Biostatistics (2009), **10**, 3, *pp.* 468–480
doi:10.1093/biostatistics/kxp005
Advance Access publication on March 29, 2009

Estimating equation–based causality analysis with application to microarray time series data

JIANHUA HU*

Department of Biostatistics, Division of Quantitative Sciences,
University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
jhu@mdanderson.org

FEIFANG HU

Department of Statistics, University of Virginia, Charlottesville, VA, USA
fh6e@virginia.edu

Summary

Microarray time-course data can be used to explore interactions among genes and infer gene network. The crucial step in constructing gene network is to develop an appropriate causality test. In this regard, the expression profile of each gene can be treated as a time series. A typical existing method establishes the Granger causality based on Wald type of test, which relies on the homoscedastic normality assumption of the data distribution. However, this assumption can be seriously violated in real microarray experiments and thus may lead to inconsistent test results and false scientific conclusions. To overcome the drawback, we propose an estimating equation–based method which is robust to both heteroscedasticity and nonnormality of the gene expression data. In fact, it only requires the residuals to be uncorrelated. We will use simulation studies and a real-data example to demonstrate the applicability of the proposed method.

Keywords: Chi-square approximation; Estimating equation; F-test; False-positive rate; Granger causality; Time-course data.

1. Introduction

Microarray technologies allow us to gain biological insight at the genomic scale by monitoring activities of thousands of genes simultaneously in a wide range of tissues, organs, and cell lines. In recent years, time-course experiments (Spellman *and others,* 1998; Cho *and others,* 2001; Whitfield *and others,* 2002) generate gene expression data measured repeatedly over a time period. Therefore, it makes it possible to explore biological functions of genes and interactions among genes and their products. Several typical types of temporal patterns and associations among genes have been studied using time-course experiments, including periodicity, co-expression detection, gene clustering, and causality. For example, Filkov *and others* (2002) and Wichert *and others* (2004) have focused on periodicity and phase

*To whom correspondence should be addressed.

detection. The correlation coefficient and its variants have been used as a primary tool of detecting gene co-expression and interactions (Schäfer and Strimmer, 2005; Zhu *and others,* 2005). There is also a large literature on model-based or nonparametric gene clustering such as Peddada *and others* (2003), Schliep *and others* (2003), and Song *and others* (2007).

Another essential problem in studying functional gene–gene interaction and constructing gene network is causality detection. The general idea is to derive pairwise causality relationship among genes based on a certain test that will be used later to construct the network. Therefore, developing an appropriate causality test is the crucial step in network construction. And it is the focus of our study in this paper. As introduced by Mukhopadhyay and Chatterjee (2007), the rough definition of causality relationship between 2 genes is that gene 1 is a cause of gene 2 if expression of gene 1 is predictive of expression of gene 2 at a future time period. In real life, 2 genes can have either a direct or an indirect causality relationship. The indirect causality implies that at least one intermediate gene exists between the 2 genes in the connection chain, which is commonly observed in gene regulatory network.

The causality relationship has been studied extensively in the literature. A class of causal models is the marginal structural models (Robins *and others*, 2000; Hernán *and others*, 2001), which are used in observational studies with exposures or treatments varying over time, for example, in the logistic regression setting. It provides the consistent inverse-probability-of-treatment-weighted estimators in the presence of unmeasured confounding factors. In our problem of studying the causality relationship between 2 time series, another class is probably more relevant that uses coherence- and partial coherence–based graphical models as in Dahlhaus (2000), Butte *and others* (2001), and Salvador *and others* (2005). However, it has been shown that the coherence-based approaches have several shortcomings: (1) it is sensitive to measurement errors (Albo *and others,* 2004) and (2) it cannot detect the time precedence relationship (Kaminski *and others,* 2001; Baccala and Sameshima, 2001) and can bear the least amount of additive random noise. In contrast, Winterhalder *and others* (2005) demonstrated that Granger causality is more appropriate for detecting the type of causality relationship of interest.

Mukhopadhyay and Chatterjee (2007) used a vector autoregressive (VAR) framework to test the Granger causality via an F-test. However, validity of the F-test requires the independently and identically normal distribution of data. It is known that the strong distributional assumptions have often been violated in gene expression data in the 2-fold way: (1) Gene expression intensities are often not normally distributed. Some researchers used data transformation procedures (Rocke and Durbin, 2003; Durbin and Rocke, 2004), typically logarithm transformation, to stabilize the variance so as to better approximate the normal distribution, which is still not satisfactory in some circumstances. (2) The errors are not necessarily homogeneous in real life or the variances of expression intensities in different groups or at different time points are not identical. For example, earlier studies (Geller *and others*, 2003; Hu and Wright, 2007) have shown that the variance is positively associated with the mean expression intensities. Another limitation of the F-test is that it is only applicable to the least square type of parameter estimates, which is restrictive in solving real problems. For example, F-test is not valid using the parameter estimates based on some L_1-distance objective function associated with robust (i.e. median) regression models, that is practically necessary in some scenarios to account for the presence of outliers.

