



Calculating p-Values in Sequence Analysis

Louis H. Y. Chen
National University of Singapore

International Conference on Bioinformatics (InCoB)
9 - 11 September 2009
Singapore



Outline

- Poisson Approximation and Extreme Values
- Two Examples
- A Poisson Approximation Theorem
- Compound Poisson Approximation
- Number of k -Word Matches between Two Sequences
- Finding Significantly Large Clusters of Palindromes in DNA
- Importance Sampling of Word Patterns in Sequences
- Conclusion



Poisson Approximation and Extreme Values

- Poisson approximation provides a method for calculating probabilities about dependent rare events through approximation by the Poisson distribution.
- The following relation between extreme values and occurrences of events makes it possible to formulate many important and interesting scientific problems in terms of occurrences of dependent rare events.
- If an observation exceeding a certain fixed threshold is regarded as the occurrence of an event, then:

Probability that the maximum of a set of observations exceeds a threshold
= Probability that at least one such event occurs.

- Poisson approximation can then be applied to calculate this probability.

Two Examples

Example 1: DNA sequence matching

- To test for similarity or homology between two DNA sequences, slide the first sequence along the second sequence. Find the alignment which gives the longest perfect match.

```
CCCAACACCCAAATATGGCTCGAGAAAGGGCAGCGACATTC
                CGGGGCAAACGAGAAAGGGCAGGTCGAGAAGAGAACC
```

An alignment that gives a perfect match of 13 letters long (see red region).

- Then use Poisson approximation to calculate the probability that a perfect match of such a length occurs.
- Probability that the length of the longest perfect match exceeds a threshold = Probability that at least one perfect match exceeds the threshold

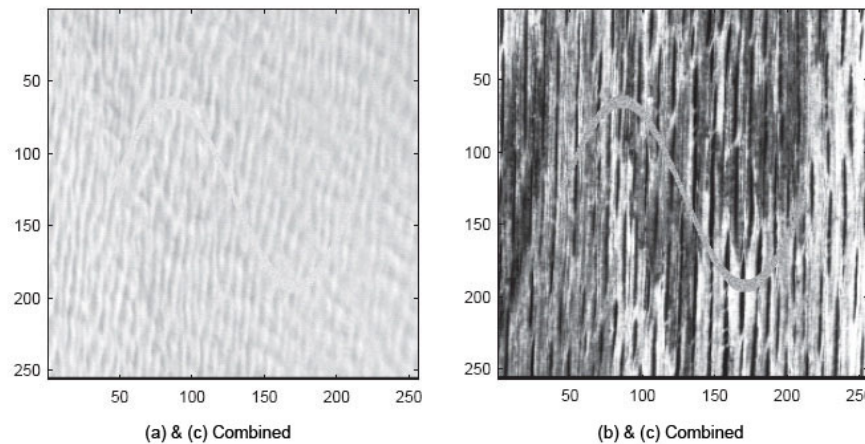
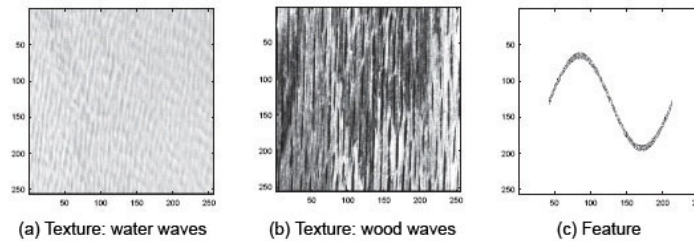
Kenneth Lange, *Mathematical and Statistical Methods for Genetic Analysis*, 2nd Edition, Springer, 2002

R. A. Lippert, H. Huang and M. S. Waterman (2002), *PNAS*

Two Examples

Example 2: Detecting structures in noisy images

Embedded curves in very noisy images

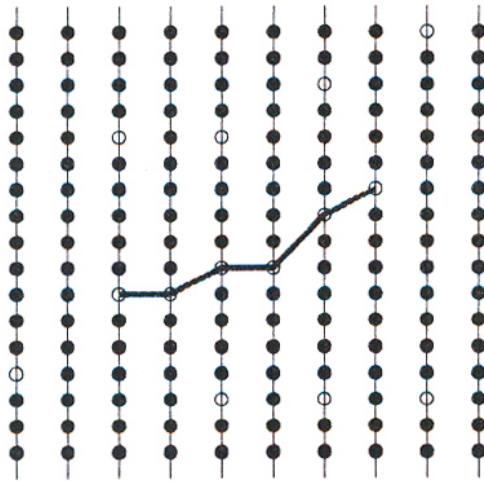


Source: Xiaoming Huo and Jihong Chen (2004)

Two Examples

Arias-Castro, Donoho and Huo (2006), *Ann. Statist.*; Chen and Huo (2006), *JASA*

- To test for existence of an embedded curve, find the longest significance run.



