Dynamic Analysis Of High Dimensional Microarray Time Series Data Using Various Dimensional Reduction Methods

Abstract
The research methodology includes two research tasks. Firstly, applying several dimension reduction methods on two microarray data sets. Secondly, applying the sparse vector autoregressive from time series microarray experiment, and performing 1-step ahead forecast to compare models with regards to performance.

Analyzing Microarray Data
RNA is extracted from the cells, labeled with different colors. DNA is hybridized on the glass slides.

Challenges In Analyzing Time Series Microarray Data
Challenges in using dimension reduction techniques for microarray time series data
- Microarray data are noisy
- The system is underdetermined
- High correlation between genes
- Select a single technique
- Select an efficient model for such data
High quality dimension reduction technique
- Least squares estimators have large variances

Motivation
Significance of dimension reduction techniques in microarray data
- Data laying down on a low-dimensional manifold
- Monitoring a gene’s behavior at multiple points over time
- Preserving almost all of the original information
- A more quantitative knowledge of dynamic regulatory gene
- Computational cost will significantly reduce
- Provides an understanding of gene activities
- It can provide an understanding of drug effectiveness over time
Research Objectives
- Applying various dimension reduction methods
- Producing orthogonal variables
- Resulted in reasonable least square estimation with low standard errors
- With an outstanding performance
- A low Dimensional manifold

Application Of Various Dimension Reduction Techniques
- Pre processing step
  - Filling missing values
  - Normalization
  - Log-transformation
- The second data used is Drosophila lifecycle gene expression data (Arbeitman et al., 2002).
- The first data set is a Human cancer cell line (HeLa) cell cycle gene expression dataset collected by Whitfield et al., 2002.

Results

R-squared value for and HeLa data

<table>
<thead>
<tr>
<th>Dimension reduction Method</th>
<th>Number of genes</th>
<th>Method</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>500</td>
<td>LASSO</td>
<td>0.59</td>
</tr>
<tr>
<td>Ridge</td>
<td>250</td>
<td>LASSO</td>
<td>0.56</td>
</tr>
<tr>
<td>SCAD</td>
<td>250</td>
<td>LASSO</td>
<td>0.50</td>
</tr>
</tbody>
</table>


R-squared value for and Drosophila data

<table>
<thead>
<tr>
<th>Dimension reduction Method</th>
<th>Number of genes</th>
<th>Method</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>1000</td>
<td>LASSO</td>
<td>0.68</td>
</tr>
<tr>
<td>Ridge</td>
<td>500</td>
<td>LASSO</td>
<td>0.55</td>
</tr>
<tr>
<td>SCAD</td>
<td>250</td>
<td>LASSO</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Dynamic Analysis Of Various Dimension Reduction Methods

Pre processing step
- Normalization
- Log-transformation
- Test of stationarity
- Eliminate the trend

Multivariate time series model
Multivariate vector autoregressive model (MVAR) with order p is:

\[ x(t) = \sum_{j=1}^{p} A_j x(t-j) + \epsilon(t), \]

\( x(t) \) is the vector of gene expression at previous time-points
\( x(t-j) \) is the vector at past time points
\( p \) is the known order of the model
\( \epsilon(t) \) denotes residuals
\( A \) is the unknown matrix of autoregressive coefficients

We can write the matrix notation as,

\[ X = E(t) \times 1 \]

Where,

\( X \) is the response variable,
\( E \) is the explanatory variable,
\( A \) is matrix of regression coefficients.

Regression
It should be noted that each gene may depend not only on its past values, but also on the past values of the other genes.

OLS estimators will be estimated imprecisely

- Large variances and covariance
- The least square estimator of regression coefficient
- Large standard errors

Conclusion
1. The results show that explanatory power under factor analysis and principle component analysis is better than the multiple linear regression for both data sets. This indicates that any method producing orthogonal dependent variables.
2. It can be seen that factor analysis has the highest r-squared value among all the of the dimension reduction techniques.
3. It can be inferred among all the regression methods, lasso has a better performance.
4. It is expected that the relative performances of reduced dimension models that provide orthogonal variables should be better with regards to low variance estimators because orthogonality leads to a reasonable coefficient estimation with low standard errors.