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

Theoretical Analysis

Large deviations theory essentials

Large deviations

- Bounds the probability of rare events - sums of random variables deviating far from the mean.
- Define $Z_n = \sum_{i=1}^{n} f(X_i)$
- $(X_i)$ from a Markov process with transition matrix $P$ with state space $\mathcal{S}$ and $f(.)$ is real valued (can be R 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:
  - $m \sup_{y \in \mathcal{S}} (\gamma - \log\lambda(\gamma))$, where
  - $\lambda$ is a non-negative measure defined as $\lambda(\gamma) = P(\gamma)$
  - $\gamma(P_j)$ is the principal eigenvalue of $P_j$

Log-likelihood and its modification

- Usual hypothesis testing consists of two steps: (i) compute a log-likelihood score, given $\theta$, and (ii) compare it against a fixed threshold $\tau$, if the score is less than $\tau$ null hypothesis is accepted.

The log-likelihood score is computed as:

$$\hat{S}_m := \sum_{i \in [m]} \left( \frac{\mu_m}{\mu_{P_2}} \cdot 1_{i_1} \cdot \log \frac{\mu_{P_2} 1_{i_1}}{\mu_{P_1} 1_{i_1}} \right).$$

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

$$\hat{S}_m := \sum_{i \in [d]} \left( Z^T \mu_{P_1} 1_{i_1} \cdot \log \frac{Z^T \mu_{P_2} 1_{i_1}}{Z^T \mu_{P_1} 1_{i_1}} \right).$$

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

Experiments

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

- Sample Cpe RNA-seq data 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?

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 be computed only once, based on which one can do hypothesis testing for diseased vs normal tissue.

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 $n \times n$ transition matrices $P_1$ and $P_2$
- What is observed: empirical distribution $\mu_{\theta}$ from a Markov chain for $\theta = \theta(\theta)$ steps
- What is unknown: the initial state

Challenges:

- Q. How to decide whether $\mu_{\theta}$ came from $P_1$ or $P_2$ 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, Diconis, 2006: The Fastest Mixing Markov Process on a Graph and a Connection to a Maximum Variance Unfolding Problem.