Granger Causality for Ill-Posed Problems: Ideas, Methods, and Application in Life Sciences

Kateřina Hlaváčková-Schindler[1], Valeriya Naumova[2], Sergiy Pereverzyev Jr.[3]

[1]Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic, e-mail: katerina.schindler@gmail.com

[2]Center for Biomedical Computing, Simula Research Laboratory, Lysaker, Norway, e-mail: valeriya@simula.no

[3]Applied Mathematics Group, Department of Mathematics, University of Innsbruck, Austria, e-mail: sergiy.pereverzyev@uibk.ac.at

Abstract

Granger causality, based on a vector autoregressive model, is one of the most popular methods for uncovering the temporal dependencies between time series. The application of Granger causality to detect inference among a large number of variables (such as genes) requires a variable selection procedure. To address the lack of informative data, so-called regularization procedures are applied. In this chapter, we review current literature on Granger causality with Lasso regularization techniques for ill-posed problems (i.e., problems with multiple solutions). We discuss regularization procedures for inverse and ill-posed problems and present our recent approaches. These approaches are evaluated in a case study on gene regulatory networks reconstruction.

Granger Causality for Ill-Posed Problems: Ideas, Methods, and Application in Life Sciences

Introduction

Causality describes the relation between a cause and its effect (its consequence). One can say that the *inverse problems*, where one would like to discover unobservable features of a cause from the observable features of an effect [24], can be seen as causality problems. When several elements or phenomena are considered and the causal relationships among them are questioned, we talk about the so-called causality network. A *causality network* can be seen as a directed graph with nodes, which correspond to the variables $\{x^j,\ j = 1, \ldots, p\}$ and directed edges, which represent the causal influences between variables. The variables represent entities or objects, for example, genes. We write $x^i \leftarrow x^j$ if the variable x^j has a causal influence on the variable x^i.

Causality Problems in Life Sciences

Causality networks arise in various scientific contexts. For example, in cell biology, one considers causality networks which involve sets of active genes of a cell. An active gene produces a protein. In biological experiments it has been observed that the amount of the protein, which is produced by a given gene, may depend on or may be *causally* influenced by the amount of proteins produced by other genes. In this way, causal relations between genes and corresponding causality network arise. These causality networks are also called *gene regulatory networks.* In cell biology, these networks are used in the research of causes of genetic diseases.

In neuroscience, causality networks are widely used to express the temporal interactions between various regions of the brain. Knowledge of these interactions can help to understand the human cognition or neurological diseases [54, 64, 51].

In practice, the first importan information that can be observed about a network is the temporal evolution (time series) of the involved variables $\{x_t^j,\ t = 1, \ldots, T\}$, where t is the index of time and j is the index of the concrete variable in the network. How can this information be used for inferring causal relations between variables?

The statistical approach to deriving causal relations between a target variable y and potential predictor variables $\{x^j,\ j = 1, \ldots, p\}$ using the known temporal evolution of their values $\{y_t, x_t^j,\ t = 1, \ldots, T,\ j = 1, \ldots, p\}$ consists of specifying a model of the relations between y and $\{x^j,\ j = 1, \ldots, p\}$. As a first step, one can consider a linear model for variable y_t:

$$y_t \approx \sum_{j=1}^{p} \beta^j x_t^j, \quad t = 1, \ldots, T.$$

The coefficients $\{\beta^j,\ j = 1, \ldots, p\}$, which can be estimated using the least-squares method, serve as indicators of causal relations. For instance, in statistics [82] by fixing the value of a threshold parameter $\beta_{\text{tr}} > 0$, one says that there is a causal relationship $y \leftarrow x^j$ if $|\beta^j| > \beta_{\text{tr}}$.

The goal of this chapter is to overview existing approaches for the reconstruction of the causal relations and to present novel techniques, originating from regularization theory, that allow for a more accurate and robust reconstruction of causality networks.

Outline of the Chapter

In the Section titled "Granger Causality and Multivariate Granger Causality", we continue our discussion of causality in general terms and introduce the notion of Granger Causality and Multivariate Granger Causality. We also discuss some methods for the reconstruction of causalities in gene regulatory networks. Consequently, we present the concept of gene regulatory networks and some recent approaches for their reconstruction. Since we consider causality problems as a special case of inverse problems, in the Section "Regularization of Ill-Posed Inverse Problems", we introduce the state of the art in the regularization theory for treating inverse ill-posed problems. We will mainly focus on approaches for treating problems with incomplete, high-dimensional, and noisy data, because of their high relevance to real-life applications. The Section "Multivariate Granger Causality" describes the state of the art for its analysis. Further, we discuss quality measures, which are used in numerical experiments for checking the performance of the methods. Finally, we discuss novel regularization techniques for the reconstruction of causal relationships and present

results of numerical experiments on gene regulatory network reconstruction using the classical approaches such as Lasso and our novel methods.

Notation

First, we introduce some standard notation that will be used in this paper. The entries of a matrix X are denoted by lower case letters and the corresponding indices, i.e., $X_{i,j} = x_{i,j}$. We define the Frobenius norm of a matrix X as

$$\|X\|_F := \left(\sum_{i,j} |x_{i,j}|^2 \right)^{\frac{1}{2}},$$

where $x_{i,j}$ is the entry (i, j) of the matrix X. It is also convenient to introduce the ℓ_p^n vector norms

$$\|x\|_{\ell_p^n} := \|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 0 < p < \infty,$$

and $\|x\|_0 := \#\{i : \ x_i \neq 0\}$ as usual. This notation will be used in Section "Multivariate Granger Causality Approaches using ℓ_1 and ℓ_2 Penalties". More specific notations will be defined in the paper, where they turn out to be useful.

Granger Causality and Multivariate Granger Causality

In 1969, the econometrist Clive Granger introduced a method to quantify temporal-causal relations among time series measurements [28], which gained great success across many scientific domains and in a variety of applications. In 2003, Granger was for his achievements awarded with the Nobel Prize in Economics. He introduced Wiener's concept of causality into the analysis of time series [80] and the notion of the "computationally measurable" causality. His method is usually referred to as *Granger Causality.* Granger causality is based on the statistical predictability of one time series using knowledge from one or more other time series. The basic idea of the method is straightforward.

Consider two simultaneously measured signals x and y, and examine two predictions of the values of y: the first one uses only the past values of the signal y, and the second one uses the past values of the both signals y and x. If the second prediction

is significantly better than the first one, then we call x to be *causal* to y [80]. Note that the contemporaneous effects are not considered. The standard test developed by Granger is based on linear regression models and leads to the two well-known alternative test statistics, the Granger-Sargent and the Granger-Wald (discussed in detail below) tests [2]. The probabilistic nature of Granger Causality leads to uncertainty concerning the relationship of cause and effect, which are fundamentally deterministic. The efficient applicability of the original Granger causality is impaired by several crucial problems of discovering latent confounding effect, missing counterfactual reasoning and capturing instantaneous and non-linear causal relationships [72], [56], [45]. Nevertheless, due to its simplicity and scalability, Granger causality remains a popular method for uncovering temporal dependencies and for detecting interactions between time series.

Rather than referring to Granger causality as a causal analysis tool, we will define it in our paper as a temporal dependence discovery method. Being aware of the above mentioned criticism, we will use the terms "G-causality", "G-causal" or "Granger Causality" in terms of temporal dependency or inference.

Granger Causality

Granger causality, GC thereafter, characterizes the extent to which a process x_t influences another process y_t, and builds upon the notion of incremental predictability. It is said that the *process* x_t *Granger causes another process* y_t if future values of y_t can be better predicted using the past values of x_t and y_t rather then only past values of y_t.

The standard test of Granger Causality is based on a linear regression model

$$y_t = a_0 + \sum_{l=1}^{L} b_{1l} y_{t-l} + \sum_{l=1}^{L} b_{2l} x_{t-l} + \xi_t \tag{1}$$

where ξ_t are uncorrelated random variables with zero mean and variance σ^2, L is the time lag, which denotes the maximum number of the considered past values of a variable, and $t = L+1, \ldots, N$. The null hypothesis that x_t does not Granger cause y_t is supported when $b_{2l} = 0$ for $l = 1, \ldots, L$, reducing (1) to

$$y_t = a_0 + \sum_{l=1}^{L} b_{1l} y_{t-l} + \xi_t'. \tag{2}$$

This model leads to the two well-known test statistics, the Granger- Sargent (GS) and the Granger-Wald (GW) test. The Granger-Sargent test is defined as

$$GS = N\frac{(R_2 - R_1)/L}{R_1/(N-2L)}, \tag{3}$$

where R_1 is the residual sum of squares in Equation (1) and R_2 is the residual sum of squares in Equation (2). The GS test statistic has an F-distribution with L and $N-2L$ degrees of freedom. The Granger-Wald test is defined as

$$GW = N\frac{\hat{\sigma}^2_{\xi'_t} - \hat{\sigma}^2_{\xi_t}}{\hat{\sigma}^2_{\xi_t}} \tag{4}$$

where $\hat{\sigma}^2_{\xi'_t}$ is the estimate of the variance of ξ'_t from model (2) and $\hat{\sigma}^2_{\xi_t}$ is the estimate of the variance of ξ_t from model (1). The GW statistic follows the χ^2_L distribution under the null hypothesis of no causality.

Multivariate Granger Causality

The bivariate Granger Causality can straightforwardly be extended to p-dimensional multivariate time series represented by $x_t \in R^{p\times 1}$.

