Comparing Exact Bayesian and BIC Markov Order Classifiers

H. S. Bhat[*][†] N. Kumar[*]

January 31, 2011

Abstract

We use an exact Bayesian calculation to design classifiers that distinguish whether a finite sequence drawn from a finite alphabet is a sample path of a Markov chain of order $k = 0$ or of order $k > 0$. Three exact Bayes (EB) classifiers are derived, each corresponding to a different prior. We also include a classifier based on the Bayesian Information Criterion (BIC), a popular technique for Markov order estimation. Using thousands of random Markov chains of known order, we test the performance of the classifiers. In both average accuracy and ROC analyses, we find that EB classifiers with informative priors perform better than the BIC classifier, with the difference becoming strikingly large when either the size of the alphabet is large or the length of the sequence is small. We also test the classifiers on five real-world data sets and find that the EB classifications, unlike the BIC classifications, match the orders of the models with highest out-of-sample predictive accuracies.

Keywords: Markov chains, order estimation, classifier, Bayesian, computational statistics

1 Introduction

Markov chains are a natural class of models for sequences of symbols drawn from a finite alphabet. Suppose we are given such a sequence $\{s_n\}_{n=1}^N$. Assuming that the sequence is a sample path of some Markov chain of order k, the problem of Markov order estimation is to use data to estimate k. Note that there is a large conceptual difference between order $k = 0$ and order $k > 0$. In the $k = 0$ case, the data is modeled by an IID process, while in the $k > 0$ case, the data is modeled by a process $\{S_n\}_{n=1}^N$ where the distribution of the random variable S_n depends on past observations of $S_{n-k}, \ldots, S_{n-1}$.

In this paper, we view the problem of using data to distinguish between Markov chains of order $k = 0$ and $k > 0$ as a classification problem. We perform an exact Bayesian (EB) calculation for a Markov chain with q symbols and order k, using three priors: the Dirichlet

[*] School of Natural Sciences, University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343

[†] Corresponding author, email: hbhat@ucmerced.edu

prior and two types of uniform prior. These derivations may be found in Section 2. In Section 3, we introduce a *margin function* whose sign classifies the sequence—positive margin corresponds to $k = 0$, and negative margin corresponds to $k > 0$. We then test our EB classifier against a classifier based on the Bayesian information criterion (BIC), a popular Markov order estimator [Guttorp, 1995]. The BIC penalty term is an approximation of the exact Bayesian calculation carried out here.

Our tests involve generating many thousands of sequences from randomly generated Markov chains, and then using both BIC and EB margin functions to classify these sequences as either $k = 0$ or $k > 0$. We use the magnitude of the margin function as a proxy for the confidence of the order estimator. The margin function enables us to compare BIC and EB estimators using receiver operating characteristic (ROC) curves, which account for both the true positive and false positive rates. The ROC curves conclusively show that the EB classifier outperforms the classifier based on the BIC estimator, especially for data sets that are either small in length or feature symbols drawn from a large alphabet.

In Section 4, we demonstrate the utility of the EB classifier on five real data sets: (1) daily air quality index data, (2) weekly corn export data, (3) hourly energy price data, (4) major league baseball attendance data, and (5) weekly unemployment data. In each case, we find that BIC classifies the sequence as $k = 0$, while the EB classifiers all indicate $k > 0$. Furthermore, in all of these cases, we find that a first-order Markov chain model has significantly higher out-of-sample predictive accuracy than a zeroth-order Markov model. These tests with real data confirm that in situations where either the alphabet is large or the data set is not large, the EB classifier is to be preferred over the BIC classifier.

1.1 Past Work

The literature on Markov order estimation is primarily concerned with asymptotic properties, especially consistency—see [Katz, 1981], [Csiszár and Shields, 2000], [Zhao et al., 2001], [Csiszár, 2002], [Morvai and Weiss, 2005], [Peres and Shields, 2005], and [Dorea and Lopes, 2006]. Consistency of a Markov order estimator means that in the limit where the length of the data sequence goes to infinity, the estimator returns the exact order of the Markov chain. Aside from consistency, there are other properties of order estimators that hold in the limit of infinite data [Finesso et al., 1996]. This asymptotic point of view goes back at least to [Billingsley, 1961].

In this work, when we concern ourselves with the performance of classifiers based on Markov order estimators, we are concerned only with their behavior for sequences of finite length. We are interested in the following question: if we have a sequence of length n and the size of the alphabet is q, then among the BIC and EB classifiers, which one is best? Asymptotic properties of the order estimators will be of limited importance, at best, on practical problems of this kind, a fact noted by [Csiszár and Shields, 2000].

Furthermore, [Dalevi et al., 2006] show that consistent order estimators such as BIC perform relatively poorly when the amount of data is not large. This explains the continuing use of the Akaike Information Criterion (AIC) order estimator, proven inconsistent 30 years ago [Katz, 1981]. The results of [Dalevi et al., 2006] on biological data sets show that, in certain contexts, the inconsistent AIC estimator performs *better* than the consistent BIC estimator. In short, if the amount of data is not large, then using a consistent estimator simply because it is consistent may be unwise.

Bayesian methods for Markov order estimation are discussed in two papers of which we are aware. In the first, [Csiszár and Shields, 2000] mention the Bayesian approach only to point out a situation in which it is provably inconsistent. Specifically, they show that if the EB with a Dirichlet prior of $\boldsymbol{\alpha} = (1/2, 1/2, \ldots, 1/2)$ is maximized over *all* candidate orders $k \geq 0$, then there exists an IID sequence (i.e., of true Markov order zero) such that in the $n \to \infty$ limit of an infinite-length sequence, the EB order estimate $\hat{k}$ goes to ∞ as well.

In our work, we include as one of our tests the situation described by Csiszár and Shields. That is, we apply the EB classifier to IID sequences of increasing lengths n. Note that our EB classifier tests a given sequence for candidate orders $0 \leq k \leq 5$. The tests show that, as n increases, the error rate of the EB classifier goes to zero. Again, this shows that consistency need not be the primary consideration for using or rejecting a given order estimator.

In the second paper [Strelioff et al., 2007], exact Bayesian calculations for Markov order estimation are pursued. The problem is not viewed as a classification problem, and the paper does not include tests of which order estimator performs better as a function of (1) alphabet size q, (2) sequence length n, and (3) the closeness of the prior used in the Bayesian calculation to the true distribution used to generate the transition matrices for the Markov chains under consideration. The last point deserves further comment.

Let us assume that one has a sequence generated by a Markov chain of some unknown order k. Let us further assume that the transition matrix for this Markov chain is itself drawn from a distribution of transition matrices. One will never know with certainty what this distribution is. In our work, we show that when using the EB classifier with Dirichlet prior, even if the parameters $\boldsymbol{\alpha}$ are not known perfectly, the classifier is nearly indistinguishable from an EB classifier with perfect knowledge of the prior. Furthermore, the EB classifier with approximate prior is far better than the BIC classifier.

2 Exact Bayesian Calculation

The Bayesian approach to model selection [Ghosh et al., 2006] is to maximize the posterior probability of a model (M_i) given the data $\{y_j\}_{j=1}^{n}$. Applying Bayes theorem to calculate the posterior probability of a model given the data, we get

$$P(M_i|y_1,\dots,y_n) = \frac{P(y_1,\dots,y_n|M_i)P(M_i)}{P(y_1,\dots,y_n)}, \tag{1}$$

where $P(y_1,\dots,y_n|M_i)$ is called the marginal likelihood of the model M_i.

If all candidate models are equally likely, then maximizing the posterior probability of a model given the data is the same as maximizing the marginal likelihood

$$P(y_1,\dots,y_n|M_i) = \int_{\boldsymbol{\Theta}_i} L(\boldsymbol{\theta}_i|y_1,\dots,y_n)g_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i, \tag{2}$$

where $\boldsymbol{\theta}_i$ denotes the vector of parameters in the model M_i, L is the likelihood function and $g_i(\boldsymbol{\theta}_i)$ is the p.d.f. of the distribution of parameters $\boldsymbol{\theta}_i$. *We refer to (2) as the Exact Bayesian (EB) marginal likelihood and evaluate it for a few priors.*

2.1 BIC

Let $|\boldsymbol{\theta}_i|$ denote the number of free parameters $\boldsymbol{\theta}_i$ in model M_i. Let $\hat{\boldsymbol{\theta}}_i$ denote the maximum likelihood estimate of the parameters $\boldsymbol{\theta}_i$. Then, instead of maximizing the marginal likelihood directly, a method that is commonly used is to maximize $\log L(\hat{\boldsymbol{\theta}}_i|y_1,\dots,y_n) - (|\boldsymbol{\theta_i}|/2)\log n$ over a set of candidate models. The penalty term $(|\boldsymbol{\theta_i}|/2)\log n$ is called the Bayesian Information Criterion (BIC) penalty. In the Appendix, we provide a careful derivation of this penalty term starting from (2) and applying, in turn, Laplace's method, a flat prior, and the weak law of large numbers.

In our tests, the k-th model will be a k-th order Markov chain over an alphabet of size q, for which the number of free parameters is $(q-1)q^k$. We define

$$B(k) = \log L(\hat{\boldsymbol{\theta}}_k|y_1,\dots,y_n) - \frac{(q-1)q^k}{2}\log n, \tag{3}$$

the BIC-penalized log likelihood for candidate order k.