To overcome the shortcomings of an F-test, we propose a test based on estimating equations. The proposed method only requires data to be uncorrelated, which allows making valid statistical inference and hypotheses test robust to a wide range of data distribution forms and the heteroscedasticity of the errors that may be related to the mean function of the models. Moreover, it is also generally applicable to parameter estimates obtained from a wide range of objective functions. Aside from the capability of maintaining the appropriate significance level, empirical studies also illustrate its advantage in terms of false-positive (FP) rate in the comparison to F-test. The method and theory will be described in Section 2. We will discuss simulation studies extensively in Section 3. A real human cell cycle time-course data will be used for demonstration in Section 4.

2. Motivation and method

We consider the following autoregressive model:

$$y_{1(t)} = c + \alpha_1 y_{1(t-1)} + \cdots + \alpha_q y_{1(t-q)} + \beta_1 y_{2(t-1)} + \cdots + \beta_q y_{2(t-q)} + \epsilon_t, \tag{2.1}$$

where q is the autoregressive lag length and ϵ_t, $t = 1, \ldots, n$, are uncorrelated and have mean 0 and variance σ_t^2. The gene 2 is said to Granger-cause gene 1 if at least a $\beta_i \neq 0$, $i = 1, \ldots, q$. The task becomes testing

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_q = 0. \tag{2.2}$$

Hamilton (1994) and Mukhopadhyay and Chatterjee (2007) used F-test in a VAR framework assuming the independent and identically distributed ϵ_t. In gene expression time-series data, it is observed that the variance of expression levels of a gene varies with the mean expression intensity at stages (or time points) in the cell cycle. In this case, F-test would lead to inconsistent results in the presence of the nonhomogeneous errors, mainly due to inconsistent variance estimation of the parameters.

To overcome the drawbacks of F-test, we propose a test based on estimating equations, which only requires ϵ_t in model (2.1) to be uncorrelated, without additional assumptions on the distributional form. In addition, it also allows the unequal variance σ_t^2. Let $Y_{1(t)} = (y_{1(t)}, \ldots, y_{1(t-q)})$ and $Y_{2(t)} = (y_{2(t)}, \ldots, y_{2(t-q)})$. We define the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, c, \boldsymbol{\alpha})$. The estimate of $\boldsymbol{\theta}$ is generally considered to be the solution of the following estimating equation:

$$S(y, \boldsymbol{\theta}) = n^{-1/2} \sum_{t=1}^{n} g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) = 0 \tag{2.3}$$

for some given function g satisfying

$$Eg(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) = 0.$$

The normalization constant $n^{-1/2}$ is chosen for the convenience of expression asymptotic results. The estimating equation (2.3) is typically obtained from minimization (maximization) of some objective function $SS(y, \boldsymbol{\theta}) = \sum_{t=1}^{n} G_t(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})$, that is, if the likelihood or least squares is used, then we have $g_t(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) = \partial G_t(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})/\partial \theta$.

It is worthwhile emphasizing that the inference development on estimating equations in our problem is different from the main existing research which focuses on the independent data (see Liang and Zeger, 1986; Godambe and Kale, 1991; Boos, 1992; Hu and Kalbfleisch, 2000). In our problem, the time-series data are dependent where some explanatory variables are in fact also some forms of the outcome variable. Based on model (2.1), the least square method results in the estimating function

$$g(Y_{1(t)}, Y_{2(t)}) = (y_{2(t-1)}, \ldots, y_{2(t-q)}, 1, y_{1(t-1)}, \ldots, y_{1(t-q)})^T \times r_t,$$

a $2q + 1$ vector, where

$$r_t = y_{1(t)} - (c + \alpha_1 y_{1(t-1)} + \cdots + \alpha_q y_{1(t-q)} + \beta_1 y_{2(t-1)} + \cdots + \beta_q y_{2(t-q)}).$$

We define the following notations prior to making the inference of $\boldsymbol{\theta}$. Let

$$V(\boldsymbol{\theta}) = \operatorname{Var} S(y, \boldsymbol{\theta}) = n^{-1} \sum_{t=1}^{n} \operatorname{Var}(g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}))$$

and

$$W(\boldsymbol{\theta}) = E\left\{n^{-1}\sum_{t=1}^{n}\frac{\partial g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\}.$$

Both $V(\boldsymbol{\theta})$ and $W(\boldsymbol{\theta})$ are $(2q+1) \times (2q+1)$ matrices. In practice, a reasonable estimate of $V(\boldsymbol{\theta})$ and $W(\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$ is

$$V(y, \boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n}(g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) - \bar{g}(y, \boldsymbol{\theta}))(g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) - \bar{g}(y, \boldsymbol{\theta}))^T,$$

where $\bar{g}(y, \boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n} g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})$. And

$$W(y, \boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n}\frac{\partial g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}.$$

Consequently, we have

$$V(y, \boldsymbol{\theta}) \to V(\boldsymbol{\theta})$$

and

$$W(y, \boldsymbol{\theta}) \to W(\boldsymbol{\theta})$$

in probability for both homogeneous and nonhomogeneous errors, which can be straightforwardly obtained by the weak law of large number.

