Generalized ROC function and time-series probabilistic classification: Application to business cycle identification *

Maximo Camacho † Andres Romeu ‡ Salvador Ramallo §

June 17, 2024

Abstract

We introduce the three-dimensional Generalized ROC (GROC) function, which extends the conventional two-dimension components of the ROC function by incorporating threshold values (α) along with False Positive Rates (FPR) and True Positive Rates (TPR) of the classifier. To evaluate classifier performance comprehensively, we propose the Area of the GROC (AGROC) function, which quantifies performance by measuring the difference between the area under the GROC projection on the TPR$\times\alpha$ plane and the area under the GROC projection on the FPR$\times\alpha$ plane. Our simulations show that AGROC outperforms the standard AUROC in evaluating classifier performance when dealing with time-series probabilistic classifiers. In our empirical analysis, we showcase the accuracy of our approach in determining which recession probabilities, computed from recently proposed Markov-switching specifications using US GDP growth rate data, most accurately align with the NBER-referenced business cycle phases.

Keywords: Business cycles, ROC curve, Area under the ROC curve
JEL classification: C14, C18, C38, E32.

*The authors are grateful for the support of grant PID2022-136547NB-I00 founded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE. All remaining errors are our responsibility. Data and codes that replicate our results are available from the authors' websites.

†Métodos Cuantitativos para la Economía y la Empresa, Universidad de Murcia, 30100 Murcia, Spain. Email: mcamacho@um.es

‡Fundamentos del Análisis Económico, Universidad de Murcia, 30100 Murcia, Spain. Email: aromeu@um.es

§Universidad de Murcia and UNIR, 30100 Murcia, Spain. Email: salvador.ramallo@um.es

1 Introduction

The Receiver Operating Characteristic (ROC), as originally introduced by Peterson and Birdsall (1953), found its roots in World War II as a tool used by the US army to enhance the detection rate of enemy aircraft through radar systems. Its straightforward interpretation played a pivotal role in the rapid adoption of this technique. The ROC curve, specifically, provides a visual representation of the trade-offs between true positives and false positives in a classifier. Thus, the Area Under the ROC curve (AUROC) serves as a natural measure of classification performance in this context. An AUROC value of 1 signifies perfect classification, while 0.5 indicates no discriminatory ability.

Over the past few decades, ROC analysis has become widely employed for assessing the performance of classification models across various fields, including radiology (Obuchowski, 2003), meteorology (Harvey et al., 1992), medicine (Kumar and Indrayan, 2011), machine learning (Bradley, 1997), psychology (Swets, 1996), and economics (Berge and Jordà, 2011). In this paper, the analysis is specifically limited to probabilistic classification models, which do not directly assign instances to classes but instead provide probabilities associated with each class. Furthermore, the focus is on time series applications that involve only two distinct classes.[1]

In practice, the effectiveness of probabilistic classification in guiding decisions relies on its ability to accurately differentiate between the two corresponding classes. Probabilistic differentiation is achieved when the probabilities closely approach zero for one class and stay near one for the other. Classification accuracy pertains to the performance of the classification, aiming to maximize true positives while minimizing false positives. While ROC curve metrics could serve as a suitable tool for evaluating the accuracy of classification probabilities, they may not fully capture how effectively these probabilities differentiate the actual classes.

One of the reasons for this limitation of the AUROC, as well as other related threshold-moving type measures such as Precision-Recall curves, stems from the scale invariance of ROC metrics and the Area Under the curve (Lobo et al, 2007). Consequently, the actual difference in the probabilities between the two classes has minimal impact on the AUROC value. Even if the probability assigned to one class only slightly exceeds that of another class, it is still counted as a success, thereby contributing to the overall AUROC score. While this property is beneficial for ranking classifiers with different units, it may not adequately address the requirement for accurately calibrated probability outputs (Cook, 2007). Other measures that are not threshold-moving type, such as the mean square error or the log loss function, however, are not specifically tailored for ranking the performance of classifiers in classification tasks..

To illustrate this limitation with an example, let us consider evaluating the performance of two probabilistic classifiers. The first classifier assigns probabilities of 1 to one class and 0 to the other. The second classifier is derived from the first one by scaling its probabilities by 1e-10. According to ROC metrics, both classifiers achieve a perfect AUROC of 1 since they correctly classify instances into both classes. Nevertheless, only the first classifier holds practical value for real-world decision-making. This is because decision-makers rely on informative

[1]The implications derived from our study can readily apply to cross-sectional classification probabilities and multiple classes as well.

probability estimates that accurately reflect the underlying class probabilities. Therefore, the minimal changes in probabilities observed in the second classifier, ranging from 1e-11 to 1e-9, lack meaningful information for real-time decision-making.

In this paper, we aim to overcome this limitation by presenting an alternative measure of classification performance. Rather than introducing additional metrics based on the averaging of previous measurements, as proposed by Caruana and Niculescu-Mizil (2004), we offer a more streamlined solution by introducing a novel metric that mitigates the shortcomings of threshold-moving measurements. Our proposed metric not only takes into account True Positive Rates (TPR) and False Positive Rates (FPR), similar to ROC metrics, but also incorporates the sequence of threshold levels (α) to weight them in its calculation. Consequently, we extend the two-dimensional (FPR×TPR) ROC curve into a three-dimensional (FPR × TPR ×α) Generalized ROC (GROC) curve. In this context, the ROC curve is a specific case of the GROC curve when projected onto the FPR × TPR plane, with the AUROC representing the area under this projection.

Introducing a novel metric for evaluating classifier performance, we present the Area of the Generalized ROC (AGROC). This metric comprises two key components. Firstly, we calculate the area under the projection of the GROC curve on the TPR×α plane, denoted as AU_{TPR}. This component quantifies the effectiveness of the classifier in accurately identifying one specific class when it occurs. Secondly, we compute the area under the projection of the GROC curve on the FPR×α plane, labeled as AU_{FPR}. This component indicates the tendency of the classifier to make errors by incorrectly signaling the occurrence of that class when it does not actually happen.

Subsequently, AGROC is determined as the difference between these two components, $AU_{TPR} - AU_{FPR}$, representing the equilibrium between classifier accuracy in identifying a specific class and the potential for false alarms during the other class. A perfect classifier achieves the maximum value of 1, while a non-informative classifier scores 0.[2] Unlike AUROC, AGROC is not scale-invariant, providing more nuanced assessment of probabilistic classifier performance. As a consequence AGROC provides a solution to the previously mentioned challenge, approaching almost 0 for classifiers generated by multiplying a perfect probabilistic classifier by 1e-10.

We validate the reliability of this framework using simulated classifiers that emulate some challenges encountered in time-series probabilistic classification. Through MonteCarlo simulations with different persistence of the distinct classes, we observed how, when faced with predictors close to perfect classification, scaled, noisy, or even with good classification ability in a limited threshold, the AUROC measure tends to be generous when evaluating the ability of the predictors, while in opposition, AGROC is able to discern these inaccuracies that move the predictors away from being a perfect classifier.

To underscore the empirical utility of our approach, we align with the direction set forth by Berge and Jordà (2011), from which the ROC metrics have garnered attention in economic contexts. Specifically, ROC metrics have been instrumental in assessing the effectiveness of state probabilities in accurately discerning business cycle turning points, which holds immense

[2]AGROC values could range between 0 and -1 in cases where classifications are inverted.

significance for academics, policymakers, and practitioners. Notable contributions using ROC measures to assess the extent to which recession probabilities align with the chronology of business cycle phases provided by official dating committees include studies by Lahiri and Wang (2013), Owyang et al. (2015), Lahiri and Yang (2016), Pönkä (2017), Pönkä and Stenborg (2019), Camacho et al. (2018), Piger (2020), Leiva-Leon et al. (2020), Ercolani and Natoli (2020), Galvão and Owyang (2020), Ferrari and Le Mezo (2021), and Lahiri and Yang (2022, 2023).

In our empirical analysis, we employ AGROC to rank various classifiers based on their alignment with the business cycle phases determined by the National Bureau of Economic Research (NBER) Business Cycle Dating Committee. Specifically, we employ different versions of a 2-state Markov-switching model to generate filtered recession probabilities from US real GDP growth rate data. While not exhaustive, our study encompasses the seminal proposal by Hamilton (1989) alongside extensions by Eo and Kim (2016), Eo and Morley (2022, 2023), and Leiva-Leon et al. (2024). Importantly, all models have been re-estimated using a sample covering the period from $1947Q4$ to $2023Q4$ to ensure the comparability of the results.

Our findings indicate that, when excluding Covid-19 data, all recession probabilities closely align with the NBER chronology. As a result, the models yield high AUROC and AGROC values, suggesting that both of which serve as valid metrics for assessing classification performance. However, the AUROC values are exceptionally high for all classifiers, despite variations in their ability to classify certain business cycle episodes, resulting in limited differentiation among them. In contrast, AGROC provides a more nuanced perspective on classifier performance, enabling a clearer ranking.

The inclusion of pandemic-related data significantly alters this landscape. This is particularly evident in the seminal proposal by Hamilton (1989), where recession probabilities are near-zero for all periods except during the pandemic recession. Despite this, the probabilities of recessions can still classify dates into the two distinct business cycle phases, but this is true only for extremely tiny thresholds like 5e-13. Due to the scale-invariant property of ROC metrics, AUROC remains close to one. By contrast, AGROC approaches zero, signaling the expected decline in business cycle classification accuracy.

In light of these considerations, we rely on GROC to conduct a comprehensive evaluation of the performance of all proposals as classifiers of the NBER business cycles. Our findings highlight the significance of departing from the assumption of constant regime-specific mean output growth rates in probabilistic classification of business cycles. Notably, the approach proposed by Eo and Kim (2016), which relaxes this assumption, emerges as the most accurate alternative for computing business cycle probabilities in the sample analyzed.

Our paper is structured as follows. In Section 2, we begin by describing the challenge of ROC metrics to measure the performance of time-series probabilistic classification models and propose the GROC alternative. Section 3 is devoted to Monte Carlo simulations, which serve as the basis to show the advantages of GROC over ROC metrics. In Section 4, we discuss the computation of filtered recession probabilities using several versions of Markov-switching models and use our new metric to assess which of these probabilities best align with the NBER-referenced state of

the business cycle. Finally, Section 5 presents our conclusions.

2 A new metric of probabilistic classification performance

2.1 Time-series probabilistic classification

To represent probabilistic classification, we assume the existence of only two distinct classes, each of which characterizing particular regimes or states. In time series analyses, the evolution of classes categorizes the dates of a calendar time period $\{1, \ldots, T\}$ into these two regimes. To characterize the class occurrence, we introduce the discrete class-indicator variable, denoted as S_t, taking values in the set $\{0, 1\}$, where $S_t = s$ indicates class s occurring at time t, with $s = 0, 1$. Consequently, the sequence $S = \{S_1, \ldots, S_T\}$ encompasses the complete array of all conceivable classification outcomes.

A key task in time-series probabilistic classification involves assigning class labels to sequentially ordered data by estimating the probability of belonging to one of the two classes at a specific time period, denoted here as $\xi_t \in [0, 1]$. Assessing the classification ability of these class, state or regime probabilities commonly involves comparing their historical performance in classifying the data. A good classifier should provide high probability of a particular class when time t is within that class, juxtaposed with a low probability of that class when time t is in the other class.

