時間変動するデータから
変動しない情報を取り出す数学

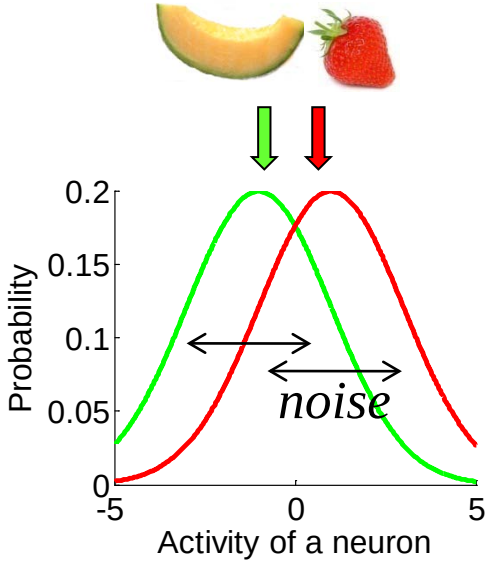
さきがけ数学塾@JST
3月8日

東北大学 情報科学研究科 数学連携推進室
三浦 佳二

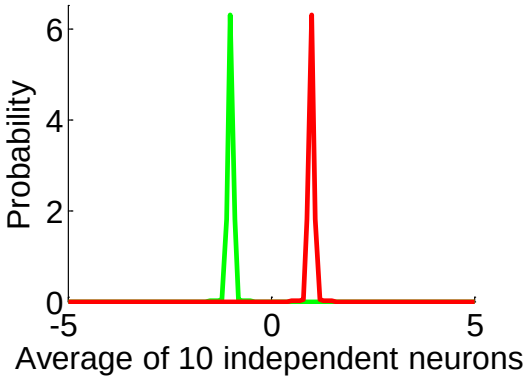
目次

- 神経電気生理学入門（1部削除しました）
- 神経細胞の活動の相関の推定

ノイズ相関の理論

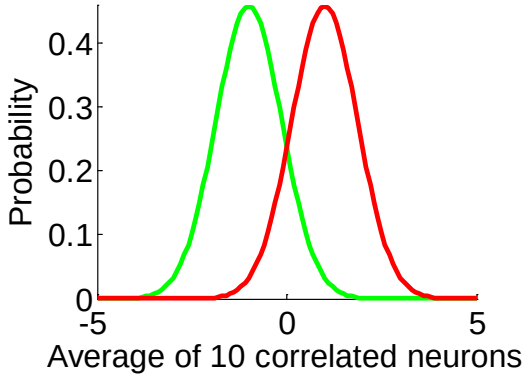


r=0



A. ノイズが細胞ごとに独立ならば、細胞平均をとれば減らせる。

r=0.1

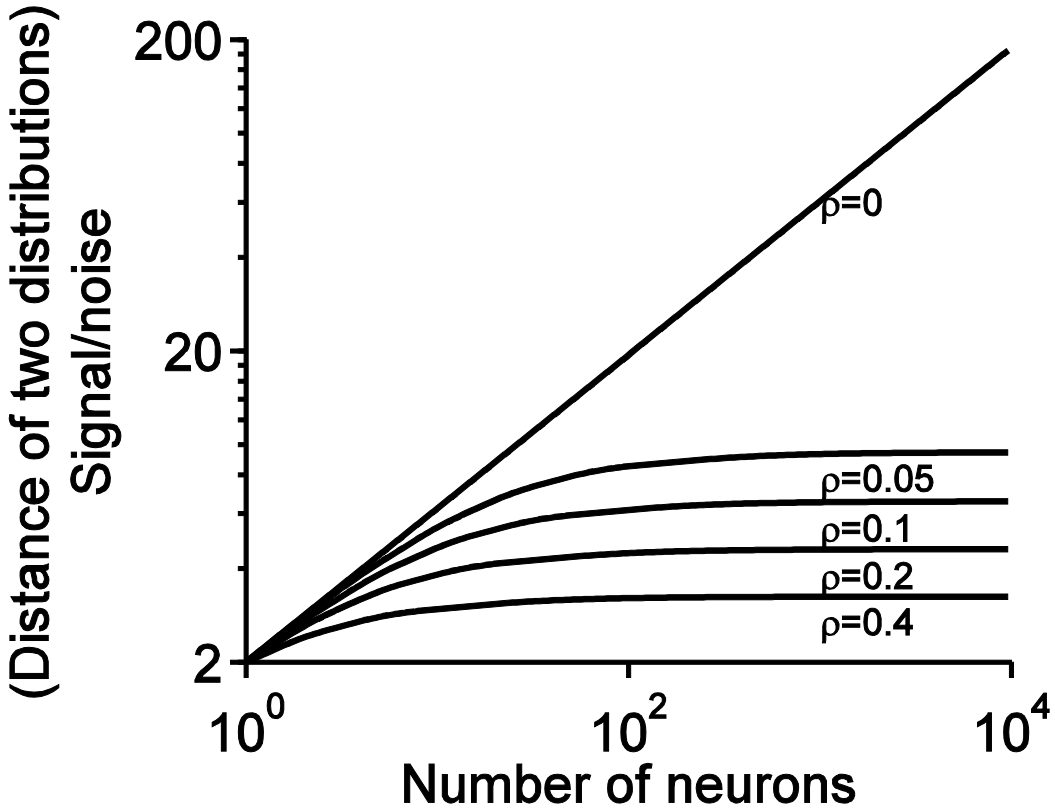


B. ノイズに相関があるならば、平均をとっても減らせない。

ノイズ: 試行毎の平均値まわりのばらつき。(神経細胞は、同じ刺激を繰り返しても、毎回異なる反応をする。)

