Microarray time course data analysis

Jianhua Hu
UT MD Anderson Cancer Center

March 23, 2008

Types and features

1. Typically short series: k=4-10 time points for shorter series; 11-50 time points for longer series.
2. Often irregularly spaced.
3. With no or few (<5) replications.
4. Can be periodic, as in the cell cycle: Cho et al. (1998), Spellman et al. (1998), or circadian rhythms (Storch et al., 2002).
5. May have no particular pattern, as in developmental time courses; Chu et al. (1998), Wen et al. (1998), Tamayo et al. (1999).

1. May be longitudinal, where mRNA samples at different times are extracted from the same unit (cell line, tissue or individual), but more commonly cross-sectional, where mRNA samples are from different units.
2. Gene expression values at different time points may be correlated, especially in a longitudinal study, or when a common reference design is used for a cross-sectional study. At other times, the experimental design induces correlations in cross-sectional studies.

1. Several general types of hypotheses of interest:
    1. the one-sample (or one-class) problem: which genes are changing in time?
    2. the 2 or >2 sample (or class) problem: which genes are changing differently in time across the samples (or classes)?
    3. gene-gene association: coexpression, causality, etc.
2. Two broad types of mRNA samples: from cells or cell lines which give reasonably repeatable responses within classes, and whole organism (mice, humans), where there is a lot of response variability within classes.

Design of microarray time course experiments

1. The first issue is: longitudinal or cross-sectional? The question revolves on whether it is important to measure change within units.
2. For two-channel (cDNA or long oligo) arrays, a major question is whether or not to use a reference design. Most frequently, the answer is yes.
3. For very short two-channel time courses, the possibility arises of optimizing the design for contrasts of interest.
4. Important design issues include not just assignment of mRNA to arrays, but also the actual conduct of the experiment, including preparation of the sample mRNA, the times of hybridizations, and the equipment, reagents and personnel used.

Replication

1. We can have biological, technical and probe set (spot) replicates.
2. Replication is a good thing. With it we get estimates of variability relative to which temporal changes and/or condition differences can be assessed.
3. Biological replicates are best, as they permit conclusions to be extrapolated, something not possible with tech. reps.
4. With unreplicated experiments, inference to a wider population is not possible, and analysis is less straightforward, being dependent on unverifiable assumptions, as no estimate of pure error is available.
5. When we do have replicates, it is better to use the variation between them in the analysis, and not simple average them.

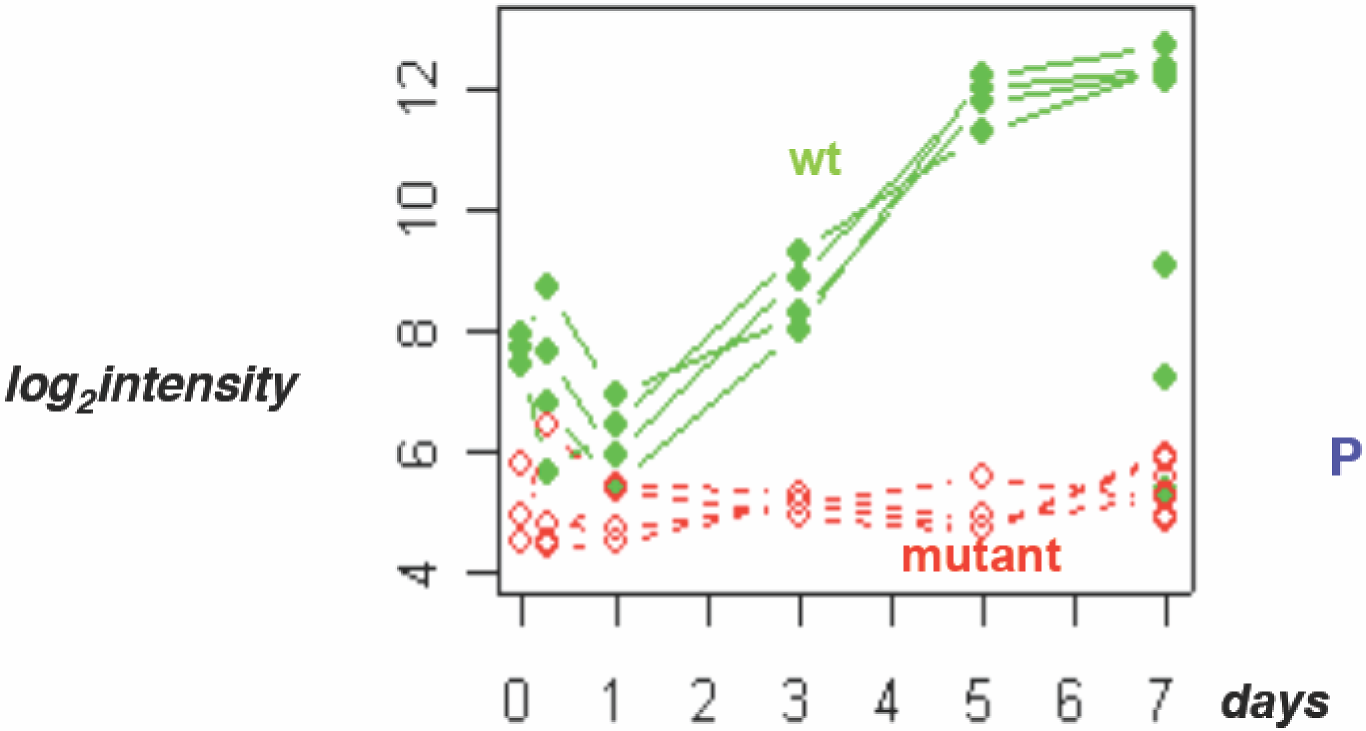
Replication in two experimental examples

1. A. thaliana (At) experiment:
    1. Two lines of plants: Columbia, Col-0=widetype (wt), and an enhanced disease susceptibility line eds 16 (mutant).
    2. Objective: to identify genes whose temporal expression patterns following infection differ between wt (Col-0) and mutant (eds 16).
    3. Three experiments-effectively biological replicates-were conducted using the wt and mutant lines, and within each, 3 technical replicate series. Not all have b een hybridized to chips. Later we use one series from experiments I and III, and two from experiment II.
    4. These experiments are longitudinal at the level of experiment, but cross-sectional at the level of mRNA sample (from separate leaves).