Assuming, without loss of generality, that the interest lies in classifying class 1, we proceed as follows. Given a threshold, α, we classify time t as class 1 when $\xi_t \geq \alpha$, and as class 0 when $\xi_t < \alpha$. A true positive event occurs when $\xi_t \geq \alpha$ and $S_t = 1$, while a false positive event occurs when $\xi_t \geq \alpha$ and $S_t = 0$. The probabilities of these two events are termed True Positive Rate (TPR) and False Positive Rate (FPR), which can be expressed as follows:

$$TPR(\alpha) = p(\xi_t \geq \alpha | S_t = 1) \tag{1}$$

$$FPR(\alpha) = p(\xi_t \geq \alpha | S_t = 0). \tag{2}$$

In practical applications, $TPR(\alpha)$ is estimated as the ratio of the number of times that $\xi_t \geq \alpha$ over the number of class 1 events. Besides, $FPR(\alpha)$ is estimated as the ratio of the number of times that $\xi_t \geq \alpha$ over the number of class 0 events.[3]

2.2 ROC metrics

ROC curves offer a graphical representation of the trade-off between $TPR(\alpha)$ and $FPR(\alpha)$ for every possible threshold α. As α approaches 0, both $TPR(\alpha)$ and $FPR(\alpha)$ tend towards one, and as α approaches 1, both tend towards zero. Thus, the ROC curve is depicted as the

[3]In the literature, the $TPR(\alpha)$ is often denoted as "sensitivity" while the $p(\xi_t < \alpha | S_t = 0) \equiv 1 - FPR(\alpha)$ is known as "specificity".

trajectory of the coordinates $\mathcal{A}(\alpha) = (FPR(\alpha), TPR(\alpha))$ in the xy-plane. When the state probabilities fail to provide informative classification of the sequence S, $TPR(\alpha) = FPR(\alpha)$ for all possible thresholds. This implies that the ROC curve aligns with the 45-degree line that connects the origin to the point (1,1).[4] Conversely, a perfect classifier situates the ROC curve in the upper-left section of the unit quadrant.

The classification performance of a vector of state probabilities is often evaluated based on the position of the ROC curve within the unit quadrant. As a consequence, a standard measure of the overall classification ability is the Area Under the ROC (AUROC) curve. In the case of a perfect classifier, AUROC equals 1, while any deviation from this perfect classification gradually reduces AUROC until it reaches 0.5, which is the expected value for a random classifier.[5] Therefore, AUROC is typically used to rank a set of classifiers by the quality of their respective classifications.

In practice, an approximation of AUROC can be obtained using a sorted grid of thresholds, α_r, where r ranges from 1 to R. The grid starts with $\alpha_1 = 0$ and ends with $\alpha_R = 1$. The ROC curve is then represented as a plot of points (x_r, y_r) in the first quadrant of the $[0,1] \times [0,1]$ coordinate plane. Here, x_r is placed on the x-axis and represents the true positive rate at threshold α_r, while y_r is placed on the y-axis and represents the false positive rate at the same threshold. Using this notation, AUROC is approximated by the following formula:

$$AUROC = \sum_{r=2}^{R} |x_r - x_{r-1}| \frac{(y_r + y_{r-1})}{2}. \tag{3}$$

This equation essentially calculates the area under the ROC curve in the xy-plane by summing the areas of trapezoids. The width of each trapezoid is equal to the difference between consecutive sampling points on the FPR axis, and the average height is determined by the corresponding points on the TPR vertical axis.

To perform inference, a bootstrapping method can be employed to obtain an empirical approximation of the distribution of AUROC, as proposed by Bertail et al. (2008). However, applying bootstrapping to the ROC can be challenging due to the presence of data from two distinct classes. To address this issue, we can use stratified bootstrapping. This involves performing resampling with replacement of the state probabilities separately from subsamples, each of which corresponding to one of the two regimes. For each of these resampled probabilities, ξ_t^m, a new value of $AUROC^m$ is calculated, with m ranging from 1 to a large value M.[6]

[4]For example, the ROC from a random classifier and a classifier with constant probabilities of class 1 events ($\xi_t = c$, for all $t = 1, \ldots, T$, and $0 \leq c \leq 1$) is the 45 degrees line.

[5]In cases where classifiers incorrectly classify class 0 as class 1, AUROC values may fall between 0 and 0.5. To address this issue, reversing the labels can be considered.

[6]An alternative approach is to represent the AUROC measure as a weighted average of correlation coefficients between a set of binary indicators and the target event as shown in Yang et al. (2024). This representation paves the way for inference development based on parametric methods under serial dependence.

2.3 Generalized ROC metrics

This section extends the traditional ROC function, defined in the xy-plane by coordinate points $\mathcal{A}(\alpha) = (FPR(\alpha), TPR(\alpha))$, to the Generalized ROC (GROC) function, defined in the xyz-space with triple-coordinate points $\mathcal{R}(\alpha) = (FPR(\alpha), TPR(\alpha), \alpha)$. This expansion adds a third coordinate, determined by the threshold α, to the standard ROC analysis.

Figure 1 illustrates our proposal with an example of ordered triples $\mathcal{R}(\alpha)$ for a typical classifier. The figure includes a three-dimensional plot of $\mathcal{R}(\alpha)$, along with three associated coordinate planes: the FPR×TPR plane, representing the FPR- and TPR-axes; the TPR×α plane, representing the TPR- and α-axes; and the FPR×α plane, representing the FPR- and α-axes.

The traditional ROC curve, depicted in the figure as line L1, represents the graphical plot of $\mathcal{A}(\alpha)$, and can be obtained from the data provided by $\mathcal{R}(\alpha)$ as the projection of the triples onto the FPR×TPR plane, where $\alpha = 0$. Consequently, the standard AUROC, indicated by the red area in Figure 1, can be calculated as the area under this projection, as explained in (3). Because the ROC is a two-dimensional projection on the bottom plane, it overlooks the magnitude of the thresholds represented by the slope of the GROC. This limitation results in a scale-invariant metrics that may yield unreliable rankings of classifiers, especially when threshold magnitudes are important such as in the case of probability classifiers.

To address this issue, we introduce a new metric for classification ability called the Area of the Generalized ROC (AGROC). Our approach focuses on the projections of $\mathcal{R}(\alpha)$ onto the TPR×α and FPR×α planes, and the areas beneath these projections, represented by curves L2 and L3 in Figure 2. Using the notation from expression (3), the approximations to the areas under these curves for a given classifier are as follows:

$$AU_{TPR} = \sum_{r=1}^{R} |y_r - y_{r-1}| \frac{(\alpha_r + \alpha_{r-1})}{2}, \tag{4}$$

$$AU_{FPR} = \sum_{r=1}^{R} |x_r - x_{r-1}| \frac{(\alpha_r + \alpha_{r-1})}{2}, \tag{5}$$

respectively. The first expression calculates the area AU_{TPR} by adding trapezoids, each with a base calculated as the difference between consecutive sampling points on the TPR axis, and height computed as the average of their associated thresholds. This calculation method measures the cumulative gains of true positives, weighted by their corresponding thresholds. Similarly, the second expression computes the area AU_{FPR} by summing trapezoids, with bases representing the differences in consecutive false positives multiplied by the average of their associated thresholds. Here, the calculation captures the cumulative losses incurred when generating false positives, weighted by their corresponding thresholds. These two metrics are visually represented by the purple and green areas in Figure 1, respectively.

In this extension of ROC measures, we define the Area of the Generalized ROC as:

$$AGROC = AU_{TPR} - AU_{FPR}. \tag{6}$$

This metric quantifies the trade-offs between the cumulative gains and the cumulative losses achieved by the classifier. Mimicking the approach used for AUROC, we can additionally approximate the sample distribution of AGROC using stratified sampling with replacement, providing an empirical estimate of the standard deviation.

3 Performance evaluation via simulations

3.1 Illustrations

Consider the sequence of 100 dates belonging to class 0 and class 1, denoted as $S = S_1, \ldots, S_{100}$, which is generated by assuming that S_t follows a 2-state Markov-switching process of order one taking values in $\{0, 1\}$, with transition probabilities $p_{00} = p(S_t = 0/S_{t-1} = 0) = 0.9$ and $p_{11} = p(S_t = 1/S_{t-1} = 1) = 0.8$. In Panel A of Figures 2 to 4, the occurrences of $S_t = 1$ are represented with shaded areas. Let us begin to elucidate our proposal with intuition by examining the classification performance of two extreme classifiers in relation to that of standard ROC metrics.

Figure 2 illustrates a classifier that impeccably distinguishes between class 0 and class 1 dates, i.e. $\xi_t = S_t$. In Panel A, the red line represents the probabilities of class 1 for this classifier, perfectly aligning with the generated class 1 dates. Panel B displays the ROC curve, obtained from the projection of the triples $\mathcal{R}(\alpha)$ onto the FPR×TPR plane. For this classifier, this curve consists of three points: (1,1) for a threshold of 0, (0,1) for all thresholds between 0 and 1, and (0,0) for a threshold of 1. The green area in Panel B represents the AUROC, which reaches its maximum value of one.

In the case of this perfect classifier, Panel C presents the projection of $\mathcal{R}(\alpha)$ on the TPR×α plane. Here, TPR remains at 1 for thresholds within the interval [0,1) and decreases to 0 when the threshold reaches 1. The red area represents AU_{TPR}, which attains the maximum value of 1. Meanwhile, Panel D displays the projection of $\mathcal{R}(\alpha)$ on the FPR×α plane, showing that FPR equals 1 for a threshold of 1 and drops to 0 for thresholds within the interval (0,1]. In this projection, AU_{FPR} is 0. AGROC, measured as the difference between AU_{TPR} and AU_{FPR}, reaches its maximum value of 1. Mirroring AUROC, AGROC signifies the highest possible discrimination ability for this classifier.

Contrary to the perfect classifier, Figure 3 focuses on classifiers that offer no meaningful information, such as when providing constant class 1 probabilities, $\xi_t = c$, where $0 \leq c \leq 1$. To illustrate, Panel A displays a classifier that provides class 1 probabilities of 0.5, $\xi_t = 0.5$, for all t. Panel B shows the ROC curve, which comprises only two points: (1,1) for $\alpha \leq 0.5$ and (0,0) for $\alpha > 0.5$, resulting in an AUROC of 0.5, represented by the green area. Panels C and D show the projections of $\mathcal{R}(\alpha)$ on the TPR×α and FPR×α planes and are represented by red and blue areas, respectively. They reveal that TPR and FPR are both 1 for $\alpha \leq 0.5$ and 0 for $\alpha > 0.5$, resulting in AU_{TPR} and AU_{FPR} both being 0.5. Consequently, AGROC is 0, indicating, like AUROC, that the classifier lacks informativeness.

Moving to more realistic classifiers, Figure 4 presents a comparative assessment of three alternative classifiers. In panel A, classifier A (red line) and classifier B (green line) were

generated as follows: at each time t, the absolute value of a random number is drawn from the Gaussian distribution $N(0, 0.05)$ for classifier A and $N(0, 0.5)$ for classifier B. This value is then added to S_t when $S_t = 0$ and subtracted from S_t when $S_t = 1$. Finally, classifier C is obtained by multiplying classifier A by 0.1.

As expected, Panel A shows a close alignment between classifier A and the sequence of states S, resulting in a ROC curve (depicted in Panel B as a red area) that closely follows the upper part of the unit quadrant, yielding an AUROC of 0.99. Similarly, Panel C reveals that TPR is 1 for all thresholds except for those very close to 1, resulting in a high AU_{TPR} of 0.96, as indicated by the red area. For this classifier, FPR is positive only for thresholds lower than 0.1 but becomes 0 for thresholds greater than 0.1, leading to a low AU_{FPR} of 0.04, as indicated by the red area in Panel D. This classifier exhibits the largest AGROC, reaching the value of 0.92.

The increased noise in generating classifier B significantly diminishes its classification accuracy compared to the sequence of states. Consequently, its ROC curve, depicted by the green area in Panel B, is closer to the 45-degree line, resulting in a decreased AUROC of 0.72, indicating lower classification performance. Using GROC metrics, classifier B, exhibits some positive signals that align with dates of $S_t = 1$, resulting in a AU_{TPR} of 0.60, represented by a green area in Panel C. However, this classifier also reports significant probabilities of class 1 dates when $S_t = 0$, as evidenced in Panel A, leading to the highest AU_{TPR} of 0.36, represented by the green area in Panel C. Subtracting AU_{TPR} from AU_{FPR} yields an AGROC of 0.24, indicating lower classification performance than classifier A.

Of particular relevance to this paper is the comparison between classifiers A and C. In Panel A, we observe that classifier A demonstrates significantly stronger probabilistic differentiation compared to classifier C. This distinction is evident as classifier A approaches near-certainty when $S_t = 1$ and near-zero probability when $S_t = 0$, whereas classifier C consistently maintains very low probabilities for state 1.