2.2 Likelihood Function

Let us consider a k-th order Markov chain with q symbols where $k \geq 0$ and $q \geq 2$. Let us first account for the different probabilities (model parameters) in this model. Every outcome in a k-th order Markov chain depends on k previous outcomes. Thus, for each sequence of length k in the Markov chain, there are q different probabilities (one for each symbol) that sum to 1. Since there are such q^k distinct sequences, we get a total of q^{k+1} transition probabilities. Let us denote these transition probabilities as

$$p_1,\dots,p_q,p_{q+1},\dots,p_{2q},\dots,p_{q^{k+1}-q+1}\cdots,p_{q^{k+1}}$$

with the probabilities arranged such that $\sum_{i=mq+1}^{(m+1)q} p_i = 1$ for $m = 0,\ 1,\ \ldots,\ q^k - 1$. Let us denote the q^k distinct sequences as $S_0,\ S_1,\ \ldots,\ S_{q^k-1}$. Let us denote $n_{mq+1},\ n_{mq+2},\ \ldots,\ n_{(m+1)q}$ to be the number of times the sequences $S_m 1,\ S_m 2,\ \ldots,\ S_m q$ are observed in the given Markov chain for all of $m = 0,\ 1,\ \ldots,\ q^k - 1$. Then the likelihood function is

$$L(p_1, \ldots, p_{q^{k+1}} | y_1, \ldots, y_n) = P(y_1, \ldots, y_k) \prod_{j=1}^{q^{k+1}} p_j^{n_j}. \tag{4}$$

Hence, for a Markov chain, the marginal likelihood (2) is

$$P(y_1, \ldots, y_n | M_{k;q}) = P(y_1, \ldots, y_k) \int_{\boldsymbol{D}} \prod_{j=1}^{q^{k+1}} p_j^{n_j} g(\mathbf{p}) d\mathbf{p},$$

where

$$\boldsymbol{D} = \Big\{ \mathbf{p} \in \mathbb{R}^{q^{k+1}} \Big| \sum_{i=mq+1}^{mq+q} p_i = 1 \text{ for } m = 0, 1, \ldots, q^k - 1, \forall p_i \geq 0 \Big\} \subset \mathbb{R}^{q^{k+1}}. \tag{5}$$

$\boldsymbol{D}$ can be factored as $\boldsymbol{D} = \boldsymbol{D}_0 \times \boldsymbol{D}_1 \times \ldots \times \boldsymbol{D}_{q^k-1}$ where

$$\boldsymbol{D}_m = \Big\{ \mathbf{p} \in \mathbb{R}^q \Big| \sum_{i=mq+1}^{mq+q} p_i = 1, \quad \forall p_i \geq 0 \Big\}.$$

Each $\boldsymbol{D}_m$ is a $(q-1)$-simplex. Since $g = g_0 g_1 \cdots g_{q^k-1}$, we can express the marginal likelihood as

$$P(y_1, \ldots, y_n | M_{k;q}) = P(y_1, \ldots, y_k) \prod_{m=0}^{m=q^k-1} \int_{\boldsymbol{D}_m} \prod_{i=mq+1}^{(m+1)q} p_i^{n_i} g_m d\mathbf{p}. \tag{6}$$

2.2.1 Remark

Our expression for (6) differs from that given by [Strelioff et al., 2007] in that we include the factor $P(y_1, \ldots, y_k)$, the probability of observing the first k symbols in the sequence. In a k-th order Markov chain, the model requires k past symbols to predict the next one; in a sequence of length $n > k$, the first symbol that could have been predicted is y_{k+1}. This means that some assumption must be made about $P(y_1, \ldots, y_k)$. We assume it equals q^{-k}, i.e., the symbols y_1 through y_k are generated by k rolls of a fair die with q faces.

2.3 Dirichlet Prior

Since the functional form of the likelihood function is the same as the Dirichlet distribution, the Dirichlet distribution is its conjugate prior. We show how to evaluate (6) using this prior.

In what follows, we make use of the following definition:

$$\lambda(m,q,\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=mq+1}^{(m+1)q} \alpha_i)}{\prod_{i=mq+1}^{(m+1)q} \Gamma(\alpha_i)}.$$

The Dirichlet prior with parameter $\boldsymbol{\alpha}$ is

$$g_m(\mathbf{p}) = \lambda(m,q,\boldsymbol{\alpha}) \prod_{i=mq+1}^{(m+1)q} p_i^{\alpha_i - 1}, \tag{7}$$

where $\sum_{i=mq+1}^{(m+1)q} p_i = 1$ and $0 \leq p_i \leq 1$. Inserting (7) in (6), we find

$$\begin{aligned} P(y_1,\ldots,y_n|M_{k;q}) &= P(y_1,\ldots,y_k) \prod_{m=0}^{m=q^k-1} \lambda(m,q,\boldsymbol{\alpha}) \prod_{m=0}^{m=q^k-1} \int_{\boldsymbol{D}_m} \prod_{i=mq+1}^{(m+1)q} p_i^{n_i+\alpha_i-1} d\mathbf{p} \\ &= P(y_1,\ldots,y_k) \prod_{m=0}^{m=q^k-1} \lambda(m,q,\boldsymbol{\alpha}) \prod_{m=0}^{m=q^k-1} (\lambda(m,q,\mathbf{n}+\boldsymbol{\alpha}))^{-1} \end{aligned}$$

As remarked above, we assume $P(y_1,\ldots,y_k) = q^{-k}$, so

$$\mathrm{EB}_\alpha(k) = \frac{1}{q^k} \prod_{m=0}^{m=q^k-1} \frac{\lambda(m,q,\boldsymbol{\alpha})}{\lambda(m,q,\mathbf{n}+\boldsymbol{\alpha})}. \tag{8}$$

2.4 Uniform Prior

As stated before, $\boldsymbol{D}_m$ is a $(q-1)$-simplex and the volume of this simplex is $1/(q-1)!$. By definition the p.d.f. of the uniform distribution over this simplex is

$$g_m(\mathbf{p}) = \Gamma(q) = (q-1)!.$$

Plugging this uniform prior into (6) yields

$$\begin{aligned} \mathrm{EB}_1(k) &= \frac{1}{q^k} (\Gamma(q))^{q^k} \prod_{m=0}^{m=q^k-1} \int_{\boldsymbol{D}_m} \prod_{i=mq+1}^{(m+1)q} p_i^{n_i} d\mathbf{p} \\ &= \frac{1}{q^k} (\Gamma(q))^{q^k} \prod_{m=0}^{m=q^k-1} (\lambda(m,q,\mathbf{n}+\mathbf{1}))^{-1} \end{aligned} \tag{9}$$

As the notation suggests, we could have obtained (9) by inserting $\alpha = 1$ into (8). This is because the uniform prior $g_m(\mathbf{p})$ is equivalent to the Dirichlet prior with $\boldsymbol{\alpha} = (1,1,\ldots,1)$.

2.5 Uniform Prior on a Subset of the Simplex

We revisit the case of the uniform prior with $q = 2$. Let us consider a k-th order Markov chain. The prior defined in Section 2.4 assumes that the transition probabilities are uniformly distributed in all of the 1-simplex, i.e., $p_{2m+1} \sim U(0, 1)$ under the constraint that $p_{2m+1} + p_{2m+2} = 1$ for all $m = 0, 1, \ldots, 2^k - 1$.

However, when $q = 2$, it is possible to consider the case where the transition probabilities are uniformly distributed on only a subset of the the 1-simplex, i.e., $p_{2m+1} \sim U(a, b)$ with either $a > 0$ or $b < 1$. The constraint $p_{2m+1} + p_{2m+2} = 1$ forces $p_{2m+2} \sim U(1 - b, 1 - a)$. Let $\hat{\boldsymbol{D}}_m$ denote the subset of the 1-simplex defined by the constraints imposed on the probabilities. Then the EB marginal likelihood of a Markov chain with such a prior is

$$\begin{aligned} \mathrm{EB}_{U(a,b)}(k) &= P(y_1, \ldots, y_k) \left(\frac{1}{a-b}\right)^{2^k} \prod_{m=0}^{2^k-1} \int_{\hat{\boldsymbol{D}}_m} (p_{2m+1})^{n_{2m+1}} (p_{2m+2})^{n_{2m+2}} d\mathbf{p} \\ &= \frac{1}{2^k} \left(\frac{1}{a-b}\right)^{2^k} \prod_{m=0}^{2^k-1} (\beta_1 - \beta_2), \end{aligned} \tag{10}$$

with

$$\begin{aligned} \beta_1 &= \beta(b; n_{2m+1} + 1, n_{2m+2} + 1) \\ \beta_2 &= \beta(a; n_{2m+1} + 1, n_{2m+2} + 1), \end{aligned}$$

where β is the incomplete beta function, which can be evaluated using standard numerical algorithms.

2.5.1 Remark

The above choice of prior was not considered by [Strelioff et al., 2007]. We have included this prior for two reasons:

1. In our past work on option pricing using a Markovian tree [Bhat and Kumar, 2010], we found that the empirical transition probabilities for a a two-state Markov chain model of the underlying stock were often close to 0.5.

