Entropy and information estimation: An overview

Ilya Nemenman

December 12, 2003

Workshop schedule: Morning

7:30 – 7:55 Ilya Nemenman, *Entropy and information estimation: A general overview.*

7:55 – 8:20 Liam Paninski, *Estimating entropy on m bins with fewer than m samples.*

8:20 – 8:45 Jose Costa, *Applications of entropic graphs in nonparametric estimation.*

8:45 – 8:55 *Coffee break.*

8:55 – 9:20 Ronitt Rubinfeld, *The complexity of approximating the entropy.*

9:20 – 9:45 Jonathan Victor, *Metric-space approach to information calculations, with application to single-neuron and multineuronal coding in primary visual cortex.*

9:45 – 10:30 *Discussion.*

Workshop schedule: Evening

4:00 – 4:25 William Bialek, *Entropy and information in spike trains: Why are we interested and where do we stand?*

4:25 – 4:50 Jon Shlens, *Estimating Entropy Rates and Information Rates in Retinal Spike Trains.*

4:50 – 5:15 Yun Gao, *Lempel-Ziv Entropy Estimators and Spike Trains.*

5:15 – 5:25 *Coffee break.*

5:25 – 5:50 Pamela Reinagel, *Application of some entropy estimation methods to large experimental data sets from LGN neurons.*

5:50 – 6:15 Gal Chechik, *Information bearing elements and redundancy reduction in the auditory pathway.*

6:15 – 7:00 *Discussion.*

Why is this an interesting problem?

- information content of (symbolic) sequences
  - spike trains
  - bioinformatics
  - linguistics
  - prediction games (Cover)
  - . . .
- dimensions of strange attractors (Grassberger et al.)
- complexity of dynamics

Leave aside average vs. single sequence problem.

Why is this a difficult problem? (first try)

$$\lim_{p \to 0} \frac{p \log p}{p} \quad = \quad \infty$$

Why is this a difficult problem? (first try)

$$\begin{aligned}
\lim_{p\to 0}\frac{p\log p}{p} &= \infty \\
S(\hat{p}) &\equiv -\hat{p}\log\hat{p} - (1-\hat{p})\log(1-\hat{p}) \text{ is convex} \\
\Longrightarrow & \; E\,S(\hat{p}) < S(E\,\hat{p}) = S(p)
\end{aligned}$$

Why is this a difficult problem? (first try)

$$\begin{aligned}
\lim_{p\to 0} \frac{p\log p}{p} &= \infty \\
S(\hat{p}) &\equiv -\hat{p}\log\hat{p} - (1-\hat{p})\log(1-\hat{p}) \text{ is convex} \\
&\Longrightarrow E\,S(\hat{p}) < S(E\,\hat{p}) = S(p)
\end{aligned}$$

- events of negligible probability may have large entropy [Rubinfeld]
- small errors in $p \Longrightarrow$ large errors in S
- negative bias (more later) [all]

$$S(\text{best } p) \neq \text{ best } S(p)$$

Entropy vs. information

$$I(X,Y) = S(X) + S(Y) - S(X,Y) = S(X) - S(X|Y)$$

Entropy vs. information

$$I(X,Y) = S(X) + S(Y) - S(X,Y) = S(X) - S(X|Y)$$

- we are interested in information
- no context–free information (information *about* something)
- entropy has no continuous limit

Different entropies

Shannon	$S = -\sum p_i \log p_i$
Renyi	$R_\alpha = \frac{1}{1-\alpha} \log \sum p_i^\alpha$
Burg	$B = \sum \log p_i$
. . .	. . .

Different entropies

Shannon	$S = -\sum p_i \log p_i$
Renyi	$R_\alpha = \frac{1}{1-\alpha} \log \sum p_i^\alpha$
Burg	$B = \sum \log p_i$
. . .	. . .

Easier to estimate R_α, $\alpha \geq 2$ (Grassberger, 2003).

Different entropies

Shannon	$S = -\sum p_i \log p_i$
Renyi	$R_\alpha = \frac{1}{1-\alpha} \log \sum p_i^\alpha$
Burg	$B = \sum \log p_i$
...	...

Easier to estimate R_α, $\alpha \geq 2$ (Grassberger, 2003).

