Entropy Estimation-based Assessment of Physiological Networks

J Randall Moorman, MD
Professor of Medicine, Physiology, BME
Director, Center for Advanced Medical Analytics
University of Virginia
Editor-in-chief, *Physiological Measurement*

The story so far

- The nature and degree to which organs are networked has information about clinical status of patients.
- Mathematical analysis of physiologic time series can detect neonatal sepsis early, and can save lives.
- While the autonomic nervous system is a means for networking, it is complicated, and resists easy, linear interpretation.
- This opens the door to non-linear approaches to analysis.
- We have used entropy estimation in our neonatal sepsis detection scheme.

Clausius 1864

- Early figure in thermodynamics
- In combustion, all the heat generated is not used for work
- He **coined the term ENTROPY** from “energy” plus “tropos” (transformation); the concept is that energy is lost.
- Overall S does not decrease = 2nd law

Boltzmann, Gibbs 1870s

- Showed the relationship of entropy (*S*) to the number of states of a system is logarithmic:

$$S = kB \log W$$

which is a special case of the more general form:

$$S = -kB \sum p_i \log p_i$$

when all the states are equally likely. Here, p_i is $1/W$

Shannon 1948

We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is reasonable to require of it the following properties:

1. H should be continuous in the p_i.

2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three

Theorem 2: *The only H satisfying the three above assumptions is of the form:*

$$H = -K\sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

Quantities of the form $H = -\sum p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[8] where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$. If x is a chance variable we will write $H(x)$ for its entropy; thus x is not an argument of a function but a label for a number, to differentiate it from $H(y)$ say, the entropy of the chance variable y.

The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.

1. $H = 0$ if and only if all the p_i but one are zero, this one having the value unity. Thus only when we are certain of the outcome does H vanish. Otherwise H is positive.

2. For a given n, H is a maximum and equal to $\log n$ when all the p_i are equal, i.e., $\frac{1}{n}$. This is also intuitively the most uncertain situation.

3. Suppose there are two events, x and y, in question, with m possibilities for the first and n for the second. Let $p(i,j)$ be the probability of the joint occurrence of i for the first and j for the second. The entropy of the joint event is

$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j)$$

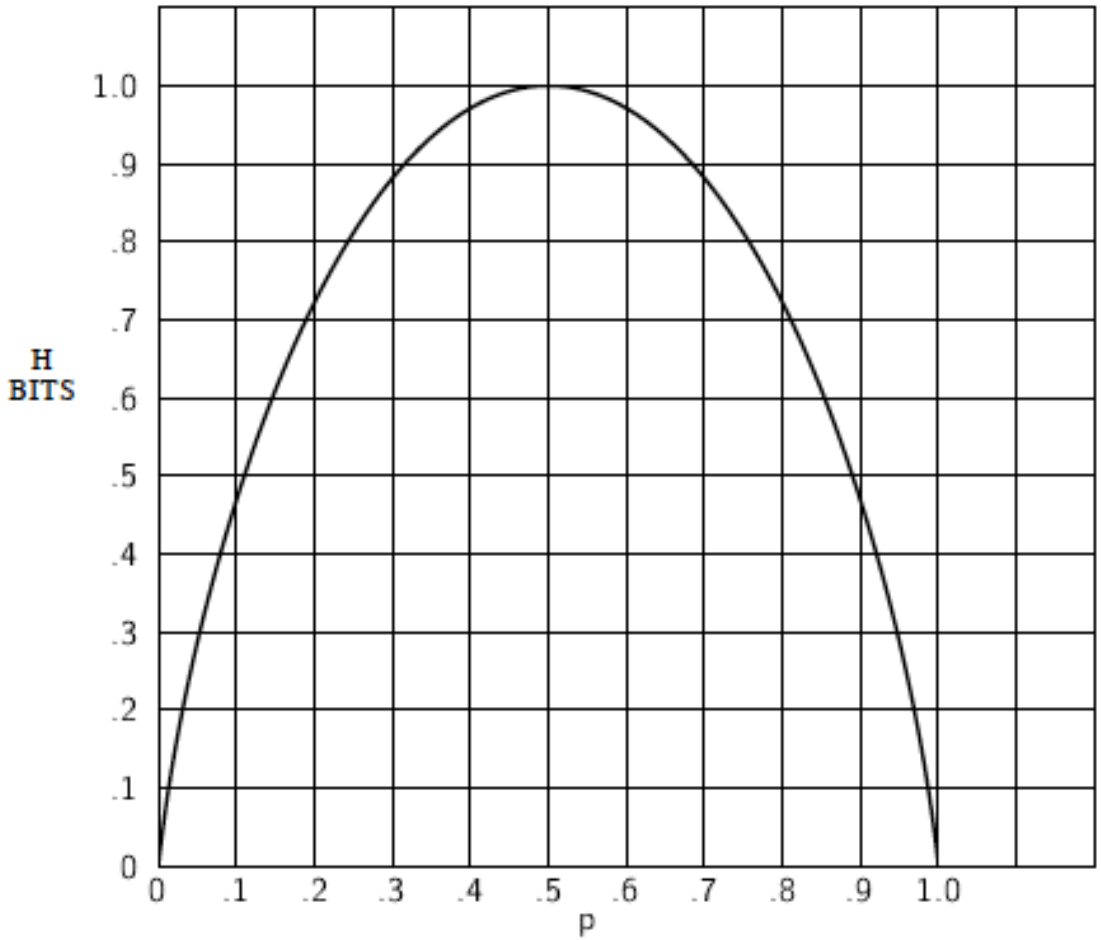


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1-p)$.

Shannon 1930s

- My greatest concern was what to call it. I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons: In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.

- M. Tribus, E.C. McIrvine, "Energy and information", *Scientific American*, 224 (September 1971)

- Shannon named it *H* after Boltzmann’s H-theorem

An intuitive feeling for $-p(x_i) \log p(x_i)$

- We wish to have a measure of the surprise that we feel when we see the next point in a time series, x_i
- One way is to measure surprise as the inverse of the probability $p(x_i)$ or $1/ p(x_i)$. Low probability points generate big surprise.
- But suppose we want to think about the surprise of the next 2 points – multiplying the 2 probabilities seems extreme. Rather, it seems we should be adding them.
- Thus let's use the log $p(x_i)$, or, in this case, - log $p(x_i)$ for the inverse
- We can then estimate the surprise of the entire time series as the sum of all the – log $p(x_i)$.
- And to estimate the average, we can take the expectation, or

$$H(X) = -\mathbb{E}[\log p(x_i)] = -\sum_{i}^{n} p(x_i) \log p(x_i)$$

Kolmogorov and Sinai 1958 and 1959

- Employed Shannon's entropy as an invariant measure of an ergodic *dynamical* system – a new concept was that new values of an dynamical process could be estimated with certainty that was characteristic of the system itself
- Thus the entropy of K and S is:

$$H_{KS} = -\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n\delta}\sum_{k_1,\ldots k_n} p(k_1,\ldots,k_n)\log p(k_1,\ldots,k_n)$$

$$H_{KS} = \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{n\to\infty}(H_{n+1} - H_n)\,.$$

Kolmogorov and Sinai 1958 and 1959

- The intuitive interpretation is that each new state in the evolving dynamical system can be expected with greater or lesser uncertainty if one knows the preceding states
- This degree of uncertainty is a invariant measure or characteristic of the system
- The concept is very naturally applied to time series data, but the limits make it impractical in its full form
- Next and most relevant steps were to cast the idea into approximations that could be used in experimental data, but first a word about estimating fractional dimensions

Renyi 1970

Gave a general form for a family of entropies of order *q*, where Shannon entropy is the case for q approaching 1. *Note that the log is now outside the* Σ.

$$H(X) = -\frac{1}{1-q}\log\sum_{i}^{n} p^{q}(x_i)$$

Takens 1981

- Embedding theorem allows reconstruction of an attractor from a time series by

$$\mathbf{x}_i = (x_{i-m+1}, x_{i-m+2}, \ldots x_i) \in R^m$$

where *m* is the embedding dimension of the attractor.

This method opened the door for non-linear dynamical analyses of experimental time series data.
(For the practitioner, there are issues about the values of the lag, especially in over-sampled data.)

Grassberger and Procaccia 1983

- Put together the idea of KS entropy and Takens embedding theorem to develop a method for determining the fractional dimension of an attractor reconstructed from an experimental time series.
- Two fundamental tools were the correlation sum and the K2 entropy, or KS entropy, or Renyi entropy of order 2.

