Ch. 14: Statistical Hypothesis Testing for Dependence in Binary Sequences

I. Introduction

Assume that a node in a communications network makes measurements of successes and failures of packets over N packet intervals. Let these measurements be defined as the binary sequence $\{y_i\}_{i=1}^N$ where $y_i = 1$ and $y_i = 0$ denote success and failure, respectively, of the i-th packet. A typical sequence could be

$$111110001100100010111111\,1$$

One objective could be to decide whether the failure (dropped packet) states follow an independent model or not. Another objective could be to decide whether the failure states follow a 1st order Markov chain model or not. Failure to follow these hypotheses could indicate an out-of-link failure mode or an in-link failure mode, respectively. More generally, lack of first order Markov behavior could indicate long range dependency in the network. Extensions of this problem to multi-channel and multi-dimensional measurements have application to network tomography, non-parametric factor analysis, Gibbs random field models, and order/disorder prediction in financial time series. Hypothesis testing and other types of inference for Markov dependency models has an extensive literature in engineering and statistics, both for hidden Markov models [1], [2], [3], notably applied to speech [4] and seismology [5], and for directly observed Markov models [6], [7].

In this chapter we present solutions to the dependency testing problem using the hypothesis testing framework discussed in Chapter 6, Section 9. The main differences between the work [6] and the work discussed here are: (1) we assume a kth order Markov chain alternative model; and (2) a large sample asymptotic regime is adopted at the outset. This simplifies the likelihood ratio to one that depends only on k-th order transition probabilities; the k-th order joint initial probabilities can be neglected. We will see that application of the GLRT methodology can be directly applied to testing for dependency with concurrent advantages of asymptotic distribution theory of the GLRT discussed in Chapter 6. The resulting tests have very intuitive structure related to information theoretic measures of dependency.

II. The Markov Dependency Model

To capture dependency we use the k-th order Markov chain model

$$P(y_i|y_{i-1},\ldots,y_{i-k}) = p^{y_i}(1-p)^{1-y_i} \tag{1}$$

where $p = p(y_{i-1},\ldots,y_{i-k})$ is the success probability

$$\begin{aligned} p &= P(y_i=1|y_{i-1},\ldots,y_{i-k}) = \begin{cases} r_{0\ldots0}, & y_{i-1}=0,\ldots,y_{i-k}=0 \\ r_{10\ldots0}, & y_{i-1}=1, y_{i-2}=0,\ldots,y_{i-k}=0 \\ \vdots & \vdots \\ r_{1\ldots1}, & y_{i-1}=1,\ldots,y_{i-k}=1 \end{cases} \\ &= r_{0\ldots0}^{(1-y_{i-1})\cdots(1-y_{i-k})}\, r_{10\ldots0}^{y_{i-1}(1-y_{i-2})\cdots(1-y_{i-k})} \cdots r_{1\ldots1}^{y_{i-1}\cdots y_{i-k}}. \end{aligned}$$

Note that for sufficiently large k any random sequence can be described by a k-th order Markov model. When the transition probabilities $P(y_i|y_{i-1},\ldots,y_{i-k})$ are independent of the time index i the quantities $r_{0\ldots0},\ldots,r_{1\ldots1}$ are constant over time and the sequence has a steady state stationary distribution. By making $r_{0\ldots0},\ldots,r_{1\ldots1}$ change in some simple way over time, e.g. a single step change occurs at an unknown time instant, the model can easily be generalized to accomodate changepoint and slippage problems. We do not investigate these extensions here.

