A simple hidden Markov model for Bayesian modeling with time dependent data

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Stephen Vardeman
Department of Statistics
Iowa State University
Ames, Iowa 50011

August 1997

*Research supported in part by NSF Grant SES 9201718

A simple hidden Markov model for Bayesian modeling with time dependent data

SUMMARY

Consider a set of real valued realizations of an observable quantity collected over time. We propose a simple hidden Markow model for these realizations in which the the predicted distribution of the next future observation given the past is easily computed. The hidden or unobservable set of parameters is assumed to have a Markov structure of a special type. The model is quite flexible and can be used to incorporate different types of prior information in straightforward and sensible ways.

Key Words: Hidden Markov model, Bayesian modeling, prediction, time series.

1 Introduction

A hidden Markov model arises in the following manor. A hidden or unobservable sequence of "parameters" $(\theta_1, \theta_2, \ldots,)$ is generated according to some Markov process. Then conditioned on these values we observe $(Y_1, Y_2, \ldots)$ where it is assumed that the Y_i's are independent given the θ_i's and the distribution of Y_t depends on the θ_i's only through θ_t.

Recently hidden Markov chains have been used to model dependent observations in a variety of fields. See for example Elliot et al. (1995). Statistical inference for such models is more complicated than for the standard statistical models and so there has been less written about them in the statistical literature. Leroux and Puterman (1992) and Rydén (1994) discuss parameter estimation for such models. Formally hidden Markov models are very similar to state space models. West and Harrison (1997) is a good discussion of these dynamic models from a Bayesian prospective.

The Kalman filter, is a hidden Markov model of a particularly nice type. One of its most convenient features is that at each stage it can be computed recursively. This computational simplicity comes at the cost of a lack flexibility that limits the type of data that can be modeled. The more recent hidden Markov models are more versatile but require more computing for their study. Two recent examples of state space models in the statistical literature which relax the assumptions of the Kalman filter are Carter and Kohn (1994) and Carlin et al. (1992). These models however require Markov chain Monte Carlo methods to study. Such models yield useful and powerful techniques but it would still be interest to find some hidden Markov models

which retain the the ease of computation of the Kalman filter but can be applied in a variety of situations. Meinhold and Singpurwalla (1983) emphasize that the Kalman filter can be easily understood as a problem in Bayesian modeling. The family of models proposed here are given in that spirit. It is a large class which will allow for the incorporation of prior information in many different situations. Like the Kalman filter the models will have the property that they can be updated recursively. At each stage, the conditional distribution of the current parameter given the data will be a mixture of known probability distributions, where the mixture weights just depend on the value of the mixture weights for the previous stage. The predictive distribution for the observation for the next observation has a similar simple form.

In section 2 the model is formally developed. In section 3 we show how the model can be updated recursively at each stage. In section 4 we show how the modeling can be done in practice for assorted types of prior information. In section 5 we consider several examples. In section 6 we give some concluding remarks.

2 The Model

Let $\mathbf{Y}_t = (Y_1, \ldots, Y_t)$ be the vector of observable real valued random variables for the successive times $1, \ldots, t$ where $t > 1$ is some fixed positive integer. Let $(\theta_1, \ldots, \theta_t)$ be the corresponding vector of parameters or the unknown and unobservable successive states of nature. Each θ_s is assumed to belong to Θ some interval of real numbers. We now give the underlying assumptions

for our model of the observables and the parameters.

We begin with our assumptions about the parameters. Let $r > 1$ be a fixed positive integer and $\pi_1, \ldots, \pi_r$ be a fixed collection of r known probability density functions on Θ. Let $\mathbf{w} = (w_1, \ldots, w_r)$ be a vector of r nonnegative weights which sum to one. For $s = 2, \ldots, t$ let $W_1^s, \ldots, W_r^s$ be nonnegative functions define on Θ such that for each s and each $\theta \in \Theta$, $\sum_{j=1}^r W_j^s(\theta) = 1$. In what follows $p(\cdot|\cdot)$ will denote a generic conditional probability density function. We now define the joint distribution of the parameters. The marginal density of θ_1 is

$$\sum_{j=1}^{r} w_j \pi_j(\theta_1) \tag{1}$$

while for $s = 2, \ldots, t$ the conditional density of θ_s given $\theta_1, \ldots, \theta_{s-1}$ is

$$p(\theta_s|\theta_1, \ldots, \theta_{s-1}) = p(\theta_s|\theta_{s-1}) = \sum_{j=1}^{r} W_j^s(\theta_{s-1}) \pi_j(\theta_s) \tag{2}$$

We next give our assumptions about the observables. For a fixed value of θ let $f(\cdot|\theta)$ be a known probability function or density function for a real valued random variable. This denotes the conditional distribution of a typical observable given a value of θ. We assume that given θ_1 the distribution of Y_1 is given by $f(\cdot|\theta_1)$, while for $s = 2, \ldots, t$ the conditional density of Y_s given $\theta_1, \ldots, \theta_{s-1}, y_1, \ldots, y_{s-1}$ is

$$p(y_s|\theta_1, \ldots, \theta_{s-1}, y_1, \ldots, y_{t-1}) = p(y_s|\theta_s) = f(y_s|\theta_s) \tag{3}$$

In the following $\mathbf{Y}_s$ will denote the observables $(Y_1, \ldots, Y_s)$ and $\mathbf{y}_s = (y_1, \ldots, y_s)$ an actual realization of this random vector.

Note that these assumptions are quite similar to those underlying the Kalman filter. We have a sequence of unobservable parameters whose distribution at each stage depends only on the value of the parameter at the previous stage. The distribution of the observable at each stage depends only on the value of the parameter at that stage. On the other hand we are not restricted to assuming that our distributions are normal, as is the case with the Kalman filter. At first glance, the form of $p(\theta_s|\theta_{s-1})$, a mixture of the π_j's with weights $W_j^s(\theta_{s-1})$'s may seem surprising. However we shall see that the flexibility inherent in this structure allows us to model many different situations, without losing the recursive property of the Kalman filter. This will be considered in more detail in section 4. Now we will consider a simple example that should help to clarify the role of the W_j^s's and the π_j's.

Consider a process which is producing large batches of some organism over time. As each batch is produced we take a small random sample, say of size n, from the batch and observe the number of "defectives". If Y_s denotes the number of defectives in the sample from batch s and θ_s the true number of defectives in the batch then we may assume that the conditional distribution of Y_s given θ_s is Binomial(n, θ_s). This much is very standard. Next we consider modeling the behavior of the θ_s's.

