Time-Course Data Prediction for Repeatedly Measured Gene Expression

Atanu Bhattacharjee
Assistant Professor
Centre for Cancer Epidemiology
Tata Memorial Centre,India

29th November 2018

Contents

Introduction

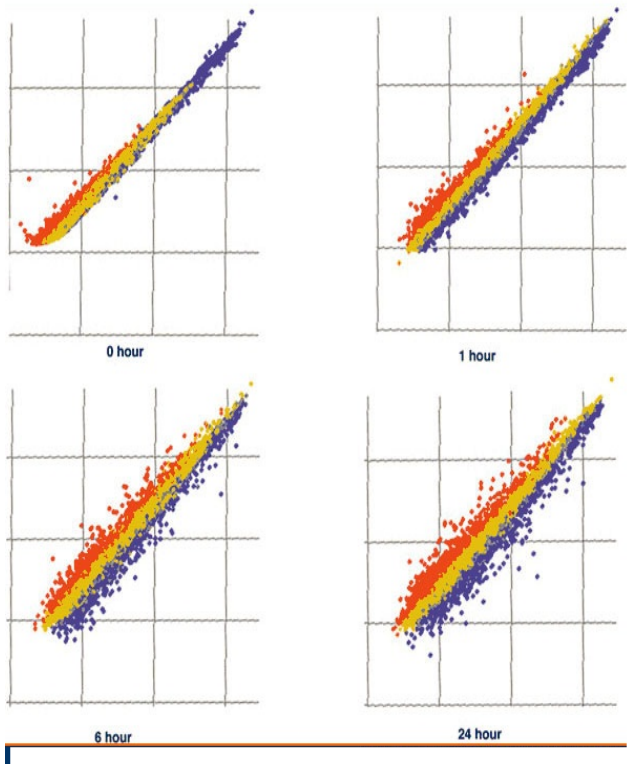
There is a rapid success in gene expression microarray data in last two decades.

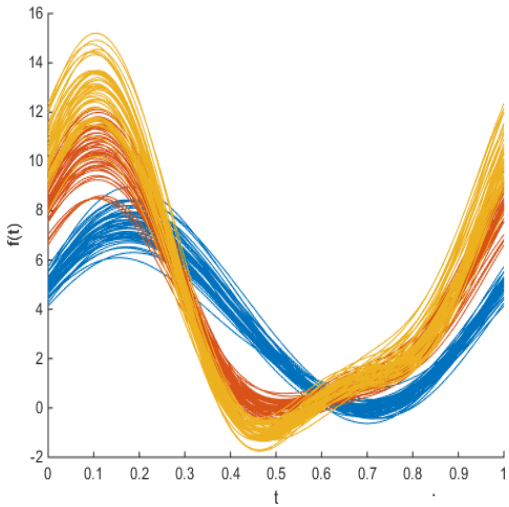
It is a procedure where DNA molecules are fixed into an order and their locations are defined as probes or spots.

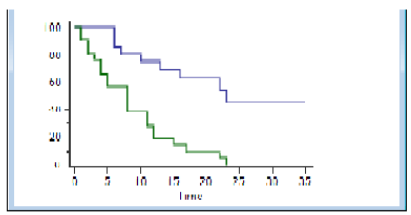
The microarray provides expression of many genes in a single and efficient manner.

It generates with big data.

Introduction







Introduction

Let S be a gene of interest. We start by the case of a one group experiment, where each patient act as her/his own respective control, her/his condition changing over time. The expression of genes inside S is modeled over time according to a function f as: for all the genes $g \in S$,

$$y_{git} = \mu + \beta_g + c_{gi} + f_g(t_i) + \epsilon_{gi} \quad (1)$$

Introduction

Every time coefficient of the trend $f_g(t_i)$ is actually divided into a fixed effect η(representing the average trend in the gene set S) and a random effect $h_{g,\cdot} \sim N(0, \sigma_h)$ grouped on the gene g, accounting for the possible heterogeneity between the genes in the gene set S.

Challenge

The presences of high variability in gene expression microarray data is obvious. But the presences of variability prepares the data more complex and interesting procedure.

The application of statistical analysis is challenging to deal with large numbers of genes data set with multi-collinearity and high correlations gene data structure considered for temporal changes.