We start with a general testing problem that has a wide application,

$$H_0\colon h(\boldsymbol{\theta}) = 0$$

for a set of differentiable functions h, with the length of the vector h denoted by r. Note that the test in (2.2) is just a special case. Let

$$H(\boldsymbol{\theta}) = \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}.$$

We define

$$U(\boldsymbol{\theta}) = H^T(HW^{-1}H^T)^{-1}HW^{-1}VW^{-1}H^T(HW^{-1}H^T)^{-1}H,$$

and its general inverse (Moore–Penrose) is then

$$U^{-}(\boldsymbol{\theta}) = W^{-1}H^T(HW^{-1}VW^{-1}H^T)^{-1}HW^{-1},$$

where the arguments y and $\boldsymbol{\theta}$ are suppressed.

Let $\tilde{\boldsymbol{\theta}}$ be the estimate of $\boldsymbol{\theta}$ under the restriction $h(\boldsymbol{\theta}) = 0$ obtained from estimating equations. The test statistics based on the estimating equation is then

$$\tilde{Q}_{h=0} = S(y, \tilde{\boldsymbol{\theta}})^T U^{-}(\tilde{\boldsymbol{\theta}}) S(y, \tilde{\boldsymbol{\theta}}).$$

We derived its asymptotic distribution property in the following theorem, with the proof shown in the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org).

THEOREM 2.1 If the Lindeberg condition holds for $g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})$. That is, for every ℓ,

$$\lim_{n\sim\infty}\sum_{t=1}^{n} E\left(\frac{g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta})^2}{s_n^2} I(\mid g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}) \mid > \ell s_n)\right) = 0,$$

where $s_n^2 = \sum_{t=1}^{n} \mathrm{Var}(g(Y_{1(t)}, Y_{2(t)}, \boldsymbol{\theta}))$. Then, $\tilde{Q}_{h=0}$ follows a chi-square distribution with the degree of freedom r under H_0. Therefore, H_0 can be rejected at the significance level α if $\tilde{Q}_{h=0} > \chi^2_{r,\alpha}$.

Coming back to the test (2.2), we have the $q \times (2q+1)$ matrix

$$H(\boldsymbol{\theta}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ & & & \vdots & & & \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix},$$

where the element in the ith row and ith column equals 1, $i = 1, \ldots, q$, corresponding to $\boldsymbol{\beta}$. So $\tilde{Q}_{h=0}$ follows the chi-square distribution with the degree of freedom q under H_0.

Because our focus is on the casuality test, we will not discuss selection of the autoregressive lag length in this paper. Throughout the empirical investigation, we will take $q = 1$ without the loss of generality. We will also study the application of the gene-pair causality relationship in network construction, adopting the same strategy in Mukhopadhyay and Chatterjee (2007) that is to first perform pairwise causality tests and then use multiple testing adjustment to select the most associated gene pairs for constructing the network. Regardless of the *ad hoc* way of network construction, it serves sufficiently the main purpose of making the comparison between the estimating equation–based test and the F-test in the high-dimensional application.

3. SIMULATION STUDY

We conducted extensively a series of simulation studies to evaluate the performance of the estimating equation–based causality detection method (EE) and the F-test (F).

It is essential for a test method to preserve the appropriate type-I error rate that is investigated in the first set of simulations. Additionally, we empirically examine the distribution of the test statistic $\tilde{Q}_{h=0}$ to verify the theoretical results stated in Theorem 1. We focus on testing the causality relationship between 2 variables X_1 and X_2 that $x_{1t} = 0.7x_{1(t-1)} + \epsilon_{1t}$ and $x_{2t} = 0.3x_{2(t-1)} + \beta x_{1(t-1)} + \epsilon_{2t}$. The interest is to test the null hypothesis H_0: $\beta = 0$. We conducted 5000 simulations with the data generated under null hypothesis H_0. We considered the number of time points to be $n = 30$ or 100. We conduct the 2 methods to test H_0 in each simulation. Three homogeneous distributions of both ϵ_{1t} and ϵ_{2t} are considered: $N(0, 1)$, t-distribution with degree of freedom 3, and centered mean-1 exponential distribution. We report the proportion of the p-value no larger than 0.05 among the 5000 simulations in the 3 columns to the left in Table 1. In most cases, the 2 methods yield similarly reasonable results that are very close to 0.05. However, EE method has discernibly better performance when the t-distribution is posed on the residuals.

We also considered a nonhomogeneous case where $\epsilon_{1t} \sim N(0, 1)$ and the variance of ϵ_{2t} is positively associated with the absolute mean value of $x_{1(t-1)}$. Specifically, $\epsilon_{2t} \sim N(0, x^2_{1(t-1)})$. It serves sufficiently as an example to demonstrate the inconsistency of F-test shown in the right column of Table 1. We can see that EE method performs consistently well while F-test yields much inflated type-I error rate.

Table 1. *Type-I error rate results in simulation study*

n	Method	Homogeneous ϵ_{2t}			Nonhomogeneous ϵ_{2t}
		$N(0, 1)$	$t(3)$	Centered exp(1)	$N(0, x^2_{1(t-1)})$
30	F	0.060	0.062	0.054	0.241
	EE	0.060	0.049	0.050	0.060
100	F	0.053	0.062	0.053	0.242
	EE	0.052	0.050	0.053	0.050

We observed that $n = 30$ and 100 generally yield the compatible results where the latter may perform slightly better in some cases, implying that the tests are also applicable when the sample size is not large.

In addition, we compared the distribution of the test statistic $\tilde{Q}_{h=0}$ to its theoretical distribution $\chi^2(1)$ in each residual distribution case. The density histogram plots in Figure 1 of the supplementary material available at *Biostatistics* online support the validity of the theoretical results.