Suppose as the growing process begins it is assumed to be in control. When it is in control about 20% of the organisms produced are defective. As long as the process remains in control the θ_s's remain in the neighborhood of .20, although they can vary slightly from batch to batch. However it is possible that that the process can go out of control by producing batches where the probability of a defective is significantly larger than .2. Suppose that the

process should be stopped once the percentage of defectives increases above 30%. Here is one possible model for the θ_s's. Let π_1 be a Beta(α_1, β_1) probability density function and π_2 be a Beta(α_2, β_2) density function. Assume the mean of π_1 is .20 and the mean of π_2 is somewhere to the right of .30. We will let the initial weights $(w_1, w_2) = (.95, .05)$ and let $W_1 = W_1^s$ and $W_2 = W_2^s$ for all s, i.e. the weight functions used to define $p(\theta_s|\theta_{s-1})$ will not depend on s. It remains to define $W_1(\theta) = 1 - W_2(\theta)$. One possible choice is to take $W_1(\theta) = 1 - \theta/\lambda$ where $\lambda \geq 1$ is some fixed real number. The size of λ reflects how likely the process is to jump out of control. As λ increases it becomes more probable that the process remains in control. Note that the probability of remaining in control from batch to batch also depends on the variances of the π_j's.

Although it is somewhat naive, the above model for the θ_s's does seem to catch the most important parts of the scenario. It has the advantage of using the available prior information in straight forward ways to define a complicated joint prior distribution. As we shall see, it is easy to find $p(\theta_s|y_1, \ldots, y_s)$. So at any stage one can compute the posterior probability that θ_s exceeds .30, given the observed number of defectives of all stages up to the present. Because of the recursive nature of the computations this can essentially be done in real time. The model assumes that with high probability the next value of θ will be close to the present value. It does not make any explicit prediction about when a jump in the percentage of defectives per batch will occur. But if such a jump does occur it will recognize it immediately, with high probability.

In section 4 we will consider several examples and show how the model

can be adapted to different kinds of prior information. In particular, we will give models that can be used when it is know that the values of underlying parameters can change dramatically, in contrast to the assumed steady state model described above. It will also be seen that the results are fairly robust against the choice of the π_j's and the W_j^s's. In the next section we will show how the necessary updating can be done recursively at each step.

3 The Updating Step

In this section we will show how the model can be updated recursively at each step. This means that it has much of the computational convenience that is also present in the Kalman filter.

We will assume that the model, as described in the previous section, is in place. We will need some more notation. If the distribution of Y given θ is given by $f(\cdot|\theta)$ and the distribution of θ is given by π_j, we let $p_j(\theta|y)$ be the conditional density of θ given $Y = y$ and $p_j(y)$ be the marginal density of Y. In other words, this is just the conditional and marginal density when a stage is considered by itself with only π_j as the prior.

We will see in what follows that at stage $t-1$ the posterior distribution of θ_{t-1} given $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1} = (y_1, \ldots, y_{t-1})$ can be represented as

$$p(\theta_{t-1}|\mathbf{y}_{t-1}) \quad = \quad \sum_{j=1}^{r} wt_j^{t-1} p_j(\theta_{j-1}|y_{t-1}) \tag{4}$$

Here the wt_j^{t-1}'s are non-negative weights which sum to one. They depend on the corresponding wt_j^{t-2}'s and y_{t-1}. We will now show the explicit relationship by demonstrating how the wt_j^{t-1}'s and y_t can be used to obtain the

wt_j^t's. Before doing that we want to make one observation about equation 4. Note that equation 4 is very similar to equation 2. In both the density for θ_{t-1} is a mixture of densities. However in the later $\pi_j(\cdot)$ has been replaced by $p_j(\cdot|y_{t-1})$ and the prior weights by their data updated versions.

We begin by finding an expression for the joint density of θ_t and $\mathbf{Y}_t$.

$$\begin{aligned} p(\theta_t, \mathbf{y}_t) &= \int p(\mathbf{y}_{t-1}) p(\theta_{t-1}|\mathbf{y}_{t-1}) p(\theta_t|\theta_{t-1}, \mathbf{y}_{t-1}) p(y_t|\theta_{t-1}, \theta_t, \mathbf{y}_{t-1})\, d\nu(\theta_{t-1}) \\ &= p(\mathbf{y}_{t-1}) f(y_t|\theta_t) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|\mathbf{y}_{t-1})\, d\nu(\theta_{t-1}) \\ &= p(\mathbf{y}_{t-1}) f(y_t|\theta_t) \sum_{j=1}^{r} E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} \pi_j(\theta_t) \end{aligned}$$

From this it follows that

$$p(\mathbf{y}_t) = p(\mathbf{y}_{t-1}) \sum_{j=1}^{r} E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_j(y_t)$$

and so we have that

$$p(\theta_t|\mathbf{y}_t) = \frac{\sum_{j=1}^{r} E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_j(y_t) p_j(\theta_t|y_t)}{\sum_{j=1}^{r} E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_j(y_t)} \tag{5}$$

If we let

$$\begin{aligned} wt_j^t &= \frac{E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_j(y_t)}{\sum_{i=1}^{r} E\{W_i^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_i(y_t)} \\ &= \frac{E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\} p_j(y_t)}{p(y_t|\mathbf{y}_{t-1})} \end{aligned} \tag{6}$$

then equation 5 is of the form of equation 4 ,that is,

$$p(\theta_t|\mathbf{y}_t) = \sum_{j=1}^{r} wt_j^t p_j(\theta_j|y_t) \tag{7}$$

These equations make the updating process perfectly clear. If the wt_j^{t-1}'s are the weights at stage $t-1$, then the updated weights, the wt_j^t's, which

are used after we observe $Y_t = y_t$, are found as follows. For each j we find $E\{W_j^t(\theta_{t-1})|\mathbf{y}_{t-1}\}$. This is straightforward using the wt_j^{t-1}'s and equation 4. Recall from equation 2 that $W_j^t(\theta_{t-1})$ is just the weight assigned to prior π_j in the mixture which generates θ_t given θ_{t-1}. Hence this expected value is just our best guess for this weight given $\mathbf{y}_{t-1}$. We next compute the marginal probability of y_t under the assumption that θ_t was just generated from the prior π_j and conditional on this value of θ_t, y_t was generated from $f(\cdot|\theta_t)$. (We call this model the *single prior model*.) We then form the product of this expectation and marginal probability, for each j, and normalize these r products to sum to one, to get our new weights in $p(\theta_t|\mathbf{y}_t)$. In this mixture, for each j, the corresponding new weight is applied to the single prior model for prior π_j.

