ISSN 1936-5330

THE FEDERAL RESERVE BANK *of* KANSAS CITY
RESEARCH WORKING PAPERS

Recession Forecasting Using Bayesian Classification

Troy Davig and Aaron Smalter Hall

August 2016; Revised February 2017
RWP 16-06
https://dx.doi.org/10.18651/RWP2016-06

RESEARCH WORKING PAPERS

Recession Forecasting Using Bayesian Classification*

Troy Davig† Aaron Smalter Hall‡

This version: February 9, 2017

Abstract

We demonstrate the use of a Naïve Bayes model as a recession forecasting tool. The approach has a close connection to Markov-switching models and logistic regression, but also important differences. In contrast to Markov-switching models, Naïve Bayes treats National Bureau of Economic Research business cycle turning points as data, rather than hidden states to be inferred by the model. Although Naïve Bayes and logistic regression are asymptotically equivalent under certain distributional assumptions, the assumptions do not hold for business cycle data. As a result, Naïve Bayes has a larger asymptotic error rate, but converges to the error rate faster than logistic regression, resulting in more accurate recession forecasts with limited data. We show Naïve Bayes consistently outperforms logistic regression and the Survey of Professional Forecasters for real-time recession forecasting up to 12 months in advance. These results hold under standard error measures, and also under a novel measure that varies the penalty on false signals depending on when they occur within a cycle. A false signal in the middle of an expansion, for example, is penalized more heavily than one occurring close to a turning point.

JEL Classification: C11, C5, E32, E37

---

*We thank Travis Berge, James Hamilton, Jeremy Nalewaik, Glenn Rudebusch and Jonathan Wright for helpful comments and suggestions. The views expressed herein are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

†Research Department, Federal Reserve Bank of Kansas City, Kansas City, MO 64198, USA.

‡Corresponding author. Research Department, Federal Reserve Bank of Kansas City, Kansas City, MO 64198, USA. Email: aaron.smalterhall@kc.frb.org

1 Introduction

The onset of a recession is one of the most significant macroeconomic events, as unemployment often rises sharply and output declines, yet forecasting their occurrence remains a challenge. A common approach is to treat the turning point dates set by the National Bureau of Economic Research's (NBER) Business Cycle Dating Committee as data, then use various explanatory variables in a binary response framework. For example, Rudebusch and Williams (2009) show that the slope of the US Treasury yield curve in a probit model can outperform the recession predictions from the Survey of Professional Forecasters. The tradition of this line of literature is clearly one of turning point prediction. Another approach is to use a framework that can capture important shifts in the data, allowing the framework to identify turning points rather than be told when they occur. Markov-switching models following the approach of Hamilton (1989) are the leading examples, as they have been shown to identify turning points similar to NBER dates.

In this paper, we take elements from both approaches by using a Naïve Bayes (NB) framework as a recession prediction tool. The approach has precedent in Neftci (1982) and Diebold and Rudebusch (1989), which also have close connections to both Markov-switching and logistic regression (LR) models. We use Bayes Theorem in a manner similar to these papers, except incorporate a richer data set, lag structure and capture the persistence of business cycle phases by using Markov-switching transition probabilities. In this respect, our approach also closely connects to Markov-switching time-series models, except we treat NBER turning points as data when identifying past recessions and expansions rather than something to be inferred. The logistic approach also treats the NBER business cycle dates as data and under certain assumptions, is equivalent to the NB approach. For business cycle data, however, the assumptions generating equivalence do not hold and as a result, recession forecasting accuracy diverges across the two approaches, with NB being considerably better under a range of criteria.

At first, the more accurate forecast from NB relative to LR may be surprising, as LR has a lower asymptotic classification error. Of particular relevance to business cycle forecasting, however, is that NB reaches its asymptotic error rate much faster than LR. As a result, the potential exists that NB can outperform LR when limited data is available. Given the relatively small number of recession observations, we find this difference is important when forecasting business cycle turning points and results in NB outperforming LR using either revised or real-time data.

The NB framework can also easily incorporate a large amount of data with a rich lag structure. While we use a large amount of data, we find a relatively narrow set performs quite well when incorporating a relatively long lag structure. Consequently, most of results focus on nonfarm payroll growth, the Institute of Supply Managements (ISM) Manufacturing Index, the change in the S&P 500 and the term spread (i.e. the difference between yields on the 10-year and 2-year Treasury securities). For payrolls, we look separately at both real-time and revised data.[1]

To evaluate forecasting performance, we use the F-measure under a zero-one loss. This approaches penalizes trivial classifiers, such as one that might predict the economy never enters a recession. Given the relatively infrequency of recessions, such an approach would be correct most of the time, but lacks true predictive ability. For comparison to other work, we also evaluate outcomes under mean absolute error. In addition to weighting each observation uniformly under these criteria, we also construct a novel weighting criteria that varies the weight of a forecast error depending on when it occurs within a business cycle. For example, indications of recession that occur in the middle of an expansion are penalized more heavily than if a signal is given closer to the time when a recession occurs.

2 Background

The Naïve Bayes approach connects two strands of literature on business cycle turning points.[2] The first focuses on using a set of data to predict whether the economy will be in a recession at some point in the future using a binary response framework. The second focuses on nowcasting and classification of business cycle states. Success of a model in this context is often how well it can identify past periods the NBER defined as recessions without using the NBER turning points in the estimation.

2.1 Turning Point Prediction

Diebold and Rudebusch (1989) use an approach closely connected to the one in this paper. Using a composite index of leading indicators within a Naïve Bayes framework, the authors produce probabilistic forecasts of future business cycle turning points and

[1]The ISM index and financial variables are not subject to revision.

[2]Naïve Bayes has also been applied in other areas, such as to infer monetary policy regimes. For example, see Jefferson (1998)

evaluate how much lead time the index can provide in terms of signalling a peak or trough. The approach works well at near-term horizons and for comparison, we also use a version of the model with only a leading economic index. However, as will be discussed in subsequent sections, our approach differs in a few dimensions, such as the construction of prior probabilities, data and lag structure, that improves forecasting performance.

In addition to using data on macroeconomic activity, several studies such as Estrella and Hardouvelis (1991), Hamilton and Kim (2002), Ang, Piazzesi and Wei (2006), Rudebusch and Williams (2009) and Liu and Moench (2016) show the spread between yields on longer- (e.g. 10-year) and shorter-term (e.g. 3-month) US Treasury securities provides valuable information for predicting future real GDP growth and recessions. In terms of the binary response approach, Estrella and Mishkin (1998) use a probit framework including financial variables to predict whether the economy will be in recession at some point in the future. Chauvet and Potter (2002) and Chauvet and Potter (2005) also us a probit framework including term spread and highlight that breakpoints are likely to have occurred, which alters the mapping between the term spread and likelihood of entering a recession.

Wright (2006) adds the term premium and level of the federal funds rate to a probit model and modifies the dependent variable. Instead of predicting whether the economy will be in a recession at a particular point in time, say at $t + h$, Wright evaluates whether a recession is likely to begin at any point between time $t + 1$ and $t+h$. Other approaches include Kauppi and Saikkonen (2008) that develop a dynamic binary probit model and Chen, Iqbal and Lai (2011) that incorporates a broad set of data by using principal component analysis within a probit-dynamic factor model. More recently, Ng (2014) and Berge (2015), use machine-learning algorithms and model averaging using relatively large data sets to predict recessions.

2.2 Business Cycle Classification

For nowcasting and classification, Markov-switching frameworks are the leading examples. Starting with Hamilton (1989), then including Hamilton (2011) and many others, these approaches use macroeconomic data to estimate models that will assign a probability that the economy is in any number of predefined states. The number of states is at the researcher's discretion, although is often set to two in the business cycle context to correspond to expansions and recessions. At a given point in time, probabilities reflect either all the information up until that time (i.e. filtered) or the entire sample (i.e. smoothed). One issue is that estimates of the underlying parameters governing the time-series process in each regime are based on the entire sample, so even filtered probabilities use information that is not available at the time an estimate is made.

As a classification tool, the early success of Markov-switching models was based on how well the different regimes corresponded to NBER-defined recessions and expansions. As a business cycle forecasting tool, they face a few shortcomings. First, all probabilistic estimates are contemporaneous, so only provide an indicator as to the state of the business cycle as of the last available data point. Still, these models can supply signals useful from a nowcasting perspective. Second, many papers may use a limited set of data, such as only GDP.

The Markov-switching literature has addressed some of these issues. For example, Chauvet and Hamilton (2006) recursively estimate a state-space Markov-switching model using real-time data, so provide an assessment of how well this class of models perform as a nowcasting tool. In general, filtered probabilities often send a signal in real-time when the economy may be slipping into a recession. As a business cycle classification tool, however, waiting at least a few quarters using smoothed probabilities produce estimates more in line with the NBER chronology. In a related approach Nalewaik (2011*a*) shows Gross Domestic Income (GDI), rather than GDP, improves the signal sent by Markov-switching models near the onset of a recession. Nalewaik (2011*b*) also highlights the importance of using real-time data within a Markov-switching model by showing it improved the ability to detect the onset of the 2001 recession. In terms of the ability to predict recessions further in the future, Nalewaik (2012) uses a three-state model with a "stall" state that is specified to precede recessions. When the economy enters the stall phase, the economy is then more likely to enter a recession in the subsequent periods. From this standpoint, the framework is a useful tool for forecasting recessions. One aspect, however, is that the model is at a quarterly frequency and the most promising specification uses GDI, which is not released until about two months after the end of a quarter. As a result, the lag between the possible onset of a recession and a signal from the model is likely to be several months. In terms of incorporating a larger set of data at a higher-frequency, Davig (2008) uses the first principal component of a large set of monthly macroeconomic data maintained by the Federal Reserve Bank of Chicago within a Markov-switching framework. Another example includes Giusto and Piger (2014) who use a machine-learning algorithm (i.e. Learning Vector Quantization) to classify recession and expansions in real-time using mixed-frequency data. The approach allows updating with each incoming data point, so can fully incorporate information even from the ragged edge of the data. As a result, the framework is useful as a nowcasting tool, but is limited if the interest is in forecasting a recession several periods into the future.

3 Methods

This section provides an overview of the NB algorithm and connection to Markov-switching and binary response models, along with a discussion of asymptotic properties of NB that provides some intuition why it outperforms LR in a business cycle forecasting context.