(Zohary et al., 1994, 三浦2011)

ノイズ相関の理論（続き）



- もし、視覚野でみられるように0.1程度のノイズ相関が存在するならば、S/Nは、100個の神経細胞で飽和してしまう！

目次

- 神経電気生理学入門
- 神経細胞の活動の相関の推定

相関関数の定義

- ２つの時系列が与えられたとする（平均0,分散1に標準化）:

  $X(t)$, $Y(t)$ (t=1,2,…,T)

- 通常の相関係数（XとYの類似度）は、

$$\boxed{\rho(0) = \frac{1}{T}\sum_{t=1}^{T} X(t)Y(t)}$$

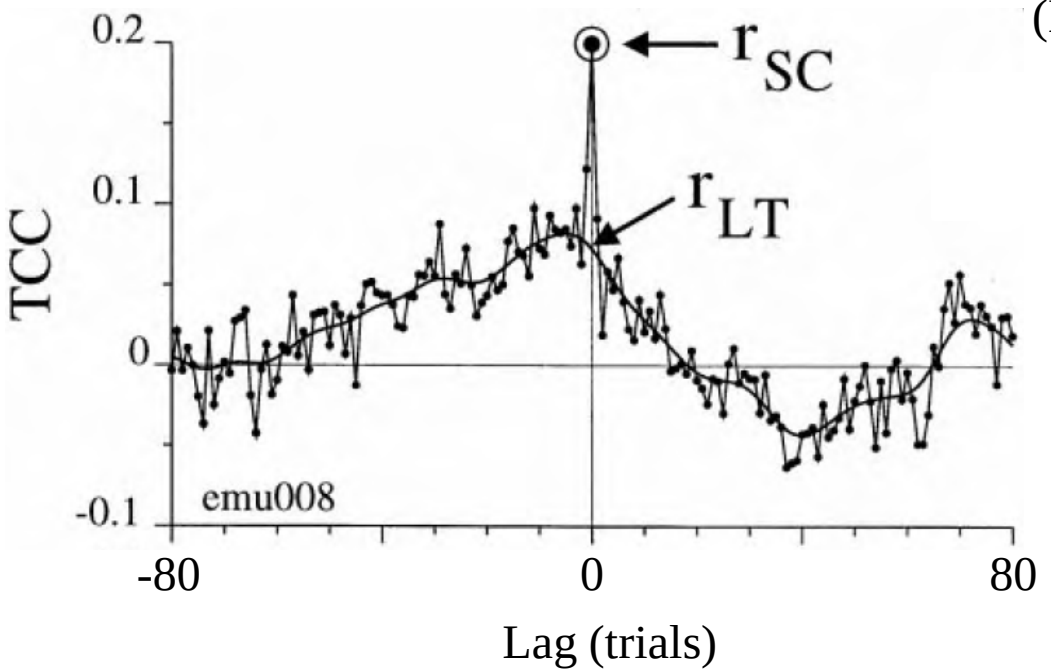
  で定義される。

- 様々な実験データのサマリーとして、<u>時間シフトした相関係数</u>（相関関数）がよく用いられる:

$$\boxed{\rho(\tau) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau} X(t)Y(t+\tau)}$$

- ただし、時間シフトτはデータ長Ｔと比べて十分短いとする。

2つの神経細胞の活動の相関関数



(Bair *et al.*, 2001)

神経科学で知りたいこと:
1. 神経細胞間の結合
2. 情報の冗長性(三浦,2011)

- 非定常なデータだとブロードな山がピークを隠してしまう！
  （脳活動のベースラインが時間変動している！？）
- この論文では、r_{SC}-r_{LT} が指標とされたが、意味が不明瞭。

→この現象をきちんとモデル化し、正しい相関の指標を提案したい

やりたいこと

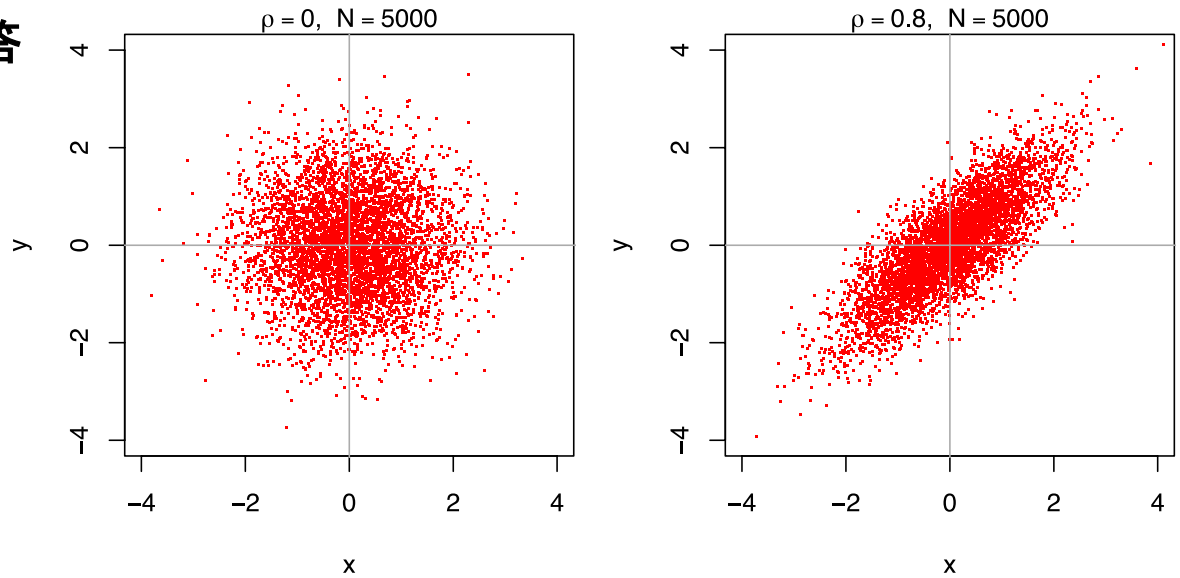
- 実験で見られる相関関数を再現する確率モデルの作成
- そのモデルのパラメータを推定するという形での、新しい指標の提案

