

The Hunting of the Bump

On Maximizing Statistical Discrepancy

Suresh Venkatasubramanian

AT&T, Nov 11, 2005

Outline

1 Introduction

2 Method

Bump Hunting

Problem: Find the region that maximizes the discrepancy function

What Kinds Of Regions ?

What Kinds Of Functions?

Let $\mathcal{M} = \{m_1, \dots\}$ be the measurements and $\mathcal{B} = \{b_1, \dots\}$ be the baseline values. For a region R ,

$$m_R = \frac{\sum_{m_i \in R} m_i}{\sum m_i}, b_R = \frac{\sum_{b_i \in R} b_i}{\sum b_i}$$

Discrepancy function f defined in terms of m_R, b_R . For example

Linear Discrepancy $f := \alpha m_R + \beta b_R + \gamma$

Combinatorial Discrepancy Difference between measurement and baseline,

$$f := \left| \sum_{m_i \in R} m_i - \sum_{b_i \in R} b_i \right| = |m_R * M - b_R * B|$$

Maximum Likelihood-based Discrepancy

Assume a parametrized baseline distribution. Consider two hypotheses:

\mathcal{H}_0 Data inside region R reflects same distribution parameters as data outside R .

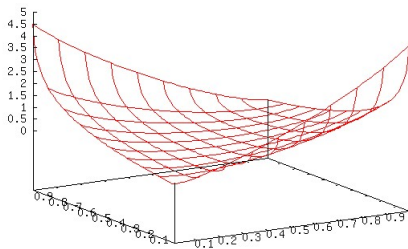
\mathcal{H}_1 Data inside region R fits *different* parameters than data outside R .

Likelihood ratio test: Compute (log of) ratio of probabilities of \mathcal{H}_1 and \mathcal{H}_0 .
The larger this number, the more likely it is that \mathcal{H}_1 reflects reality.

Example: The Kulldorff Scan Statistic

A well-known discrepancy function in biosurveillance. Assumes an underlying Poisson distribution.

$$\begin{aligned} d_K(m_R, b_R) &= \text{KL}(\{m_R, 1 - m_R\}, \{b_R, 1 - b_R\}) \\ &= m_R \log \frac{m_R}{b_R} + (1 - m_R) \log \frac{1 - m_R}{1 - b_R} \end{aligned}$$



How hard is it to maximize discrepancy?

Main Results

Outline

1 Introduction

2 Method

Main Idea

Approximately solve the (f, \mathcal{R}) problem, all we need is a solution to the $(ax + b, \mathcal{R})$ problem.

Step 1: Linearize f approximately, replacing it by a collection of $k(\epsilon)$ linear functions.

Step 2: Solve the problem on each of the k linear functions.

Step 3: Take the best solution.

Running time: $k(\epsilon) * c(\ell, \mathcal{R})$.