3.1 The Naïve Bayes Model

Naïve Bayes is a straightforward supervised model that can be surprisingly effective.[3] NB uses Bayes theorem to learn the conditional probability that observed data is drawn from a certain class of observations. Bayes theorem for k classes is given as

$$P(C_k|\mathbf{x}) = \frac{P(C_k)P(\mathbf{x}|C_k)}{P(\mathbf{x})}, \tag{1}$$

where C_k is the kth class and $\mathbf{x} = (x_1, ..., x_m)$ is a unit of observed data with m variables. The $P(C_k)P(\mathbf{x}|C_k)$ term in the numerator of the right-hand side is equivalent to the joint probability $P(C_k, x_1, ..., x_m)$ and so, assuming conditional independence of the variables, is decomposed as

$$P(C_k, x_1, ..., x_m) = P(C_k)P(x_1|C_k)...P(x_m|C_k) = P(C_k)\prod_{i=1}^{m} P(x_i|C_k). \tag{2}$$

Variable independence is a strong assumption, which is often violated, and would seem to undermine the utility of NB. In practice though, violating this assumption

[3] In the machine learning literature, "supervised" refers to the labeling of specific classes of outcomes, such as periods that correspond to recessions and expansions. Alternatively, Markov-switching time series models would be "unsupervised," since the algorithm assigns each set of observations to different regimes without the investigator *a priori* assigning observations to a regime or defining the characteristics of the regimes.

has surprisingly little impact on predictive ability. A number of theoretical results, discussed briefly in the next subsection, show cases in which conditional independence can be violated while preserving forecasting accuracy.[4]

Given the equations above, the decision rule for the NB model which assigns an outcome class $\hat{y}$ to observation $\mathbf{x}$ is written as

$$\hat{y} = \underset{k\in\{1,\ldots,k\}}{argmax}\ P(C_k)\prod_{i=1}^{m} P(x_i|C_k). \tag{3}$$

3.2 Naïve Bayes' Connection to Markov Switching

In a business cycle context, we can affix time subscripts and rewrite (1) to reflect the probability the economy is currently in a recession at time t as

$$P(R_t|\mathbf{x}_t) = \frac{P(R_t)f(\mathbf{x}_t|R_t)}{P(R_t)f(\mathbf{x}_t|R_t) + P(E_t)f(\mathbf{x}_t|E_t)}, \tag{4}$$

where we replace the arbitrary class label C_k with two state symbols $C \in \{R, E\}$ and subscript them by time to denote that either a recession or expansion is in place at t. $\mathbf{x}_t$ is the vector of observed data $\mathbf{x}$ at time t.

Chauvet and Hamilton (2006) illustrate how using the unconditional probability that the economy is in recession is the first step toward making the NB algorithm useful as a business cycle classification tool. For example, from January 1959 to until June 2016, the economy has been in recession 13.5% of the time, implying $P(R_t) = .135$ and $P(E_t) = .865$. The second step is to parameterize the conditional density as

$$f(x_{i,t}|R_t) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-\left(x_{i,t} - \mu_{i,R}\right)}{2\sigma_i^2}\right), \tag{5}$$

where σ_i is the standard-deviation of x_i, which is specified to be independent of the business cycle, and $\mu_{i,R}$ is the mean of x_i conditional on being in a recession. The conditional density for expansions is defined analogously. The conditional density in (5) can incorporate a variety of dynamics, such as lagged endogenous and exogenous variables.

Chauvet and Hamilton (2006) also highlight that accounting for the persistence of recessions and expansions can improve the inference from (4). For example, of the

[4] For example, see Domingos and Pazzani (1997) and Zhang (2004).

597 months the economy has been in an expansion starting in 1959 until June 2016, 98.5% of the time the NBER has declared the economy to also be in an expansion the following month. When in recession, 91.4% of the time the economy is also in a recession the following month. Combining an estimate regarding the current business cycle state with the Markov transition probabilities provides an estimate that the economy will still be in recession the following period. For example,

$$P(R_{t+1}|\mathbf{x}_t) = P(R_{t+1}|R_t, \mathbf{x}_t)P(R_t|\mathbf{x}_t) + P(R_{t+1}|E_t, \mathbf{x}_t)P(E_t|\mathbf{x}_t), \tag{6}$$

where the Markov-switching model sets $P(R_{t+1}|R_t, \mathbf{x}_t) = P(R_{t+1}|R_t)$ and $P(R_{t+1}|E_t, \mathbf{x}_t) = P(R_{t+1}|E_t)$. If the NBER dates are taken as data, then $P(R_{t+1}|R_t) = .913$ and $P(E_{t+1}|E_t) = .985$. When the next period arrives, combining the new data with (6) as follows

$$P(R_{t+1}|\mathbf{x}_{t+1}) = \frac{P(R_{t+1}|\mathbf{x}_t)f(\mathbf{x}_{t+1}|R_{t+1})}{P(R_{t+1}|\mathbf{x}_t)f(\mathbf{x}_{t+1}|R_{t+1}) + P(E_{t+1}|\mathbf{x}_t)f(\mathbf{x}_{t+1}|E_{t+1})}, \tag{7}$$

provides an update to the conditional probability that the economy is in a recession.

In practice, Markov-switching models do not take NBER-defined business cycle states as data and instead, identify when shifts occur from the data and treat the transition probabilities as parameters to estimate. Hamilton (1989) shows how to combine the various elements above to construct the likelihood function and then estimates a model using real GNP, which produces states that align well with NBER-defined dates. From this standpoint, the Markov-switching framework is an unsupervised model that is effective at modeling historical data and finding points at which the economy moves from expansion to recession and back again. It naturally accounts for the persistence of recession/expansion periods and temporal nature of the data.

For our approach, the key point of departure from the Markov-switching framework is that we treat NBER-defined business cycles as data. As a result, we can use Bayes Theorem to not only classify current data as belonging to a recession or expansion, but whether current data is informative about the occurrence of a recession at some point in the future. For example, assume we want to make an inference on the state of business cycle six months ahead using current data, so want to calculate

$$P(R_{t+6}|\mathbf{x}_t) = \frac{P(R_{t+6})f(\mathbf{x}_t|R_{t+6})}{P(R_{t+6})f(\mathbf{x}_t|R_{t+6}) + P(E_{t+6})f(\mathbf{x}_t|E_{t+6})}, \tag{8}$$

which requires

$$f(x_{i,t}|R_{t+6}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(x_{i,t} - \mu_{i,R})}{2\sigma_i^2}\right). \tag{9}$$

Treating the NBER data as observed makes these calculations possible. In the Markov-switching setting, (9) cannot be computed because future regimes at time t are inherently unknown. Of course, Markov-switching frameworks can easily generate $P(R_{t+6}|\mathbf{x}_t)$ by iterating on the transition probabilities, as in (6). Instead, NB directly incorporates the conditional density (9) over the data six periods ahead into (8), rather than the conditional density over the current data combined with transition probabilities.

An important difference of this approach with the work of Neftci (1982), Palash and Radecki (1985), Diebold and Rudebusch (1989) and Diebold and Rudebusch (1991) that use a Naïve Bayes approach is this lag structure. These papers report results that use a formulation similar to (4), but instead of utilizing a lag structure as in (8), use a composite index of leading indicators. When these frameworks signal a turning point, it is a sign on an imminent recession and evaluation of the model is in terms of how long before a recession actually transpired. To connect and compare to these earlier results, we will also use a composite of leading indicators, as well as other data, within the lag structure illustrated in (8). Another difference in our approach to these papers is the Markov transition probability we incorporate into the model, as we illustrate next.

3.3 Incorporating State Persistence into Naïve Bayes

Markov-switching models have the capability of enforcing persistence for recession/expansion states, which is an important part of forecasting turning points because we know that business cycle phases tend to persist over time. This aspect is not present in the standard NB model, since forecasts for all periods are treated as independent from each other. We therefore modify the NB model by replacing the unconditional prior with the empirical Markov transition probabilities. Again, this is possible because we treat the NBER-defined turning points as data, whereas they are objects to estimate in the Markov-switching setting.

In the recession forecasting context, the decision function then becomes

$$\hat{y}_t = \underset{C\in\{E,R\}}{argmax} \, (P(C_{t+j}|R_{t+j-1})P(R_{t+j-1}) + ... \\ P(C_{t+j}|E_{t+j-1})P(E_{t+j-1})) \prod_{i=1}^{m} P(x_{i,t}|C_{t+j}), \tag{10}$$

where j is the recession forecasting horizon. We refer to this model as the Markov-Switching Naïve Bayes (MS-NB) model. This approach captures persistence, but in terms of future forecasts. For example with $j = 6$, the model provides a prediction of the business cycle state six months ahead conditional on data at time t. If the model forecasts a high probability the economy is in a recession at $t + 5$, the prediction for $t + 6$ will incorporate this information rather than treat it as an independent observation.

3.4 Naïve Bayes' Connection to Probit and Logit Models

Another approach to forecasting business cycle turning points is to directly estimate (8). Estrella and Mishkin (1998) take this approach using a probit model, so estimate

$$P(R_{t+h}|\mathbf{x}_t) = \Phi\left(\alpha_0 + \alpha_1 \mathbf{x}_t\right),$$

where h corresponds to the forecast horizon.

Probit and logit models are quite similar, as both fit sigmoid-style functions to the observed data. For LR, the conditional probability of a recession is written as

$$P(R_t|\mathbf{x}_t) = \frac{1}{1 + exp(w_0 + \sum_{i=1}^{m} w_i x_{i,t})}. \tag{11}$$

and has a direct connection to the NB model under three assumptions: 1) the class variable C takes on a binary value - in the current setting, either R or E, 2) the continuous random variables $x_1, ..., x_m$, x_i and x_j are conditionally independent for all i and j in $\{1, ..., m\}$ for $i \neq j$, 3) $x_{i,t}$ is drawn from $N(\mu_{i,C}, \sigma_i)$, so that the mean depends on the class, but not the variance. Given these assumptions, the parametric form of the conditional distribution, $P(R|\mathbf{x}_t)$ as in (4) or $P(R|\mathbf{x}_{t+6})$ as in (8), learned indirectly by NB corresponds exactly to the conditional distributions given for LR in equation (11) as shown in Ng and Jordan (2002).

The assumptions that establish equivalence between NB and LR, however, do not always hold. First, the explanatory variables may not be drawn from a Gaussian distribution, which destroys equivalence between the two. Second, violating the assumption of equal class variance leads to different decision boundaries. A decision boundary is the threshold in the variable space that separates observations of one class from another. The boundary is used to make predictions as to the class corresponding to a new set of observations. For example, in the business cycle context, the model will label a set of observations corresponding to a recession or expansion depending on which side of the boundary the observations fall. For NB, the boundary has quadratic curvature in the presence of heteroscedasticity, whereas the boundary remains linear for the LR model. As we will illustrate in the case of business cycle data, equal class variance does not hold and the curvature in the decision boundary for NB produces more accurate predictions. Finally, as a further note, the LR model does not require specifying a distribution over the explanatory variables, while NB does. While this may appear to disadvantage NB, it can actually be a benefit when using smaller samples as explored in the next section.

3.5 Asymptotic Properties of Naïve Bayes and Logistic Regression

By estimating the conditional distribution directly, LR has a lower asymptotic classification error than NB. NB can, however, reach the asymptotic error rate faster with respect to the number of training observations.[5] NB converges to the asymptotic error in $O(log\ m)$ observations, compared to $O(m)$ for LR, where m is the number of variables. This point is relevant for forecasting recessions, as the economy has experienced only eight recession since 1959, providing relatively few observations for model estimation. Further, when performing forecasting using a rolling sample window, forecasts which occur earlier in the timeline have even fewer past observations with which to estimate a model. This creates a challenging situation for the LR model as it struggles to find a robust decision boundary with relatively few observations, while the NB model can leverage a specification of the variable distributions to more efficiently estimate the decision boundary.

Appendix 1 gives an overview and discussion of the proof of asymptotic properties for these two models. This asymptotic analysis and the corresponding proof assumes

[5]For example, see Ng and Jordan (2002)

certain distributional conditions of the data which may not always hold. While we point out that the data set we use empirically violates these assumptions, we also show that in practice the asymptotic convergence of NB is superior to LR, as expected from the theoretical results. This result is consistent with Domingos and Pazzani (1997) and Zhang (2004) that show violating the distributional assumptions may not adversely impact the performance of NB.