In summary, $p(\theta_t|\mathbf{y}_t)$ is just a mixture of the posteriors from all the single prior models, which are computed using y_t. The weights in the mixture are just the updated weights which depend only on the weights from the previous stage, the W_j^t and y_t. It is easy to see from the previous equations that the density of the predictive distribution of Y_{t+1} given $\mathbf{y}_t$ is given by

$$p(y_{t+1}|\mathbf{y}_t) = \sum_{j=1}^{r} E\{W_j^{t+1}(\theta_t)|\mathbf{y}_t\} p_j(y_{t+1}) \tag{8}$$

In principle, this can be easily computed from $p(\theta_t|\mathbf{y}_t)$. It is also a mixture, in this case a mixture of the marginals of the future observation under all the single prior models. Hence if the model, i.e. $f(\cdot|\cdot)$, the π_j's, and the W_j^s's, is such that all the single prior model computations can be done conveniently, then at each stage $p(\theta_t|\mathbf{y}_t)$ and $p(y_{t+1}|\mathbf{y}_t)$ can also be found easily. It is exactly these quantities that are usually of most interest at each stage of the

experiment.

In what follows we will assume that $f(\cdot|\cdot)$ belongs to the one parameter exponential family. In addition we will assume that the π_j's belong to the appropriate family of conjugate priors. This means that all the necessary calculations can easily be implemented using a standard statistical package. Despite their computational simplicity we will see in the next section that these models are still flexible enough to realistically model quite different types of prior information.

4 Selecting a Model

When determining a model one must select $f(\cdot|\cdot)$, the π_j's, and the W_j^s's. Since $f(\cdot|\cdot)$ represents the conditional distribution of an observable given a parameter, it is most convenient to follow standard statistical practice and select one of the usual distributions. In the same spirit the π_j's will be selected from the usual conjugate families so that the computations within all the single family models will be straight forward. It remains to select the W_j^s's. As we shall see the choice of these W_j^s's and the assumed Markov structure of the parameters, given in equation 2, allow for the representation of a wide variety of types of prior information.

For definiteness suppose that given θ, $f(\cdot|\theta)$ is a normal density with mean θ and variance σ^2. Let $r \geq 2$ be fixed and let π_j be a normal density with mean μ_j and variance τ^2, where $\mu_1 < \cdots < \mu_r$. In the discussion that follows we assume that the μ_j's are equally spaced, although this is technically unnecessary. Let the *cut points* be a vector of real numbers of

length $r-1$ whose coordinates are strictly increasing. We denote this vector by $\mathbf{u} = (u_1, \ldots, u_{r-1})$. Let $u_0 = -\infty$ and $u_r = \infty$. Now given a vector of cut points $\mathbf{u}$ and $h \in [0,1]$ we can define a family of weight functions $W_j(\theta)$ as follows,

$$W_j(\theta) = \begin{cases} (1-h)/r & \text{for } \theta \notin (u_{j-1}, u_j) \\ h + (1-h)/r & \text{for } \theta \in (u_{j-1}, u_j) \end{cases} \tag{9}$$

Suppose in the definition of $p(\theta_s|\theta_{s-1})$ for some integer s, W_j^s is taken to be the W_j defined above for some vector of cut points $\mathbf{u}$ and $h \in [0,1]$. Note that if $h = 0$ the W_j's are all a constant function of θ and are all identically equal to $1/r$. In this case the model for the parameters assumes that they are just a random sample from the distribution which is the equal weight mixture of the π_j's. If $h = 1$ then W_j is one on the interval (u_{j-1}, u_j) and zero else where. In this case θ_s is generated from a single prior, the one that is associated with the cut point interval defined by $\mathbf{u}$ which contained θ_{s-1}. Hence by letting h vary and selecting the vector of cut points judiciously we can model a variety of structures for the parameters. We will next consider two different scenarios. In each case we will assume the normal distributions discussed above, but this is only for convenience.

First, suppose we know that the parameters belong to some known interval, say (a, b) and with high probability will change little from stage to stage. A big change is possible, although not likely, and the direction and time of such a change is unknown. However if such a change does occur, succeeding values of the parameter will tend to be nearly constant until another large change. Here we are assuming that our process will only change significantly when a random shock affects the system. We are not attempting to model

the random shocks but want a structure that allows us to check after the fact that a large shock has occurred. We first select the r equal spaced means, the μ_j's, for the π_j's. We will take μ_1 close to a and μ_r close to b. We will also choose an appropriate value of the variance τ^2. As will shall see, for most purposes our inferences are quite robust against these choices. We will take the vector of cut points to be the midpoints of the μ_j's. For each s we take W_j^s to be the W_j defined in equation 9 with $h = 1$. Under such a model there is high probability that successive parameter values will be close together. The actual size of this probability and how close they tend to be will depend on the actual choices of the μ_j's and τ^2. This model will never predict a large jump, but once such a jump has occurred and we have observed the corresponding y value it will give a high posterior probability to the event that such a jump did indeed occur. It will also predict succeeding observations well until another big change occurs. This model is similar in spirit to the beta-binomial model discussed in section 2. These type of models are easily defined for all the standard families of distributions belonging to the exponential family, along with their conjugate family of priors. They would be appropriate when one is monitoring a process which is, for the most part, in a nearly steady state over time and there is little prior information about how the system could change.

In our second scenario we consider a situation where it is known that the parameters tend to grow with time and we have some idea about the average or expected change from stage to stage. As in the first example we will assume the parameters are known to belong to the interval (a, b). The μ_j's will be selected in the same way and we let h be close to one. In the previous

example, we saw that if we took the cut points to be the midpoints of the μ_j's then the π_j with the largest weight at stage s is going to be the one whose mean is closest to the value of y_{s-1}. Now if the cut points, given by $\mathbf{u}$, were moved far enough to the left, i.e. decreased by a sufficient amount, and the W_j^s's are the W_j's defined in equation 9 using $\mathbf{u}$ then the intervals which get the most probability are not those close to θ_{s-1} but those that are farther to the right. For such a choice the model predicts that θ_s will tend to be larger than θ_{s-1}. How much larger depends on the values of the model parameters. Because of the form of $p(y_{t+1}|\mathbf{y}_t)$ given in equation 8 the predicted value of Y_s given $\mathbf{y}_{s-1}$ will be larger that y_{s-1}. As we shall see it is the proper choice of the cut points, at each stage, which will allow us to model a variety of phenomena. Note that if the parameters tend to grow smaller over time we would need to shift the cut points to the right to account for the decrease. Cyclic behavior of the parameters could be handled as well, if the periods of up swing and down turn are known. Although this is not the case in general, there are instances where this type of prior information is available. In the next section we will consider three examples where these different types of prior information are known.