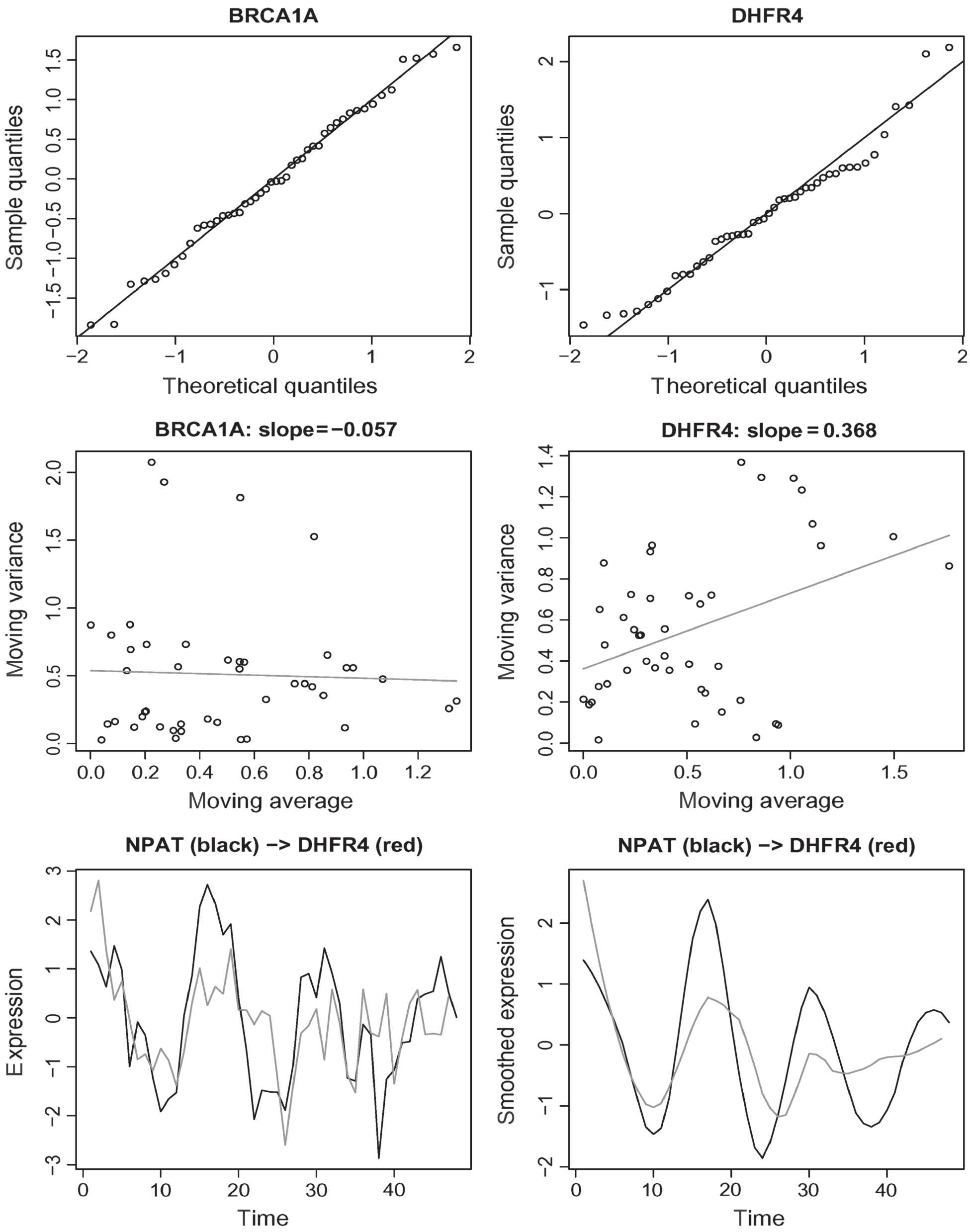

Fig. 1. In the first row, the Q–Q plots of the expression intensities of the 2 genes *BRCA1A* and *DHFR4* are shown. In the second row, the plots of the moving variance versus the moving average intensities of *BRCA1A* and *DHFR4* are shown. In the third row, the expression intensities along time for the gene pair (*NPAT*, *DHFR4*) are shown.

Next, we simulate a network of 14 genes in a similar way as in Mukhopadhyay and Chatterjee (2007) where the 6 independent genes x_1, x_7, x_8, x_9, x_{11}, and x_{14} are all AR(1) process with autocorrelation <1. To mimic the real situation, we considered the independent genes to come from the stationary or nonstationary time-series processes, where the latter is to account for the periodicity pattern of gene expression in the cell cycle regulation. The other 8 dependent genes are simulated from the Granger-caused time-series processes associated with one or more other genes. The models to generate the whole network are described in Table 2.

The goal of simulating such a gene network is to mimic the realistic problem, taking into consideration different complexity in the dependent structure, including the indirect causalities (i.e. $x_1 \to x_3$), multiple-to-one prediction (i.e. $(x_7, x_8, x_9) \to x_6$), one-to-multiple prediction (i.e. $x_{11} \to x_{10}$ and $x_{11} \to x_{12}$), and the disconnect component.