Grassberger and Procaccia 1983

The correlation sum is the fraction of pairs whose distance is smaller than a tolerance *r*

$$\hat{C}(r) = \frac{2}{N(N-1)} \sum_{i<j} \theta(r - |\mathbf{x}_i - \mathbf{x}_j|)$$

$$K_2 = \lim_{N\to\infty} \lim_{m\to\infty} \lim_{r\to 0} - \ln[C^{m+1}(r) - C^m(r)].$$

Where K_2 is a lower bound for KS entropy

Eckmann and Ruelle 1985

Define:

$C_i^m(r)$ is the probability that points in the signal stay within a ball for *m* points

$$\phi^m(r) = \frac{1}{N}\sum_i \log C_i^m(r)$$

$$\Phi^{m+1}(r) - \Phi^m(r) \approx \sum_{i=1}^{N-m+1} \ln[C_i^m(r) / C_i^{m+1}(r)]$$

$$H_{\mathrm{ER}} = \lim_{N\to\infty} \lim_{m\to\infty} \lim_{r\to 0} [\Phi^m(r) - \Phi^{m+1}(r)]$$

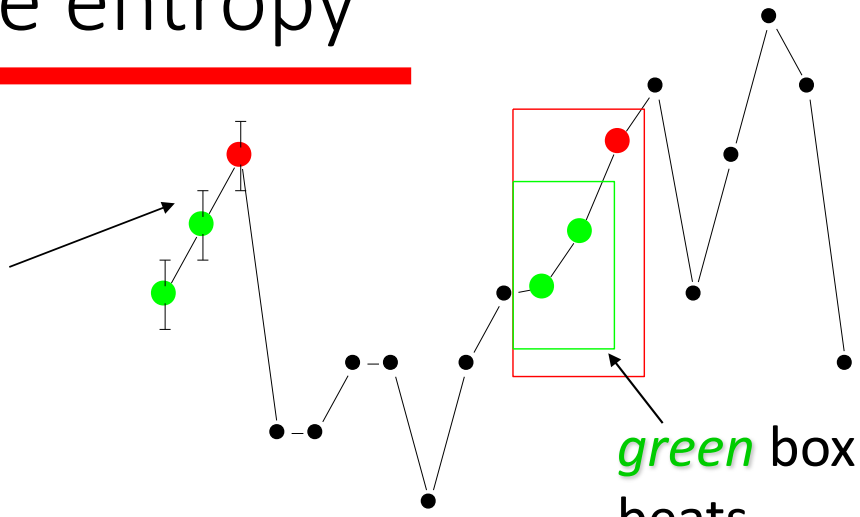
Pincus 1991

$$A_E(m,r,N) = \Phi^m(r) - \Phi^{m+1}(r)$$

$$A_E(m,r,N) \approxeq \frac{1}{N-m} \sum_{i=1}^{N-m} \ln \frac{n_i^m}{n_i^{m+1}}$$

Approximate entropy

bars are *r* msec



green box is *m* beats

red box is *m*+1 beats

A = match of length *m+1*
B = match of length *m*

Approximate Entropy ≈ Σ -ln (1+Σ**A**) / (1+Σ**B**)

For regular, repeating data, Σ A / ΣB nears 1 and entropy nears 0.

Pincus 1991

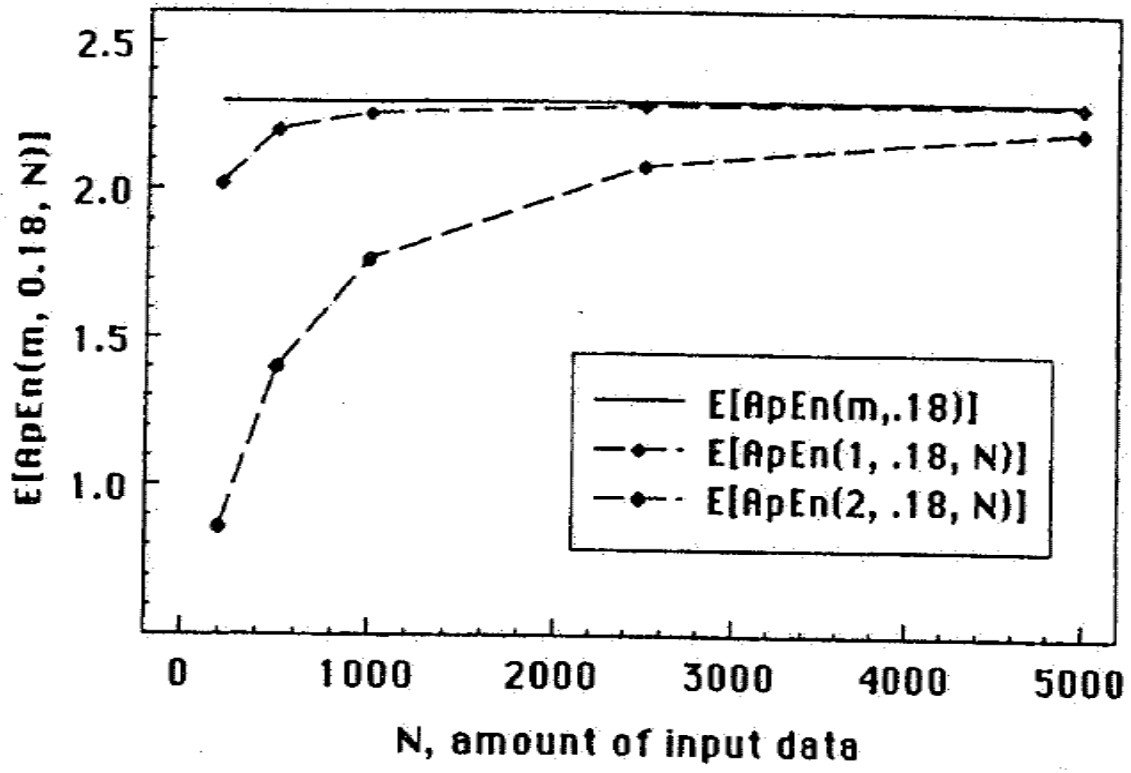
On the basis of calculations that included the above theoretical analysis, I drew a preliminary conclusion that, for $m = 2$ and $N = 1000$, choices of r ranging from 0.1 to 0.2 SD of the $u(i)$ data would produce reasonable statistical validity of ApEn(m, r, N). For smaller r values, one usually achieves poor conditional probability estimates in Eq. 8, while for larger r values, too much detailed system information is lost. To avoid a significant contribution from noise in an ApEn calculation, one must choose r larger than most of the noise.

Pincus and Huang 1992

C. *Parameter Choices*

The selection of values of m and r for the ApEn(m,r,N) statistic should depend on the amount of available data. We generally would like to choose r as small and m as large as possible. The tradeoff is given by the requirement of statistical validity, given by a small ApEn standard deviation, for a specified amount of data. In many applications, we anticipate between 100 and 5000 input data points. Based on calculations that included theoretical analyses of deterministic and stochastic processes (Pincus, 1991; Pincus and Keefe, 1992) and clinical applications (Pincus, Gladstone and Ehrenkranz, 1991; Kaplan et. al., 1991), we have concluded that for m=2 and N=1000, values of r between 0.1 to 0.25 standard deviations of the u(i) data produce good statistical validity of ApEn(m,r,N). For smaller r values, one usually achieves poor conditional probability estimates, while for larger r values, too much detailed system information is lost.

Pincus 1991: ApEn is biased



Problems with ApEn

- Bias: arises from allowing pairs to match themselves (as allowed by Eckmann and Ruelle, though explicitly excluded by Grassberger and Procaccia) so as to avoid log 0/ log 0

- Leads to error of unknown magnitude (larger for fewer matches) and threatens relative consistency
- How to choose *r*?
- How to choose *m*?

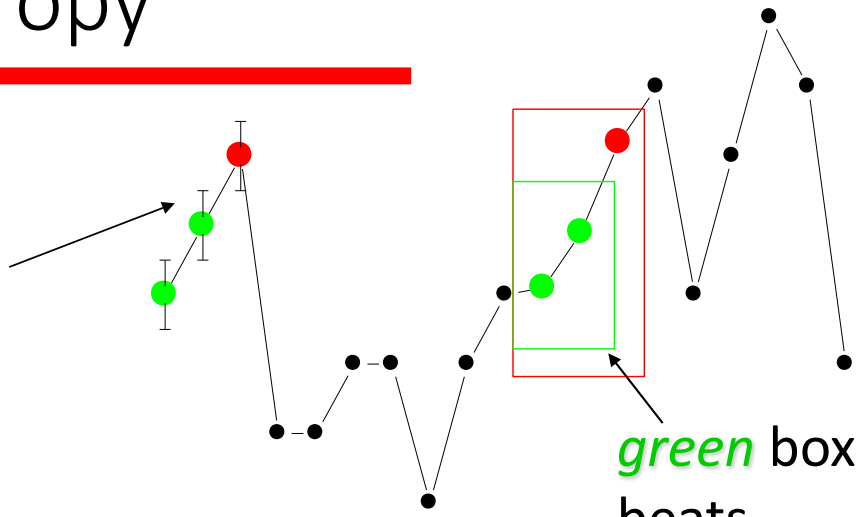
Richman and Moorman 2000

- We wished to apply ApEn to the problem of neonatal sepsis
- We sought to remove bias by removing the template-wise approach to counting

$$S_E(m,r,N) = \ln \frac{\sum_{i=1}^{N-m} n_i'^{m}}{\sum_{i=1}^{N-m} n_i'^{m+1}}$$