Testing for independence (nominal 0th order Markov) against dependency reduces to testing the two hypotheses

$$H_0: p(y_{i-1},\ldots,y_{i-k}) = p_o, \quad H_1: p(y_{i-1},\ldots,y_{i-k}) \neq p_o, \text{ for some } \{b_i\} \neq \{c_i\}. \tag{2}$$

where p_o is a constant which may be either known or unknown. Note that under H_0 $\{y_i\}_{i=1}^N$ is Bernoulli with parameter p_o. An equivalent way of expressing these hypotheses (for unknown p_o) is

$$H_0 : r_{0\ldots 0} = \cdots = r_{1\ldots 1}, \quad H_1 : r_{c_1\ldots c_k} \neq r_{b_1\ldots b_k}, \text{ for some } \{b_i\} \neq \{c_i\}. \tag{3}$$

In H_1, $\{b_i\}$ and $\{c_i\}$ are length k binary sequences.

Likewise the hypotheses for testing for 1st order Markov dependency against kth order dependency ($k > 1$) can be stated as

$$H_0 : p(y_{i-1}, \ldots, y_{i-k}) = p(y_{i-1}), \quad H_1 : p(y_{i-1}, \ldots, y_{i-k}) \neq p(y_{i-1}),. \tag{4}$$

In the sequel we derive likelihood ratio tests for independence (0th order Markov) and tests for 1st Markov dependence of $\{y_i\}_{i=1}^N$.

III. Tests for Independence

Let $P_{H_1,\underline{r}}(\{y_i\})$ and $P_{H_0,p_o}(\{y_i\})$ denote the probability mass function of $\{y_i\}_{i=1}^N$ under H_1 (parameterized by the 2^k-element vector $\underline{r} = [r_{0\ldots 0}, \ldots, r_{1\ldots 1}]^T$), and under H_0 (parameterized by p_o). When the values of $\underline{r}, p_o$ are known the hypotheses in (3) and (9) are simple and the optimal test of level α is given by the Neyman-Pearson likelihood ratio test (LRT) [8]

$$\Lambda(\underline{r}, p_o) = \frac{P_{H_1,\underline{r}}(\{y_i\})}{P_{H_0,p_o}(\{y_i\})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5}$$

where threshold η is selected such that the specified level $\alpha \in [0, 1]$ of false alarm is attained $P(\Lambda > \eta | H_0) = \alpha$. This test is optimal in the sense of attaining the highest power $P(\Lambda > \eta | H_1)$ among all tests of level α.

Under the Markov chain model the LRT reduces to the equivalent form

$$\Lambda(\underline{r}, p_o) = \frac{P_{H_1;\underline{r}}(y_k, \ldots, y_1)}{P_{H_0;p_o}(y_k, \ldots, y_1)} \prod_{j=k+1}^{N} \frac{P_{H_1;\underline{r}}(y_j | y_{j-1}, \ldots, y_{j-k})}{P_{H_0;p_o}(y_j | y_{j-1}, \ldots, y_{j-k})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{6}$$

For $N \gg k$ the ratio involving the joint probability of the initial states $y_k, \ldots, y_1$ can be neglected and the LRT takes the approximate form:

$$\ln \Lambda(\underline{r}, p_o) = \sum_{j=k+1}^{N} \ln \frac{P_{H_1;\underline{r}}(y_j | y_{j-1}, \ldots, y_{j-k})}{P_{H_0;p_o}(y_j | y_{j-1}, \ldots, y_{j-k})} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \tag{7}$$

where γ is set to satisfy the constraint that the false alarm probability equals α.

We will focus in the sequel on the cases of first and second Markov chain models ($k = 1, 2$) and large N.

A. First Order Model

For $k = 1$ the model for conditional probability of success p in (1) reduces to

$$\begin{aligned} p &= p(y_{i-1}) = \begin{cases} r_1, & y_{i-1} = 1 \\ r_0, & y_{i-1} = 0 \end{cases} \\ &= r_1^{y_{i-1}} r_0^{1-y_{i-1}} \end{aligned}$$

Hence, testing for independence reduces to testing the two hypotheses

$$H_0 : r_1 = r_0 = p_o, \quad H_1 : r_1 \neq r_0. \tag{8}$$

If it is known a priori that dependent failures tend to occur in bursts then a more appropriate set of hypotheses is

$$H_0 : r_1 = r_0 = p_o, \quad H_1 : r_1 > r_0, \tag{9}$$

since this H_1 corresponds to making y_i a failure more likely given y_{i-1} is a failure.

