Filtering: The Case for "Noisier" Data

B. Lucena* I. Kontoyiannis†

August 24, 2005

Abstract

Suppose there is some discrete variable X of interest, which can only be observed after passing through 2 channels (Q and R). You are limited to n noisy observations of X and then must estimate the value of X. Your one control is a parameter k which determines the level of correlation in your observed data. Specifically, X is first transmitted through the channel Q n/k times to yield variables $Y_1, \ldots, Y_{n/k}$, and then each Y_i is transmitted through channel R k times to yield variables $Z_{i,1}, \ldots, Z_{i,k}$. How should k be chosen to maximize the probability of successfully guessing the correct value of X? While $k = 1$ yields data points $Z_{i,1}$ which are conditionally independent given the value of X, we find that this does not always mean that $k = 1$ is the optimal choice. In fact, many simple situations yield cases where the optimal value of k is greater than 1. We explore this phenomenon and present both theoretical and empirical results.

Keywords: Discrete memoryless channels, binary symmetric channel, Potts channel, Z-channel, mutual information, maximum likelihood, filtering.

1 Introduction

The ideas in this paper can be largely motivated by the following two biological questions.

1. An Epidemiology Problem. Suppose we want to test for the presence of a disease in a certain population, and we do this by selecting members of the population at random and then testing them. Assuming that the test is imperfect (there is a certain chance we might get a false positive or a false negative), is it better to test as many people as possible once each, or to test fewer people with more than one test?

2. Ancestral Reconstruction. Given an evolutionary tree, suppose we want to reconstruct a character at the root of the tree from the information at the leaves. However, we

*Division of Computer Science, University of California, Berkeley, Soda Hall, Berkeley, CA 94720, USA. Email: lucena@cs.berkeley.edu. Supported in part by an NSF Mathematical Sciences Postdoctoral Research Fellowship.

†Division of Applied Mathematics and Dept of Computer Science, Brown Univ, 182 George St., Providence, RI 02912, USA. Email: yiannis@dam.brown.edu Web: www.dam.brown.edu/people/yiannis/. Supported in part by NSF grant #0073378-CCR and by a research fellowship from the Sloan Foundation.

can only observe a fixed proportion of the leaves. Which leaves should we choose? Intuition suggests that, if possible, we should choose our leaves to induce a "star-topology", that is, choose leaves which are conditionally independent of each other given the value of the root. However, as we will see this intuition is often wrong.

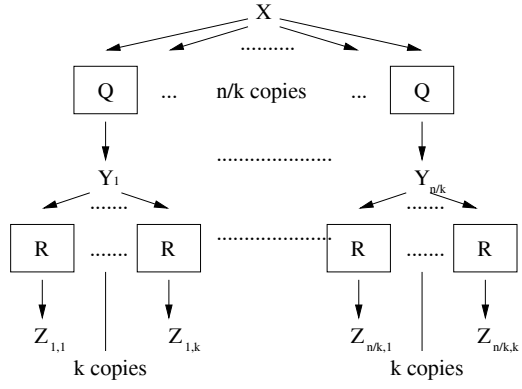

Figure 1: The channel network.

We abstract these two problems into the setting depicted in Figure 1; we assume that n/k is an integer. A variable X is chosen according to some distribution, then transmitted through the channel Q n/k times independently to yield the variables $Y_1, Y_2, \ldots, Y_{n/k}$. Then, each Y_i variable is transmitted through channel R k times independently to yield variables $Z_{i,1}, Z_{i,2}, \ldots, Z_{i,k}$. When $k = 1$ the leaves are conditionally independent given the root. For larger values of k there is a greater level of correlation in the leaves, since leaves in the same subtree are not conditionally independent given the root. If we assume that the total number of observations we are allowed to make, or equivalently the total number n of leaves in the tree, the main question we consider is *what is the optimal value of* k? That is, given that the variable of interest X has some fixed distribution, and given the two channels Q and R, how much correlation do we want to have in our observations? Naturally, the answer depends on the definition of what we mean by "optimal". Much recent work on this topic [2] has focused on the *mutual information* criterion, that is, on maximizing $I(X; \bar{Z}^k)$ where $\bar{Z}^k = (Z_{1,1}, \ldots, Z_{1,k}, Z_{2,1} \ldots, Z_{n/k,1}, \ldots, Z_{n/k,k})$. In this paper, we use an alternate criterion; we want to maximize the probability of guessing the correct value of X after observing $\bar{Z}^k$. In this sense, this is a *filtering* problem. For trees of infinite depth, these criteria are roughly equivalent (see, e.g., [2]), whereas, for finite trees the choice of criterion can make a significant difference. This is an interesting issue which we will not consider futher in this paper; we note, however, that for a variety of different criteria, somewhat surprisingly we find that some $k > 1$ is often preferred to $k = 1$.