Sample entropy

bars are *r* msec



green box is *m* beats

red box is *m*+1 beats

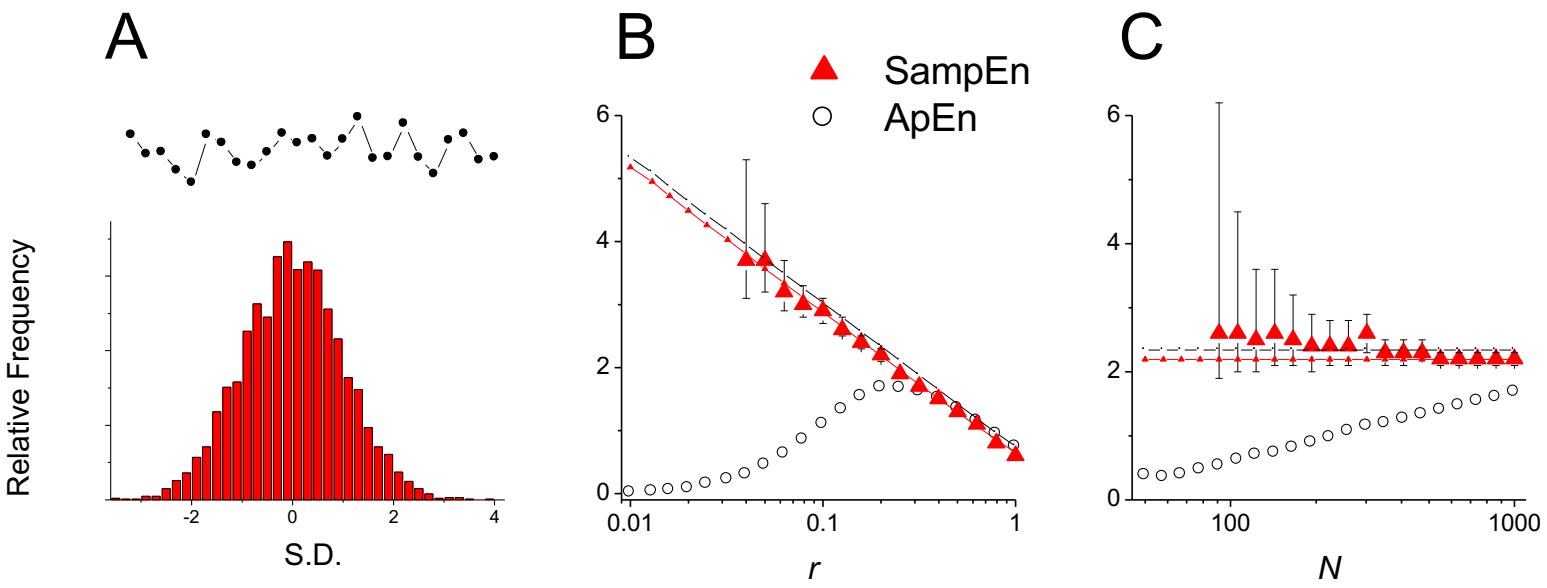
A = match of length *m+1*
B = match of length *m*

Sample Entropy = -ln Σ A / ΣB

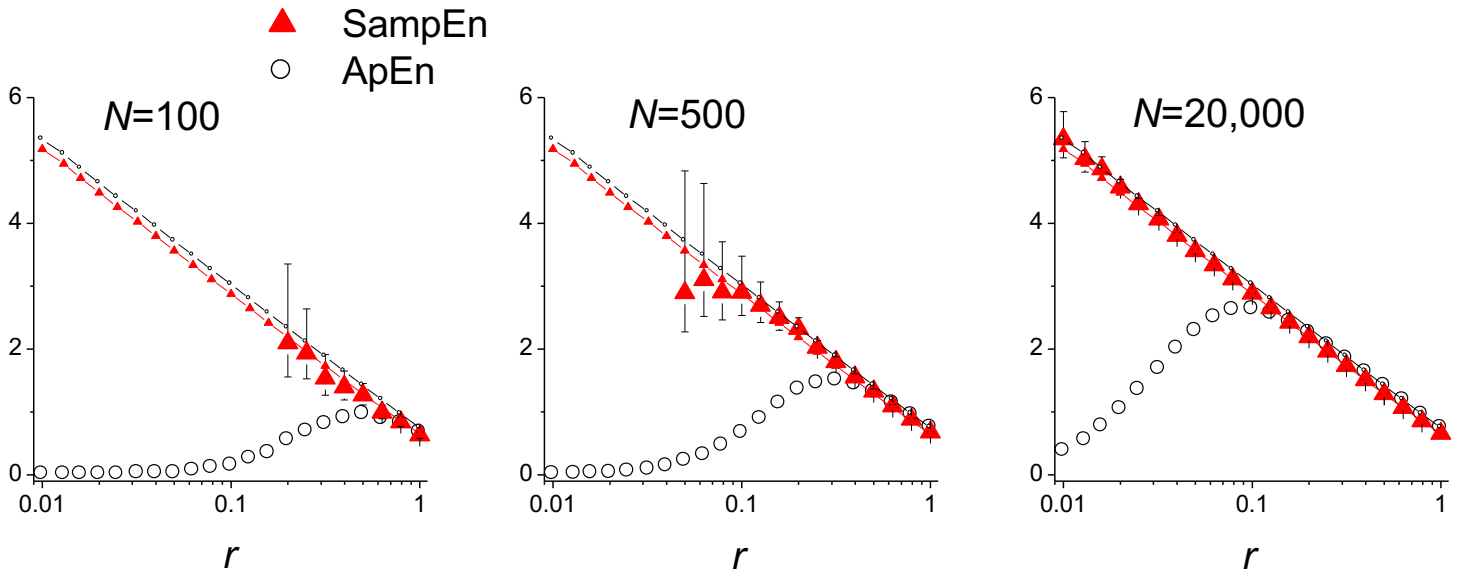
Approximate Entropy ≈ Σ -ln (1+ΣA) / (1+ΣB)

For regular, repeating data, Σ A / ΣB nears 1 and entropy nears 0.

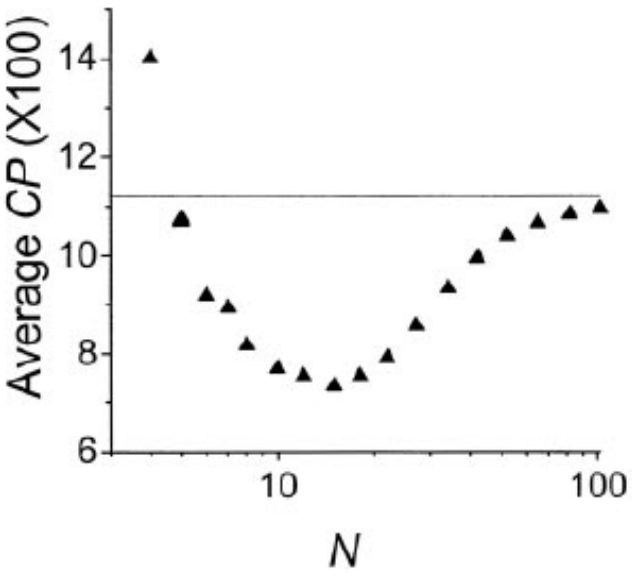
Bias in entropy estimates



Bias in entropy estimates



Non-independence of templates leads to bias



Templates are allowed to overlap, and data points can appear as part of the template or as the m+1st point.

This is relieved if templates are disjoint.

For long time series, the bias is small.

Richman 2006

Proposition 4.1. *The asymptotic variance of SampEn(m, r, n). is given by*

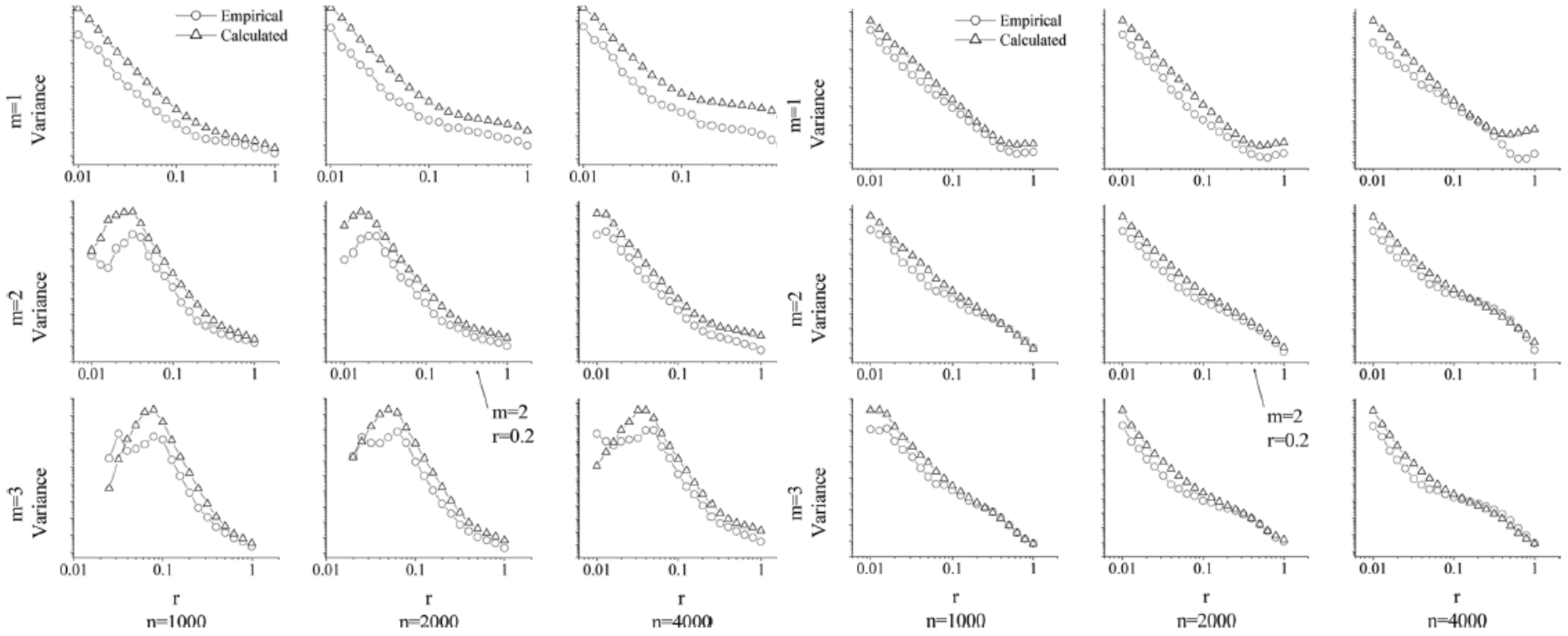
$$\sigma^2_{S(m,r,n)} \to \frac{\sigma^2(m)}{p_m^2} - 2\frac{\sigma^2(m \mid m+1)}{p_m p_{m+1}} + \frac{\sigma^2(m+1)}{p_{m+1}^2}$$