1. Replication in the OPC/OL experiment:
    1. Oligodendrocytes (OL) myelinate Central Nervous System axons...... and develop from migrating oligodendrocyte precursor cells (OPC).
    2. Broad purpose: to examine gene regulation in cultured oligodendrocyte precursor cells (OPC) as they develop into oligodendrocytes (OL). item Specific purpose: to identify a subset of genes with up-regulated time courses. Candidate genes predicted to be secreted will be assayed for their ability to cluster sodium channels along cultured retinal ganglion axons.
    3. 4 independent preparations were performed, each of which generated mRNA for every time point. We view this as 4 biological replicates of a longitudinal study. Again, it is not clearcut. For each biological replicate, a dye-swap pair of technical replicates was done.

Hybridizations for the At experiments

1. Initially we hybridized mRNA from just one of the technical replicate series from experiments I and III, and two from experiment II.
2. The Affymetrix Arabidopsis 24K GeneChip was used. In total of 2 (genotypes) × 6 (times) × 4 (experiments)=48 chips were hybed.
3. Preprocessing steps (background, normalization, probe set summarization) were done by RMA.
4. Question: find genes whose expression profiles differ between genotypes.

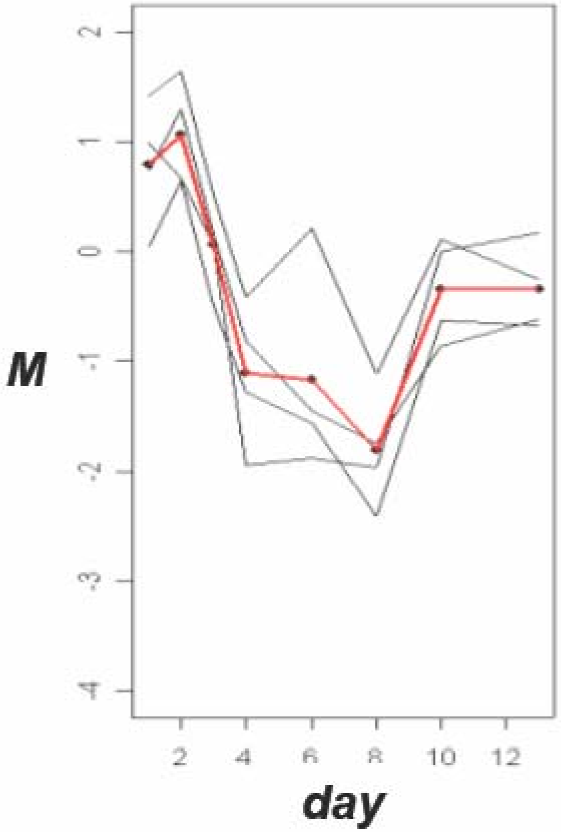


Profiles differ

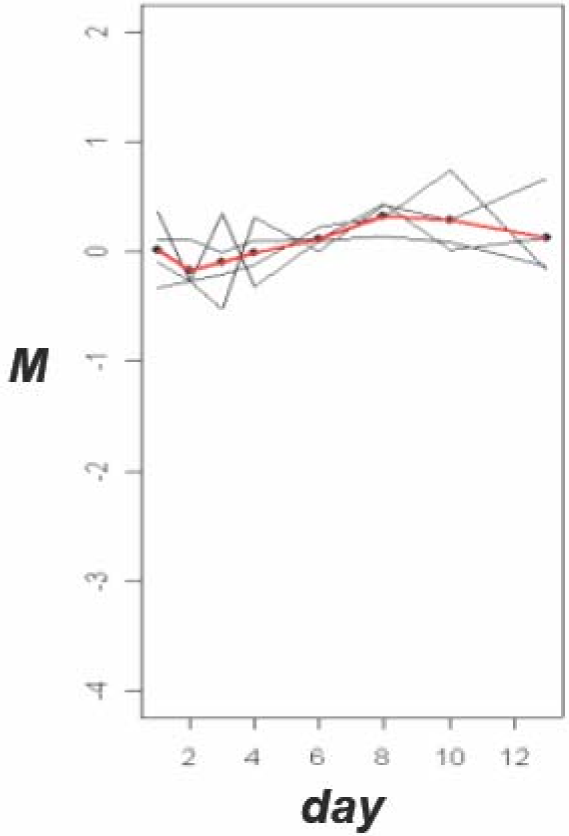
Extra dots at day 7 from uninfected samples.

Pooling in the OPC/OL experiment

1. Each prep has its own reference pool, which is the pool of all the individual time point mRNA samples of that prep.
2. Question: find genes whose expression levels change over time.



Change over time



No change over time

Approaches to solve the problems

1. Clustering
2. pairwise comparisons
3. ANOVA
4. Empirical Bayes methods

Custering with tiem-course data

1. Clustering methods have been widely used in this context to find groups of genes with interesting and similar patterns.
2. Hierarchical clustering: Eisen et al. (1998)
3. Self-organizing maps: Tamayo et al. (1999), Saban et al. (2001), Burton et al. (2002).
4. k-means clustering: Tavazoie et al. (1999)
5. Bayesian model-based clustering: Bar-Joseph et al. (2002, 2003), Ramoni et al. (2002)
6. HMM clustering: Schliep et al. (2003).

