ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200252

Graphical Granger Causality by Information-Theoretic Criteria

Kateřina Hlaváčková-Schindler[1] and **Claudia Plant**[2]

Abstract. Causal inference by a graphical Granger model (GGM) among p variables is typically solved by p penalized linear regression problems in time series with a given lag. In practice however, the estimates of a penalized linear regression after a finite number of steps can be still far from the optimum. Furthermore, the selection of the regularization parameter, influencing the precision of the model is not trivial, especially when the corresponding design matrix is super-collinear. In this paper, for the first time we concept a graphical Granger model as an instance of combinatorial optimization. Computing maximum likelihood (ML) estimates of the regression coefficients and of the variance for each of p variables we propose an information-theoretic graphical Granger model (ITGGM). In the sense of information theory, the criterion to be minimized is the complexity of the class of the selected models together with the complexity of the data set. Following this idea, we propose four various information-theoretic (IT) objective functions based on stochastic complexity, on minimum message length, on Akaike and on Bayesian information criterion. To find their minima we propose a genetic algorithm operating with populations of subsets of regressor variables. The feature selection by the ITGGM with any of the functions is parameter-free in the sense that beside the ML estimates which are for each and within the model constant, no adjustable parameter is added into these objective functions. We further provide a theoretical analysis of the convergence properties of the GGM with the proposed IT functions. We test the performance of the functions in terms of F_1 measure with respect to two common penalized GGMs on synthetic and real data. The experiments demonstrate the significant superiority of the IT criteria in terms of F_1 measure over the two alternatives of the penalized GGM for Granger causal inference.

1 INTRODUCTION

Granger causality is a popular method for causal inference in time series due to its computational simplicity. The assumption of this approach is that knowing a cause helps to predict its effects in the future. Over the last decade, graphical Granger models, i.e. the multivariate Granger causality based on vector autoregressive regression (VAR) extended the Granger concept for more than two time series. Due to the high number of involved variables, the corresponding optimization problems are ill-posed and a penalization criterion can be enforced, e.g. in [2] and [18]. The solutions of the corresponding penalized problem ideally converge to the solution of the original constrained problem. This is however not guaranteed since the Lasso variable selection by the Gaussian regression can be without additional condition inconsistent [41]. Consequently, there may be multiple solutions that minimize the Lasso loss function. Another issue is the selection of the regularization parameter.

Instead of searching for an optimum in an infinite continuous search space with additional penalties, in this paper we present the graphical Granger problem as an instance of the combinatorial optimization in a finite search space. In the sense of information theory, the criterion to be minimized is the complexity of the class of the selected models together with the complexity of the data set. By means of ML inference, we compute the estimates of the variance and of the regression coefficients of the GGM for each of p variables. Based on this pair of statistics we propose the IT criteria for GGM. The key idea of the IT criteria is that if a statistical model compresses data, then the model captures (with high probability) the regularities in the data. Due to the bias-variance tradeoff problem, a similar tradeoff applies to the selection of an appropriate criterion for a data set for the graphical Granger model with assumptions on a model. Consequently, one cannot set one IT objective function as a universal one. The main contributions of the paper are as follows:

- For the first time, we concept the feature selection problem by a graphical Granger model as a problem of combinatorial optimization.
- We convert the matrix of lagged time series in GGM into a fixed design matrix form for each of p variables. By maximum likelihood we computed a pair of statistics describing each of p variables in the GGM with the super-collinear design matrix.
- We propose four IT objective functions for GGM with these statistics.
- We provide a theoretical analysis of convergence properties of the GGM with the proposed IT functions.
- We design the information-theoretic genetic algorithm ITGA, operating with populations of subsets of regressor variables.
- Our results on causal inference with synthetic and real data demonstrate the significant superiority of the IT criteria based on stochastic complexity, on minimum message length, on Akaike and on Bayesian information criterion, in terms of F_1 measure over two common penalized GGMs for Granger causal inference.

The paper is organized as follows. Section 2 presents the necessary background. Related work and research gaps are discussed in Section 3. Section 4 brings the problem of graphical Granger model into the form suitable for the IT criteria, further the derivation of the IT criteria for GGM as well as theoretical analysis of convergence properties of the GGM with the IT criteria. Our optimization procedure ITGA is proposed in Section 5. Experiments on synthetic and real data are in Section 6. Section 7 is devoted to conclusions.

[1] Faculty of Computer Science, University of Vienna, Vienna, Austria, email: katerina.schindlerova@univie.ac.at
[2] Faculty of Computer Science, University of Vienna, Vienna, Austria and ds:UniVie, University of Vienna, Vienna, Austria email: claudia.plant@univie.ac.at

2 PRELIMINARIES

2.1 Granger causality

Granger causality [9], has been for decades a popular tool to discover temporal relationships among variables. Assume two stationary time series $x = \{x^t | t = 1, \ldots, n\}$ and $y = \{y^t | t = 1, \ldots, n\}$. Let the following two models represent two autoregressive models corresponding to time series y with and without the past observations of x taken into consideration:

$$\begin{aligned} y^t &= \alpha_1 y^1 + \cdots + \alpha_{t-1} y^{t-1} + \beta_1 x^1 + \ldots \\ &\quad + \beta_{t-1} x^{t-1} + \varepsilon^t \end{aligned} \tag{1}$$

$$y^t = \alpha_1 y^1 + \cdots + \alpha_{t-1} y^{t-1} + \varepsilon^t. \tag{2}$$

Following the principle of Granger causality, x Granger-causes y if the model (1) significantly improves the predictability of y comparing to the model (2). This concept can be extended to $p > 2$ time series and time lag $d \geq 1$. The corresponding model to this is called graphical Granger model. For simplicity, the lag d is assumed for all time series the same. Let $x_1^t, \ldots, x_p^t$ be p time series up to time n. The vector-autoregressive (VAR) model is for $i = 1, \ldots, p$ given by:

$$x_i^t = X_{t,d}^{Lag} \beta_i' + \varepsilon_i^t \tag{3}$$

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' is the transpose of the matrix of the regression coefficients and ε_i^t is white noise. The time series x_j Granger–causes the time series x_i for the given lag d, denote $x_j \to x_i$ 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 (3) is non-zero [2].