Based on the intuition that the cause should precede its effect (ie. following Hume's definition of causality), in multivariate Granger causality one states that a (vector) variable x^i can be potentially G-caused by the past versions of the involved variables $\{x^j,\ j = 1,\dots,p\}$. Then, in the spirit of the statistical approach described above and using a (multivariate) vector auto-regressive model (VAR) for the *G-causal relations among p (scalar)* variables x^j_t, we consider the following approximation problem for the scalar values:

$$x^i_t \approx \sum_{j=1}^{p}\sum_{l=1}^{L}\beta^j_l x^j_{t-l}, \quad t = L+1,\dots,T, \tag{5}$$

where L is the *maximal time lag*. The approximation problem (5) can be specified using the least-squares approach:

$$\sum_{t=L+1}^{T}\left(x^i_t - \sum_{j=1}^{p}\sum_{l=1}^{L}\beta^j_l x^j_{t-l}\right)^2 \to \min_{\beta^j_l}.$$

Then, the coefficients $\{\beta^j_l\}$ can be determined from a system of linear equations. In the following sections, we denote the coefficient matrix by $A^{est} = (\beta^1, \beta^2, \dots, \beta^p)$, where the

coefficients are obtained by an approximation method. Performing a statistical significance test on the value of coefficients, one identifies the Granger causes of the target series. As in the statistical approach, one can now fix the value of the threshold parameter (i.e. of the substantive cut-off) $\beta_{\text{tr}} > 0$ and say that

$$x^j \text{ Granger causes } x^i, \text{ denoted by } x^i \leftarrow x^j \quad \text{if} \quad \sum_{l=1}^{L} |\beta_l^j| > \beta_{\text{tr}}. \tag{6}$$

It is well-known from the literature (see, e.g., [46]) that an application of Granger causality on gene regulatory networks with a large p may not lead to satisfactory results. This poor performance is reflected in the non-uniqueness of the solution of the the corresponding minimization problem and the potentially large number of reconstructed spurious relations. Actually, in practice one would expect to have only a few causal relations for a given gene, which means that the vector $\left(\beta_l^j\right)$ is sparse. In this case, the statistical significant tests are inefficient, while they lead to higher chance of spurious correlations. Moreover, the high dimensionality of biological data leads to further challenges. To address this issue, various *variable selection procedures* can be applied. Most of them are extensions of 'classical' variable selection procedures such as Lasso [75], LARS [23], and elastic nets [90].

Lasso (least absolute shrinkage and selection operator) is an alternative regularised version of least squares, which, in addition to the minimization of the residual sum of squares, imposes an ℓ_1 norm on the coefficients $\{\beta_l^j\}$. Due to the nature of ℓ_1 norm, Lasso shrinks the regression coefficients towards 0 and returns some coefficients which are exactly 0, implementing variable selection in this way. In the following, we will refer to *Lasso-Granger* as to an algorithm for learning the temporal dependency among multiple time-series based on variable selection using Lasso.

LARS (least angle regression) is a less greedy version of traditional forward selection method. A simple modification of the LARS algorithm is computationally less intensive compared to Lasso. The efficiency of the LARS algorithm makes it widely used in variable selection problems.

However, for highly correlated variables, Lasso tends to select only one variable instead of the whole group. To overcome this challenge, the elastic net method, which

combines ℓ_2 and ℓ_1 penalties on the coefficients was proposed. The convex function induced by the ℓ_2 penalty helps elastic net to achieve a grouping effect where strongly correlated predictors tend to be in or out of the model together. Elastic net often outperformed Lasso in terms of prediction error for correlated data.

In the following sections we continue our discussion on existing variable selection procedures with an emphasis on methods for discovering causal relations in gene regulatory networks.

Gene Regulatory Networks

Biomolecular interactions in a cell, called transcriptional regulation, show a complex non-linear dynamics. Models of transcriptional regulation are commonly depicted in form of networks, where directed connections between nodes represent regulatory interactions. The goal of these models is to infer on (or to reconstruct) the structure of gene regulatory networks from experimental data. Biological samples are usually profiled using the so-called gene expression microarrays, which correspond to the vector measurements and provide quantitative information to assess molecular control mechanisms. An experiment as sample, y, is a result of a single microarray experiment corresponding to a single column in the matrix of gene expressions, $y = (x_j^1, \ldots, x_j^n)'$ where n is the number of genes in the data set. A gene expression profile from microarrays has typically 5000 to 100000 variables (genes) and just $15 - 100$ measurements.

The detection of causality in a gene regulatory network from gene expression measurements, is a challenging problem, being solved by various computational methods with various success.

The most popular methods to model interactions in gene regulatory networks from experimental data are so-called *Dynamic Bayesian networks* (see, for example, [83]]. The application of ordinary differential equations is also popular in biological modeling (see, for example, [17] or [11] and [88]). These methods are reliable for modeling the local kinetics among a small number of genes; however, for larger gene

regulatory networks are these approaches computationally intensive.

Several other methods for modeling interactions among genes have been recently proposed and applied to gene expression data, such as *Structural Equation Models, Probabilistic Boolean Networks* and *Fuzzy Controls* (see, for example, [17, 73], just to mention a few). These methods are mainly applied to small genetic networks to study the dynamics of adjacent genes, and will not be discussed in this paper.

Taking into account the increasing interest of biologists in investigating interactions among large number of genes together with the scalability and simplicity of Granger Causality methods, we focus in this Chapter on these methods together with various ℓ_1 and ℓ_2 penalties.

Regularization of Ill-Posed Inverse Problems

The problem of the reconstruction of a gene regulatory network belongs to the class of inverse problems with high-dimensional dataset and sparse number of measurements. Recently, this problem attracted increasing attention from various scientific communities in inverse problems, machine learning, and approximation theory.

A general inverse problem (see, e.g., [24, 34, 62, 36, 48]) can be seen as an operator equation

$$\mathbf{y} = A\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{y}$ represents the data obtained in observational experiments, in other words the *effect*, $\boldsymbol{\beta}$ is the solution to be reconstructed, the *cause*, and the operator A represents the *model* between the cause and its effect. The approximation problem (5) can be seen as a problem of form (7).

In practice, one has take into account that the data $\mathbf{y}$ in (7) are *noisy*. Ideally, one assumes that there is a hidden cause $\boldsymbol{\beta}^\dagger$ with corresponding *ideal* data $\mathbf{y}^\dagger$ such that $\mathbf{y}^\dagger = A\boldsymbol{\beta}^\dagger$. The data $\mathbf{y}$ deviate from $\mathbf{y}^\dagger$, and the norm $\delta := \|\mathbf{y}^\dagger - \mathbf{y}\|$ is referred to as noise. Typically, the sources of the noise are imperfect measurements and modeling errors.

Inverse problems are often *ill-posed*, which means that equation (7) using the

noisy data $\mathbf{y}$, may have no solution, or the solution of (7) may be arbitrarily far away from the expected cause $\boldsymbol{\beta}^\dagger$. So-called *Regularization Methods* are proposed to deal with the ill-posedness of inverse problems.

A well-known class of regularization methods is the so-called *Tikhonov-type* regularization (see, e.g., [76, 77, 79]), where the solution of (7) is constructed as the minimizer $\boldsymbol{\beta}(\lambda)$ of the following functional:

$$\|\mathbf{y} - A\boldsymbol{\beta}\|^2 + \lambda\,\rho(\boldsymbol{\beta}) \to \min_{\boldsymbol{\beta}}. \tag{8}$$

In (8), λ is the so-called *regularization parameter*, and $\rho(\cdot)$ is a functional that is often similar to a norm functional. The methods (11), (12), (13) which are discussed below have form (8).

The appropriate choice of the regularization parameter λ is very important for the successful application of regularization methods. The goal is to choose λ such that the reconstruction error $\|\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}(\lambda)\|$ is minimal. This choice has to be made without knowledge of $\boldsymbol{\beta}^\dagger$. From a theoretical viewpoint [10], the choice of λ has to be coupled to the noise level δ and to the data $\mathbf{y}$.

In the theoretical analysis of choice rules, one tries to obtain an estimate for the reconstruction error $\|\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}(\lambda)\|$ that converges to zero as the noise level tends to zero. Also, the rate of convergence of $\|\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}(\lambda)\|$ is of interest, and one tries to design choice rules such that the convergence rate of the error is optimal over a class of solutions $\boldsymbol{\beta}^\dagger$. In this respect, the so-called *balancing principle* [52, 57, 42] is highly important.

Although knowledge of the noise level is important from the theoretical point of view, in practice it is either unknown, or it is challenging to estimate its value reliably. This is the case, for example, for inverse problem (5). In this case, one uses *heuristic choice rules*. In this paper, we use the so-called *quasi-optimality criterion*, which we present in Section "Granger Causality with Multi-Penalty Regularization". This choice rule has a close connection to the above-mentioned balancing principle, providing a certain reliability in its results.

In the context of causality detection, the concept of *consistency* of the reconstruction methods, which is discussed in the next section, seems to be similar to

the concept of the above mentioned *convergence* of regularization methods. However, clear links between these two concepts seem to be missing, and consideration of these links is an interesting subject for future research.

Multivariate Granger Causality Approaches using ℓ_1 and ℓ_2 Penalties

In [44] and [70] statistical properties of the Lasso-Granger methods were reviewed. Prior to these papers, Arnold et al. [4] and Fujita at al. [27] discussed the consistency of the Lasso-Granger algorithm and proved that the learned temporal dependencies will converge to the ground truth exponentially fast, if the time series data are generated from linear Gaussian models.

Inspired by [44], the common objectives in the analysis of Lasso-Granger methods are showing that the method is consistent in terms of three performance metrics:

1. *Prediction consistency* states that the estimation matrix A^{est} by Lasso-Granger can be used to accurately predict the future values of the time series. Formally, an estimation A_T^{est} obtained from a time series of length T is a consistent estimator if

$$\frac{1}{T}\sum_{t=1}^{T}\|\sum_{l=1}^{L}(A_T^{est,l}-a^{true}l)\mathbf{x}(t-l)\|_2^2\to 0,\ \text{for } T\to\infty, \tag{9}$$

where A^{true} is the true coefficient matrix.

2. *Parameter estimation consistency* states that the estimated coefficients should be close to the true coefficients:

$$\|A_T^{est,l}-A^{true,l}\|_F\to 0,\ T\to\infty,\ \text{for } l=1,\dots,L. \tag{10}$$

3. *Support recovery* states that the non-zero pattern of the estimate A_T^{est} matches the non-zero pattern of the true coefficient matrix with a high probability as $T\to\infty$.