2. As $\epsilon > 0$ decreases to zero, it becomes increasingly difficult for any order estimator to distinguish the sample paths of a first-order Markov chain with transition probabilities drawn from a $U(0.5 - \epsilon, 0.5 + \epsilon)$ distribution from the sample paths of a fair coin-tossing (IID Bernoulli with $p = 0.5$) process. We would like to determine whether an EB classifier that "knows" the distribution from which the transition probabilities are drawn performs better than a classifier that is ignorant of this distribution.

2.6 EB with Model Bias

Let us restate the problem of order estimation and our attack thus far. We are given a sample path of a Markov chain with q symbols and asked to find the order of the Markov chain. Even though q figures in our calculations, it is given to us and we do not have to estimate it. Our task is to calculate the posterior probability (1) for different candidate orders $k = 0$ to K. We then set $\hat{k}$ equal to the candidate order that maximizes the posterior probability. By assuming that all candidate models are equally likely, we conclude that maximizing the marginal likelihood of the model is the same as maximizing the posterior probability of the model given the data.

In the spirit of parsimony, let us consider a modification of the EB that penalizes for a higher order Markov chain. We assume that $P(M_i)$ in (1) is different for different candidate orders. Specifically, for any model $M_{k;q}$ with q symbols and candidate order k, we assume

$$P(M_{k;q}) = \frac{\Gamma(q)^{-q^k}}{N_{M_q}}, \tag{11}$$

where $N_{M_q} = \sum_{k=0}^{K} \Gamma(q)^{-q^k}$, a normalizing constant. Since the different candidate models are no longer equally likely, maximizing the marginal likelihood is not the same as maximizing the posterior probability of the model given the data. Thus, the order estimate $\hat{k}$ is the value of k for which the following quantity is maximized among all candidate k's,

$$P(y_1, \ldots, y_n | M_{k;q}) P(M_{k;q}) = \frac{\Gamma(q)^{-q^k}}{N_{M_q}} \int_{\boldsymbol{D}} L(p_1, \ldots, p_{q^{k+1}} | y_1, \ldots, y_n) g(\mathbf{p}) d\mathbf{p}.$$

Assuming a uniform prior for the probabilities, the above quantity becomes

$$\begin{aligned} P(y_1, \ldots, y_n | M_{k;q}) P(M_{k;q}) &= \frac{1}{N_{M_q}} \int_{\boldsymbol{D}} L(p_1, \ldots, p_{q^{k+1}} | y_1, \ldots, y_n) d\mathbf{p} \\ &= \frac{1}{N_{M_q}} P(y_1, \ldots, y_k) \prod_{m=0}^{m=q^k-1} \int_{\boldsymbol{D}_m} \prod_{i=mq+1}^{(m+1)q} p_i^{n_i} d\mathbf{p} \end{aligned}$$

Since N_{M_q} is same for all candidate orders, maximizing the above quantity over all candidate orders is the same as maximizing

$$\mathrm{EB}^M(k) = \frac{1}{q^k} \prod_{m=0}^{m=q^k-1} (\lambda(m, q, \mathbf{n} + \mathbf{1}))^{-1} \tag{12}$$

Note that the expression for $\mathrm{EB}^M(k)$ is what would have been obtained if we had integrated (4) over the domain $\mathbf{D}$ defined in (5). In other words, the expression for $\mathrm{EB}^M(k)$ is algebraically equivalent to evaluating

$$\int_{\boldsymbol{\Theta}_i} L(\boldsymbol{\theta}_i|y_1,\ldots,y_n)d\boldsymbol{\theta}_i,$$

the right-hand side of (2) with the g_i term deleted. For this reason, we think of $\mathrm{EB}^M(k)$ as an exact Bayesian result with a flat prior [Wasserman, 2003, Chap 11.6]. This EB with model bias does not appear in earlier works such as [Strelioff et al., 2007].

3 Classification and Results

We can now convert the derivations from the previous section into classifiers for distinguishing between Markov chains of order $k = 0$ and $k > 0$. We accomplish this by introducing a margin function. Different Markov order estimation techniques yield different results for the same data set, making it difficult to decide the optimal order. By simulating Markov chains of known orders, and generating sequences of varying lengths and alphabet sizes, we compare the outputs of different classifiers against the ground truth. Through extensive testing, we obtain practical guidance on which classifier performs the best in a particular set of conditions.

3.1 Margin Function

The EB marginal likelihood (6) gives us the the probability of candidate order k. We have denoted it as $\mathrm{EB}(k)$—with descriptive subscripts—since it is a function of k with q held fixed. Recall that $B(k)$ is the BIC estimate (3) of the log probability of candidate order k.

Generally, given $E(k)$, the difference of log likelihoods $\log E(k_1) - \log E(k_2)$ conveys the relative confidence that the true order is k_1 versus k_2. Thus we define the EB and BIC margins ψ_{EB} and ψ_{BIC}, respectively, to be

$$\psi_{\mathrm{EB}} = \log \mathrm{EB}(0) - \max_{1\le k\le K} \log \mathrm{EB}(k), \tag{13a}$$

$$\psi_{\mathrm{BIC}} = B(0) - \max_{1\le k\le K} B(k). \tag{13b}$$

In this paper, we choose $K = 5$. Let us also define

$$\hat{k}_{\mathrm{EB}} = \operatorname*{arg\,max}_{0\le k\le K} \mathrm{EB}(k), \quad \hat{k}_{\mathrm{BIC}} = \operatorname*{arg\,max}_{0\le k\le K} B(k). \tag{14}$$

Now note that if $\hat{k}_{\mathrm{EB}} = 0$, then $\psi_{\mathrm{EB}} > 0$, and if $\hat{k}_{\mathrm{EB}} > 0$, then $\psi_{\mathrm{EB}} < 0$. The same relationship holds between $\hat{k}_{\mathrm{BIC}}$ and ψ_{BIC}. We therefore use the sign of the margin function to classify a given sequence. Negative (respectively, positive) margin means that we classify the sequence as having arisen from a Markov chain of order $k > 0$ (respectively, $k = 0$).

3.2 Tests

We now test the performance of the classifiers on random sequences generated using random Markov chains of known order k. For each fixed value of the order k and the alphabet size q, we sample a distribution F of transition probability matrices to generate a random Markov chain. We then use this Markov chain to generate a sequence of length n.

For the tests in Sections 3.2.1, 3.2.2, and 3.2.3, the transition matrix distribution F is, respectively, a uniform distribution (equivalent to a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (1, \ldots, 1)$), a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$, and a uniform distribution on $[0.4, 0.6]$ for the special case of $q = 2$ symbols.

We set, in turn, $k = 0$, $k = 1$, and $k = 2$. For each (q, n, F), we generate 1000 sequences (i.e., instances) of each order k, so that we have a total of 3000 instances. We apply the EB and BIC classifiers to all of these instances and save all resulting margin function values. This procedure is repeated 10 times.

The results of the tests are displayed in two forms, as tables showing the average percentage of correctly classified instances, and also as ROC curves, where classified instances are ranked/scored by margin function as in [Fawcett, 2006]. We define the positive class label to be $k = 0$ and the negative class label to be $k > 0$, again corresponding to the sign of the margin function. The true positive rate is the number of correctly classified positive instances divided by the number of positive instances. The false positive rate is the number of correctly classified negative instances divided by the number of negative instances. Note that each ROC curve we display is the average of ten raw ROC curves.

3.2.1 F = Uniform

We follow the procedure in Section 3.2 with F equal to the uniform distribution, i.e., the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (1, \ldots, 1)$. Tables 1, 2, 3 and 4 show the overall percentage of correctly classified instances for each of the following four classifiers in turn: EB_1, EB_α, BIC, and EB^M. Tests involving the EB_α classifier are intended to simulate the situation where we assume a Dirichlet prior but do not know that the true $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$ is $\alpha = 1$. For these tests, $\alpha \sim U(0.5, 1.5)$ distribution.

Table 1 shows that for any alphabet size $3 \leq q \leq 10$, as the sequence length n increases, the EB_1 classifier achieves near-perfect classification. The classifier performs better when q is larger.

Table 2 shows that imprecise knowledge of the parameter α in the Dirichlet prior does not hurt the classifier's success rate. Again, for each $q \geq 3$, as n increases, the EB_α classifier achieves either perfect or near-perfect classification.

For both the EB_1 and EB_α classifiers, the $q = 2$ case is where the classifier displays its worst performance, relative to larger values of q. However, if we focus only on the $q = 2$ results, it is clear that the classifier improves as n increases. Indeed, in further tests, we can confirm that as n increases beyond the upper limit of 500 in the tables, the classifiers achieve near-perfect classification.

Table 3 shows the results for the BIC classifier. It is apparent that for each fixed q, as n increases, the classification success rate improves. For the BIC classifier, we know based on [Katz, 1981] and [Csiszár and Shields, 2000] that as $n \to \infty$, the percentages must converge to 100. However, our tests show two phenomena that consistency theory does not bear out. First, for each fixed n, as q increases from 4, performance worsens. Second, if we compare row q of Table 3 to row q of either Tables 1 or 2, for $q \geq 4$, we see that the performance of the BIC classifier is dramatically worse than that of either the EB_1 or EB_α classifiers.