2.2 Causal inference by penalization

Since (3) is in general an ill-posed problem, it is for each $i = 1, \ldots, p$ reformulated for numerical treatment by adding a penalty, serving as a variable selection method estimating the effects of variables on each other, i.e.

$$\hat{\beta}_i = \arg\min_{\beta_i} \sum_{t=d+1}^{n} \left(x_i^t - X_{t,d}^{Lag} \beta_i'\right)^2 + \lambda P(\beta_i) \tag{4}$$

for a given d and all $t = d+1, \ldots, n$ and $i = 1, \ldots, p$ and a penalty $P(\beta_i)$ and λ a regularization parameter. Granger model assumes the same lag d for each i which we also consider. Similarly, the time series x_j Granger–causes the time series x_i for the given lag d, denote $x_j \to x_i$ for $i, j = 1, \ldots, p$ if and only if at least one of the d coefficients in the $j-th$ row of $\hat{\beta}_i$ of the solution of (4) is non-zero. Most papers use Lasso or elastic net penalty, e.g. [2],[18].

3 RELATED WORK

Granger causality based on vector autoregression with Lasso penalty was introduced by Arnold et al. in [2] under the name Temporal causal modeling with Lasso penalty (TCML). During the last decade, other penalty constraints have been employed, e.g. elastic net and adaptive Lasso in [18] or truncated Lasso in [38]. As Lasso penalty together with linear regression is in general inconsistent, these methods do not have to provide a high precision in terms of common evaluation measures (e.g. F_1, precision or recall). Furthermore, the precision of the penalized GGM depends on the selection of λ parameter. For λ close to zero, the Lasso regression estimate is close to the maximum likelihood estimate, whereas for large λ, the penalty term can worsen the fitting of the least squares. Therefore choices of λ should compromise between those two extremes and it is usually done iteratively. By adding one variable at a time, an early choice of one variable may influence when other variables correlated to it enter later in the relaxation process. The penalty parameter should compromise the above extremes and is in practice chosen by K-fold cross-validation. However even this does not guarantee that the optimal solution for coefficients β_i will be found. Zou in [41] gave examples when (a general) Lasso regression is inconsistent and derived the so called oracle properties as a necessary condition under which it is consistent. The adaptive Lasso, which he introduced, fulfills the oracle properties. It uses adaptive weights for penalizing different coefficients in the l_1 penalty. The graphical Granger model with adaptive Lasso (we denote it ADTCML) is under the oracle conditions consistent. Although the number of iterations in which Lasso and adaptive Lasso optimization algorithm achieve a solution satisfactorily close to the optimum can be estimated under some conditions (e.g. in [25], [41]), for a general objective function it is still an open problem. Therefore in practice, after a finite number of steps, the achieved estimators can be still far from the optimum in both methods. Our experiments in this paper with TCML and ADTCML demonstrated this scenario.

There are papers dealing with causality detection by information-theoretic criteria. Among them, compression-based algorithms apply the Kolmogorov complexity and define so called causal indicators by mean of the minimum description length (MDL) for numeric data [23], [5] and for mixed data types (i.e. numeric or discrete) in [24]. However, the algorithms in [23] and [24] do not use time series. The algorithm in [5] uses event sequences, however from a discrete space of observations. The Granger causality here is not defined by a VAR model but by prediction probabilities of the event sequences. Moreover, all these methods are designed to infer the pairwise causal relations. A direct application of them for discovery in causal networks can lead to the decrease in precision, especially when the number of processes increases. Our paper deals with VAR-based graphical Granger causality. Peters et al. [28] introduces the method TiMINo testing causal inference on time series using restricted structural equation models. Since TiMINo is not a VAR based causality we did not use it for comparison in our experiments. In our preliminary results comparing our criteria experimentally with Bayesian networks (BN) by IT from [17] we observed that our criteria outperform BN in precision for 'long' time series, and vice versa.

Criteria AIC or BIC have been used already in connection with Granger causality, however only to determine the lag d, e.g in [40], which was done by the pairwise causal testing. To our best knowledge, we are not aware of any application of compression schemes on causality for multiple time series by a graphical Granger method. By deriving the IT criteria for a graphical Granger method as in the following sections, our paper fills this research gap.

3.1 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 of counterfactuals, [21], [26]. As its name implies, Granger causality is not necessarily true causality. In defense of his method, Granger in [10] 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 whether a causal relation exists 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 [22], if the theo-

retical 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 [14], since they are much less costly than common epidemiological or marketing research approaches.

3.2 Variable selection in the general Gaussian regression by information-theoretic criteria

Rissanen in [30] proposed to use normalized maximum likelihood (NML) as a variable selection method for linear regression (i.e. no time series and no lag was considered). The normalized maximum likelihood formulation of stochastic complexity contains two components, the maximized log likelihood and the component called parametric complexity of the model. The stochastic complexity for the data, relative to a suggested model, serves as a criterion for model selection. The calculation of the stochastic complexity can be understood as an implementation of the minimum description length principle (MDL) [32]. To obtain an NML based model selection criterion for the Gaussian linear regression, [31] defines appropriate constraints on the data space.

The idea of using an IT criterion as a variable selection method for linear regression is to choose a subset of the regressor variables with indices from γ which is a subset of the indices of all regressors and in this way to eliminate the regression variables that are not relevant for the expression of the independent variable [29]. Based on the probability distribution of data matrix X_γ corresponding to model γ, maximum likelihood estimates $\hat{\theta}_\gamma = (\hat{\beta}_\gamma, \hat{\sigma}^2_\gamma)$ are computed, where β_γ is the vector of regression coefficients and σ^2_γ is the variance of the model. The goal is to pick the model γ with least NML costs. Two the most well-known IT criteria for model selection applying maximum likelihood principle are AIC [1] and BIC [35]. Other IT based criterion for variable selections uses minimum message length principle and is from in [34]. All these criteria assume the design matrix of a full column rank (which is not the case of GGM).

4 GRAPHICAL GRANGER MODEL AS p MULTIPLE LINEAR REGRESSIONS

In this paper we present the graphical Granger model problem among p variables as a combinatorial optimization problem in a finite search space of an exponential size for each of p variables. The criterion to be minimized is the complexity of the class of selected models and of the data for each of p variables. Corresponding to (3) we propose several objective functions as information-theoretic functionals and compare their performance on synthetic and real data. The IT criteria mentioned above cannot be applied to graphical Granger model (3) directly due to that they are designed for designed matrices with a full column rank and they do not consider a vector-autoregressive process of a given lag parameter. In the following we convert problem (3) into an appropriate form so that we can then use these criteria.