The consistency of the Lasso Granger method in terms of the first two performance metrics under some additional assumptions on the matrix A^{true} has been shown in a recent paper [70]. We refer to [70] for further discussion on consistency of the method and the corresponding error estimates. Unfortunately, no consistency for support recovery and asymptotic normality are ensured for the Lasso-Granger method.

These results, however were derived for special modifications of Lasso, such as adaptive Lasso [70], [89].

The multivariate Granger causality methods that apply Lasso to the problem of reconstruction of gene regulatory networks were first proposed by Arnold et al. [4]. This method and its variations belong to *Graphical Lasso Granger (GLG)* methods. The model of the Graphical Lasso Granger method has the form

$$\sum_{t=L+1}^{T}\left(x_t^i-\sum_{j=1}^{p}\sum_{l=1}^{L}\beta_l^j x_{t-l}^j\right)^2+\lambda\|\boldsymbol{\beta}\|_1\to\min_{\beta_l^j}. \tag{11}$$

where $\lambda>0$ and L denotes the lag of the time series.

The solution of (11) for each variable $\{x^i,\ i=1,\dots,p\}$ with the causality rule (6) defines an estimator of the causality network between the variables $\{x^i\}$. Although method (11) enjoys great computational advantages and excellent performance, it is a well-known fact that the Lasso has a tendency to over-select the variables, i.e. reconstruct spurious causation.

In many situations, natural groupings exist between variables, and variables belonging to the same group should be either selected or eliminated as a group. Yuan and Lin [84] proposed an extension of Lasso, the so-called *Group Lasso*, to address this issue. This approach was used in Lozano et al. [46] to develop a novel methodology, termed *Graphical Group Lasso Granger* (GgrLG), which overcomes the limitations mentioned above for detection of causal relations. In particular, given J groups of variables which partition the set of predictors, the so-called *group Lasso estimate* $\hat{\boldsymbol{\beta}}_{group}(\lambda)$ [84] is defined as the minimizer of

$$\sum_{t=L+1}^{T}\left(x_t^i-\sum_{j=1}^{p}\sum_{l=1}^{L}\beta_l^j x_{t-l}^j\right)^2+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_{G_j}\|_2, \tag{12}$$

where $\boldsymbol{\beta}_{G_j}=\{\boldsymbol{\beta}_k : k\in G_j\}$, $\lambda>0$ is a regularization parameter. This form of the functional presupposes that the groups are of equal length, which is a quite natural assumption in this case since they correspond to the number of sampling points realized in regression. It is worthwhile to mention that the use of the ℓ_2-norm as a penalty norm enforces the coefficients $\boldsymbol{\beta}_{G_j}$ within a given group to be similar in amplitude (as

opposed to using the ℓ_1 norm). A limitation of the group Lasso is that it requires a priori information of group structures, which is often unavailable. Moreover, the procedures of minimizing (12) are nonlinear and require the solution of $O(pL)$ equations on each iteration step. This can be computationally intensive for large numbers of genes.

Extending upon these results, Zeng and Xie in [85] proposed two new methods to select variables in correlated data, the so-called gLars and gRidge. These methods conduct grouping and selecting at the same time and therefore work well when prior information of group structures is not available. Simulations and real examples show that the proposed methods often outperform the existing variable selection methods, including least angle regression (LARS) and elastic net, in terms of both reducing prediction error and preserving sparsity of representation. Another method based on group Lasso penalty with a linear autoregressive model was proposed and applied to gene regulatory networks by Kojima et al. [41].

Analysis of Granger causality between two groups of time series was also applied in brain functional connectivity analysis, where the functional connection between two regions of brain is investigated by analyzing multiple time series representing each region. Using the concept of canonical correlation [71], canonical Granger causality is proposed to be calculated between two time series representing the groups of times series, which are linear combinations of the time series in each group [6].

Rajapakse and Mundra in [60] experimentally tested the stability of multivariate vector autoregressive methods (MVAR) with ridge, lasso, and elastic net penalties by simulation on synthetic data and on gene expression data sets gathered over the Hela cell cycle. The stability of these MVAR methods with Lasso and with elastic net were comparable, and their accuracies were much higher than the MVAR with the ridge function.

Other methods to infer causal relationships, the so-called *Adaptive Thresholding Lasso Granger* (AtrLG) [67] and *Graphical Truncating Lasso Granger* (GtrLG) [65], were proposed by Shojaie and Michailidis and their consistency was proven in [66]. Let $\mathbf{x}$ be the $n \times p$ matrix of observations and let x_t denote the matrix corresponding to the

$t-th$ time point, and x_t^j be its $j-th$ column. The truncating lasso estimate of the graphical Granger model is found by solving the following estimation problem for $i = 1, \ldots, p$

$$\underset{\beta^t \in R^p}{\operatorname{argmin}} \frac{1}{n} \|x_T^i - \sum_{j=1}^{p} \sum_{l=1}^{L} x_{T-l}^j \beta_l^j\|_2^2 + \lambda \sum_{l=1}^{L} \psi^t \sum_{j=1}^{p} |\beta_l^j| w_j^l \tag{13}$$

$$\psi^1 = 1, \psi^l = M^{I\{\|a^{(l-1)}\|_0 < p^2\beta/(T-l)\}}, l \geq 2,$$

where M is a large constant, β is the allowed false negative rate, determined by the user, and $a^{l-1} = (\beta_{l-1}^1, \ldots, \beta_{l-1}^p)$ is a vector of coefficients, estimated at $(l-1)$. In practice is selected $M = g \exp n$ for g a large positive number. Selection of β can be based on the cost of false negatives in the specific problem at hand, as well as the sample size; as sample size increases, smaller values of β can be considered. A practical strategy for selecting β is to first find the lasso (or adaptive lasso) estimate and select β so that the desired false negative rate is achieved.

The truncating effect of the proposed penalty (imposed by ψ^l) is motivated by the rationale that the number of effects (edges) in the graphical model decreases as the time lag increases. Consequently, if there are fewer than $\frac{p^2\beta}{T-l}$edges in the $(l-1)$ estimate, all the later estimates are forced to zero. Hence, the Truncating Lasso penalty provides an estimate of the order of the underlying VAR model. In addition, by applying this penalty, the number of covariates in the model is reduced as the coefficients for effects of genes on to each other after the estimated time lags are forced to zero. Shojaie and Michailidis showed in [65] that the resulting estimate is consistent for variable selection (i.e. the correct edges are estimated with increasing probability, as the sample size increases) in the high dimensional sparse setting. With high probability, the signs of the effects are consistently estimated and the order of the underlying VAR model is correctly estimated.

Similar to GtrLG, AtrLG method attempts to simultaneously estimate the order of the VAR model and the structure of the network. While the truncating Lasso estimate is based on the assumption that the effects of genes on to each other decay over time, the adaptively thresholded Lasso estimator relies on a less stringent

structural assumption that sets a lower bound on the number of edges in the adjacency matrix of the graphical Granger model at each time point. The relaxation of the decay assumption allows the new estimator to correctly estimate the order of the time series in a broader class of models. The GtrLG may fail in situations where the decay assumption is violated. The method has two more drawbacks. First, the order of the VAR model d is often unknown and is, therefore, set to $T-1$, resulting in $p(T-1)$ covariates in the weighted Lasso estimation problem. Moreover, the weighted Lasso estimate may potentially include edges from different time points of variable x_j to any given variable x_i, $i \neq j$. We also refer the reader to the recent work of Shojaie, where reconstruction of gene regulatory networks by regularization techniques was addressed, for more detailed analysis of the above presented methods and their extensions [68].

As another application of a Lasso-Granger method, Bahadori and Liu in [8] used the copula approach and proposed a semi-parametric algorithm (*Granger Non-paranormal (G-NPN)*) for dependency analysis of time series with non-Gaussian marginal distributions, called *Copula Granger* method. Modelling of the dependency relations requires p time series $O(p^2)$ parameters, which can lead to high dimensionality and inconsistency of the non-parametric methods. The goal of the copula approach is to separate the marginal properties of the data from their dependency structure. The marginal distribution of the data can efficiently be estimated using non-parametric techniques with exponential convergence rate. The ℓ_1 regularization technique could be used to estimate the dependency structure in high dimensional settings.

The learning G-NPN model involves three steps:

(i) Find the empirical marginal distribution for each time series $\hat{F}_i$.

(ii) Map the observations into the copula space as $\hat{f}_i(x_t^i) = \hat{\mu}_i + \hat{\sigma}_i \Phi^{-1}(\hat{F}_i(x_t^i))$ where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the mean and standard deviation of the original time series. Φ^{-1} is the inverse cumulative distribution function of a standard normal.

(iii) Find the Granger Causality among $\hat{f}_i(x_t^i)$.

In practice, the Winsorized estimator of the distribution function is used, to avoid the large numbers $\Phi^{-1}(0^+)$ and $\Phi^{-1}(1^-)$, [8]:

$$\tilde{F}_j = \begin{cases} \delta_n, & \text{if } \hat{F}(x^j) < \delta_n \\ \hat{F}(x^j) & \text{if } \delta_n \leq \hat{F}(x^j) < 1 - \delta_n \\ (1 - \delta_n) & \text{if } \hat{F}(x^j) > 1 - \delta_n. \end{cases}$$

The Winsorized estimator is the transformation of statistics by limiting extreme values in the statistical data with the goal of reducing the effect of possibly spurious outliers, see for example [30]. Bahadori and Liu in [8] proved that the convergence rate for Copula Granger method is the same as the one for Lasso.

The Copula Granger method was tested with respect to the Granger method and the Lasso Granger method on synthetic and experimental data (Twitter applications) with the best precision for Copula Granger Lasso method [8]. We compared the Copula Granger method to the Lasso-Granger in [31] on the network of nineteen genes with better results for the Copula Granger method.

In our recent paper [58] we focused on an important tuning possibility of the Lasso, namely an appropriate choice of the *threshold parameter* β_{tr} and introduced the so called *Graphical Lasso Granger method (GLG) with two-level-thresholding*. This method is equipped with an appropriate *thresholding strategy* and an appropriate *regularization parameter choice rule*.