“混合”正規分布モデル

2つの神経細胞の応答として，2変数の正規分布を考える:

$$q(X;\xi,\Sigma^{-1}) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{X}-\xi)'\Sigma^{-1}(\mathbf{X}-\xi)\}$$

- X={x, y} は2細胞の応答

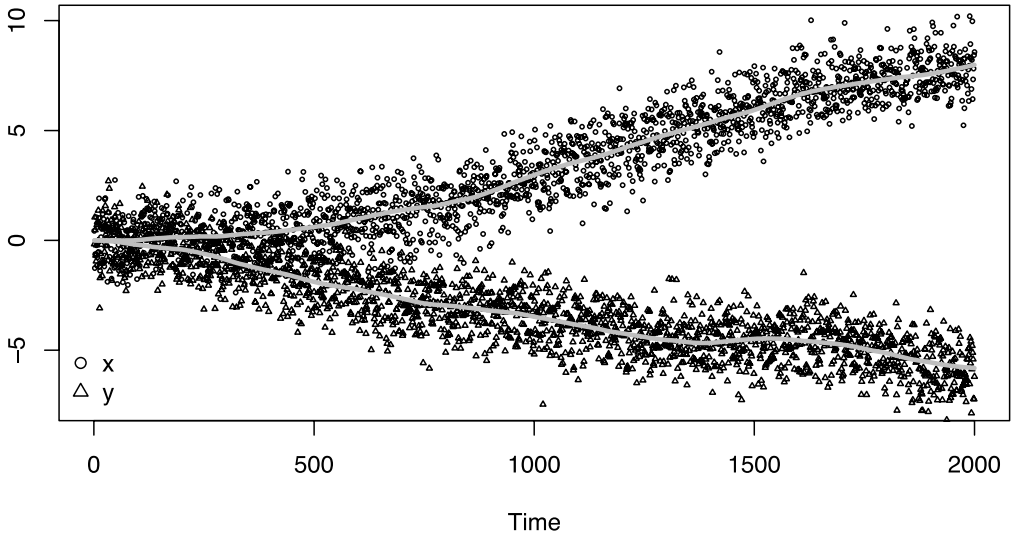


→このモデルでは「トレンド」を表現できない！平均値ξは「トレンド」として任意変動する！

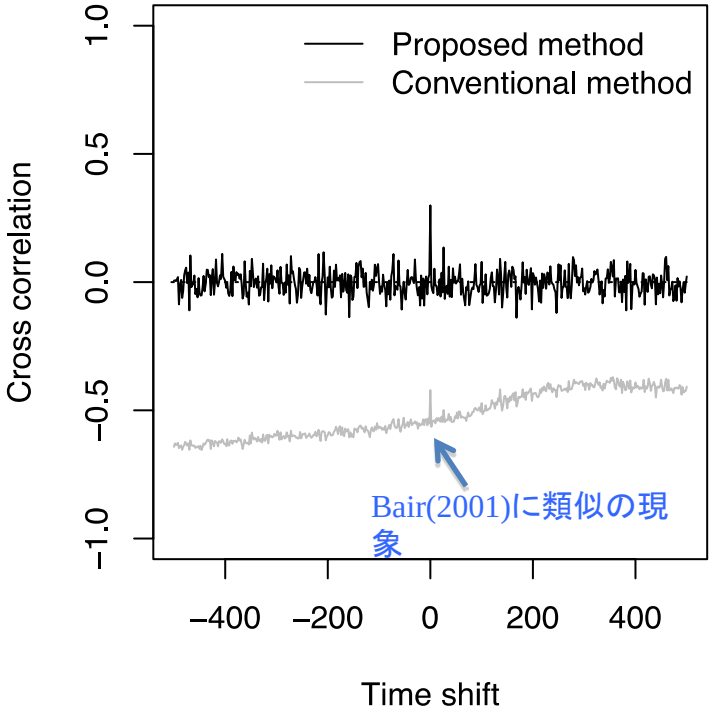
シミュレーションによる現象の再現と提案法による相関関数の推定

1. 今回,トレンド$\xi_1(t)$,$\xi_2(t)$はそれぞれARIMA(0,2,1)モデルで生成
2. そのξ_1,ξ_2を平均値として2変数正規分布で(x,y)を生成
3. ξの変動に関しては全く何も仮定せず(後述)共分散行列$\sum$を推定