- Can we use $\lim_{\alpha \to 1} R_\alpha$ to estimate S? [Costa]

- Can we use R_α to bound S? [Bialek]

Types of convergences

(Beirlant et al. 1997)

- weak: $S_N \longrightarrow S$ in probability
- mean square: $E(S_N - S)^2 \longrightarrow 0$
- strong: $S_N \longrightarrow S$ a. s.
- asymptotic normality: $\lim \sqrt{N}(S_N - S) \sim \mathcal{N}(0, \sigma^2)$ (Gabrielli et al., 2003)
- distribution (L_2): $NE(S_N - S)^2 \rightarrow \sigma^2$

Continuous variables

(Beirlant et al. 1997)

Differential entropy $S[P] = -\int dx P(x) \ln P(x)/Q(x)$

Continuous variables

(Beirlant et al. 1997)

Differential entropy $S[P] = -\int dx P(x) \ln P(x)/Q(x)$

Conditions on $P(x)$ (w. r. t. Q):

- smoothness
- light tails
- small peaks (bounded)

Continuous variables

(Beirlant et al. 1997)

Differential entropy $S[P] = -\int dx P(x) \ln P(x)/Q(x)$

Conditions on $P(x)$ (w. r. t. Q):

- smoothness
- light tails
- small peaks (bounded)

Always undersampled, but convergence (and rates) are calculable.

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\mathrm{est}} \log P_{N,\mathrm{est}}$

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\mathrm{est}} \log P_{N,\mathrm{est}}$
- (1d) spacing entropy estimators

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\mathrm{est}} \log P_{N,\mathrm{est}}$
- (1d) spacing entropy estimators
- nearest neighbor distances, any dimension

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\mathrm{est}} \log P_{N,\mathrm{est}}$
- (1d) spacing entropy estimators
- nearest neighbor distances, any dimension
- average number of neighbors in the vicinity of points (Grassberger and Procaccia, 1983) [Costa]

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\text{est}} \log P_{N,\text{est}}$
- (1d) spacing entropy estimators
- nearest neighbor distances, any dimension
- average number of neighbors in the vicinity of points (Grassberger and Procaccia, 1983) [Costa]
- sieving [Victor, Bialek]

Methods for continuous distributions

- plug-in estimates: $S_N = -\int dx\, P_{N,\mathrm{est}} \log P_{N,\mathrm{est}}$
- (1d) spacing entropy estimators
- nearest neighbor distances, any dimension
- average number of neighbors in the vicinity of points (Grassberger and Procaccia, 1983) [Costa]
- sieving [Victor, Bialek]

Metric is very important!

Discrete variables

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

Discrete variables

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

(light tails, small peaks) $\longrightarrow$ (rank ordered form)

Discrete variables

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

(light tails, small peaks) $\longrightarrow$ (rank ordered form)

(smoothness) $\longrightarrow$???

Discrete variables

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

(light tails, small peaks) $\longrightarrow$ (rank ordered form)

(smoothness) $\longrightarrow$???(maybe also rank plots)

Why is this a difficult problem? (second try) *No go* theorems

For N samples from an i. i. d. distribution over K bins (Note: non-i. i. d. $= K \to \infty$):

- finite alphabets: plug-in and LZ asymptotically consistent, convergence rate (bias) $\sim K/N$

Why is this a difficult problem? (second try)
No go theorems

For N samples from an i. i. d. distribution over K bins
(Note: non-i. i. d. $= K \to \infty$):

- finite alphabets: plug-in and LZ asymptotically consistent, convergence rate (bias) $\sim K/N$
- countable alphabets (or $K \gg N$) (Antos and Kontoyiannis, 2002; Wyner and Foster, 2003)
  - plug-in and LZ are universally consistent (under mild conditions)
  - no universal rate–of–convergence results exist for either

Why is this a difficult problem? (second try) *No go* theorems

For N samples from an i. i. d. distribution over K bins (Note: non-i. i. d. $= K \to \infty$):

- finite alphabets: plug-in and LZ asymptotically consistent, convergence rate (bias) $\sim K/N$

- countable alphabets (or $K \gg N$) (Antos and Kontoyiannis, 2002; Wyner and Foster, 2003)
  - plug-in and LZ are universally consistent (under mild conditions)
  - no universal rate–of–convergence results exist for either

- for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$

  - for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$

- no finite variance unbiased entropy estimators; huge variance, small bias, but nonmonotonic is possible (Grassberger, 2003)

  - for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$

- no finite variance unbiased entropy estimators; huge variance, small bias, but nonmonotonic is possible (Grassberger, 2003)

- no universally consistent multiplicative entropy estimator for $N/K \to 0$, $K \to \infty$ [Rubinfeld]

  - for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$

- no finite variance unbiased entropy estimators; huge variance, small bias, but nonmonotonic is possible (Grassberger, 2003)

- no universally consistent multiplicative entropy estimator for $N/K \to 0$, $K \to \infty$ [Rubinfeld]

- universal consistent entropy estimation is possible only for $K/N \to \text{const}$, $K \to \infty$ [Paninski]

What to do?