For the case of unknown r_1, r_0 or p_o the hypotheses in (8) (9) are composite and most powerful tests of level α generally do not exist. An alternative is provided by the likelihood principle [8], [9], known as the generalized likelihood ratio test (GLRT) in the engineering community [10].

Several cases will be of interest: (1) known p_o under H_0 and unknown r_1, r_0 under H_1 (nominal i.i.d. situation has known success probability); (2) known p_o, r_0 and unknown r_1 (probability of recovery from failure state is also known); and (3) unknown r_0, r_1, p_o. For case (2) the GLRT for H_0 against the two sided hypothesis H_1 in (8) is

$$\hat{\Lambda}(r_0, p_o) = \frac{\max_{r_1 : r_1 \neq r_0} P(\{y_i\}|H_1, r_1, r_0)}{P(\{y_i\}|H_0, p_o)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{10}$$

and similarly for the one sided H_1 in (9) except that the maximization is performed over the set $\{r_1 : r_1 > r_0\}$. The intuition behind the above GLRT is that in the absence of precise knowledge of r_1, the most likely value (the maximum likelihood estimate under H_1) is substituted into the LRT (7) to provide the strongest evidence for validity of H_1. For case (3) the GLRT is analogously

$$\hat{\Lambda} = \frac{\max_{r_1, r_0 : r_1 \neq r_0} P(\{y_i\}|H_1, r_1, r_0)}{\max_{p_o} P(\{y_i\}|H_0, p_o)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{11}$$

For both of these GLRT's invocation of the asymptotic Chi-square distribution of $2 \ln \hat{\Lambda}$ [9] can be used to set the thresholds. More details below.

It will turn out that all of the tests for the case $k = 1$ will reduce the functions of the measurements:

$$\begin{aligned}
N_0 &= \#\text{ ones in } \{y_i\}_{i=1}^N \\
N_1 &= \#\text{ zeros in } \{y_i\}_{i=1}^N \\
N_{01} &= \#\text{ transitions from } y_{i-1} = 1 \text{ to } y_i = 0 \text{ in } \{y_i\}_{i=1}^N \\
N_{10} &= \#\text{ transitions from } y_{i-1} = 0 \text{ to } y_i = 1 \text{ in } \{y_i\}_{i=1}^N \\
N_{11} &= \#\text{ transitions from } y_{i-1} = 1 \text{ to } y_i = 1 \text{ in } \{y_i\}_{i=1}^N \\
N_{00} &= \#\text{ transitions from } y_{i-1} = 0 \text{ to } y_i = 0 \text{ in } \{y_i\}_{i=1}^N
\end{aligned}$$

B. Case of Known r_1, r_0, p_o

Define

$$Z \stackrel{\text{def}}{=} \ln \Lambda(r, q).$$

Substitution of the model (1) into (7) and specializing to $k = 1$ yields the LRT

$$\begin{aligned}
Z &= \ln\left(\frac{r_1}{p_o}\right) \sum_{i=1}^N y_{i-1} y_i + \ln\left(\frac{1 - r_1}{1 - p_o}\right) \sum_{i=1}^N y_{i-1}(1 - y_i) \\
&\quad + \ln\left(\frac{1 - r_1}{1 - p_o}\right) \sum_{i=1}^N (1 - y_{i-1})(1 - y_i) + \ln\left(\frac{r_1}{p_o}\right) \sum_{i=1}^N (1 - y_{i-1}) y_i \\
&= aN_{11} + bN_{01} + cN_{00} + dN_{10} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma
\end{aligned} \tag{12}$$

To set the threshold γ we make use of the large N CLT to obtain

$$\gamma = E[Z|H_0] + \sqrt{\text{var}(Z|H_0)}\,\Phi^{-1}(1 - \alpha)$$

where $\Phi(u)$ is the standard normal cdf and Φ^{-1} denotes its functional inverse.