トレンドを持つ2つの時系列を生成：



- データにはトレンド（灰色の線、平均値のドリフト）が存在する。

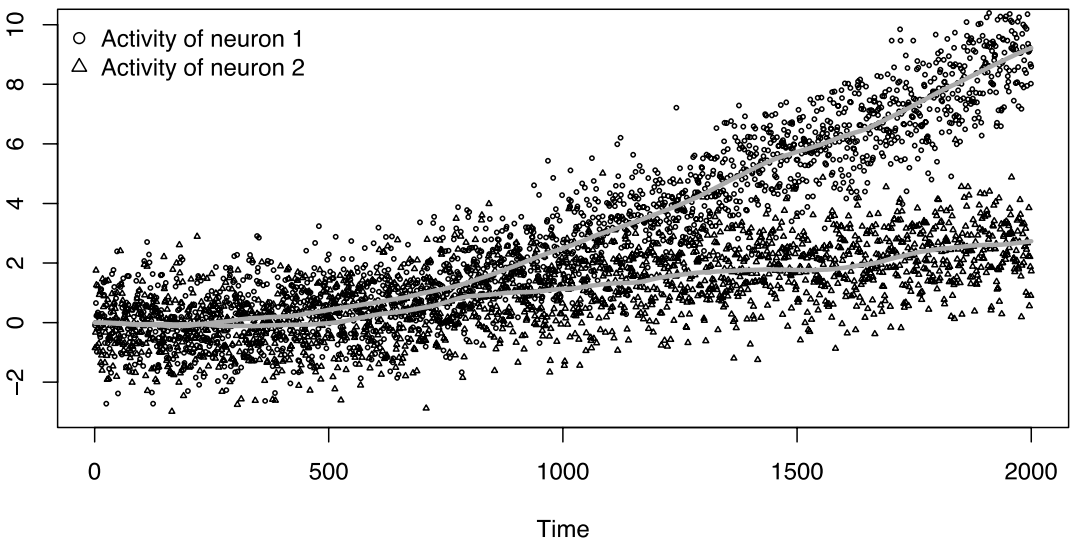


- 共分散行列は $\Sigma_{11}==\Sigma_{22}=1$, $\Sigma_{12}=\Sigma_{21}=0.3$。
- 長い単一の時系列から推定(T=2000)

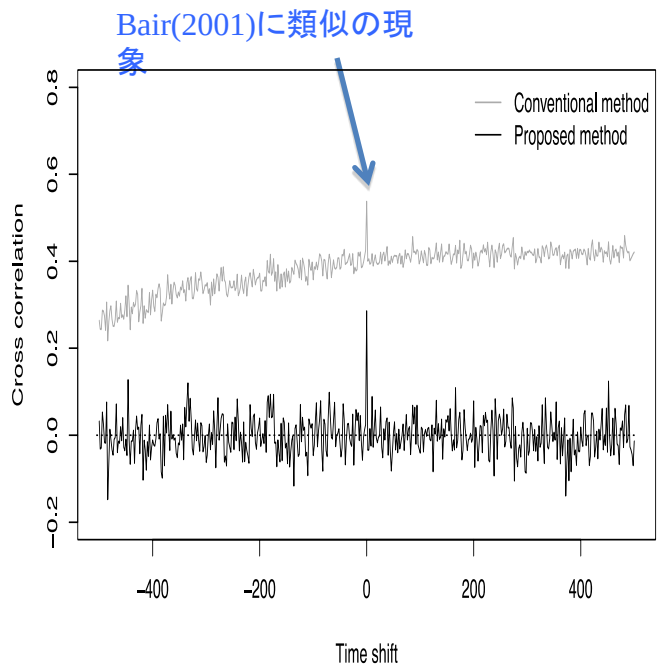
別の例（乱数の異なる実現）

1. 今回,トレンド$\xi_1(t)$,$\xi_2(t)$はそれぞれARIMA(0,2,1)モデルで生成
2. そのξ_1,ξ_2を平均値として2変数正規分布で(x,y)を生成
3. ξの変動に関しては全く何も仮定せず(後述)共分散行列$\sum$を推定