As discussed in Section 2, the distribution of ϵ_t can be either nonnormal (symmetric or asymmetric) or nonhomogeneous (variance dependent on time) in the real life. First, we considered 2 homogeneous ϵ_t cases: (1) $\epsilon_t \sim N(0, 1)$, which is the only case discussed in Mukhopadhyay and Chatterjee (2007), and (2) $\epsilon_t \sim$ centered $\exp(1)$, representing an asymmetric distribution case.

We generate all the time series in various scenarios for 100 equidistant time points. We then conduct the tests at time points $t = 10, 20, 40, 60, 80$, and 100 to assess the influence of the number of observations (or sample size) on the results. Between a pair of genes, we only take the direction with smaller p-value from the causality test to construct the gene network. We understand that it is possible for 2 genes to have the bidirectionally causal relationship in the either direct or indirect fashion.

For the high-throughput data, it is well acknowledged that multiple testing issue needs to be taken into account. There is a large literature on this subject, among which false discovery rate (FDR; Benjamini and Hochberg, 1995; Storey, 2002) has been the most commonly used. Although applying FDR to our problem is straightforward, we note that different p-value cutoff can be incurred using a common FDR threshold for different tests because of the unequal estimate of the distribution and the proportion of null genes produced by permutation procedures, that is, implemented in Q-value software (Storey and Tibshirani, 2003). As a consequence, it may hinder the direct and square comparison between the 2 tests. Therefore,

Table 2. *Data generation of the network of* 14 *genes*

Independent series	
Case 1: stationary	Case 2: nonstationary
$x_{1t} = 0.7x_{1(t-1)} + \epsilon_t$	$x_{1t} = \sin\frac{\pi t}{40} + 0.7x_{1(t-1)} + \epsilon_t$
$x_{7t} = 0.8x_{7(t-1)} + \epsilon_t$	$x_{7t} = 0.8x_{7(t-1)} + \epsilon_t$
$x_{8t} = 0.7x_{8(t-1)} + \epsilon_t$	$x_{8t} = \cos\frac{\pi t}{40} + 0.7x_{8(t-1)} + \epsilon_t$
$x_{9t} = 0.77x_{9(t-1)} + \epsilon_t$	$x_{9t} = 0.77x_{9(t-1)} + \epsilon_t$
$x_{11t} = 0.7x_{11(t-1)} + \epsilon_t$	$x_{11t} = \cos\frac{\pi t}{40} + 0.7x_{11(t-1)} + \epsilon_t$
$x_{14t} = 0.65x_{14(t-1)} + \epsilon_t$	$x_{14t} = 0.65x_{14(t-1)} + \epsilon_t$
Dependent series	
$x_{2t} = 0.29x_{2(t-1)} + 0.65x_{1(t-1)} + \epsilon_t$	
$x_{3t} = 0.15x_{3(t-1)} + 0.29x_{2(t-1)} + 0.65x_{14(t-1)} + \epsilon_t$	
$x_{6t} = 0.12x_{6(t-1)} + 0.3x_{7(t-1)} + 0.3x_{8(t-1)} + 0.3x_{9(t-1)} + \epsilon_t$	
$x_{4t} = 0.17x_{4(t-1)} + 0.4x_{3(t-1)} + 0.7x_{6(t-1)} + \epsilon_t$	
$x_{5t} = 0.6x_{5(t-1)} + 0.8x_{4(t-1)} + \epsilon_t$	
$x_{10t} = 0.4x_{10(t-1)} + 0.3x_{11(t-1)} + \epsilon_t$	
$x_{12t} = 0.4x_{12(t-1)} + 0.4x_{11(t-1)} + \epsilon_t$	
$x_{13t} = 0.4x_{13(t-1)} + 0.4x_{11(t-1)} + \epsilon_t$	

we instead take the common significance level for the 2 methods throughout the simulations and real-data investigation. Without the loss of generality, we use the p-value cutoff value of 0.01 in the simulation study.

Now, we describe a summary measure to assess the performance of the 2 methods. Mukhopadhyay and Chatterjee (2007) introduced an accuracy summary statistic as follows. Let G denote the number of genes and e_i denote the edge connecting 2 genes, $i = 1, \ldots, \frac{G(G-1)}{2}$. Define $l(e_i) = 0$ if the decision based on the test is accurate; otherwise, $l(e_i) = 1$. The accuracy summary statistic is defined as

$$A = 100\left[1 - \frac{2}{G(G-1)} \sum_{e_i} l(e_i)\right].$$

Although the summary statistic measures the overall decision accuracy, we note that it cannot deliver the important information on the FPs or false negative (FNs). In exploring a biological problem using the high-through put data such as microarray experiment, it is more important to control for FP rate since the detected candidate markers selected from statistical analysis will be focused for further biological validation procedure which is costly and time consuming, aside from that biologists accept the reality of not being able to detect all the true biomarkers, which is related to FNs. Therefore, we also record the number of FPs (the detected edges that are not true edges) and the number of FNs (the true edges that are not detected).