Consider the full model for p variables x_i^t and (integer) lag $d \geq 1$ corresponding to the optimization problem (3) with $\{x_i^t, i = 1, \ldots, p, t = 1, \ldots, n\}$. We put the time series into the $p \times n$ matrix X. Let ε_i^t be Gaussian distributed errors with zero-mean and $\beta_i \in \mathbb{R}^{1\times(p\times d)}$ for all $t = 1, \ldots, n$, $i = 1, \ldots, p$; For fixed $d \geq 1$ can (3) be rewritten as

$$x_i^t = \sum_{j=1}^{p} \sum_{l=1}^{d} x_j^{t-l} \beta_j^l + \varepsilon_i^t \tag{5}$$

for $t = d+1, \ldots, n$ and ε_i is a $(n-d)$ dimensional vector of Gaussian noise with zero mean and unit variance. To be able to use the ML estimation, in the following we convert the matrix of lagged time series in (3) into a fixed design matrix form of the multiple vector regression. Denote $x_i = (x_i^{d+1}, x_i^{d+2}, \ldots, x_i^n)$. Assume $n - d > pd$. We construct the $(n-d) \times (d \times p)$ design matrix $X =$

$$\begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_p^d & \ldots & x_p^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_p^{d+1} & \ldots & x_p^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_p^{n-1} & \ldots & x_p^{n-d+1} \end{bmatrix} \tag{6}$$

and a $1 \times (d \times p)$ vector $\beta_i = (\beta_1^1, \ldots, \beta_1^d, \ldots, \beta_p^1, \ldots, \beta_p^d)$ and $\varepsilon_i = (\varepsilon_i^1, \ldots, \varepsilon_i^{n-d})$. We can see that

$$x_i' = X\beta_i' + \varepsilon_i' \tag{7}$$

is the (design) matrix form expressing the problem (5). Denote now by $\gamma_i \subset \{1, \ldots, p\}$ the subset of indices of regressor variables to infer x_i and $k_i := |\gamma_i|$ its cardinality. Let $\beta_i(\gamma_i) \in \mathbb{R}^{1\times(d\times k_i)}$ be the vector of unknown regression coefficients with a fixed ordering within the γ_i subset. Assume for illustration purposes and without lack of generality that the first k_i indices out of p vectors belong into γ_i. Considering only the columns from the matrix X in (6) corresponding to γ_i, we define the $(n-d) \times k_i$ matrix of lagged vectors with indices from γ_i as $X(\gamma_i) :=$

$$\begin{bmatrix} x_1^d & \ldots & x_1^1 & \ldots & x_{k_i}^d & \ldots & x_{k_i}^1 \\ x_1^{d+1} & \ldots & x_1^2 & \ldots & x_{k_i}^{d+1} & \ldots & x_{k_i}^2 \\ x_1^{d+2} & \ldots & x_1^3 & \ldots & x_{k_i}^{d+2} & \ldots & x_{k_i}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & x_1^{n-d+1} & \ldots & x_{k_i}^{n-1} & \ldots & x_{k_i}^{n-d+1} \end{bmatrix} \tag{8}$$

The problem (7) for regressor variables with indices from γ_i is expressed as

$$x_i' = X(\gamma_i)\beta_i'(\gamma_i) + \varepsilon_i'(\gamma_i), \tag{9}$$

with $\beta_i(\gamma_i)$ to be a $1 \times (d \times k_i)$ matrix of unknown coefficients and $\varepsilon_i(\gamma_i)$ is a $(n-d)$-dimensional Gaussian noise vector with zero mean and unit variance. Wherever it will be clear from context, we write β_i instead $\beta_i(\gamma_i)$, X_i instead of $X(\gamma_i)$ and σ_i^2 instead of $\sigma_i^2(\gamma_i)$.

4.1 The normalized maximum likelihood criterion for graphical Granger model

4.1.1 Derivation of the ML estimates

We assume that in model (9) each x_i follows Gaussian distribution with the density function $f(x_i, \gamma_i, \beta_i, \sigma_i^2) =$

$$\frac{1}{(2\pi\sigma_i^2)^{(n-d)/2}} \exp\Big[-\frac{1}{2\sigma_i^2} \sum_{t=d+1}^{n} (x_i^t - X_i\beta_j^t)^2\Big]. \tag{10}$$

ML estimates of the parameters of model (9) are then

$$\hat{\beta}_i = [X_i'X_i]^{-1}X_i'x_i' \text{ and } \hat{\sigma}_i^2 = \frac{\|x_i' - X_i\hat{\beta}_i'\|^2}{(n-d-k_i)} \tag{11}$$

where $k_i = |\gamma_i|$ is the number of elements of γ_i, i.e. the number of regressors in X_i, each with lag d and of length $n - d$. Denote $\hat{\theta}_i := (\hat{\beta}_i, \hat{\sigma}_i^2)$. This pair is a sufficient statistic for estimating $\theta_i = (\beta_i, \sigma_i^2)$.

4.1.2 Selection by stochastic complexity

The MDL principle for a model selection is based on the idea to capture regular features in data by constructing a model in a certain class which allows the shortest description of the data and the model itself. For each x_i satisfying (5) we consider the family of models $M_i := M(\gamma_i) = \{f(x_i,\theta_i), \gamma_i \in \Gamma\}$ defined by densities in (10) where Γ is the set of all subsets of $\{1,\dots,p\}$. Denote the ML estimate $\hat{\theta}_i$ of θ_i such that $f(x_i,\hat{\theta}_i) = \max_{\theta_i} f(x_i,\theta_i)$.

By constructing the SC based objective function, we use the structure of the normalized maximum likelihood (NML) density function, which was introduced in [31] for a Gaussian linear regression. We apply this construction for each GGM with $i = 1,\dots,p$ and get

$$\hat{f}(x_i,\hat{\theta}_i) = \frac{f(x_i,\hat{\theta}_i)}{C(\gamma_i)}, \quad C(\gamma_i) = \int f(x_i,\hat{\theta}_i)dx_i \tag{12}$$

where $\hat{f}(x_i,\hat{\theta}_i)$ is a density function provided that $C(\gamma_i)$ is bounded. The NML density function provides a general technique to apply the MDL minimum description length (MDL) principle for statistical model selection. Thus the derivation of the NML density of the model γ_i and $C(\gamma_i)$ is crucial for practical implementation of the MDL principle. Applying logarithm on both sides of (12) we get the expression

$$-\log \hat{f}(x_i,\gamma_i) = -\log f(x_i,\hat{\theta}_i) + \log C(\gamma_i) \tag{13}$$

where $C(\gamma_i)$ is a normalizing term. This is the "shortest code length" for x_i that can be obtained by γ_i and is called stochastic complexity of x_i given γ_i. We denote it by $SC(x_i,\gamma_i)$. [31] suggested for a general linear regression to constrain the data space, since C can be infinite, Applying these constraints appropriately for GGM we get the NML density function