In the special case that $r_1 = p_o$, $c = d = 0$ and we have the simplification

$$Z = aN_{11} + bN_{01} \overset{H_1}{\underset{H_0}{\gtrless}} \gamma \tag{13}$$

A straightforward but tedious calculation establishes that for the above test

$$\begin{aligned} E[Z|H_0] &= Nap_o^2 + Nbp_o(1-p_o) \\ \mathrm{var}(Z|H_0) &= a^2[Np_o^2(1-p_o^2) + 2(N-1)p_o^3(1-p_o)] + b^2[Np_o(1-p_o)(1-p_o(1-p_o)) - 2(N-1)p_o^2(1-p_o^2)] \\ &\quad +2ab[Np_o(1-p_o)(1-p_o^2) + (N-1)(1-p_o)p_o^2(1-2p_o)]. \end{aligned}$$

Although the calculations are somewhat messier, the power function of this test can also be calculated under the large N assumption.

C. Case of Unknown r_1, r_0, Known p_o

Since the unknowns r_1, r_0 only enters through the numerator of $\Lambda(\underline{r}, p_o)$ the maximization in (10) is equivalent to the maximization of Z defined in (12) over r_1, r_0.

C.1 Unknown r_1, r_0, Known p_o: Two Sided Composite Hypotheses

Here, except for the measure-zero set $\{r_1, r_0 : r_1 = r_0\}$, the maximization over r_1, r_0 of Z is over a domain unconstrained by p_o. As Z is convex in r_1, r_0 the optimal test is simple to derive

$$\begin{aligned} \max_{r_1, r_0} Z &= \ln\left(\frac{\hat{r}_1}{p_o}\right) \sum_{i=1}^N y_{i-1}y_i + \ln\left(\frac{1-\hat{r}_1}{1-p_o}\right) \sum_{i=1}^N y_{i-1}(1-y_i) \\ &\quad + \ln\left(\frac{1-\hat{r}_0}{1-p_o}\right) \sum_{i=1}^N (1-y_{i-1})(1-y_i) + \ln\left(\frac{\hat{r}_0}{p_o}\right) \sum_{i=1}^N (1-y_{i-1})y_i \\ &= \hat{a}N_{11} + \hat{b}N_{01} + \hat{c}N_{00} + \hat{d}N_{10} \overset{H_1}{\underset{H_0}{\gtrless}} \gamma \end{aligned} \tag{14}$$

where $\hat{r}_1, \hat{r}_0$ are the MLEs (under H_1) of r_1, r_0

$$\hat{r}_1 = \frac{N_{11}}{N_{11} + N_{01}}, \quad \hat{r}_0 = \frac{N_{10}}{N_{10} + N_{00}}$$

For large N, under H_0 the statistic $2\max_{r_1, r_0} Z$ is distributed as Chi-square with 1 degree of freedom (df). This fact gives the asymptotic threshold value for γ in (14)

$$\gamma = \tfrac{1}{2}\mathcal{X}_1^{-1}(1-\alpha)$$

where $\mathcal{X}_k(u)$ is the cdf of the Chi square distribution with k df.