In [33], we compared our method to other Lasso Granger methods for gene regulatory network reconstruction, namely to the Lasso Granger method from Arnold et al., [4], Graphical Truncating Lasso from Shojaie and Michailidis, to the Copula Granger method from Bahadori and Liu [8], and to a method not using Lasso, i.e. a modification of a Bayesian network method from Äijö and Lähdesmäki [1]. As in [65] and [46], we used the gene expression data for the set of selected genes from the data basis of genes active in human cancer (HeLa), analyzed by Whitfield et al. in [81]. Our method was superior in this comparison. Details are discusses below in Section "Novel Regularization Techniques".

Despite the computational benefit and simplicity of the linear regression, model (5) could be too simple to appropriately match the underlying dynamics of the

phenomena and may sometimes lead to misspecifications. A more realistic situation would be to assume that the target function depends nonlinearly on relevant variables. This situation is much less studied and in the vast majority of the literature is restricted to the so called *additive model,* where the target function is assumed to be the sum

$$f(x) = \sum_{j=1}^{p} f_j(x_j) \tag{14}$$

of nonlinear univariate functions f_j in some Reproducing Kernel Hilbert Spaces (RKHS) $\mathcal{H}_j$ such that $f_j \equiv 0$ for $j \notin \{\nu_i\}_{i=1}^{l}$. For the sake of brevity, we omit the discussion on Reproducing Kernel Hilbert Spaces, and refer the reader to the seminal paper [5] on a comprehensive theory of RKHS.

Several authors, e.g., [7], [50], just to mention a few, observed that detection of relevant variables in the model (14) can be performed using multi-parameter regularization with special regularization terms: partial derivatives, different regularization spaces, etc. However, the application of the proposed multi-parameter methods on the real-life problems can be a non-trivial task due to several important reasons. First of all, the authors do not address the issue of selecting regularization parameters, which is a challenging and tedious task when there are more than two or three parameters. Second, the above mentioned approached can be computationally demanding and is, therefore, are not always suitable for problems with higher dimensions.

In the context of regularization theory, the multi-parameter regularization has been broadly studied as a mechanism to achieve the theoretically optimal rate of reconstruction without an a priori knowledge of relevant information on the solution. We refer the interested reader to recent papers on multi-parameter and multi-penalty regularization [47], [53], [26]. Taking our inspiration from these recent works and the above-mentioned findings in learning theory community, we propose in the Subsection "Granger Causality with Multi-Penalty Regularization" a novel multipenalty regularization approach for detecting relevant variables from a priori given high-dimensional data under the assumption that the input-output relation is described by a nonlinear function depending on few variables. Different than the above mentioned

work on detection of relevant variables, the method we propose is simple and fast to implement, i.e., there is no need for any sophisticated parameter choice rules.

Applied Quality Measures

Let a causality network among n elements be given by a directed graph with the nodes given by these elements. A *graphical method* is a method that reconstructs the causality network with the variables $\{x^j\}$ by means of a directed graph. Graphical metohds are frequently used in biology, see for example [38], [86].

Intuitively, the quality of a graphical method can be evaluated by the ability of the method to reconstruct the *known* causality network. It can be tested by various ways, for example by using the adjacency matrix. An *adjacency matrix* $A = \{a_{i,j} \mid \{i,j\} \subset \{1,\dots,p\}\}$ for the causality network has the following elements:

$$a_{i,j} = 1 \quad \text{if} \quad x^i \leftarrow x^j; \quad a_{i,j} = 0 \quad \text{otherwise.}$$

Assume that there is a *true* adjacency matrix A^{true} of the true causality network, and its *estimator* A^{est}, which is produced by a graphical method. The elements of the adjacency matrix A^{est} can be classified as follows.

- If $a_{i,j}^{est} = 1$ and $a_{i,j}^{true} = 1$, then $a_{i,j}^{est}$ is called true positive. The number of all true positives of matrix A^{est} will be denoted as TP.
- If $a_{i,j}^{est} = 0$ and $a_{i,j}^{true} = 0$, then $a_{i,j}^{est}$ is called true negative. The number of all true negatives of matrix A^{est} will be denoted as TN.
- If $a_{i,j}^{est} = 1$ and $a_{i,j}^{true} = 0$, then $a_{i,j}^{est}$ is called false positive. The number of all false positives of matrix A^{est} will be denoted as FP.
- If $a_{i,j}^{est} = 0$ and $a_{i,j}^{true} = 1$, then $a_{i,j}^{est}$ is called false negative. The number of all false negative of matrix A^{est} will be denoted as FN.

The following quality measures of the estimator A^{est} will be considered:

- *Precision* (also called positive predictive value) of A^{est}:

$$P = \frac{TP}{TP + FP}, \ 0 \leq P \leq 1. \tag{15}$$

- *Recall* (also called sensitivity) of A^{est}:

$$R = \frac{TP}{TP + FN},\ 0 \leq R \leq 1. \tag{16}$$

Since it is possible to have a high precision and low recall, and vice versa, one considers also an average between these two measures.

The so called F_1-*score* is defined as the harmonic mean of precision and recall:

$$\frac{1}{F_1} = \frac{1/P\ +\ 1/R}{2}. \tag{17}$$

The computational complexity of Lasso Granger methods (i.e. including the above mentioned one) is $O(nd^2p^2)$, where n is the number of observations (i.e. the length of the time series), p is number of genes, and d is the order of the corresponding VAR model. The computational complexity of Graphical Truncated Lasso is $O(n\hat{d}^2p^2)$, where $\hat{d}$ is the estimate of the order d of VAR model (i.e. the effective number of time lags in VAR, noted L elsewhere) from the truncated Lasso penalty [65].

Novel Regularization Techniques with a Case Study of Gene Regulatory Networks Reconstruction

Optimal Graphical Lasso Granger Estimator

Assume that the true causality network with the variables $\{x^j\}$ is given by the adjacency matrix A^{true}. Assume further that the observation data $\{x_t^j\}$ are given. The *best* reconstruction of A^{true} that can be achieved by the so-called *optimal* Graphical Lasso Granger estimator and we proposed in [58]. For brevity, the abbreviation GLG method will be used for a Graphical Lasso Granger method.

Let $\boldsymbol{\beta}_i(\lambda)$ denote the solution of the minimization problem (11) in the GLG-method, and $\boldsymbol{\beta}_i^j(\lambda) = (\beta_{1,i}^j, \ldots, \beta_{L,i}^j)$. Then, the Graphical Lasso Granger estimator $A^{\text{GLG}}(\lambda, \beta_{\text{tr}})$ of the adjacency matrix A^{true} is defined as follows:

$$A_{i,j}^{\text{GLG}}(\lambda, \beta_{\text{tr}}) = 1 \ \text{ if } \quad \|\boldsymbol{\beta}_i^j(\lambda)\|_1 > \beta_{\text{tr}};$$

$$A_{i,j}^{\text{GLG}}(\lambda, \beta_{\text{tr}}) = 0 \quad \text{otherwise.}$$

Let $A^{\mathrm{GLG}}_{i,*}(\lambda, \beta_{\mathrm{tr}})$ denote the i-th row of the Graphical Lasso Granger estimator. For the given regularization parameter λ, let $\beta^i_{\mathrm{tr}}(\lambda)$ be the threshold parameter that minimizes the number of false entries in the row $A^{\mathrm{GLG}}_{i,*}(\lambda, \beta_{\mathrm{tr}})$, i.e. the threshold parameter that solves the following minimization problem:

$$\|A^{\mathrm{true}}_{i,*} - A^{\mathrm{GLG}}_{i,*}(\lambda, \beta_{\mathrm{tr}})\|_1 \to \min_{\beta_{\mathrm{tr}}}. \tag{18}$$

Then, we consider the minimization of the number of false entries with respect to the regularization parameter λ, i.e. let $\lambda_{\mathrm{opt},i}$ solve

$$\|A^{\mathrm{true}}_{i,*} - A^{\mathrm{GLG}}_{i,*}(\lambda, \beta^i_{\mathrm{tr}}(\lambda))\|_1 \to \min_{\lambda}. \tag{19}$$

In this way, we obtain, what we call, the *optimal* Graphical Lasso Granger estimator $A^{\mathrm{GLG,opt}}$ of the true adjacency matrix A^{true}:

$$A^{\mathrm{GLG,opt}}_{i,j} = A^{\mathrm{GLG}}_{i,j}(\ \lambda_{\mathrm{opt},i}, \beta^i_{\mathrm{tr}}(\lambda_{\mathrm{opt},i})\ ).$$

Note that the optimal Graphical Lasso Granger estimator minimizes the following quality measure, which we call Fs-measure:

$$Fs = \frac{1}{p^2}\|A^{\mathrm{true}} - A^{\mathrm{est}}\|_1,\ 0 \le Fs \le 1. \tag{20}$$

Fs-measure represents the number of *false* elements in the estimator A^{est} that is scaled with the total number of elements in A^{est}.

In practice, the minimization problems (18) and (19) can be approximated by the corresponding minimization problems over finite sets of parameters β_{tr}, λ. If we consider a set with N_{tr} values for β_{tr}, and a set with N_λ values for λ, then, in order to determine $A^{\mathrm{GLG,opt}}$, one needs to use $N_{\mathrm{tr}} \cdot N_\lambda$ Lasso Granger solvers. The computational complexity of one Lasso Granger solver was discussed in the previous section.

In the networks created by nature, the true causal relations among selected genes are often unknown. One can use the detected relations from available genetic databases, for example from frequently updated gene and protein interactions data base Biogrid [15]. The Biogrid tool "Genemania" is a graphical databasis of detected interactions among genes by experimenting in genetic laboratories all over the world. The biological

experiments are expensive, and, therefore, the knowledge of a reliable computational method is of high importance.

To approach the problem of how close one can get to $A^{\mathrm{GLG,opt}}$ without the knowledge of A^{true}, let us first focus on the choice of the threshold parameter β_{tr}.