Table 4 shows results for the EB classifier with model bias, EB^M. The results are broadly similar to those for the BIC classifier. The conclusions of the previous paragraph hold with the critical row $q = 4$ replaced by $q = 5$.

For six (n, q) pairs, averaged ROC curves for all four classifiers are plotted in Figure 1. Magenta, black, red, and blue curves correspond to the EB_1, EB_α, EB^M, and BIC classifiers, respectively. In the legend, we report the areas under the curves.

Across all of the plots in Figure 1, we see that the EB_1 and EB_α classifiers achieve areas under the ROC curve greater than 0.99; these areas are, for each (n, q) pair, strictly greater than the areas under the ROC curves for the EB^M and BIC classifiers. From these curves, we form the following hypothesis.

> Hypothesis P: using an approximately correct prior distribution of Markov transition probabilities increases the performance of a Markov order classifier.

This hypothesis will be revisited below.

3.2.2 F = Dirichlet

We follow the procedure in Section 3.2 with F equal to a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$ for different values of α. For these tests, the first two classifiers are now EB_α, a classifier whose Dirichlet prior uses the true value of α, and $\text{EB}_{\widetilde{\alpha}}$, a classifier whose Dirichlet prior uses an approximation of the true α given by $\widetilde{\alpha} = \gamma\alpha$, where $\gamma \sim U(0.1, 1.9)$.

We do not include tables of overall classifier success rates; if were to include these tables, they would show the same trends and lead to the same conclusions discussed above.

For six (n, q, α) tuples, we plot averaged ROC curves for all four classifiers in Figure 2. Magenta, black, red, and blue curves correspond to the EB_α, $\text{EB}_{\widetilde{\alpha}}$, EB^M, and BIC classifiers, respectively. In the legend, we report the areas under the curves.

Unlike the uniform distribution (i.e., when $\alpha = 1$), the distributions featured in these tests are peaked, more strongly so as α increases. This feature of the distributions makes it harder to distinguish sequences generated by $k = 0$ and $k > 0$ Markov chains. As can be seen in the six plots in Figure 2, the EB_{α} and $\mathrm{EB}_{\widetilde{\alpha}}$ classifiers perform significantly better than the EB^{M} and BIC classifiers. This adds weight to Hypothesis P.

In particular, note that when $\alpha = 20$, the areas under the ROC curves for the EB^{M} and BIC classifiers is closer to 0.5 (the area under the ROC curve for a classifier that guesses randomly) than it is to 1. Also, note that in all six cases, the magenta and black curves nearly coincide, leading one to believe that approximate knowledge of the parameter α in the Dirichlet prior is just as good as perfect knowledge of α.

3.2.3 F = Uniform(0.4,0.6)

Now we restrict our attention to an alphabet of size $q = 2$. We follow the procedure in Section 3.2 with F equal to a uniform distribution on the interval $(0.4, 0.6)$; the EB marginal likelihood using this distribution as a prior was discussed in section 2.5.

For these tests, the first two classifiers are $\mathrm{EB}_{U(0.4,0.6)}$ and EB_{37}. The parameter $\alpha = 37$ was chosen by matching the mean and variance of the Dirichlet distribution with that of the uniform distribution on $(0.4, 0.6)$.

For six values of n, we plot averaged ROC curves for all four classifiers in Figure 3. Magenta, black, red, and blue curves correspond to the $\mathrm{EB}_{U(0.4,0.6)}$, EB_{37}, EB^{M}, and BIC classifiers, respectively. In the legend, we report the areas under the curves.

As was remarked at the end of Section 2.5, we chose to examine this distribution because it leads to sequences for which it is very difficult to determine the true order k. As we might have expected, the ROC curves show that none of the classifiers does exceptionally well at distinguishing $k = 0$ sequences from $k > 0$ sequences.

However, as we found in previous tests, the magenta and black curves nearly coincide. This shows that even if the prior is chosen from a family of distributions that is different from the distribution F, matching the mean and variance of F yields a classifier that is, for all practical purposes, indistinguishable from a classifier that uses a prior distribution identical to F.

Also, as we found in previous tests, the areas under the magenta and black curves are always the largest of the four areas. There is consistent agreement that $\mathrm{EB}_{U(0.4,0.6)}$ and EB_{37} are the best-performing classifiers on this difficult set of tests. This adds further weight to Hypothesis P.

3.3 Csiszár/Shields Inconsistency Example

In [Csiszár and Shields, 2000], it is shown that the order estimator

$$\hat{k}^{\mathrm{CS}}_{\mathrm{EB}} = \arg\max_{0 \le k} \mathrm{EB}(k) \tag{15}$$

is inconsistent. Note that (15) differs from (14) in that the set of candidate orders k is no longer bounded above by K. With the order estimator given by (15), inconsistency is shown explicitly for IID sequences over an alphabet of size q where the probabilities $p_1, \ldots, p_q$ are drawn uniformly on $\boldsymbol{D}$ defined in (5). The EB considered in [Csiszár and Shields, 2000] is, in our notation, $\mathrm{EB}_{0.5}$.

In order to demonstrate that the inconsistency result from [Csiszár and Shields, 2000] does not detract from the EB classifiers considered in this paper, we run the following test. For each fixed pair (q, n), we take 30000 samples from the uniform distribution on $\boldsymbol{D}$ and use each sample to generate an IID sequence of length n. Each sequence generated in this way is an instance. We then classify all 30000 instances using 5 different classifiers: EB_1, $\mathrm{EB}_{0.5}$, EB_{10}, EB^M, and BIC. For each classifier, and for each (q, n) pair, we report the average percentage of correctly classified instances in Table A.

The tests demonstrate three trends. First, as we found in earlier tests, for fixed q, as n increases, the EB_α success rate approaches 100%. This remains true regardless of whether we use the true value $\alpha = 1$, or whether we use $\alpha = 0.5$ or $\alpha = 10$.

Second, the BIC classifier has the best percentages and, moreover, these percentages approach 100 very quickly as n increases from 50. We can now return to our earlier results in Table 3; we see that when $(q, n) = (6, 50)$, the BIC classifier correctly classifies 33.33% of all instances. In light of the results in Table A, we see that these correct classifications must consist almost exclusively of correctly identifying $k = 0$ sequences. The BIC classifier's accuracy for $k > 0$ sequences improves very gradually as n increases.

Finally, we note that for fixed n, as q increases, the BIC classifier's performance on $k = 0$ instances improves. This is the opposite behavior of what we found in our earlier tests where $k = 1$ and $k = 2$ instances were included. This leads us to believe that the BIC classifier heavily prefers a prediction of $\hat{k}_{\mathrm{BIC}} = 0$. For large q, or for small n, this preference is so strong that it causes the BIC classifier to have a large error rate on $k > 0$ sequences.

4 Applications

In this section, we apply the EB and BIC order estimators to five different real-world data sets. In raw form, each data set we consider is a real-valued time series. For the first example, that of the Air Quality Index (AQI), we use a table provided by the US Environmental Protection Agency (EPA) to convert the real-valued time series into a sequence over an alphabet of size $q = 6$.

For the other examples, we use quantiles to convert the real-valued time series into a sequence over a finite alphabet. For a given time series $\{s_n\}_{n=1}^N$, let $s_{\min}$ and $s_{\max}$ be the minimum and maximum values encountered in the time series. We define the empirical quantile $Q(x)$ as the smallest y such that

$$\frac{\#\text{ of } s_n \text{ such that } s_n \leq y}{N} = x.$$

Now consider the $q - 1$ quantiles $Q(j/q)$, where $j = 1, 2, \ldots, q - 1$. Let $\{\sigma_1, \sigma_2, \ldots, \sigma_q\}$ be an alphabet of size q. Define the mapping

$$\phi(x) = \begin{cases} \sigma_1 & s_{\min} \leq x < Q(1/q) \\ \sigma_2 & Q(1/q) \leq x < Q(2/q) \\ \vdots & \vdots \\ \sigma_j & Q((j-1)/q) \leq x < Q(j/q) \\ \vdots & \vdots \\ \sigma_q & Q((q-1)/q) \leq x \leq s_{\max}. \end{cases}$$

Applying ϕ to each element in the time series, we obtain a sequence of length n over the alphabet of size q. We then feed this sequence into each of the four classifiers considered in this paper, resulting in four margin function values that we record. Recall that the margin function's sign indicates a classification of $k = 0$ (positive margin) or $k > 0$ (negative margin).

For each data set, we then analyze the out-of-sample predictive accuracy of both zeroth- and first-order Markov models. The procedure for doing this for a k-th order Markov model and a sequence $\{\phi(s_n)\}_{n=1}^{N}$ is as follows: for each m from k to $N-2$,

1. Train a k-th order Markov chain model using $T_m = \{\phi(s_n)\}_{n=1}^{m+1}$. By "train," we mean that one uses the training set T_m to calculate maximum likelihood estimates of all transition probabilities. The trained model is a matrix of transition probabilities.
2. Use the trained model and present/past data points $U_k = \{\phi(s_n)\}_{n=m+2-k}^{m+1}$ to form a one-step ahead prediction ξ_{m+2}: use the k data points U_k to single out a row of the transition probability matrix, let r be the column at which the maximum probability is found, and then set $\xi_{m+2} = \sigma_r$. Compare ξ_{m+2} against the true symbol $\phi(s_{m+2})$.