5 Three Examples

5.1 Percentage of Defectives

For the first example we consider again the one discussed in section 2. Recall that we assumed there is a process producing organisms which begins in

control, i.e. only 20% of the organism produced are defective. At each stage we are interested in the posterior probability that the percentage of defectives exceeds 30%. We let $r = 2$ and select just two π_j's. The first is Beta(6.2,18.8) and the second is Beta(24.8,28.2). Their means are .2 and .4 respectively and their common variance is .005. The probabilities assigned to the interval (0,.3) are .74 and .01. The initial weight assigned to the first was .95 and to the second was .05. These choices are consistent with the underlying scenario. At each stage we take as the two weight functions, the W_j's defined in equation 9 with cut point .3. As described in the first scenario of the last section this choice is appropriate when the θ_s's are believe to more or less constant over time.

To study the performance of this model in detecting the process moving out of control, we considered two different situations. Each had 12 stages and in each case, at stage s, we observed Y_s a Binomial(20,θ_s) random variable. In the first case the first four θ_s's equaled .2, the next four equaled .3 and the last four equaled .4. In the second case the θ_s's were a sequence of length 12 beginning at .2 and ending at .4. For each case we observed the sequence $Y_1, \ldots, Y_{20}$ 500 times and for each stage computed the posterior probability that θ exceeds .3 and then took the average. The results for the odd numbered stages for the two cases are given in Table 1. It is important to note that this calculation is straight forward and is of the type that is often of interest in practice. Moreover, it can be found at each stage and does not need a set of preliminary data before it can be used. Hence it can used for small data sets. The specific values of the answers does indeed depend on the choice of the π_j's and W_j^s's. However, additional calculations demonstrated

that the answers are reasonably robust against model specification.

Put table 1 about here

5.2 Growth Data for Mice

For the next example we will consider a data set analyzed in Rao (1987) using growth curve models. The data are the weights of 13 mice taken at seven time points (Table 2). Rao was interested in the problem of predicting the weight on the seventh or last period given all the previous weights. A naive predictor would be to fit a straight line to the fifth and sixth period measurements and extrapolate to predict the seventh period. For these data the sum of the squared errors for this method is 0.055. Rao tried a variety of different methods and the best fit he found had a sum of squared errors of 0.031 (Table 7 of Rao (1987)). Because of the nature of the data, the earlier periods seem to carry little information about the later periods and so the linear extrapolation method seems to work reasonably well. For this reason will consider the problem of using periods 5 and 6 to to predict the weight for the last period.

For such an experiment the statistician will have prior information about a typical growth curve for a typical mouse. The average differences in the weight of the 13 mice for the six successive periods are 0.175, 0.165, 0.135, 0.119, 0.058 and 0.088. That is, the mice are gaining less weight on the average over each successive three day period, until the last period when the average gain increases by 0.03. Suppose that one knew a priori that the gain from period 6 to 7 was in fact on the average 0.03 more then the gain from

the previous period. Then another possible predictor at period 7 is just the weight of a mouse at period 6 plus the amount gained from period 5 to 6 plus 0.03. For this predictor the sum of the squared errors for the 13 mice is 0.044. In what follows we will assume that the statistician knows that the average gain from period 6 to 7 is larger than the weight gain from period 5 to 6 and has a reasonable prior guess for the difference between the two average gains.

Let $\theta_{s,i}$ denote the parameter for mouse i at period s. $Y_{s,i}$ will denote the random variable which is the weight of mouse i at period s. The values of the $Y_{s,i}$'s are what is recorded in Table 2. We can think of $\theta_{s,i}$ as the idealized true weight of mouse i at period s and $Y_{s,i}$ as representing the variability about this true weight. In the previous example the parameter at each period was the the percentage of defective organisms in the batch under consideration. It seems reasonable to assume that in some instances this parameter could be of real interest. In this example however the real interest seems to lie with the observables, i.e. the $y_{s,i}$'s. The existence of the $\theta_{s,i}$'s seem more problematical. They are useful however in modeling the observables and for this reason we find it helpful to consider them, even though some might question their existence. We will discuss this issue further in the next section. In any case we will consider the problem of using the values from periods 5 and 6 to predict the values of period 7. Since we are ignoring the first four periods we will denote these last three columns as columns 1, 2 and 3. Hence we will be using $y_{1,i}$ and $y_{2,i}$ to predict the value of $Y_{3,i}$ for $i = 1, \ldots, 13$.

We begin by selecting a model. We will use the normal-normal model described in the previous section. That is, we assume that the distribution

of $Y_{s,i}$ given $\theta_{s,i}$ is normal with mean $\theta_{s,i}$ and variance σ^2=.001. Let $\boldsymbol{\mu}$ be the vector of length 8 given by $\boldsymbol{\mu}$=(.6,.7,. . . ,1.2,1.3). We will select the $r = 8$ distributions π such that π_j is a normal density with mean μ_j and variance $\tau^2 = .01$. We also need to select values for $\mathbf{w}$ the initial weight vector. We chose $\mathbf{w} = (0, .1, .8, .1, 0, 0, 0, 0)$ which gives a prior expectation of the set of $\theta_{1,i}$'s equal to 0.8 which is consistent with the data since the average weight for this period is 0.805. As we shall see these choices of $\boldsymbol{\mu}$, σ^2, τ^2 and $\mathbf{w}$ just don't matter much when we are calculating our predictors of the weights of the mice at the last period. The most important fact about these choices is that the range of $\boldsymbol{\mu}$ covers the appropriate values and the difference between its successive values are roughly of the same order of magnitude as the average change in weight from stage to stage, that is the μ_j's are consistent with the scale of the problem.

