A Rank-Based Estimate for Cell Lineage Data

Tamer Elbayoumi and Jeffrey Terpstra — Department of Statistics, Western Michigan University, Kalamazoo, Michigan.

Abstract

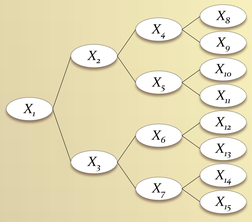
The presence of aberrant observations (i.e. outliers) in cell lineage data is quite common. As such, it is desirable to have an outlier-resistant estimation procedure as an alternative to least squares estimation (maximum likelihood estimation under normality). In this work, we consider rank-based estimates of the parameters of a first order bifurcating autoregressive [BAR(1)] model. The BAR(1) model was proposed by Cowan and Staudte (1986) for cell lineage data. In it, each line of descendents follows a first order autoregressive [AR(1)] model and allows sister cells from the same mother to be correlated. Real examples and a simulation study are performed in order to examine the behavior of these rank-based estimation procedures. More specifically, we compute finite sample relative efficiencies with respect to least squares estimate. The results indicate that the rank-based estimation procedures are more efficient when outlying observations are present.

Introduction

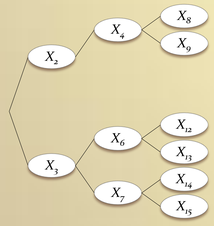
Bifurcating Autoregressive (BAR) Model is an adaptation of autoregressive (AR) model to binary tree structured data. This model is developed by Cowan and Staudte (1986) for cell lineage data, where each individual cell in any generation gives rise to two offspring in the next generation. Cowan and Staudte (1986) proposed the first-order Bifurcating Autoregressive BAR (1) model for cell lineage data, where each line of descendent follows an AR(1) model. This model allows to the sister cells be correlated.

Cell lineage data consists of observations of some quantitative characteristics of the cells over several generations of descendents from an initial cell, such as cell lifetime or final cell volume at cell division. BAR model takes into account both inherited and environmental effects to explain the development of the quantitative characteristic under study.

The purpose of studying the cell lineage process is of particular interest whether the observed correlations between related cells are due to similarities in the environments in which the cell grow, inherited effects, or a combination of both.



Fig(1): Symmetric cell lineage tree, no missing values. Note that the initial cell X_1 gives rise to two offspring in the next generation X_2 and X_3, by the same manner X_2 gives rise to two offspring in the next generation X_4 and X_5, and so on.



Fig(2): Asymmetric cell lineage tree. Note that the initial cell, cell number 5 and its offspring were missing, and then the other cells would retain their same numbering; in this way the structure of the tree may be identified by the indices of the observed values.

First-Order Bifurcating Autoregressive Model

Let X_t be an observed cell in a culture of some quantitative characteristic at time t. starting with the initial value X_1, the zero mean BAR(1) model which proposed by Cowan and Staudte (1986), is defined as

$$X_t = \phi\, X_{[t/2]} + \varepsilon_t \text{ , for all } t \geq 2$$

Where $[u]$ denotes the largest integer less than or equal to u. it is assumed that $\{(\varepsilon_{2t}, \varepsilon_{2t+1}), t \geq 1\}$ is a sequence of independent and identically bivariate random vectors with common mean zero, and common variance-covariance structure

$$\begin{pmatrix} \mathbf{1} & \theta \\ \theta & \mathbf{1} \end{pmatrix} \sigma^2$$

Where θ is the correlation between $(\varepsilon_{2t}, \varepsilon_{2t+1})$ they follow a distribution F. $|\phi_1| < 1$ denotes the mother-daughter correlation coefficient which need to be estimated. The sister-sister correlation is defined as $\rho = \phi^2 + (1-\phi^2)\theta$.

Comments on The Previous Works

Most of previous works used estimation methods such as Maximum Likelihood estimation, Modified Maximum Likelihood estimation to robust against outliers, and Least squares estimation.

The presence of aberrant observations in cell lineage data is quite common. The existence of outliers in data makes the data deviate from normality. Classical methods often have very poor performance in present of outliers. Then outliers have a highly affect on estimated BAR parameters using these estimations.

The Goal of This Work

This work aims to propose a robust estimation to the BAR model depending on Rank-Based procedure, or sometimes is called Weighted Wilcoxon (WW) procedure. Beyond of the theoretical proves and their complications, we will consider an empirical study depends on real examples and a simulation study in order to examine the behavior of these rank-based estimation procedures. More specifically, we will compute finite sample relative efficiencies with respect to least squares estimate.