Drawbacks of clustering methods

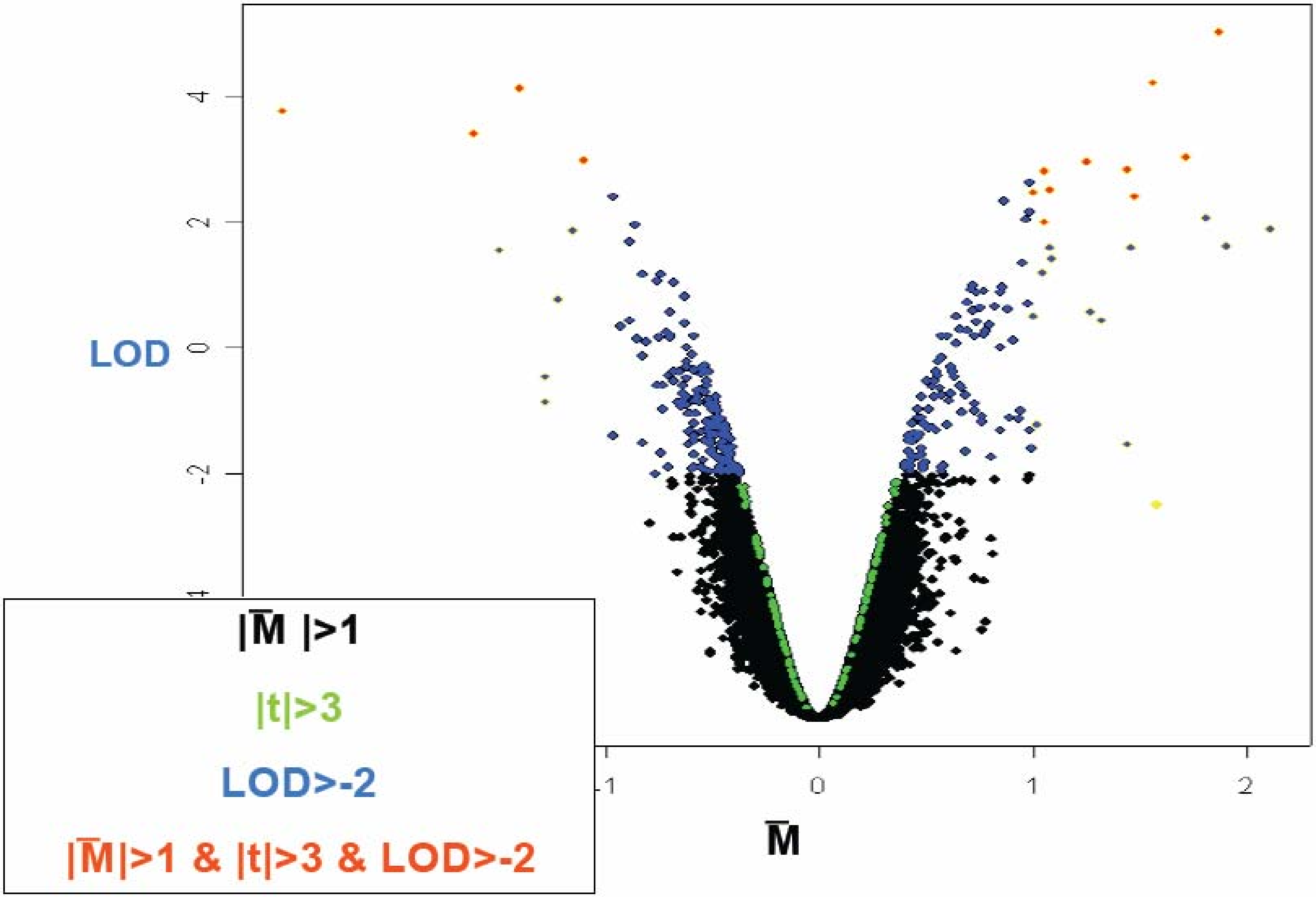
1. They make no explicit use of the replicate inforamtion. They either use all the slides or means of the replicates.
2. Clustering does not provide a ranking for the individual genes based on the magnitude of change in expression levels over time.
3. When the number of genes becomes large, clustering methods may not provide clear group patterns.
4. Cluster analysis may fail to detect changing genes that belong to clusters for which most genes do not change (Bar-Joseph et al. 2003).
5. Question: How many clusters?

Pairwise comparisons

1. One strategy is to make many or all univariate pairwise comparisons, e.g., of consecutive times: days 1 vs 2, 2 vs 3, etc.
2. Illustration on the OPC/OL data: t-tests, univariate posterior odds: the LOD statistic (Lonnsted et al., 2002; Smyth, 2004); the moderated t statistic (Smyth, 2004).
3. Moderated s^2 of m values: $s_*^2 = \frac{(n-1)s^2+\nu\lambda^2}{n-1+\nu}$
4. Moderated t: $t_*^2 = \frac{\bar{M}^2}{s_*^2/n}$
5. Log$_{10}$ of posterior odds against differential expresssion:

$$LOD = c + (\frac{n+\nu}{2}) log_{10} \frac{t_*^2 + n - 1 + \nu}{t_*^2 \frac{1/n}{1/n+1/k} + n - 1 + \nu}$$

OPC/OL experiment: day 6 vs. day 4



Drawbacks of pairwise comparisons

1. If involves a large number of tests for each gene, and there are over 10,000 genes in a typical microarray experiment: a two-way multiple testing problem.
2. Merging all the lists of genes can be difficult.
3. Cannot rank the genes according to the overall amount of change.

Drawbacks with ANOVA

1. ANOVA includes time as a factor.
2. This approach does not deal adequately with correlations across time.
3. An element of moderation just as with the t-statistic in the pairwise comparisons is desirable.

The goal in general

We prefer a formula to rank genes, in order to

1. Find those changing.
2. The formula should be (1) t-like or F-like; (2) multivariate; (3) moderated.
3. Moderation is required because variance and covariances are poorly estimated.
4. Some sort of smoothing, borrowing strength, or empirical Bayes approach is desirable to improve the identification of genes of interest.
5. Along the line of empirical Bayes approach, Tai and Speed (2006) used multivariate normals with conjugate priors, without the need of using MCMC.

Multivariate empirical Bayes approach

1. We denote by $X_{g,1}, \cdots, X_{g,n}$ the replicate random k-vectors representing the observed time series for a single gene.
2. For the At data, $n = 4$ and $k = 6$, and the $X_{g,i,t}$ are differences of log intensities, i.e. log ratios of mutant to wt.
3. For the OPC/OL data, $n = 4$ and $k = 8$, and the $X_{g,i,t}$ are log ratios of experimental to reference pool intensities.
4. the underlying model is that these $X_{g,i}$ are i.i.d. $N(\mu_g, \Sigma_g)$, and we make different assumptions about μ_g and Σ_g.