and estimated by

$$\hat{\sigma}^2_{S(m,r,n)} \cong \frac{\hat{\sigma}^2(m)}{\hat{p}_m^2} - 2\frac{\hat{\sigma}^2(m \mid m+1)}{\hat{p}_m \hat{p}_{m+1}} + \frac{\hat{\sigma}^2(m+1)}{\hat{p}_{m+1}^2},$$

where the individual components are as defined above.

Observed vs expected variances



Hypothesis testing using SampEn

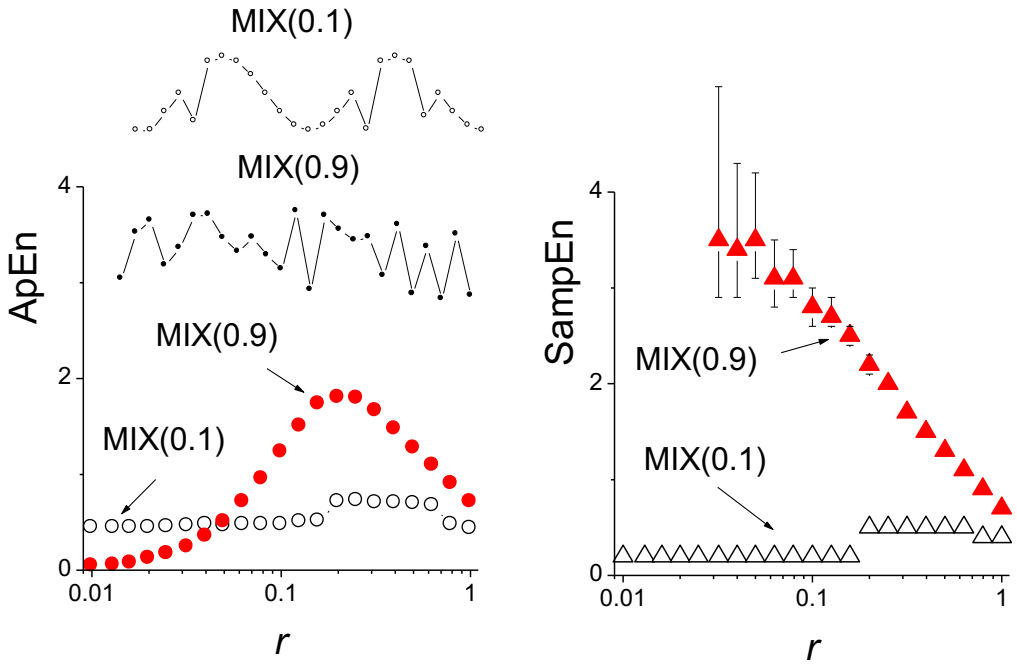
Proposition 5.1. *The test statistic* $SampEn(m-1, r, n) - SampEn(m, r, n)$ *is asymptotically normal with approximate variance*

$$\frac{\sigma^2(m-1)}{p_{m-1}^2} + 4\frac{\sigma^2(m)}{p_m^2} + \frac{\sigma^2(m+1)}{p_{m+1}^2} - 4\left(\frac{\sigma^2(m-1\,|\,m)}{p_{m-1}p_m} + \frac{\sigma^2(m\,|\,m+1)}{p_m p_{m+1}}\right) + 2\frac{\sigma^2(m-1\,|\,m+1)}{p_{m-1}p_{m+1}}.$$

Under the null hypothesis that there is no significant difference between $SampEn(m-1, r, n)$ *and* $SampEn(m, r, n)$*, the statistic is expected to have a mean of zero.*

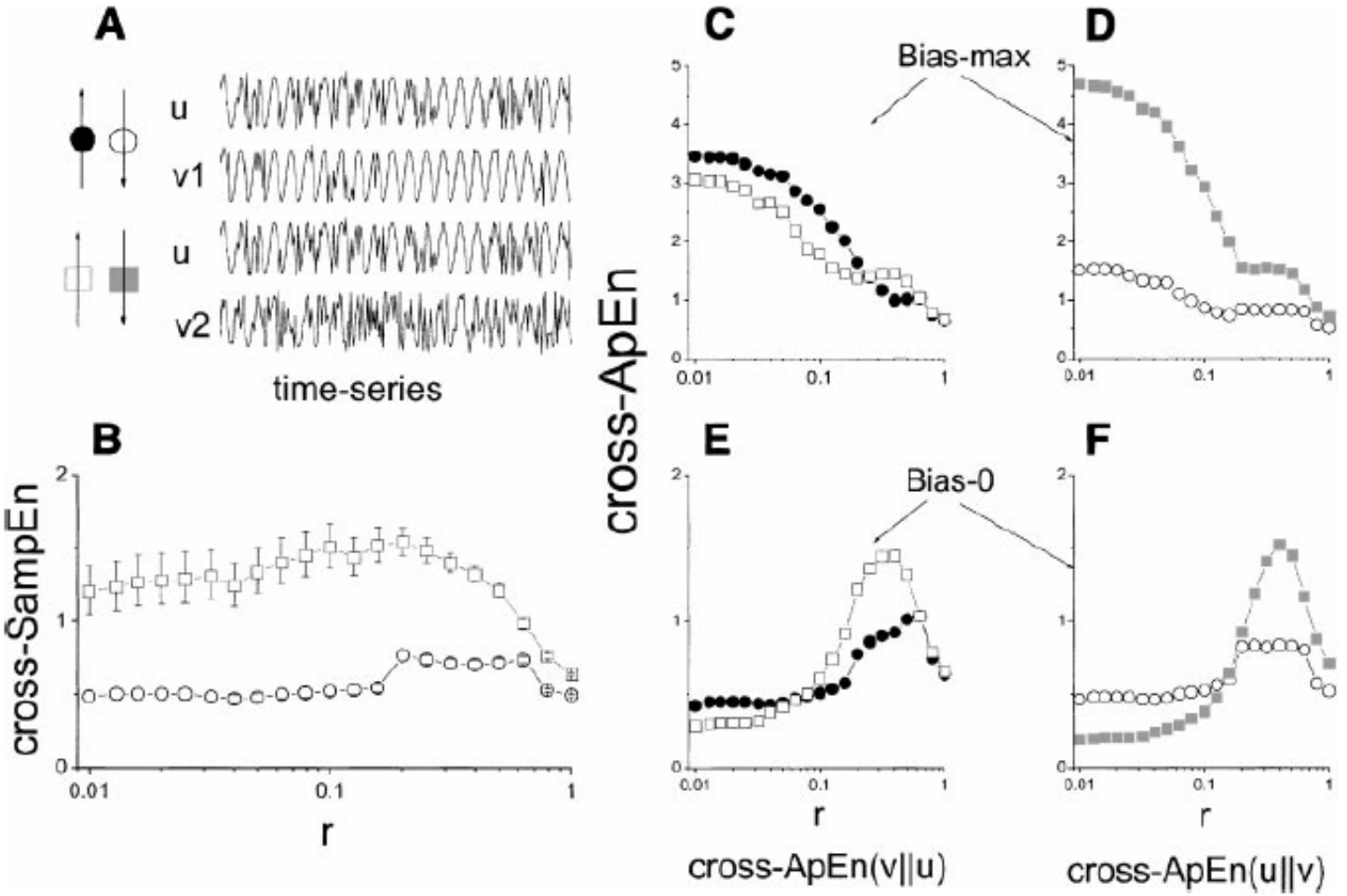
Note: use this as a means of picking *m*

Cross-entropy



Now think about 2 time series, say, from different organs that are networked together, and calculate the entropy using 1 series for the original templates and the other for the possible matches.