4 Data and Real-Time Recession Forecasts

In producing recession forecasts, we assess the value of using a broad set of data compared to a narrower set, while also evaluating the benefit of using a rich lag structure. We also compare forecasting performance using real-time and revised data.

4.1 Data

We begin with 135 macroeconomic variables, at monthly intervals, recorded in the FRED-MD data set.[6] The observation period is from January 1959 through June 2016. An overview of variables are give in McCracken and Ng (2015). In most versions of the model, however, we use the following "core" set of four variables from these 135 variables (FRED series identifier is in parentheses):

1. ISM Manufacturing: Production Index (NAPMPI)
2. Total non-farm payroll growth (PAYEMS)
3. 10-year treasury rate minus Fed funds rate (T10YFFM)
4. S&P's Common Stock Price Index: Composite (SP500)

The raw values of all the variables are transformed according to the FRED-MD methods described in McCracken and Ng (2015). For the core set, the transformations are as follows: 1) NAPMPI: no transformation, 2) PAYEMS: first difference of consecutive values, 3) T10YFFM: no transformation, 4) S&P500: first difference of natural log of consecutive values.

[6] See McCracken and Ng (2015) for details of the data.

The FRED-MD data set is a useful resource for of macroeconomic data, but only provides revised data. As a result, evaluating the forecasting performance of various models using only this data does not accurately reflect the real-time information set at the time a prediction is made. To supplement this data, we use the Federal Reserve Bank of Philadelphia's Real-time Data Set for Macroeconomists for real-time values of total non-farm payrolls in the core set of four variables. The other three variables in this core set are not revised. Using the unrevised data allows us to produce forecasts using the information set available at the time a forecast would have been originally made.

To identify business cycle turning points, we use the recession dates identified by the National Bureau of Economic Research's Business Cycle Dating Committee (BCDC). Although business cycle classification is not entirely straightforward, as Berge and Jordá (2011) discuss, the NBER chronology is still commonly viewed as the gold standard. The intention of the BCDC, however, is not to be timely in calling the onset of a recession, but instead to compile an accurate historical time series of business cycle durations and turning points. Additionally, we construct forecasts to indicate if a recession will *begin* within some number of months, and not simply if the economy will be in a recession at some number of months in the future. This detail is subtle but important, as historically there have been a few NBER recession periods shorter than the longest horizon we will attempt to forecast (i.e. 12 months). Consequently, we follow Wright (2006) and define R_{t+h} to equal unity if a recession exists at any time from t until $t+h$, inclusively.

4.2 Incorporating Lags into Naïve Bayes

To incorporate a broad set of variables and their lags, we rewrite the conditional density from the right hand side in (1) as

$$f\left(\mathbf{x}_t|R_{t+j}\right)=\frac{1}{\sqrt{(2\pi)^m\left|\boldsymbol{\Sigma}\right|}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_t-\mu\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_t-\mu\right)'\right), \tag{12}$$

where all off-diagonal elements of the $m\times m$ matrix $\boldsymbol{\Sigma}$ are zero. When using all 135 variables with no lags, $\mathbf{x}_t=(x_{1,t},x_{2,t},...,x_{m,t})$ where $m=135$. Incorporating lags when using the full set of 135 variables is an option, though poses practical difficulties as the number of parameters grows exponentially and quickly exceeds the number of total observations in the data set. For this reason, we do not incorporate lags when using all 135 observations.

When using the four-variable core set of variables, however, incorporating lags is straightforward. Define the core set as $\widetilde{\mathbf{x}}_t$, then use $\mathbf{x}_t = (\widetilde{\mathbf{x}}_t, \widetilde{\mathbf{x}}_{t-1}, ..., \widetilde{\mathbf{x}}_{t-k})$ in (12), where k denotes lag length. Throughout the remainder of the paper, we will set $k = 10$ and be clear when using either real-time or revised data when reporting results using the core set of variables. The predictive value of NAPMPI and PAYEMS variables peaks at $k = 0$, then descends smoothly to near zero at $k = 10$. For the other two variables, the $k = 0$ data actually have limited predictive value, though rises and peaks at $k = 4$ for T10YFFM and at $k = 6$ for SP500, while retaining some predictive value out to $k = 10$. Extending beyond 10 lags would provide more predictive value from the T10YFFM and SP500 variables, but introduce extraneous variables and noise for NAPMPI and PAYEMS.

4.3 Real-Time Recession Forecasts

To evaluate forecasting performance, we replicate as closely as possible the information set of a forecaster in real time at each point in the sample. We use an expanding window starting in January 1973, which allows the initial set of parameter estimates to be based on the first two recessions after 1959. After the initial forecast in January 1973, we advance one month to generate forecasts for February and again use observations beginning in 1959. This process continues sequentially through the set of monthly observations. Forecast horizons, h, are set to $\{0, 3, 6, 9, 12\}$ months ahead.

When estimating the models, we do not include all observations up to the month prior to when the forecast is made. The reason is because NBER recessions are retroactively determined, so information about the business cycle state at time t would not have been available to update the parameter estimates at t. For example, the 2001 recession is identified to have started in March 2001, though the NBER did not make the announcement of a business cycle peak until November 26, 2001. Consequently, we would be unable to update the model between these dates under an assumption on the business cycle state, since that state was unknown during this period. As a result, using this data to update parameter estimates amounts to using information from the future, and so contaminates the real-time nature of the forecasting exercise. To address this issue, we introduce a "blind" of 18 months, so that we do not use data from the most recent 18 months to update the parameter estimates. Of course, we use real-time macroeconomic and financial data to update the recession probabilities for each month and over the various horizons, it is only the parameter estimates that do not incorporate data from the previous 18 months.

The recession probability estimates from the MS-NB model for forecasts over horizons $h = \{0, 3, 6, 9, 12\}$ months ahead are shown in Figure 1. The model uses real-time values of the core variables with 10 lags, including contemporaneous values, for a total of 44 variables. Note that forecasts at the different horizons for the same time period were made at different points in time. For example, the $h = 1$ horizon forecast for time period t was made in time period $t - 1$, while the $h = 6$ horizon forecast for t was made in period $t - 6$.

Figure 2 displays results for the MS-NB, standard NB, and LR models in a more consolidated format, where recession probabilities are not represented using a separate axis, but rather as a color coding. The heatmap figures can be read similarly to the separated line plots: the x-axis corresponds to date, and the y-axis consists of several forecast "bands" each corresponding to a different horizon. The bottom band is the current quarter forecast, ranging to the twelve-month ahead forecast at the top. Vertical dotted lines demarcate the NBER-defined recession periods. Color intensity in the forecast bands corresponds to recession probability, ranging from white (zero probability) to yellow (low probability) to orange (elevated probability) to red (highest probability). For each model and forecast horizon, we again use real-time values of the core variable set with 10 lags.

The forecast probability figures reveal several qualitative differences in the predictive performance of NB and MS-NB over LR. First, at all forecast horizons, the MS-NB forecasts exhibit the characteristic persistence in recession/expansion periods, whereas the LR forecasts do not. This persistence is due in part to the modified priors introduced exactly for this purpose, however the results for the standard NB show a tendency toward state persistence as well. In contrast, the results for LR, particularly at longer forecast horizons, fluctuate between high and low probability forecasts during even short periods of time.

For MS-NB and NB, forecasts of a recession are sometimes too early, other times too late, but a recession is never forecasted to start in the middle of a long expansion period, and in only a handful of cases is an expansion forecasted in the middle of a recession period. In contrast, LR does make these errors.

When comparing the MS-NB and standard NB forecasts, the effect of the modified priors is evident. The weaker signals at the leading and lagging edges of several recession periods are sharpened and several low probability predictions during expansion periods are flattened toward zero. Also, while some MS-NB forecasts show elevated probabilities for non-recession periods, these probabilities usually stay below 0.5. The exception is that sometimes a recession is forecast a few periods too early or too late, which is more likely to occur for longer horizon forecasts.

Finally, the forecast quality of LR begins to break down substantially at longer forecast horizons, providing much weaker signals, or missing recession periods entirely by failing to render a forecast probability above 0.5 at any time during the recession period.

5 Evaluating Recession Forecasting Performance

This section presents the results of an empirical evaluation of the asymptotic properties of NB compared to LR. This includes the variation in prediction errors as a function of the data used in model estimation, as well as a visualization and analysis of variation in the decision boundaries with respect to random samplings of data. We also present the results on rolling forecast experiments which aim to replicate the process of real world recession forecasting at multiple horizons as closely as possible.

5.1 Evaluation Criteria

Evaluating binary predictions differs in some respects from evaluating predictions from regression models for continuous variables. For binary outcomes, a prediction is either correct or incorrect, and therefore evaluating those predictions using a zero-one loss can be more informative than a real-valued loss. Still, a real-valued error criteria, such as mean-absolute error, will appropriately give more credit to a model that assigns, say, a 0.3 to a recession than a model assigning 0.1 if one materializes. Consequently, we consider the various models under both a zero-one loss and mean-absolute error.

In terms of definitions, take $\hat{y} = \hat{f}(\mathbf{x})$ as the binary prediction of classifier $\hat{f}$ for observation $\mathbf{x}$, with y as the true binary class for $\mathbf{x}$. The zero-one loss L for a set of observations $\{\mathbf{x}_i : i = 1...n\}$ is then defined as

$$L = \frac{1}{n}\sum_{i=1}^{n} I[\hat{y}_i \neq y_i], \tag{13}$$

where I is the indicator function that returns unity when the condition is true and zero otherwise. While MS-NB, NB and LR models all yield probabilities for their outcome predictions, these probabilities can be mapped into a binary space. For LR, the mapping is as follows

$$\hat{y}_i^{LR} = \begin{cases} 1, & P(R|\mathbf{x}_i) >= 0.5 \\ 0, & P(R|\mathbf{x}_i) < 0.5 \end{cases}. \tag{14}$$

and zero-one the mapping for NB (and MS-NB) is

$$\hat{y}_i^{NB} = \begin{cases} 1, & P(R|\mathbf{x}_i) >= P(E|\mathbf{x}_i) \\ 0, & P(E|\mathbf{x}_i) > P(R|\mathbf{x}_i) \end{cases}. \tag{15}$$

If focusing on the zero-one loss, the best classifier may be viewed as the one that maximizes accuracy, defined as

$$Accuracy = 1 - L = \frac{1}{n}\sum_{i=1}^{n} I[\hat{y}_i = y_i]. \tag{16}$$

For many classification problems, however, this criteria is misleading. Consider a classification problem where the number of observations belonging to each class $k \in \{0, 1\}$ are unbalanced, so $\sum_{i=1}^{n} I[y_i = 0] > \sum_{i=1}^{n} I[y_i = 1]$. Assume an extreme case where $\frac{1}{n}\sum_{i=1}^{n} I[y_i = 0] = 0.9$ and $\frac{1}{n}\sum_{i=1}^{n} I[y_i = 1] = 0.1$. In this example, a trivial classifier $\hat{f}(\mathbf{x}) = 0$ would yield $L = 0.1$ and $Accuracy = 0.9$. On the surface, a 90% prediction accuracy seems favorable, but we know that this trivial classifier has no predictive value as it gives all observations the same classification no matter what. To address this issue and better assess the performance of a classifier, we also evaluate *precision* and *recall*. Precision tells us the fraction of positive predictions which are truly positive, and recall tells us the fraction of truly positive observations that were predicted positive. Let true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) be defined as

$$TP = \sum_{i=1}^{n} I[\hat{y}_i = 1] * I[y_i = 1], \tag{17}$$

$$FP = \sum_{i=1}^{n} I[\hat{y}_i = 1] * I[y_i = 0], \tag{18}$$

$$TN = \sum_{i=1}^{n} I[\hat{y}_i = 0] * I[y_i = 0], \tag{19}$$

$$FN = \sum_{i=1}^{n} I[\hat{y}_i = 0] * I[y_i = 1]. \tag{20}$$

Precision and recall are then given by

$$Precision = \frac{TP}{TP + FP}, \tag{21}$$

$$Recall = \frac{TP}{TP + FN}. \tag{22}$$

Taking the harmonic mean of precision and recall yields the F-measure, also referred to as the F-score or F_1-measure, as follows

$$F = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{23}$$

Thus, $F = 1$ when $FP = 0$ and $FN = 0$, and $F = 0$ when $TP = 0$. In the illustrative example of the unbalanced panel and trivial classifier, there are no true positives because every prediction was for 0 - so $F = 0$, which better reflects the trivial classifier as having no predictive ability.[7]

In terms of real-valued criteria, we use mean absolute error, defined as

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |P(R|\mathbf{x}_i) - y_i|. \tag{25}$$

[7] The F-measure can be generalized to other weightings of precision and recall as follows

$$F_\beta = (1 + \beta^2) \frac{Precision * Recall}{\beta^2 * Precision + Recall}, \tag{24}$$

though we found $\beta = 1$ is adequate in the case of evaluating business cycle data.