We will report the average success rate

$$\mathcal{S}(k) = \nu/(N-1-k), \tag{16}$$

where ν is the number of times ξ_{m+2} equals $\phi(s_{m+2})$ across all the tests from $m = k$ to $m = N-2$.

4.1 Descriptions of the Data Sets

Here we describe the five real-world data sets that were used in this study:

1. **Air Quality Index.** We downloaded daily ozone data for Merced County (California) in 2009 from the US EPA AirExplorer web site[1]. Ground-level ozone is one of the contaminants that have an associated Air Quality Index (AQI). We used an EPA-provided table[2]

[1] http://www.epa.gov/airexplorer/index.htm

[2] http://www.airnow.gov/index.cfm?action=aqibasics.aqi

to convert the real-valued ozone data into a sequence over an alphabet of size $q = 6$. The length of the data set is $n = 343$.

2. **Corn Exports.** We downloaded weekly data on corn exports from Jan. 4, 1990 to Oct. 21, 2010 from the US Department of Agriculture[3]. The data is in metric tons. We used 9 quantiles to convert the data into a sequence over an alphabet of size $q = 10$. The length of the data set is $n = 1103$.

3. **Energy Prices.** We downloaded hourly energy price data from Jan. 28, 2009 to Jan. 31, 2009 from the California Independent System Operator[4]. The raw data has units of dollars per megawatt-hour. We used 4 quantiles to convert the data into a sequence over an alphabet of size $q = 5$. The length of the data set is $n = 96$.

4. **Baseball Attendance.** We downloaded attendance data for all 2010 regular season home games at AT&T Park, home of the San Francisco Giants, from baseball-reference.com[5]. We used 4 quantiles to convert the data into a sequence over an alphabet of size $q = 5$. The length of the data set is $n = 81$.

5. **Unemployment.** We downloaded data on non-seasonally adjusted, federal initial unemployment claims from Nov. 17, 2007 to Sept. 25, 2010 from the US Department of Labor[6]. We used 7 quantiles to convert the data into a sequence over an alphabet of size $q = 8$. The length of the data set is $n = 150$.

Markov models for pollution [Dong et al., 2009], commodity exports [Mahadevaiah et al., 2005], energy prices [Rajaraman et al., 2001], and unemployment [Neftçi, 1984] have been considered previously by practitioners in the respective disciplines.

4.2 Discussion of Results

The results of applying four classifiers and two predictive models to our five data sets are summarized in Table A. Each of the first four columns contains the values of the margin function for a classifier applied to our five data sets. The final two columns report the average success rate $\mathcal{S}(k)$—defined in (16)—for $k = 0$ and $k = 1$.

Note that in all rows of the table, the sign of the margin function for the EB_1 and EB_5 classifiers is always negative. This means that, for all five data sets, both of these classifiers give a classification of $\hat{k} > 0$.

Note also that, in all but the first row, the success rate $\mathcal{S}(1)$ for a first-order model is approximately twice the success rate $\mathcal{S}(0)$ for a zeroth-order model. In the first row, we find

[3] `http://www.fas.usda.gov/export-sales/wkHistData.htm`

[4] `http://oasishis.caiso.com/`

[5] `http://www.baseball-reference.com/`

[6] `http://workforcesecurity.doleta.gov/unemploy/claims.asp/`

that $\mathcal{S}(1)$ is greater than $\mathcal{S}(0)$, but the ratio between the two values is not as large as in the last four rows.

Putting the above two paragraphs together, we see that the EB_1 and EB_5 classifiers both yield a classification that is in agreement with the model that gives better out-of-sample predictive accuracy. Except for the baseball attendance data, the EB^M classifier is in agreement with other EB classifiers.

At the same time, the BIC classifier yields a positive margin in all five cases, meaning that it classifies each data set as $\hat{k} = 0$, the model that gives worse out-of-sample predictive accuracy.

We hypothesize that the above results are due to a combination of two factors. First, the length N of all five data sets ranges from 81 to 1103, none of which are large enough to justify the law of large numbers step in the derivation of the BIC penalty term. Second, the number of symbols q ranges from 5 to 10. Our extensive tests in Section 3.2 have shown that the BIC classifier yields classifications very similar to those of the EB classifiers when $q = 2$. As q increases, the BIC classifier's error rate increases significantly, as shown in Table 3, while the EB_1 and EB_α classifiers' error rates decrease, as shown in Tables 1 and 2.

5 Conclusion

Overall, the results on both synthetic (Section 3.2) and real-world data sets (Section 4) lead to the conclusion that the EB classifier with an informative prior is to be preferred over the BIC classifier. On synthetic data sets, we found that an EB classifier with an informative prior gives higher percentage accuracy and larger areas under the ROC curves—see Hypothesis P in Section 3.2. On real-world data sets, we found that the EB classifications match the orders of the models that gave highest out-of-sample predictive accuracy. Across the whole study, we found that for sequences of a fixed length n, the gap between EB and BIC performance widens when the size of the alphabet q increases. We also found that using a prior that approximates the true distribution from which Markov transition matrices are drawn leads to results that are just as good as using the true distribution as the prior.

Besides the EB classifiers with informative priors, we derived in Section 2.6 the EB with model bias that, effectively, has a flat prior of $g \equiv 1$. Throughout all of our tests, we found that the EB with model bias classifier yields results very similar to those of the BIC classifier. When we examine carefully the derivation of the BIC (see Appendix), we find that the BIC also assumes a flat prior. We hypothesize that this is the main reason why EB with model bias and BIC classifiers lead to similar results, and also the main reason why EB classifiers with informative priors do so much better than the BIC classifier.

A Derivation of the BIC Penalty Term

We begin by defining our notation:

$$\begin{aligned}
\mathbf{y} &: \text{observed data } y_1, \dots, y_n \\
M_i &: \text{candidate model} \\
P(\mathbf{y}|M_i) &: \text{marginal likelihood of the model } M_i \text{ given the data} \\
\boldsymbol{\theta}_i &: \text{vector of parameters in the model } M_i \\
g_i(\boldsymbol{\theta}_i) &: \text{the prior density of the parameters } \boldsymbol{\theta}_i \\
f(\mathbf{y}|\boldsymbol{\theta}_i) &: \text{the density of the data given the parameters } \boldsymbol{\theta}_i \\
L(\boldsymbol{\theta}_i|\mathbf{y}) &: \text{the likelihood of } \mathbf{y} \text{ given the model } M_i \\
\hat{\boldsymbol{\theta}}_i &: \text{the MLE of } \boldsymbol{\theta}_i \text{ that maximizes } L(\boldsymbol{\theta}_i|\mathbf{y})
\end{aligned}$$

Laplace's Method. Laplace's method for approximating an integral is

$$\int_a^b e^{Mf(x)}\,dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)} \text{ as } M \to \infty.$$

For this approximation to hold, the function f should have one global maximum and it should decay rapidly to zero away from the maximum.

Let us now calculate

$$P(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i = \int \exp\Big(\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)\Big)d\boldsymbol{\theta}_i.$$

We expand $\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)$ about its posterior mode $\tilde{\boldsymbol{\theta}}_i$ where $f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)$ attains its maximum and, consequently, $\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)$ also attains its maximum. Thus, we approximate

$$\overbrace{\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)}^{Q} \approx \log\big(f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_i)g_i(\tilde{\boldsymbol{\theta}}_i)\big) + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}Q|_{\tilde{\boldsymbol{\theta}}_i} + \frac{1}{2}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T H_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i),$$

where $H_{\boldsymbol{\theta}_i}$ is a $|\boldsymbol{\theta}_i| \times |\boldsymbol{\theta}_i|$ matrix such that $H_{mn} = \partial^2 Q/\partial\theta_m\partial\theta_n|_{\tilde{\boldsymbol{\theta}}_i}$. Since Q attains its maximum at $\tilde{\boldsymbol{\theta}}_i$, the Hessian matrix $H_{\boldsymbol{\theta}_i}$ is negative definite. Let us denote $\tilde{H}_{\boldsymbol{\theta}_i} = -H_{\boldsymbol{\theta}_i}$, and then approximate $P(\mathbf{y}|M_i)$:

$$P(\mathbf{y}|M_i) \approx \int \exp\bigg\{Q|_{\tilde{\boldsymbol{\theta}}_i} + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}Q|_{\tilde{\boldsymbol{\theta}}_i} - \frac{1}{2}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \tilde{H}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)\bigg\}\,d\boldsymbol{\theta}_i.$$

Since Q attains its maximum at $\tilde{\boldsymbol{\theta}}_i$, we see that $\nabla_{\boldsymbol{\theta}_i} Q|_{\tilde{\boldsymbol{\theta}}_i} = 0$. Hence

$$\begin{aligned} P(\mathbf{y}|M_i) &\approx \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \tilde{H}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)\right\} d\boldsymbol{\theta}_i \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{-\frac{1}{2}X^T \tilde{H}_{\boldsymbol{\theta}_i} X\right\} dX. \end{aligned}$$