Richman, Moorman 2000



Richman, Moorman 2000

Lake 2011

Took a stochastic point of view, and converted the conditional probability to a density by normalizing for the matching volume $(2r)^m$

$$QSE = -\log\left(\frac{CP}{2r}\right) = -\log CP + \log 2r = SampEn + \log 2r$$

Note: use this as a means of adjusting for different values of *r*

PHYSICAL REVIEW E 95, 062114 (2017)

Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations

Wanting Xiong,[1,2] Luca Faes,[3] and Plamen Ch. Ivanov[2,4,5,*]

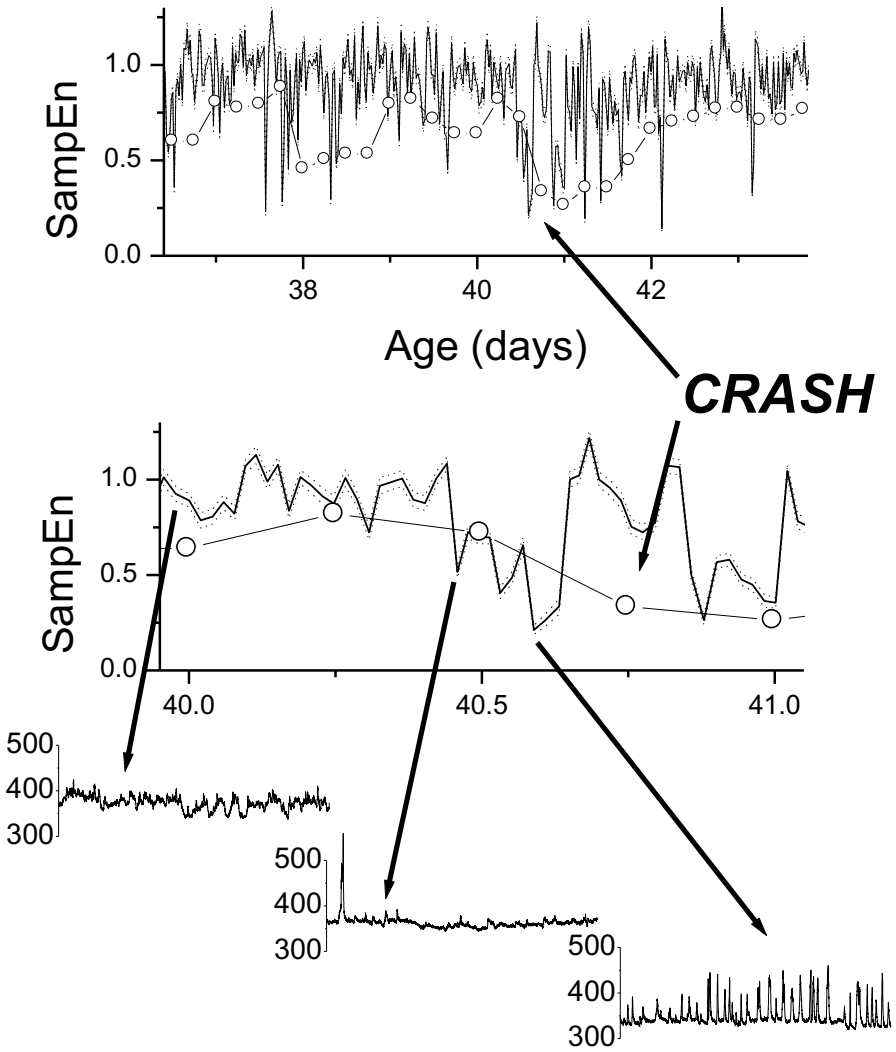
Required reading!

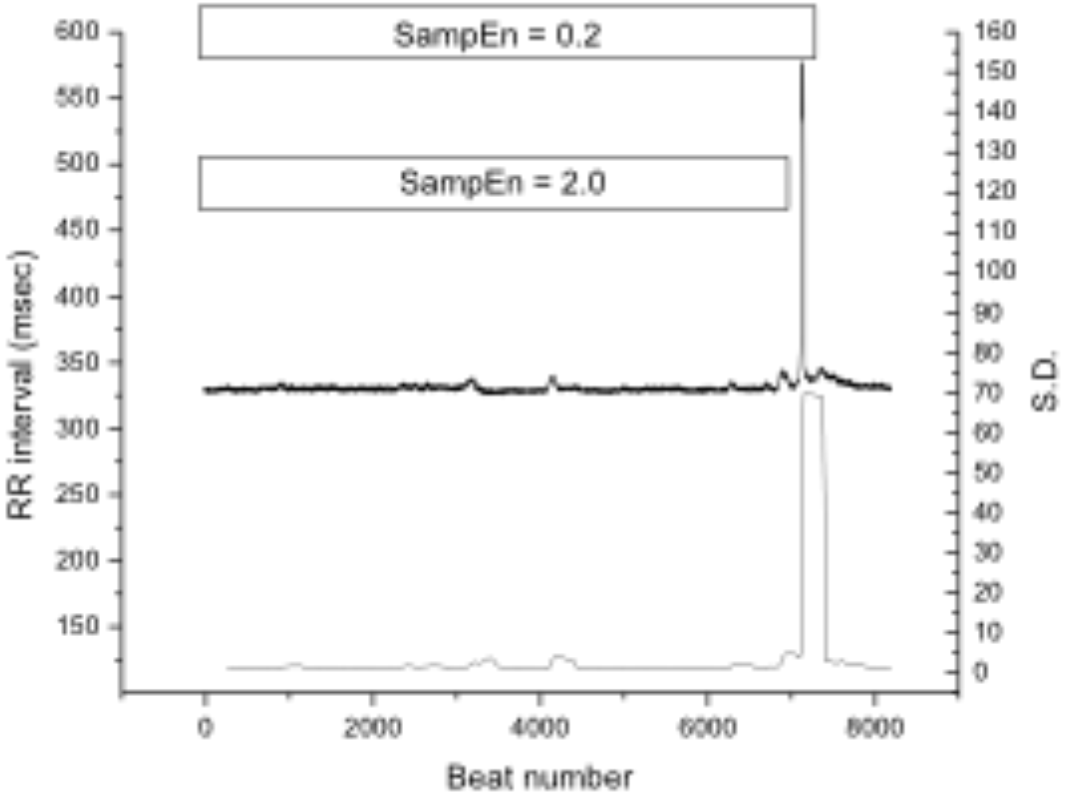
Systematically examines, among other things:

- parameter selection
- non-stationarities, like spikes
- long-range correlations

UVa group

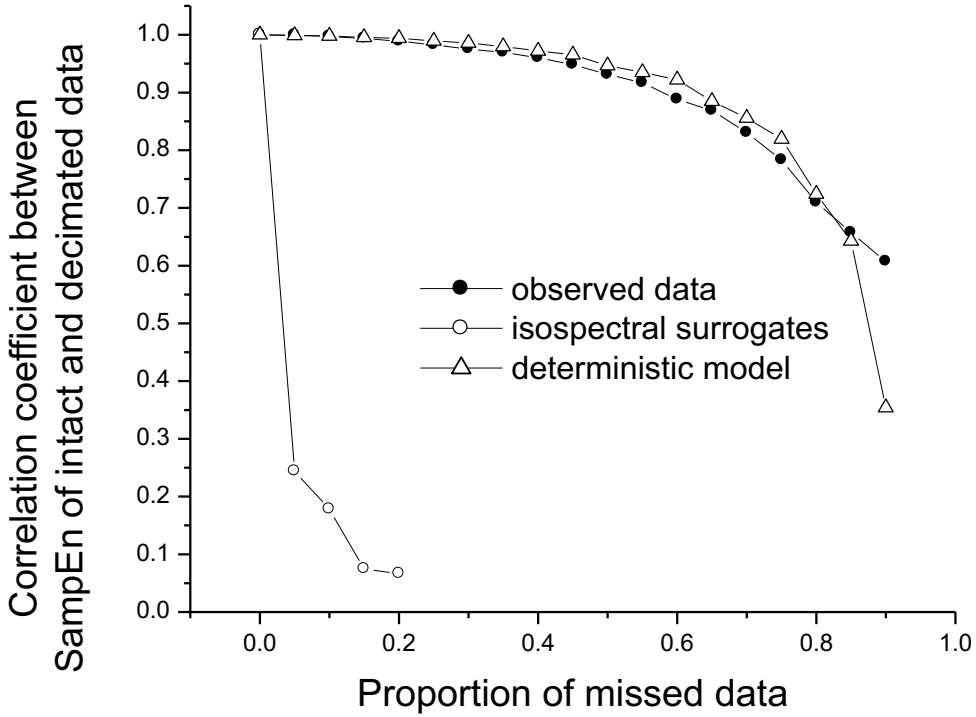
- Since 2000, we have been pondering some of this.
- In particular, we have been interested in why it is that entropy falls before neonatal sepsis, and how to pick *r* and *m*
- In brief:
  - Entropy falls before neonatal sepsis because the data are non-stationary and have spikes, not because of a change in order or regularity
  - Stationarity of heart rate is, in fact, quite elusive
  - We pick *r* such that the numerator count is sufficient, and then adjust the entropy estimate for the *r* that we chose
  - We pick *m* based on autocorrelation, Richman suggested a more elegant way based on the difference between entropy estimates