Despite these disparities, both classifiers yield identical ROC curves, reflected by AUROC values of 0.99, as depicted by the green area in Panel B, matching for both classifiers. However, upon employing GROC metrics, classifier C presents relatively diminished probabilities for class 1 when $S_t = 1$, resulting in a considerable reduction in AU_{TPR} to 0.09, as indicated by the blue area in Panel C. Conversely, the probabilities of class 1 when $S_t = 0$ for classifier C are exceedingly low, implying that false signals only emerge for extremely minuscule thresholds, thus yielding a low AU_{FPR} of 0.01. Consequently, the AGROC value for classifier C decreases to 0.08, accentuating its inferior probabilistic differentiation performance compared to classifier A.

This example illustrates how AUROC may not be a reliable metric for ranking classification probabilities when assessing their relative classification performance. This is particularly evident in scenarios where the scale-invariance property of AUROC fails to adequately differentiate between classifiers with varying levels of probabilistic differentiation.

3.2 Monte Carlo analyses

In order to assess the robustness and reliability of the AUROC and AGROC metrics, a Monte Carlo experiment was conducted. Inspired by the illustration in the previous subsection, this experiment comprised nine unique scenarios for the data generating process of the classes S_t. These scenarios resulted from the simulation of three binary Markov processes for S_t: first, a high-persistence process (HI) with transition probabilites $p_{00} = p_{11} = 0.9$, second, a low-persistence (LO) with transition probabilites $p_{00} = p_{11} = 0.2$ and thrid, an unbalanced persistence (UNB) process, with high 0-state persistence ($p_{00} = 0.9$) and low 1-state ($p_{11} = 0.2$). The analysis was performed across three sample sizes, $T = 200, 500, 1000$, resulting the mentioned nine scenarios. For each scenario and size, a single run of S_t was generated.

Our aim is to test the performance of distinct time-series probabilistic classification models that were representative of different situations and that illustrated the different capacity of the two measures analyzed. In particular, we tested up to seven distinct classification models:

1. (U) Uninformative signal: $\xi_t^1 \sim U(0,1)$
2. (AT) Almost True signal: $\xi_t^2 = S_t(1 - |\varepsilon_t|) + (1 - S_t)|\varepsilon_t|$, where $\varepsilon_t \sim N(0, 0.1)$
3. (N) Noisy signal: $\xi_t^3 = S_t(1 - |\varepsilon_t|) + (1 - S_t)|\varepsilon_t|$, where $\varepsilon_t \sim N(0, 0.5)$
4. (HS) Heavily Shrinked signal: $\xi_t^4 = 0.05\xi_t^2$
5. (SS) Softly Shrinked signal: $\xi_t^5 = 0.5\xi_t^2$
6. (MA) Moving-Average fat tails signal: $\xi_t^6 = S_t + (1 - S_t)P(S_t, 5)$ where $P(S_t, 5)$ is a moving average of order 5 centered in t.
7. (MAAT) Moving-Average and Almost True signal: $\xi_t^7 = S_t\xi_t^2 + (1 - S_t)\xi_t^6$

For each of the seven distinct classification models, 1000 samples of one of the simulated classifier above were obtained and the corresponding corresponding AUROC and AGROC metrics are computed and recorded. The whole process is repeated for each classifier and the results are summarized in Table 1, where average values over the 1000 replicas are reported. We proceed to break down the different signals and the results for the measures under study.

The first signal U is considered a *lower-boundary* test case. Note that U simply random noise drawn from a uniform distribution over the interval (0,1). Panel A of Figure 5 shows the poor performance of one simulation of this classifier, which is entirely naive and fails to convey any information about the true state of nature, S_t, with $S_t = 1$ represented by shaded areas. Consequently, the fourth column of the table shows that the average AUROC is around 0.5 and AGROC is close to zero, regardless of the persistence of the binary Markov processes and the sample size of the simulation.

The AT classifier, an example of which is depicted in Panel B of Figure 5, represent the prototype of a *very accurate signal*: it almost mimics the true state and takes the value of one when $S_t = 1$ except for a small error. Consequently, high values from the AUROC and AGROC metrics would be expected. Fifth column in Table 1 confirms that this is indeed the case. However, we observe a differentiated behavior between both metrics: the AUROC scores a

perfect classifier value of 1, while the AGROC metrics is more conservative with a more modest 0.84 value.

To investigate if this deterioration pattern persists for higher levels of classifier error, we introduced the N classifier, an example of which is displayed in Panel C of Figure 5. The sixth column in Table 1 shows that the N noisy signal reduces the AUROC approximately by a 30% but decreases the AGROC by almost a 75%, further with respect to the AT. This result suggests that AGROC penalizes the amount of noise conveyed in the classifier more strictly than AUROC does.

Classifiers HS and SS explore the effect of the scale-invariance property on classification performance. Both mimics the AT classifier although they apply shrinkage proportions of 95% and 50%, respectively. The representative illustrations of these two classifiers, displayed in Panels D and E of Figure 5, illustrate the performance deterioration experienced by HS and, to a lesser extent, by SS compared to AT. Despite this evident deterioration, columns seven and eight of Table 1 show that AUROC is unaffected by the changes in units and remains at a very high measure of 1 in both cases, as it did in the original units of the AT case. However, AGROC is more sensitive to the shrinkage factor: the value of this metrics is close to zero when the shrinkage is significant and close to 0.4 for the softer shrinkage.

Interestingly, we have found that the classification metrics are almost insensitive to the choice of sample size or the persistence of the binary Markov processes S_t. This is unsurprising, as changing these parameters only affects the frequency of regime changes in the processes. Consequently, the predictive ability of the probabilistic classifier across different scenarios is expected to differ only when the signal displays some form of memory or dynamics, which is not the case for any of the five previous ξ_t's considered.

We further explore the influence of the persistence of the regimes in on classification performance by introducing the MA classifier. This classifier exhibits short-run memory, as illustrated in Panel F of Figure 5. Consequently, the figure shows that classification performance deteriorates around the regime-change dates, and we expect higher deterioration when the regimes are less persistent, as the number of turning points increases. However, column nine of Table 1 shows that AUROC does not deteriorate when the persistence of the regimes falls. By contrast, AGROC does deteriorate when the persistence of the states falls.

To further explore the differences in both metrics and their responses to signal noise and scaling, we introduce a supplementary classifier denoted as MAAT. This classifier combines features of moving average, as in MA, and good classification performance, as in AT. MAAT behaves like MA within 0-states and like an AT signal within 1-states. An illustrative example is shown in Panel G of Figure 5. This classifier effectively predicts state 1 despite noise and follows a moving average of order 5 when outside state 1. Consequently, it represents a signal of lower predictive quality than MA, a characteristic that the metrics should be able to discern. The results in the last column of Table 1 indicate that AUROC fails to identify the lower performance of the signal, while AGROC successfully reflects this fact by yielding markedly lower values than for the MA scenario.

4 Empirical application

Berge and Jordà (2011) played a pivotal role in popularizing the use of ROC curves in the realm of business cycle analysis. They conceptualized dating procedures as a distinct instance of a standard time-series classification problem, thus highlighting the applicability of ROC curves within this framework. Their work has inspired numerous empirical economic applications utilizing ROC metrics. In this section, we aim to compare our GROC metrics with the traditional ROC approach in assessing the performance of classifiers in business cycle analysis.

Specifically, our study examines the business cycle classification performance of various versions of univariate Markov-switching models for US GDP growth rates. These models have gained widespread acceptance within economic research circles due to their robustness in providing a probabilistic classification framework that accurately identifies the different phases of the business cycle.

4.1 Markov-switching models

To address the complexity of identifying the observed tendency of economic activity to exhibit markedly different behaviors during downturns, reflecting the nonlinear dynamics inherent in economic cycles, Hamilton (1989) introduced the regime-switching specification. Specifically, the author proposed that the real output growth, denoted as y_t, for $t = 1, ..., T$, can be represented as a non-linear process controlled by an unobserved regime-switching two-state variable, S_t, taking values in $\{0, 1\}$. The model is expressed as:

$$y_t = \mu_{s_t} + \epsilon_t, \tag{7}$$

where $\epsilon_t \sim iidN(0, \sigma^2_\epsilon)$. Here, we abstract from autoregressive parameters, following the empirical results by Camacho and Perez-Quiros (2007) and the recent specifications by, among others, Eo and Kim (2016) and Eo and Morley (2022, 2023). If we assume $\mu_0 > 0$ and $\mu_1 < 0$, we can label $S_t = 0$ and $S_t = 1$ as representing the states of economic expansion and recession at time t, respectively.

Hamilton (1989) assumed that the state variable S_t evolves as an irreducible 2-state Markov chain. The transition probabilities for this Markov chain are defined as:

$$p(S_t = j|S_{t-1} = i, S_{t-2} = h, ..., \Psi_{t-1}) = p(S_t = j|S_{t-1} = i) = p_{ij} \tag{8}$$

where i and j take on values of 0 or 1, and Ψ_t represents the information set up to period t, collected in $\{y_1, ..., y_t\}$. Additionally, the author described a forward filter that calculates the filtered probabilities of being in a recession at time t given Ψ_t. Here, we call this model M_H.

Several extensions of this seminal work have been proposed in the literature, aiming to incorporate empirical features such as time-dependency in regime-specific means, time-dependency in the long-run mean growth rate, or stochastic volatility. One of the most significant contributions was proposed by Eo and Kim (2016), who were the first to relax the assumption of constant regime-specific growth rates made by Hamilton (1989). Specifically, let μ_{0,τ_0} and μ_{1,τ_1} be the

growth rates during the τ_0th episode of expansion or the τ_1th episode of recession in the sample, respectively, and denote the long-run growth rate as δ_{D_t}. Consider the Markov-switching model of the business cycle:

$$y_t = \delta_{D_t} + (1 - S_t)\mu_{0,\tau_0} + S_t\mu_{1,\tau_1} + \epsilon_t, \tag{9}$$

where the variance of the disturbance term is specified as a stochastic volatility process, implying $ln(\sigma^2_{\epsilon,t}) = ln(\sigma^2_{\epsilon,t-1}) + \eta_t$, with $\eta_t \sim iidN(0, \sigma^2_\eta)$.

In order to allow for different episodes of expansions and recessions, the dynamics of the regime-specific mean growth rates are assumed to follow a random walk. Specifically, if there are N_0 episodes of expansion and N_1 episodes of recession, the dynamics of the random walks are

$$\mu_{0,\tau_0} = \mu_{0,\tau_0-1} + \omega_{0,\tau_0}, \quad \omega_{0,\tau_0} \sim N(0, \sigma^2_{0,\tau_0}) \tag{10a}$$

$$\mu_{1,\tau_1} = \mu_{1,\tau_1-1} + \omega_{1,\tau_1}, \quad \omega_{1,\tau_1} \sim N(0, \sigma^2_{0,\tau_1}) \tag{10b}$$

where $\tau_0 = 1, 2, \ldots, N_0$ and $\tau_1 = 1, 2, \ldots, N_1$.

Referring to the unconditional state probabilities as $\pi_i = P(S_t = i)$, with $i = 0, 1$, Eo and Kim (2016) assumed that both the long-run condition $E[\pi_0\mu_{0,\tau} + \pi_1\mu_{1,\tau}] = 0$ and identifying restrictions $\mu_{0,\tau} > 0$, and $\mu_{1,\tau} < 0$ hold. In regards to the long-run mean growth rate, the authors considered two possibilities. Firstly, they proposed that the long-run mean growth rate is a random walk process determined by:

$$\delta_t = \delta_{t-1} + e_t, \tag{11}$$

with $e_t \sim iidN(0, \sigma^2_e)$. Alternatively, they suggested an abrupt structural break, modeled as a Markov-switching process with an absorbing state:

$$\delta_{D_t} = \delta_0(1 - D_t) + \delta_1 D_t, \tag{12}$$

with D_t following a Markov chain with two states, $D_t = 0$ and $D_t = 1$, whose the transition probabilities are $(D_t = 0|D_{t-1} = 0) = q_{00}$ and $p(D_t = 1|D_{t-1} = 1) = q_{11} = 1$.