トレンドを持つ2つの時系列を生成：



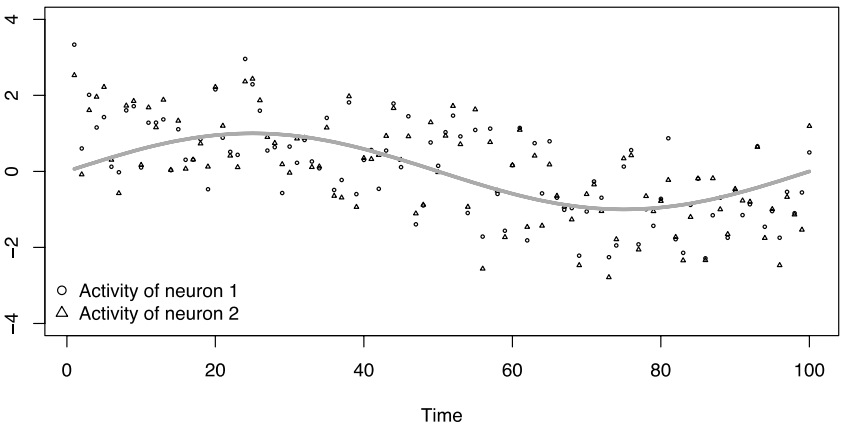
- データにはトレンド（灰色の線、平均値のドリフト）が存在する。



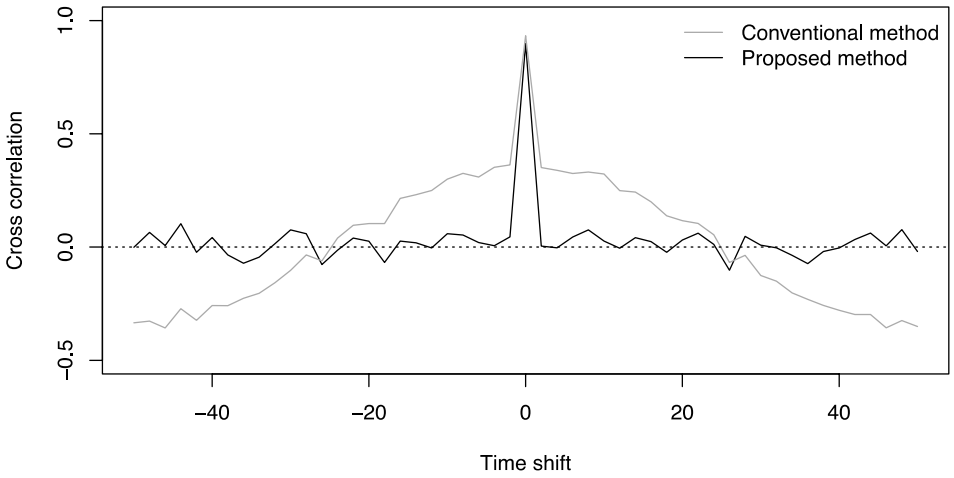
- 共分散行列は $\Sigma_{11}==\Sigma_{22}=1$, $\Sigma_{12}=\Sigma_{21}=0.3$。
- 長い単一の時系列から推定(T=2000)

トレンドの形によらず正しく推定可能

- このスライドの例ではトレンド（平均値のドリフト）はサイン波として生成

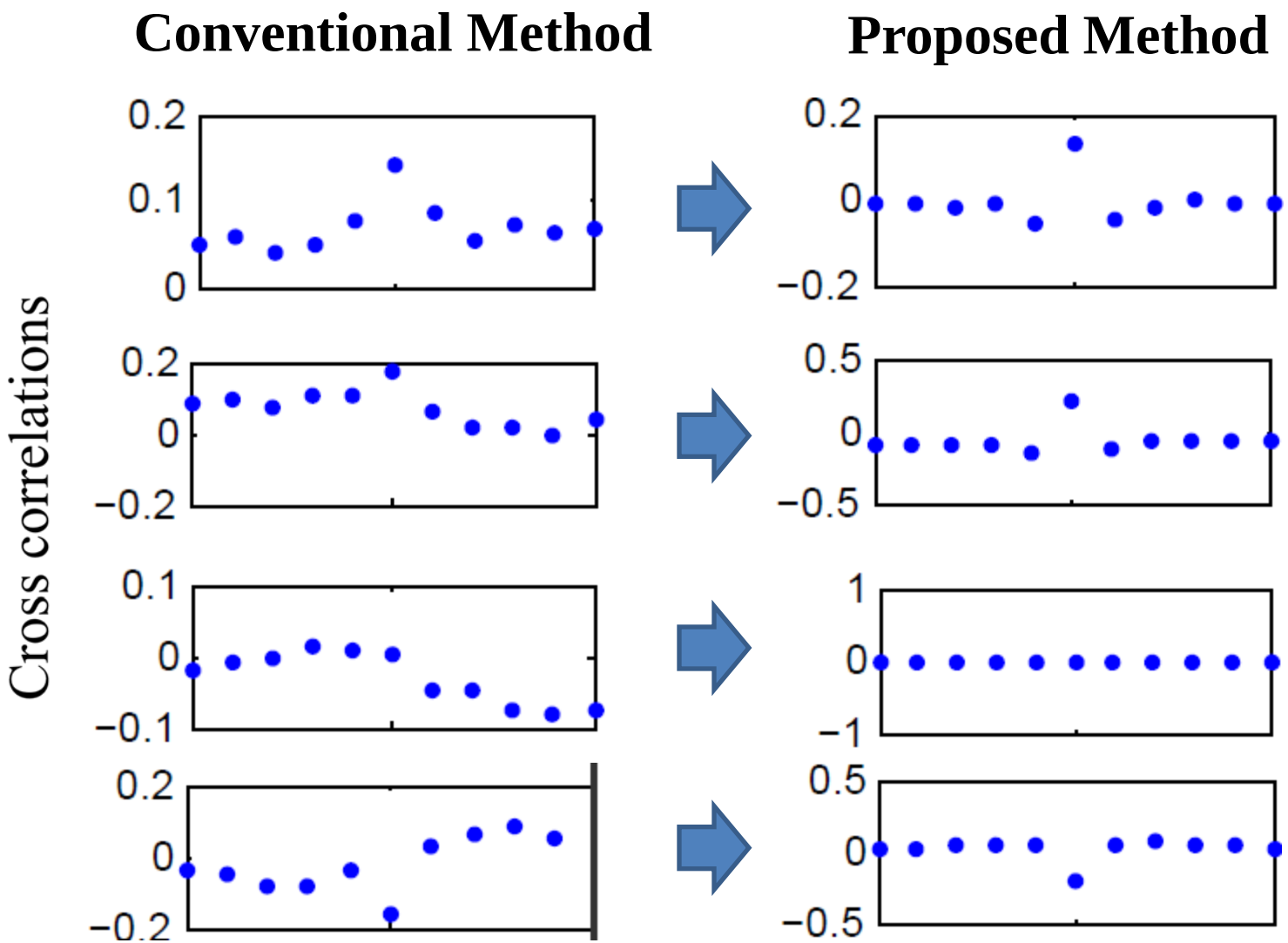


- サイン波のトレンド（灰色の線）をもつ2本の時系列



- 共分散行列は $\Sigma_{11}==\Sigma_{22}=1, \Sigma_{12}=\Sigma_{21}=0.9$
- 長い単一の時系列から推定(T=1000)