It remains to select the W_j^1's and W_j^2's. Both will be of the form given in equation 9. In both cases we will take $h = .8$, again this particular choice has little affect on the calculations that follow and any other choice in this neighborhood would perform similarly. The key choice is the selection of the cut points for each stage. Let $\boldsymbol{\mu}_{-1}$ be the vector which consists of the first seven coordinates of $\boldsymbol{\mu}$. In each case the cut points will be of the form $\mathbf{u} = \boldsymbol{\mu}_{-1} - \gamma$ where γ is a real number. If we take $\gamma = -0.05$ then the cut points just become the midpoints between the successive μ_j's and as we saw in the discussion following equation 9 this would be correct if we believed that for mouse i the $\theta_{s,i}$'s, for $s = 1, 2$ and 3 were in a steady state. Since we know that for each mouse the $\theta_{s,i}$'s are increasing we need to increase the value of γ. In fact, since the average difference between stage 1 and 2 is 0.058, γ_1

should satisfy the relationship $-0.05 + .058 \doteq \gamma_1$ and so we take $\gamma_1 = .01$. In the same way, since the average difference between stage 2 and 3 is 0.088, γ_2 should satisfy the relationship $-0.05 + 0.088 \doteq \gamma_2$ and so we take $\gamma_1 = .04$. In our model $\gamma_2 - \gamma_1 = 0.088 - 0.058$ represents how much larger we expect the weight gain to be on average from stage 2 to 3 than it was from stage 1 to 2. Since this agrees exactly with the observed data we would expect this model to do well. In practice we would not expect to have such good information but it will be a useful test case and we will call it the bench mark model.

We have now defined a probability model for the weight of the 13 mice over the last three periods. The model treats each mouse separately and our prediction for a given mouse depends just on the model and the two previous observations for the given mouse. As we have noted this predicted value is easy to calculate through the updating procedure. For the above bench mark model we computed the prediction of the weight of each mouse at the last stage and found that the sum of the squared errors, say SSE, to be 0.022, a significant reduction over Rao's best fit which yielded an SSE = 0.031.

To justify our claim that the above result is reasonably robust against the choice of our model parameters we present some additional calculations where we change some of the values of the parameters in the bench mark model. We made two other choices of $\mathbf{w}$ which made the prior expectation of the $\theta_{1,i}$'s 0.75 and 0.84 respectively. In both cases the SSE was 0.022. The values of $\mathbf{w}$ just don't matter very much after the first stage because of the Markov nature of the model. Next we took $h = .95$ and $h = .7$, the resulting values of the SSE were 0.026 and 0.025. Clearly the results are more sensitive to the choice of h and it is a mistake to take h to small. We next let $\sigma^2 = .01$

and $\tau^2 = .1$ and $h = .95$ and found SSE $= 0.024$. Two other models, both with $h = .8$ where used. In the first we let $\sigma^2 = .001$ and $\tau^2 = .0025$ and in the second $\sigma^2 = .00025$ and $\tau^2 = .0025$, the resulting SSE's were 0.026 and 0.024. In the usual modeling situation σ^2 would represent the error in the scale used to measure the mice. The same is true here and in practice the choice of σ^2 should reflect this. On the other hand, τ^2 is seemingly related to the amount of variation in a typical mouse's weight from stage to stage. This latter relationship is not completely straightforward, because in most cases these changes from stage to stage depend more strongly on the W_j^t's. If τ^2 is chosen to be too large then the influence of the individual π_j's are minimized because then become to similar. But if it is chosen too small then certain intervals on the line will not be given much probability and the parameter space for the θ's will have "holes" in it. Hence if we avoid these two extremes the results of point prediction are fairly robust over a wide range of choices of σ^2 and τ^2. Next we shifted μ to the right by 0.03 and to the left by 0.03. The resulting SSE's were 0.023 and 0.021. This demonstrates the the choice of the μ_j's are not very important as long as they cover the range of possible values.

The parameters which most strongly affect the analysis are γ_1 and γ_2 which define the cut points for the two stages. We computed the SSE for 12 variations of the bench mark model for different choices of the γ's. The results are given in Table 3. Note that the choice of γ_1 is not very important. This is because of the Markov nature of our model and the fact that we are just considering the error in our predictions of the last stage. On the other hand SSE is sensitive to the value of γ_2. This is not surprising since it is

impossible to predict the future unless one is willing to make assumptions about how the future is related to the past. In this model this is exactly the role that γ_2 plays when we are predicting the last day.

Put table 3 about here

Up until now we have been choosing all the parameters in our model on the basis of assumed prior information. A more traditional Bayesian approach would be to consider some of the parameters to be unknown and then specify a prior distribution for them. For example after we have observed stages 5 and 6 the likelihood function of the data will include γ_1, but unfortunately not γ_2. If put a prior distribution over γ_1, then we could find its posterior distribution given all the observations from stages 5 and 6. Under this scenario the predictive distributions of the weights of the mice at stage 7 are no longer independent. However because of the form our model they are conditionally independent given the value of γ_1. If we assume some nice relationship between γ_2 and γ_1 then the predictive distribution of each mouse is just a mixture of the individual predictive distributions when γ_1 is known. One reasonable assumption, that could be sensible in practice, is to assume that the difference $\gamma_2 - \gamma_1$ is known. This is a slightly weaker assumption than assuming the cut points are known for both stages. Basically, one needs to have some idea how the average weight gain from stage 6 to stage 7 compares to the gain from stage 5 to stage 6. Recall in our bench mark model we took $\gamma_1 = .01$ and the difference $\gamma_2 - \gamma_1 = .03$ because that was indicated by the data. In Table 4 we give the results for three different possible sets of values for γ_1 and three different choices for the difference. In every case we put

the uniform prior of (1/3,1/3,1/3) on the three possible values of γ_1. Notice, that the first set is centered at .01, the "correct" value, while the second is shifted to the right and the third to the left. The three assumed values for the difference $\gamma_2 - \gamma_1$ are 0.01, 0.03 and 0.05. Again we have the "correct" value and an under estimate and an over estimate. Remember because of the scale of the problem these are fairly significant under and over estimates. It is not unreasonable to expect that in situations such as this one where much prior information is available one's prior guess should be quite close to the truth. It is interesting to note that in every case but one we have a smaller SSE than Rao's 0.031. This seems to suggest that the flexibility of this model allows one to incorporate prior information into a problem in a straightforward way that is more difficult in other setups.

Robert et al. (1993) considered a full Bayesian analysis for a hidden Markov model. Since the computational problems for such an approach are difficult they proposed an approximate method which relied on Markov chain Monte Carlo methods to study the model. Robert (1994) gives some additional discussion of a Bayesian approach to such problems.