A model such as that in expression (9), where the stochastic long-run growth follows a random walk as in (11), with the corresponding long-run and identifying restrictions, is referred to as model II in Eo and Kim (2016). Conversely, when the stochastic long-run growth, as defined in (12), is permitted to exhibit one structural break, they referred to it as model IV. Here, we denote these two extensions of model (7) as M^2_{EK} and M^4_{EK}, respectively.

As competitors, Eo and Kim (2016) also allowed constant regime-specific mean growth rates, meaning that $\mu_{0,\tau} = \mu_0$ and $\mu_{1,\tau} = \mu_1$ in (9). Additionally, they also allow for potential bounce-back effects. Specifically, the model becomes

$$y_t = \delta_{D_t} + (1 - S_t)\mu_0 + S_t\mu_1 + \lambda \sum_{m=1}^{M} (y_{t-m} - \delta_{D_{t-m}}) + \epsilon_t, \tag{13}$$

where the variance of the disturbance term is specified as a stochastic volatility process as in (9).

The third addend refer to the bounce-back effect and allows for long and deep recessions to be followed by stronger recoveries, where the authors set $M = 6$ for the length of the bounce-back effect.

Considering the specifications of the long-run mean growth rate and the presence of bounce-back effects in (13), Eo and Kim (2016) examined four alternative specifications. Not allowing bounce-back effects ($\lambda = 0$) and modeling the long-run mean growth rate as a random walk as in (11) resulted in model I, while modeling the latter as having a structural shift as in (12) led to model V. Conversely, permitting bounce-back effects ($\lambda \neq 0$) and modeling the long-run mean growth rate as a random walk as in (11) led to model III, while modeling the latter as having a structural shift as in (12) resulted in model VI. Here, we denote these models as M^1_{EK}, M^5_{EK}, M^3_{EK}, and M^6_{EK}, respectively.

Closely related to model (9), Leiva-Leon at al. (2024) considered time-varying output growth means associated with expansions and recessions that do not follow random walks. In particular, their specification assumes $\delta_{D_t} = 0$ and constant variance of errors as in (7) but does not allow for bounced-back effects. Additionally, viewing explicitly each regime-specific episode as a function of time, the authors proposed the following transition between expansion and recession episodes, marked by τ_{0t} and τ_{1t}:

$$\tau_{0t} = \begin{cases} 1 - S_1, & \text{if } t = 1 \\ (1 - S_1) + \sum_{j=2}^{t} max(0, S_{j-1} - S_j), & \text{otherwise,} \end{cases} \tag{14a}$$

$$\tau_{1t} = \begin{cases} S_1, & \text{if } t = 1 \\ S_1 + \sum_{j=2}^{t} \ max(0, S_j - S_{j-1}), & \text{otherwise,} \end{cases} \tag{14b}$$

where $\tau_{0t} = 1, 2, \ldots, N_0$ and $\tau_{1t} = 1, 2, \ldots, N_1$. When a transition between states occurs, such as from expansion to recession, $\mu_{0,\tau_{0t}}$ remains unchanged during the new phase of the state, while $\mu_{1,\tau_{1t}}$ is derived based on the output growth rates observed during the corresponding recessionary period. We call this model M_{DGE}.

The latest extensions of the Markov-switching model that we consider here were proposed by Eo and Morley (2022, 2023). The authors introduced a three-state Markov-switching latent variable where $S_t = 0$ represents expansions, $S_t = 1$ indicates L-shaped recessions, and $S_t = 2$ signifies U-shaped recessions. Notably, the transition probabilities are constrained to prevent transitions between the two types of recessions, i.e., $p_{12} = p_{21} = 0$. Utilizing the indicator function $I(\cdot)$, the extended model for output growth incorporates the following conditional mean based on the three regimes:

$$y_t = \mu_0 + \mu_1 \chi_t I(S_t = 1) + \mu_2 \chi_t I(S_t = 2) + \varphi \sum_{k=1}^{K} \chi_{t-k} I(S_{t-k} = 2) + \epsilon_t, \tag{15}$$

with $\mu_0 > 0$, $\mu_1 < 1$, and $\mu_2 > 2$.

In this equation, the variance of the disturbance term allows to account for a breakdate volatility reduction, expressed as $\epsilon_t \sim (0, \sigma_t^2)$, with $\sigma_t^2 = \sigma_0^2 I(t \leq 1984Q2) + \sigma_1^2 I(t > 1984Q2)$. Enforcing an U-shaped recession to have only transitory effects implies $\mu_2 + M\varphi = 0$, where M is

set to 5. Lastly, to accommodate the extreme outliers during the pandemic, the model integrates the volatility decay function χ_t, where $\chi = 1$ for $t < 2020Q2$ and $\chi_t = 1 + (c-1) + \rho^{(t-2020Q2)}$ otherwise. The probability of recession can be derived by summing the probability of being in either of the two recession types, either L-shaped or U-shaped. When $\rho = 0$, the model reduces to Eo and Morley (2022), denoted as M_{EM}^{22}, while it is labeled as M_{EM}^{23} otherwise.[7]

4.2 Classification ability

This section is dedicated to illustrating the benefits of GROC-type metrics compared to standard ROC-type metrics for assessing the classification performance of recession probabilities derived from the aforementioned versions of an univariate Markov-switching for US GDP growth rates. To do this, we evaluate the alignment of the filtered probabilities of recession with the NBER-referenced dates.[8]

To ensure comparability, all of these models are estimated using the same sample, which starts in $1947Q4$ following Eo and Kim (2016) and ends in $2023Q4$.[9] However, to examine the potential biases in classification performance caused by the extreme values in output growth rates during the first quarters of 2020 due to the pandemic, we also conduct the analysis using a sample ending in $2019Q4$.

Figures 6 and 7 provides an illustrative example of how ROC and GROC metrics gauge the performance of an inferred probability of recession from Hamilton's (1989) Markov-switching model as a business cycle classifier for two different samples. Starting from Figure 6, Panel A presents the quarterly US GDP growth rates spanning from $1951Q1$ to $2019Q4$, while Panel B depicts the corresponding filtered probabilities of the negative state. The classification probabilities roughly align with the official chronology of business cycle expansions and recessions as provided by the NBER, represented by shaded areas.

The correspondence is evident in Panel C, where the ROC curve is prominently located near the upper left corner. Quantitative support for this classification effectiveness is provided by an AUROC value of 0.989, as reported in Table 2 for M_H, along with a standard deviation of 0.005. Nonetheless, the near-perfect AUROC value may overstate the performance in classification, given its failure, for example, to promptly identify the recession in 2000 or its delayed recognition of the 1970 recession. Moving to GROC metrics, the projections of the ROC curves on the TPR$\times\alpha$ and FPR$\times\alpha$ planes in Panels D and E reveal the large cumulative gains of the true positives (AU_{TPR}=0.607) and small cumulative losses when generating false positives (AU_{TPR}=0.003). As a result, Table 2 shows that AGROC is large, 0.604, although significantly below the value of 1 for a perfect classifier, leaving room for ranking in better positions classifiers with larger AGROC and overcoming the drawback of ROC metrics.

Examining the performance of M_H when including pandemic data, provides an illustrative

[7]Readers interested in detailed derivations and estimation procedures for the models mentioned in this paper are encouraged to refer to the original works cited in the text.

[8]Importantly, our aim is not to provide here the best Markov-switching model, but rather to illustrate how some alternative specifications in the literature can be ranked using GROC metrics according to their ability to classify dates into expansions and recessions.

[9]The models were estimated using codes obtained from the authors' websites.

example of the failure of ROC metrics for probabilistic classifiers due to their scale-invariant property. Panel A of Figure 7, which extends the sample up to 2023$Q4$, shows that the US economy experienced, in 2020, the deepest U-shaped recovery in recent history, with the sharpest downturn, falling by -8.93% in the second quarter, and rebounding at a record rate of 7.55% in the third quarter. As a consequence, Panel B illustrates that the very large drop in 2020 is identified by the likelihood as the low-growth state, and the model apparently relegates all the previous recessions of normal falls to the high-growth state. Thus, all the probabilities of recession are nearly reduced to zero for the period prior to the COVID-19 pandemic. Intuitively, this should significantly reduce the ability of the recession probabilities to classify the business cycle.

The unexpected outcome is reported in Panel D, where the ROC curve continues to exhibit a high classification performance, closely tracking points near the top-left corner. The quantitative evidence of this surprising effectiveness is reflected in an AUROC value of 0.914, as detailed in Table 2, with a standard deviation of 0.024. The reason for this seemingly counterintuitive result can be attributed to the scale invariance property of AUROC. This issue is clearly illustrated in Panel C, where we zoomed in on the probabilities of recession reported in Panel B by reducing the y-axis scale. This panel show that for extremely tiny thresholds like 5e-13 the probabilities of recession still provide classification ability.

Consequently, as AUROC assesses true positives and false positives for all conceivable thresholds of the recession probabilities, including almost negligible thresholds, it still yields near-perfect classifier values. However, despite the large AUROC values for both samples, the probabilities derived from the larger sample yield a classifier with significantly less economic significance than that of the shorter sample.

Fortunately, our proposed GROC metrics effectively address this limitations of ROC measures. Panels E and F display the projections of the GROC curve on the TPR$\times\alpha$ and FPR$\times\alpha$ planes, respectively. As the probabilities of recession obtained from M_H when including pandemic data only identify recessions for extremely small thresholds, it results in a substantially reduced AU_{TPR} of 0.028. Additionally, the metrics reveal a limited number of false signals across different thresholds, leading to a small AU_{FPR} of 0.005. With AGROC falling to 0.023 and a standard error of 0.022, barely surpassing the value of an uninformative classifier, the GROC metric accurately reflects the considerably superior classifier performance of the Markov-switching model that excludes Covid-19 data.

After illustrating the weaknesses of ROC metrics and the strengths of our GROC extension, Table 2 helps us rank recently proposed extensions of Hamilton's (1989) seminal Markov-switching model for the samples ending in 2019$Q4$ and 2023$Q4$. Ideally, the models would be ranked to closely align with the NBER business cycle classification, with the top-ranking model expected to achieve the closest correspondence. To facilitate understanding, Figures 8 and 9 display the probabilities of recession provided by these models for both samples, respectively.

Starting with the sample ending in 2019$Q4$, the AUROC values in the table reveal a strikingly high classification performance across all models, with scores often approaching values close to one. This result is in line with Saito and Rehmsmeier (2015) and Lahiri and Yang (2023),

who suggest that in cases of severe class imbalance the performance metrics under ROC get overloaded by the success in predicting the dominant event, even when the rare object interest is not predicted well, as illustrated in Figure 8. This characteristic of AUROC metrics poses a challenge in distinguishing between good classifiers, as they tend to provide little room for differentiation across them, ranging from 0.945 to 0.989. However, the AGROC values offer a more nuanced perspective on classifier performance, taking into account both true positive and false positive rates across all thresholds. Notably, AGROC designates M^2_{EK} as the model with the highest discriminatory power in accurately classifying business cycle phases, with a value of 0.731 (captures pretty accurately every recession and expansions), and ranging the values for the different model in this case from 0.489 to that value.

Evaluating potential changes in ranking classification performance with the sample ending in $2023Q4$ is of particular interest, given that the models must contend with the extreme outliers observed during the pandemic. Some of these models have the necessary flexibility to account for recessions and expansions of various magnitudes, such as the time-varying regime-specific means of M^2_{EK}, M^4_{EK} and M_{DGE}. Notably, M^{23}_{EM} is specially designated to incorporate pandemic data by introducing a decay function for volatility from $2020Q2$.

As indicated in Table 2, among the models considered, M^2_{EK} once again emerges as the top performer, achieving the highest AGROC score of 0.616 (only fails partially for the last recession period). This underscores the superior business cycle classification performance of Markov-switching models that allow different episodes of expansion and recession to have varying output mean growth rates, enabling accurate classification of business cycle phases compared to other models included in the analysis. Again, the range of differentiation is different in both measures, being in the case of AUROC between 0.914 and 0.977, and between 0.023 and 0.616 in the case of AGROC, again highlighting the greater distinguishing ability of the latter measure.