- look for unbiased estimators for special cases

What to do?

- look for unbiased estimators for special cases
- look for multiplicative estimation for $S \ll 1$, $1 - S \ll 1$, otherwise for additive (nonuniform in S)

What to do?

- look for unbiased estimators for special cases
- look for multiplicative estimation for $S \ll 1$, $1 - S \ll 1$, otherwise for additive (nonuniform in S)
- "almost" good is enough, especially for $K \gg N$

Methods

asymptotic corrections to maximum likelihood (plug-in, naive); Miller, jackknife, Panzeri–Treves, Grassberger, Paninski

coincidence based Lempel–Ziv (Grassberger), Ma, NSB, Jimenez–Montano et al., [Bialek, Gao, Shlens]

Asymptotic corrections

$$\begin{aligned} S(N) &= S_{\rm ML}(N) + \frac{K^*(\{p\})}{2N} + O\left(\frac{K^*}{N}\right) \\ S_{\rm ML} &= -\sum \hat{p}_{\rm ML} \log \hat{p}_{\rm ML} \end{aligned}$$

Asymptotically, $K^* \to K-1$, otherwise *effective number of bins*.

Estimate: $K^* \geq 2^S \Longrightarrow$

Methods can succeed only for $N \gg 2^S$!

(Some) coincidence–based methods

Ma's (1981) argument, the birthday problem

For uniform K–bin distribution: for $N_c \sim \sqrt{K}$, probability of coincidences ~ 1.

$$S = \log K \approx \log N_c^2 = 2 \log N_c$$

Works better then it should!

(Some) coincidence–based methods

Ma's (1981) argument, the birthday problem

For uniform K–bin distribution: for $N_c \sim \sqrt{K}$, probability of coincidences ~ 1.

$$S = \log K \approx \log N_c^2 = 2 \log N_c$$

Works better then it should!

Works in nonasymptotic regime $N \sim 2^{1/2S}$.

Extensions?

- good entropy estimator $\neq$ good distribution estimator

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]
- imagine sampling sequences of length $m \gg 1$ from N_c samples

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]
- imagine sampling sequences of length $m \gg 1$ from N_c samples
  - $\sim N_c^m$ different sequences

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]
- imagine sampling sequences of length $m \gg 1$ from N_c samples
  - $\sim N_c^m$ different sequences
  - uniformly distributed (equipartition), $p \approx 2^{-mS}$

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]
- imagine sampling sequences of length $m \gg 1$ from N_c samples
  - $\sim N_c^m$ different sequences
  - uniformly distributed (equipartition), $p \approx 2^{-mS}$
  - if i. i. d., then (Ma) $mS = 2 \log N_c^m \Longrightarrow S = \log N_c$

Extensions?

- good entropy estimator $\neq$ good distribution estimator
- NSB method [Bialek]
- imagine sampling sequences of length $m \gg 1$ from N_c samples
  - $\sim N_c^m$ different sequences
  - uniformly distributed (equipartition), $p \approx 2^{-mS}$
  - if i. i. d., then (Ma) $mS = 2\log N_c^m \Longrightarrow S = \log N_c$
  - what happens earlier: *non–independence* or *equipartition*?

Final comments

- Good estimators exist for $N \sim 2^S$.

Final comments

- Good estimators exist for $N \sim 2^S$.

- For $N \ll K$ (or 2^S) no–go theorems exist, . . .

Final comments

- Good estimators exist for $N \sim 2^S$.
- For $N \ll K$ (or 2^S) no–go theorems exist, . . .
- . . . but coincidences can save us for special cases.

Final comments

- Good estimators exist for $N \sim 2^S$.
- For $N \ll K$ (or 2^S) no–go theorems exist, . . .
- . . . but coincidences can save us for special cases.
- Let's search for entropy (not distributions!) estimates and . . .

Final comments

- Good estimators exist for $N \sim 2^S$.

- For $N \ll K$ (or 2^S) no–go theorems exist, . . .

- . . . but coincidences can save us for special cases.

- Let's search for entropy (not distributions!) estimates and . . .

- SEARCH FOR THESE SPECIAL CASES!