Hypotheses

1. With the At data, we are interested in testing the null hypothesis $H_g : \mu_g = 0, \Sigma_g > 0$, against the alternative $K_g : \mu_g \neq 0, \Sigma_g > 0$.
2. With the OPC/OL data, we are interested in testing the null hypothesis $H_g : \mu_g =$constant, $\Sigma_g > 0$, against the alternative $K_g : \mu_g$ not constant, $\Sigma_g > 0$.

1. The priors for μ_g and Σ_g are set to reflect the indicator status $I = I_g$ of the gene, where $I_g = 1$ if H_g is true, and $I_g = 0$ otherwise, i.e. if K_g is true.
2. We suppose that $Pr(I_g = 1) = p$, independently for every gene, for a hyperparameter p, $0 < p < 1$. (the subscripts g will be dropped afterwards.)
3. our prior for Σ is inverse Wishart with degrees of freedom ν and matrix parameter $(\nu\Lambda)^{-1}$, where $\Lambda > 0$ is positive definite. When we dealing with a variance σ^2, we use an inverse Gamma prior with analogous paramters λ and ν.
4. Our priors for μ will be different depending on whether $I = 0$ or $I = 1$. In all cases, it is multivariate normal and involves Λ.
5. Finally, the data $X_1, \cdots, X_n$ are supposed i.i.d. given I,Σ, and μ, with $X_i \mid I, \Sigma, \mu \sim N(\mu, \Sigma)$.

Summary results for the At experiment

1. The moderated S is:

$$\tilde{S} = [E(\Sigma^{-1} \mid S)]^{-1} = \frac{(n-1)S + \nu\Lambda}{n - 1 + \nu},$$

2. The moderated t-statistic is

$$\tilde{t} = n^{1/2}\tilde{S}^{-1/2}\bar{X}.$$

3. The posterior odds is

$$O = \frac{P(I = 1 \mid data)}{P(I = 0 \mid data)} = (\frac{p}{1-p})\frac{P(\tilde{t} \mid I = 1)}{P(\tilde{t} \mid I = 0)}$$

4. The multivariate B-statistic $MB = log_{10}O$.

Likelihood Ratio (LR) statistic

1. LR test simply tests the null H against the alternative K in the usual way. It is obtained

$$LR = 2(l_K^{max} - l_H^{max}) = n \log(1 + \frac{n}{n-1}\bar{X}^T S^{-1} \bar{X})$$

$$= n \log(1 + T^2/(n-1))$$

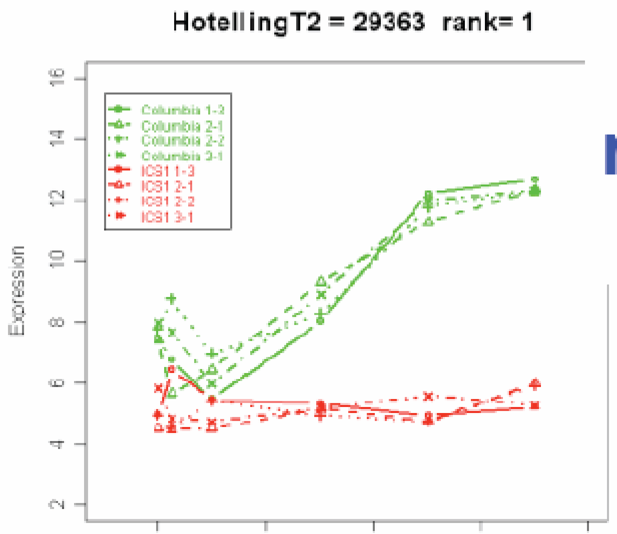
where S is assumed nonsingular. T^2 is Hotelling's statistic. We can plug in the moderated statistic $\tilde{T}^2$.

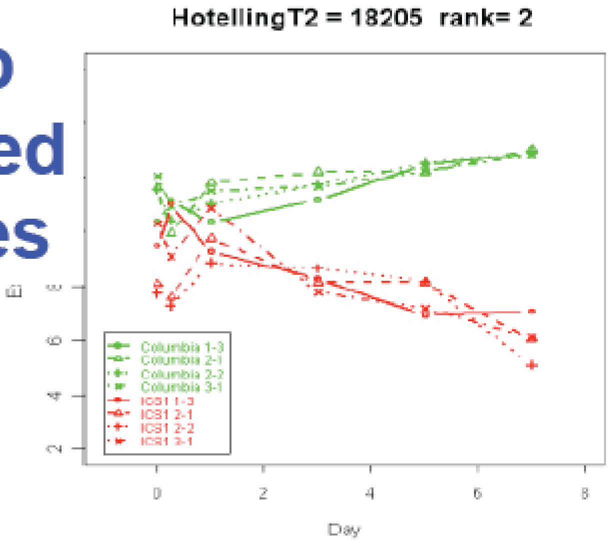
Hyperparameter estimation

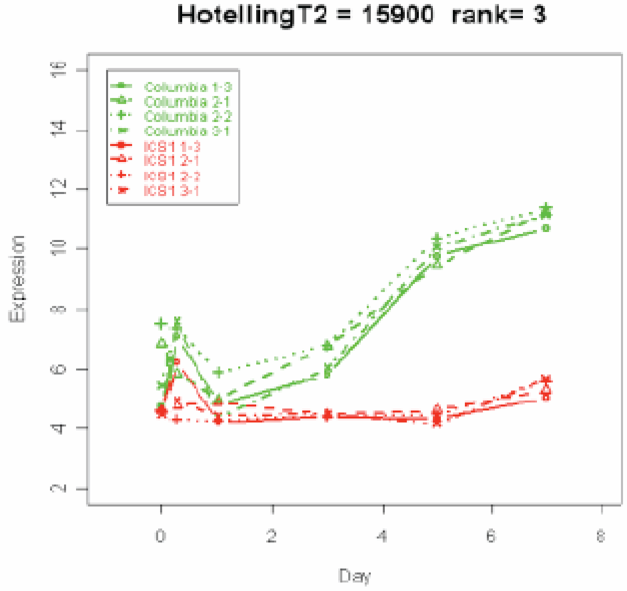
1. The prior $p = 0.02$.
2. Estimates of ν and η are developed using the univariate approach of Smyth (2004).
3. Λ is estimated by the method of moments using the formula