$$\hat{f}(x_i,\gamma_i) = \frac{f(x_i,\gamma_i;\hat{\beta}_i,{\hat{\sigma}_i}^2)}{\int_{P(R_i,s_i)} f(x_i,\gamma_i;\hat{\beta}_i,{\hat{\sigma}_i}^2)dx_i} \text{ with constraints}$$

$$P(R_i,s_i) := P(x_i,\gamma_i,R_i,s_i) = \{x_i; \|X_i\hat{\beta}_i'\|^2 \le (n-d)R_i; {\hat{\sigma}_i}^2 \ge s_i\} \tag{14}$$

where $R_i > 0$ and $s_i > 0$ are positive constants. The NML density under the constraints (14) is

$$\hat{f}(x_i,R_i,s_i) = f(x_i,\hat{\theta}_i)/C(s_i,R_i) \tag{15}$$

where the normalizing constant $C(s_i,R_i)$ depends on two hyperparameters R_i and s_i, similarly as in [31]. To get rid of the two hyperparameters in $C(s_i,R_i)$, we maximize (15) with respect to s_i,R_i and get the ML estimates of the mean and variance of the GGM corresponding to (9) as $\hat{R}_i = \frac{\|X_i\hat{\beta}_i'\|^2}{n-d}$ and $\hat{s} = \hat{\sigma}_i^2$. If we substitute them into (15), we get the maximized NML function (mNML). In the second stage, the normalization of the data space is constrained such that $P(x_i) =$

$$\{x_i; (n-d)R_{i,1} \le \|X_i\hat{\beta}_i'\|^2 \le (n-d)R_{i,2}; s_i^1 \le \hat{\sigma}_i^2 \le s_i^2\} \tag{16}$$

where $0 \le R_{i,1} \le R_{i,2}$, $0 \le s_i^1 \le s_i^2$ are given positive constants. By normalizing function $\hat{f}(x_i,\hat{s}_i,\hat{R}_i)$ we get the normalized mNML function $\hat{f}(x_i)$. The stochastic complexity (13) has then the form

$$\begin{aligned} SC(x_i,\gamma_i) = {} & \frac{n-d-k_i}{2}\log\hat{\sigma}_i^2 + \frac{k_i}{2}\log\hat{R}_i \\ & - \log\Gamma\left(\frac{n-d-k_i}{2}\right) - \log\Gamma\left(\frac{k_i}{2}\right) + c, \end{aligned} \tag{17}$$

where $\Gamma(.)$ is the gamma function. Since c is a constant for all models γ_i, we can omit c from minimization of SC.

4.2 Other information-theoretic criteria for GGM

Due to the bias-variance tradeoff problem, selecting an appropriate criterion for a data set for the GGM with assumptions on a model brings a similar tradeoff. Consequently, one cannot set one criterion as a universal one. In this paper we are inspired by IT criteria that were derived for a general linear regression problem with different additional assumptions, concretely Bayesian information criterion (BIC) in [35], Akaike information criterion (AIC) in [36], the minimum message length criterion (MML) in [34], and with three criteria SC_i proposed for collinear design matrices in [15], [8]. By plugging the above derived statistic $\hat{\theta}_i = (\hat{\beta}_i,\hat{\sigma}_i^2)$ into these criteria appropriately enables us to propose the following criteria for each variable x_i, $i = 1,\dots,p$ in model (9): The BIC, AIC and MML criterion for GGM, respectively have for $\hat{\theta}_i$ the form

$$BIC(x_i,\gamma_i) = \frac{n-d}{2}\log\hat{\sigma}_i^2 + \frac{k_i}{2}\log(n-d). \tag{18}$$

$$AIC(x_i,\gamma_i) = \frac{n-d}{2}\ln\hat{\sigma}_i^2 + \frac{(k_i+1)(n-d+k_i+2)}{(n-d-k_i-2)} - \frac{k_i}{n-d-k_i} \tag{19}$$

$$\begin{aligned} MML(x_i,\gamma_i) = {} & \tfrac{n-d-k_i}{2}\ln(2\pi) + \tfrac{n-d-k_i}{2}\left(\ln\left(\tfrac{(n-d)\hat{\sigma}_i^2}{n-d-k_i}\right)+1\right) \\ & + \frac{k_i}{2}\ln(\pi x_i x_i') - \ln\Gamma\left(\frac{k_i}{2}+1\right) + \frac{1}{2}\ln(k_i+1). \end{aligned} \tag{20}$$

For (9) and statistic $\hat{\theta}_i$ we consider for GGM IT functions penalizing the collinear design matrix

$$\begin{aligned} SC_1(x_i,\gamma_i) = {} & (n-d)\ln\hat{\sigma}_i^2 + k_i\ln\frac{\|X(\gamma_i)\hat{\beta}_i\|^2}{k_i(n-d)} - \\ & - \ln(k_i(n-d-k_i)), \end{aligned} \tag{21}$$

$$\begin{aligned} SC_2(x_i,\gamma_i) = {} & \frac{n-d-k_i}{2}\ln\frac{\hat{\sigma}_i^2}{n-d-k_i} + \frac{k_i}{2}\ln\frac{\|\hat{\beta}_i\|^2}{k_i(n-d)} \\ & - \ln(k_i(n-d-k_i)) + \ln|X_iX_i'|, \end{aligned} \tag{22}$$

$$\begin{aligned} SC_3(x_i,\gamma_i) = {} & \frac{n-d-k_i}{2}\ln\frac{\hat{\sigma}_i^2}{n-d-k_i} + \frac{k_i}{2}\ln\frac{\|X_i\hat{\beta}_i\|^2}{k_i(n-d)} \\ & - \ln(k_i(n-d-k_i)) - \ln|X_iX_i'|. \end{aligned} \tag{23}$$

4.3 Convergence properties of the objective functions

As our design matrix X_i is constructed, its column vectors are highly correlated, i.e. X_i is super-collinear, or in other words its condition number is high. In the following we address the convergence questions of the considered estimators.

4.3.1 Penalized GGM under super-collinearity

Lasso regression is known to yield a sparse solution, in which many regression coefficients are equal to zero. Since the predictors x_i and their lagged version are highly correlated in X_i, the estimator achieved by TCML or ADTCM $\hat{\beta}_i(\lambda)$ may contain "too many" zeroes. This is problematic if the true GGM model is not sparse.