Since the matrix $\tilde{H}_{\boldsymbol{\theta}_i}$ is symmetric (by virtue of being the negative of the Hessian matrix), we can diagonalize it as $\tilde{H}_{\boldsymbol{\theta}_i} = S^T \Lambda S$. Let us make a substitution $X = S^T U$ to evaluate the integral above. The Jacobian matrix $J_{mn}(U) = \partial X_m / \partial U_n \Rightarrow J(U) = S^T$. Thus $\det J(U) = 1$, and

$$\begin{aligned} P(\mathbf{y}|M_i) &\approx \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{-\frac{1}{2}U^T \Lambda U\right\} (\det J(U)) dU \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{-\frac{1}{2}\sum_{j=1}^{|\boldsymbol{\theta_i}|} \lambda_j U_j^2\right\} dU \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \prod_{j=1}^{|\boldsymbol{\theta_i}|} \sqrt{\frac{2\pi}{\lambda_j}} \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \frac{(2\pi)^{|\boldsymbol{\theta_i}|/2}}{\prod_j^{|\boldsymbol{\theta_i}|} \lambda_j^{1/2}} \\ &= f(y|\tilde{\boldsymbol{\theta}}_i) g_i(\tilde{\boldsymbol{\theta}}_i) \frac{2\pi^{|\boldsymbol{\theta_i}|/2}}{|\tilde{H}_{\boldsymbol{\theta}_i}|^{1/2}}, \end{aligned} \tag{17}$$

where λ_j is the j-th eigenvalue of the matrix $\tilde{H}_{\boldsymbol{\theta}_i}$. Taking log of (17), we get

$$2 \log P(\mathbf{y}|M_i) = 2 \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_i) + 2 \log g_i(\tilde{\boldsymbol{\theta}}_i) + |\boldsymbol{\theta}_i| \log(2\pi) + \log |\tilde{H}_{\boldsymbol{\theta}_i}^{-1}|. \tag{18}$$

Flat Prior and the Weak Law of Large Numbers. Since the observed data $\mathbf{y}$ is given, $f(\mathbf{y}|\boldsymbol{\theta}_i)$ is the likelihood $L(\boldsymbol{\theta}_i|\mathbf{y})$ and L attains its maximum at the maximum likelihood estimate $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$. If we set $g_i(\boldsymbol{\theta}_i) = 1$, an uninformative, flat prior, then each element in the matrix, $\tilde{H}_{\boldsymbol{\theta}_i}$ can be expressed as

$$\tilde{H}_{mn} = -\frac{\partial^2 \log L(\boldsymbol{\theta}_i|\mathbf{y})}{\partial\theta_m \partial\theta_n}\Big|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i}.$$

The matrix $\tilde{H}_{\boldsymbol{\theta}_i}$ is the observed Fisher information matrix. We can further represent every entry in the observed Fisher information matrix as

$$\begin{aligned}\tilde{H}_{mn} &= -\frac{\partial^2 \log(\prod_{j=1}^n L(\boldsymbol{\theta}_i|y_j))}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -\frac{\partial^2 \sum_{j=1}^n \log L(\boldsymbol{\theta}_i|y_j)}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -\frac{\partial^2 \left(\frac{1}{n}\sum_{j=1}^n n\log L(\boldsymbol{\theta}_i|y_j)\right)}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i}.\end{aligned}$$

At this point, we assume that the observed data $y_1, \ldots, y_n$ is IID and that n is large. This allows us to invoke the weak law of large numbers on the random variable $X_j = n\log L(\boldsymbol{\theta}_i|y_j)$. We obtain

$$\frac{1}{n}\sum_{j=1}^n n\log L(\boldsymbol{\theta}_i|y_j) \xrightarrow{\text{P}} E[n\log L(\boldsymbol{\theta}_i|y_j)]. \tag{19}$$

Using (19), every element in the observed Fisher information matrix is

$$\begin{aligned}\tilde{H}_{mn} &= -\frac{\partial^2 E[n\log L(\boldsymbol{\theta}_i|y_j)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -n\frac{\partial^2 E[\log L(\boldsymbol{\theta}_i|y_j)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -n\frac{\partial^2 E[\log L(\boldsymbol{\theta}_i|y_1)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= nI_{mn},\end{aligned}$$

so

$$|\tilde{H}_{\boldsymbol{\theta}_i}| = n^{|\boldsymbol{\theta}_i|}|I_{\boldsymbol{\theta}_i}|, \tag{20}$$

where $I_{\boldsymbol{\theta}_i}$ is the Fisher information matrix for a single data point y_1. Plugging the result from (20) into (18), we obtain

$$2\log P(\mathbf{y}|M_i) = 2\log L(\hat{\boldsymbol{\theta}}_i|\mathbf{y}) + 2\log g_i(\tilde{\boldsymbol{\theta}}_i) + |\boldsymbol{\theta}_i|\log(2\pi) - |\boldsymbol{\theta}_i|\log n - \log|I_{\boldsymbol{\theta}_i}|. \tag{21}$$

For large n, keeping the terms involving n and ignoring the rest, we find

$$\log P(\mathbf{y}|M_i) = \log L(\hat{\boldsymbol{\theta}}_i|\mathbf{y}) - \frac{|\boldsymbol{\theta}_i|}{2}\log n. \tag{22}$$

The right-hand side of (22) is the BIC estimate for a model M_i.

References

H. S. Bhat and N. Kumar. Markov tree options pricing. In *Proceedings of the Fourth SIAM Conference on Mathematics for Industry (MI09)*, San Francisco, CA, 2010.

P. Billingsley. *Statistical Inference for Markov Processes.* University of Chicago Press, Chicago, 1961.

I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Transactions on Information Theory*, 48(6):1616–1628, 2002.

I. Csiszár and P. C. Shields. The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28:1601–1619, 2000.

D. Dalevi, D. Dubhashi, and M. Hermansson. A new order estimator for fixed and variable length Markov models with applications to DNA sequence similarity. *Statistical Applications in Genetics and Molecular Biology*, 5:1–24, 2006.

M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, and D. Kenski. PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications*, 36(5):9046–9055, 2009.

C. C. Y. Dorea and J. S. Lopes. Convergence rates for Markov chain order estimates using EDC criterion. *Bulletin of the Brazilian Mathematical Society, New Series*, 37(4):561–570, 2006.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

L. Finesso, C. C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Transactions on Information Theory*, 42(5):1488–1497, 1996.

J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Methods.* Springer-Verlag, New York, 2006.

P. Guttorp. *Stochastic Modeling of Scientific Data.* Stochastic Modeling. Chapman and Hall, London, 1995.

R. W. Katz. On some criteria for estimating the order of a Markov chain. *Technometrics*, 23 (3):243–249, 1981.

G. S. Mahadevaiah, P. C. Ravi, and P. G. Chengappa. Stability analysis of raw cotton export markets of India—Markov chain approach. *Agricultural Economics Research Review*, 18: 253–259, 2005.

G. Morvai and B. Weiss. Order estimation of Markov chains. *IEEE Transactions on Information Theory*, 51(4):1496–1497, 2005.

S. H. Neftçi. Are economic time series asymmetric over the business cycle? *The Journal of Political Economy*, 92(2):307–328, 1984.

Y. Peres and P. C. Shields. Two new Markov order estimators. Preprint, 2005.

R. Rajaraman, L. Kirsch, F. L. Alvarado, and C. Clark. Optimal self-commitment under uncertain energy and reserve prices. In B. F. Hobbs, M. H. Rothkopf, R. P. O'Neill, and H.-P. Chao, editors, *The Next Generation of Electric Power Unit Commitment Models*. Kluwer, Boston, April 2001.

C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E*, 76: 011106–1–14, 2007.

L. A. Wasserman. *All of Statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2003.

L. C. Zhao, C. C. Y. Dorea, and C. R. Gonçalves. On determination of the order of a Markov chain. *Statistical Inference for Stochastic Processes*, 4:273–282, 2001.

		Sequence length n									
		50	100	150	200	250	300	350	400	450	500
Alphabet size q	2	77.38	83.67	86.48	88.12	89.58	90.28	90.93	91.26	91.95	92.57
	3	87.35	93.32	96.06	97.46	98.25	98.74	99.05	99.28	99.36	99.40
	4	92.07	96.73	98.18	98.93	99.31	99.48	99.64	99.75	99.83	99.89
	5	93.80	97.88	98.97	99.46	99.66	99.75	99.84	99.86	99.91	99.92
	6	94.80	98.54	99.44	99.70	99.82	99.87	99.88	99.94	99.93	99.97
	7	95.45	98.87	99.55	99.79	99.85	99.95	99.97	99.96	99.96	99.98
	8	96.08	99.18	99.69	99.86	99.95	99.95	99.97	99.98	99.99	100.00
	9	96.04	99.18	99.76	99.92	99.94	99.94	99.98	99.98	99.99	99.99
	10	96.24	99.28	99.79	99.94	99.97	99.98	99.99	99.99	99.99	100.00

Table 1: Percentage of correctly classified instances using the EB_1 classifier. Instances generated from random Markov chains with uniformly distributed transition matrices.