Hollow nodes are significant. A significance run is a chain with all its nodes being significant. In the figure is a significance run of length 5, where $C = 2$.

- Then use Poisson approximation to calculate the probability that a significance run of such a length occurs.
- Probability that the longest significance run exceeds a threshold = Probability that at least one significance run exceeds the threshold

A Poisson Approximation Theorem

Theorem

Chen (1975), *Ann. Probab.*;

Arratia, Goldstein and Gordon (1989), *Ann. Probab.*; (1990) *Statist. Sci.*

Suppose there are n possibly dependent events A_1, \dots, A_n .

Let B_i be a subset of $\{1, 2, \dots, n\}$.

Then

$$\begin{aligned} \sum_{k=0}^{\infty} \left| P(k \text{ events occur}) - \frac{\lambda^k e^{-\lambda}}{k!} \right| \\ \leq 2 \left[(1 \wedge \lambda^{-1})(b_1 + b_2) + (1 \wedge 1.4\lambda^{-1/2})b_3 \right] \end{aligned}$$

and

$$\begin{aligned} |P(\text{At least one event occurs}) - (1 - e^{-\lambda})| \\ \leq (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3) \end{aligned}$$



A Poisson Approximation Theorem

where

$$b_1 = \sum_{i=1}^n \sum_{j \in B_i} p_i p_j, \quad b_2 = \sum_{i=1}^n \sum_{i \neq j \in B_i} p_{ij},$$

$$b_3 = \sum_{i=1}^n E|P(A_i | A_j : j \in B_i^c) - p_i|$$

with

$$p_i = P(A_i \text{ occurs}), \quad p_{ij} = P(A_i \text{ and } A_j \text{ occur}).$$

- Local dependence - For each i , the event A_i is independent of the events outside the set B_i . In this case, $b_3 = 0$.
- If we assume that the base pairs in an DNA are independent random variables, then we have local dependence for most problems concerning word patterns.

A Poisson Approximation Theorem

Idea of proof: (Stein (1972), *Proc. Sixth Berkeley Symp.*; Chen (1975), *Ann. Probab.*)

- If W represents the number of occurrences of dependent rare events and Z represents a Poisson random variable such that their expected values satisfy $EW = EZ = \lambda$, then for any A set of non-negative integers,

$$P(W \in A) - P(Z \in A) = E \{ \lambda f_A(W + 1) - W f_A(W) \}$$

where f_A is a bounded solution of the difference equation

$$\lambda f(w + 1) - w f(w) = I(w \in A) - P(Z \in A).$$

- Thus in approximating $P(W \in A)$ by $P(Z \in A)$, the error is

$$E \{ \lambda f_A(W + 1) - W f_A(W) \}$$

which can then be estimated (or bounded) by bounding $f_A(w)$ and $f_A(w + 1) - f_A(w)$ and by exploiting the dependence structure of the rare events.

- Use $\sum_{k=0}^{\infty} |P(W = k) - P(Z = k)| = 2 \sup_{A \in \mathcal{Z}^+} |P(W \in A) - P(Z \in A)|$.



Compound Poisson Approximation

- Often the rare events of interest occur in clumps with the number of clumps approximately Poisson distributed.
- In such instances, the compound Poisson distribution is a better approximation for the number of occurrences of the events.
- A compound Poisson distribution is that of $Y_1 + Y_2 + \dots + Y_N$ where Y_1, Y_2, \dots are independent and identically distributed and N is Poisson distributed independent of the Y_i 's.
- Compound Poisson approximation can be achieved via Poisson approximation using a declumping technique.
(Arratia, Goldstein and Gordon (1990), *Statist. Sci.*)
- It can also be done directly using Stein's method (more general than the method of Poisson approximation).
(Barbour, Chen and Loh (1992), *Ann. Probab.*; Barbour and Xia (1999), *ESAIM*; Barbour and Chryssaphinou (2001), *Ann. Probab.*)

Number of k -Word Matches between Two Sequences

R. A. Lippert, H. Huang and M. S. Waterman (2002), *PNAS*

Compare two sequences by counting the number of k -letter words the two sequences have in common.

(Comparison time is linear with sequence length as no positional information is used in the count.)

Assume that the letters in the two sequences of lengths m and n are i.i.d. random variables taken from a finite alphabet.

Let $\bar{m} = m - k + 1$ and $\bar{n} = n - k + 1$.