5 Conclusions

The Area Under the ROC (AUROC) has become a common metric for assessing the performance in probabilistic classification. However, although it is simple to calculate, its interpretation in this context could be meaningless for classifier comparison because the ROC metric does not depend on the scale of the probabilities. To address this limitation, we introduce a Generalized ROC (GROC) function, which extends the traditional ROC curve by incorporating a third coordinate, represented by the thresholds (α) used to assign categories to probabilites.

In our setup, the ROC curve is essentially a projection of the GROC function onto the FPR×TPR plane, with AUROC quantifying the classification performance of a given classifier. Furthermore, we introduce the projection of the GROC function onto the TPR×α plane and calculate the area under this curve as AU_{TPR}, measuring the ability of the classifier for correct classification. Additionally, the projection of the GROC function onto the FPR×α plane provides the basis for AU_{FPR}, which assesses the capacity of the classifier to misclassification.

Additionally, we introduce the Area of the Generalized ROC (AGROC), calculated as the difference between AU_{TPR} and AU_{FPR}. This innovative metric effectively captures the balance

between false positives and true positives, offering a reliable measure for evaluating classifier performance. Unlike AUROC, AGROC is not scale-invariant, enhancing its robustness in ranking classifiers for classifying trough probabilities. The robustness of AGROC in ranking classifiers has been extensively assessed through simulations, which have been designed to replicate certain challenges encountered in probabilistic classification.

To offer an empirical illustration, we turn to Berge and Jordà (2011), who notably introduced ROC curves in business cycle analysis, treating dating procedures as a standard classification problem. Utilizing US GDP growth rate data, we employ our metric to rank classifiers derived from different versions of Markov-switching models. The rankings are based on how well the estimated probabilities of recession align with the NBER-determined business cycles. Our results underscore the importance of allowing for different episodes of expansion and recession to have varying output mean growth rates to address the challenge posed by the large outliers observed during the 2020 pandemic.

We look forward to future work addressing the following issues. Following Saito and Rehmsmeier (2015) and Lahiri and Yang (2023), we see a natural extension of our approach generalizing PRC measures of classification performance for classifiers that are applied to strongly imbalanced datasets in which the number of negatives outweighs the number of positives significantly. In addition, our GROC can be re-expressed in terms of the correlation coefficient between a binary outcome and an indicator as in Yang et al. (2024). This could be a technical convenience to perform statistical inference that is robust to the serial correlation in the data.

References

[1] Berge, T., and Jordà, O. 2011. Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics* 3: 246-277.

[2] Bertail, P., Clémençon, S., and Vayatis, N. 2008. On bootstrapping the ROC curve. *Advances in Neural Information Processing Systems* 21.

[3] Bradley, A. 2011. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30: 1145–1159.

[4] Camacho, M., and Perez-Quiros, G. 2007. Jump-and-rest effect of U.S. business cycles. *Studies in Nonlinear Dynamics and Econometrics* 11(4): Article 3.

[5] Camacho, M., Perez-Quiros, G., and Poncela, P. 2018. Markov-switching dynamic factor models in real time. *International Journal of Forecasting* 34: 598-611.

[6] Caruana, R., and Niculescu-Mizil, A. 2004. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and DataMining*, KDD'2004: 69–78.

[7] Cook, N. R. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7): 928-935.

[8] Eo, Y., and Kim, Ch. 2016. Markov-switching models with evolving regime-specific parameters: Are post-war booms or recessions all alike? *The Review of Economics and Statistics* 98: 940-949.

[9] Eo, Y., and Morley, J. 2022. Why has the U.S. economy stagnated since the Great Recession? *The Review of Economics and Statistics* 104: 246-258.

[10] Eo, Y., and Morley, J. 2023. Does the Survey of Professional Forecasters help predict the shape of recessions in real time? *Economics Letters* 233: 111419.

[11] Ercolani, V., and Natoli, 2020. Forecasting US recessions: The role of economic uncertainty. *Economics Letters* 193: No 109302.

[12] Ferrari, M., and Le Mezo, H. 2021. Text-based recession probabilities. European Central Bank Working Paper No. 2516.

[13] Galvão, A., Owyang, M. 2020. Forecasting low frequency macroeconomic events with high frequency data. Federal Reserve Bank of St. Louis Working Paper No. 2020-028A.

[14] Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357-384.

[15] Harvey L., Hammond K., Lusk C., and Mross E. 1992. Application of signal detection theory to weather forecasting behavior. *Monthly Weather Review* 120: 863-883.

[16] Kumar, R., and Indrayan, A. 2011. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics* 48: 277-287.

[17] Lahiri, K., and Wang, J. 2013. Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting* 29: 175-190.

[18] Lahiri, K., and Yang, L. 2016. A non-linear forecast combination procedure for binary outcomes. *Studies in Nonlinear Dynamics and Econometrics* 20: 421-440.

[19] Lahiri, K., and Yang, Ch. 2022. ROC approach to forecasting recessions using daily yield spreads. *Business Economics* 57: 191-203.

[20] Lahiri, K., and Yang, Ch. 2023. ROC and PRC approaches to evaluate recession forecasts. *Journal of Business Cycle Research* 19: 119-148.

[21] Leiva-Leon, D., Perez-Quiros, G., and Rots, E. 2024. Real-time weakness of the global economy. *Journal of Applied Econometrics*, forthcoming.

[22] Lobo, J.M., Jiménez-Valverde, A., and Real, R. 2087. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17(2): 145-151.

[23] Obuchowski, N. 2003. Receiver operating characteristic curves and their use in radiology. *Radiology* 229: 3-8.

[24] Owyang, M., Piger, J., and Wall, H. 2015. Forecasting national recessions using sate-level data. *Journal of Money, Credit and Banking* 47: 847-866.

[25] Peterson, W., and Birdsall. T. 1953. The theory of signal detectability: Part I. The general theory. University of Michigan, Department of Electrical Engineering, Electronic Defense Group, Technical Report 13. Ann Arbor, June.

[26] Piger, J. 2020. Turning points and classification. In: Fuleky, P. (Ed.), *Macroeconomic Forecasting in the era of big data*. Springer International Publishing, pp. 585–624.

[27] Pönkä, H. 2017. The role of credit in predicting U.S. recessions. *Journal of Forecasting* 36: 221–247.

[28] Pönkä, H., and Stenborg, M. 2019. Forecasting the state of the Finnish business cycle. Publications of Ministry of Finance, 2019:13, 1-16.

[29] Saito, T., and Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10(3): e0118432

[30] Swets, J. 1996. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Mahwah, NJ: Lawrence Erlbaum Associates.

[31] Yang, L., Lahiri, K., and Pagan , A. 2024. Getting the ROC into Sync. *Journal of Business and Economic Statistic* 42: 109-121.

Table 1: Comparison of AUROC and AGROC measures for different time-series probabilistic classication models

Metrics	T	Persistence	U	AT	N	HS	SS	MA	MAAT
	200	Low	0.527	1.000	0.709	1.000	1.000	1.000	1.000
	200	High	0.531	1.000	0.710	1.000	1.000	1.000	0.991
	200	Unb.	0.539	1.000	0.708	1.000	1.000	1.000	1.000
	500	Low	0.519	1.000	0.710	1.000	1.000	1.000	0.999
AUROC	500	High	0.518	1.000	0.710	1.000	1.000	1.000	0.996
	500	Unb.	0.528	1.000	0.708	1.000	1.000	1.000	1.000
	1000	Low	0.513	1.000	0.709	1.000	1.000	1.000	1.000
	1000	High	0.513	1.000	0.711	1.000	1.000	1.000	0.996
	1000	Unb.	0.520	1.000	0.710	1.000	1.000	1.000	1.000
	200	Low	-0.003	0.840	0.217	0.039	0.420	0.596	0.439
	200	High	-0.001	0.840	0.220	0.039	0.420	0.906	0.730
	200	Unb.	0.000	0.840	0.217	0.039	0.420	0.868	0.715
	500	Low	-0.000	0.840	0.219	0.039	0.420	0.555	0.384
AGROC	500	High	0.001	0.840	0.219	0.039	0.420	0.885	0.691
	500	Unb.	0.000	0.840	0.217	0.039	0.420	0.914	0.763
	1000	Low	0.000	0.840	0.219	0.039	0.420	0.559	0.394
	1000	High	-0.001	0.840	0.220	0.039	0.420	0.871	0.659
	1000	Unb.	0.001	0.840	0.219	0.039	0.420	0.909	0.762

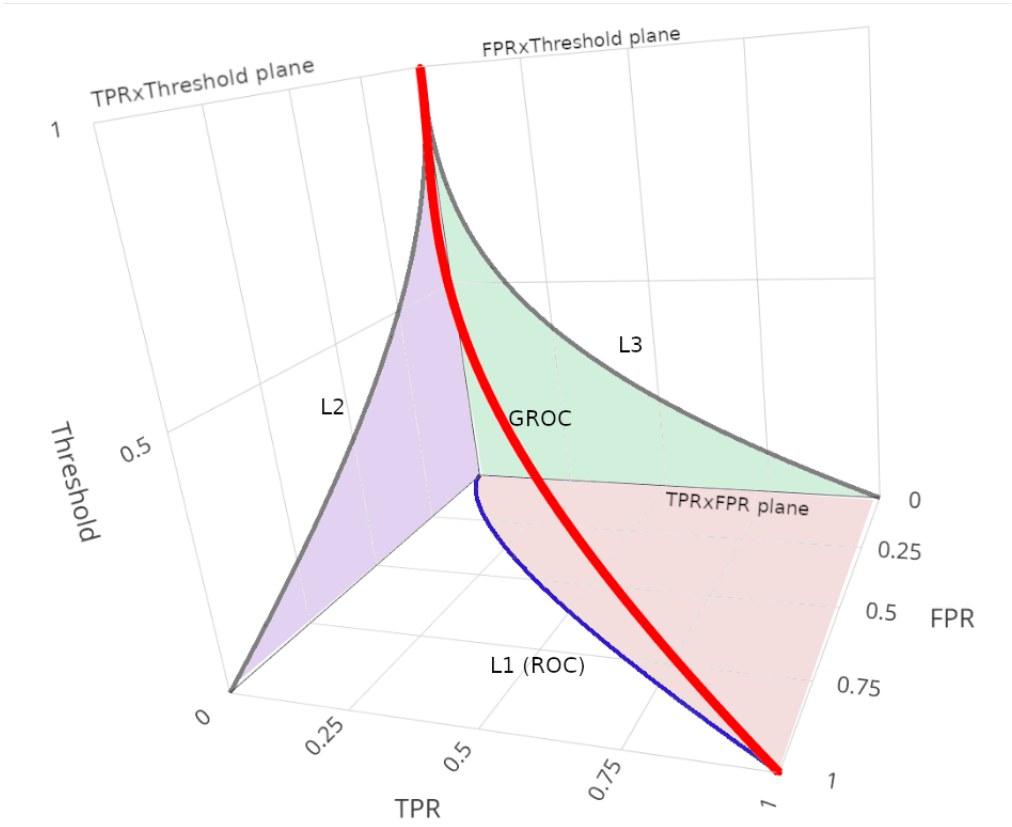
Notes. The table displays AUROC and AGROC values for various versions of predictors for 200, 500 and 1000 time-series length, with a low, high and an unbalanced level of persistence. The table display the results for an uninformative distribution as a classifier (U), and almost true perfect classifier (AT), a noisy classifier (N), a classifier softly shrinkaged (SS), a heavily shrinkaged one (HS), a classifier with fat bumps based on a moving average prediction (MA), and a classifier based on an hybrid almost true and a fat bumps situation (MAAT).

