(\boldsymbol{\beta}_i | \mathbf{x}_i, \mathbf{X}_i))$ may be obtained by computing the second order partial derivatives of L_i for $r, s = 1, \ldots, k_i$. This gives

$$\begin{aligned} \frac{\delta^2 L_i(\boldsymbol{\beta}_i|\mathbf{x}_i,\mathbf{X}_i)}{\delta^2 \beta_i^r \beta_i^s} &= \frac{\delta L_i}{\delta \beta_i^s} \sum_{t=d+1}^n [x_i^t \sum_{l=1}^d x_r^{t-l} - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{t-l} \beta_j^l) \sum_{l=1}^d x_r^{t-l}] \\ &= -\sum_{t=d+1}^n \exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{t-l} \beta_j^l)(\sum_{l=1}^d x_s^{t-l})(\sum_{l=1}^d x_r^{t-l}). \end{aligned} \tag{A1}$$

If we denote $\mathbf{W}_i := diag(\exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{d+1-l} \beta_j^l), \ldots, \exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{n-l} \beta_j^l))$ then we have Fisher information matrix $J(\boldsymbol{\beta}_i) = (\mathbf{X}_i)' \mathbf{W}_i \mathbf{X}_i$. Alternatively, $\mathbf{W}_i$ can be obtained by plugging values into formula (13) from [23]. Value of $\mathbf{W}_i$ from (22) is obtained by plugging values for log link corresponding to the Poisson case as $\eta_i^t = [\mathbf{X}_i\boldsymbol{\beta}_i]^t = \log(\mu_i^t)$ and $\frac{\delta \eta_i^t}{\delta \mu_i^t} = \frac{1}{\mu_i^t}$ into Formula (13) from [23].

Derivation of matrix $\mathbf{W}_i$ for $\mathbf{x}_i$ with overdispersed Poisson distribution: Assume now the dispersion parameter $\phi_i > 0, \neq 1$. The variance of the overdispersed Poisson distribution is $\phi_i \boldsymbol{\mu}_i$. We know that the Poisson regression model can be still used in overdispersed settings and the function L_i is the same as $L_i(\boldsymbol{\beta}_i)$ derived above. We use the robust sandwich estimate of covariance of $\hat{\boldsymbol{\beta}}_i$ as it was proposed in [42] for general Poisson regression. The Fisher information matrix of overdispersed problem is $J_i = J(\boldsymbol{\beta}_i) = (\mathbf{X}_i)' \mathbf{W}_i \mathbf{X}_i$ where $\mathbf{W}_i$ is constructed for $\mathbf{x}_i$ Poisson based on [42] and has the form $\mathbf{W}_i = diag([x_i^{d+1} - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{d+1-l} \beta_j^l)]^2, \ldots, [x_i^n - \exp(\sum_{j=1}^{k_i} \sum_{l=1}^d x_j^{n-l} \beta_j^l)]^2)$.

Case $\mathbf{x}_i$ is gamma L_i in (25) is obtained directly from (24) by applying logarithm on it. By plugging values for log link corresponding to the gamma case as $\eta_i^t = \frac{1}{\mu_i^t}$ and $\frac{\delta \eta_i^t}{\delta \mu_i^t} = \frac{1}{(\mu_i^t)^2}$ into Formula (13) from [23], matrix $\mathbf{W}_i$ from (26) is directly obtained.

Case $\mathbf{x}_i$ is inverse-Gaussian L_i in (28) is obtained directly from (24) by applying logarithm on it. By plugging values for log link corresponding to the inverse-Gaussian case as $\eta_i^t = [\mathbf{X}_i\boldsymbol{\beta}_i]^t = \log(\mu_i^t)$ and $\frac{\delta \eta_i^t}{\delta \mu_i^t} = \frac{1}{\mu_i^t}$ into Formula (13) from [23], matrix $\mathbf{W}_i$ from (29) is directly obtained.

References

1. Behzadi, S.; Hlaváčková-Schindler, K.; Plant, C. Granger Causality for Heterogeneous Processes. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Springer: Cham, Switzerland, 2019.
2. Zou, H. The adaptive lasso and its oracle property. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
3. Hryniewicz, O.; Kaczmarek, K. Forecasting short time series with the bayesian autoregression and the soft computing prior information. In *Strengthening Links Between Data Analysis and Soft Computing*; Springer: Cham, Switzerland, 2015; pp. 79–86.
4. Bréhélin, L. A Bayesian approach for the clustering of short time series. *Rev. D'Intell. Artif.* **2006**, *20*, 697–716. [CrossRef]
5. Wallace, C.S.; Boulton, D.M. An information measure for classification. *Comput. J.* **1968**, *11*, 185–194. [CrossRef]

