

# 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  $y$ .

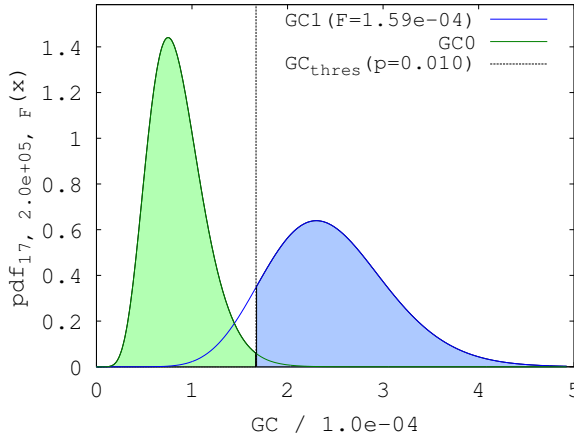
Recall: Distribution of GC obey

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

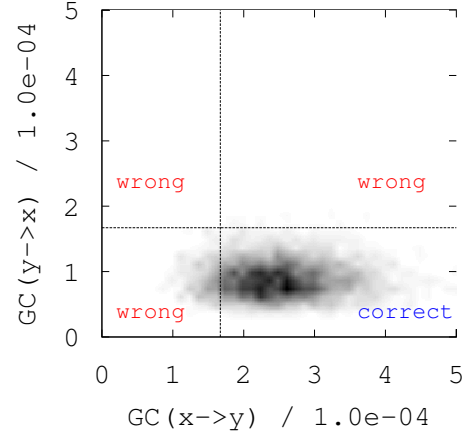
in the large  $L$  limit ( $\stackrel{a}{\sim}$ ). Where  $L$  is number of data samples,  $\hat{F}_{x \rightarrow y}$  is calculation value of true GC  $F_{x \rightarrow y}$ ,  $m$  is fitting order.  $\chi'^2$  is Noncentral chi-squared distribution (see ref. [1] for definition and properties).

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, without any other knowledge about the two neuron network).



(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 ( $\text{GC}_{\text{thres}}$ ) that we used to judge whether there is connection or not. Here  $\text{GC}_{\text{thres}}$  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  $p_{\text{correct}}$  will be very simple:

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

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?

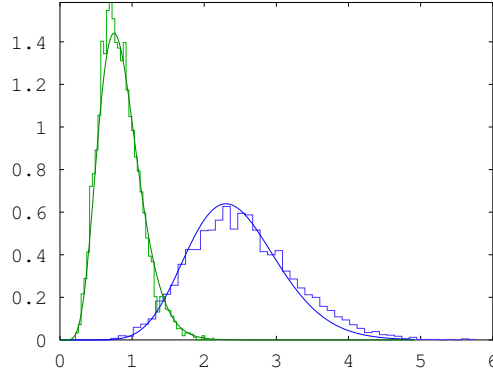
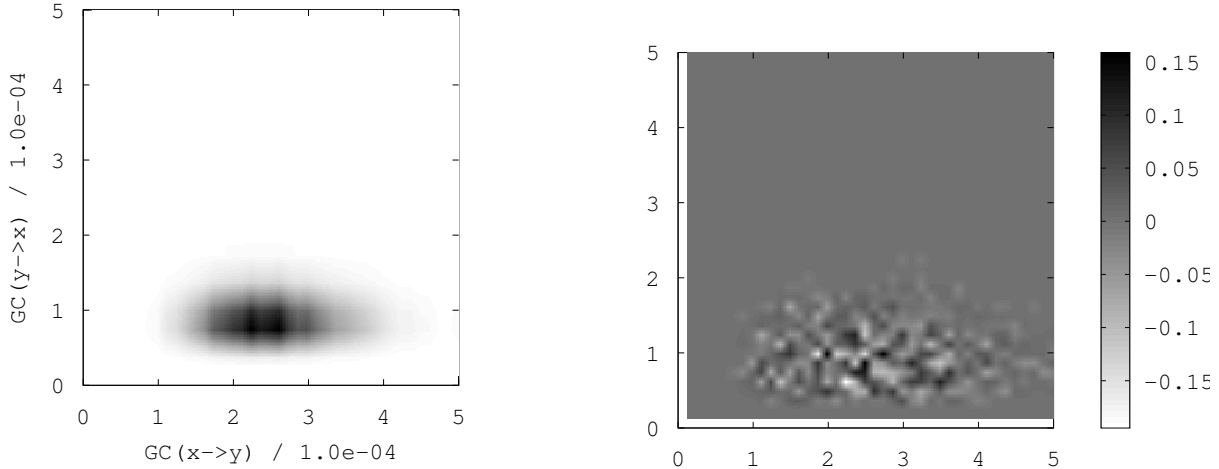