In Table 1 of the supplementary material available at *Biostatistics* online, we show in each case the mean accuracy summary statistic with its standard error, the mean FP, and the mean FN over all the 100 simulations at the significance level of 0.01. In general, the 2 methods are very compatible in terms of both the average accuracy measure and the standard error. But the difference between the two is that EE tends to yield less FPs while F-test tends to yield less FNs. More detailed discussions are provided in the supplementary material available at *Biostatistics* online. In addition, it is worth pointing out that F-test is in fact theoretically inconsistent, regardless of its seemingly reasonable empirical performance under some situations.

Next, we examined the data in the nonhomogeneous normal distribution case where the variance of ϵ_{it} in the model of predicting gene i is positive in absolute expression intensities of another gene (or other genes) that interacts with gene i at time $t-1$. The data generation details are described in Table 3. Explicitly, we assume the variance of the genes in the dependent series is directly related to the genes predicting them. For the genes in the independent series, we also assume their variance is associated with another (arbitrarily selected) gene's expression intensity due to indirect interaction between the two in the same network. Although the data generation seems arbitrary, it is sufficient to illustrate how the irregular variance structure could affect the causality test. Nevertheless, the true variance structure is unknown in the real cell cycle time-series gene expression data. Therefore, a method robust to the data distribution form is desirable.

Table 3. *Standard deviation of* ϵ_{it}

Independent series	Dependent series
$\mathrm{sd}(x_{1t}) = 1$	$\mathrm{sd}(x_{2t}) = 5 \mid x_{1(t-1)} \mid$
$\mathrm{sd}(x_{7t}) = 5 \mid x_{1(t-1)} \mid$	$\mathrm{sd}(x_{3t}) = 5 \mid x_{2(t-1)} + x_{14(t-1)} \mid$
$\mathrm{sd}(x_{8t}) = 5 \mid x_{7(t-1)} \mid$	$\mathrm{sd}(x_{6t}) = 5 \mid x_{7(t-1)} + x_{8(t-1)} + x_{9(t-1)} \mid$
$\mathrm{sd}(x_{9t}) = 5 \mid x_{8(t-1)} \mid$	$\mathrm{sd}(x_{4t}) = 5 \mid x_{3(t-1)} + x_{6(t-1)} \mid$
$\mathrm{sd}(x_{11t}) = 5 \mid x_{9(t-1)} \mid$	$\mathrm{sd}(x_{5t}) = 5 \mid x_{4(t-1)} \mid$
$\mathrm{sd}(x_{14t}) = 5 \mid x_{11(t-1)} \mid$	$\mathrm{sd}(x_{10t}) = 5 \mid x_{11(t-1)} \mid$
	$\mathrm{sd}(x_{12t}) = 5 \mid x_{11(t-1)} \mid$
	$\mathrm{sd}(x_{13t}) = 5 \mid x_{11(t-1)} \mid$

Table 4. *Heterogeneous* ϵ_t

Method	Time	5	10	15	20	40	60	80	100
F	Avg	85.14	86.52	87.35	88.14	88.42	88.64	88.67	89.25
	SE	4.74	4.98	3.72	3.10	3.21	3.03	3.13	3.14
	FP	7.50	6.24	5.44	5.15	4.74	4.61	4.69	4.49
	FN	6.02	6.03	6.07	5.64	5.80	5.73	5.62	5.29
EE	Avg	90.68	91.26	91.88	92.26	91.92	91.86	91.58	91.63
	SE	2.52	2.18	2.34	2.27	2.07	2.38	2.22	2.12
	FP	1.98	1.31	1.16	0.93	1.09	1.11	1.16	1.18
	FN	6.50	6.64	6.23	6.11	6.26	6.30	6.50	6.44

SE, standard error.

The results with the sample size varying between 5 and 100 are shown in Table 4. It is clear that EE outperforms F-test in terms of the overall accuracy measure because of its obvious advantage in FP with not much sacrifice of FN. We also note that the standard error of F-test is larger than EE, indicating the inconsistent performance of F.

4. A real-data example

We study the applicability of the proposed method using the human cancer cell cycle data (Whitfield *and others*, 2002) available at http://genome-www.stanford.edu/Human-CellCycle/Hela. Li *and others* (2006) studied gene regulatory network on 20 genes, represented by 23 probe sets, using one experiment in this data with a double thymidine block for cell synchronization which consisted of 48 time points. We focus on the same set of genes to study the gene–gene causal relationship. A simple scaling normalization was applied to the arrays to make their overall expression levels comparable prior to data analysis.

We implemented both F and EE on this set of genes. At the p-value threshold of 0.001, F detected 104 gene pairs that showed the significant causal relationship, while EE detected 78, among which 70 pairs were detected by both. It is worthwhile investigating the gene pairs that are detected by only a method in order to understand what causes the difference. The theoretical development and simulation study tell us that the performances of 2 methods are supposed to be similar if the expression intensities of a gene to be predicted is homogeneously and normally distributed. It motivates us to take a look at the distribution of the expression intensities of each individual gene first.

An intuitive way is comparing the sample quantiles to the theoretical quantiles of the i.i.d. standard normal distribution. A large discrepancy implies that at least one of the following distribution assumptions does not hold: normality and homogeneity that in particular indicates the common variance at different time points in our case. We empirically observed that the majority of the 23 probe sets manifests the fairly small discrepancy between the sample and the theoretical quantiles, except for 6 genes. For a specific gene, we also kept the count of gene pairs involving it that was only detected by a method. Not surprisingly, we found that the most difference is associated with these 6 genes. Hereafter, we will focus our discussion on a few genes for the purpose of demonstration, with more details provided in the supplementary material available at *Biostatistics* online.