Table 2: Classification performance

	$1947Q4 - 2019Q4$		$1947Q4 - 2023Q4$	
Model	AUROC	AGROC	AUROC	AGROC
M_H	0.989	0.604	0.914	0.023
	(0.005)	(0.042)	(0.024)	(0.022)
M^1_{EK}	0.975	0.578	0.974	0.431
	(0.008)	(0.042)	(0.008)	(0.034)
M^2_{EK}	0.981	0.731	0.970	0.616
	(0.007)	(0.026)	(0.009)	(0.031)
M^3_{EK}	0.975	0.680	0.976	0.524
	(0.009)	(0.041)	(0.007)	(0.033)
M^4_{EK}	0.979	0.678	0.977	0.529
	(0.007)	(0.036)	(0.007)	(0.033)
M^5_{EK}	0.970	0.524	0.963	0.304
	(0.009)	(0.045)	(0.011)	(0.036)
M^6_{EK}	0.975	0.668	0.966	0.487
	(0.008)	(0.044)	(0.009)	(0.029)
M_{DGE}	0.984	0.685	0.975	0.562
	(0.006)	(0.027)	(0.008)	(0.031)
M^{22}_{EM}	0.945	0.489	0.968	0.348
	(0.023)	(0.049)	(0.011)	(0.041)
M^{23}_{EM}	-	-	0.931	0.390
	(-)	(-)	(0.023)	(0.049)

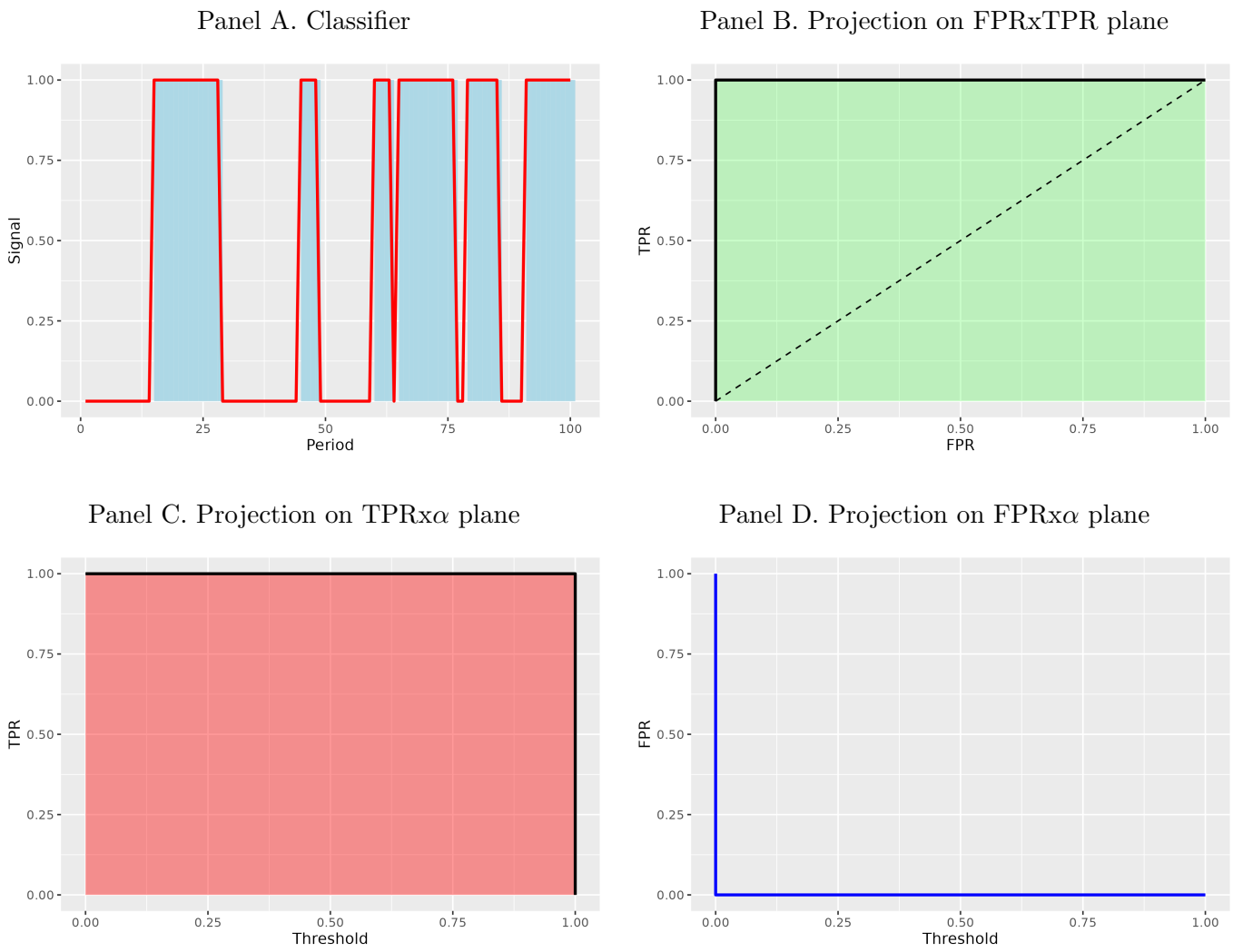
Notes: The table displays AUROC and AGROC values for various versions of univariate Markov-switching models applied to US GDP growth rates, along with their respective standard deviations in parentheses. The model references are as follows: M_H refers to Hamilton (1989), M^j_{EK} denotes model j in Eo and Kim (2016), M_{DGE} refers to Leiva-Leon et al. (2024), and M^{22}_{EM} and M^{23}_{EM} represent models from Eo and Morley (2022) and (2023), respectively. The samples periods span from $1947Q4$ to $2019Q4$ and from $1947Q4$ to $2023Q4$.

Figure 1: Three-dimension representation



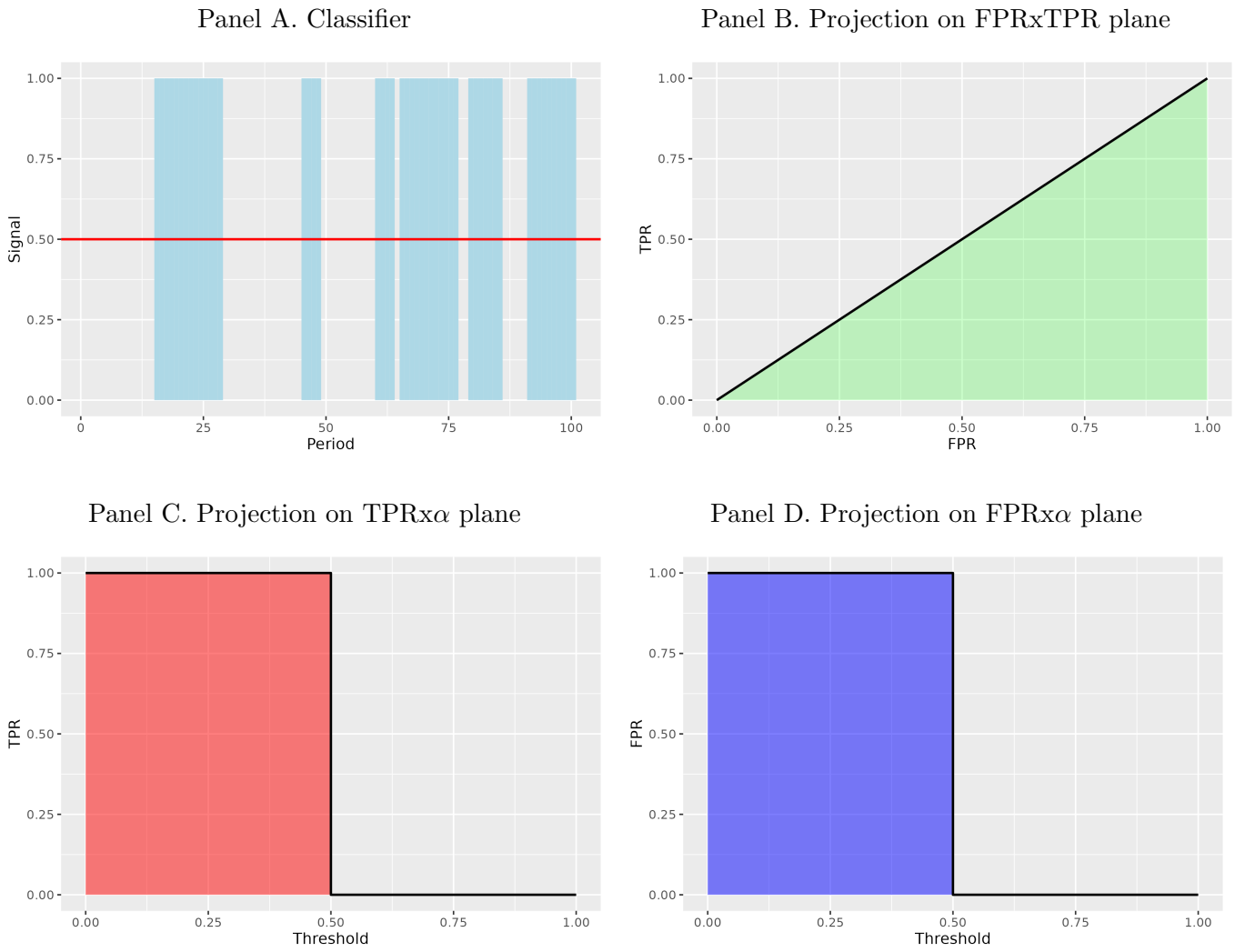
Notes. The red line refers to the ordered triples $\mathcal{R}(\alpha) = (FPR(\alpha), TPR(\alpha), \alpha)$ of a classifier, where α is the threshold. The figure shows the three coordinate planes: the TPR×FPR plane, the TPR×α plane, and the FPR×α plane. The projections of $\mathcal{R}(\alpha)$ on these three planes are L1 (ROC curve), L2 and L3, respectively.

Figure 2: Projections of GROC ($\mathcal{R}(\alpha)$) for a perfect classifier



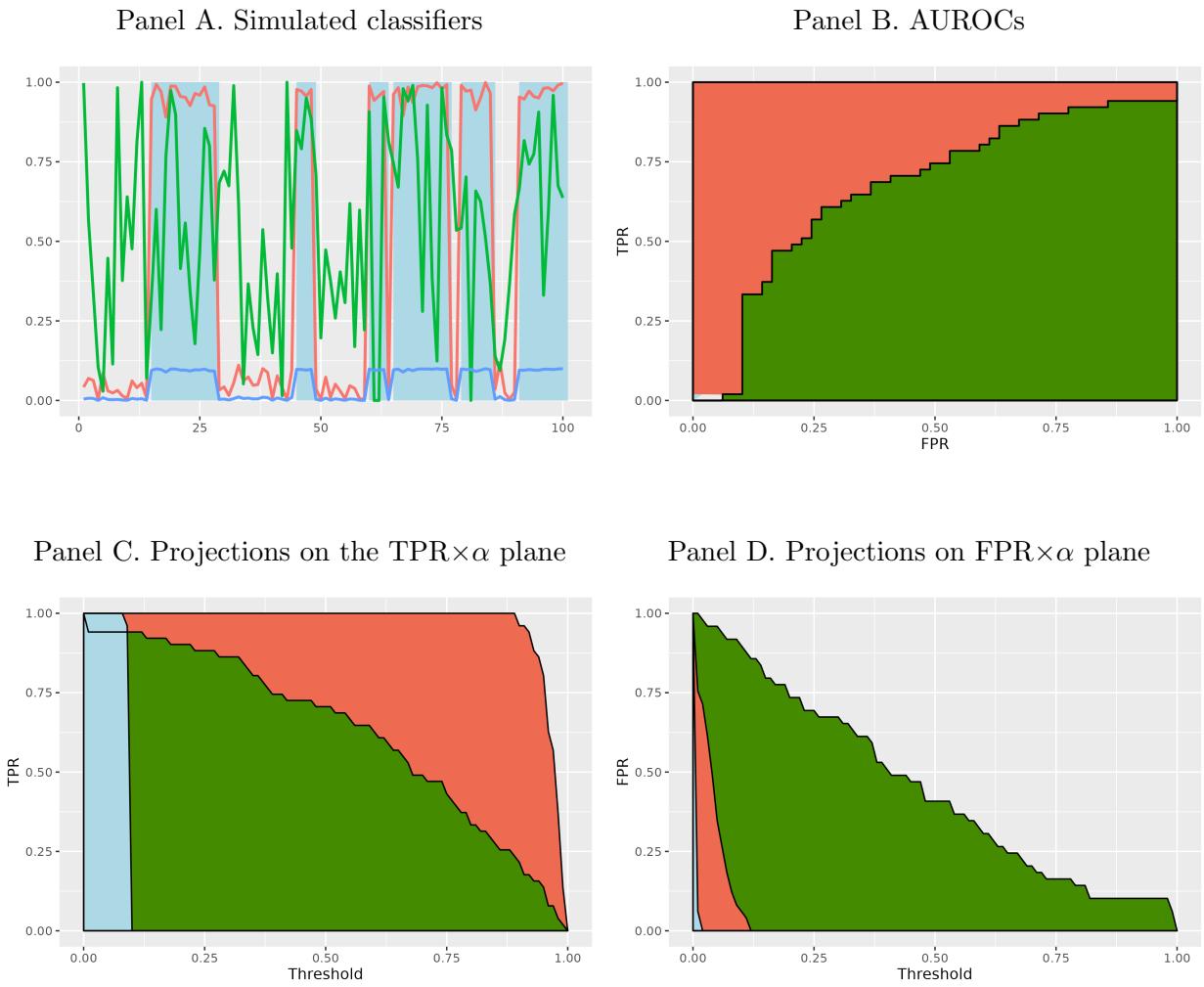
Notes. Panel A represents the occurrence of class 1 with shaded areas and plots the probability of class 1 provided by a perfect classifier. Panels B, C and D display the projections of GROC on the FPRxTPR, TPRxα and FPRxα planes, respectively.