実データへの応用 (preliminary)



推定の定式化：混合正規分布モデル

２つの神経細胞の応答として，２変数の正規分布を考える：

$$q(X;\xi,\Sigma^{-1}) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{X}-\xi)'\Sigma^{-1}(\mathbf{X}-\xi)\}$$

- X={x, y} は2細胞の応答
- 正規分布の平均値としての「トレンド」$\xi=\{\xi_1,\xi_2\}$ は未知分布$k(\xi_1,\xi_2)$に従って分布：

$$p(X_1,X_2;\Sigma^{-1},k) = \int k(\xi)q(X_1;\xi,\Sigma)q(X_2;\xi,\Sigma)d\xi$$

- 連続性から，M(=2)個の，時間的に隣り合うξは同じという最低限の仮定を課す．
- $k(\xi_1,\xi_2)$がどのような形でも、Σ^{-1}を正しく推定したい（最も誤差の少ない方法で）．

トレンドの分布k(ξ_1,ξ_2)の形が何であっても共分散行列∑を正しく推定できる！

$$\hat{\Sigma}_{11} = \frac{1}{N}\sum_{t=1}^{N}\left\{(x_1^{(t)} - \overline{x^{(t)}})^2 + (x_2^{(t)} - \overline{x^{(t)}})^2\right\}/(2-1)$$

$$\overline{x^{(t)}} = \frac{x_1^{(t)} + x_2^{(t)}}{2}$$

$$\hat{\Sigma}_{12} = \frac{1}{N}\sum_{t=1}^{N}\left\{(x_1^{(t)} - \overline{x^{(t)}})(y_1^{(t)} - \overline{y^{(t)}}) + (x_2^{(t)} - \overline{x^{(t)}})(y_2^{(t)} - \overline{y^{(t)}})\right\}/(2-1)$$

$$\hat{\rho} = \frac{\hat{\Sigma}_{12}}{\sqrt{\hat{\Sigma}_{11}\hat{\Sigma}_{22}}}$$

・時系列データ{x(t),y(t)}を右辺に代入するだけ！
・ξ(t)の無限通りの可能性に対応できる！
・トレンドの分布k(ξ)によらず正しく推定可能な推定量の中でも、推定誤差が最小となる！
・n-1(=1)で割っているため、うまくいく。

情報幾何学とは？

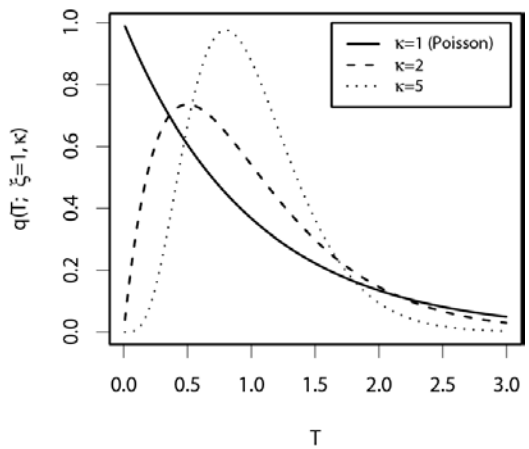
- 統計モデルに含まれるパラメタの「直交性」を利用する！

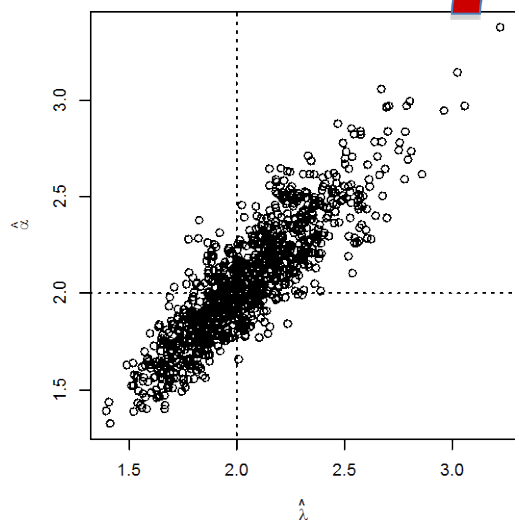
<u>ガンマ分布での例：</u>

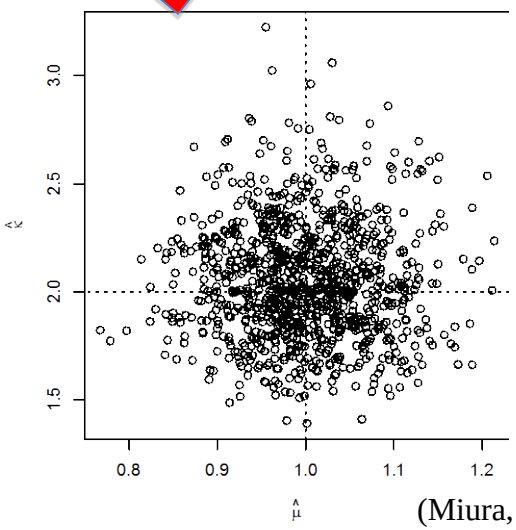
$$q(T;\xi,\kappa)=\frac{(\xi\kappa)^{\kappa}}{\Gamma(\kappa)}T^{\kappa-1}e^{-\xi\kappa T}$$