		Sequence length n									
		50	100	150	200	250	300	350	400	450	500
Alphabet size q	2	77.20	83.51	86.31	88.06	89.49	90.23	90.91	91.27	91.91	92.53
	3	87.44	93.26	96.08	97.44	98.19	98.68	99.04	99.24	99.36	99.37
	4	92.07	96.51	98.15	98.89	99.33	99.48	99.64	99.72	99.84	99.92
	5	93.53	97.84	98.90	99.44	99.66	99.75	99.84	99.86	99.92	99.92
	6	94.71	98.51	99.41	99.69	99.82	99.88	99.88	99.94	99.94	99.97
	7	95.13	98.84	99.58	99.79	99.86	99.95	99.96	99.96	99.95	99.99
	8	96.00	99.17	99.63	99.84	99.94	99.95	99.97	99.98	100.00	100.00
	9	95.71	99.11	99.73	99.91	99.96	99.95	99.97	99.98	100.00	100.00
	10	96.09	99.11	99.77	99.94	99.98	99.98	99.98	99.99	100.00	100.00

Table 2: Percentage of correctly classified instances using the EB_α classifier with $\alpha \sim U(0.5, 1.5)$. Instances generated as in Table 1.

		Sequence length n									
		50	100	150	200	250	300	350	400	450	500
Alphabet size q	2	76.62	82.95	85.69	87.36	88.81	89.78	90.50	90.88	91.72	92.12
	3	65.12	78.37	87.73	92.75	95.47	97.23	97.92	98.26	98.69	98.96
	4	47.82	60.26	67.85	72.79	77.51	82.11	86.36	91.00	94.31	96.40
	5	35.31	46.14	56.03	61.63	65.01	67.25	69.14	70.81	72.76	74.75
	6	33.33	35.07	42.69	52.02	58.47	62.35	64.39	65.62	66.29	66.81
	7	33.33	33.34	34.18	38.32	45.20	53.00	58.25	62.05	64.23	65.44
	8	33.33	33.33	33.33	33.51	34.86	38.86	44.85	51.43	56.85	60.60
	9	33.33	33.33	33.33	33.33	33.35	33.58	34.67	37.45	42.00	47.68
	10	33.33	33.33	33.33	33.33	33.33	33.33	33.36	33.48	34.01	35.57

Table 3: Percentage of correctly classified instances using the BIC classifier. Instances generated as in Table 1.

		Sequence length n									
		50	100	150	200	250	300	350	400	450	500
Alphabet size q	2	77.38	83.67	86.48	88.12	89.58	90.28	90.93	91.26	91.95	92.57
	3	76.93	91.52	95.63	97.29	98.10	98.64	98.98	99.08	99.20	99.37
	4	53.98	69.69	79.73	89.80	96.11	98.33	99.25	99.64	99.80	99.88
	5	35.78	53.46	63.61	68.08	71.13	74.47	77.92	83.45	89.39	94.16
	6	33.33	37.25	51.21	60.58	64.52	66.19	67.10	67.85	68.94	69.99
	7	33.33	33.35	35.99	46.04	56.25	62.14	64.56	65.91	66.35	66.56
	8	33.33	33.33	33.35	34.17	39.62	48.78	56.99	61.97	64.61	65.78
	9	33.33	33.33	33.33	33.33	33.48	35.23	40.21	48.17	55.57	60.76
	10	33.33	33.33	33.33	33.33	33.33	33.34	33.56	34.98	38.64	44.78

Table 4: Percentage of correctly classified instances using the EB^M classifier. Instances generated as in Table 1.

			Sequence length n									
			50	100	150	200	250	300	350	400	450	500
Alphabet size q	2	EB_1	70	84	89	91	93	93	94	94	95	95
		$\mathrm{EB}_{0.5}$	77	89	92	93	94	95	95	95	96	96
		EB_{10}	73	79	82	84	86	88	89	90	91	91
		EB^M	70	84	89	91	93	93	94	94	95	95
		BIC	86	91	93	94	95	95	95	96	96	96
	4	EB_1	84	92	96	97	98	98	99	99	99	99
		$\mathrm{EB}_{0.5}$	81	92	96	98	98	99	99	99	99	99
		EB_{10}	94	96	97	98	98	98	99	99	99	99
		EB^M	99	99	99	100	99	99	100	100	100	100
		BIC	99	100	100	100	100	100	100	100	100	100
	6	EB_1	90	96	98	99	99	99	99	99	99	99
		$\mathrm{EB}_{0.5}$	86	95	97	98	99	99	99	99	99	99
		EB_{10}	97	99	99	99	99	99	99	99	99	99
		EB^M	100	100	100	100	100	100	100	100	100	100
		BIC	100	100	100	100	100	100	100	100	100	100

Table 5: Percentage of correctly classified instances using five different classifiers. Instances considered here are always IID (i.e., $k = 0$) sequences generated from probabilities chosen from the uniform distribution on $\boldsymbol{D}$.

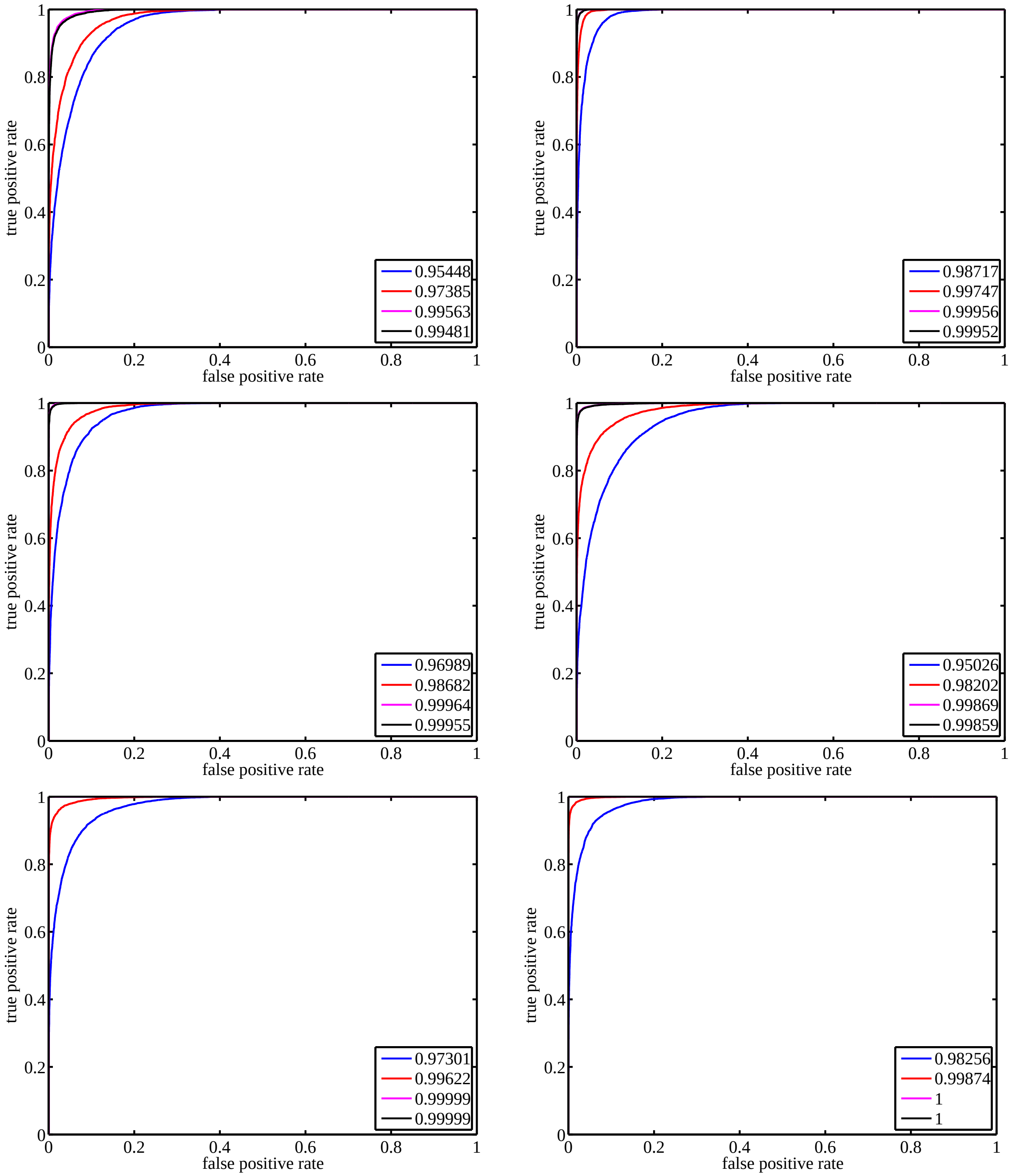

Figure 1: Here we report tests using random sequences of length n from random Markov chains over q symbols with transition probabilities drawn uniformly. From left to right, top to bottom, we plot ROC curves for $(n, q) = (100, 4)$, $(200, 4)$, $(150, 5)$, $(100, 6)$, $(200, 9)$, $(400, 10)$. Areas under ROC curves are reported in the legend. Magenta, black, red, and blue correspond to the EB_1, EB_α, EB^M, and BIC classifiers, respectively. See Section 3.2.1 for further details.