Figure 1 showed the plot of the empirical quantiles of the standardized expression intensities versus the theoretical quantiles of the standard normal distribution for 2 genes breast cancer 1, early onset isofom A (*BRCA1A*) and dihydrofdale reductase gene (*DHFR4*). *BRCA1A* in the top left panel showed that the 2 distributions are generally close to each other, representing the example with nearly normal distribution. Note that the inconsistent estimate of some quantiles in the extreme tail region is mainly caused by the small number of available observations, which is well recognized in statistical literature. In the contrast, *DHFR4* (one of the 6 genes) in the top right panel served as the nonnormal distribution example where the nonlinear up-tail pattern is clearly observed.

Furthermore, it is interesting to investigate if the large deviation from the i.i.d. normal distribution is caused by the nonhomogeneous distribution of the data, more explicitly, if the variance depends on the mean intensity at a time point. Although this is impossible to interrogate without the replicate samples available at a time point, we propose to use the observations at the nearby time points for the investigation. In so doing, we assume that the correlation of expression intensities across time points is high for nearby time points and low otherwise. The idea is to look at the plot of the variance versus the absolute mean intensity of the 5 observations at time point t_i up to t_{i+4}, $i = 1, \ldots, n-4$. If in truth the variance is homogeneous, we expect the moving variance to be uncorrelated with the moving mean intensities. The middle row of Figure 1 contains such plots of the *BRCA1A* and *DHFR4*. The line in red indicates the least square fit of the moving variances regressing on the moving average intensities, with the estimated slope shown in the title above a plot. It is clearly observed that *BRCA1A* generally shows no association between the moving variance and the moving average intensities with the slope close to 0. In the contrast, *DHFR4* shows some linear association between the two with the nontrivial slopes.

The next question to ask is how the non-i.i.d. distribution affects the causality test empirically. One interesting observation is that EE detected 3 genes that strongly predict gene *DHFR4* while F could not detect any. We used the gene pair nuclear protein, ataxia-telangiectasia locus (*NPAT*, *DHFR4*) as an example where EE obtained the p-value of 0.0002 while F is 0.013. The expression intensities of the 2 genes along the 48 time points are displayed in the bottom left panel of Figure 1. To show the direct prediction of *DHFR4*, the expression intensities of *DHFR4* in red at time point 2 and up were used in the plot. We also displayed the smoothed expression intensity curve in the top right panel to show the pattern more clearly. It described the predictive pattern that the change in the expression intensities of *NPAT* always triggered the change in the intensities of gene *DHFR4* of a smaller magnitude in a concordant fashion.

To investigate the impact of data distribution on the test, we standardized the expression profile of gene *DHFR4* with a monotonic normal score transformation. It is noteworthy that this transformation also removes the true nonhomogeneity in terms of the variance besides forming the normal distribution. Clearly, performing such a transformation is not recommended in real application because it erases the underlying true correlation among the data, and the result is difficult to be interpreted after the transformation. However, given that the truth is unknown in the real experiment, we take this strategy to make a comparison between these 2 methods because we expect the proposed test to be less susceptible to different distribution forms than F-test.

Specifically, we obtain the ranks of gene expression intensities $R_1, \ldots, R_n$ and use them to construct the transformed profile, $\Phi^{-1}(R_1/(n+1)), \ldots, \Phi^{-1}(R_n/(n+1))$, where $\Phi(\cdot)$ is the cumulative normal distribution. On the transformed profile of *DHFR4*, EE and F yielded the p-values of 0.00005 and 0.0003, respectively. Note that F substantially increased the test significance and became significant at the prespecified p-value threshold, while EE roughly maintained the same significance level as that without data transformation. Interestingly, Mukhopadhyay and Chatterjee (2007) also detected this causality relationship using another cell cycle experiment.

From our extensive study of this data, the general observation is that the proposed EE method yields more consistent and robust results in response to different data distribution than F-test. In particular, F-test tends to produce much more inconsistent results than EE if the homogeneous normality distribution assumption does not hold. We want to point out that it is impossible to study the variance pattern of the residuals at a time point given no replicates in the real data set. So the inconsistency of F-test may be caused by the nonnormal distribution or the possible confounding heteroscedasticity of the errors. This empirical conclusion is also in agreement with the theoretical development and simulation results described in Sections 2 and 3.

5. Discussion

We proposed a estimating equation–based test for gene–gene causality relationship in microarray time-course data. We theoretically derived that the proposed test statistic has the asymptotic chi-square distribution property under the minimum data distribution assumption. Its capability of maintaining the appropriate significance level has been empirically verified via the simulation study of a wide variety of data distribution forms. Both simulation and real-data example demonstrate the advantage of the proposed method in reducing the FP rate with a minor sacrifice of increasing FN rate of the casuality detection. In the contrast, F-test could break down in the presence of non-i.i.d normal data distribution, in particular, when the variance of the predicted genes varies with the time points and is associated with the mean expression intensities of some other genes. In addition to the advantage of robustness, the proposed method can also accommodate a wide range of objective functions, as long as the estimating equations are appropriately described. For example, the L_1-distance objective function can be solved with some modification (involving bootstrap procedures) based on the fundamental method described in this paper.