Thresholding strategy

The purpose of the threshold parameter β_{tr} is to differentiate the relations $x^i \leftarrow x^j$ with *small* values of $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$ as the non-causal ones. When can we say that $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$ is small? We propose considering the following *guiding indicators* of smallness:

$$\begin{aligned} \beta^i_{\min}(\lambda) &= \min\{\ \|\boldsymbol{\beta}_i^j(\lambda)\|_1,\ j=1,\dots,p \mid \|\boldsymbol{\beta}_i^j(\lambda)\|_1 \neq 0\ \}, \\ \beta^i_{\max}(\lambda) &= \max\{\ \|\boldsymbol{\beta}_i^j(\lambda)\|_1,\ j=1,\dots,p\ \}. \end{aligned} \tag{21}$$

In particular, we propose considering the threshold parameter of the following form:

$$\beta^i_{\mathrm{tr},\alpha}(\lambda) = \beta^i_{\min}(\lambda) + \alpha(\ \beta^i_{\max}(\lambda) - \beta^i_{\min}(\lambda)\ ). \tag{22}$$

It should be noted that $\beta^i_{\min}(\lambda)$ and $\beta^i_{\max}(\lambda)$ determine the interval of possible values for β_{tr}, namely $\beta_{\mathrm{tr}} \in [\beta^i_{\min}(\lambda) - \varepsilon_1,\ \beta^i_{\max}(\lambda)]$, where $\varepsilon_1 > 0$ is a small constant. Thus, with $\alpha \in [-\varepsilon_2, 1]$, where $\varepsilon_2 > 0$ is another small constant, $\beta^i_{\mathrm{tr},\alpha}$ covers the entire range of possible values for β_{tr}. The choice $\alpha = 1/2$ is the default. Also, it is worth noting that the choice of the threshold (22) is independent of the scaling of the data.

The optimal GLG-estimator with the threshold parameter $\beta^i_{\mathrm{tr},1/2}$ can be defined as follows. Let $\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i}$ solve the minimization problem:

$$\|A^{\mathrm{true}}_{i,*} - A^{\mathrm{GLG}}_{i,*}(\lambda, \beta^i_{\mathrm{tr},1/2}(\lambda))\|_1 \to \min_{\lambda}.$$

Then, the corresponding optimal GLG-estimator is

$$A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}(i,j) = A^{\mathrm{GLG}}_{i,j}(\ \lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i}, \beta^i_{\mathrm{tr},1/2}(\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i})\ ).$$

The choice of the threshold parameter $\beta^i_{\mathrm{tr},1/2}$ rises the following issue. A gene receives always a causal relations, unless the solution of (11) $\boldsymbol{\beta}_i(\lambda)$ is zero. But how strong are these causal relationships compared to each other? The norm $\|\boldsymbol{\beta}_i^j(\lambda)\|_1$ can be seen as an *indicator of the strength* of the causal relationship $x^i \leftarrow x^j$.

Let us now construct a matrix $A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}$, similar to the adjacency matrix $A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}$, in which the norm $\|\boldsymbol{\beta}^j_i(\lambda)\|_1$ is used instead of the value 1. i.e.

$$A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) = \|\boldsymbol{\beta}^j_i(\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i})\|_1 \quad \text{if} \quad \|\boldsymbol{\beta}^j_i(\lambda^{\mathrm{tr},1/2}_{\mathrm{opt},i})\|_1 > \beta^i_{\mathrm{tr},1/2},$$
$$A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) = 0 \quad \text{otherwise.}$$

The false causal relations of the estimator $A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2}$ showed up in the experiments on a gene regulatory network in [58] to be actually weak. This observation suggested to use a second thresholding that is done on the network, at the level of the adjacency matrix.

Thresholding on the network level is similar to thresholding on the gene level. Specifically, let us define the guide indicators of smallness on the network level in a way similar (21):

$$A_{\min} = \min_{i,j=1,\dots,p} \{\, A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) \neq 0 \,\},$$
$$A_{\max} = \max_{i,j=1,\dots,p} \{\, A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) \,\}.$$

And, similar (22), define the threshold on the network level as follows:

$$A_{\mathrm{tr},\alpha} = A_{\min} + \alpha(\, A_{\max} - A_{\min} \,). \tag{23}$$

We propose terming the described combination of two thresholdings on the gene and network levels *two-level-thresholding*. The adjacency matrix obtained by this thresholding strategy is the following:

$$A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2;\alpha_1}(i,j) = 1 \quad \text{if} \quad A^{\mathrm{GLG,opt};\boldsymbol{\beta}}_{\mathrm{tr},1/2}(i,j) > A_{\mathrm{tr},\alpha},$$
$$A^{\mathrm{GLG,opt}}_{\mathrm{tr},1/2;\alpha_1}(i,j) = 0 \quad \text{otherwise.}$$

It turned out that with $\alpha = 1/4$, in (23) the optimal GLG-estimator for the gene regulatory network in [58] can be fully recovered.

An automatic realization of the GLG-method

For an automatic realization of the GLG-method, i.e. when the true adjacency matrix A^{true} is not known, one needs in addition to a thresholding strategy, a choice

rule for the regularization parameter λ in (11). For such a choice, we proposed in [58] using the so called *quasi-optimality criterion* [78, 14, 40]. In this criterion, one considers a set of regularization parameters

$$\lambda_k = \lambda_0 q^k,\ q < 1,\ k = 0, 1, \ldots, n_\lambda, \tag{24}$$

and for each λ_k the corresponding solution of (11) $\boldsymbol{\beta}^i(\lambda_k)$ is computed. Then, the index of the regularization parameter is selected as follows:

$$k^i_{\text{qo}} = \underset{k}{\text{argmin}}\{\ \|\boldsymbol{\beta}_i(\lambda_{k+1}) - \boldsymbol{\beta}_i(\lambda_k)\|_1\ \}. \tag{25}$$

Let us note that the motivation for the choice of the set of possible regularization parameters as (24), and for the choice of the regularization parameter as (25) is discussed in [57].

In the exprerimental part of this paper we compare the GLG-method with an appropriate thresholding to other discussed methods on the network of nineteen genes given by gene expressions from biological experiments of Whitfield et al [81].

Granger Causality with Multi-Penalty Regularization

The natural groupings between the values x^t_j of variables x_j can be introduced into multivariate regression by considering, instead of (5) the following form

$$x^t_\nu \approx \sum_{j=1}^{p} f_j\left(\sum_{l=1}^{L} \beta^l_j x^{t-l}_j\right), \quad t = L+1, L+2, \ldots, T, \tag{26}$$

where f_j are univariate functions in some Reproducing Kernel Hilbert Spaces $\mathcal{H}_j$. Then, a conclusion that gene x_k causes the gene x_ν can be drawn by determining that variable x_k is a relevant variable of a function of the form (14).

In this section, we present a novel method for variable selection in (14) using the multi-penalty regularization. To our best knowledge, this is the first work in the field, which describes an application of multi-penalty regularization for inferring causal relations in gene regulatory networks.

An estimator of the target function (14) can be constructed as the sum $\sum\limits_{j=1}^{p} f^\lambda_j(x_j)$ of the minimizers of the functional

$$T_\lambda(f_1, f_2, \ldots, f_p;\ Z_N) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - \sum_{j=1}^{p} f_j(x^i_j)\right)^2 + \sum_{j=1}^{p}\lambda_j \|f_j\|^2_{\mathcal{H}_j}, \tag{27}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a vector of the regularization parameters, and $Z_N = \{ (x_1^i, x_2^i, \ldots, x_p^i, y^i) \}_{i=1}^N$ denotes a dataset of observed values y^i, $i = 1, 2, \ldots, N$, of a variable of interest y paired with simultaneously observed values x_ν^i, $\nu = 1, 2, \ldots, p$, of the variables $x_1, x_2, \ldots, x_p$ that possibly interact with y.

On first sight, it may be seen that the results of the minimization of the functional (27) do not systematic lead to sparsity as in the previously addressed approaches. The sparse structure can be reconstructed following the next three steps.

The first step consists of constructing the minimizers $f_j = f_j^{\lambda_j}(x_j)$ of the functionals $T_{\lambda_j}(f_j; Z_N)$ defined by (27) with $p = 1$, $\lambda_1 = \lambda_j$, $x_1^i = x_j^i$, $\mathcal{H}_1 = \mathcal{H}_j$, $j = 1, 2, \ldots$. Using classical results from approximation theory [39, 69], such minimization is reduced to solving systems of N linear equations. Then the minimizers $f_j^{\lambda_j}(x_j)$ are used to rank the variables x_j according to the values of the discrepancies

$$\mathcal{D}(f_j^{\lambda_j}(x_j);\ Z_N) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_j^{\lambda_j}(x_j^i) \right)^2 \right)^{1/2},\ j = 1, 2, \ldots,$$

as follows: the smaller the value of $\mathcal{D}(f_j^{\lambda_j}(x_j); Z_N)$, the higher the rank of x_j. This step can be seen as an attempt to interpret the data Z_N by using only a univariate function, and the variable with the highest rank is considered as the first relevant variable x_{ν_1}.

The next step consists of testing the hypothesis that a variable with the second highest rank, say x_μ, is also relevant. For such a test, we compute the minimizers $f_{\nu_1}^{\lambda_{\nu_1}}$, $f_\mu^{\lambda_\mu}$ of the functional

$$T_\lambda(f_{\nu_1}, f_\mu;\ Z_N) = \frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_{\nu_1}(x_{\nu_1}^i) - f_\mu(x_\mu^i) \right)^2 + \lambda_{\nu_1} \|f_{\nu_1}\|_{\mathcal{H}_{\nu_1}}^2 + \lambda_\mu \|f_\mu\|_{\mathcal{H}_\mu}^2. \tag{28}$$

Our idea is based on the observation [53] that in multi-penalty regularization with a component-wise penalization, such as (28), one requires small as well as large values of the regularization parameters $\lambda_{\nu_1}, \lambda_\mu$, i.e., both λ_{ν_1} and $\lambda_\mu << 1$, and $\lambda_\mu > 1$ respectively. Therefore, in the proposed approach, variable x_μ is considered relevant if for $\{\lambda_{\nu_1}, \lambda_\mu\} \subset (0, 1)$, the values of the discrepancy

$$\mathcal{D}(f_{\nu_1}^{\lambda_{\nu_1}}, f_\mu^{\lambda_\mu};\ Z_N) = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_{\nu_1}^{\lambda_{\nu_1}}(x_{\nu_1}^i) - f_\mu^{\lambda_\mu}(x_\mu^i) \right)^2 \right)^{1/2} \tag{29}$$

are essentially smaller than the ones for $\lambda_{\nu_1} \in (0,1)$, $\lambda_\mu > 1$. If it is not the case, the above hypothesis is rejected, and we test in the same way the variable with the third highest rank and so on.