The prediction of future time point gene expression values is an unexplored area.

Data Methodology

The proposed extensions of the structural model for time-course microarray studies were applied to real data set, publicly available in the GEO (Gene Expression Omnibus) database.

Six healthy subjects (1 female and 5 male) were selected for the study.

Bacterial endotoxin (CC-RE, lot2) was intravenously administrated to study subject with same dosage level. 0.09% Sodium Chloride was used as Placebo treatment.

The primary investigation was conducted on 3 time points (0 hrs,2 hrs and 6 hrs) in this study for modeling.

Statistical Methodology

It is assumed that

$$y_{gpi} \sim \text{Normal}(\mu_g, \sigma_g^2) \tag{2}$$

Now,

$$y_{gi} = (y_{gi1}, ..., y_{git})' \tag{3}$$

The $\hat{\mu}_g$ and $\hat{V}_g$ are the mean and variance estimates respectively.

Our aim is to estimate $\hat{\mu}_g$ and $\hat{V}_g$ about prediction on future observations.

Statistical Methodology

Let the model is defined as

$$Y = X_{t\times v}\tau_{v\times k}S_{k\times N} + \epsilon_{t\times N} \quad (4)$$

The parameter S and X are design matrix and τ is known. The rank of S and X are v<N and v<t respectively. The factor ϵ is assumed to follow Normal $(0, \sum)$. The factor ϵ is having $t-$variants. Let the term t is observation time for replicates.The terms k,N,v and t are considered to represent replicates of gene expression measured at different time points.

The intention is to obtain the predicted value of gene expression of Y through the prior measurement of X.The covariance structure is defined as $\hat{C}$. Now, $\hat{C} = (c_{ij}), c_{ij} = \rho_{|i-j|}, i \neq j, c_{ii} = 1$ for $i, j = 1, ...t, \sigma^2 > 0$ and $\rho_{ij} < 1$ subject to $\hat{C}$ being positive definite.

Statistical Methodology

The aim of this study is to estimate the future time prediction value of y i.e. time ($T^{(2)}$). The parametric estimation of y and ($T^{(1)}$) is utilized to generate the estimates about time ($T^{(2)}$).

The term τ is estimated as

$$\tau^2 = Q^{-1}(Q_1\hat{\tau}_1 + Q_2\hat{\tau}_2) \tag{5}$$

$$\sigma^2 = \frac{A + B + C}{tN + t_1K} \tag{6}$$

The terms A,B and C are defined as,

$$A = (\hat{\tau}_1 - \hat{\tau}_2)' Q_1 Q^{-1} Q_2 (\hat{\tau}_1 - \hat{\tau}_2) + tr(X' \hat{C}^{-1} X)^{-1} X' \hat{C}^{-1} S \hat{C}^{-1} X$$

Statistical Methodology

$$B = tr(Z^{'}\hat{C}Z)^{-1}Z^{'}YY^{'}Z + tr(X^{(1)^{'}}\hat{C}_{11}^{-1}X^{(1)})^{-1}X^{(1)^{'}}\hat{C}_{11}^{-1}S_1\hat{C}_{11}^{-1}X^{(1)} \tag{7}$$

$$C = (Z_1^{'}\hat{C}_{11}Z_1)^{-1}Z_1^{'}Z_1^{'}V^{(1)^{'}}Z_1 \tag{8}$$

and

$$Q_1 = AA^{'}(X^{'}\hat{C}^{-1}X), Q_2 = FF^{'}(X^{(1)}\hat{C}_{11}^{-1}X^{(1)}) \tag{9}$$

$$Q = Q_1 + Q_2 \tag{10}$$

$$\hat{\tau}_1 = (X^{'}\hat{C}^{-1}X)^{-1}X^{'}\hat{C}^{-1}YA^{'}(AA^{'})^{-1} \tag{11}$$

$$\hat{\tau}_2 = (X^{(1)}\hat{C}_{11}^{-1}X^{(1)})^{-1}X^{(1)^{'}}\hat{C}_{11}^{-1}V^{(1)}F^{'}(FF^{'})^{-1} \tag{12}$$