4.3.2 IT GGM under super-collinearity

The ML estimates of $\hat{\beta}_i$ and of $\hat{\tau}_i$ in (11) are identical to the least-squares (LS) estimators, thus $\hat{\beta}_i$ and $\hat{\tau}_i$ are consistent estimators under the condition that x_i is a stationary, stable Gaussian VAR process of order d and $\sqrt{n}(\hat{\beta}_i - \beta_i)$ and $\sqrt{n}(\hat{\tau}_i - \tau_i)$ are asymptotically normally distributed. It follows from more general results, see e.g. [20], Chapter 3.4. A general ML estimate is efficient, i.e. it achieves the Cramér–Rao lower bound when the sample size tends to infinity. This means that no consistent estimator has lower asymptotic mean squared error than the ML estimate (or other estimators attaining this bound). For a general linear regression with a regular design matrix it was shown for SC in [8], for BIC in [11] and for AIC in [37] that if the true model is finite-dimensional and is included in the set of candidates, then the probability that this model is selected goes to one as the sample size increases. However, if the true model is not finite-dimensional, then SC is asymptotically efficient in the sense that it selects the candidate model which minimizes the one-step mean squared error of prediction. The same property for AIC was proven in [37]. We explain the above statements in terms of condition numbers of the design matrices and iterative processes. A regular design matrix has condition number equal one. It is commonly known that an algorithm, which introduces no errors of its own, can in a finite number of iterations find an approximation of the solution of the regression with the design matrix with condition number exactly one, whose precision is no worse than that of the data. Our design matrices X_i have high condition numbers. In this case, fitting of the linear regression model to the data brings a large error of the estimates of the regression parameters corresponding to the collinear covariates. For a design matrix with a high condition number, including the smallest singular values in the inversion merely adds numerical noise to the solution. This can be cured with the truncated SVD approach [12], which explicitly sets all singular values below a certain threshold to zero. In this way the truncated SVD computes more stable and exact estimates of the regression coefficients for all considered IT criteria. The truncated SVD approach does not theoretically guarantee the convergence of the IT functions in a finite number of steps, however, in a sense the truncated SVD version of X_i is the closest approximation to X_i that can be achieved by a matrix of $rank(X_i)$ (the last follows from [7]). As we will demonstrate in the experiments, the SVD approximation of X_i gives for the IT criteria considerably better solutions in F_1 measure as by the rival penalization methods.

5 OPTIMIZATION OF THE INFORMATION-THEORETIC FUNCTIONS

Denote $S(\gamma_i)$ as a representative of any of the proposed IT objective functions. The selection of the best structure γ_i amounts to evaluate values of $S(\gamma_i)$ for all $\gamma_i \subset \{1, \ldots, p\}$. i.e. for all 2^p possible subsets and then to pick the subset under which the minimum of the function was achieved. Since precision is a more important issue than speed in Granger causality inference, in the following we propose a genetic algorithm type procedure to find a minimum of $S(\gamma_i)$.

5.1 Information-theoretic genetic algorithm (ITGA)

For a fixed i and $d \geq 1$, for a $\gamma_i \subset \{1, \ldots, p\}$, $|\gamma_i| = k_i$ we define a Boolean vector Q_i of length p corresponding to γ_i such that it has ones in the positions of the indices from γ_i, otherwise zeros. Genetic algorithm ITGA executes genetic operations on populations of Q_i. The procedure is summarized in Algorithm 1. 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 of length p. Then we select $m/2$ individuals in the current population with lowest $S(Q_i)$ as the elite subpopulation of parents of the next population. For a predefined number of generated populations np, 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. Mutation corresponds to a random change in Q_i and 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 $S(Q_i)$ to form the next generation. The algorithm stops when the predefined number of generations ng is achieved.

Algorithm 1 ITGA

Input: $series, d, np, ng, m, z$
m an even integer, z position for off-spring;
$series$:= matrix of $x_i^t, i = 1, \ldots, p, t = 1, \ldots, n - d$;
Output: Adj := adj. matrix of the output causal graph;
// for every x_i find Q_i with minimum $S(Q_i)$
for all x_i **do**
  Create initial population $\{Q_i^j, j = 1, \ldots, m\}$ at random; Compute $S(Q_i^j)$ for each $j = 1, \ldots, m$;
  v:=1;
  while $v \leq ng$ **do**
    u:=1;
    while $u \leq np$ **do**
Sort $S(Q_i^j)$ 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 $S(Q_i^j)$ for each j; Mutate a single parent Q_i^j at a random position; Compute $S(Q_i^j)$ for each j; Add the children with minimum $S(Q_i^j)$ 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 $S(Q_i)$
return (Adj)

6 EXPERIMENTS

To assess the similarity between the target and output causal graph by ITGA for a corresponding IT function we use the commonly defined F_1-measure. We executed experiments on four synthetic and three real-valued data sets. Based on [4], given sufficient number of observations n, a graphical Granger model is for $\frac{n}{d} > p + 1$ consistent, otherwise inconsistent. For insufficiently long time series we also observed lower F_1 in our experiments. Among the many methods mentioned in Section Related Work, for a fair comparison we use only the VAR based Granger causal methods since they are designed for Gaussian time series, similarly as our criteria. Our code in Matlab including the data sets and supplementary material (SM) are publicly available at: `http://tinyurl.com/yxvkv7ae`.

6.1 Synthetic experiments

For all seven IT objective functions we did experiments with synthetic causal networks. As comparison penalization methods we

used VAR based methods TCML and ADTCML and Matlab package $penalized$ [27]. Based on the strategy in [27], for TCML and ADTCML we varied the variable penalization parameter λ from interval $(0, \lambda_{max}]$ and took the best result with respect to F_1 measure. The other algorithms mentioned in Related Work do not use a VAR model, therefore we did not apply them as comparison methods.

6.1.1 Two causal networks with five units

The first causal network among five processes is given by Figure 2a and corresponding equations in paper [3] with $d = 3$; The second causal network among five processes is from [33] with $d = 4$. These equations in both cases generate five processes with random noise $\sim N(0, 1)$. In the first series of experiments we examined the search space of all seven constructed functions with statistic $\hat{\theta}_i$ for each process x_i and for variable n in the graphical Granger model. The goal was to find for which of the functions these statistics $\hat{\theta}_i$ are sufficiently descriptive, i.e. whether the global minimum of each function described by these statistics corresponds to the global minimum with the biggest F_1 measure. We explored the search spaces for $n = 50, \ldots, 2000$. We found by exhaustive search for each objective function a subset γ_i for each $i = 1, \ldots, 5$ for which this objective function achieved a minimum and compared it to the ground truth. In all examined cases the objective function SC had the highest F_1 value (0.98), in most cases followed by MML, BIC, AIC. As a comparison method we used TCML, where in procedure $penalized$ we set $\lambda_{max} = 5$, taking the best result with respect to F_1 from the interval $(0, 5]$. Interestingly, although designed to penalize the collinearity of X_i matrices, SC_1, SC_2 and SC_3 had for all investigated d and n F_1 values smaller than those by TCML. Therefore we excluded these three objective functions from further experiments. Table 1 shows some results of these experiments with equations from [3] with $np = 50$ each averaging over 50 random generations of time series. It demonstrates that the IT functions give significantly better F_1 than TCML. Similar results were achieved also in experiments with [33] and can be found in SM.