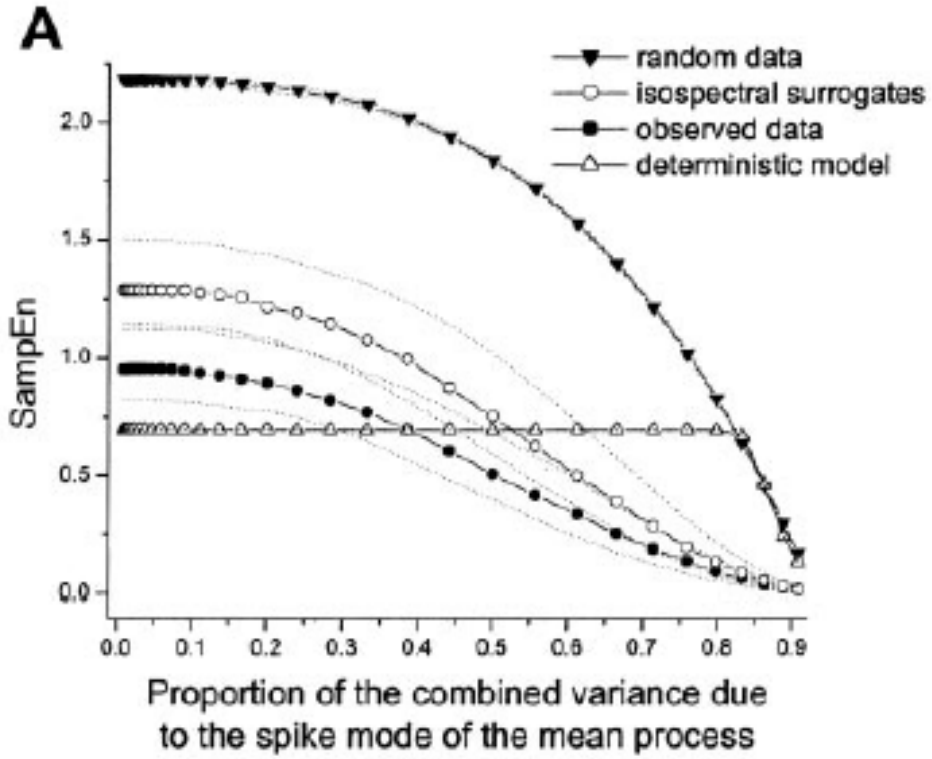


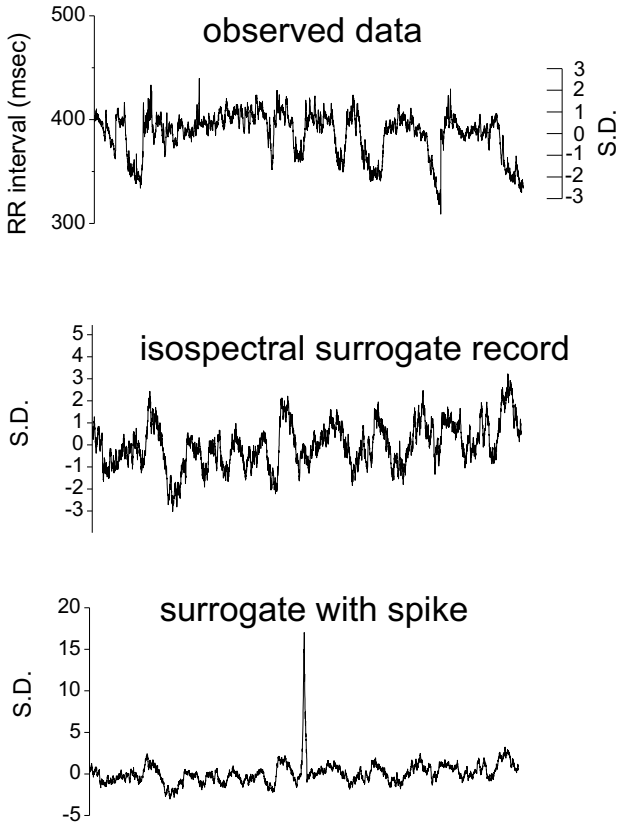
Lake…Moorman 2002

In records with spikes, SampEn is not detecting deterministic order



Spikes reduce sample entropy





Lake…Moorman 2002

Spikes reduce sample entropy

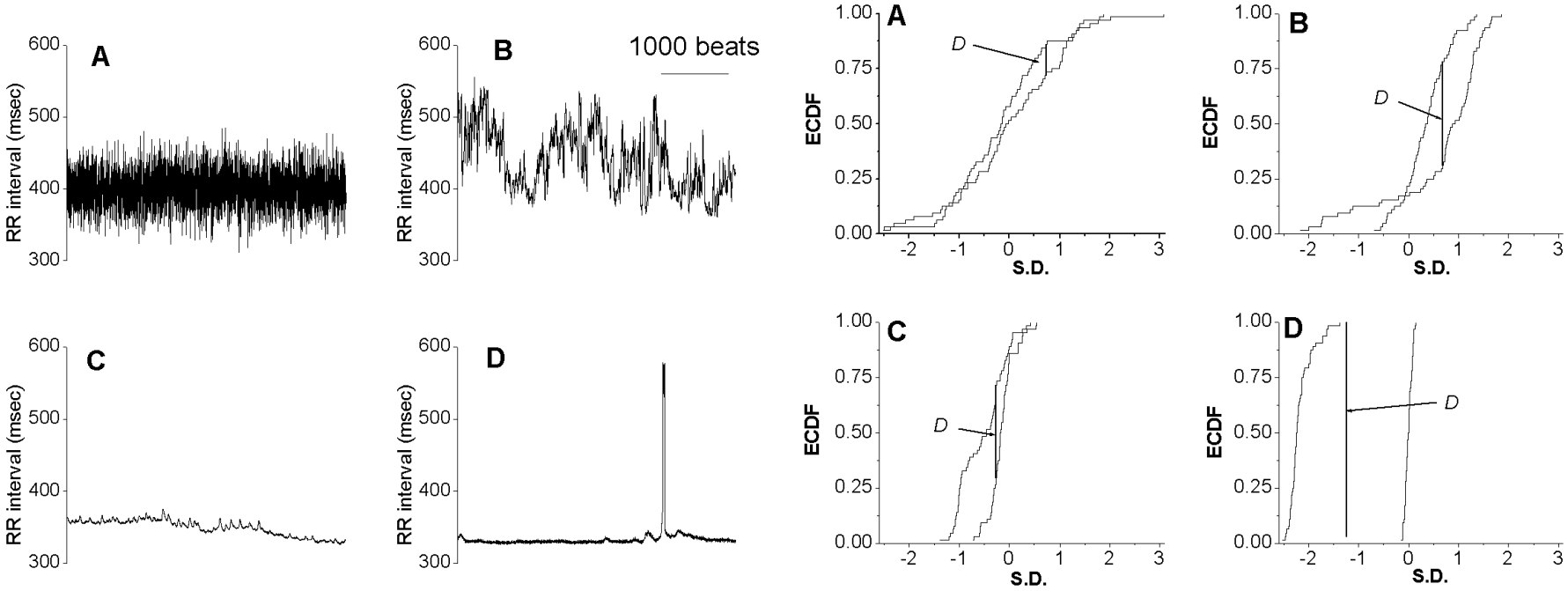
$$\mathbf{SampEn}(m,r,N) \approx -\log\left(\frac{r}{\sqrt{\pi}}\right) - \frac{\Delta^2\varepsilon(1-\varepsilon)}{2\sigma_b^2}$$

Where $\Delta =$ the height and $\varepsilon N =$ the number of beats in a spike, and