Put table 4 about here

5.3 Monthly Housing Starts

A data set available in S, see Becker, Chambers and Wilk (1988), is the US monthly housing starts form January 1966 to December 1974. The first four years of these data are presented in Figure 1 along with the values of two predictors that will be discussed in the following. A brief perusal of the

data shows what is well known a priori. That is December, January and February are the months with the fewest housing starts and housing starts are roughly constant over these months. Then the next two months show dramatic increases in the number of housing starts. This is followed by six months of relatively stable or slightly decreasing numbers of housing starts. Finally the the number of housing starts falls off more sharply from October to November and November to December. We will now show how this prior information can be adapted to the model proposed in this note in a straight forward and essential object manner.

For definiteness supposed we have in hand the data for the year 1966 and want to make a prediction for January 1967. In addition we wish to have throughout the year a new prediction for the next month once the data for the current month is available. We will use the normal-normal model that was also used for the mice data. We begin by letting r, the number of π_j's be 9. We let $\boldsymbol{\mu}$=$(50, 65, \ldots, 155, 170)$. This choice covers the range of possible values, seems to have approximately the correct scale and as before the particular values are not very important. We let the initial weight vector put weights .1, .8 and .1 on the first three coordinates of $\boldsymbol{\mu}$ and zeros elsewhere. One possible sensible choice of τ^2 is to let it be 81. This is consistent with the scale of our choice of $\boldsymbol{\mu}$. Our choice of σ^2 should possibly reflect the variability of the observables, conditioned on the values of the parameters. How this should be related to τ^2 is not completely clear. Fortunately, as is the case with some of the other parameters in the model, when we are interested in point predictors, the actual choice does not matter much. We will begin by taking $\sigma^2 = 81$. The same is true for the value of h,

used to define the weighting functions, as in the last example we let $h = .8$. It remains to select the cut points, one set for each month of the year.

As we saw in the previous example, for the sensitivity of our predictions, these are the most important parameters. There are eight months, January, February and May through October, where we believe that the housing starts for this month should be quite similar to those of the previous month. For such a steady state situation we take the cut points $\mathbf{u}$ to be just the midpoints between the successive members of $\boldsymbol{\mu}$. Using our previous notation, this is the vector $\boldsymbol{\mu}_{-1} + 7.5$. The cut points for the change from February to March in 1967 should reflect our beliefs about how much larger the March value will be than the February value. One naive estimate of this difference, which we have in hand, is just the difference between the March and February values from the previous year, 1966. This value is 43.4. Hence the cut points for the change from February to March should just be the cut points for January to February shifted to the left by the amount 43.4, i.e. $\boldsymbol{\mu}_{-1} + 7.5 - 43.4$. The cut points for the change from March to April could be handled in exactly the same way. However we will make a modification which we will allow us to take into account possible differences from year to year. We will take as our estimate of the change from March to April for the current year to be the average of the change from March to April for the past year with the change from February to March of this year. For the year 1967 this becomes $\boldsymbol{\mu}_{-1} + 7.5 - (21.4 + 29.7)/2$. The cut points for the last two months are handled similarly except the shift must be to the right since we know that housing starts are decreasing. For example in 1967 the cut points for the change from October to November would be $\boldsymbol{\mu}_{-1} + 7.5 + 4$ and for November

to December would be $\boldsymbol{\mu}_{-1} + 7.5 + (12.8 + 16.8)/2$.

We wish to emphasize that the selection of the cut points described above is essentially objective and makes use of the kinds of prior information that has been traditional used when modeling time series. Moreover the above approach could be used for a group of successive years where at the end of each year the last set of updated weights could be taken as the prior weights for the next year. For the housing start data, using the choice of cut points and other model parameters described above, we found the predictions for each month in 1967, 1968 and 1969. The average squared error of the 36 predictions, say ASE, was 100.9. To show why the actual choices of σ^2 and τ^2 don't matter much when making point predictions we report a few other results. Keeping $\tau^2 = 81$ we let $\sigma^2 = 40, 160$ and 800 and found the ASE's to be 104.7, 102.2 and 122.5. Three other cases considered were $\sigma^2 = \tau^2 = 200$, $\sigma^2 = 200$ and $\tau^2 = 400$, and $\sigma^2 = 400$ and $\tau^2 = 200$. The resulting ASE's were 101.5, 103.6 and 106.8. Other calculations show that the choices of the initial weight vector don't matter and h can range between .8 and .9 without much affect on ASE. Two naive predictors would be to predict this month by last month or by the same month of the previous year. The ASE's for these two methods are 341.8 and 540.9.

To compare our results to a more sophisticated analysis we used the general seasonal multiplicative ARIMA process to fit the data. Following the definition on page 313 of Brockwell and Davis (1987) we considered the SARIMA$(p, d, q) \times (P, D, Q)_s$ Process with $p = P = 1$, $s = 12$ and $d = D = q = Q = 0$. We fit this model to the first four years of the data. Since this is not many observations for this type of analysis we fit it again to all the

data from the eight years. We did this using an ARIMA package in S-Plus, see Statistical Sciences Inc. (1991). Since this model uses y_{t-1}, y_{t-12} and y_{t-13} to predict Y_t we have only 35 possible one step ahead test predictions, February 1967 through December 1969. For the estimated model from the two data sets the ASE's were 294.6, when just the first four years were used, and 184.4, when all eight years were used. Even though the diagnostics for this two procedures seemed reasonable, we also repeated the process when $p = 2$, keeping the other parameters unchanged. The results gave a minuscule reduction in the ASE's. In Figure 1 we have plotted the the data for 1966 through 1969 and the one step ahead predictions for our model with $\sigma^2 = \tau^2 = 81$ and from the SARIMA model which was fitted to all eight years of data, with p=1. It is clear from the plots that our method gives better predictions.

So far we have just concentrated on point predictions for this example. In such situations one is often interested in an estimate of variance for the predictor. Under our model this is easy to find, however this variance depends on the values of σ^2 and τ^2 and increase as they increase. There is no correct, objective solution for the problem of selecting these values. A sensible solution depends on your prior information and getting the scale approximately correct. That is why we used $\sigma^2 = \tau^2 = 81$ for the calculations presented here. The average value of the standard deviations for the 36 predictions was 25.3. If we used $\sigma^2 = \tau^2 = 49$ and keep the rest of the parameters unchanged, we find an ASE of 104.0 and an average value of the standard deviations of the predictions of 23.6. So again we see for this model, that although an answer does depend of the values of the parameters, it is reasonable robust

and not overly sensitive to the actual choices.