When the variable x_μ was accepted as the second relevant variable, i.e., $x_{\nu_2} = x_\mu$, we proceed with testing whether the variable with the third highest rank, say x_ν, can be taken as the third relevant variable, i.e., whether $x_{\nu_3} = x_\nu$. Thus, we compute the minimizers $f_{\nu_1}^{\lambda_{\nu_1}}$, $f_{\nu_2}^{\lambda_{\nu_2}}$, $f_\nu^{\lambda_\nu}$ of the functional

$$\begin{aligned} T_\lambda(f_{\nu_1}, f_{\nu_2}, f_\nu;\ Z_N) =& \frac{1}{N}\sum_{i=1}^{N}\Big(y^i - f_{\nu_1}(x^i_{\nu_1}) - f_{\nu_2}(x^i_{\nu_2}) - f_\nu(x^i_\nu)\Big)^2 + \\ & \lambda_{\nu_1}\|f_{\nu_1}\|^2_{\mathcal{H}_{\nu_1}} + \lambda_{\nu_2}\|f_{\nu_2}\|^2_{\mathcal{H}_{\nu_2}} + \lambda_\nu\|f_\nu\|^2_{\mathcal{H}_\nu}, \end{aligned} \tag{30}$$

where, with a little abuse of notation, we use the same symbols f_{ν_1}, $f_{\nu_1}^{\lambda_{\nu_1}}$ as in (28),(29). Then, as above, variable x_ν is considered relevant if for $\{\lambda_{\nu_1}, \lambda_{\nu_2}, \lambda_\nu\} \subset (0,1)$, the values of the discrepancy

$$\mathcal{D}(f_{\nu_1}^{\lambda_{\nu_1}}, f_{\nu_2}^{\lambda_{\nu_2}}, f_\nu^{\lambda_\nu};\ Z_N) = \left(\frac{1}{N}\sum_{i=1}^{N}\Big(y^i - f_{\nu_1}^{\lambda_{\nu_1}}(x^i_{\nu_1}) - f_{\nu_2}^{\lambda_{\nu_2}}(x^i_{\nu_2}) - f_\nu^{\lambda_\nu}(x^i_\nu)\Big)^2\right)^{1/2} \tag{31}$$

are essentially smaller than the corresponding values of (31) for $\{\lambda_{\nu_1}, \lambda_{\nu_2}\} \subset (0,1)$, $\lambda_\nu > 1$. (In our experiments, the small parameter was chosen from the interval $[0.00001, 0.3]$ and the large one from $[1, B]$ where B is some large constant, say $B = 10$.)

Otherwise, the variable with the next highest rank is tested in the same way.

If the discrepancy (31) exhibits the above mentioned property, then for testing the variable with the next highest rank in accordance with the proposed approach, we need to add to (30) one more penalty term corresponding to that variable, so that the functional $T_\lambda(f_1, f_2, \ldots, f_p; Z_N)$ of the form (27) containing the whole set of penalties may appear only at the end of the testing procedure.

For the sake of brevity, we omit the theoretical justification of the presented approach here and refer the interested reader to our recent paper [32] for a detailed mathematical description and theoretical justification. However, we note that the theoretical results do not require any strict assumptions neither on the distribution of the data points nor on the number of them.

It is important to mention that the choice of the regularization parameter(s) is not a tedious and tricky task for the proposed method, since we are not interested in the exact reconstruction of the given value y^i but in values of the discrepancies for small and large values of the regularization parameters. Monte-Carlo-type simulations are used to make such comparison. Namely, if $x_{\nu_1}, x_{\nu_2}, \ldots, x_{\nu_{l-1}}$ have been already accepted as relevant variables, then the values of $\mathcal{D}$ for the randomly chosen $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^l$ are compared to the ones for the randomly chosen $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^{l-1} \times [1, B]$, $B > 1$, and x_{ν_l} is accepted as relevant if these values are essentially dominated by the ones for $(\lambda_{\nu_1}, \lambda_{\nu_2}, \ldots, \lambda_{\nu_l}) \in (0,1)^{l-1} \times [1, B]$.

The computational complexity of multi-penalty regularization is $O(Np^2)$, where N is the number of given points and p is the number of variables.

Case Study of Gene Regulatory Network Reconstruction

We used the databasis of gene expression data from the biological experiments of Whitfield et al. [81], as in our papers [31] and [33]. We selected nineteen genes which are active in human cancer cell line, whose gene regulatory network was reconstructed based on the biological experiments by Li et al. [43]. The causal structure for these genes was adopted from [46] and is presented in Figure 1. We take this causal network as a benchmark network for a comparison of the discussed methods. The nineteen genes, which we consider, play a substantial role at the human cancer cell lines. They have the following names: PCNA, NPAT, E2F1, CCNE1, CDC25A, CDKN1A, BRCA1, CCNF, CCNA2, CDC20, STK15, BUB1B, CKS2, CDC25C, PLK1, CCNB1, CDC25B, TYMS, DHFR. The gene expressions in the database from [81] for these genes were given for 48 observations with one hour intervals.

The data values are illustrated is in Figure 2. The horizontal coordinate x indicates the 48 time measuremens, the vertical coordinate y indicates the order of the 19 genes; The color of the pixel corresponds to the value determined by the color-scale on the right hand side.

We used the following MATLAB codes: our code for GLG-method with an

appropriate thresholding which we extended with graphical outputs using MATLAB graphical software Graphviz4MATLAB Version 2.24. For experiments with Lasso Granger method we used the MATLAB code from Bahadori [9] written for the bivariate case, which we extended to the multivariate case. We extended this code also with the graphical outputs using Graphviz4MATLAB. Similarly, we extended the MATLAB code for Copula Granger method from [9]. These methods were compared to the method using dynamic Bayesian networks and ordinary differential equations from [1] in [33]. The latter method showed frequent overfitting with respect to the number of false positives and had high computationally costs. Here we compare the performance of the Lasso Granger methods with the Granger method with multipenalty regularization. The code for the multipenalty regularization method has been developed by us in MATLAB.

As quality (performance) measures we considered the number of true positive outcomes denoted by TP and the classification accuracy
$CA = (TP + TN)/(TP + TN + FP + FN)$.

The Lasso Granger method was tested in 4 variations:

- Lasso Granger with zero threshold ($\beta_{\text{tr}} = 0$ in (6)) and optimized regularization parameter λ in (11). We refer to this variation as LG.
- Lasso Granger with optimized regularization parameter and threshold, which is referred to as LG1.
- Lasso Granger with optimized regularization parameter and threshold given by formula (22) with $\alpha = 1/4$, LG2.
- And finally, Lasso Granger with regularization parameter chosen by quasi-optimality criterion and threshold given by formula (22) with $\alpha = 1/4$, LG3. This is an automatic realization of the Lasso Granger method without the knowledge of the true adjacency matrix.

We call the Granger Causality method with Multi-Penalty Regularization MPR. Let us note that in MPR, the coefficients $\left(\beta_j^l\right)$ in (26) have to be precomputed. For this purpose, one can use any (regularization) method for solving the approximation problem (5). In this case, of course, the results of MPR depend on the choice of this

method. In [32], we used the l_2-regularized least squares method for obtaining the coefficients $\left(\beta_j^l\right)$. Here, we used the Lasso, which is the l_1-regularized least squares method. The regularization parameter in both regularization methods was chosen by the quasi-optimality criterion.

The Copula Granger method (CG), all mentioned variations of the Lasso Granger method, and the Granger Causality with Multi-Penalty Regularization required only a few seconds run at a PC workstation with 64-bit processor. CA and TP quality measures of the considered methods are summarized in Table 1. In Figure 3, we present the considered gene regulatory network and its reconstructions with LG3 and MPR in the circular layout.

One observes that although the CG gives the largest number of TPs among the automatic realizations of graphical methods, i.e. CG, LG3, MPR, it gives the lowest CA, while MPR gives the highest CA together with rather high TP. This makes MPR a very promising method for the reconstruction of gene regulatory networks.

Conclusion

The results of the reconstruction of the gene regulatory network in the experimental section emphasize the importance of the thresholding strategies for the variable selection regularization methods, such as the Lasso. The newly developed MPR technique [32] can be seen as an advanced thresholding, and our experimental results show its superior behavior with respect to our method with thresholding strategy.

As we noted before, the MPR requires a method that computes the coefficients $\left(\beta_j^l\right)$ in (26). Currently, we tested the behavior of the MPR with l_2- regularization in [32], and here with l_1-regularization, which is the Lasso. In our tests, the MPR gave superior results. Also, the coupling between these Lasso-modifications and MPR is interesting to realize. It would be also interesting to see the reconstruction of the gene regulatory network in Figure 1 by means of the discussed modifications of Lasso: truncating Lasso and adaptively thresholded Lasso.

The methods proposed in this paper are written in Matlab and are available upon

request.

Acknowledgements

The first author gratefully acknowledges the partial support by the research grant GACR 13-13502S of the Grant Agency of the Czech Republic (Czech Science Foundation).