Figure 3: Projections of GROC ($\mathcal{R}(\alpha)$) for an uninformative classifier



Notes: Panel A represents the plot of the probability of class 1 provided by a constant-probability classifier. For further details, refer to the notes of Figure 2.

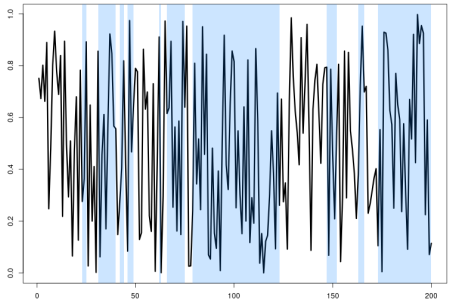
Figure 4: Comparative performance



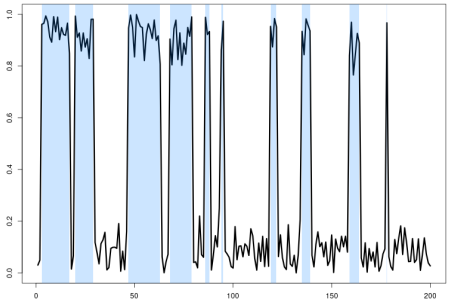
Notes: In Panel A, the red line is generated by adding random draws from $N(0, 0.05)$ to a perfect classifier when the class is 0 and subtracting them when the class is 1. The green line is generated using the same method, but with the variance of the Gaussian process increased to 0.5. The blue line is obtained by multiplying the red line by 0.1. For further details, refer to the notes of Figure 2.

Figure 5: Representative predictions from seven different predictors

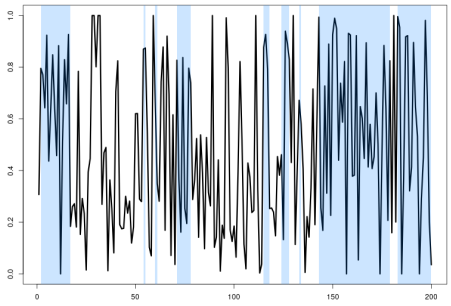
Panel A. Uninformative



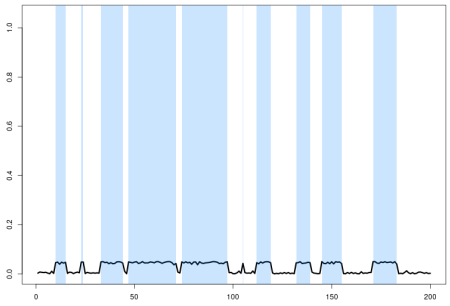
Panel B. Almost True



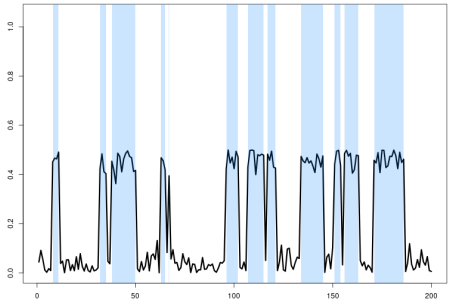
Panel C. Noisy



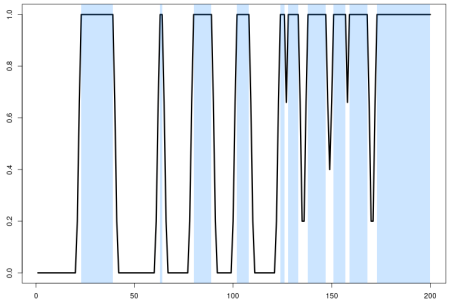
Panel D. Heavily shrinkaged



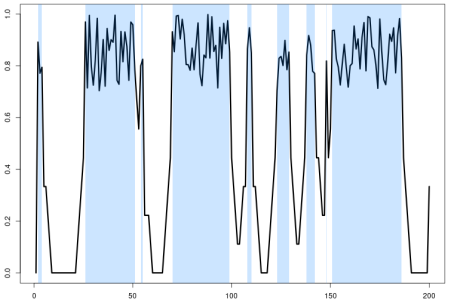
Panel E. Softly shrinkaged



Panel F. MA fat bumps



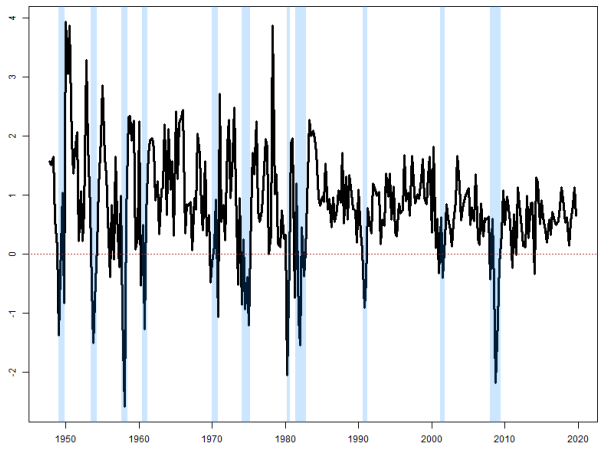
Panel G. MA fat bumps and almost true



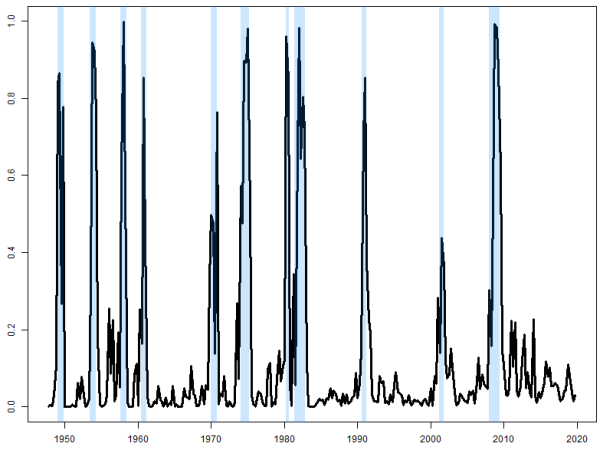
Notes. The figure displays the probability representative predictions of seven different predictors. Shaded areas represent class 1.

Figure 6: US business cycle inferences 1947Q4-2019Q4

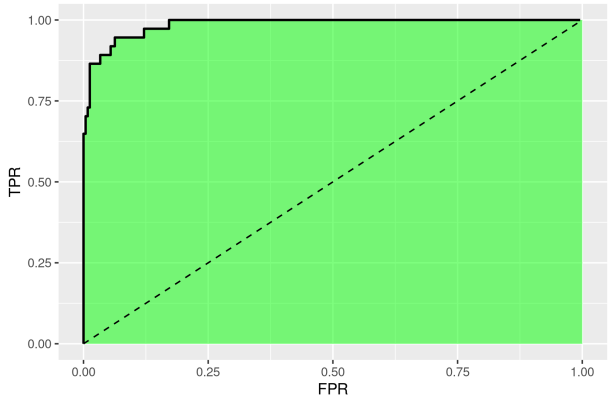
Panel A. GDP growth rates



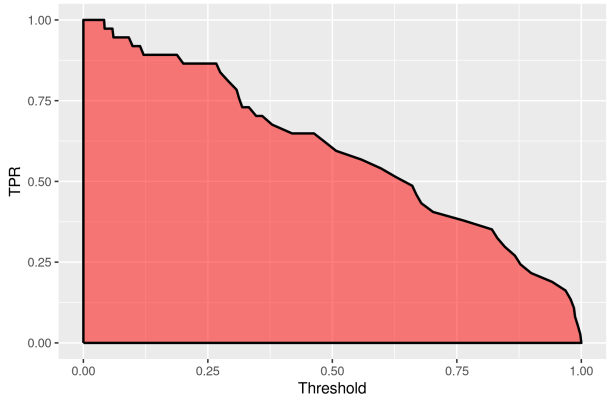
Panel B. Hamilton(1989)



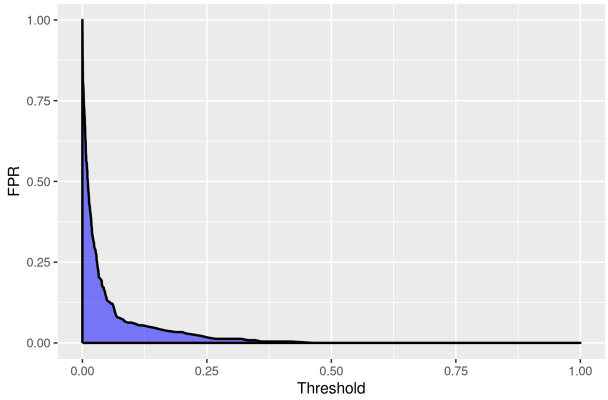
Panel C. AUROC



Panel D. Projections on TPR×α



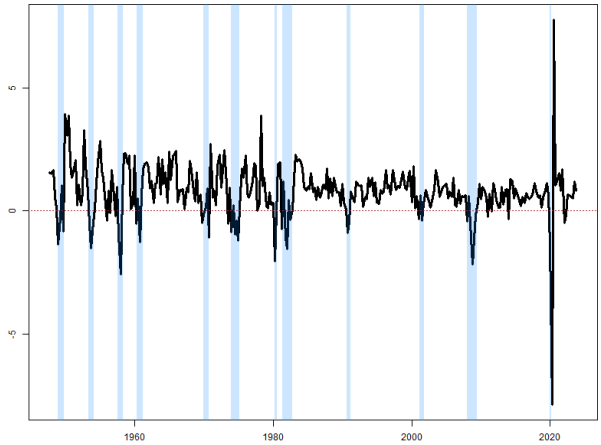
Panel E. Projections on FPR×α



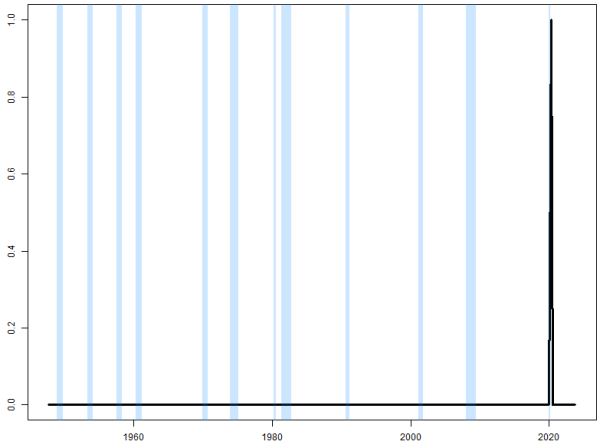
Notes: Panel A displays US GDP growth rates, and Panel B shows the probabilities of recession derived from Hamilton (1989). Panel C presents the AUROC, while Panels D and E display the projections of GROC on the TPR×α and FPR×α planes, respectively. Shaded areas refer to the NBER recessions.