Because of the beauty and simplicity of the Kalman filter there have been many attempts to generalized it. West et al. (1985) is such a generalization for Bayesian modeling and forecasting in dynamic state space models. However the updating process is more complicated then what is needed here. As we have seen both the modeling and analysis for these data is relatively straightforward.

6 Conclusions

We have presented a hidden Markov model for dependent data being observed over time. Given the many papers on probability models in the statistical literature it is perhaps surprising that the role of such models in statistical inference is an issue of some controversy. For a recent discussion see Freedman (1996) and the related discussion. Freedman notes that proposers of regression models suggest that they could be useful "for (i) description, (ii) prediction and (ii) inferring causation from association". He accepts the first, allows how the second could be true in certain situations and argues strongly against the third especially in models with lots of parameters that are difficult to interpert. Our models are certainly similar in that they have many parameters some of which may be difficult to interpret.

In the previous section we considered three different examples. In the first we argued that it made sense to us to assume that the hidden Markov parameters actually existed. For the other two examples this is not the case and we agree with Freedman's argument that our models do not demonstrate

how θ_t is "caused" by θ_{t-1}. In fact we do not claim that this is even so in the first example where we assumed the θ_t's actually exist. We are not proposing these models because we believe that they do indeed "explain" or model how the data is actually generated. For us the unobserved "parameters" are an assumed mathematical convenience which in some instances have a sensible interpretation. For a given set of data we hope to find a model which is a helpful description or summary of the data and in some cases a good predictor of future observations. Our model depends on many underlying parameters and as we have seen many different models give about the same results for the prediction problem. For us finding a reasonable model should not be thought of as a model selection problem where one is trying to find a "best" model. To use this approach one needs to have enough prior information so that one has the basic scale of the problem correct. If in addition one wants the model to yield good predictions over time then some prior information or beliefs about how future observations are related to the past is also needed.

In the first example we considered a "steady state" type of problem where random shocks are assumed to happen over time but no attempt is made to model how they occur. However because of the computational simplicity of the model it is easy to find the posterior probability that a random shock has occurred at any stage. The next two examples where selected so that our approach could be compared to other methods. In Rao (1987) and the related discussion there was some unhappiness about the performance of growth curves in explaining the data. As was noted before part of the trouble is that what happens at the later stages does not depend very strongly on what happens at the earlier stages. So using models which tie the stages

together leds to some inflexibility which hampers prediction. Since we were interested in prediction we extracted some "prior" information from the data that allowed us to relate the first two stages to the last one, the one we wished to predict. Without building such relationships into the model prediction seems impossible to us. Hopefully the information used from the data is of the type that could realistically be assumed to exist in practice and this will not be seen as just a data mining expedition. In the housing example the type of prior information used is readily available and the modeling done here seems to us tp be as "objective" as that used in other approaches.

We have just concentrated on problems where one needed to find $p(\theta_t|\mathbf{y}_t)$ and $p(y_{t+1}|\mathbf{y}_t)$. However it is easy to find $p(\theta_{t-1}, \theta_t|\mathbf{y}_t)$ and $p(y_{t+1}, y_{t+2}|\mathbf{y}_t)$ and similar quantities. We have restricted attention to problems where the Y_t's and θ_t's are real valued. This is not necessary and vector valued problems can be handled formally in exactly the same way. For such problems selecting appropriate weight functions used in the mixture of the π_j's can be more difficult. We believe that the examples considered above have demonstrated that these models should prove useful in a variety of situations. One advantaged they have over more traditional time series methods is that they can be applied in problems with only a few observations.

References

[1] Becker, Richard A., Chambers, John M. and Wilks, Allan R. (1988), *The New S Language*, Wadsworth & Brooks/Cole, Pacfic Grove, California.

[2] Brockwell, Peter J. and Davis, Richard A, (1987), *Time Series: Theory and Methods*, Springer-Verlag, New York.

[3] Carlin, Bradley P., Polson, Nicholas G. and Stoffer, David S. (1992), "A monte carlo approach to nonnormal and nonlinear state-space modeling",*Journal of the American Statistical Association*, 87 493-500.

[4] Carter, C. K. and Kohn, R. (1994), "On Gibbis sampling for state space models", *Biometrika* 81 541-553.

[5] Freedman, D. (1996), "Some issues in the foundation of statistics" (with discussion), *Foundations of Science*, 1, 19-39. Polish Academy of Science, Warsaw. Reprinted by Kluwer, Dordrecht, The Netherlands.

[6] Elliott, Robert J., Aggoun, Lakhdar and Moore, John B. (1995), *Hidden Markov models: Estimation and control*, Springer, New York.

[7] Leroux B. G. and Puterman, M. L. (1992), "Maximum-penalized- likelihood estimation for independent and Markov-dependent mixture models", *Biometrics* 48 545-558.

[8] Meinhold, Richard J. and Singpurwalla, Nozer D. (1983), "Understanding the Kalman Filter", *American Statistician*, 37 123-127.

[9] Rao, C. R. (1987), "Prediction of Future Observations in Growth Curve Models", *Statistical Science*, 2 434-471.

[10] Robert, Christian P. (1994), *The Bayesian choice: A decision-theoretic motivation*, Springer, New York.

[11] Robert, Christian P., Celeux, Gilles and Diebolt, Jean (1993), "Bayesian estimation of hidden Markov chains: A stochastic implementation" *Statistics and Probability Letters*, 16 77-83.

[12] Rydén, Tobias (1994), "Consistent and asymptotically normal parameter estimates for hidden Markov models", *Annals of Statistics*, 22 1884-1895.

[13] *S-Plus*, (1991), Statistical Sciences, Inc., Seattle, Washington.

[14] Williams, J. S. and Izenman, A. J. (1981), "A Class of Linear Spectral Models and Analyses for the Study of Longitudinal Data", Technical Report, Dept. of Statistics, Colorado State Univ.

[15] West, Mike, Harrison, Jeff and Migon, Helio S. (1997), "Dynamic generalized linear models and Bayesian forecasting" (with discussion), *Journal of the American Statistical Association*, 80 73-97.

[16] West, Mike and Harrison, Jeff (1997) *Bayesian forecasting and dynamic models*, 2nd ed. Springer, New York.

Table 1: The average probability, based on 500 repetitions, that the parameter exceeds .3 for the odd numbered stages for the Beta-Binomial example for the two different cases.