References

[1] Äijö, T., Lahdesmäki, H.: Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics, no 22, Vol. 25 (2009)

[2] Abramowitz, M., Stegun, I.A..: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing, Dover, New York (1972)

[3] Ancona, N., Marinazzo, D., Stramaglia, S.: Radial basis function approach to nonlinear Granger causality of time series, Physical Review E 70 (2004)

[4] 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 (2007)

[5] Aronszajn, N.: Theory of reproducing kernels, Trans. Am. Math. Soc. vol 68, pp. 337–404 (1950)

[6] Ashrafulla, S., Haldar, J.P. , Joshi, A.A., and Leahy, R.M. Canonical Granger causality applied to functional brain data. In ISBI, 2012.

[7] Bach, F.: Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning, Advances in Neural Information Processing Systems 21, edited by D Koller and et al, pp. 105–112 (2009)

[8] Bahadori, T., Y. Liu, Y.: An examination of large-scale Granger causality inference. SIAM Conference on Data Mining (2013)

[9] Website, http// www-scf.usc.edu/∼mohammab/codes/

[10] Bakushinskii, A.B.: Remarks on choosing regularization parameter using the quasi-optimality and ratio criterion. USSR Comp. Math. Math. Phys., 24:181âĂŞ182 (1984)

[11] Bansal M., Belcastro, V., Ambesi-Impiombato, A, di Bernardo, D.: How to infer gene networks from expression profiles. Mol. Syst. Biol., 3, 78 (2007)

[12] Bansal, M., Della Gatta, G., di Bernardo, D.: Inference of gene regulatory networks and compound mode of action from time course gene xxpression profiles. Bioinformatics, 22, 815822 (2006)

[13] Barenco, M. et al.: Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biol., 7, R25 (2006)

[14] Bauer, F, Reiß, M, Regularization independent of the noise level: An analysis of quasi-optimality, Inverse Probl., Vol 24, 5 (2008)

[15] Biological General Repository for Interaction Datasets, Biogrid 3.2

[16] Broomhead, D.S., Lowe, D.: Multivariate functional interpolation and adaptive networks, Complex Systems 2, 321 - 355 (1988)

[17] Cao, J. and Zhao, H.: Estimating dynamic models for gene regulation networks. Bioinformatics, 24, 1619-1624 (2008)

[18] Caraiani, P.: Using complex networks to characterize international business cycles, PLoS ONE, vol 8, number 3, pp. 58109 (2013)

[19] Chen, Y., Rangarajan, G., Feng, J., Ding, M.: Analyzing mulitiple non- linear time series with extended Granger causality. Phys. Lett. A 324, 26-35 (2004)

[20] Cooper, G. F.: The computational complexity of probabilistic inference using Bayesian belief networks, Artificial Intelligence, 42: 393 - 405 (1990)

[21] Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Commun. Pure Appl. Math., Vol 57, No 11, pp. 1413–1457 (2004)

[22] Ebert-Uphoff, I., Deng, Y.: Causal discovery for climate research using graphical models, J. Clim., vol. 25, pp. 5648–5665 (2012)

[23] Efron, B, Hastie, T.; Johnstone, I.; Tibshirani, R.: Least angle regression, Ann. Statist. 32, 407-499 (2004)

[24] Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer Academic Publishers, "Dordrecht (1996)

[25] Fornasier, M.: Theoretical Foundations and Numerical Methods for Sparse Recovery, Publisher Berlin: de Gruyter (2010)

[26] Fornasier, M., Naumova, V., Pereverzyev, S.V.: Parameter choice strategies for multi-penalty regularization, SINUM 52, 1770-1794 (2014)

[27] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Ferreira, C.E. Modeling gene expression regulatory networks with the sparse vector autoregressive model, BMC Systems Biology, 1:37 (2007)

[28] Granger, C.W.J.: Investigating causal relations by econometric and cross- spectral methods, Econometrica 37, 424-438 (1969)

[29] Grasmair, M., Haltmeier, M., Scherzer, O.: Sparse regularization with l^q penalty term, Journal of Inverse Problems, Vol 24, No 5, pp 13 (2008)

[30] Hasings, C., Mosteller, F., Tukey, J.W., Winsor, C.P.: Low moments for small samples: a comparative study of order statistics, Annals of Mathematical Statistics, 18, 413âĂŞ426 (1947)

[31] Hlaváčková-Schindler, K., Bouzari, H.: Granger Lasso causal models in high dimensions: Application to gene expression regulatory networks, The Proceedings of EVML/PKDD 2013, SCALE, Prague (2013)

[32] Hlaváčková-Schindler, K., Naumova, V., Pereverzyev, S., Jr.: Multi-penalty regularization for detecting relevant variables. Pre-print Nr 11, Leopold Franzens Universität Innsbruck (2014)

[33] Hlaváčková-Schindler, K., Pereverzyev, S., Jr. Lasso Granger Causal Models: Some Strategies and Their Efficiency for Gene Expression Regulatory Networks, Lecture Notes in Artificial Intelligence, Springer, in print, 2014.

[34] Hofmann, B.: Mathematik Inverser Probleme, Stuttgart, Teubner (1999)

[35] Jensen, F.V.: An Introduction to Bayesian Networks, London, UCL Press (1996)

[36] Kabanikhin, S.I.: Definitions and examples of inverse and ill-posed problems, J. Inv. Ill-Posed Problems, 16:317âĂŞ357 (2008)

[37] Kaminski, M., Ding, M.: Truccolo, W.A., Bressler, S.L.: Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biol Cybern, 85, 14557 (2001)

[38] Khanin, R., Wit, E. Construction of malaria gene expression network using partial correlations. Methods of Microarray Data Analysis V, (2007)

[39] Kimeldorf, G. S., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. The Annals of Mathematical Statistics vol 41, pp. 495–502 (1970)

[40] Kindermann, S., Neubauer, A.: On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization, Inverse Probl. Imaging, Vol 2, Nr 2, 291–299 (2008)

[41] Kojima, K., Fijita, A., Shimamura, T., Imoto, S., Miyano, S., Estimation of nonlinear gene regulatory networks via L1 regularized NVAR from time sereis gene expression data. Genome informatics. International Conference on Genome Informatics 2 ; 20:37-51 (2008)

[42] Lazarov, R.D., Lu, S., Pereverzev, S.V.: On the balancing principle for some problems of numerical analysis. Numer. Math. 106:659âĂŞ689 (2007)

[43] Li, X., Rao, S., Jiang, W., Li, C., Xiao, Y., Guo, Z., Zhang, Q., Wang, L., Du, L., Li, J., Li, L., Zhang, T., Wang, Q.K.: Discovery of time-delayed gene regulatory Networks based on temporal gene expression profiling. BMC Bioinformatics, 7, 26 (2006)

[44] Liu, H., Lafferty, J.D., Wasserman, T.: The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research, 10, 2295-2328 (2009)

[45] Liu, Y., Bahadori, M.T.: A Survey on Granger Causality: A Computational View, Technical Report, University of Southern California, June 24 (2014)

[46] Lozano, A.C., Abe, N., Liu, Y., Rosset, S.: Grouped graphical Granger modeling for gene expression regulatory networks discovery. Vol. 25 ISMB, pp. i110-i118 (2009)

[47] Lu, S., Pereverzev, S.V.: Multi-parameter regularization and its numerical realization, Numer. Math., vol 118, pp. 1–31 (2011)

[48] Lu, S., Pereverzev, S.V.: Regularization theory for ill-posed problems. Selected topics. Inverse and Ill-posed Problems Series, 58. De Gruyter, Berlin (2013)

[49] Marinazzo, D., Pellicoro, M., Stramaglia, S.: Kernel-Granger causality and the analysis of dynamic networks. Physical Review E, 77: 056215 (2008)

[50] Mosci, S., Rosasco, L., Santoro, M., Verri, A., Villa, S.: Nonparametric Sparsity and Regularization. Technical Report MIT, CSAIL, Cambridge, USA, vol 41 (2011)

[51] Marinazzo, D., Pellicoro, M., Stramaglia, S.: Causal information approach to partial conditioning in multivariate data sets, Computational and Mathematical Methods in Medicine (2012)

[52] Mathe, P., Pereverzev, S.V.: Geometry of linear ill-posed problems in variable Hilbert scales. Inverse Problems, 19:789âĂŞ803 (2003)

[53] Naumova, V., Pereverzyev, S.V.: Multi-penalty regularization with a component-wise penalization, Inverse Problems, vol 29, 075002 (2013)

[54] Paluš, M., Komárek, V., Procházka, T., Hrnčír, Z., Štěrbová, K.: Synchronization and information flow in EEGs of epileptic patients, IEEE Engineering in Medicine and Biology Magazine,Vol. 20 No 5, pp. 65–71 (2001)

[55] Pearl, J.: Probabilistic reasoning in intelligent systems. San Mateo, CA, Morgan Kaufmann (1988)

[56] Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press (2009)

[57] Pereverzev, S., Schock, E.: On the adaptive selection of the parameter in regularization of ill-posed problems, SIAM J. Numer. Anal., Vol. 43, pp. 2060–2076 (2005)

[58] Pereverzyev, S. Jr., Hlaváčková-Schindler, K.: Graphical Lasso Granger method with two-level-thresholding for recovering causality networks, Research Report, 09/13, Department of Applied Mathematics, Leopold Franzens Universität Innsbruck (2013)

[59] Purdom, E., Holmes, S.P.: Error distribution for gene expression data, Statistical Applications in Genetics and Molecular Biology, Vol. 4, 1, 16 (2005)

[60] Rajapakse, J.C., Mundra, P. A., Stability of building gene regulatory net- works with sparse autoregressive models, Bioinformatics, 12(Suppl 13):S13 (2011)

[61] Ramlau, R., Teschke, G.: A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints., Journal of Numer. Math., Vol. 104, No 2, pp 177–203 (2006)

[62] Rieder, A.: Keine Probleme mit inversen Problemen. Eine Einführung in ihre stabile Lösung. Wiesbaden, Vieweg (2003)

[63] Sambo, F., Camillo, B.D., Toffolo, G., CNET: An algorithm for reverse engineering of causal gene networks, NETTAB2008, Varenna, Italy (2008)

[64] Seth, A.K.: Causal connectivity of evolved neural networks during behavior, Network-Computation in Neural Systems, Vol. 16, No. 1, pp 35–54 (2005)

[65] Shojaie, A. Michalidis, G.: Discovering graphical Granger causality using the truncating lasso penalty, Bioinformatics 26, 18: i517-i523 (2010)

[66] Shojaie, A., Basu, S. Michalidis, G.: Adaptive thresholding for reconstructing regulatory networks from time course gene expression data, Manuscript at www.biostat.washington.edu (2011)

[67] Shojaie, A., Basu, S. Michalidis, G.: Adaptively Thresholded Lasso for Gene Regulatory Networks, Statistics In Biosciences 4(1): 66-83 (2012)

[68] Shojaie, A.: Link Prediction in Biological Networks using Penalized Multi-Mode Exponential Random Graph Models, Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA (2013)

[69] Schölkopf, B., Herbrich, R., Smola, A.: A generalized representer theorem, in Computational learning, Springer: Lecture Notes in Computer Science 2111, theory, pp. 416–426 (2001)

[70] Song, S., and Bickel, P.J. Large vector auto regressions, arXiv preprint arXiv:1106.3915 (2011).