Table 1. F_1 values for each method, variable n, $ng = 50$ random generations of series, TCML: $\lambda_{max} = 5$, equations from [3].

d=3, n=	50	100	500	1600	1800	2000
SC	0.73	0.86	0.92	**0.98**	**0.98**	**0.98**
MML	0.72	0.85	0.95	0.97	**0.98**	**0.98**
BIC	0.73	0.86	0.94	0.97	0.97	0.97
AIC	0.71	0.83	0.71	0.71	0.70	0.72
TCML	0.53	0.53	0.54	0.57	0.58	0.59

In the second series of experiments with five processes from Figure 2a from [3] we evaluated the performance of F_1 for $d = 3$, and n ranging with values from 50 to 2000 (selected values) using ITGA. We considered $m = 30$ as the size of of a genetic population and the number of generated populations $np = 30$ or 50, by TCML and ADTCML $\lambda_{max} = 5$ and considered their highest F_1 measures over interval $(0, \lambda_{max}]$. Table 2 demonstrates the superiority of all IT methods in terms of F_1 values. A similar F_1 superiority of IT methods for the second causal network with 5 units, $d = 4$ from [33] is demonstrated in SM. Figure 1 shows the superior behavior of F_1 of the IT methods for variable n from 100 to 500 and for both networks with five processes comparing to TCML and ADTCML. Similar experiments for n from 1200 to 2000 for both networks with 5 units can be found in SM. The best precision in F_1 for both causal networks with 5 units is achieved in most cases of n by SC, BIC and MML (without order); Criteria AIC and TCML and ADTCML had however considerably lower F_1. The figures of the trajectories of F_1 for all six objective functions with respect to n for both cases of 5 unit networks can be found in SM. The runtime of TCML and ADTCML with $\lambda_{max} = 5$ in Matlab and on HP notebook with Intel Core i5-7200U processor was e.g. in the first network for $n = 1400$ 0.75 seconds and 5.8 seconds, respectively. The optimization of any IT method was slower (genetic algorithm), e.g. for $n = 1400$, with $ng = 50$, $np = 30$, SC method took 55 seconds, MML 57 seconds, BIC 58 seconds and AIC 61 seconds.

Table 2. F_1 for each method, $np = 50$, $m = 30$, $ng = 20$, TCML and ADTCML with $\lambda_{max} = 5$, equations from [3].

d=3, n=	100	500	1400	1600	1800	2000
SC	0.79	0.92	**0.93**	**0.93**	0.91	0.90
MML	0.80	0.91	**0.92**	0.91	0.91	0.90
BIC	0.78	0.90	**0.92**	0.91	0.91	0.90
AIC	0.65	0.70	0.71	0.70	0.70	0.68
TCML	0.53	0.61	0.64	0.66	0.66	0.68
ADTCML	0.35	0.40	0.40	0.40	0.40	0.40

6.1.2 A causal network with ten units

The best precision in F_1 for the causal network with ten units (see SM for the equations) are given by SC, BIC and MML (in this order) in most cases of values n. The experiments for the lag $d = 4$, $ng = 10, np = 30$ and $m = 30$ for variable n for methods SC, BIC, AIC, MML, TCML and ADTCML with $\lambda_{max} = 5$ can be found in Table 3. We can conclude that in all our synthetic experiments, the highest F_1 measure was achieved by SC, MML and BIC objective functions and AIC, TCML and ADTCML had considerably lower F_1.

Table 3. F_1 values for each method, 10 series, $d = 4$, $np = 30$, $m = 30$, $ng = 10$, TCML and ADTCML: $\lambda_{max} = 5$.

d=4,n=	100	500	1400	1600	1800	2000
SC	0.59	0.69	0.71	0.70	0.70	0.71
MML	0.59	0.70	**0.72**	**0.72**	**0.72**	**0.72**
BIC	0.61	0.70	**0.73**	0.71	**0.73**	**0.73**
AIC	0.54	0.56	0.56	0.57	0.57	0.57
TCML	0.45	0.48	0.44	0.46	0.45	0.43
ADTCML	0.35	0.30	0.30	0.30	0.28	0.29

6.1.3 A causal network with twenty units

The best F_1 for the causal network with twenty units (see SM for the equations) were achieved in most cases of $n = 500, \ldots, 4000$ by MML and BIC ($F_1 = 0.64$), followed by SC ($F_1 = 0.63$) and then by TCML and ADTCML ($F_1 = 0.55$). Table 4 demonstrates the superiority of the IT methods over TCML and ADTCML. To conclude for all considered networks, one can see that F_1 of the IT methods is decreasing with the number of units, however this happens also for TCML and ADTCM, with even lower F_1. This is due to the fact that with increasing p, the condition number of matrices X_i is increasing.

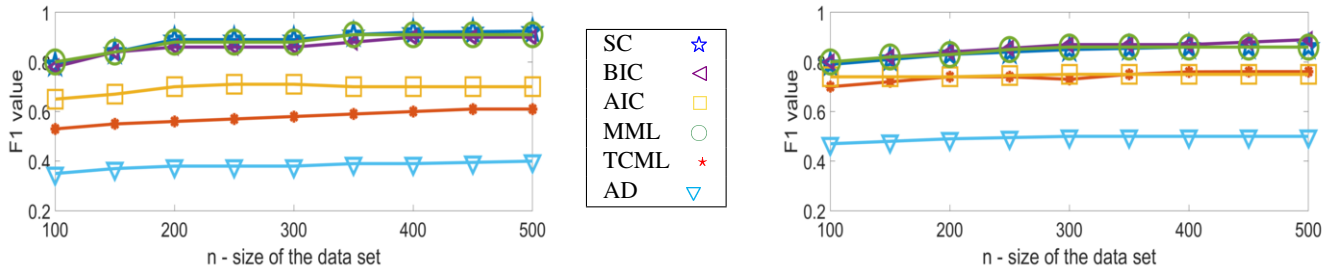

Figure 1. Experiment with 5 series, $ng = 20$, $np = 50$, $m = 30$ for $n = 100, 200, 300, 400, 500$; TCML, AD=ADTCML with $\lambda_{max} = 5$, left: network from [3] and $d = 3$; right: network from [33] with $d = 4$.