	Stage					
	1	3	5	7	9	11
case 1	.16	.19	.34	.49	.72	.83
case 2	.16	.25	.35	.49	.62	.79

Table 2: Weights of 13 male mice measured at successive intervals of 3 days over 21 days from birth to weaning (Williams and Izenman (1981)).

	Day 3	Day 6	Day 9	Day 12	Day 15	Day 18	Day 21
1	0.190	0.388	0.621	0.823	1.078	1.132	1.191
2	0.218	0.393	0.568	0.729	0.839	0.852	1.004
3	0.211	0.394	0.549	0.700	0.783	0.870	0.925
4	0.209	0.419	0.645	0.850	1.001	1.026	1.069
5	0.193	0.362	0.520	0.530	0.641	0.640	0.751
6	0.201	0.361	0.502	0.530	0.657	0.762	0.888
7	0.202	0.370	0.498	0.650	0.795	0.858	0.910
8	0.190	0.350	0.510	0.666	0.819	0.879	0.929
9	0.219	0.399	0.578	0.699	0.709	0.822	0.953
10	0.225	0.400	0.545	0.690	0.796	0.825	0.836
11	0.224	0.381	0.577	0.756	0.869	0.929	0.999
12	0.187	0.329	0.441	0.525	0.589	0.621	0.796
13	0.278	0.471	0.606	0.770	0.888	1.001	1.105

Table 3: The SSE or sum of the squared error of predictions at the last stage for the 13 mice with the bench mark model with three different choices of γ_1 and γ_2 defining the cut points of the two stages.

	Value of γ_2			
	.01	.03	.05	.07
$\gamma_1 = -.01$	.025	.021	.024	.033
$\gamma_1 = .01$	.024	.021	.024	.034
$\gamma_1 = .03$	.025	.021	.024	.035

Table 4: The SSE or sum of the squared error of predictions at the last stage for the 13 mice with the bench mark model with three different sets of possible γ_1 values under the uniform prior and three different assumed values for $\gamma_2 - \gamma_1$.

Set of Possible γ_1 Values	$\gamma_2 - \gamma_1$	SSE
	.01	.022
(-.01,.01,.03)	.03	.022
	.05	.028
	.01	.021
(.01,.04,.07)	.03	.028
	.05	.040
	.01	.030
(-.05,-.02,.01)	.03	.023
	.05	.021

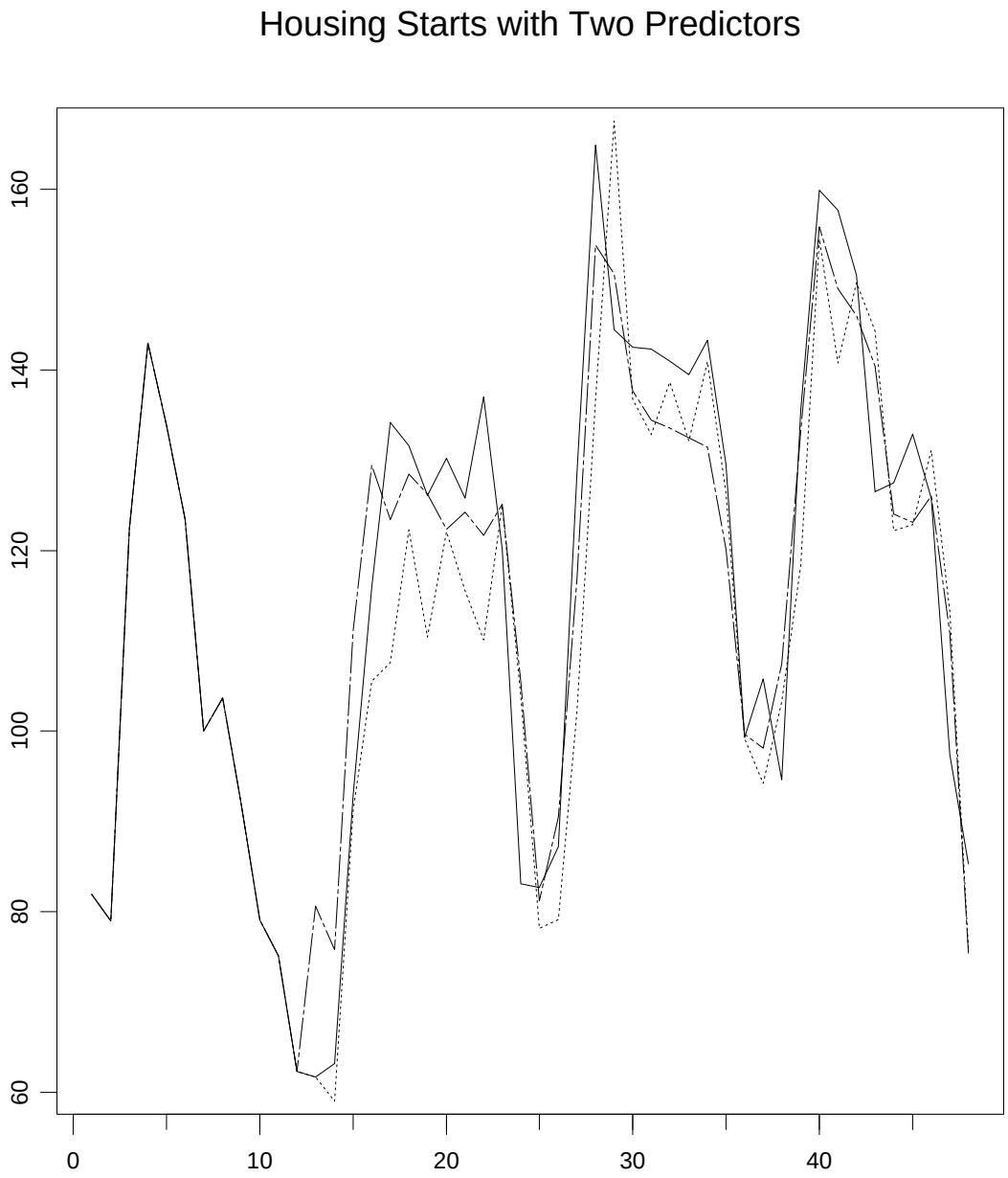

Figure 1: For the US monthly housing starts data for 1966 through 1969, the solid line is the original data, the equal sized dashed line is the 35 predicted values based on the SARIMA model and the unequal sized dashed line is the 36 predicted values for the model with $\sigma^2 = \tau^2 = 81$.