For the case that $r_0 = p_o$ is known the GLRT test statistic becomes

$$\max_{r_1} Z = \ln\left(\frac{\hat{r}_1}{p_o}\right) \sum_{i=1}^N y_i y_{i-1} + \ln\left(\frac{1-\hat{r}_1}{1-p_o}\right) \sum_{i=1}^N (1-y_i)y_{i-1} \quad = \quad \hat{a}N_{11} + \hat{b}N_{01} \overset{H_1}{\underset{H_0}{\gtrless}} \gamma$$

where $\hat{r}_1$ is the MLE of r

$$\hat{r}_1 = \frac{N_{11}}{N_{11} + N_{01}}.$$

C.2 Unknown r_1, r_0, Known p_o: One Sided Composite Hypotheses

Here, the maximization must be restricted to the set $r_1 > r_0$. This yields the test

$$\max\{0, \hat{a}N_{11} + \hat{b}N_{01}\} \overset{H_1}{\underset{H_0}{\gtrless}} \gamma$$

D. Case of Unknown r_1, r_0, p_o

For this case the maximization must be done separately over numerator and denominator of (10). When $\{y_i\}$ is not identically the zero sequence, it is easily shown that the numerator is maximized by the pair

$$\hat{r}_1 = \frac{N_{11}}{N_{11} + N_{01}}, \quad \hat{r}_0 = \frac{N_{10}}{N_{10} + N_{00}}$$

and the denominator is maximized by

$$\hat{r}_0 = \frac{N_1}{N}.$$

We have the large N identities $N_1 = N_{11} + N_{01}$, $N_0 = N_{10} + N_{00}$.

D.1 Unknown r_1, r_0, p_o: Two Sided Composite Hypotheses

It can be easily shown that the large N GLRT statistic becomes

$$\hat{\Lambda} = \frac{\left(\frac{N_{11}}{N_1}\right)^{N_{11}} \left(\frac{N_{01}}{N_1}\right)^{N_{01}} \left(\frac{N_{10}}{N_0}\right)^{N_{10}} \left(\frac{N_{00}}{N_0}\right)^{N_{00}}}{\left(\frac{N_1}{N}\right)^{N_1} \left(\frac{N_0}{N}\right)^{N_0}}$$

Taking the log of this expression and dividing by N yields the interesting result

$$\begin{aligned} \ln \hat{\Lambda} &= -\hat{p}_1 \ln \hat{p}_1 - \hat{p}_0 \ln \hat{p}_0 + \hat{p}_1(\hat{p}_{1|1} \ln \hat{p}_{1|1} + \hat{p}_{0|1} \ln \hat{p}_{0|1}) \\ &\quad + \hat{p}_0(\hat{p}_{1|0} \ln \hat{p}_{1|0} + \hat{p}_{0|0} \ln \hat{p}_{0|0}) \\ &= H(y_i) - H(y_i | y_{i-1}) \end{aligned}$$

where $\hat{p}_k = N_k/N$, $\hat{p}_{k|l} = N_{kl}/N_l$ are empirical estimates of probabilities $P(y_i = k)$ and $P(y_i = k | y_{i-1} = l)$, $H(y_i)$ is the empirical Shannon entropy function of y_i and $H(y_i|y_{i-1})$ is the empirical Shannon conditional entropy function of y_i given y_{i-1}. As the difference between these two entropies is the empirical Shannon mutual information $I(y_i, y_{i-1})$ between y_i and y_{i-1} we have the equivalent GLRT

$$\begin{aligned} \ln \hat{\Lambda} &= I(y_i; y_{i-1}) \\ &= \sum_{k,l=0,1} \hat{p}_{k|l} \hat{p}_l \ln \frac{\hat{p}_{k|l}}{\hat{p}_k} \overset{H_1}{\underset{H_0}{\gtrless}} \gamma \end{aligned}$$

For large N, $2NI(y_i, y_{i-1})$ is distributed as Chi square with 2 df, i.e. exponential with exponential parameter $\beta = 1/2$.

D.2 Unknown r_1, r_0, p_o: One Sided Composite Hypotheses

Here the test becomes decide H_1 if

$$\frac{N_{11}}{N_{11} + N_{01}} > \frac{N_{10}}{N_{10} + N_{00}} \text{ and } I(y_i, y_{i-1}) \overset{H_1}{\underset{H_0}{\gtrless}} \gamma$$

otherwise decide H_0.