Define $Y_{i,j} = I(\text{A perfect match of length } k \text{ or longer from positions } i \text{ and } j \text{ respectively of the two sequences})$

Define $X_{i,j} = I(\text{A perfect match of length } k \text{ or longer beginning from positions } i \text{ and } j \text{ respectively of the two sequences})$

Number of k -Word Matches between Two Sequences

Example: $m = 16, n = 13, k = 4$

AACTGTTACCGATTGA
GCATCCGATAACT

$$Y_{1,10} = 1, X_{1,10} = 1$$

$$Y_{9,5} = 1, X_{9,5} = 1$$

$$Y_{10,6} = 1, X_{10,6} = 0$$

$$\text{Let } D_2 = \sum_{i=1}^{\bar{m}} \sum_{j=1}^{\bar{n}} Y_{i,j}.$$

$$\text{Let } D_{2*} = \sum_{i=1}^{\bar{m}} \sum_{j=1}^{\bar{n}} X_{i,j}.$$

Then D_2 is the number of k -words the two sequences have in common while D_{2*} is the number of maximal perfect matches with length k or longer between the two sequences.

Number of k -Word Matches between Two Sequences

D_{2*} counts the “clumps” while D_2 counts the “clumps” and the events within each “clump” with the number of events within each clump approximately geometric distributed.

Both $\{X_{i,j}\}$ and $\{Y_{i,j}\}$ are locally dependent.

Assume $m = n$. For $k > 2 \log(n)$, D_{2*} is approximately Poisson distributed and D_2 is approximately compound Poisson distributed.

This method of achieving compound approximation for D_2 via Poisson approximation for D_{2*} is called the declumping technique.

Back to Example 1 (DNA sequence matching):

$$P(\text{The longest perfect match} \geq k) = P(D_{2*} \geq 1).$$

Finding Significantly Large Clusters of Palindromes in DNA

Leung, Choi, Xia and Chen (2005), *J. Computat. Bio.*

- Palindromes in DNA

A	G	T	G	A	T	A	T	C	.	.	.	G	C	G
									.	.	.			
T	C	A	C	T	A	T	A	G	.	.	.	C	G	C
1	2	3	4	5	6	7	8	9						

The position 6 is the center of a palindrome of length 6.

Assuming the base pairs are independent and identically distributed random variables,

$$P \left(\begin{array}{l} \text{The position } i \text{ is the center of} \\ \text{a palindrome of length } \geq 2L \end{array} \right) = [2(p_{AP} + p_{CT})]^L$$

If length $2L = 10$, $p_A = p_T = 0.2$ and $p_C = p_G = 0.3$,

then $P = (0.26)^5 \approx 0.001$.



Finding Significantly Large Clusters of Palindromes in DNA

Palindromes are involved in a variety of biological processes.

- Recognition sites for bacterial restriction enzymes to cut foreign DNA
- DNA gene regulation
- DNA replication
- DNA-protein binding

Problems of interest

- Over and under representation of palindromes
- Significantly large clusters of palindromes

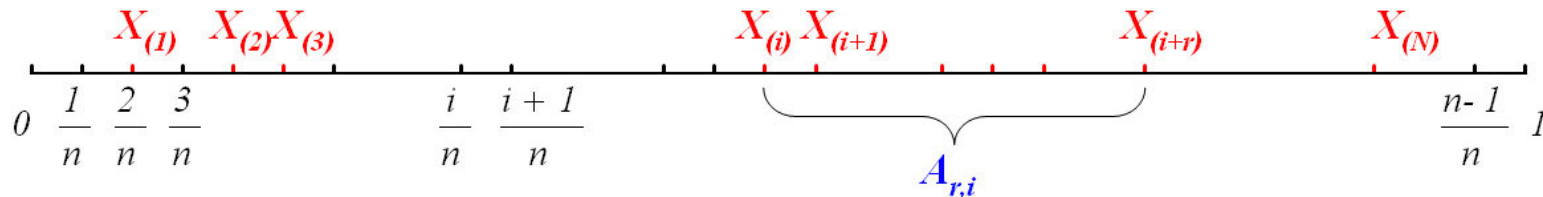
Significantly large clusters of palindromes for 16 herpesviruses were studied.

Finding Significantly Large Clusters of Palindromes in DNA

Locating significantly large clusters of palindromes.

$$\text{Length of DNA} = M \quad (M \text{ base pairs})$$

$$n = M - 2L + 1$$



$$X_{(i)} = \text{center of a palindrome of length } \geq 2L$$

$$A_{r,i} = X_{(i+r)} - X_{(i)}, \quad i = 1, \dots, N - r \quad (r\text{-scans})$$

$$A_r = \min\{A_{r,i} : i = 1, \dots, N - r\} \quad (\text{minimum } r\text{-scan})$$

$$\begin{aligned} P(A_r \leq w) &= P(\text{At least one of } A_{r,i} \leq w) \\ &= P(\text{At least one rare event occurs}). \end{aligned}$$

Finding Significantly Large Clusters of Palindromes in DNA

Assume that the centers of palindromes are independently and uniformly distributed over $[0, 1]$.