パラメタ変換

$$\begin{cases}\mu=\dfrac{\lambda}{\alpha}\\ \kappa=\lambda\end{cases}$$







(Miura, 2011)

ガンマ分布から乱数で100点観測し、分布の形のパラメタを推定

パラメタ（λとα）の推定値のばらつき（異なる乱数の実現が1つの点）

変数変換後（μとκ）は無相関になり、パラメタ毎に独立に統計検定が行える（cf. 正規分布の平均と分散は直交している）

部分的なパラメタの最適な推定

スコア関数

$$l_i \equiv \frac{d \log p(T;\vec{\theta})}{d\theta_i}$$

Fisher行列

$$I_{ij} \equiv \overline{l_i l_j} = \int p l_i l_j dT$$

誤差の下限

$$\mathrm{Cov}(\hat{\theta}) \geq \frac{1}{n} I^{-1}$$

今回のスコア関数

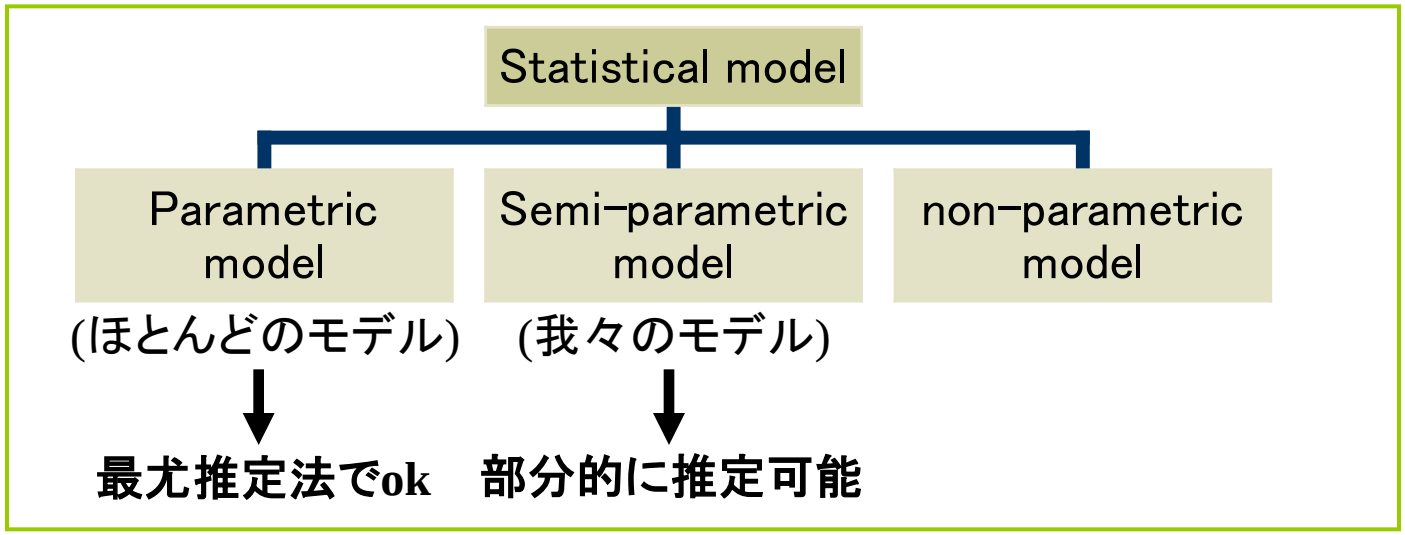
$$\begin{cases} u(X;\Sigma,k) \equiv \dfrac{d \log p(X;\Sigma,k)}{d\Sigma_{ij}} \\ v(X;\Sigma,k) \equiv \dfrac{\delta \log p(X;\Sigma,k)}{\delta k(\xi)} \end{cases}$$

ブロック対角化

$$\begin{bmatrix} I_{11} & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \xrightarrow[\text{グラムシュミットの直交化}]{y = u - \frac{u \cdot v}{v \cdot v} v} \begin{bmatrix} I'_{11} & 0 & 0 \\ 0 & - & - \\ 0 & - & - \end{bmatrix}$$

セミパラメトリックモデル

セミパラメトリックモデルとは，パラメタとして，ベクトル Σ と未知関数 $k(\xi)$ の両方を含むモデル．



最尤推定法は，セミパラメトリックモデルに対しては正しい推定を与えない
(パラメタの数よりもずっと多い数の観測データが必要．)

自由度を考えても ξ の正確な推定は困難．しかし，Σ だけでも推定したい！

最尤推定法

目的:

観測データ: $(\vec{X}^{(1)},\ldots,\vec{X}^{(N)}) \xrightarrow{\text{推定}} (\hat{\Sigma},\hat{\xi}^{(1)},\cdots,\hat{\xi}^{(N)})$