In fact this work is an extension to Terpestra and et al. (1997, 2000, 2001a, 2001b, 2001c) works. Terpestra and others proposed robust estimates of Autoregressive model based on Weighted Wilcoxon (WW) procedure.

Rank-Based Estimation Method

Consider the BAR (1) model , if the error terms are Gaussian, it is well-known that Least squares estimate of is optimal. The proposed estimate of will be a value of ϕ that minimizes the following dispersion function,

$$D(\phi) = \sum_{2 \leq i < j \leq n} b_{ij} \left| \varepsilon_i - \varepsilon_j \right|$$

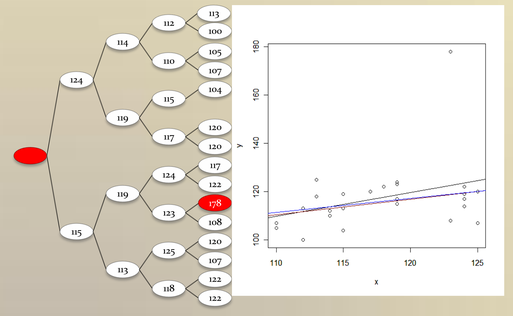
$$= \sum_{2 \leq i < j \leq n} b_{ij} \left| (x_i - x_j) - (x_{[i/2]} - x_{[j/2]})\phi \right|$$

With b_{ij} denoted weights used for $(i,j)^{th}$ the comparison .

There are three estimates based on the weights:

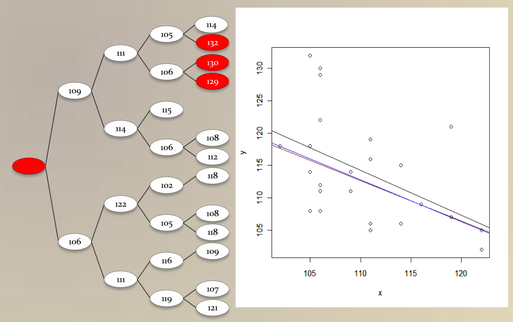
- Wilcoxon estimates, when $b_{ij} = 1$.
- Generalized Rank (GR) estimates, when weights depend on the observed values of the factor space points, i.e. $b_{ij} = b(x_{[i/2]}, x_{[j/2]})$.
- High Breakdown Robust (HBR) estimates, when weights depend on the observed values of the factor space points and the response, i.e. $b_{ij} = b(x_i, x_{[i/2]}, x_j, x_{[j/2]})$.

Example(1): EMT6 cells (tree 41 of Staudte, Guiget and Collyn d'Hooge(1984)), lifetimes in Tenths of Hours



Fig(3):
LS (——)
WIL (——)
GR (——)
HBR (——)

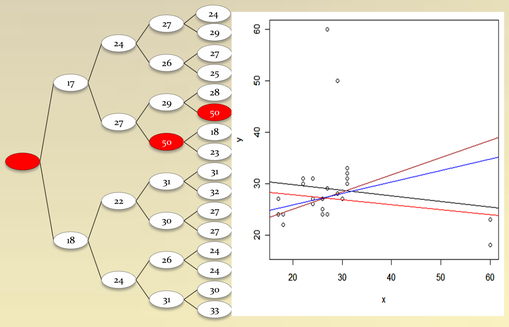
Example(2): EMT6 cells (tree 29 of Staudte, Guiget and Collyn d'Hooge(1984)), lifetimes in Tenths of Hours



Fig(4):
LS (——)
WIL (——)
GR (——)
HBR (——)

Example(3): A cell lineage tree for E.Coli taken from laboratory of E.O.Powell. The units are minutes. Cowan, R. and Staudte, R.G.(1986).

Let the observations numbers 11 and 21 are outliers comparing to the rest observations and practically the initial cell often is missing.



Fig(5):
LS (——)
WIL (——)
GR (——)
HBR (——)

Simulation Study

The behavior of the Rank-Based estimates is studied via Monte Carlo. Our primary interest is the behavior of the mother- daughter parameter, only the zero mean BAR(1) is considered. The distribution of errors is determined according to a contaminated normal distribution. 10000 realizations of balanced tree size 127 are generated, and ϕ = -0.9,-0.5, 0.5, and 0.9 are chosen. For comparison, the empirical asymptotic relative efficiency (ARE) based on the empirical mean squares errors (MSEs) relative to Least squares is computed. Only the results of mother-daughter correlation are reported.