Use compound Poisson approximation (Glaz, Naus, Roos and Wallenstein (1994)) to obtain

$$P(A_r \leq w) \approx 1 - \exp\left\{ - (N - r)\pi(1 - p + p^r(r + p - rp)) \right\}$$

where $\pi = Q_1$, $p = 1 - \frac{Q_2}{Q_1}$, etc.

Importance Sampling of Word Patterns in Sequences

Chan, Zhang and Chen (2008), *Preprint (arXiv:0811.4447)*

Suppose the sequence $\mathbf{s} = s_1 s_2 \dots s_n$ of letters from a finite alphabet is distributed according to a probability law P .

Let \mathcal{V} be a finite set of words and Let $N = N_n$ be the number of non-overlapping words from \mathcal{V} in $\mathbf{s} = s_1 s_2 \dots s_n$.

To estimate $p = P(N \geq c)$ for $c \geq 1$ where p is small.

1. Direct Monte Carlo:

Simulate K independent copies $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(K)}$ of \mathbf{s} according to the probability law P .

Let $N(\mathbf{s}^{(k)})$ be the number of non-overlapping words from \mathcal{V} in $\mathbf{s}^{(k)}$.

The estimate of p is

$$\hat{p}_D = K^{-1} \sum_{k=1}^K I(N(\mathbf{s}^{(k)}) \geq c)$$

But such estimation of p is inefficient for small p .



Importance Sampling of Word Patterns in Sequences

2. Importance sampling:

Select a probability law $Q \neq P$ for generating $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(K)}$.

The estimate of p is then

$$\hat{p}_I = K^{-1} \sum_{k=1}^K L^{-1}(\mathbf{s}) I(N(\mathbf{s}^{(k)}) \geq c) \text{ where } L(\mathbf{s}) = Q(\mathbf{s})/P(\mathbf{s}).$$

We require $Q(\mathbf{s}) > 0$ whenever $N(\mathbf{s}) \geq c$, so as to ensure that \hat{p}_I is unbiased for p .

Importance Sampling of Word Patterns in Sequences

3. An algorithm for $c \geq 1$ (defining Q):

Assume that $\mathbf{s} = s_1 s_2 \dots s_n$ is a Markov chain with stationary distribution π (defining P).

First create a word bank with each word in the word bank taking a value \mathbf{v} in \mathcal{V} with probability $q(\mathbf{v}) > 0$.

Let X_i take value 1 or 0 with probability $P(X_i = 1 | s_0 s_1 \dots s_i)$, e.g. $= c/n$. (*)

Step 1. Let $i = 0$, $s_0 \sim \pi$, X_0 satisfying (*).

Step 2. (a) If $X_i = 1$, select word \mathbf{v} from the word bank. If $\ell(\mathbf{v}) \leq n - i$, that is, if word can fit into the remaining sequence, let $s_{i+1} \dots s_{i+\ell(\mathbf{v})} = \mathbf{v}$ and generate $X_{i+\ell(\mathbf{v})}$ according to (*). Increment i by $\ell(\mathbf{v})$ and go to step 3.

(b) If the word selected cannot fit into the remaining sequence or if $X_i = 0$, generate s_{i+1} from the Markov chain and X_{i+1} according to (*). Increment i by 1 and go to step 3.

Step 3. If $i < n$, repeat step 2. Otherwise end.

Importance Sampling of Word Patterns in Sequences

4. Optimality of \hat{p}_I :

- Relative error (RE) of an estimate \hat{p} is given by $\sqrt{\text{Var}(\hat{p})}/p$.
- Under certain conditions, \hat{p}_I is asymptotically optimal as $p \rightarrow 0$, that is, $\text{RE}(\hat{p}_I) \rightarrow 0$ with $\log K = o(|\log p|)$ as $p \rightarrow 0$.
- Since $\text{RE}(\hat{p}_D) = \sqrt{(1-p)/Kp} \not\rightarrow 0$, \hat{p}_D is not asymptotically optimal.
- Simulation run time for the importance sampling algorithm is twice that of the direct Monte Carlo for each run.

5. Applications

- The algorithm is applied to palindromes, inverted repeats, position specific weight matrices and co-occurrences of motifs.
- It can be used for cases where approximations are not accurate (due to short words or short sequences) or are not available.



Conclusion

- Poisson approximation and compound Poisson approximation provide a means for calculating p -values for word patterns in sequence analysis.
- These include sequence comparison and finding significantly large clusters of palindromes.
- Importance sampling provides an alternative method for calculating p -values.
- It is applicable even when approximations are not accurate or not available.
- An asymptotically optimal algorithm is developed for a wide class of problems involving word count.
- These include palindromes, inverted repeats, position specific weight matrices and co-occurrences of motifs.
- The scope of application of the above methods is not confined to the examples given.



Thank you