Figure 2: Comparison of statistic data and asymptotic pdf of  $\rho_{m,L,F}(x)$ . Same parameter as Fig.(1).

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 the product of marginal distribution.



(a) product of marginal distribution, looks similar to Fig.(1b).

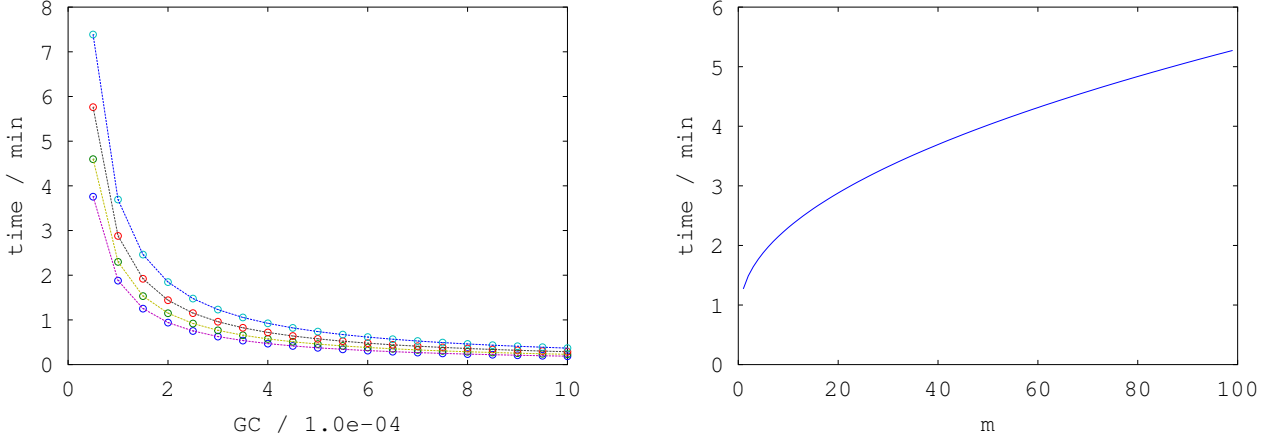
(b) pdf of  $(\hat{F}_{x \rightarrow y}, \hat{F}_{y \rightarrow x})$  subtract product of marginal distribution, normalized to distribution peak equals one.

Figure 3

In this parameter (See description of Fig.(1)), the 4000 experiements tell us the correct rate is about 87.1%, while using asymptotic pdf and assume  $\hat{F}_{x \rightarrow y}$  and  $\hat{F}_{y \rightarrow x}$  are independent, we get 88.9% from Eq.(1) through  $m, L, F_{x \rightarrow y}, F_{y \rightarrow x} = 0$  in Fig.(1). I think they're matched good enough.

## 2.2 Minimum required data length

From Eq.(1) it's now possible to solve the minimum required data length  $L_{\min}$  or data time length  $T_{\min}$  (instead of doing a lot of numerical experiments), if  $m$  and  $F_{x \rightarrow y}$  are known.



(a) Required time length v.s. GC value. Four curves from up to down corresponding to  $m = \{5, 10, 20, 40\}$ , circle dot is obtained by solving Eq.(1), dashed line is obtained from Eq.(2)(see below).