Table 4. F_1 values for each method, network with 20 units, equations in SM with lag $d = 2$, variable data set size n, $ng = 10$, $m = 30$, TCML and ADTCML with $\lambda_{max} = 5$.

d=2,n=	500	1000	2000	3000	4000	5000
SC	0.63	0.63	0.63	0.63	0.63	0.62
MML	**0.64**	0.63	**0.64**	0.63	0.63	0.63
BIC	**0.64**	**0.64**	**0.64**	**0.64**	0.63	0.63
AIC	0.61	0.61	0.63	0.61	0.61	0.61
TCML	0.50	0.49	0.50	0.51	0.54	0.55
ADTCML	0.51	0.54	0.55	0.55	0.55	0.55

6.2 Experiments with real data

6.2.1 *Gene expression time series*

In most real-data applications, the ground truth causal network is not known, so only an external knowledge (e.g. a biological experiment) can be used for assessing the performance. For real biological datasets, the number of genes p is usually far more than the number of time points n. As mentioned in the theory above and also observed in the synthetic experiments, short time series are limiting for the precision of a graphical Granger model. Therefore, one cannot expect very high F_1-values with time series created by biological experiments. We used the gene expressions time series from [39]. These are selected 19 genes (with their names in SM) that are active in human cancer cell lines, each with 47 time observations. By statistical fitting, these time series follow a normal distribution. The corresponding gene regulatory network was reconstructed based on the biological experiments by Li et al. and its figure (benchmark) can be found in [18]. The results of our experiments for $d = 3$ are in the first row of Table 5. The IT methods are again superior over the rival methods in terms of F_1.

Table 5. F_1 for each method, $m = 20$, $np = 30$, $\lambda_{max} = 5$; $19G = 19$ genes, $7PF = 7$ portfolios, $12C = 12$ climatic.

data set, n	*SC*	*MML*	*BIC*	*AIC*	*TCML*	*ADTCML*
19G, 47	0.58	0.57	**0.63**	0.61	0.35	0.11
7PF, 1400	0.56	0.62	**0.65**	0.55	0.44	0.09
12C, 500	0.74	0.78	**0.80**	**0.80**	0.53	0.78
12C, 1500	**0.80**	**0.80**	**0.80**	**0.80**	0.53	0.78

6.2.2 *Portfolio time series*

The portfolio data set and the causal graph as a benchmark are from [13]. The seven time series with $n = 1400$ are daily values of the portfolio returns, i.e. values of gain or loss realized by an investment portfolio in time. By statistical testing we confirmed the normal distribution of each time series. The lag $d = 1, 2$ was considered and the best performance of F_1 was achieved for $n = 1400$ with $d = 2$ and is in the first row of Table 5. Figure 2 shows comparison of the external knowledge with BIC and TCML method. One can see that portfolios 1,4 and 5 have for the BIC method more similar results to the external knowledge as by TCML to the external knowledge. The superiority of all IT methods in terms of F_1 over the rival methods for variable n from 1300 to 2500 is shown in SM.

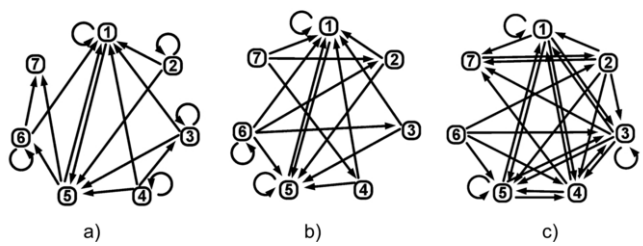

Figure 2. 7 portfolios, a) extern. knowledge, b) BIC with $m = 30$, $np = 30$, c) TCML with $\lambda_{max} = 5$, both $d = 1$.

6.2.3 *Climatic data*

The data set and external knowledge come from spatio-temporal causal modeling on climatic data from North America as described in [19]. The goal of the study was to find the variables attributing the changes in temperature (undirected dependencies). The monthly observations in high resolution from [16] were considered for 12 variables, namely methane (CH4), carbon-dioxide (CO2), hydrogen (H2), carbon-monoxide (CO), UV (AER), temperature (TMP), precipitation (PRE), vapor (VAP), cloud cover (CLD), wet days (WET), frost days (FRS), global horizontal (SOL). Figure 6d in [19] was taken as a benchmark network. It is an undirected subgraph of adjacent variables to the variable temperature (TEMP). In this task we assigned to the resulting output subgraph of each method its symmetrized version, since the benchmark graph was the graph of undirected dependencies. We did testing with $n = 500, 1500,$

$m = 20, np = 3, d = 3$. The significant superiority of all IT methods over the rival methods in terms of F_1 is demonstrated in the last two rows of Table 5.

7 CONCLUSIONS

In this paper, for the first time we proposed an information-theoretic graphical Granger model (ITGGM) together with a genetic algorithm procedure ITGA operating with subsets of possible regressors of the GGM. ITGA is scalable for both the number of time series and their length; There is no obstacle for its parallelisation. We proposed four information-theoretic criteria for GGM and discussed their convergence properties. The comparison of ITGGM with the proposed IT criteria on synthetic examples and real data demonstrated the significant superiority of the IT criteria based on stochastic complexity, on minimum message length, on Akaike and on Bayesian information criterion, in precision by F_1 over the two penalized graphical Granger models TCML and ADTCML. The idea of using IT criteria for causal inference among non-Gaussian time series is a topic of our future research. For comparison of our IT functions on multiple data sets we also intend to apply non-parametric tests for statistical comparisons of classifiers, e.g. the Wilcoxon signed ranks test for comparison of two classifiers or the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets, e.g. [6].

REFERENCES

[1] H. Akaike, 'Information Theory and an Extension of the Maximum Likelihood Principle', in *Selected Papers of Hirotugu Akaike*, 199–213, Springer, (1998).

[2] A. Arnold, Y. Liu, and N. Abe, 'Temporal Causal Modelling with Graphical Granger Methods', *ACM SIGKDD*, (2007).

[3] L.A. Baccalá and K. Sameshima, 'Partial Directed Coherence: A New Concept in Neural Structure Determination', *Biol. Cybernetics*, **84**(6), 463–474, (2001).