6. 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. *J. Mach. Learn. Res.* **2011**, *12*, 1225–1248.
7. Kim, S.; Putrino, D.; Ghosh, S.; Brown, E.N. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput. Biol.* **2011**, *7*, e1001110. [CrossRef]
8. Arnold, A.; Liu, Y.; Abe, N. Temporal causal modeling with graphical Granger methods. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2007; pp. 66–75.
9. Shojaie, A.; Michailidis, G. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **2010**, *26*, i517–i523. [CrossRef]
10. Lozano, A.C.; Abe, N.; Liu, Y.; Rosset, S. Grouped graphical Granger modeling methods for temporal causal modeling. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 577–586.
11. Nelder, J.; Wedderburn, R. Generalized Linear Models. *J. R. Stat. Soc. Ser. A (General)* **1972**, *135*, 370–384. [CrossRef]
12. Hlaváčková-Schindler, K.; Plant, C. Poisson Graphical Granger Causality by Minimum Message Length. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2020 (ECML/PKDD), Ghent, Belgium, 14–18 September 2020.
13. Granger, C.W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
14. Mannino, M.; Bressler, S.L. Foundational perspectives on causality in large-scale brain networks. *Phys. Life Rev.* **2015**, *15*, 107–123. [CrossRef]
15. Maziarz, M. A review of the Granger-causality fallacy. *J. Philos. Econ. Reflect. Econ. Soc. Issues* **2015**, *8*, 86–105.
16. Granger, C.W. Some recent development in a concept of causality. *J. Econom.* **1988**, *39*, 199–211. [CrossRef]
17. Lindquist, M.A.; Sobel, M.E. Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. *NeuroImage* **2011**, *57*, 334–336. [CrossRef] [PubMed]
18. Spirtes, P.; Glymour, C.N.; Scheines, R.; Heckerman, D. *Causation, Prediction, and Search*; MIT Press: Cambridge, MA, USA, 2000.
19. Glymour, C. Counterfactuals, graphical causal models and potential outcomes: Response to Lindquist and Sobel. *NeuroImage* **2013**, *76*, 450–451. [CrossRef] [PubMed]
20. Marinescu, I.E.; Lawlor, P.N.; Kording, K.P. Quasi-experimental causality in neuroscience and behavioural research. *Nat. Hum. Behav.* **2018**, *2*, 891–898. [CrossRef] [PubMed]
21. Wallace, C.S.; Freeman, P.R. Estimation and inference by compact coding. *J. R. Stat. Soc. Ser. B* **1987**, *49*, 240–252. [CrossRef]
22. Wallace, C.S.; Dowe, D.L. Minimum message length and Kolmogorov complexity. *Comput. J.* **1999**, *42*, 270–283. [CrossRef]
23. Schmidt, D.F.; Makalic, E. Minimum message length ridge regression for generalized linear models. In *Australasian Joint Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2013; pp. 408–420,
24. Segerstedt, B. On ordinary ridge regression in generalized linear models. *Commun. Stat. Theory Methods* **1992**, *21*, 2227–2246. [CrossRef]
25. Computational complexity of mathmatical operations. Available online: https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations (accessed on 2 October 2020). [CrossRef]
26. Rissanen, J. *Stochastic Complexity in Statistical Inquiry*; World Scientific: Singapore, 1989; Volume 15, p. 188.
27. Barron, A.; Rissanen, J.; Yu, B. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760.
28. Hansen, M.; Yu, B. Model selection and minimum description length principle. *J. Am. Stat. Assoc.* **2001**, *96*, 746–774. [CrossRef]
29. Hansen, M.H.; Yu, B. Minimum description length model selection criteria for generalized linear models. *Lect. Notes Monogr. Ser.* **2003**, *40*, 145–163. [CrossRef]
30. Marx, A.; Vreeken, J. Telling cause from effect using MDL-based local and global regression. In Proceedings of the 2017 IEEE International Conference on Data Mining, New Orleans, LA, USA, 18–21 November 2017; pp. 307–316.

31. Marx, A.; Vreeken, J. Causal inference on multivariate and mixed-type data. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Dublin, Ireland, 10–14 September 2018; Volume 2018, pp. 655–671.
32. Budhathoki, K.; Vreeken, J. Origo: Causal inference by compression. *Knowl. Inf. Syst.* **2018**, *56*, 285–307.
33. Hlaváčková-Schindler, K.; Plant, C. Graphical Granger causality by information-theoretic criteria. In Proceedings of the European Conference on Artificial Intelligence 2020 (ECAI), Santiago de Compostela, Spain, 29 August–2 September 2020; pp. 1459–1466. [CrossRef]
34. McIlhagga, W.H. Penalized: A MATLAB toolbox for fitting generalized linear models with penalties. *J. Stat. Softw.* **2016**, *72*. [CrossRef]
35. Zou, H.; Hastie, T.; Tibshirani, R. On the "degrees of freedom" of the lasso. *Ann. Stat.* **2007**, *35*, 2173–2192. [CrossRef]
36. Available online: https://meteo.boku.ac.at/wetter/mon-archiv/2020/202009/202009.html (accessed on 5 September 2020). [CrossRef]
37. Zentralanstalt für Meteorologie und Geodynamik 1190 Vienna, Hohe Warte 38. Available online: https://www.zamg.ac.at/cms/de/aktuell (accessed on 5 September 2020).
38. Alexandersson, A.; Steingrimsdottir, T.; Terrien, J.; Marque, C.; Karlsson, B. The Icelandic 16-electrode electrohysterogram database. *Nat. Sci. Data* **2015**, *2*, 1–9.
39. Available online: https://www.physionet.org (accessed on 5 September 2020). [CrossRef]
40. Mikkelsen, E.; Johansen, P.; Fuglsang-Frederiksen, A.; Uldbjerg, N. Electrohysterography of labor contractions: propagation velocity and direction. *Acta Obstet. Gynecol. Scand.* **2013**, *92*, 1070–1078.
41. Agresti, A. *Categorical Data Analysis*; Section 12.3.3.; John Wiley and Sons: Hoboken, NJ, USA, 2003; Volume 482. [CrossRef]
42. Huber, P.J. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA; 1967; Volume 1, pp. 221–233.