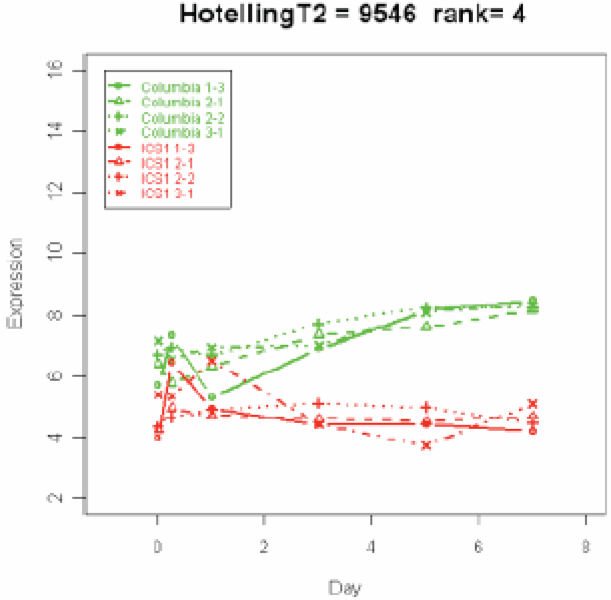
$$E(S) = (\nu - k - 1)^{-1}\nu\Lambda.$$

Top ranked genes


HotellingT2 = 29363 rank= 1


HotellingT2 = 18205 rank= 2


HotellingT2 = 15900 rank= 3


HotellingT2 = 9546 rank= 4

Conclusions

1. Methods which rank genes (e.g. the MB statistic or the moderated Hotelling T^2) perhaps provide easier access to genes whose absolute or relative expression varies over time, than do multi-gene methods (e.g. clustering).
2. Among the single-gene methods, MB performs no worse than other methods in both real data and simulated data comparisons, and better than the F.
3. Moderation is needed to improve the variance estimation.
4. The MB statistic may be able to select interesting genes which are missed by other methods.

1. The time course data can also be used to explore gene-gene association problems.

Exploration of gene-gene association

1. Cell cycle microarray experiments provide gene expression intensity data over time.
2. High dimensional time series analysis can be conducted to explore a wide range of problems. For example, periodicity (Filkov et al. 2002), gene-gene coexpression detection (Schafer and Strimmer, 2005), clustering of genes (Zhu et al. 2005a, 2005b).
3. We focus on exploring the gene-gene causal relationship. That is, if the expression of gene 1 is predictive of expression of gene 2 at a future time period.

Causality

1. Graphical models based on coherence/partial coherence. Brillinger (1996); Dahlhaus (2000); Butte et al. (2001); Salvador et al. (2005).
2. Drawback: sensitive to measurement noise (Albo et al., 2004). incapability of detecting time precedence relationship (Baccala and Sameshima, 2006).
3. Granger causality is good for detection of causalities in stationary/nonstationary time series (Winterhalder et al, 2005; Mukhopadhyay and Chatterjee, 2006).

Granger causality

1. Assuming the following autoregressive model:

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

   where $\epsilon_t \sim N(0, \sigma^2)$, $t = 1, ..., n$.

2. Granger causality can be tested using a vector autoregressive (VAR) (Hamilton, 1994; Mukhopadhyay and Chatterjee, 2006). Specifically, F test is used to test

   $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0.$$

Disadvantages

1. Required distributional assumption: $\epsilon_t \sim N(0, \sigma^2)$.
2. Real application: (i) **Nonhomogeneous errors**; (ii) **Not normal distributed**.
3. It could lead to inconsistent results.
4. Only for least squared estimate.

Why inconsistent?
The F test depends on the consistent variance estimation.
Example 1. If $q = 1$, $c = \alpha_1 = 0$, then the model is

$$y_{1(t)} = \beta_1 y_{2(t-1)} + \epsilon_t,$$

where $E\epsilon_t = 0$ and $Var(\epsilon_t) = \sigma_t^2$.

$$\hat{\beta}_1 = \frac{\sum_{t=1}^n y_{1(t)} y_{2(t-1)}}{\sum_{t=1}^n y_{2(t-1)}^2}$$

$$= \beta_1 + \frac{\sum_{t=1}^n y_{2(t-1)} \epsilon_t}{\sum_{t=1}^n y_{2(t-1)}^2}$$

The variance estimation of $\hat{\beta}_1$ is,

$$\hat{Var}(\hat{\beta}_1) = \frac{\sum_{t=1}^{n} e_t^2}{n \sum_{t=1}^{n} y_{2(t-1)}^2},$$

where e_t are the residuals, with

$$e_t = y_{1(t)} - \hat{\beta}_1 y_{2(t-1)}.$$

For homogeneous errors, $\sigma_1 = ... = \sigma_n = \sigma$, Then the variance is

$$Var(\hat{\beta}_1) \sim \frac{\sigma^2}{\sum_{t=1}^{n} y_{2(t-1)}^2}$$

In this case,

$$\hat{Var}(\hat{\beta}_1) - Var(\hat{\beta}_1) \rightarrow 0.$$

For nonhomogeneous errors, σ_t are different, then

$$Var(\hat{\beta}_1) \sim \frac{\sum_{t=1}^{n} y_{2(t-1)}^2 \sigma_t^2}{(\sum_{t=1}^{n} y_{2(t-1)}^2)^2}.$$

However,

$$\hat{Var}(\hat{\beta}_1) \rightarrow \frac{\sum_{t=1}^{n} \sigma_t^2}{n \sum_{t=1}^{n} y_{2(t-1)}^2}.$$

Therefore, the variance estimate $\hat{Var}(\hat{\beta}_1)$ is **not consistent.**

Motivation

1. We proposed a test based on estimating equations.
2. Applicable to a wide class of estimate: Least squared estimate; M-estimate; General linear model based estimate; L_1 and quantile based estimate; etc.
3. Relax of the distributional assumption: only need ϵ_t to be independent. No distribution assumption and for nonhomogeneous errors.
4. Computationally implementable.