E. Second Order Markov Model

The results generalize in an obvious way from the case $k = 1$ to the case $k = 2$. In particular, the LRT for known conditional success probabilities $r_{00}, r_{10}, r_{01}, r_{11}$ under H_1 and known success probability p under H_0 is analogous to (12)

$$\begin{aligned}\ln \Lambda &= N_{111} \ln(r_{11}/p) + N_{110} \ln(r_{10}/p) + N_{101} \ln(r_{01}/p) + N_{100} \ln(r_{00}/p) \\ &\quad + N_{011} \ln((1-r_{11})/(1-p)) + N_{010} \ln((1-r_{10})/(1-p)) N_{001} \ln((1-r_{01})/(1-p)) + N_{000} \ln((1-r_{00})/(1-p)) \mathop{\gtrless}_{H_0}^{H_1} \gamma\end{aligned}$$

where, e.g., N_{100} denotes the number of times in the sequence $\{y_i\}_{i=1}^N$ that (y_{i-1}, y_{i-2}) transition from $(0,0)$ to $y_i = 1$. The H_0 and H_1 distribution of $\ln \Lambda$ is well approximated by a Gaussian rv for $N \gg 2$. After calculating mean and variance under these hypotheses the threshold and the power function can be determined analytically.

When the r_{ij} are unknown but p is known, the GLRT is identical to the LRT above except that the success probabilities under H_1 are replaced by empirical estimates

$$\begin{aligned}\hat{r}_{11} &= \frac{N_{111}}{N_{111} + N_{011}} \\ \hat{r}_{10} &= \frac{N_{110}}{N_{110} + N_{010}} \\ \hat{r}_{01} &= \frac{N_{101}}{N_{101} + N_{001}} \\ \hat{r}_{00} &= \frac{N_{000}}{N_{000} + N_{100}}\end{aligned}$$

Under H_0 the test statistic $2 \ln \Lambda$ is distributed as Chi square with 3 df.

Finally, when both p and r_{ij} are unknown, it is easily shown that the GLRT becomes

$$H(y_i) - H(y_i | y_{i-1}, y_i) = I(y_i; y_{i-1}, y_{i-2}) \mathop{\gtrless}_{H_0}^{H_1} \gamma$$

where H and I denote empirically estimated entropies and mutual information quantities. In particular

$$I(y_i; y_{i-1}, y_{i-2}) = \sum_{k,l,m=0,1} \hat{p}_{k|l,m} \hat{p}_{l,m} \ln \frac{\hat{p}_{k|l,m}}{\hat{p}_k} \mathop{\gtrless}_{H_0}^{H_1} \gamma \qquad (15)$$

where $\hat{p}_{1|l,m} = N_{1lm}/(N_{1lm} + N_{0lm})$ is the empirical estimate of the second order transition probability and $\hat{p}_{lm} = N_{lm}/N$ is the bivariate probability estimate. Under H_0 the test statistic $2NI(y_i; y_{i-1}, y_{i-2})$ is distributed as a Chi square with 4 df.

IV. Tests for 1st Order Dependence

A completely analogous procedure to deriving the tests introduced in the last section yields simple and intuitive tests for 1st order Markov $\{y_i\}$ (given by (4) for general values of $k > 1$).

For simplicity first consider the special case of $k = 2$. When under H_0 the 1st order Markov transition probabilities $r_0 = P(y_i = 1 | y_{i-1} = 0)$, $r_1 = P(y_i = 1 | y_{i-1} = 1)$ are known and the 2nd order transition probabilities $r_{00}, \ldots, r_{11}$ are also known the log LRT is

$$\begin{aligned}&\ln \Lambda \\ =\ & N_{111} \ln(r_{11}/r_1) + N_{110} \ln(r_{10}/r_1) + N_{101} \ln(r_{01}/r_0) + N_{100} \ln(r_{00}/r_0) \\ & + N_{011} \ln((1-r_{11})/(1-r_1)) + N_{010} \ln((1-r_{10})/(1-r_1)) N_{001} \ln((1-r_{01})/(1-r_0)) + N_{000} \ln((1-r_{00})/(1-r_0)) \mathop{\gtrless}_{H_0}^{H_1} \gamma\end{aligned}$$

When $r_{00}, \ldots, r_{11}$ are unknown but r_0, r_1 are known the GLRT is given by the above test implemented with empirical estimates $\hat{r}_{00}, \ldots, \hat{r}_{11}$ defined in the previous section.