$\Delta^2\varepsilon(1-\varepsilon)$ is the variance added by the spikes

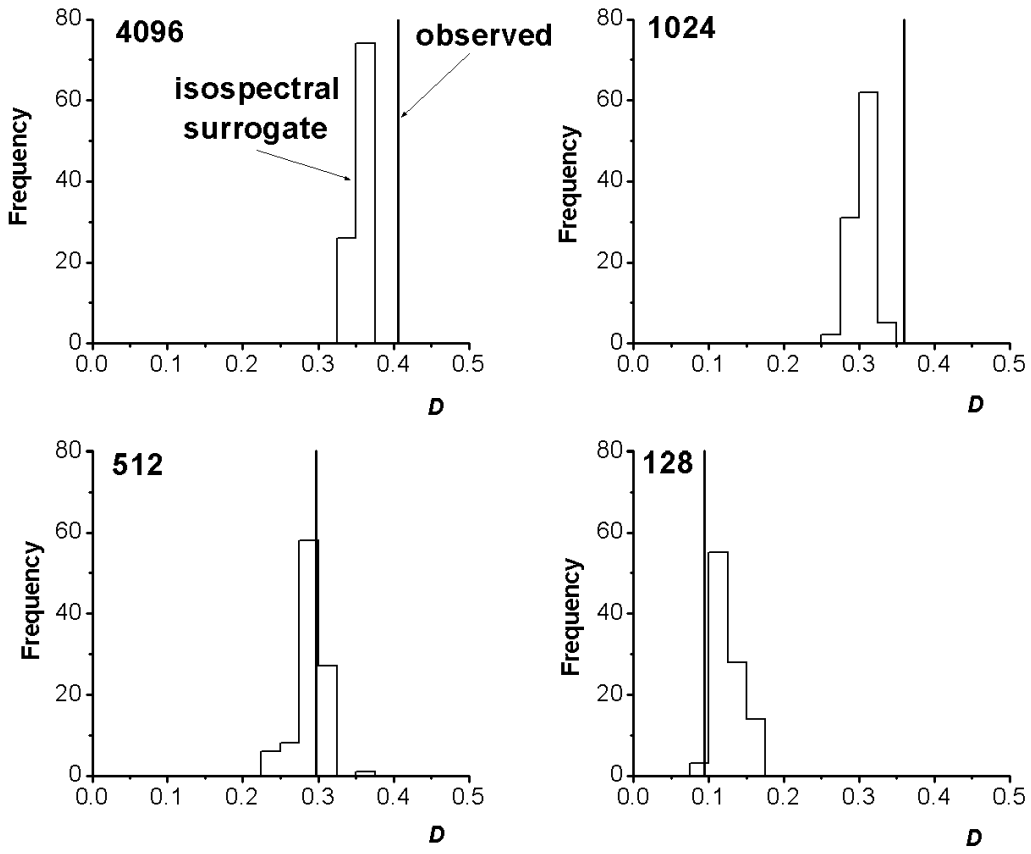
Lake…Moorman 2002

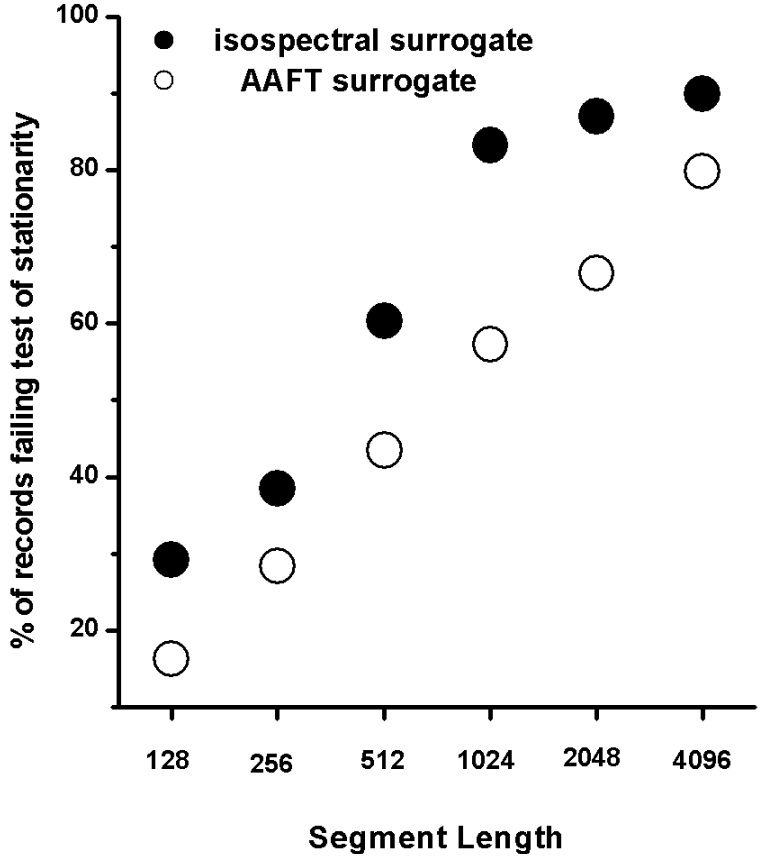
A KS test for heart rate stationarity



Cao, Lake, Moorman 2003

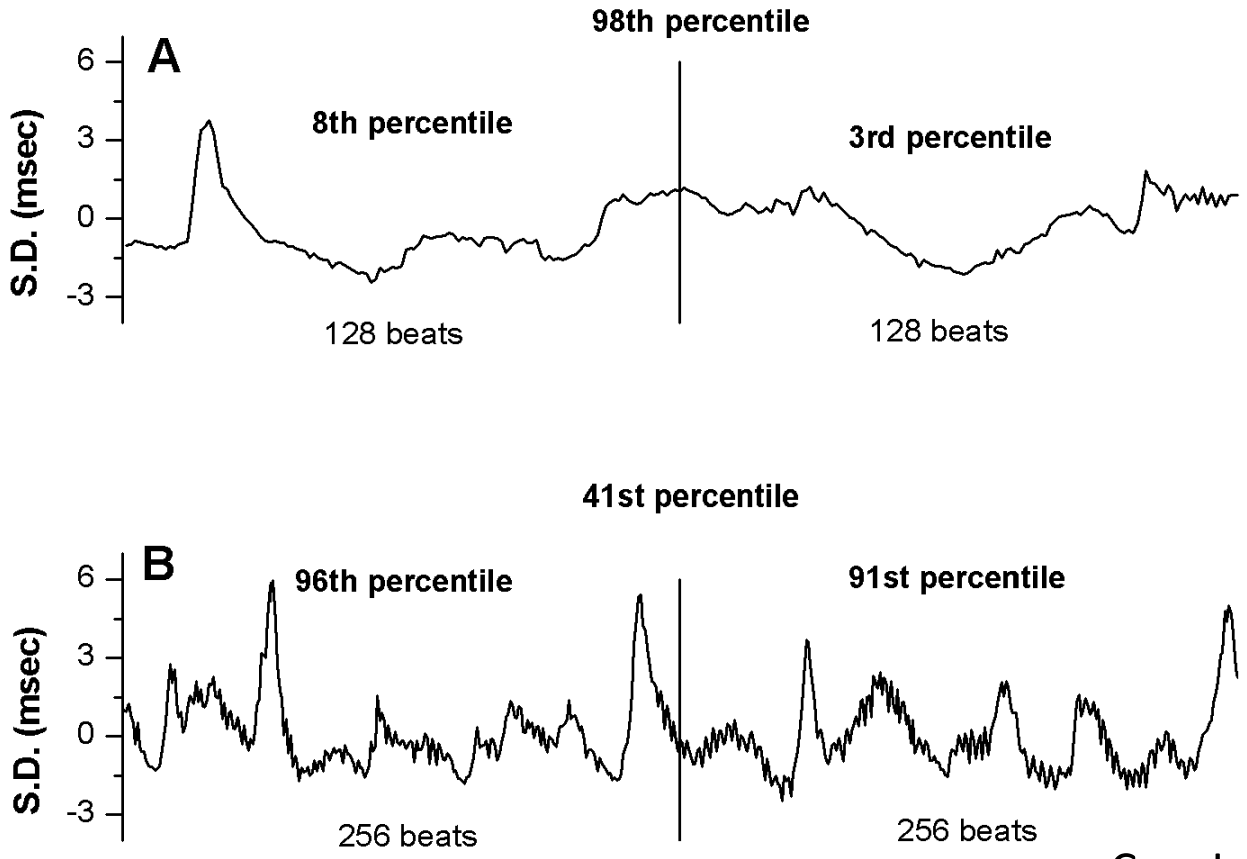
A KS test for heart rate stationarity





Cao, Lake, Moorman 2003

Heart rate stationarity is an elusive matter



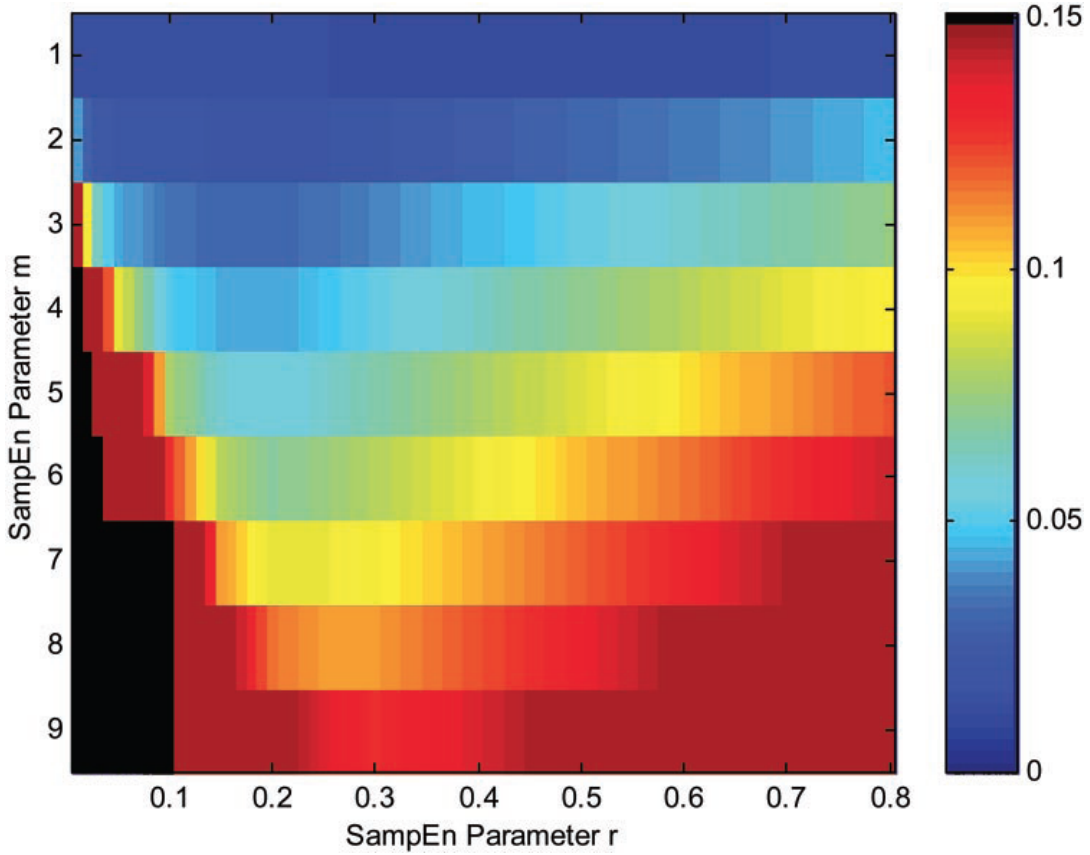
Cao, Lake, Moorman 2003

How to pick *m* and *r* with brute force

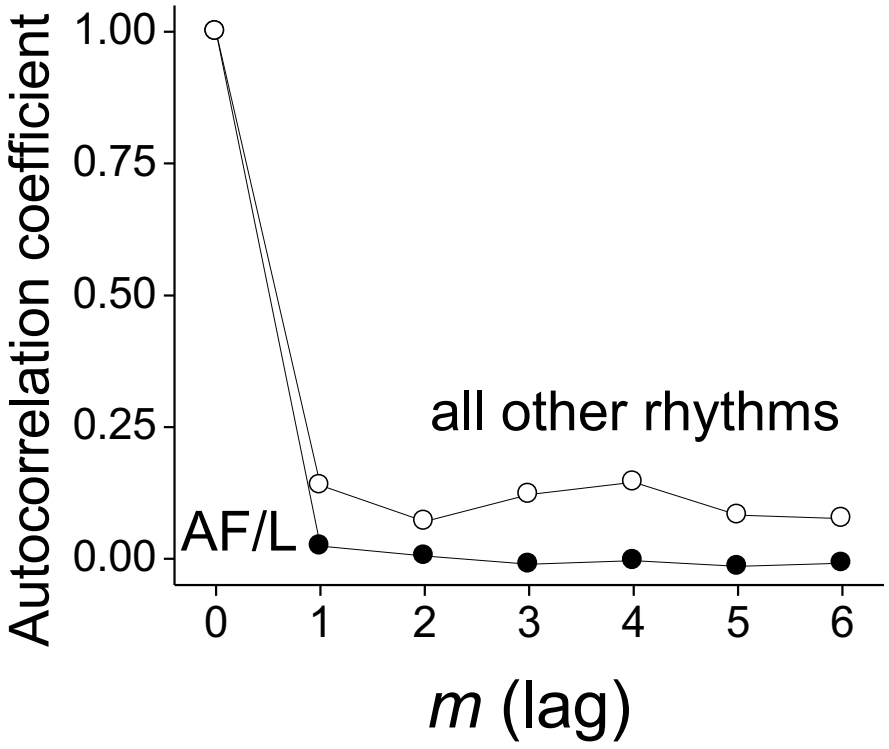
In neontatal HR data, we sought to minimize the standard error of the CP and SampEn estimates.
The heat map plots:

$$\max\left(\frac{\sigma_{CP}}{CP}, \frac{\sigma_{CP}}{-\log(CP)CP}\right)$$

Lake…Moorman 2002



How to pick *m*



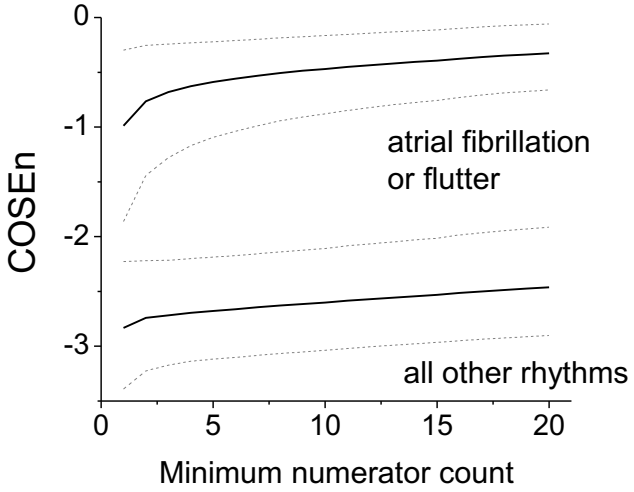
Atrial fibrillation (AF), a common cardiac arrhythmia, has uncorrelated heartbeat intervals, and *m* =1 is sensible.

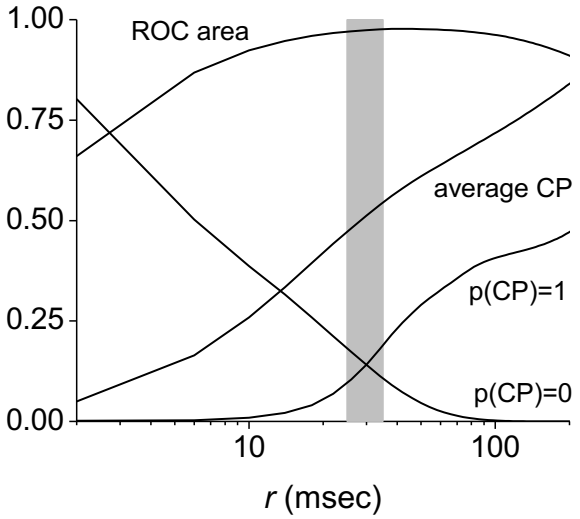
This makes QSE (a related metric called COSEn, in fact) an efficient AF detector in as few as 10 beats.

Lake, Moorman 2011

How to pick *r*

Since we can use QSE/COSEn to adjust for whatever *r* we pick, we suggest picking a value that allows enough matches of length *m*+1 so that we can have confidence in the CP statistic