(b) Required time length v.s. fitting order. Fix GC value to  $F_{\text{true}} = 1.0 \times 10^{-4}$

Figure 4: False positive error rate set to 0.01, required correct ratio set to 90%.

In the case of false positive error rate set to 0.01, required correct ratio set to 90%, there is a good approximation of minimum length (relative error of  $T_{\min}$  is about 0.1%):

$$T_{\min} \approx \frac{10.00}{F_{\text{true}}} \left( 1.153 + \frac{\sqrt{m - 0.513}}{1.917} \right) \Delta t, \quad (m \in \{1, 2, \dots, 100\}) \quad (2)$$

where  $1/\Delta t$  is sample rate,  $\Delta t = 0.5$  ms in all above cases.

Recall that  $F_{\text{true}}/\Delta t \rightarrow \text{const.}$  in the limit  $\Delta t \rightarrow 0$ , so Eq.(2) tell us that there is no benefit to use small  $\Delta t$ , because in that case  $m \propto 1/\Delta t$  which make  $T_{\min}$  larger.

Bigger positive error rate will lead to smaller  $T_{\min}$ , but effect is limited. e.g. set positive error rate to 0.05 will decrease  $T_{\min}$  about 12%, while set positive error rate to 0.005 will increase  $T_{\min}$  about 7%.

Now remaining problem is: what is the true GC and corresponding  $m$ .  
We once have done that:

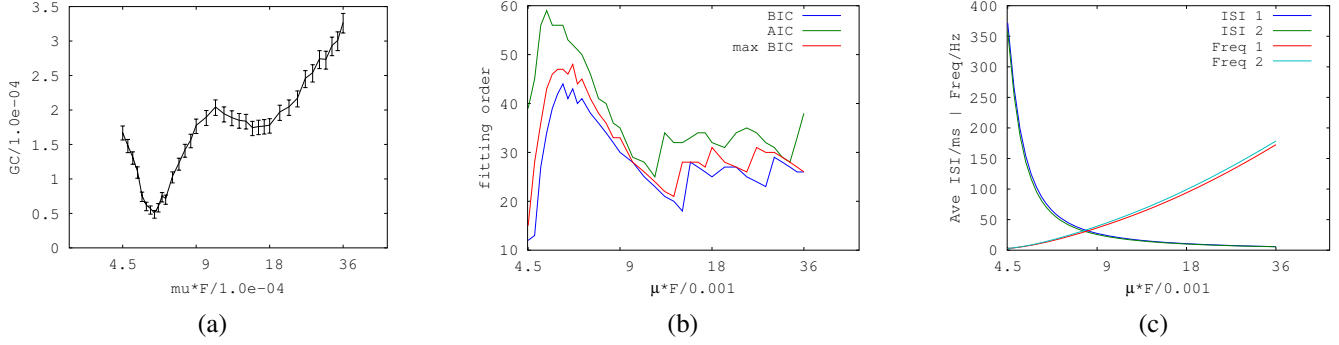


Figure 5: Scan  $F$ .  $\mu = 0.01, S = 0.01, T = 1.0 \times 10^4 \text{ sec}, \Delta t = 0.5 \text{ ms}$