We model Problem 1 by letting X, Y_i, $Z_{i,j}$ be binary random variables, where the marginal distribution of X is Bernoulli(p). Here, X is 1 if the disease exists in the population and 0 if not. We let Q be a *Z-channel* with parameter δ. That is, 0 is always transmitted accurately while 1 maps to 0 with probability δ and to 1 with probability $(1 - \delta)$. Each Y_i variable

represents the status of a different person chosen from the population, and the choice of the Z-channel for Q reflects the fact that if the disease does not exist in the population (i.e. $X = 0$) then no person can have the disease (i.e. $Y_i = 0$ for all i). On the other hand, if the disease does exist, only some fraction $(1 - \delta)$ of the population actually has it. We then choose a binary symmetrical channel with error probability ϵ for channel R. This models a test with equal probabilities of false positives and false negatives. We will refer to this setting as the *Z-BSC setting*. It has three model parameters: δ, ϵ, and p.

We model Problem 2 by letting X, Y_i, $Z_{i,j}$ be variables on the state space $\{1, 2, \ldots, m\}$, where the marginal distribution of X is discrete uniform. We then choose channels Q and R to be *Potts channels* with the same parameter α. This means that each value is accurately transmitted with probability α and switched to a particular different value with probability $(1 - \alpha)/(m - 1)$. This is a reasonable model of evolution and is equivalent to the well-known *Jukes-Cantor* model when $m = 4$. In general, we refer to this as the *Potts setting*. It has two model parameters: α and m. Clearly, more complicated models could be used for both of these problems but even these simple cases display interesting phenomena. Our results fall into two categories. In the first, we look at problems where n, the number of leaves, is small (i.e. $n = 2$ or 3). These cases can be thoroughly understood theoretically and exhibit surprising phenomena which motivate the study of the more difficult case, which is to analyze what happens when n is large. Here we present a combination of theoretical and simulation results, which demonstrate that the correct amount of correlation can improve accuracy significantly.

2 The Z-BSC setting for $n = 2$

In this section we examine the Z-BSC setting for the simplest non-trivial case $n = 2$. The question is whether $k = 2$ is better than $k = 1$ for different choices of the parameters δ, ϵ, and p. In the language of Problem 1, we are asking whether testing two people once each or testing one person twice gives us a higher probability of making the correct guess about the presence of the disease in the population.

For a fixed $\delta \in (0, 1)$ and $\epsilon \in (0, 1/2)$, let $P_k(p)$ = the probability of making the correct guess (via maximum likelihood) about the root. We prove the following result:

Theorem Fix $\delta \in (0, 1)$ and $\epsilon \in (0, 1/2)$. The following statements are true:

- For $p < \frac{\epsilon^2}{(1+\delta)\epsilon^2+(1-\delta)(1-\epsilon)^2}$ or $p > \frac{(1-\epsilon)^2}{(\delta(1-\epsilon)+(1-\delta)\epsilon)^2}$ we have that $P_2(p) = P_1(p)$.
- For $\frac{\epsilon^2}{(1+\delta)\epsilon^2+(1-\delta)(1-\epsilon)^2} < p < \frac{2\epsilon(1-\epsilon)}{3\epsilon(1-\epsilon)+(\delta\epsilon+(1-\delta)(1-\epsilon))(\delta(1-\epsilon)+(1-\delta)\epsilon)}$ we have that $P_2(p) > P_1(p)$.
- For $\frac{2\epsilon(1-\epsilon)}{3\epsilon(1-\epsilon)+(\delta\epsilon+(1-\delta)(1-\epsilon))(\delta(1-\epsilon)+(1-\delta)\epsilon)} < p < \frac{(1-\epsilon)^2}{(\delta(1-\epsilon)+(1-\delta)\epsilon)^2}$ we have that $P_1(p) > P_2(p)$.

This demonstrates that we sometimes "learn more" about the population from testing one person twice than from testing two people once each. This suggests the possibility that,

even for large values of n, it is possible that some $k > 1$ is better than $k = 1$. That is, there is some optimal amount of correlation than can help us in this decision procedure.