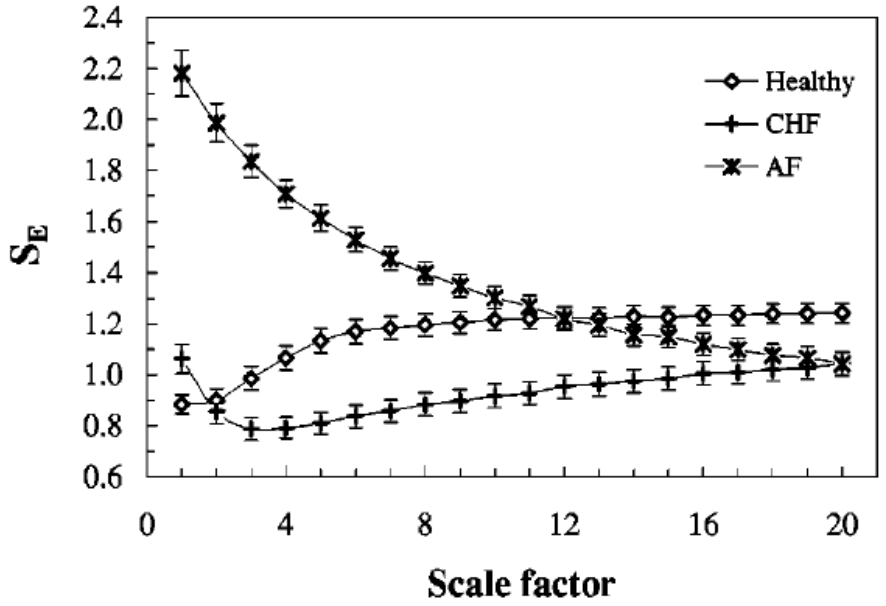




Lake, Moorman 2011

Costa, Goldberger, Chen 2005

Multiscale entropy – sample entropy meets detrended fluctuation analysis (DFA)



Lee, Nemati, others 2013

Transfer entropy measures the reduction in uncertainty in yi given past x_i and y_i compared to only y_i.

$$T_{X\to Y}(\tau) = \sum_{y_i, y_{i-1}, x_{i-\tau}} p(y_i,\ y_{i-1},\ x_{i-\tau}) \log \frac{p(y_i|y_{i-1}, x_{i-\tau})}{p(y_i|y_{i-1})}$$

and determines changes in coupling between two time series

Other newer versions

- …where MSE = multiscale entropy
- Refined composite MSE (rcMSE)
- $MSE_{moments}$
- *M*ultivariate MSE (*M*MSE)
- *M*ultivariate refined composite MSE (*M*rcMSE)
- *M*ultivariate *G*eneralized MSE (*MG*MSE)
- *M*ultivariate *G*eneralized refined composite MSE (*MG*rcMSE)
- Cross-entropy versions? Your name here…

Conclusions

- The concept of entropy of physiological time series has a very interesting and non-linear history.
- Entropy estimates such as sample entropy have been very widely applied, sometimes sensibly.
- Before interpreting results, it is important to consider non-entropy causes for changes in the results of entropy estimates.
- A low value of approximate entropy means bias, spikes or order
- A low value of sample entropy means spikes or order