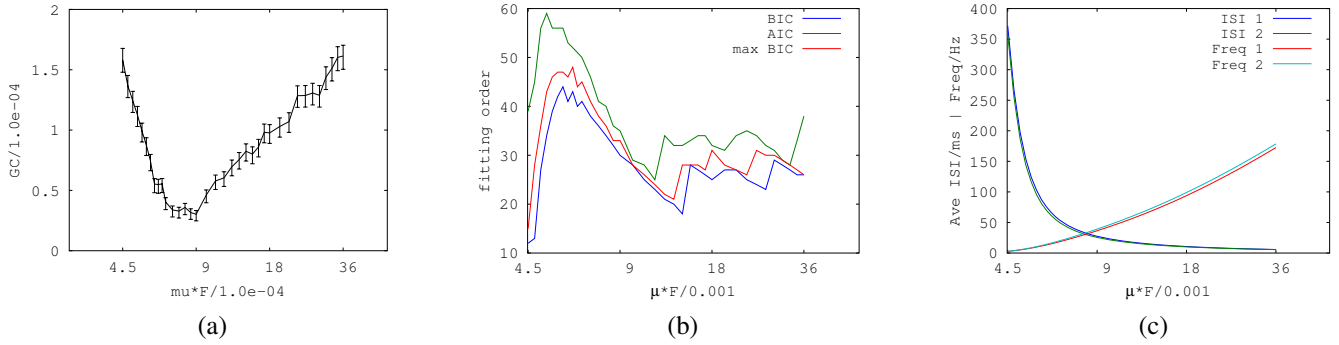


Figure 6: Scan  $F$ .  $\mu = 0.02, S = 0.01, T = 1.0 \times 10^4 \text{ sec}, \Delta t = 0.5 \text{ ms}$

Importing these data to Eq.(2) we get:

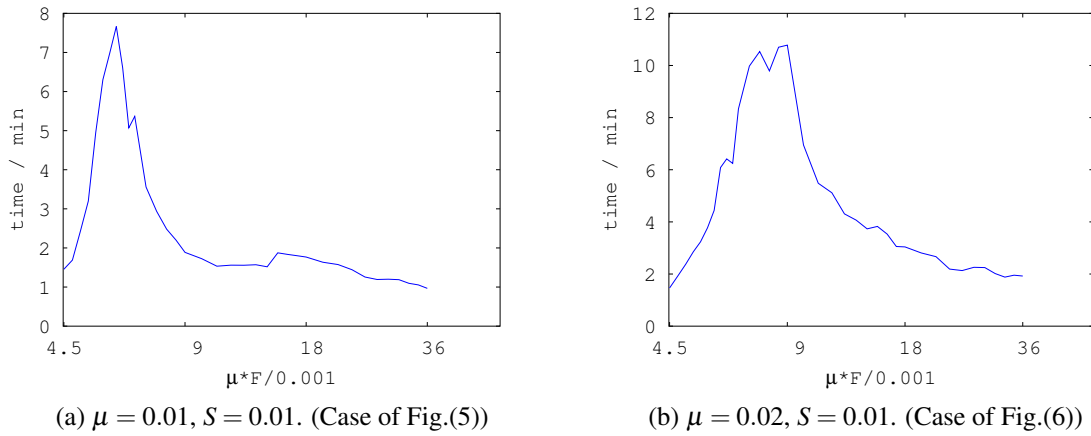


Figure 7: Required time length

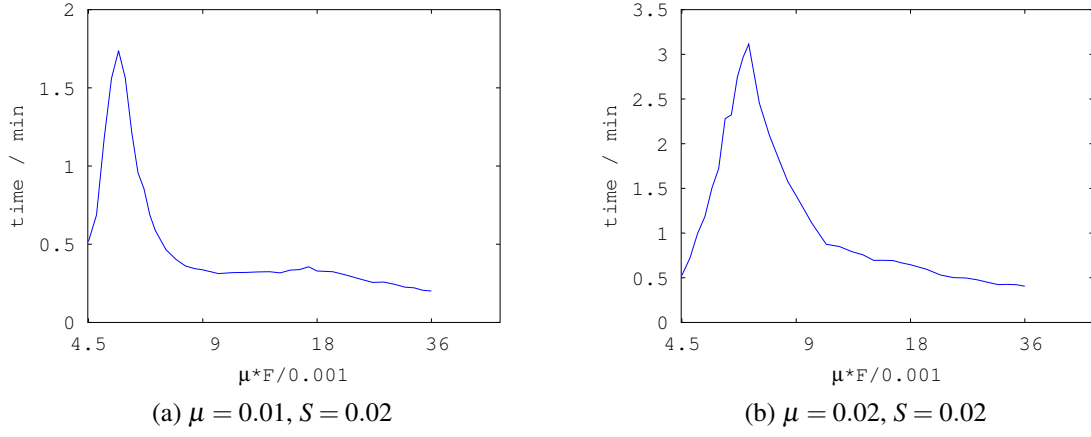


Figure 8: Required time length

As what we expected, twice the  $S$ , roughly quarter the required time.