Finally, when in addition r_0, r_1 are unknown, the GLRT takes the form

$$H(y_i|y_{i-1}) - H(y_i|y_{i-1}, y_{i-2}) = I(y_i; y_{i-2}|y_{i-1}) \quad \overset{H_1}{\underset{H_0}{\gtrless}} \quad \gamma$$

where the empirical conditional entropies $H(y_i|y_{i-1})$ and $H(y_i|y_{i-1}, y_{i-2})$ and the conditional mutual information $I(y_i; y_{i-2}|y_{i-1})$ are defined via empirical transition probability estimates as in (15).

For $k > 2$ it is easily shown that the above GLRT's extend to

$$\begin{aligned} & \ln \hat{\Lambda} \\ = & \sum_{b_3, \ldots, b_k = 0,1} [N_{11b_3\ldots b_k} \ln(\hat{r}_{1b_3\ldots b_k}/r_1) + N_{01b_3\ldots b_k} \ln(1 - \hat{r}_{1b_3\ldots b_k})/(1 - r_1)] \\ & + \sum_{b_3, \ldots, b_k = 0,1} [N_{10b_3\ldots b_k} \ln(\hat{r}_{0b_3\ldots b_k}/r_0) + N_{00b_3\ldots b_k} \ln(1 - \hat{r}_{0b_3\ldots b_k})/(1 - r_0)] \quad \overset{H_1}{\underset{H_0}{\gtrless}} \quad \gamma \end{aligned}$$

for known r_0, r_1. The statistic $2 \ln \hat{\Lambda}$ is asymptotically Chi square with 2^k df under H_0.

For unknown r_0, r_1

$$H(y_i|y_{i-1}) - H(y_i|y_{i-1}, \ldots, y_{i-k}) = I(y_i; y_{i-2}, \ldots, y_{i-k}|y_{i-1}) \quad \overset{H_1}{\underset{H_0}{\gtrless}} \quad \gamma,$$

The test statistic $2I(y_i; y_{i-2}, \ldots, y_{i-k}|y_{i-1})$ has an asymptotic Chi square distribution under H_0 with 2^{k-1} degrees of freedom.

We remark that, due to bias-variance tradeoffs, there is probably an optimal value of k which maximizes the power of this GLRT test. This remains to be investigated.

References

[1] L. E. Baum and Welch, "A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, no. 2, pp. 164–171, 1970.
[2] R. J. Elliott, L. Aggoun, and J. Moore, *Hidden Markov models*, Springer-Verlag, New York, 1995.
[3] O. Zeitouni and A. Dembo, "Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes," *IEEE Trans. on Inform. Theory*, vol. IT-34, no. 4, pp. 890–893, 1988.
[4] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations," *IEEE Trans. on Inform. Theory*, vol. IT-32, no. 3, pp. 307–309, 1986.
[5] J. Goutsias and J. M. Mendel, "Optimal simultaneous detection and estimation of filtered discrete semi-Markov chains," *IEEE Trans. on Inform. Theory*, vol. IT-34, no. 3, pp. 551–568, 1988.
[6] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, no. 1, pp. 12–40, 1961.
[7] J. L. Denny and A. L. Wright, "On tests for Markov dependence," *Z. Wahrsch. Verw. Gebiete Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, no. 4, pp. 331–338, 1978.
[8] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, 1959.
[9] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.
[10] H. L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.