Inference Based on Estimating Equation

For simplicity, we consider two genes with observations as $y_{1(0)}, ..., y_{1(n)}$ for the first gene and $y_{2(0)}, ..., y_{2(n)}$ as the second gene.
Assume the following autoregressive model holds:

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

where ϵ_t, $t = 1, ..., n$, are independent and satisfy $E\epsilon_t = 0$ and q is the autoregressive lag length.

The gene two is said to *Granger cause* gene one if $\beta_i \neq 0$ for at least one i. The main statistical problem is to test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0.$$

Let $Y_{1(t)} = (y_{1(t)}, ..., y_{1(t-q)})$ and $Y_{2(t)} = (y_{2(t)}, ..., y_{2(t-q)})$.
$\theta = (c, \alpha, \beta)$.

General we consider that the estimate of θ is from following estimating equation

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

for some given function g satisfying

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

For the estimating equation, see Boos (1992), Godambe and Kale (1992), Liang and Zeger (1986), Hu and Kalbfleisch (2000) for details. In literature, inference based on estimating equation is mainly for **independent data**. Here we consider **dependent Microarray data**.

When the least square method is used to estimate the parameters, then the estimating function

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

$$[y_{1(t)} - (c + \alpha_1 y_{1(t-1)} + \ldots + \alpha_q y_{1(t-q)} + \beta_1 y_{2(t-1)} + \ldots + \beta_q y_{2(t-q)})].$$

which is a $2q + 1$ vector.

In Example 1, the corresponding estimating equation is

$$\sum_{t=1}^{n} y_{2(t-1)}(y_{1(t)} - \beta_1 y_{2(t-1)}) = 0.$$

To do the inference of θ, we need following notations. Let

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

and

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

Both $V(\theta)$ and $W(\theta)$ are $(2q+1) \times (2q+1)$ matrix.

In application, we use the following two estimates for given θ, which is

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

Where

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

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

By the law of large number, for both **homogeneous and nonhomogeneous errors**,

$$V(y,\theta) \to V(\theta)$$

and

$$W(y,\theta) \to W(\theta).$$

Now we consider a general testing problem

$$H_0 : h(\theta) = 0$$

for some differentiable function h. Let r be the dimension of h. Let

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

To test $H_0 : h(\theta) = 0$, we define

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

and its general inverse (Moore-Penrose) is then

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

Let $\tilde{\theta}$ be the estimate of θ under the restriction $h(\theta) = 0$. The test statistics is then

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

We can show that $Q_{h=0}$ is a chi-square distribution with degree of freedom r under H_0. We will reject H_0, if

$$Q_{h=0} > \chi^2_{r,\alpha}.$$

In our application,

$$H_0 : \beta_1 = ... = \beta_q = 0$$

It is very easy to calculate the matrices:

$$H(\theta),\ U(\theta),\ U^{-}(\theta),\ \text{and } Q_{h=0} = S(y, \tilde{\theta})^T U^{-}(\tilde{\theta}) S(y, \tilde{\theta}).$$

$Q_{h=0}$ is a chi-square distribution with degree of freedom q under H_0. We will reject H_0, if

$$Q_{h=0} > \chi^2_{q,\alpha}.$$

Simulation studies

1. We generated two variables X_1 and X_2 where $x_{1t} = 0.7x_{1(t-1)} + \epsilon_{1t}$ and $x_{2t} = 0.3x_{2(t-1)} + \beta x_{1(t-1)} + \epsilon_{2t}$.
2. The interest is to test the null hypothesis $H_0 : \beta = 0$.
3. We conducted 5000 simulations with the data generated under null hypothesis H_0.
4. We considered the number of time points to be $n = 30$ or 100.

Table: Type-I error rate results in simulation study.

n	Method	Homogenous ϵ_{2t}			Non-homogenous 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

1. In addition, we made comparisons of the distribution of the test statistic $\tilde{Q}_{h=0}$ to its theoretical distribution $\chi^2(1)$.

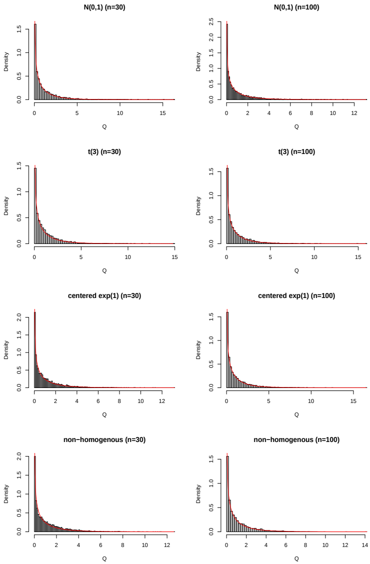


Figure: Distribution of $\tilde{Q}_{h=0}$ versus $\chi^2(1)$.

1. We simulated a network of 14 genes in various cases.
2. Independent genes are x_1,x_7,x_8,x_9,x_{11}, and x_{14}. All are AR(1) processes with autocorrelation< 1.
3. All the other genes are generated from dependent series, with lag 1 and autocorrelation< 1.
4. The series are generated at 100 equidistant time points.