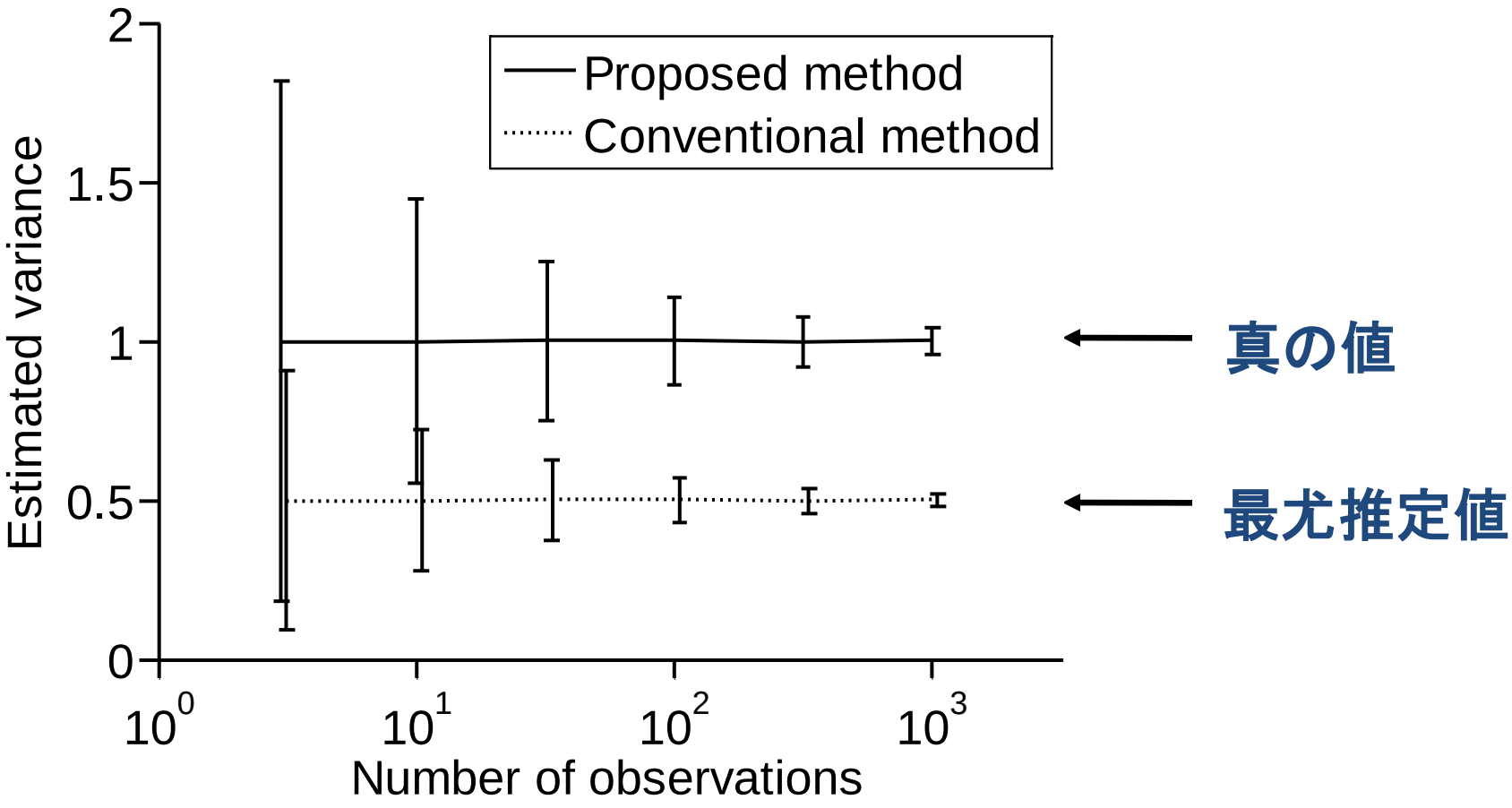
尤度: $L=\prod_{t=1}^{N} q(X_1^{(t)};\hat{\xi}^{(t)},\hat{\Sigma})q(X_2^{(t)};\hat{\xi}^{(t)},\hat{\Sigma})$ **<u>尤度を最大化するパラメタを求める</u>**

$\dfrac{d\log L}{d\hat{\Sigma}}=0$ などより、次の連立方程式を解けばよい:

$$\begin{cases} \hat{\xi}_x^{(t)} = \dfrac{x_1^{(t)}+x_2^{(t)}}{2} \\ \hat{\Sigma}_{11} = \dfrac{1}{N}\displaystyle\sum_{t=1}^{N}\dfrac{(x_1^{(t)}-\hat{\xi}_x^{(t)})^2+(x_2^{(t)}-\hat{\xi}_x^{(t)})^2}{2} \\ \ldots \\ \text{etc.} \end{cases}$$

⇨ **うまくいかない！**
直感的に考えても、分散は2で割ってはダメ

最尤推定法ではΣを正しく推定できない



使用法

仮定:

- ２つの時系列データ
- 平均値がどのようにドリフトしていても良い
- ガウスノイズには、同時刻の相関が存在する

目的:

- 相関（ノイズの共分散行列）を推定

まとめと議論

- どのようなトレンドがあっても、共分散行列を最適に推定する方法を提案した。
- トレンドを推定するにはデータ数が不十分な状況であるため、最尤法ではうまくいかない。
- 可能なトレンドの形が「無限」にあるにも関わらず、うまく推定できる。有限個のパラメタを推定する問題とは異なり、∞次元の問題解決には数学の力が必須！
- ちなみに、従来の相関関数は平均が一定だという誤った仮定をして共分散を推定するために間違えてしまう。一般に、間違った仮定は推定に大きな悪影響を及ぼす。