5.2 Timing-dependent Error Weights

Both F-measure and MAE treat each prediction separately and as uniformly important. In a real-time forecasting context, however, errors may not be viewed equally. For example, predicting a recession to begin in what turns out to be the middle of an expansion may be viewed as considerably worse than predicting a recession a few months prior to when a recession actually begins. A uniform weighting metric makes no distinction between these types of errors. To address this issue, we construct timing-dependent error weights. This weighting scheme penalizes prediction errors differently depending on where they occur in the business cycle, where errors in the middle of cycles are penalized more than errors near the edge of cycle transitions.

Let t_{start} be the first period in any expansion/recession cycle, t_{end} be the last period in that same cycle, and t_{mid} be the cycle midpoint. We fit a monotonically non-decreasing function to the range $\{t_{start}, ..., t_{mid}\}$, and a symmetric, monotonically non-increasing function to $\{t_{mid}, ..., t_{end}\}$. The choice of functions can be made in a wide variety of ways, but we have chosen to use a truncated Gaussian curve with several modifications.

We first fit a Gaussian function with μ equal to t_{mid} and σ equal to the number of periods in $\{t_{start}, ..., t_{end}\}$. Let $s = \sum_{t_{start}}^{t_{end}} 1$ be the number of periods in the cycle delimited by t_{start} and t_{end}. Let w_i be the weight at t_i, $t_{start} \leq t_i \leq t_{end}$, then

$$w_i = \frac{1}{\sqrt{2s^2\pi}} e^{-\frac{(t_i - t_{mid})^2}{2s^2}}. \tag{26}$$

Next, we apply an initial scaling to w_i such that that the error weights at cycle transitions are zero and error weights at cycle midpoints are one. Let $min_w = min(\{w_{start}, ..., w_{end}\})$ and $max_w = max(\{w_{start}, ..., w_{end}\})$, then

$$w_i^{scaled} = \frac{w_i - min_w}{max_w - min_w}. \tag{27}$$

For the terminal cycles at the beginning and ending of the sample, we do not have complete information about starting/ending periods of these cycles. So in the final cycle, we replace all weights after the midpoint with the weight at the midpoint, and likewise for the first cycle, we replace all weights prior to the midpoint with the weight at the midpoint. Finally, we apply another normalization to ensure that across the entire range of a cycle, the total mass of error weights are the same as in the uniformly weighted case, and therefore total error penalization during a cycle does not change based on cycle length. These final weights are given as

$$w_i^{norm} = \frac{s * w_i^{scaled}}{\sum_{j=start}^{end} w_j^{scaled}}, \tag{28}$$

where $\sum_{i=start}^{end} w_i^{norm} = s$. Figure 3 shows a plot of the timing-dependent weights across the sample period. When computing timing-dependent errors for F-measure and MAE, the forecast errors in each period are multiplied by the corresponding error weight.

5.3 Recession Classification Performance

Tables 1-6 show quantitative results for various forecasting methods, horizons, and evaluation measures. Tables 1 and 3 show the mean absolute error and F-measure when using a uniform error penalty. Tables 4 and 6 show the same results using a timing-dependent error penalty as discussed in the previous section. Each table compares forecasts from MS-NB and LR, as well as those from the Survey of Professional Forecasters (SPF), using the average of all reported estimates.[8] All three methods are evaluated at 5 forecast horizons (current, 3-, 6-, 9- and 12-month ahead). Results are further divided by the variable set used in the estimation, with four options: 1) the core four-variable set only, 2) term-spread only, 3) the Conference Board's Leading Economic Indicator (LEI) variable and 4) all 135 variables. Comparisons are also performed across revised and real-time data for the core variable set. In each of the tables, the best result for each forecast horizon and variable set/real time combination is highlighted in bold.

Table 1 shows based on MAE under uniform weights, MS-NB generally outperforms LR. In some cases, LR outperforms for current and 3-month forecasts, though the MS-NB has better performance in all cases for forecasts 6-month ahead and longer. MS-NB is better at all future horizons (i.e. 3-months and longer) using real-time values for the core set of variables, which is perhaps most relevant from a forecasting perspective.

The values in Table 1 show which model has the lowest MAE, though does not

[8]The SPF has regularly asked professional forecasters for about 40 years their assessment of whether real GDP will decline in the quarter the survey is taken, as well as in the subsequent four quarters. In this section, we compare these prediction from those of the various models, similar to the approach of Rudebusch and Williams (2009).

indicate whether the differences are statistically significant. To assess whether the performance across models is statistically significant, Table 7 shows p-values using a two-sided Diebold-Mariano test for equality of forecast errors. The table also includes the standard NB model for comparison.[9] Note that each box corresponds to a particular forecast horizon and then each element shows the p-value from testing one model against another. For example, the 3x2 element in the third box (i.e. Horizon = 6 mo) is .02, indicating that null hypothesis that forecasts errors from the MS-NB are equal to those from LR can be rejected at the .05 level of significance. All p-values testing equality of forecasts errors from the SPF survey against each econometric model can be rejected at the .01 significance level, indicating that they all outperform the SPF forecasts. The last row of each box, which corresponds to a particular forecast horizon, shows that for 6-months and longer, the NB and MS-NB models significantly outperform the LR model.

Table 3 shows the results using the F-measure and uniform weights, where again the MS-NB model generally outperforms the LR model, particularly at longer horizons. Table 8 shows the zero-one loss differences are statistically significant at the 12-month horizon, but not as significant at the shorter horizons. In addition, the differences between the MS-NB and NB are not statistically significant.

Tables 4 and 6 show across all forecast horizons, MS-NB always outperforms LR when using the core variable set with real-time data and timing-dependent weights. MS-NB also outperforms LR in nearly every other case as well. In terms of significance, Tables 9 and 10 show MS-NB and NB outperform LR. In addition, the MS-NB significantly outperforms NB at many horizons using MAE under the time-varying weight.

Additionally, tables 2 and 5 show the results for an alternative performance measure which also uses a zero-one loss, the Area Under receiver operating characteristic Curve, or AUC. The AUC measure has been advocated by Berge and Jordá (2011) and an empirical distribution for AUC has been described in Liu and Moench (2016). The AUC is constructed by varying the zero-one loss threshold over a range, and plotting the corresponding receiver operating characteristic (ROC) curve, then summing the area under the ROC curve. The ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis, for each of the zero-one loss thresholds. Having an empirical distribution for the AUC makes it statistically more attractive than the F-measure, however as a realistic performance measure it is more

[9]SeeDiebold and Mariano (1995) for details.

cumbersome because of the need to decide on a specific zero-one loss threshold for the out-of-sample predictions. Careful attention must be paid to selecting the best threshold using only in-sample data, as using the best threshold with respect to out-of-sample data will lead to overfitting. Overall results for AUC performance largely mirrors the overall results for F-measure.

Finally, Tables 11 and 12 present F-measure and MAE results, with timing-dependent weightings, for MS-NB with different numbers of variable lags. We report results for lag numbers in the set $\{1, 3, 6, 10\}$. These results indicate that a larger number of lags is preferable for near and mid-term forecasts, but performance can improve at 9- and 12- month horizons with fewer variable lags. It is possible that as the horizon increases, the extra variable lags lose much of their predictive ability and turn into noise which obscures the predictive ability of the remaining variables.

5.4 Empirical Evaluation of Asymptotic Properties

Despite the theoretical equivalence of NB and LR under certain distributional assumptions about the data, our results illustrate that they do not perform equivalently in practice, as the underlying assumptions are often violated.[10] Despite the violation of these assumptions, our results demonstrate the advantage of NB for business cycle data. In particular, the NB model converges to its asymptotic error rate more quickly than LR, a feature most apparent when forecasting recessions at longer forecast horizons. This is an important aspect of the comparison in a recession forecasting context, as there are relatively few past recessions with which to estimate a decision boundary.

To illustrate, Figure 4 plots the predictive ability in F-measure for the NB and LR model as a function of the fraction of the data used in estimation. Each row shows a different forecast horizon, with the top two charts showing contemporaneous predictions and the bottom showing 12-month ahead predictions. To generate the subsamples, data is sampled at equidistant intervals so that the entire sample space is evenly covered. The subsample includes current observations and the relevant lags. The x-axis of each plot shows the amount of data used to estimate each model, starting at 5% and increasing in 1 percentage point increments. On the y-axis, the F-measures are based on predictions for the entire data set, so include both in-sample

[10] In this section, we focus on NB, rather than MS-NB, since the theoretical equivalence to LR is with respect to the NB classifier.

and out-of-sample predictions. A larger value for the F-measure indicates greater predictive ability. When using 100% of the data, all predictions become in-sample predictions. Plots in the left column show raw results, while plots in the right column are smoothed for easier interpretation. Real time data is used for these experiments, and the plots are smoothed using a centered 5-period moving average window.

In the top row showing F-measures for current period estimates, the asymptotic differences between models are clearly evident. The predictive performance of NB plateaus before LR, while LR performs better using larger subsamples. However, as an illustration of how quickly NB converges to its asymptotic limit, note that it reaches an F-measure of 0.6 using only 5% of the data and converges to its maximal F-measure of nearly 0.8 with less than 10% of the data. By 20% of data, LR has caught up and begins to surpass NB. This advantage, however, diminishes for forecasts at longer horizons. For the 3-month ahead forecasts, the LR model requires over 65% of the data to reach the performance of NB and at the 6-month ahead horizon, it requires nearly all of the data to perform at parity. Further ahead, the 9- and 12-month ahead forecasts from the LR model are unable to consistently reach the predictive performance of NB regardless of the amount of data used in estimation.

5.5 Visualizing Data and Decision Boundaries

While the results above provide an empirical summary of NB forecasting performance relative to the LR model, they do not provide a reason why NB performs so much better at longer forecast horizons. To provide some intuition, Figure 5 highlights differences in the decision boundaries for the two models. Each chart shows a scatter plot of the entire data set after compressing the original 44 variables (i.e. 4 contemporaneous values, plus 10 lags) into two dimensions using principal component analysis.