1. The gene-gene dependent series:
$$x_{2t} = 0.29x_{2(t-1)} + 0.65x_{1(t-1)} + \epsilon_t;$$
$$x_{3t} = 0.15x_{3(t-1)} + 0.29x_{2(t-1)} + 0.65x_{14(t-1)} + \epsilon_t;$$
$$x_{6t} = 0.12x_{6(t-1)} + 0.3x_{7(t-1)} + 0.3x_{8(t-1)} + 0.3x_{9(t-1)} + \epsilon_t;$$
$$x_{4t} = 0.17x_{4(t-1)} + 0.4x_{3(t-1)} + 0.7x_{6(t-1)} + \epsilon_t;$$
$$x_{5t} = 0.6x_{5(t-1)} + 0.8x_{4(t-1)} + \epsilon_t;$$
$$x_{10t} = 0.4x_{10(t-1)} + 0.3x_{11(t-1)} + \epsilon_t;$$
$$x_{12t} = 0.4x_{12(t-1)} + 0.4x_{11(t-1)} + \epsilon_t;$$
$$x_{13t} = 0.4x_{13(t-1)} + 0.4x_{11(t-1)} + \epsilon_t.$$

1. The independent series (stationary):
$$x_{1t} = 0.7x_{1(t-1)} + \epsilon_t;$$
$$x_{7t} = 0.8x_{7(t-1)} + \epsilon_t;$$
$$x_{8t} = 0.7x_{8(t-1)} + \epsilon_t;$$
$$x_{9t} = 0.77x_{9(t-1)} + \epsilon_t;$$
$$x_{11t} = 0.7x_{11(t-1)} + \epsilon_t;$$
$$x_{14t} = 0.65x_{14(t-1)} + \epsilon_t;$$

1. The independent series (non-stationary):
$x_{1t} = sin\frac{\pi t}{40} + 0.7x_{1(t-1)} + \epsilon_t$;
$x_{7t} = 0.8x_{7(t-1)} + \epsilon_t$;
$x_{8t} = cos\frac{\pi t}{40} + 0.7x_{8(t-1)} + \epsilon_t$;
$x_{9t} = 0.77x_{9(t-1)} + \epsilon_t$;
$x_{11t} = cos\frac{\pi t}{40} + 0.7x_{11(t-1)} + \epsilon_t$;
$x_{14t} = 0.65x_{14(t-1)} + \epsilon_t$;

1. Several homogeneous residual distributions.
    1. $\epsilon_t \sim N(0, 1)$.
    2. $\epsilon_t \sim$ centered *exp*(1), representing an asymmetric distribution case.
2. We applied algorithms at time points $t =$10, 20, 40, 60, 80, and 100.
3. Between a pair of genes, we only consider the direction with lower p-value of the test for causality.

1. Accuracy summary statistics (Mukhopadhyay and Chatterjee, 2006): Let G denote the number of genes, $A = 100[1 - \frac{2}{G(G-1)} \sum_{e_i} I(e_i)]$.
2. For each edge e_i that connects two genes, $i = 1, \cdots, \frac{G(G-1)}{2}$, $I(e_i) = 0$ if the conclusion drawn from a test if a edge is statistically significant agrees with the truth; $I(e_i) = 1$ otherwise.
3. 100 simulations were performed. And the threshold of p-value=0.01 is used to take into account multiple testing issue.
4. The average accuracy summary statistics (together with standard error) over the simulations are reported next.
5. we also record the number of false positives (FP, the detected edges that are not true edges); and the number of false negatives (FN, the true edges that are not detected).

Table: Homogeneous ϵ_t case 1 with stationary and non-stationary independent series.

$\epsilon_t \sim N(0,1)$	Stationary						
Method	time	10	20	40	60	80	100
F	Avg	89.45	89.78	92.26	93.48	94.12	94.70
	SE	2.48	2.46	2.81	2.46	2.44	2.34
	FP	2.88	3.04	2.73	2.94	3.12	3.32
	FN	6.72	6.26	4.31	2.99	2.23	1.50
EE	Avg	89.54	88.79	91.64	92.75	93.69	94.54
	SE	2.97	2.55	2.60	2.55	2.26	2.24
	FP	2.64	2.61	2.38	2.88	3.10	3.24
	FN	6.88	7.59	5.23	3.72	2.64	1.73
	Non-stationary						
Method	time	10	20	40	60	80	100
F	Avg	89.16	89.70	91.37	88.86	87.53	84.58
	SE	2.82	2.94	3.07	4.04	4.05	4.92
	FP	2.88	3.36	5.01	8.18	9.92	12.91
	FN	6.98	6.01	2.84	1.96	1.43	1.12
EE	Avg	88.90	88.30	90.40	88.84	87.86	84.95
	SE	2.55	3.08	2.70	4.47	4.02	4.99
	FP	2.65	3.13	4.75	7.69	9.51	12.37
	FN	7.45	7.52	3.99	2.47	1.54	1.33

Table: Homogeneous ϵ_t case 2 with stationary and non-stationary independent series.

$\epsilon_t \sim exp(1)$ (centered)	Stationary						
Method	time	10	20	40	60	80	100
F	Avg	89.46	90.29	92.59	93.90	94.49	94.89
	SE	2.64	2.34	2.33	2.66	2.53	2.31
	FP	2.99	3.05	2.86	2.89	2.92	3.25
	FN	6.60	5.79	3.88	2.66	2.09	1.40
EE	Avg	89.49	89.73	91.05	93.21	94.47	95.16
	SE	2.46	2.41	2.72	2.36	2.24	2.21
	FP	2.15	2.13	2.38	2.17	2.33	2.47
	FN	7.41	7.22	5.76	4.01	2.70	1.93
	Non-stationary						
Method	time	10	20	40	60	80	100
F	Avg	89.29	89.68	90.59	88.64	87.43	85.12
	SE	2.85	2.58	3.19	4.18	4.14	4.72
	FP	2.94	3.45	5.78	8.45	9.98	12.45
	FN	6.81	5.94	2.78	1.89	1.46	1.09
EE	Avg	89.20	89.04	89.92	88.89	87.62	85.10
	SE	2.66	2.51	3.04	3.90	4.02	4.54
	FP	2.14	2.57	4.84	7.33	9.21	11.94
	FN	7.69	7.40	4.33	2.78	2.06	1.62

Next, we examined the stationary data with non-homogenous normal distribution case where the variance of ϵ_{it} in the model of predicting gene i is associated with expression intensities of other interacting genes at the previous time points.

Table: Standard deviation of ϵ_{it}.