Statistical Methodology

The mean estimation of $(\hat{T}^{(2)})$ is observed from the distribution of $(T^{(2)})$ given $(\hat{T}^{(1)})$ and Y as

$$\hat{T}^{(2)} = X^{(2)}\hat{\tau}F + \sum_{21}\sum_{11}^{-1}(\hat{T}^{(1)} - X^{(1)}\hat{\tau}F) \tag{13}$$

where $X = (X^{(1)'}, X^{(2)'})', \hat{\sum} = \hat{\sigma}^2\hat{C} = (\hat{\sum}_{ij}), \hat{\tau}$ and σ^2 are detailed above, ,$X^{(i)}$ is $p_i \times m$, and $\hat{\sum}_{ij}$ is of dimension of $p_i \times p_j, p_1 + p_2 = p$.

Statistical Methodology

The likelihood of τ, σ and ρ is defined as

$$L(\sigma^2, \tau, \rho | Y) \propto \sigma^{-tN} |C|^{-\frac{N}{2}} \exp\{-\frac{1}{2\sigma^2} \text{trC}^{-1}[Y-(X,E)_0^{\tau}A][Y-(X,E)_0^{\tau}]'\} \tag{14}$$

The non-informative prior is used to generate the posteriors with

$$g(\tau, \sigma^2, \rho) \propto \frac{1}{\sigma^2} \tag{15}$$

It is assumed that τ, σ^2 and ρ have independent prior distribution and with no available information about the parameters . Hence the posterior density of $\tau, \sigma^2$and ρ given Y is

$$P(\sigma^2, \tau, \rho | Y) \propto \sigma^{-(tN+2)} |C|^{-\frac{N}{2}} \exp\{-\frac{1}{2\sigma^2} trC^{-1}[Y-(X,E)_0^{\tau}A][Y-(X,E)_0^{\tau}A]'\} \tag{16}$$

Statistical Methodology

The equation can be summarized as

$$p(\tau, \rho|Y) \propto |C|^{-\frac{N}{2}} S_1^{-\frac{tN}{2}} \tag{17}$$

where

$$S_1 = tr(X^{'} C^{-1} X)^{-1} (\hat{\tau} - \hat{\tau}_0) A A^{'} \hat{\tau} + b_0 (18)$$

$$b_0 = tr(X^{'} C^{-1} X)^{-1} X^{'} C^{-1} S C^{-1} X + tr(E^{'} C E)^{-1} E^{'} Y Y^{'} E \tag{19}$$

and $\hat{\tau}_0$ is same as $\hat{\tau}_1$ with $\hat{C}_1$ replaced by C.

Statistical Methodology

$$P(\tau|Y,\rho) = tr(\hat{\tau}_0, AA^{'}, b_0, X^{'}C^{-1}X, tN) \tag{20}$$

Thus the approximately, the posterior distribution of τ can be obtained from the following inequality:

$$\hat{b}_0^{-1} tr(X^{'}C^{-1}X)(\tau - \hat{\tau}_0^*)AA^{'}(\tau - \hat{\tau}_0^*) \leq \frac{vk}{tN - vk} F_\alpha(vk, tN - vk) \tag{21}$$

Result

The estimates of $\hat{\tau}_0^*, \hat{b}_0^*, \hat{C}$ are measured through $\hat{\tau}_0, \hat{b}_0$ and C at $\rho = \hat{\rho}$ and $F_\alpha(v_1, v_2)$ is upper 100α percent point of the F distribution.

The posterior estimates of gene expression values for 6th patients 6 hours time point estimates fitted by MCMC sampling.

The estimates is obtained through observed value of total 5 patients at 3 time points observation i.e.(0hours,2hours and 6hours) and 6th patients (0hours,2hours) observations.

Result

The precision of priors is assumed to follow as $\tau \sim \text{dgamma}(0.0001, 0.0001)$.