[71] Soto, J., Pantazis, D., Jerbi, K., Baillet, S. and Leahy, R. Canonical correlation analysis applied to functional connectivity in MEG. In ISBI, 2010.

[72] Spirtes, P., Glymour, C., Scheines, R.:Causation, Prediction, and Search, 2nd edition. The MIT Press (2001).

[73] Shmulevich, I., Dougherty, E.R. Kim, S., Zhang, W.: Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks, Bioinformatics 18 (2): 261-274 (2002).

[74] Steinhaeuser, K., Ganguly A.R., Chawla, N.V.: Multivariate and multiscale dependence in the global climate system revealed through complex networks, Clim. Dyn., 39, pp. 889–895 (2012)

[75] Tibshirani, R., Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. B, Vol. 58, pp. 267-288 (1996)

[76] Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics 4:1035âĂŞ1038 (1963)

[77] Tikhonov, A.N., Arsenin, V.Y.: Solution of Ill-posed Problems. Washington, Winston & Sons (1977)

[78] Tikhonov, A. N., Glasko, V.B. Use of the regularization method in non-linear problems, USSR Comp. Math. Math. Phys. 5:93–107 (1965)

[79] Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., Yagola A.G.: Numerical Methods for the Solution of Ill-Posed Problems, Kluwer Academic Publishers (1995)

[80] Wiener, N.: The Theory of Prediction, In: Modern Mathematics for Engineers, Ed. E.F. Beckenbach, McGraw-Hill, New York (1956)

[81] Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown. P.O., Botstein, D.: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Moll.Biol.Cell., 13(6):1977-2000 (2002)

[82] Wikipedia, Causality, The Free Encyclopedia (2013)

[83] Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics, 20, 35943603 (2004)

[84] Yuan, M., and Lin, Y., Model selection and estimation in regression with grouped variables, Journal of The Royal Statistical Society Series B, 68(1): 49-67 (2006).

[85] Zeng, L., Xie, J. Group Variable Selection for Data with Dependent Structures, Journal of Statistical Computation and Simulation, DOI:10.1080/00949655.2010.529812 (2011)

[86] Zhang, L., Kim, S.: Learning Gene Networks under SNP Perturbations Using eQTL Datasets, DOI: 10.1371/journal.pcbi.1003420 (2014)

[87] Zou, C., Feng, J.: Granger causality vs dynamic Bayesian network inference: A comparative study. BMC Bioinformatics, 10:122 (2009)

[88] Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics, 21, 7179 (2005)

[89] Zou, H. The adaptive lasso and its oracle properties. JASA, 101(476):1418âĂŞ1429 (2006)

[90] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J.R.Statist. Soc. B, 67, Part 2, pp.301-320 (2005)

	CG	LG	LG1	LG2	LG3	MPR
CA	0.80	0.58	0.88	0.85	0.81	0.88
TP	58	38	63	51	42	53

Table 1

Quality measures for the considered methods. The number of the causal links in the considered gene regulatory network from Figure 1 is 95. This number can be seen as the maximal possible value for TP.

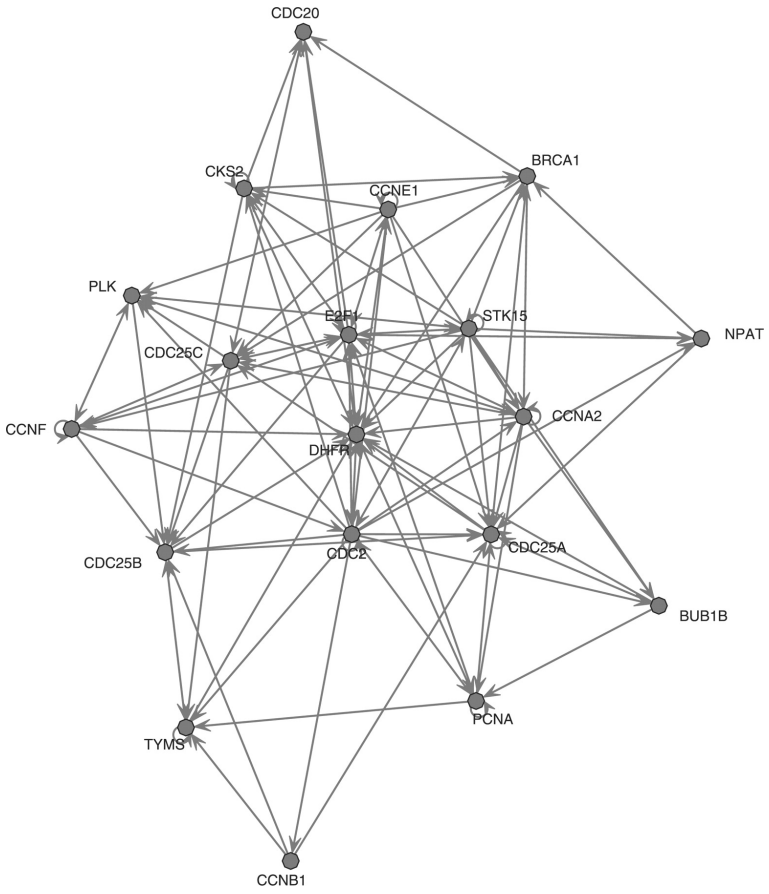

Figure 1. Causal structure from biological experiments for nineteen selected genes (adopted from [46])

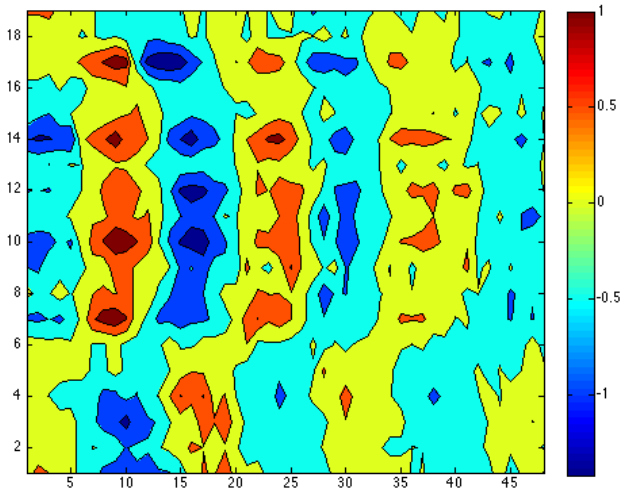

Figure 2. The horizontal coordinate x indicates the 48 time measurements, the vertical coordinate y indicates 19 genes ordered; The color of the pixel corresponds to the value determined by the color-scale on the right hand side.

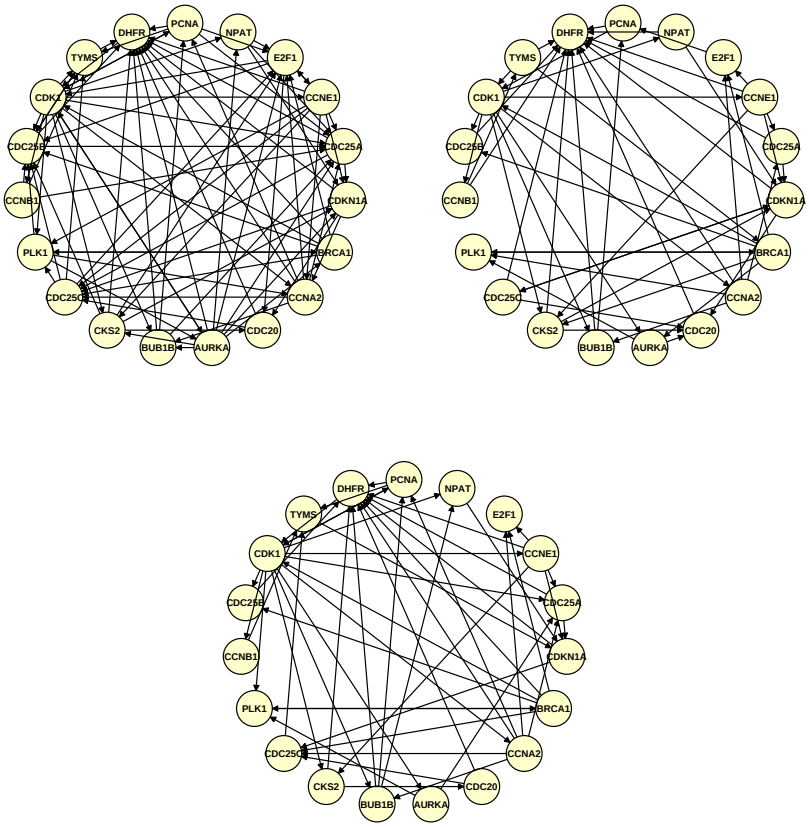

Figure 3. The considered gene regulatory network (top left) and its reconstructions with LG3 (top right) and MPR (bottom) in the circular layout.