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

Table: 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

A real example

1. We used the human cancer cell cycle data (Whitfield et al., 2002) available at http: //genome-www.stanford.edu/Human-CellCycle/Hela.
2. Li et al. (2006) studied gene regulatory network on 20 genes, represented by 23 probe sets, using one experiment in this data which consisted of 48 time points.
3. We focus on the same set of genes.

1. At the p-value threshold of 0.001, F detected 104 gene pairs that showed significant causal relationship while EE detected 78. And 70 pairs were detected by both.
2. The performances of two methods should be similar if expression intensities of a gene to be predicted is homogenously and normally distributed.
3. It motivates checking the distribution of expression intensities of each individual gene.
4. We observed that the majority of genes are nearly normally distributed except for 6 genes.
5. Observation: the most difference in the results between F and EE is associated with these 6 genes.

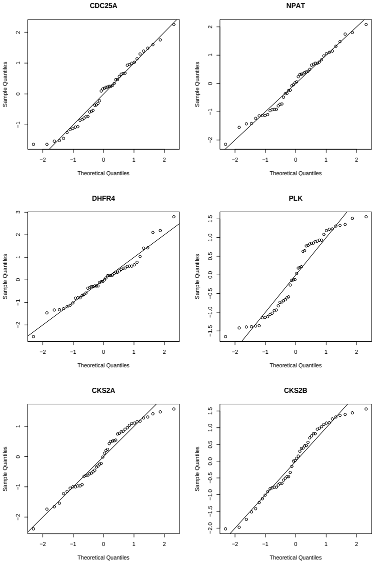


Figure: Q-Q plots of the expression intensities of 6 genes.

1. One interesting observation: EE detected 3 genes that strongly predict gene *DHFR4* while F could not detect any.
2. We used the gene pair (*NPAT*, *DHFR4*) as an example where EE obtained the p-value of 0.0002 while F is 0.013.
3. We standardized the expression profile of gene *DHFR4* with a monotonic normal score transformation.
4. Specifically, we obtain the ranks of gene expression intensities $R_1, \cdots, R_n$ and use them to construct the transformed profile, $\Phi^{-1}(R_1/(n+1)), \cdots, \Phi^{-1}(R_n/(n+1))$, where $\Phi(.)$ is the cumulative normal distribution.
5. On the transformed profile of *DHFR4*, EE and F yielded the p-values of 0.00005 and 0.0003, respectively.

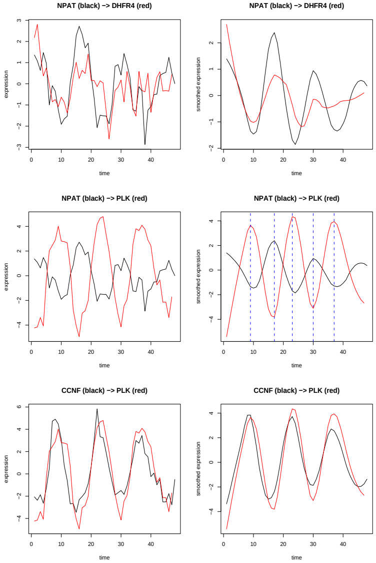


Figure: expression intensities along time for 3 gene pairs.

1. The second gene pair example: gene *NPAT* is tested to predict *PLK* which has the non-normal distribution.
2. F claimed the association to be significant with p-value of 0.0005 while EE yielded nonsignificant result (p-value=0.003).
3. The normal score transformation resulted in the p-value of 0.002 using F and p-value of 0.006 using EE.

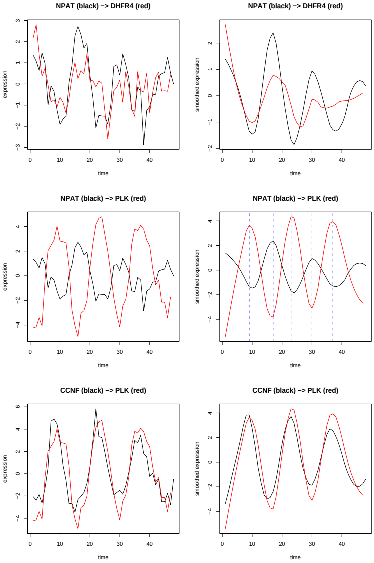


Figure: expression intensities along time for 3 gene pairs.

1. In the third example, the gene-gene causality relationship between *CCNF* and *PLK* is detected by both methods.
2. Normality transformation on *PLK* only slightly altered the significance levels of both F test and EE test.
3. This relationship has been validated in Li et al. (2006).

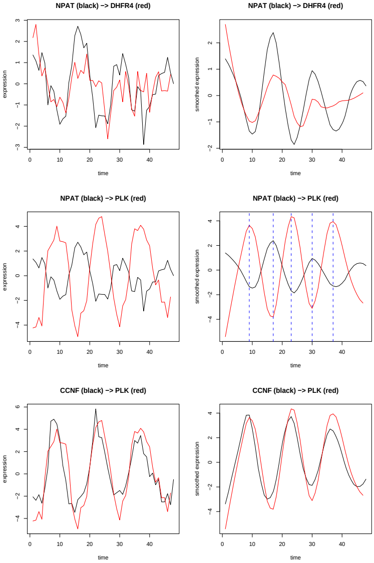


Figure: expression intensities along time for 3 gene pairs.

Remarks

1. The proposed test based on estimation equation has the appealing theoretical property: valid and consistent regardless of the data distribution form.
2. The estimation equation method is computationally simple.
3. The possible un-equidistant time points can be taken into account.
4. Missing data problem is often seen in the real data.
5. Nonlinear causality???

References

1. Li (2002). Genome-wide coexpression dynamics: theory and application. PNAS 99, 16875-16880.
2. Albo et al. (2004). Is partial coherence a viable technique for identifying generators of neural oscillations? Biological Cybernetics 90, 318.
3. Winderhalder et al. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. Signal processing, 85, 2137-2160.
4. Mukhopadhyay and Chatterjee (2007). Causality and pathway search in microarray time series experiment. Bioinformatics, 23, 442-449.
5. Hu and Kalbfleisch (2000). The estimating function bootstrap (with discussions). Canadian Journal of Statistics, 28, 449-499.