In Figure 5, the chart on the left shows the NB decision boundary using the entire data sample (i.e. the red curved line), along with contours of the conditional distributions of recession/expansion periods. The decision boundary is the set of points where $P(R|\mathbf{x}) = P(E|\mathbf{x})$. For NB, the the class distributions do not have equal variance, which creates curvature in the boundary. The chart on the right shows the decision boundary for the LR model using the entire data set (i.e. thick blue line).

For both models, a point lying above the decision boundary is predicted as an expansion period, and any point (i.e. "+") lying below the decision boundary is predicted as a recession period. Next, we randomly sample a 50% subset of data points 20 times, and construct decision boundaries using these subsamples and plot the decision boundaries (i.e. the thin black lines). This procedure generates a distribution of decision boundaries. An important aspect of the boundaries is that the region where the data density is the highest, near the sample centroid, the deviation of the NB boundaries is smaller compared to the LR model. Away from the centroid, the boundaries for the NB model show more variation, but this is where the data density is much lower, so has little impact on overall classification accuracy. The next section examines these properties quantitatively.

5.6 Analysis of Decision Boundaries

To give a quantitative assessment of decision boundary properties, Figure 6 shows a plot of the mean deviation in decision boundaries using random 50% samples of the data, along with the local data densities along the boundaries. The x-axis in Figure 6 denotes the position along the decision boundary, with 0 corresponding to the lowest point on the boundary in Figure 5. The left y-axis is the mean deviation in decision boundaries at each position, while the right y-axis reflects the local data density along the boundaries.

More precisely, let $B = \{\mathbf{b}_1, ..., \mathbf{b}_g\}$ be a set of g points spaced at regular interval along a decision boundary. Let B^{NB} and B^{LR} be the sets of points along the decision boundaries for NB and LR using 100% of the data. Let $\hat{B}_1^{NB}, ..., \hat{B}_{20}^{NB}$ and $\hat{B}_1^{LR}, ..., \hat{B}_{20}^{LR}$ be the decision boundaries estimated for the 20 different 50% random samples. Let $\hat{B}_j^{NB} = \{\hat{\mathbf{b}}_{1,j}^{NB}, ..., \hat{\mathbf{b}}_{g,j}^{NB}\}$ be the set of g points along boundary $\hat{B}_j^{NB}$, and likewise let $\hat{B}_j^{LR} = \{\hat{\mathbf{b}}_{1,j}^{LR}, ..., \hat{\mathbf{b}}_{g,j}^{LR}\}$ be the set of points along $\hat{B}_j^{LR}$. Let $D_B(\hat{B}_1, ..., \hat{B}_{20})$ be the set of mean distances between a boundary B that uses 100% of the data and the boundaries $\hat{B}_1, ..., \hat{B}_{20}$ that use 50% of the data, given by

$$D_B(\hat{B}_1, ..., \hat{B}_{20}) = \{\frac{1}{20}\sum_{j=1}^{20} ||\mathbf{b}_i - \hat{\mathbf{b}}_{i,j}|| : i = 1, ..., g\}. \tag{29}$$

Figure 6 shows the set of mean deviations for the g points along the decision boundaries. For NB, the set $D_B^{NB}(\hat{B}_1^{NB}, ..., \hat{B}_{20}^{NB})$ is the thick red line and for LR, $D_B^{LR}(\hat{B}_1^{LR}, ...\hat{B}_{20}^{LR})$ is the thick blue line. This figure shows that as we move along the decision boundaries, the decision boundary of the NB model is more stable around the center of the data distribution, but is less stable near the tails.

This feature could be advantageous or not, depending on the data distribution. To illustrate that indeed the density of the data is greater near the region of the boundary that is most stable, let $S(B, r)$ be the set of g local data densities along boundary B, where the local density is the fraction of data points within a radius r of boundary point b_i

$$S(B, r) = \{\frac{1}{n}\sum_{j=1}^{n} I[||\mathbf{x}_j - \mathbf{b}_i|| < r] : i = 1, ..., g\} \tag{30}$$

The sets of data densities along the NB and LR decision boundaries are from the full set of data. Figure 6 then shows the densities $S(B^{NB}, 20)$ and $S(B^{LR}, 20)$, normalized to unit interval, as a thin red line for NB and a thin blue line for LR.

The key observation in Figure 6 is that the boundaries are more stable where the data density is the highest, particularly for NB. The implication is that new data points are much less likely to be misclassified by NB compared to LR, an advantage confirmed by the empirical results in the previous section.

6 Conclusions

The NB model for business cycle turning point forecasting outperforms LR in almost every situation and has provided relatively clear recessionary signals up to 12 months in advance. One reason, as we highlight, is that NB converges to its asymptotic error rate faster than LR and given the limited sample for business cycles, is advantageous for forecasting recessions. While NB can easily incorporate a large amount of data, the results in this paper suggest a more limited set of variables with a rich lag structure performs best. We chose the four variables to align with previous work, though future work can develop methods to select variables and the number of lags under an optimality criteria. In addition, we provide a time-varying weight to forecasting errors, such as penalizing recessionary signals that occur in the middle of an expansion.

A Asymptotic Comparison of NB and LR

Here we give an overview of the proof for the asymptotic properties of NB and LR which appears in Ng and Jordan (2002). The asymptotic comparison makes two key points:

1. The asymptotic error of LR is less than or equal to the asymptotic error of NB.
2. The error rate of NB converges to the asymptotic error rate in order *logm* samples, where m is the number of variables, while LR converges linearly (order m).

The first point is straightforward to show, relying on the result that the error of LR converges to that of the best linear model, and therefore must be no worse than NB. Said another way, as the amount of observations approach infinity, LR will learn a linear decision boundary that gives the best possible error rate. NB is not guaranteed to so, and so the error rate of LR can never be asymptotically worse than NB.

To show this first point, Ng and Jordan (2002) prove two propositions: 1) the asymptotic error of the LR model is less than or equal to that of the NB model, because the error rate of LR converges to the error of the best linear model (assuming finite VC dimension)Vapnik (2000), then it must be asymptotically no worse than that of NB; and 2) the result of applying Vapnik's uniform convergence bounds to LR is that the number of examples needed to approach the asymptotic error is $O(m)$.

The second point is more complex and contains statements about the rate of asymptotic convergence for both NB and LR, which are independently proved. The convergence rate of LR can be analyzed by applying Vapnik's uniform convergence bounds along with the concept of *VC dimension*.[11] VC dimension characterizes the complexity or discriminative power of a model. The VC dimension for a model f parameterized with θ is the maximum number of observations $(\mathbf{x}_1, ..., \mathbf{x}_n)$ in some variable space with dimensionality m such that the points can be arranged and θ selected so that f can correctly classify all $(x_1, ..., x_n)$ for every possible binary label set $(y_i, ..., y_n)$. The VC dimension of a linear model is $m + 1$. The VC dimension is

[11] From **V**apnik-**C**hervonenkis theory described in detail in *Vapnik, Vladimir N (2000). The Nature of Statistical Learning Theory. Information Science and Statistics. Springer-Verlag. ISBN 978-0-387-98780-4.*

used to provide a probabilistic upper bound on the out-of-sample classification error of f. For linear models (including LR) with VC dimension $m+1$, the number of data observations required to approach the asymptotic error rate is at most order m.

The proof that NB has logarithmic asymptotic convergence is given in two steps: 1) the distribution parameters of an estimated NB model are shown to converge logarithmically in m (in probability) to the asymptotic parameters for the true data distribution, where m is the number of variables; and 2) when the distribution parameters of an estimated NB model are close to the true parameters of the distribution, the error of the estimated model is close to the error obtained when using the true parameters. The first step is proven by applying Chernoff bounds to the difference between estimated parameters and true parameters, which gives a bound on the probability that a random variable (the difference between estimated and true parameters, in our case) exceeds some value.

The second step, that convergence in parameters implies convergence in the predictions and therefore error, can be shown as follows. First, observe that that the decision boundary corresponding to the estimated parameters will give the same prediction for any sample x as a decision boundary using the true parameters as long as the x is on the same side of both boundaries. Given that the difference between the estimated decision boundary and the true decision boundary is proportional to the difference between estimated and true parameters, when the deviation of the estimated boundary is small compared to the expected shortest distance between any x and the true decision boundary, it is unlikely that the estimated model will render a prediction different from when using the asymptotic parameters.

Establishing bounds on the error rate as a function of the deviation in estimated decision boundary is key, and can be done loosely with the Chebyshev inequality, or more strongly (assuming variable independence) using the Chernoff bound again. This concludes our overview of the asymptotic comparison between NB and LR.

References

Ang, Andrew, Monika Piazzesi and Min Wei. 2006. "What does the yield curve tell us about GDP growth?" *Journal of Econometrics* 131:359–403.

Berge, Travis J. 2015. "Predicting Recessions with Leading Indicators: Model Averaging and Selection over the Business Cycle." *Journal of Forecasting* 34:455–471.

Berge, Travis J. and Óscar Jordá. 2011. "Evaluating the Classification of Economic Activity into Recessions and Expansions." *American Economic Journal: Macroeconomics* 3(2):246–277.

Chauvet, Marcelle and James D. Hamilton. 2006. Dating Business Cycle Turning Points. In *Nonlinear Time Series Analysis of Business Cycles*, ed. Costas Milas, Phillip Rothman and Dick van Dijk. Amsterdam: North-Holland.

Chauvet, Marcelle and Simon Potter. 2002. "Predicting a Recession: Evidence from the yield curve in the presence of structural breaks." *Economic Letters* 77:245–253.

Chauvet, Marcelle and Simon Potter. 2005. "Forecasting Recessions Using the Yield Curve." *Journal of Forecasting* 24:77–103.

Chen, Zhihong, Azhar Iqbal and Huiwen Lai. 2011. "Forecasting the Probability of US Recessions: A Probit and Dynamic Factor Modelling Approach." *Canadian Journal of Economics* 44(2):651–672.

Davig, Troy. 2008. "Detecting Recession in the Great Moderation: A Real-Time Analysis." *Economic Review* pp. 5–33. Federal Reserve Bank of Kansas City.

Diebold, Francis X. and Glenn D. Rudebusch. 1989. "Scoring the Leading Indicators." *Journal of Business* 62(3):369–391.

Diebold, Francis X. and Glenn D. Rudebusch. 1991. Turning point prediction with the composite leading index: An ex ante analysis. In *Leading Economic Indicators: New Approaches and Forecasting Records*, ed. Kajal Lahiri and Geoffrey H. Moore. Cambridge: Cambridge Univesity Press pp. 231–256.

Diebold, Francis X. and Roberto Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 13:253–263.

Domingos, Pedro and Michael Pazzani. 1997. "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss." *Machine Learning* 29:103–130.

Estrella, Arturo and Frederic S. Mishkin. 1998. "Predicting US Recessions: Financial Variables as Leading Indicators." *The Review of Economics and Statistics* 80(1):45–61.

Estrella, Arturo and Gikas A. Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 46:555–576.

Giusto, Andrea and Jeremy Piger. 2014. "Identifying Business Cycle Turning Points in Real Time with Vector Quantization.". Working Paper.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57(2):357–384.