There are at least several interesting open problems to be explored in this area. It is necessary to take into account the nonstationary component in the mean expression which is typically a function of the time points using the estimating equation approach. As the simulation study illustrated, it may cause both the methods to break down if the mean expression function at a time point is not correctly specified. In addition, a large portion of the missing observations are found in the real-data example. In this paper, we only focused on the genes with no missing observations, which implies a large loss of information. How to handle the missing data in the model is worthwhile investigating, which is not a trivial problem since the outcome variable and the predictors are dependent. Besides, we only studied lap-1 association here. So how to choose the laps that best reflect the association between 2 genes is interesting and can also be considered as a variable selection problem.

Acknowledgments

Conflict of Interest: None declared.

Funding

National Science Foundation (DMS-0706818 to J.H., DMS-0349048 to F.H.); National Institutes of Health (R01 RGM080503A to J.H.).

Supplementary material

Supplementary material is available at http://www.biostatistics.oxfordjournals.org.

References

Albo, Z., Prisco, G. V., Chen, Y., Rangarajan, G., Truccolo, W., Feng, J., Vertes, R. P. and Ding, M. (2004). Is partial coherence a viable technique for identifying generators of neural oscillations? *Biological Cybernetics* **90**, 318–326.

Baccala, L. A. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics* **84**, 463–474.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

BOOS, D. D. (1992). On generalized score tests. *The American Statistician* **46**, 327–333.

BUTTE, A. J., BAO, L., REIS, B. Y., WATKINS, T. W. AND KOHANE, I. S. (2001). Comparing the similarity of time-series gene expression using signal processing metrics. *Journal of Biomedical Informatics* **34**, 396–405.

CHO, R. J., HUANG, M., CAMPBELL, M. J., DONG, H., STEINMETZ, L., SAPINOSO, L., HAMPTON, G., ELLEDGE, S. J., DAVIS, R. W. AND LOCKHART, D. J. (2001). Transcriptional regulation and function during the human cell cycle. *Nature Genetics* **27**, 48–54.

DAHLHAUS, R. (2000). Graphical interaction models for multivariate time series. *Metrika* **51**, 157–172.

DURBIN, B. AND ROCKE, D. M. (2004). Variance stabilizing transformations for two-color microarrays. *Bioinformatics* **20**, 660–667.

FILKOV, V., SKIENA, S. AND ZHI, J. (2002). Analysis techniques for microarray time-series data. *Journal of Computational Biology* **9**, 317–330.

GELLER, S. C., GREGG, J. P., HAGERMAN, P. AND ROCKE, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics* **19**, 1817–1823.

GODAMBE, V. P. AND KALE, B. K. (1991). Estimating functions: an overview. In: Godambe, V. P. (editor), *Estimating Functions*. Oxford: Clarendon Press, pp. 3–20

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

HERNÁN, M. A., BRUMBACK, B. AND ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96**, 440–448.

HU, F. AND KALBFLEISCH, J. D. (2000). The estimating function bootstrap. *The Canadian Journal of Statistics* **28**, 449–499.

HU, J. AND WRIGHT, F. A. (2007). Assessing differential gene expression with small sample sizes in oligonucleotide arrays using a mean-variance model. *Biometrics* **63**, 41–49.

KAMINSKI, M., DING, M., TRUCCOLO, W. A. AND BRESSLER, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics* **85**, 145–157.

LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LI, X., RAO, S., JIANG, W., LI, C., XIAO, Y., GAO, Z., ZHANG, Q., WANG, L., DU, L., LI, J. *and others* (2006). Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics* **7**, 26.

MUKHOPADHYAY, N. D. AND CHATTERJEE, S. (2007). Causality and pathway search in microarray time series experiment. *Bioinformatics* **23**, 442–449.

PEDDADA, S., LOBENHOFER, E. K., LI, L., AFSHARI, C. A., WEINBERG, C. R. AND UMBACH, D. M. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834–841.

ROBINS, J. M., HERNÁN, M. A. AND BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

ROCKE, D. M. AND DURHIN, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**, 966–972.

SALVADOR, R., SUCKLING, J., SCHWARZBAUER, C. AND BULLMORE, E. (2005). Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B* **360**, 937–946.

Schliep, A., Schönhuth, A. and Steinhoff, C. (2003). Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19**, i255–i263.

Schäfer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764.

Song, J. J., Lee, H., Morris, J. S. and Kang, S. (2007). Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry* **31**, 265–274.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B* **64**, 479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

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. *and others* (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell* **13**, 1977–2000.

Wichert, S., Fokianos, K. and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* **20**, 5–20.

Winterhalder, M., Scheltera, B., Hessec, W., Schwabc, K., Leistritzc, L., Klanc, D., Bauerd, R., Timmera, J. and Witte, H. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing* **85**, 2137–2160.

Zhu, D., Hero, A. O., Qin, Z. S. and Swaroop, A. (2005). High throughput screening of co-expressed gene pairs with controlled false discovery rate (FDR) and minimum acceptable strength (MAS). *Journal of Computational Biology* **12**, 1029–1045.

[*Received March 30, 2008; revised November 24, 2008; accepted for publication December 18, 2008*]