	Mother- Daughter correlation ϕ																			
	-0.9					-0.5					0.5					0.9				
$\theta(\varepsilon_{2t}, \varepsilon_{2t+1})$	-0.9	-0.5	0	0.5	0.9	-0.9	-0.5	0	0.5	0.9	-0.9	-0.5	0	0.5	0.9	-0.9	-0.5	0	0.5	0.9
N(0,1)																				
LS/WIL	0.860	0.927	0.951	0.965	0.960	0.851	0.915	0.953	0.963	0.956	0.852	0.913	0.949	0.970	0.963	0.857	0.922	0.953	0.966	0.955
LS/GR	0.813	0.867	0.907	0.918	0.915	0.804	0.862	0.903	0.917	0.900	0.800	0.854	0.894	0.923	0.906	0.776	0.831	0.847	0.868	0.850
LS/HBR	0.845	0.919	0.947	0.963	0.956	0.835	0.909	0.950	0.961	0.951	0.837	0.906	0.945	0.969	0.959	0.832	0.914	0.947	0.961	0.947
0.95N(0,1)+0.05N(0,25)																				
LS/WIL	1.386	1.491	1.622	1.610	1.563	1.384	1.459	1.530	1.592	1.533	1.341	1.458	1.509	1.573	1.535	1.421	1.543	1.625	1.642	1.549
LS/GR	1.196	1.266	1.363	1.320	1.308	1.092	1.113	1.149	1.184	1.150	1.037	1.097	1.111	1.142	1.146	1.165	1.176	1.205	1.193	1.154
LS/HBR	1.285	1.465	1.629	1.619	1.554	1.183	1.389	1.491	1.574	1.516	1.111	1.388	1.472	1.560	1.520	1.288	1.495	1.615	1.630	1.520
0.95N(0,1)+0.05N(0,100)																				
LS/WIL	3.041	3.341	3.527	3.599	3.303	2.766	2.985	3.293	3.441	3.288	2.719	3.046	3.230	3.253	3.305	3.159	3.472	3.538	3.718	3.386
LS/GR	2.229	2.291	2.317	2.358	2.103	1.712	1.725	1.899	2.016	1.941	1.679	1.758	1.834	1.797	1.873	2.062	2.105	2.042	2.085	1.926
LS/HBR	2.381	3.196	3.516	3.580	3.187	1.805	2.655	3.190	3.367	3.136	1.769	2.763	3.100	3.179	3.157	2.383	3.186	3.404	3.674	3.267

Conclusions

The idea behind this work is to propose a rank-based estimation as a robust estimation against aberrant observations (i.e. outliers) for Bifurcating Autoregressive (BAR) Model to fit cell lineage data. After reviewing the previous works in this topic, many authors recommended to find out a robust estimation method for this sort of data in present of aberrant observations. After introducing many examples of cell linage data, estimating, and comparing to Least squares estimate, Wilcoxon, Generalized Rank, and High Breakdown robust parameters of BAR(1) model, it is found that the Rank-based estimates are fitting the data well in present of aberrant observations and missing values. A simulation study are performed to examine the behavior of rank-based estimation comparing to least squares estimation, and Empirical Asymptotic Relative Efficiency (ARE) are computed based on the empirical mean squares errors (MSEs) relative to Least squares. It is found that the high efficiency properties of the Wilcoxon estimation to LS estimation for the independent errors case extend to the Wilcoxon estimation for BAR model. It is found also that the rank-based estimates are more efficient than the Least squares estimate in all cases for contaminated normal data. So it can be conclude that the rank-based estimation is promising to introduce a robust estimation to fit Bifurcating Autoregressive Model for the cell linage data in presence of outliers.

References:

Cowan, R. and Staudte, R.G.(1986), The bifurcating autoregression model in cell lineage studies. Biometrics 42, 769-783.

Staudte, R.G., Guiguet, M., and d'Hooghe, M.C. (1984), Additive models for dependent cell populations. J. Theoret. Biol. 109, 127-146.

Terpstra, J.T. (1997), A robust estimate for an autoregressive time series. PhD Dissertation, Western Michigan University.

Terpstra, J.T., McKean, J.W. and Naranjo, J.D. (2001a), GR-estimates for an autoregressive time series, Statist. Prob. Letter. 51, 165–172.

Terpstra, J.T., McKean, J.W., and Naranjo, J.D. (2000). Highly efficient weighted wilcoxon estimates for autoregression. Stat., 35, 45–80.

Terpstra, J.T., McKean, J.W., and Naranjo, J.D. (2001c). Weighted wilcoxon estimates for autoregression. Aust. N. Z. J. Stat, 43 (4), 399–419.

Terpstra, J.T., and Rao, M.B. (2001b), General rank estimates for an autoregressive time series: A U statistics approach, Statist. Infer. Stoch. Proc. 4, 155–179.