Hamilton, James D. 2011. "Calling Recessions in Real Time." *International Journal of Forecasting* 27(4):1006–1026.

Hamilton, James D. and Dong Heon Kim. 2002. "A Reexamination of the Predictability of Economic Activity Using the Yield Spread." *Journal of Money, Credit and Banking* 34(2):340–360.

Jefferson, Philip N. 1998. "Inference Using Qualitative and Quantitative Information with an Application to Monetary Policy." *Economic Inquiry* 36(1):108–119.

Kauppi, Heikki and Pentti Saikkonen. 2008. "Predicting U.S. Recessions with Dynamic Binary Response Models." *Review of Economics and Statistics* 90(4):340–360.

Liu, Weiling and Emanuel Moench. 2016. "What Predicts U.S. Recessions?" *International Journal of Forecasting* 32(4):1138–1150.

McCracken, Michael W. and Serena Ng. 2015. "FRED-MD: A Monthly Database for Macroeconomic Research.". Working Paper, 2015-012B.

Nalewaik, Jeremy J. 2011*a*. "Forecasting Recessions Using Stall Speeds.". Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System.

Nalewaik, Jeremy J. 2011*b*. "Incorporating Vintage Differences and Forecasts into Markov Switching Models." *International Journal of Forecasting* 27(2):281–307.

Nalewaik, Jeremy J. 2012. "Estimating Probabilities of Recession in Real Time Using GDP and GDI." *Journal of Money, Credit and Banking* 44(1):235–253.

Neftci, Salih N. 1982. "Optimal Prediction of Cyclical Downturns." *Journal of Economic Dynamics and Control* 4:225–241.

Ng, Andrew Y. and Michael I. Jordan. 2002. "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes.".

Ng, Serena. 2014. "Boosting Recessions." *Canadian Journal of Economics* 47(1):1–34.

Palash, Carl J. and Lawrence J. Radecki. 1985. "Using Monetary and Financial Variables to Predict Cyclical Downturns." *Federal Reserve Bank of New York Review* 10.

Rudebusch, Glenn D. and John C. Williams. 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business & Economic Statistics* 27(4):492–503.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.

Wright, Jonathan H. 2006. "The Yield Curve and Predicting Recessions.". Finance and Economics Discussion Series, Working Paper, 2006-07.

Zhang, Harry. 2004. "The Optimality of Naive Bayes.". American Association for Artificial Intelligence.

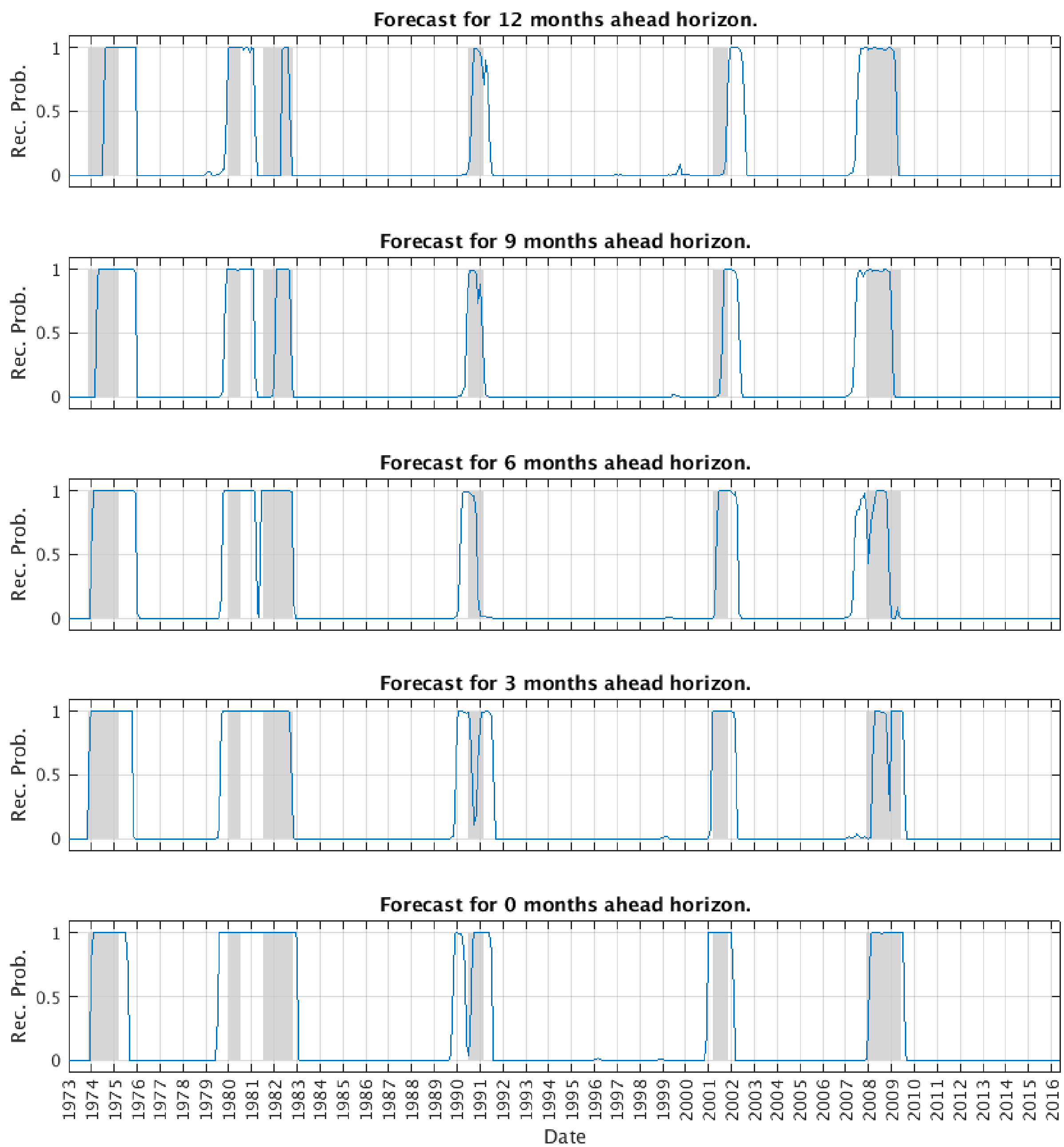

Figure 1: Recession probability forecasts by MS-NB at various horizons. Vertical dotted black lines indicate start/end of NBER recession periods. Models use real-time values of the four-variable core set.

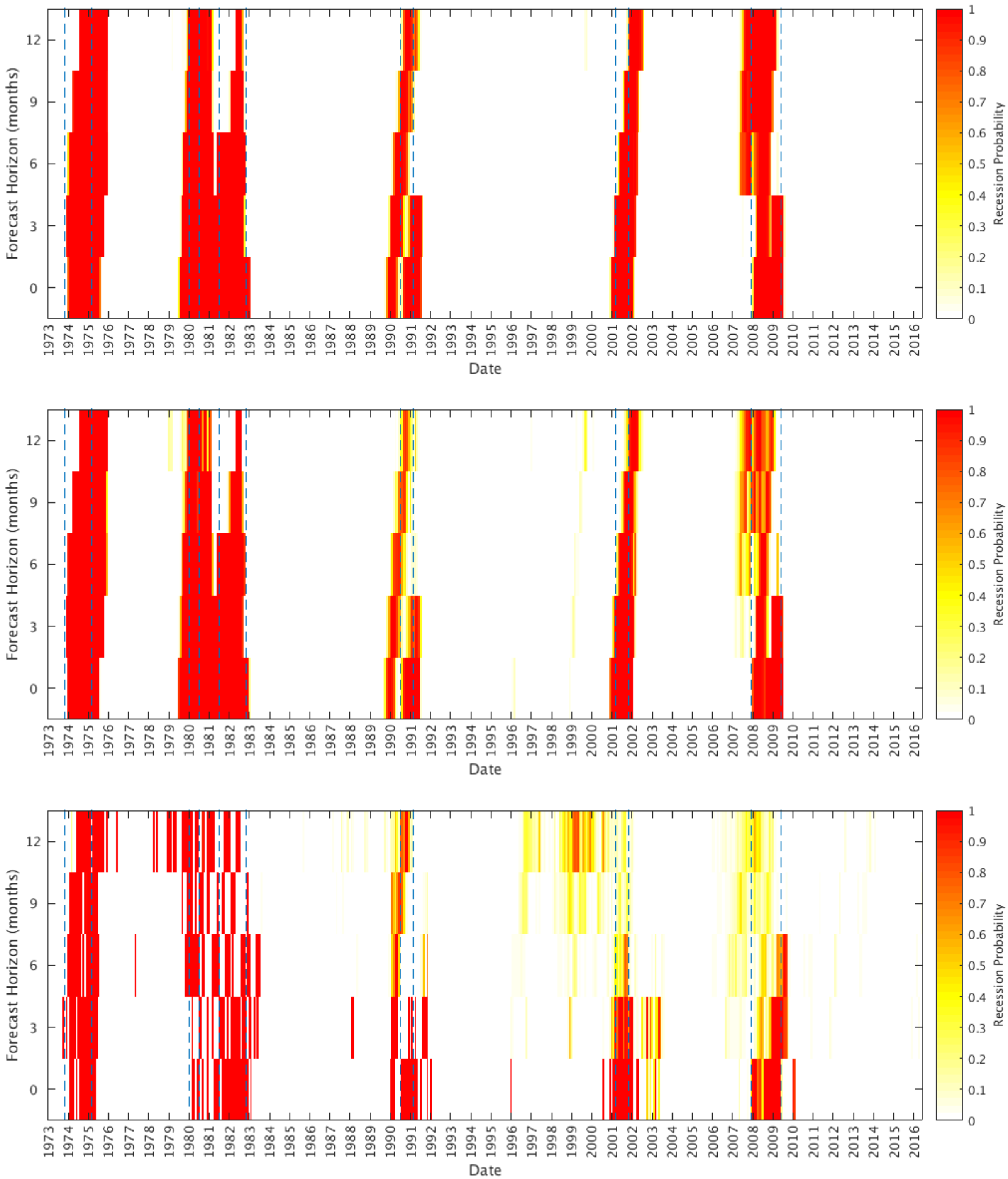

Figure 2: Recession probabilities forecasts of MS-NB (top), NB (middle), and LR (bottom) for different horizons. Vertical dotted black lines indicate start/end of NBER recession periods. Models use real-time values of the four-variable core set.

