

Minimum Required Data Length to Reconstruct GC Network

November 9, 2013

1 Task

Compute the minimum required data length when calculating GC. Use IF neural model as an example.

2 Analysis

Suppose there are two random variables x and y .

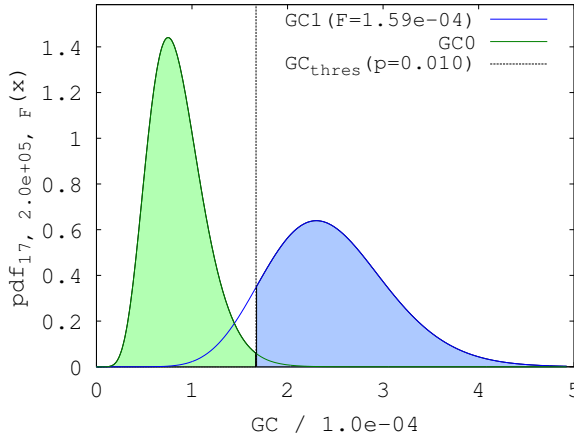
Recall: Distribution of GC obey

$$L \cdot \hat{F}_{x \rightarrow y} \stackrel{a}{\sim} \chi'^2(m, L \cdot F_{x \rightarrow y})$$

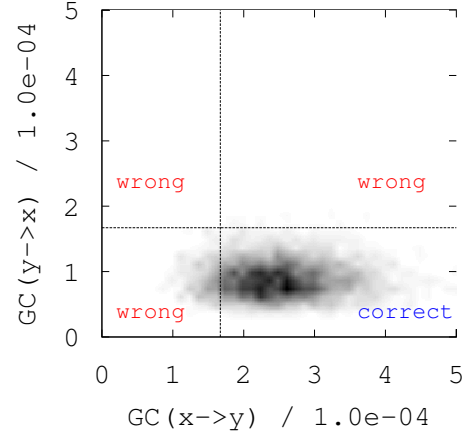
in the large L limit ($\stackrel{a}{\sim}$). Where L is number of data samples, $\hat{F}_{x \rightarrow y}$ is calculation value of true GC $F_{x \rightarrow y}$, m is fitting order. χ'^2 is Noncentral chi-squared distribution (see ref. [1] for definition and properties).

For convenience, we denote the probability density function (pdf) of \hat{F} as $\rho_{m,L,F}(x)$.

Now we want to know: for a given L , m and true $F_{x \rightarrow y}$, $F_{y \rightarrow x}$, what's the expected correct rate? Obviously, 25% will be the lowest bound (random guess).



(a) theoretical asymptotic pdf of $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$ (separately).



(b) density of simultaneous distribution of $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$. Obtained from 4000 experiment and counting by divide each axis into 40 uniform bins.

Figure 1: GC pdf under one set of typical parameter: $\mu = 1.0\text{kHz}$, $f = 0.012$, $S = 0.01$, using $L = 2.0 \times 10^5$ ($T = 1 \times 10^5\text{ms}$), $m = 17$, the true GC is $F_{x \rightarrow y} \approx 1.592 \times 10^{-4}$ and $F_{y \rightarrow x} \approx 0.006 \times 10^{-4}$ (obtained by $L = 1 \times 10^8$). The black line represent the GC thresholding value (GC_{thres}) that we used to judge whether there is connection or not. Here GC_{thres} satisfies $P(\hat{F}_{y \rightarrow x} < \text{GC}_{\text{thres}}) = 0.01$ and $F_{y \rightarrow x} = 0$ (our null hypothesis), i.e. false positive error rate is 1%.

If $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$ are independent, then the expression for correct rate $p_{correct}$ will be very simple:

$$p_{correct} = \int_0^{F_{thres}} \rho_{m,L,F0}(F) dF \left(1 - \int_0^{F_{thres}} \rho_{m,L,F1}(F) dF \right), \quad (1)$$

that is the product of areas of green and blue region in Fig.(1a). Otherwise, we have to count the volumn of lower right part of Fig.(1b).

2.1 Is $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$ are independent?

Geweke said (Ref.[2]) $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$ are asymptotically independent. But how “asymptotically”.

First, is the asymptotic pdf of $\rho_{m,L,F}(x)$ accurate?

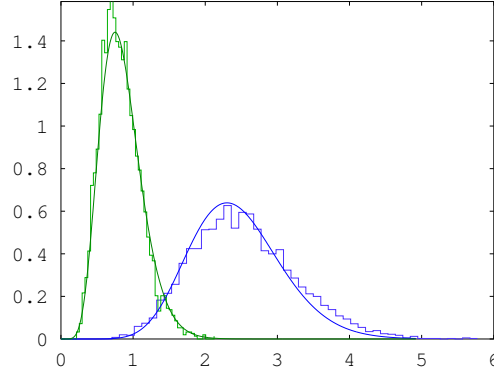
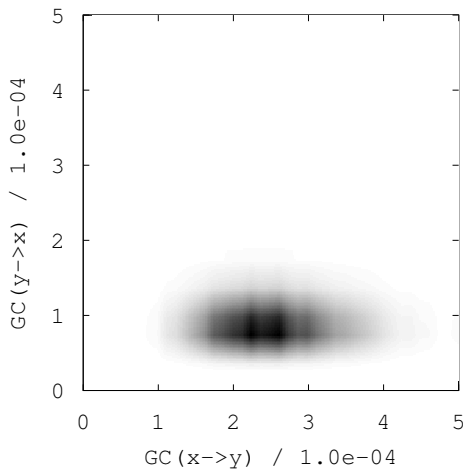


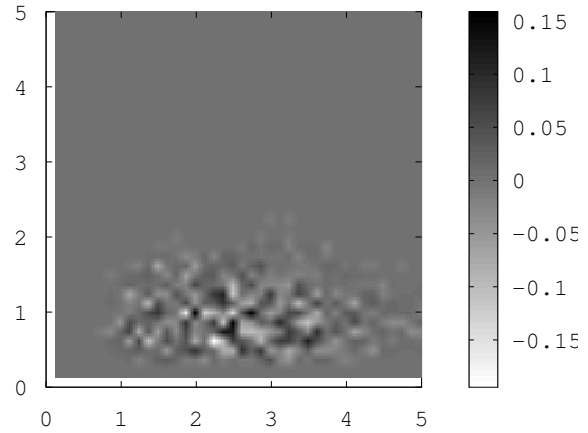
Figure 2: Comparison of statistic data and asymptotic pdf of $\rho_{m,L,F}(x)$

Second, by the 4000 GC data point mentioned in Fig.(1b), we can calculate the correlation of $\hat{F}_{x \rightarrow y}$ and $\hat{F}_{y \rightarrow x}$. The result is -0.012 , which can be explained by statistic error ($1/\sqrt{4000} \approx 0.016$).

Further, we compare the joint distribution to product of marginal distribution.



(a) product of marginal distribution



(b) pdf of $(\hat{F}_{x \rightarrow y}, \hat{F}_{y \rightarrow x})$ subtract product of marginal distribution, normalize to distribution peak equals one.

Figure 3

References

- [1] http://en.wikipedia.org/wiki/Non-central_chi-square_distribution
- [2] Measurement of Linear Dependence and Feedback Between Multiple Time Series, John Geweke, Journal of the American Statistical Association, Vol.77, No.378 (1982)