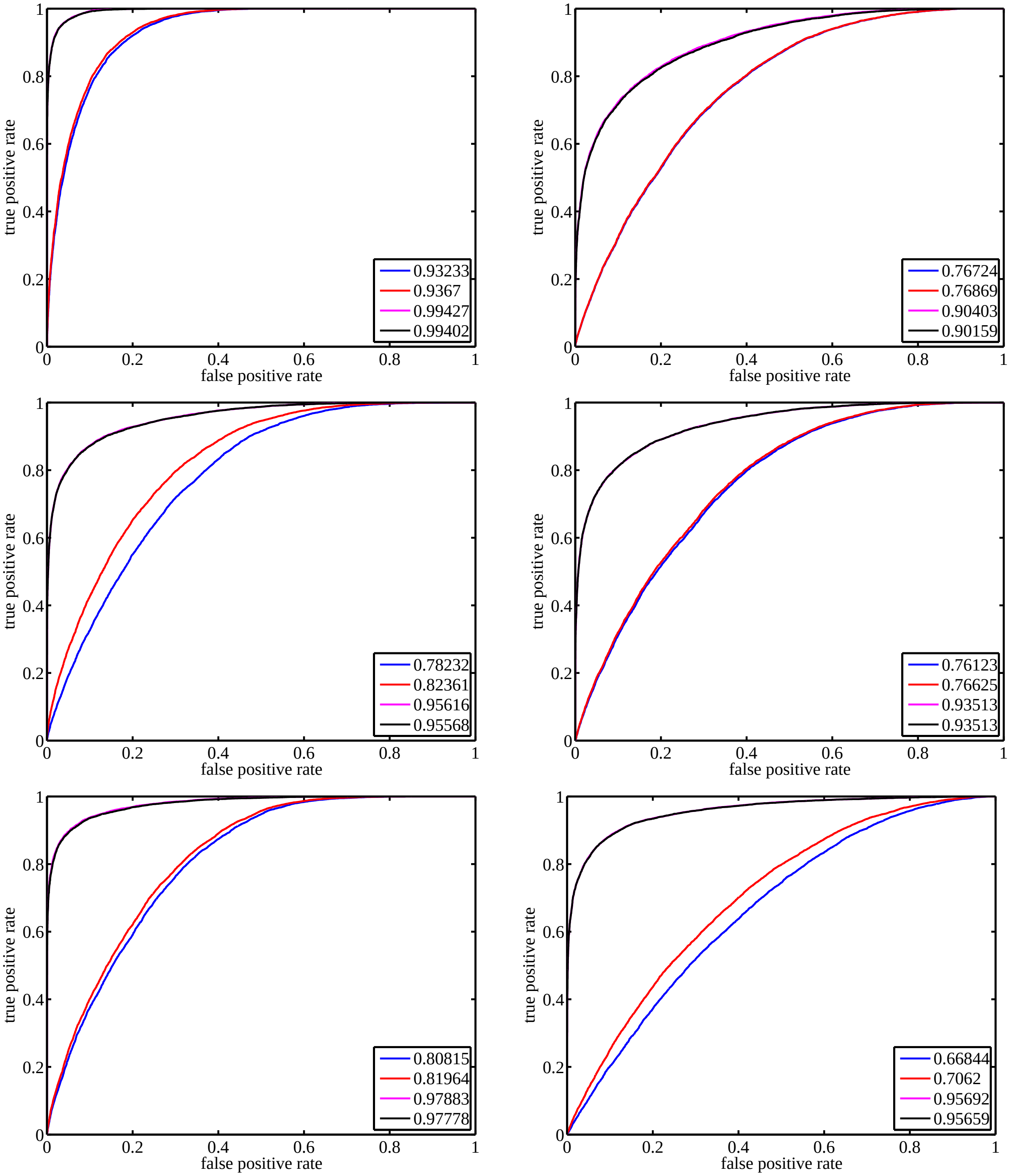

Figure 2: Here we report tests using random sequences of length n from random Markov chains over q symbols with transition probabilities drawn from a Dirichlet $(\alpha, \ldots, \alpha)$ distribution. From left to right, top to bottom, we plot ROC curves for $(n, q, \alpha) = (400, 4, 5)$, $(400, 4, 20)$, $(150, 5, 5)$, $(350, 5, 15)$, $(400, 6, 10)$, $(500, 10, 20)$. Areas under ROC curves are reported in the legend. Magenta, black, red, and blue correspond to the EB_α, $\text{EB}_{\widetilde{\alpha}}$, EB^M, and BIC classifiers, respectively. See Section 3.2.2 for further details.

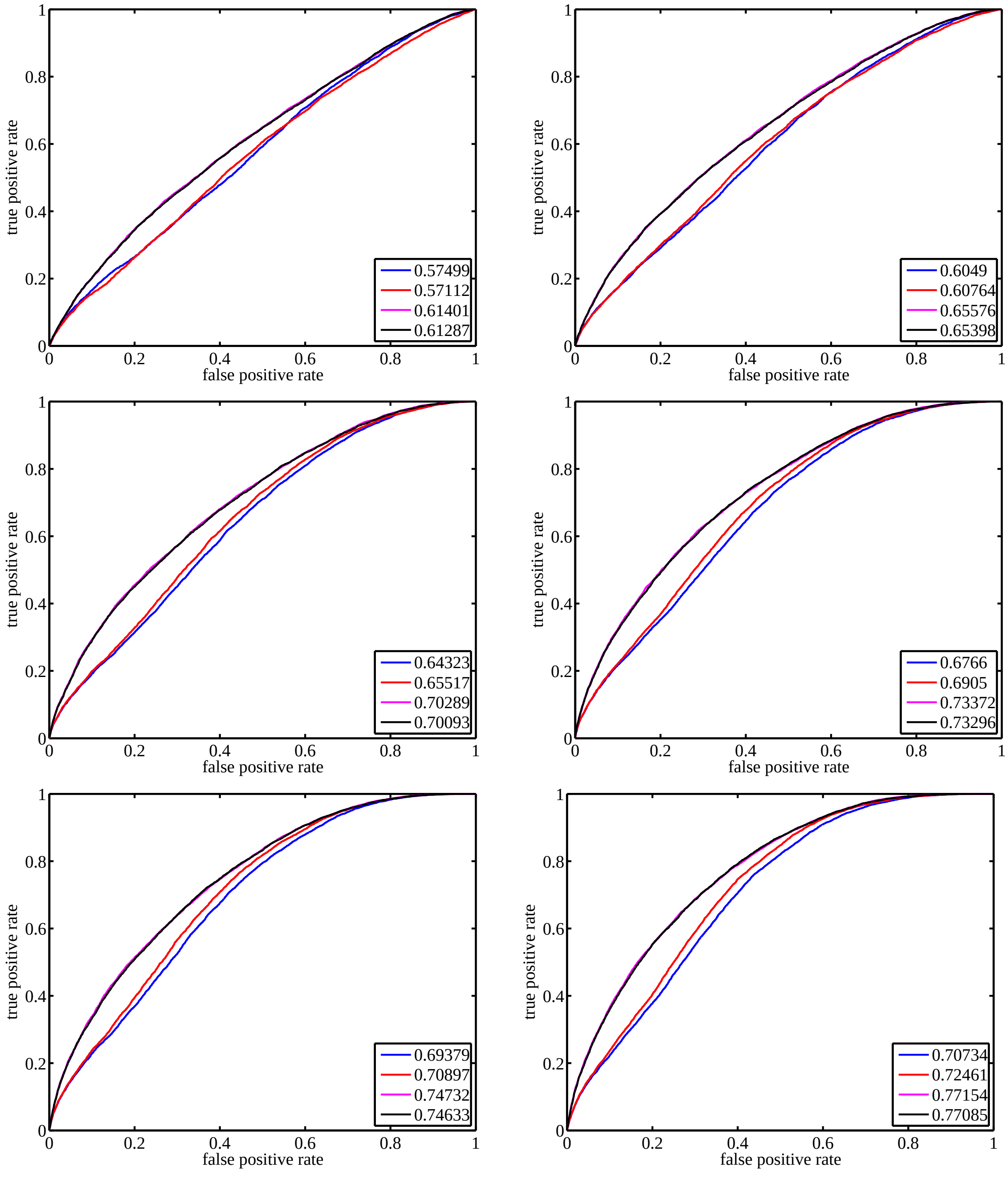

Figure 3: Here we report tests using randomly generated sequences of length n from randomly generated Markov chains over 2 symbols with transition probabilities drawn from the uniform distribution on $[0.4, 0.6]$. From left to right, top to bottom, we plot ROC curves for $n = 100$, 150, 250, 350, 400, 500. Areas under ROC curves are reported in the legend. Magenta, black, red, and blue correspond to the $\text{EB}_{U(0.4,0.6)}$, EB_{37}, EB^{M}, and BIC classifiers, respectively. See Section 3.2.3 for further details.

	EB_1	EB_5	EB^M	BIC	$\mathcal{S}(0)$	$\mathcal{S}(1)$
Air Quality Index	-27.5	-47.6	-23.7	3.9	0.7638	0.8192
Corn Exports	-132.7	-127.1	-17.4	36.6	0.0907	0.1976
Energy Prices	-13.1	-9.3	-0.4	5.4	0.1667	0.3542
Baseball Attendance	-10.1	-7.6	2.5	9.7	0.1481	0.3951
Unemployment Claims	-67.4	-31.0	-7.7	0.4	0.3067	0.5667

Table 6: Margin function values of four different classifiers applied on five different data sets are reported under their respective classifier columns. The column $\mathcal{S}(0)$ and $\mathcal{S}(1)$ report the average out-of-sample success rates of zeroth- and first-order models on these data sets. See Section 4 for further details.