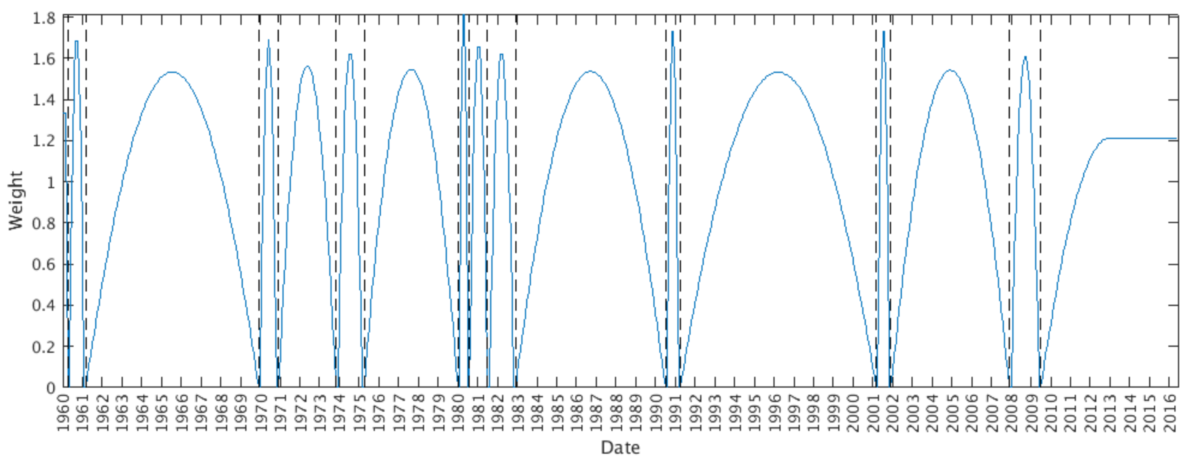

Figure 3: Timing-dependent error weights for predictions showing higher penalties for errors in the middle of expansion/recession cycles versus at the beginning/end of cycles. Vertical dashed lines indicate start/end of NBER-defined recessions.

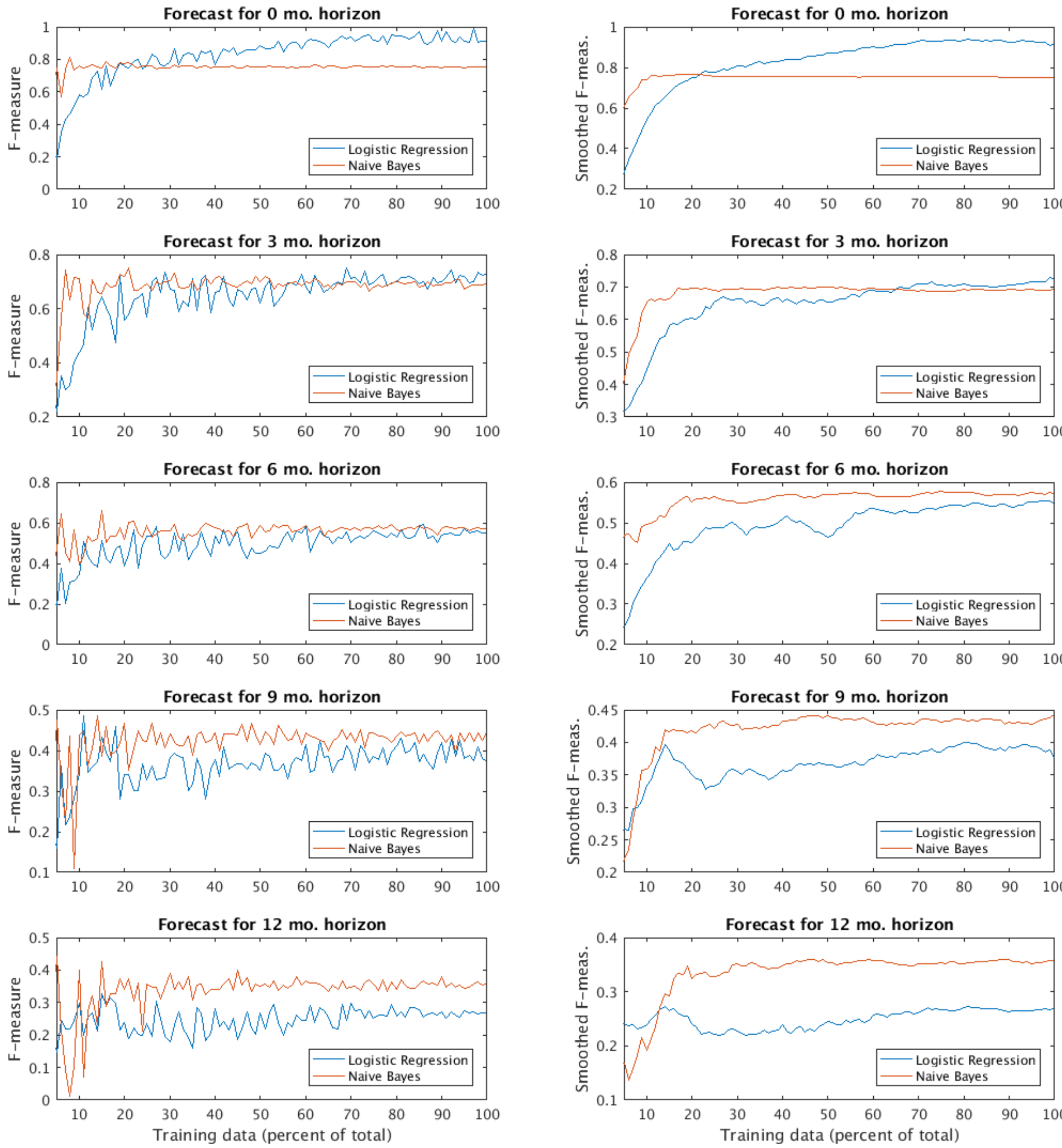

Figure 4: F-measure comparisons of NB and LR performance across different forecast horizons using real-time values of the core set of variables. Each model uses contemporaneous values, plus 10 lags, as explanatory variables.

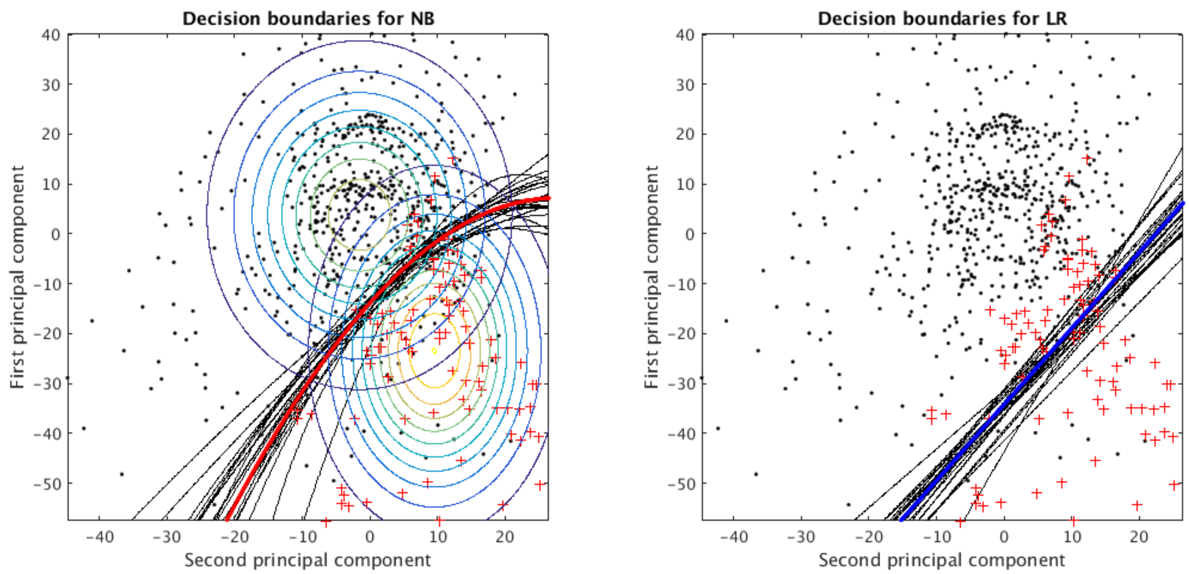

Figure 5: Comparison of decision boundaries from NB and LR.

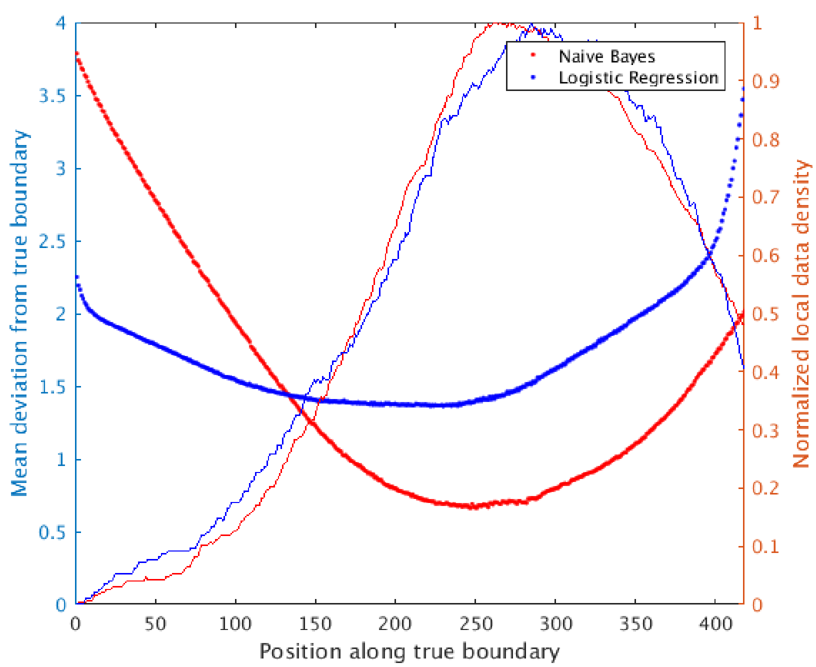

Figure 6: Comparison of data density and decision boundaries from NB and LR.

Table 1: Forecasting Results in mean absolute error for SPF, MS-NB, NB, and LR under various conditions.

MAE, uniform weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.160	0.211	0.239	0.253	0.254	0.224
MS-NB	Core	Yes	0.087	**0.099**	**0.107**	**0.103**	**0.115**	**0.102**
NB	Core	Yes	0.080	0.103	**0.107**	0.109	0.122	0.104
LR	Core	Yes	**0.079**	0.119	0.140	0.138	0.176	0.130
MS-NB	Core	No	0.092	0.105	**0.107**	**0.107**	**0.127**	**0.107**
LR	Core	No	**0.076**	**0.103**	0.129	0.138	0.168	0.123
MS-NB	Spread	No	0.150	**0.126**	**0.110**	**0.099**	**0.104**	**0.118**
LR	Spread	No	**0.144**	0.150	0.147	0.134	0.155	0.146
MS-NB	LEI	No	0.132	0.137	**0.137**	**0.186**	0.237	0.166
LR	LEI	No	**0.113**	**0.125**	0.154	0.202	**0.236**	0.166
MS-NB	All 135	No	**0.140**	**0.160**	**0.173**	**0.175**	**0.168**	**0.163**
LR	All 135	No	0.386	0.399	0.390	0.382	0.363	0.384

Table 2: Forecasting Results in AUC for SPF, MS-NB, NB, and LR under various conditions.

AUC, uniform weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.971	0.931	0.845	0.665	0.507	0.784
MS-NB	Core	Yes	0.967	**0.953**	**0.944**	**0.913**	**0.795**	**0.914**
NB	Core	Yes	**0.969**	0.951	0.940	0.909	0.789	0.911
LR	Core	Yes	0.917	0.789	0.711	0.690	0.692	0.760
MS-NB	Core	No	**0.969**	**0.956**	**0.944**	**0.891**	**0.773**	**0.906**
LR	Core	No	0.901	0.861	0.767	0.772	0.731	0.806
MS-NB	Spread	No	0.839	**0.876**	**0.908**	**0.930**	**0.921**	**0.895**
LR	Spread	No	**0.864**	0.842	0.845	0.816	0.789	0.831
MS-NB	LEI	No	**0.972**	0.960	**0.936**	**0.874**	**0.737**	**0.896**
LR	LEI	No	0.968	**0.961**	0.924	0.802	0.692	0.869
MS-NB	All 135	No	**0.746**	**0.738**	**0.757**	**0.712**	**0.651**	**0.721**
LR	All 135	No	0.566	0.539	0.553	0.509	0.518	0.537

Table 3: Forecasting Results in F-measure for SPF, MS-NB, NB, and LR under various conditions.

