Designing optimal tests for slow converging Markov chains

Cliff Stein, Pratik Worah

Overview

Neyman-Pearson (NP) hypothesis test consists of comparing the empirical log-likelihood (equivalently the hypothesis test score) with a fixed constant, and accepting or rejecting the null hypothesis based on the outcome.

Note: assumes that large number of samples are present - much larger than the mixing time

Goal: design a modified NP test that requires few samples, and yet an error bound holds

Motivation

Why small number of samples?

- Often we don't have access to a large number of test samples.
- For example, medical tests that compare healthy and diseased tissues that have access to a small number of cells relative to the dimension of the quantity being tested.
- Testing for glucose levels with a small number of cell sample may not have a large error as it is a one dimensional quantity. But, testing for the distribution of RNA expression levels in a heterogeneous tissue may have a large error, since the tissue may have many different types of cells, leading to a relatively high dimensional hypothesis testing problem.

Why Markov Chains?

- Many important natural processes are known to be Markov, i.e., their next state depends only on their current state. For example, *DNA transcription has been modeled as a Markov process*.

Problem description

- ○ **What is known:** two n X n transition matrices P1 and P2
- ○ **What is observed:** empirical distribution μ_m from a Markov chain for m = o(n) steps
- ○ **What is unknown:** the initial state
- **Challenges**:
  - **Q**. How to decide whether μ_m came from P1 or P2 and bound error?
  - **Q**. How to account for the effect of initial state?
  - **Q**. How to carry out the large deviation analysis in the small sample i.e., non-asymptotic, case?
- **Related work**:
  - Sun, Boyd, Xiao, Diaconis, 2006: The Fastest Mixing Markov Process on a Graph and a Connection to a Maximum Variance Unfolding Problem

.

Theoretical Anaylsis

Large deviations theory essentials

Large deviations

- Bounds the probability of rare events - sums of random variables deviating far from the mean.
- **Define**: $Z_m := 1/m\sum_{i=1}^{m} f(X_i)$;
  - $\{X_i\}$ from a Markov process with transition matrix P with state space [n] and
  - f(.) is real valued (can be R^d as well)

Gartner-Ellis Theorem:

- Let A(q) be a small ball around point q, then for large enough m, $\log P(Z_m \in A(q))$ can be upper (and lower) bounded by:
  - $m \sup_x (xq - \log\rho(P_x))$, where
  - P_x is a non-negative matrix defined as $P_x(i,j) := P(i,j)e^{(xf(j))}$
  - $\rho(P_x)$ is the principal eigenvalue of P_x

Log-likelihood and its modification

- Usual hypothesis testing consists of two steps: (i) compute a log-likelihood score, given μ_m, and (ii) compare it against a fixed threshold τ, if the score is less than τ null hypothesis is accepted.
- The log-likelihood score is computed as:

$$\hat{S}_m := \sum_{i\in[m]} \langle \mu_m, 1_i \rangle \cdot \log \frac{\langle \mu_m P_2, 1_i \rangle}{\langle \mu_m P_1, 1_i \rangle}.$$

- We modify the log-likelihood score computation using a d X n dimensional projection matrix Z as follows:

$$\hat{S}_m := \sum_{i\in[d]} \langle Z^T \mu_m, 1_i \rangle \cdot \log \frac{\langle Z^T \mu_m P_2, 1_i \rangle}{\langle Z^T \mu_m P_1, 1_i \rangle}.$$

- The score is then compared to a threshold $\tau \mp \beta$, where β is a positive quantity that depends on the height of the principal eigenvector.
- If score less than $\tau - \beta$, null hypothesis is accepted, and if score is greater than $\tau + \beta$ alternative hypothesis is accepted.

Experiments

Using single cell RNA-seq data from Bastidas-Ponce et al, 2019.

- Sample Cpe expressions from ~20 beta cells.
- Compute transition matrices P1 and P2 for Cpe expression, modeled as a Markov process, from beta cells and alpha cells respectively (see below).
- Compute ordinary and modified log likelihood and try to decide whether the sample came from beta or alpha cells?

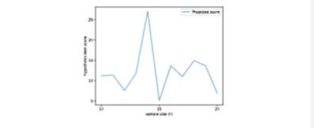
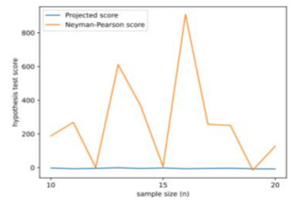


Figure 2. High positive values of the score suggest that the empirical distribution is from beta cells (the ground truth). Here we can distinguish between the two cell types (beta cells vs ductal cells) using just 10-20 cell samples of Cpe expressions.



- Computed log likelihoods in Figure 1 (above) and 2 (left).
- Red curve is original log likelihood.
- Blue curve (magnified on the left) is the modified log likelihood.
- Higher values indicate sample came from beta cells.
- ***Red curve fluctuates but Blue curve is stabler, suggests modified log likelihood is more reliable***

Transition matrices

- The transition matrices P1 and P2 for Cpe expressions, measure the probability that a cell transitions from Cpe expression level c1 to c2 in a small time interval.
- They can be computed by measuring the relative transition frequency of cells sorted according to their latent time (see for example, Bergen et al. 2020).
- Note that transition matrices need to computed only once, based on which one can do hypothesis testing for diseased vs normal tissue.

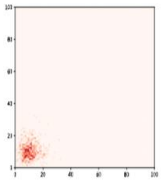


Figure 3. Heat map for the transition matrix of Cpe expression in beta cells. Darker areas indicate higher values.

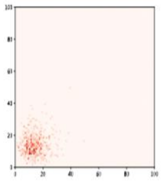


Figure 4. Heat map for the transition matrix of Cpe expression in alpha cells. Darker areas indicate higher values.