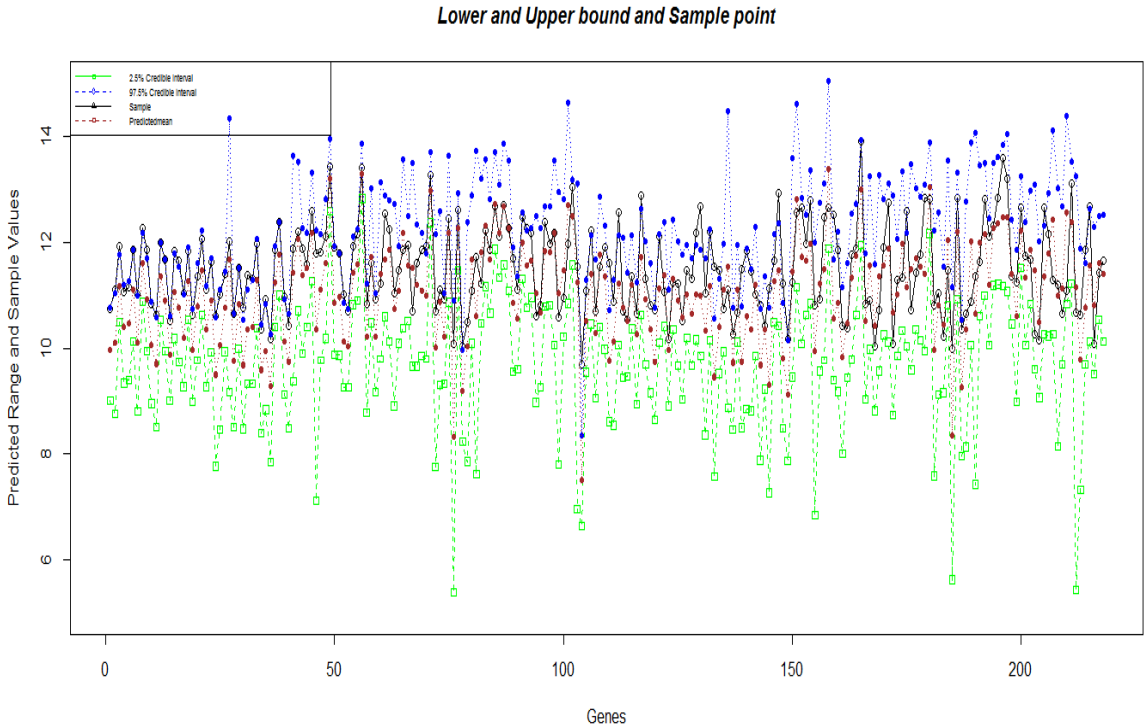
The distance matrix is assumed to follow $\text{dnorm}(\text{gene}_{ij}, \tau)$ as gene_{ij} gives the ith genes j^{th} time point measurements.

Result

The variation is assumed through

$$\sigma \sim \sqrt{(\frac{1}{\tau^2})} \tag{22}$$

The deviation estimate i.e. $\frac{(expected - observed)^2}{\text{expected}}$ is presented by acuracy plot.


Lower and Upper bound and Sample point
2.5% Credible interval
97.5% Credible interval
Sample
Predictedmean
Predicted Range and Sample Values
6
8
10
12
14
0
50
100
150
200
Genes

Result

This analysis was performed using R2OpenBugs.

The predicted mean, credible intervals $(2.5\%, 97.5\%)$ and observed values for 218 genes are presented.

It shows that the predicted mean values and observed range are in the range with credible intervals.

Result

The precision (reciprocity of the variance) was used in the normal distribution (dnorm), and later on transformed back to variance to generate posterior estimation. All starting values of parameters were put together as list( ).

The mean score at the first occasion of first individuals as the staring value is adopted.

Result

The prediction of specific time point of specific individual is formulated through consideration of data at that time point as "NA" in OpenBUGS.

Convergence is checked visually through traceplots and by using several starting points for each gene.The 20,000 posterior samples are thinned to 100 for estimating the posterior distribution of quantities of interest.

Result

For all genes, the MC errors for these quantities are less than 0.05%. For this full model processing 3924 data points, 20000 iterations take 3 hours 35 minutes on a dual processor 2.4 GHz machine running version 1.4 of OpenBUGS under Window Vista. It is aimed to develop faster, idea build code.

Conclusion

This study is dedicated to explore the application of longitudinal data analysis in gene expression data prediction.

The intention is about prediction of how the process will evolve in the future.


Thank
you