[4] M.T. Bahadori and Y. Liu, 'An Examination of Practical Granger Causality Inference', in *Proceedings of SIAM ICDM*, (2013).

[5] K. Budhathoki and J. Vreeken, 'Causal Inference on Event Sequences', in *Proceeding of the SIAM ICDM*, (2018).

[6] J. Demšar, 'Statistical comparisons of classifiers over multiple data sets', *Journal of Machine learning research*, **7**(Jan), 1–30, (2006).

[7] M. Frank and J.M. Buhmann, 'Selecting the rank of truncated svd by maximum approximation capacity', in *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 1036–1040. IEEE, (2011).

[8] C.D. Giurcaneanu, S.A. Razavi, and A. Liski, 'Variable Selection in Linear Regression: Several Approaches Based on Normalized Maximum likelihood', *Signal Processing*, **91**(8), 1671 – 1692, (2011).

[9] C.W.J. Granger, 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods', *Econometrica*, 424–438, (1969).

[10] C.W.J. Granger, 'Some Recent Development in a Concept of Causality', *Journal of Econometrics*, **39**(1-2), 199–211, (1988).

[11] M.H. Hansen and B. Yu, 'Minimum description length model selection criteria for generalized linear models', *Lecture Notes-Monograph Series*, 145–163, (2003).

[12] P.C. Hansen, 'Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank', *SIAM Journal on Scientific and Statistical Computing*, **11**(3), 503–518, (1990).

[13] S. Kleinberg, *Causality, Probability, and Time*, Cambridge University Press, 2013.

[14] H. Kwak, C. Lee, H. Park, and S. Moon, 'What Is Twitter, a Social Network or a News Media?', in *Proceedings of the 19th International Conference on World Wide Web*, pp. 591–600. ACM, (2010).

[15] E.P. Liski and A. Liski, *Minimum Description Length Model Selection in Gaussian Regression under Data Constraints*, 201–208, Physica HD, Heidelberg, 2009.

[16] Y. Liu. USC Melady Lab. `http://www-bcf.usc.edu/~liu32/melady.html`, 2019. Approached 2019-03-12.

[17] Z. Liu, M.B. Malone, and C. Yuan, 'Empirical evaluation of scoring functions for Bayesian network model selection', *Bioinformatics*, **13**(15), 1–16, (2012).

[18] A.C. Lozano, N. Abe, Y. Liu, and S. Rosset, 'Grouped Graphical Granger Modeling for Gene Expression Regulatory Networks Discovery', *Bioinformatics*, **25**(12), i110–i118, (2009).

[19] A.C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe, 'Spatial-Temporal Causal Modeling for Climate Change Attribution', in *ACM SIGKDD KDD*, pp. 587–596, (2009).

[20] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, Springer Science and Business Media, 2005.

[21] M. Mannino and S.L. Bressler, 'Foundational Perspectives on Causality in Large-Scale Brain Networks', *Physics of Life Reviews*, **15**, 107–123, (2015).

[22] I.E. Marinescu, P.N. Lawlor, and K.P. Kording, 'Quasi-experimental Causality in Neuroscience and Behavioural Research', *Nature Human Behaviour*, 1, (2018).

[23] A. Marx and J. Vreeken, 'Telling Cause from Effect Using MDL-Based Local and Global Regression', in *IEEE ICDM*, pp. 307–316, (2017).

[24] A. Marx and J. Vreeken, 'Causal Inference on Multivariate and Mixed-Type Data', in *ECML PKDD 2018*, pp. 655–671, (2018).

[25] P. Massart and C. Meynet, 'Some rates of convergence for the selected Lasso estimator', in *International Conference on Algorithmic Learning Theory*, pp. 17–33. Springer, (2012).

[26] M. Maziarz, 'A Review of the Granger-causality Fallacy', *The Journal of Philosophical Economics: Reflections on Economic and Social Issues*, **8**(2), 86–105, (2015).

[27] W. McIlhagga, 'penalized: A MATLAB Toolbox for Fitting Generalized Linear Models with Penalties', *Journal of Stat. Software, Articles*, **72**(6), (2016).

[28] J. Peters, D. Janzing, and B. Schölkopf, 'Causal inference on time series using restricted structural equation models', in *Advances in Neural Information Processing Systems*, pp. 154–162, (2013).

[29] J. Rissanen, 'Modeling by Shortest Data Description', *Automatica*, **14**(5), 465–471, (1978).

[30] J. Rissanen, 'Fisher Information and Stochastic Complexity', *IEEE Trans. Inf. Theor.*, **42**(1), 40–47, (September 1996).

[31] J. Rissanen, 'MDL Denoising', *IEEE Transactions on Information Theory*, **46**(7), 2537–2543, (2000).

[32] J. Rissanen, *Information and complexity in statistical modeling*, Springer Science & Business Media, 2007.

[33] B. Schelter, M. Winterhalder, B. Hellwig, B. Guschlbauer, C.H. Lücking, and J. Timmer, 'Direct or Indirect? Graphical Models for Neural Oscillators', *Journal of Physiology-Paris*, **99**(1), 37–46, (2006).

[34] D.F. Schmidt and E. Makalic, 'MML Invariant Linear Regression', in *Advances in Artificial Intelligence*, pp. 312–321, (2009).

[35] G. Schwarz, 'Estimating the Dimension of a Model', *The Annals of Statistics*, **6**(2), 461–464, (1978).

[36] A. Seghouane, 'Asymptotic Bootstrap Corrections of AIC for Linear Regression Models', *Signal Processing*, **90**(1), 217–224, (2010).

[37] R. Shibata, 'Asymptotically efficient selection of the order of the model for estimating parameters of a linear process', *The annals of statistics*, 147–164, (1980).

[38] A. Shojaie and G. Michailidis, 'Discovering Graphical Granger Causality Using the Truncating Lasso Penalty', *Bioinformatics*, **26**(18), i517–i523, (2010).

[39] M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, and P. Brown, 'Identification of Genes Periodically Expressed in the Human Cell Cycle and their Expression in Tumors', *Molecular Biology of the Cell*, **13**(6), 1977–2000, (2002).

[40] D. Zhou, Y. Xiao, Y. Zhang, Z. Xu, and D. Cai, 'Granger Causality Network Reconstruction of Conductance Based Integrate and Fire Neuronal Systems', *PloS One*, **9**(2), e87636, (2014).

[41] H. Zou, 'The Adaptive Lasso and Its Oracle property', *Journal of the American Statistical Association*, 1418–1429, (2008).