Case of using spike train:

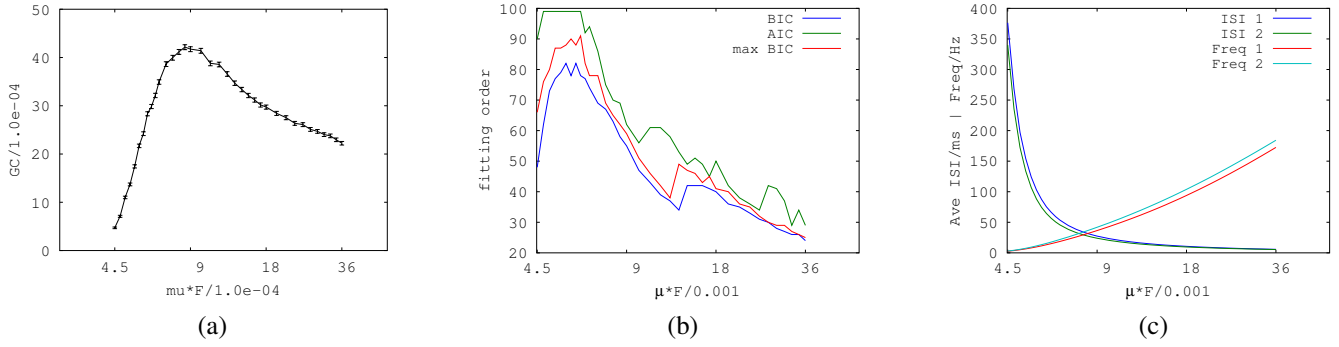


Figure 9: Using spike train. Scan  $F$ .  $\mu = 0.01, S = 0.02, T = 1.0 \times 10^4$  sec,  $\Delta t = 0.5$  ms

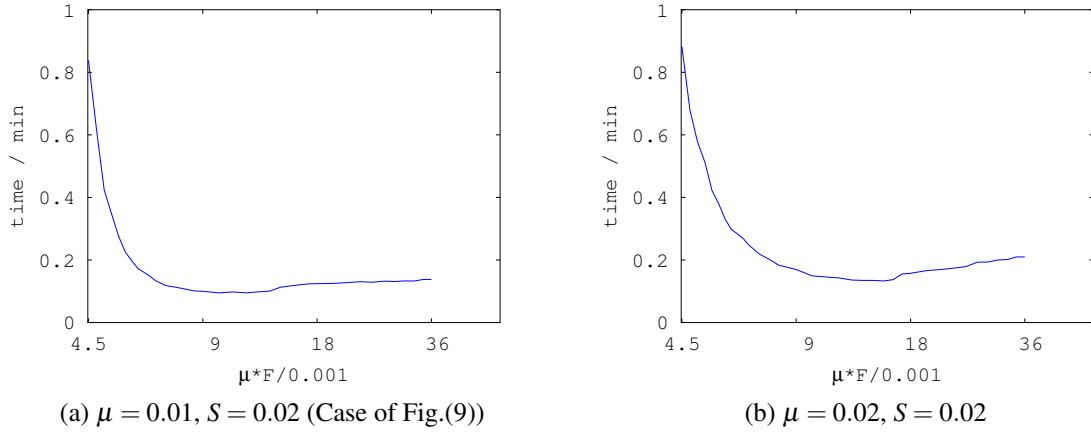


Figure 10: Required time length. (Use spike train)

Note:  $EPSP \approx 107S$  mV or  $EPSP \approx 7S$  in the normalized unit.

## References

- [1] [http://en.wikipedia.org/wiki/Non-central\\_chi-square\\_distribution](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)