3 The Potts setting for $n = 2$ and $n = 3$

We now take the same approach to the Potts setting. Recall that in this setting we assume the root has a discrete uniform distribution. For a given n and m, let $P_k(\alpha) =$ the probability of accurately reconstructing the root using maximum likelihood using $\bar{Z}^k$. We consider only $\alpha \in (1/m, 1)$. The first result applies to any value of n when $m = 2$.

Theorem Let $m = 2$, then for all $n \geq 1$ and for all $\alpha \in (1/2, 1)$ we have that $P_1(\alpha) \geq P_k(\alpha)$ for $k > 1$.

The above theorem is derived as a consequence of the work in [1]. It says that if all our observations are binary ($m = 2$), then $k = 1$ is *always* the optimal choice. We also show that if we are only allowed $n = 2$ observations, then there is no difference between $k = 1$ and $k = 2$.

Theorem Let $n = 2$. Then for all $\alpha \in (1/m, 1)$ we have $P_2(\alpha) = P_1(\alpha)$.

This is a consequence of the fact that the second leaf does not affect your decision at all. If the second leaf has the same value as the first, then you choose the same as if you had only seen the first leaf. If the second leaf has a different value, since the marginal distribution of the root is discrete uniform, it is still optimal to guess the value of the first leaf.

Therefore, next we turn to the more interesting case when both $n, m \geq 3$. This next result describes what happens when $n = 3$ for any $m \geq 3$ by comparing $P_1(\alpha)$ to $P_3(\alpha)$.

Theorem Let $m \geq 3$. Then there exist values $\alpha_*, \alpha^* \in (1/m, 1)$ such that $P_3(\alpha) > P_1(\alpha)$ for $\alpha < \alpha_*$, while $P_3(\alpha) < P_1(\alpha)$ for $\alpha > \alpha^*$.

This suggests that when the channels are very "noisy," that is, where α is close to $1/m$, then the redundancy of the leaves present for $k > 1$ can be useful as a means of "strengthening" the signal.

4 Results for large values of n

We first focus on the Z-BSC setting. For large values of n, the expressions for P_k become more difficult to analyze, particularly for values of k other than 1. That is, it becomes difficult to evaluate the probability of successfully guessing the right answer, since the optimal guessing procedure itself involves a fairly complex computation, relying on a belief propagation/ dynamic programming algorithm. The one case that is still easy is when $k = 1$, since then, in the Z-BSC setting, the sum of the $Z_{i,j}$ variables is a sufficient statistic. There is also a simple formula for determining the *threshold* T such that we should guess $X = 1$ if the sum exceeds T and guess $X = 0$ otherwise. So it is easy to simulate with a computer, and can be analyzed theoretically for large n using limit theorems from probability. For $k > 1$, it is considerably more difficult even to merely simulate the optimal procedure.

We work around these problems by considering a different decision procedure. Though this procedure is not optimal, it is close to optimal. It is also equivalent to the optimal procedure for $k = 1$. Therefore, it provides a means of showing that some $k > 1$ performs better than $k = 1$ if a suboptimal procedure for the larger value of k is better than the optimal procedure for $k = 1$. Most conveniently, the probability of success of the suboptimal procedure for a given k equals exactly the probability of success of the optimal procedure for different values of n and ϵ. This makes a direct comparison possible and enables us to present some theoretical results.

The first theorem below shows that there exist situations where some correlation is desirable, even when the total number n of observations is large. Therefore, the phenomenon observed in the case where $n = 2$ does not disappear as n gets big but persists.

The second theorem shows that, as the number of observations grows, the optimal value of k remains bounded.

Theorem There exists $\delta \in (0, 1), \epsilon \in (0, 1/2)$, and $p \in (0, 1)$ and N_0 such that for $n > N_0$, $P_3(n) > P_1(n)$ where $P_k(n) =$ the probability of correctly guessing the root using $\bar{Z}^k$ as a function of n for δ, ϵ, and p fixed.

Theorem For a fixed $\delta \in (0, 1), \epsilon \in (0, 1/2)$, and $p \in (0, 1)$, let $k^*(n)$ be the value of k which maximizes $P_k(n)$. There exists some number K such that $k^*(n) < K$ for all n.

We also have extensive simulation results which highlight these theorems and provide a sense of what values of k are best for different values of the parameters. We also have corresponding preliminary results on the Potts setting, where we also show simulations which demonstrate similar phenomena as in the Z-BSC setting. The discrepancy is somewhat less striking, however.

References

[1] W. Evans, C. Kenyon, Y. Peres, and L. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Prob.*, 10(2):410–433, 2000.

[2] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, 11(1):285–300, 2001.