Figure 7: US business cycle inferences 1947Q4-2023Q4

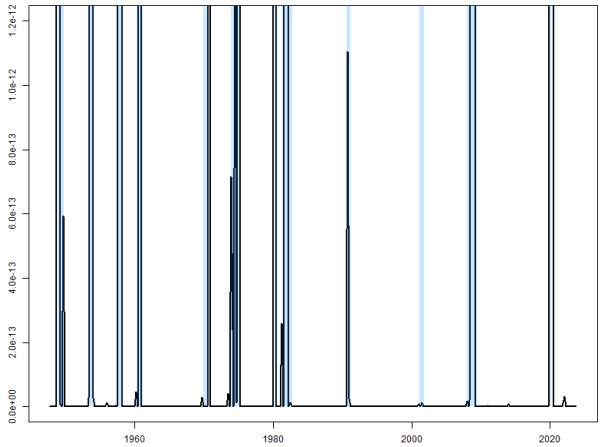
Panel A. GDP growth rates



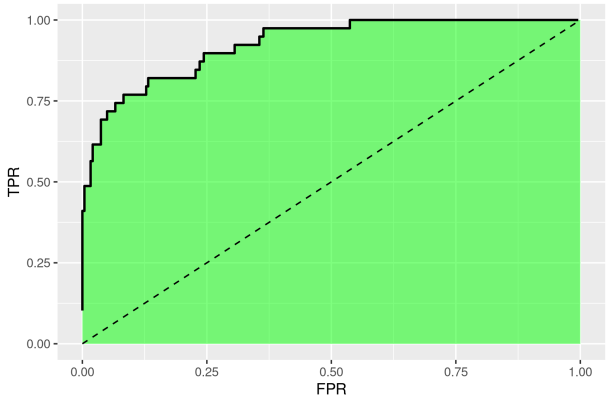
Panel B. Hamilton(1989)



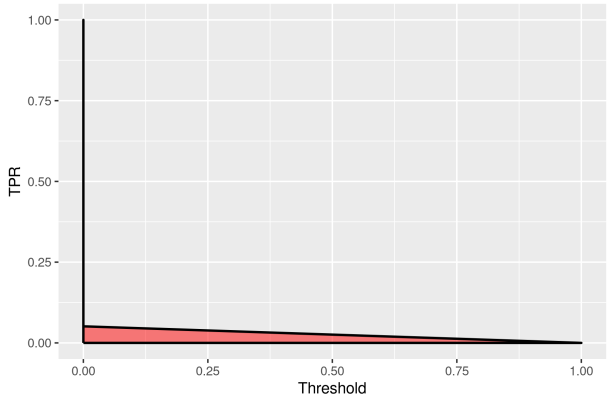
Panel C. Hamilton(1989) zoomed



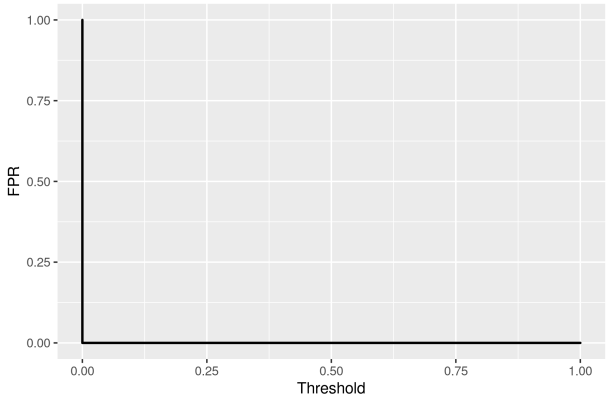
Panel D. AUROC



Panel E. Projections on TPR×α

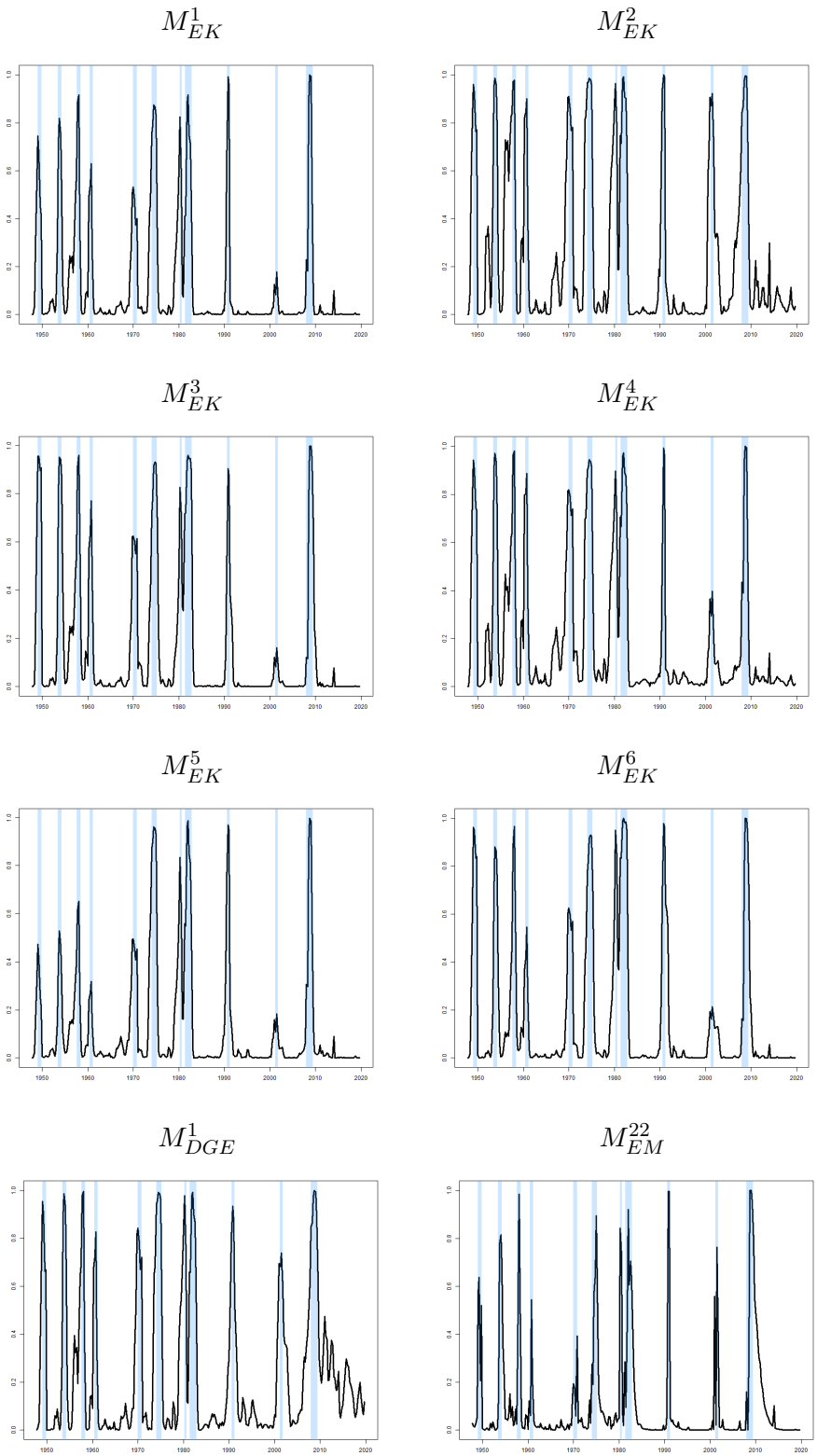


Panel F. Projections on FPR×α



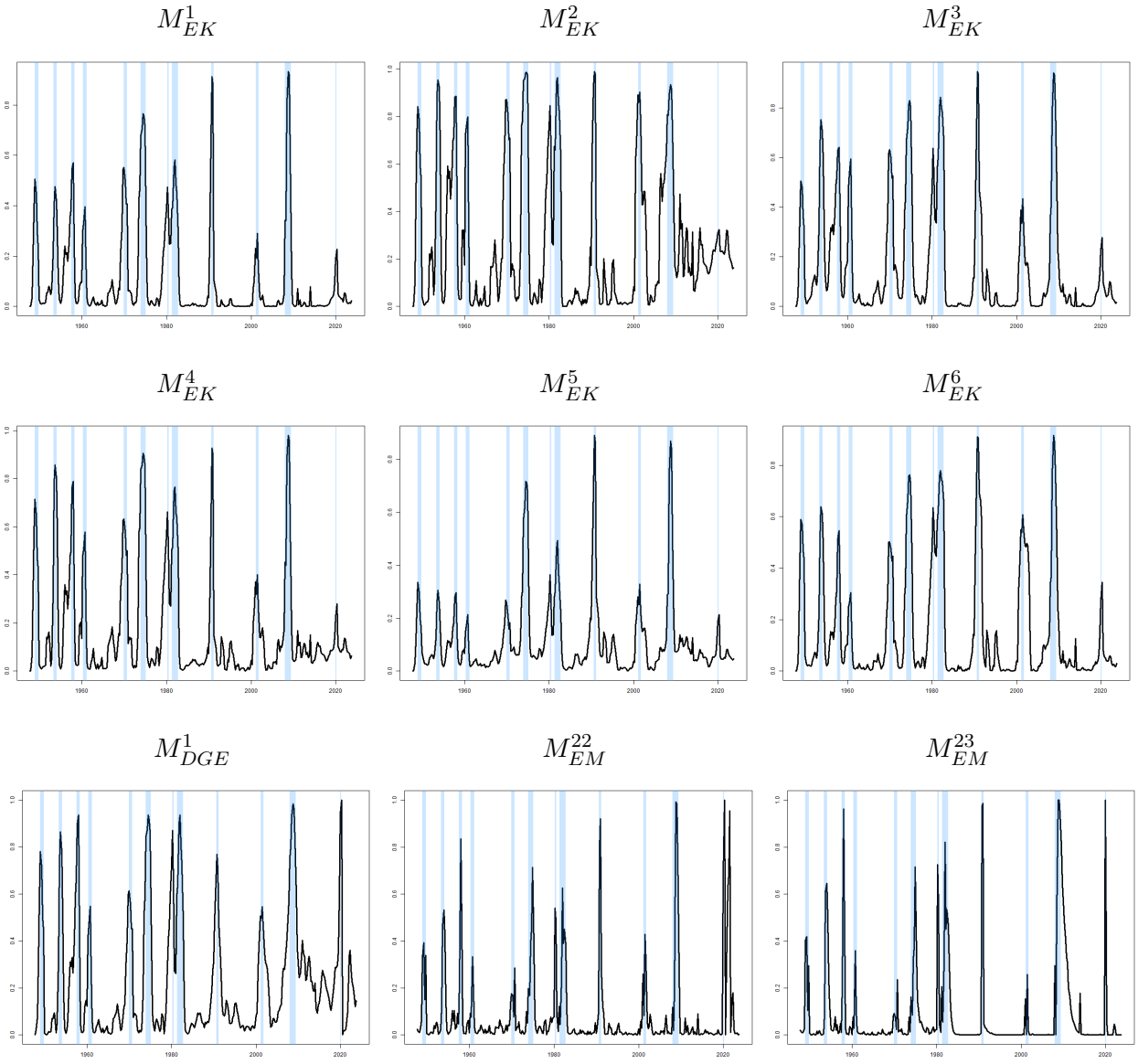
Notes: Panel A displays US GDP growth rates. Panel B shows the probabilities of recession derived from Hamilton (1989), which is zoomed in Panel C. Panel D presents the AUROC, while Panels E and F display the projections of GROC on the TPR×α and FPR×α planes, respectively. Shaded areas refer to the NBER recessions.

Figure 8: Probability classifiers 1947Q4-2019Q4



Notes: The figure displays probabilities of US recessions provided by alternative Markov-switching specifications to Hamilton (1989). Shaded areas refer to the NBER recessions. For further details, refer to the notes of Table 2.

Figure 9: Probability classifiers 1947$Q4$-2023$Q4$



For clarification, please refer to the details provided in Figure 8.