F-measure, uniform weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.672	0.440	0.145	0.000	0.000	0.251
MS-NB	Core	Yes	0.754	**0.713**	**0.658**	**0.634**	**0.574**	**0.667**
NB	Core	Yes	**0.777**	0.686	0.625	0.610	0.504	0.640
LR	Core	Yes	0.764	0.632	0.460	0.360	0.325	0.508
MS-NB	Core	No	0.746	**0.720**	**0.671**	**0.630**	**0.533**	**0.660**
LR	Core	No	**0.759**	0.653	0.496	0.442	0.339	0.538
MS-NB	Spread	No	**0.519**	**0.609**	**0.650**	**0.684**	**0.617**	**0.616**
LR	Spread	No	0.427	0.430	0.422	0.431	0.333	0.409
MS-NB	LEI	No	0.676	0.670	**0.663**	**0.537**	**0.398**	**0.589**
LR	LEI	No	**0.775**	**0.738**	0.635	0.496	0.291	0.587
MS-NB	All 135	No	**0.598**	**0.464**	**0.356**	**0.279**	**0.303**	**0.400**
LR	All 135	No	0.273	0.234	0.233	0.220	0.196	0.231

Table 4: Forecasting Results in mean absolute error for SPF, MS-NB, NB, and LR under various conditions.

MAE, timing-dependent error weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.135	0.190	0.223	0.244	0.247	0.208
MS-NB	Core	Yes	**0.034**	**0.047**	**0.054**	**0.057**	**0.070**	**0.052**
NB	Core	Yes	**0.034**	0.055	0.062	0.067	0.079	0.060
LR	Core	Yes	0.039	0.081	0.104	0.115	0.146	0.097
MS-NB	Core	No	0.035	**0.044**	**0.053**	**0.057**	**0.082**	**0.054**
LR	Core	No	**0.034**	0.073	0.098	0.120	0.138	0.092
MS-NB	Spread	No	**0.103**	**0.082**	**0.070**	**0.060**	**0.066**	**0.076**
LR	Spread	No	0.128	0.129	0.120	0.113	0.129	0.124
MS-NB	LEI	No	**0.062**	**0.059**	**0.053**	**0.088**	**0.150**	**0.082**
LR	LEI	No	0.064	0.077	0.107	0.157	0.197	0.120
MS-NB	All 135	No	**0.088**	**0.119**	**0.147**	**0.148**	**0.154**	**0.131**
LR	All 135	No	0.373	0.382	0.393	0.376	0.359	0.377

Table 5: Forecasting Results in AUC for SPF, MS-NB, NB, and LR under various conditions.

AUC, timing-dependent error weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.991	0.972	0.913	0.707	0.536	0.824
MS-NB	Core	Yes	**0.986**	**0.980**	**0.976**	**0.967**	**0.841**	**0.950**
NB	Core	Yes	**0.986**	0.979	0.973	0.966	0.840	0.949
LR	Core	Yes	0.956	0.804	0.738	0.708	0.767	0.795
MS-NB	Core	No	**0.988**	**0.981**	**0.976**	**0.955**	**0.814**	**0.943**
LR	Core	No	0.946	0.863	0.807	0.789	0.804	0.842
MS-NB	Spread	No	0.884	**0.928**	**0.955**	**0.965**	**0.958**	**0.938**
LR	Spread	No	**0.897**	0.890	0.888	0.844	0.864	0.877
MS-NB	LEI	No	**0.994**	0.989	**0.974**	**0.945**	**0.800**	**0.940**
LR	LEI	No	0.989	**0.990**	**0.974**	0.875	0.757	0.917
MS-NB	All 135	No	**0.759**	**0.762**	**0.803**	**0.796**	**0.689**	**0.762**
LR	All 135	No	0.576	0.569	0.541	0.523	0.534	0.549

Table 6: Forecasting Results in F-measure for SPF, MS-NB, NB, and LR under various conditions.

F-measure, timing-dependent error weights

Method	Variables	Real Time?	Cur. Mo.	3 Mo.	6 Mo.	9 Mo.	12 Mo.	Avg. All
SPF	N/A	Yes	0.757	0.454	0.149	0.000	0.000	0.272
MS-NB	Core	Yes	**0.893**	**0.848**	**0.812**	**0.786**	**0.697**	**0.807**
NB	Core	Yes	0.890	0.818	0.752	0.738	0.647	0.770
LR	Core	Yes	0.874	0.720	0.534	0.386	0.376	0.578
MS-NB	Core	No	**0.889**	**0.869**	**0.813**	**0.787**	**0.634**	**0.798**
LR	Core	No	0.882	0.741	0.606	0.473	0.402	0.621
MS-NB	Spread	No	**0.622**	**0.715**	**0.750**	**0.791**	**0.732**	**0.722**
LR	Spread	No	0.413	0.471	0.492	0.487	0.373	0.447
MS-NB	LEI	No	0.816	0.826	**0.843**	**0.728**	**0.495**	**0.742**
LR	LEI	No	**0.894**	**0.861**	0.775	0.638	0.361	0.706
MS-NB	All 135	No	**0.711**	**0.556**	**0.381**	**0.302**	**0.297**	**0.449**
LR	All 135	No	0.286	0.265	0.223	0.243	0.208	0.245

Table 7: P-values for significance of forecasting results using Diebold-Mariano test for equality of forecast errors under a real-valued loss (as used in calculation of MAE) using uniform weights for SPF, MS-NB, NB, and LR under various forecast horizons. The models use the core set of real-time variables with 10 lags.

Method	SPF	MS-NB	NB	LR
0 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.09	0.55
NB	<0.01	0.09	1.00	0.93
LR	<0.01	0.55	0.93	1.00
3 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.27	0.11
NB	<0.01	0.27	1.00	0.19
LR	<0.01	0.11	0.19	1.00
6 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.97	0.02
NB	<0.01	0.97	1.00	0.01
LR	<0.01	0.02	0.01	1.00
9 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.23	0.03
NB	<0.01	0.23	1.00	0.04
LR	<0.01	0.03	0.04	1.00
12 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.16	<0.01
NB	<0.01	0.16	1.00	<0.01
LR	<0.01	<0.01	<0.01	1.00

Table 8: P-values for significance of forecasting results using Diebold-Mariano test for equality of forecast errors under a zero-one loss (as used in calculation of F-measure and AUC) using uniform weights for SPF, MS-NB, NB, and LR under various forecast horizons. The models use the core set of real-time variables with 10 lags.

Method	SPF	MS-NB	NB	LR
0 mo. Horizon				
SPF	1.00	0.79	0.67	0.61
MS-NB	0.79	1.00	0.10	0.38
NB	0.67	0.10	1.00	0.88
LR	0.61	0.38	0.88	1.00
3 mo. Horizon				
SPF	1.00	0.11	0.17	0.57
MS-NB	0.11	1.00	0.48	0.32
NB	0.17	0.48	1.00	0.45
LR	0.57	0.32	0.45	1.00
6 mo. Horizon				
SPF	1.00	0.05	0.02	0.43
MS-NB	0.05	1.00	0.64	0.15
NB	0.02	0.64	1.00	0.06
LR	0.43	0.15	0.06	1.00
9 mo. Horizon				
SPF	1.00	0.10	0.05	0.34
MS-NB	0.10	1.00	0.56	0.26
NB	0.05	0.56	1.00	0.14
LR	0.34	0.26	0.14	1.00
12 mo. Horizon				
SPF	1.00	0.30	0.54	0.11
MS-NB	0.30	1.00	0.25	0.01
NB	0.54	0.25	1.00	0.02
LR	0.11	0.01	0.02	1.00

Table 9: P-values for significance of forecasting results using Diebold-Mariano test for equality of forecast errors under a real-valued loss (as used in calculation of MAE) using time-varying weights for SPF, MS-NB, NB, and LR under various forecast horizons. The models use the core set of real-time variables with 10 lags.

Method	SPF	MS-NB	NB	LR
0 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.43	0.63
NB	<0.01	0.43	1.00	0.69
LR	<0.01	0.63	0.69	1.00
3 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.01	0.01
NB	<0.01	0.01	1.00	0.04
LR	<0.01	0.01	0.04	1.00
6 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.10	<0.01
NB	<0.01	0.10	1.00	<0.01
LR	<0.01	<0.01	<0.01	1.00
9 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.04	<0.01
NB	<0.01	0.04	1.00	<0.01
LR	<0.01	<0.01	<0.01	1.00
12 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.05	<0.01
NB	<0.01	0.05	1.00	<0.01
LR	<0.01	<0.01	<0.01	1.00

Table 10: P-values for significance of forecasting results using Diebold-Mariano test for equality of forecast errors under a zero-one loss (as used in calculation of F-measure and AUC) using time-varying weights for SPF, MS-NB, NB, and LR under various forecast horizons. The models use the core set of real-time variables with 10 lags.

Method	SPF	MS-NB	NB	LR
0 mo. Horizon				
SPF	1.00	0.10	0.08	0.13
MS-NB	0.10	1.00	0.64	0.76
NB	0.08	0.64	1.00	0.71
LR	0.13	0.76	0.71	1.00
3 mo. Horizon				
SPF	1.00	<0.01	<0.01	0.05
MS-NB	<0.01	1.00	0.14	0.04
NB	<0.01	0.14	1.00	0.11
LR	0.05	0.04	0.11	1.00
6 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.19	<0.01
NB	<0.01	0.19	1.00	0.01
LR	<0.01	<0.01	0.01	1.00
9 mo. Horizon				
SPF	1.00	<0.01	<0.01	<0.01
MS-NB	<0.01	1.00	0.58	<0.01
NB	<0.01	0.58	1.00	<0.01
LR	<0.01	<0.01	<0.01	1.00
12 mo. Horizon				
SPF	1.00	<0.01	<0.01	0.95
MS-NB	<0.01	1.00	0.32	<0.01
NB	<0.01	0.32	1.00	<0.01
LR	0.95	<0.01	<0.01	1.00

Table 11: Forecasting Results in F-measure for MS-NB with different number of variable lags.

F-Measure, time-dependent weights

Number of Lags	Cur. Mo.	3 Mo. Ahead	6 Mo. Ahead	9 Mo. Ahead	12 Mo. Ahead
10	**0.890**	**0.843**	**0.817**	0.809	0.672
6	0.849	0.800	0.805	**0.848**	0.768
3	0.817	0.792	0.760	0.835	**0.813**
1	0.793	0.789	0.749	0.827	0.811

Table 12: Forecasting Results in MAE for MS-NB with different number of variable lags.

MAE, time-dependent weights

Number of Lags	Cur. Mo.	3 Mo. Ahead	6 Mo. Ahead	9 Mo. Ahead	12 Mo. Ahead
10	**0.038**	**0.050**	**0.055**	0.053	0.077
6	0.051	0.062	0.060	**0.048**	0.062
3	0.059	0.064	0.067	0.052	**0.058**
1	0.072	0.064	0.071	0.052	0.062