Board of Governors of the Federal Reserve System

International Finance Discussion Papers

Number 1307

January 2021

Optimizing Credit Gaps for Predicting Financial Crises: Modelling Choices and Tradeoffs

Daniel O. Beltran, Mohammad R. Jahan-Parvar, and Fiona A. Paine

Please cite this paper as:
Beltran, Daniel O., Mohammad R. Jahan-Parvar, and Fiona A. Paine (2021). “Optimizing Credit Gaps for Predicting Financial Crises: Modelling Choices and Tradeoffs,” International Finance Discussion Papers 1307. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2021.1307.

Optimizing Credit Gaps for Predicting Financial Crises: Modelling Choices and Tradeoffs

Daniel O. Beltran Mohammad R. Jahan-Parvar Fiona A. Paine*

November 2020

Abstract

Credit gaps are good predictors for financial crises, and banking regulators recommend using them to inform countercyclical capital buffers for banks. Researchers typically create credit gap measures using trend-cycle decomposition methods, which require many modelling choices, such as the method used, and the smoothness of the underlying trend. Other choices hinge on the tradeoffs implicit in how gaps are used as early warning indicators (EWIs) for predicting crises, such as the preference over false positives and false negatives. We evaluate how the performance of credit-gap-based EWIs for predicting crises is influenced by these modelling choices. For the most common trend-cycle decomposition methods used to recover credit gaps, we find that optimally smoothing the trend enhances out-of-sample prediction. We also show that out-of-sample performance improves further when we consider a preference for robustness of the credit gap estimates to the arrival of new information, which is important as any EWI should work in real-time. We offer several practical implications.

JEL Classification: C22; E39; G28.

Keywords: Credit, Credit Gap, Optimization, Predictive Power, Robustness, Trend-cycle decomposition.

*Beltran and Jahan-Parvar are affiliated with the Board of Governors of the Federal Reserve System, and Paine is affiliated with the MIT Sloan Finance Group, e-mails are: daniel.o.beltran@frb.gov (Beltran), mohammad.jahan-parvar@frb.gov (Jahan-Parvar), and painef@mit.edu (Paine). We thank Dan Cao, Matthew Canzoneri, Behzad Diba, Luca Guerrieri, Seung Lee, John Rogers, Tim Schmidt-Eisenlohr, Frank Warnock, Molin Zhong and seminar participants at Federal Reserve Board, Gerogetown University, Midwest Macroeconomic Meetings 2019 (UGA Terry), Georgetown CER conference 2019, and CFE 2020 for comments and discussions. Jahan-Parvar acknowledges the hospitality of the finance group, Carey School of Business, the Johns Hopkins University during the fall of 2019. This paper was previously circulated under the title "Designing Gaps as Early Warning Indicators: Accuracy, Predictive Power, and Robustness." The analysis and the conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors. Any remaining errors are ours.

1 Introduction

Studies such as Lowe and Borio (2002b), Borio and Drehmann (2009), and Drehmann and Juselius (2014) show that credit growth and credit gaps –deviations of aggregate credit to GDP ratio from its long-run trend–are good predictors of systemic banking crises, often referred to as financial crises. They thus support the view that such financial crises are often "credit booms gone bust" (Minsky, 1977; Kindleberger, 1978; Schularick and Taylor, 2012). In particular, since the 2008 Global Financial Crisis (GFC), researchers and policy makers have used credit gaps as inputs for designing early warning indicators (EWI) for financial crises. A reasonable EWI is expected to be accurate (yield minimal false positive and false negative signals), and have good predictive power (predict a reasonable percentage of events). More practically, policymakers, including central banks, use credit gaps for assessing risks to financial stability. Indeed, the Basel Committee for Banking Supervision (BCBS) recommends using credit gaps for determining countercyclical regulatory capital buffers for banks, and makes their baseline gap estimates publicly available.

Estimating credit gaps for macroprudential policy, and using them to design financial crises EWIs inherently requires making many modelling choices, namely: 1) the method used to decompose the credit-to-GDP series into trend and cycle, 2) how much to smooth the trend in doing this decomposition, and 3) how to specify the preferences of the policymaker regarding false positives and false negatives when designing EWIs. Previous studies have constructed EWIs by taking into account a subset of the modelling choices, for example, by considering a single method for trend-cycle decomposition, or by fixing the preference over false positives and false negatives, or by taking a strong stance on how much to smooth the underlying trend when constructing credit gaps. Furthermore, existing studies do not consider the ability of EWIs to predict crises in advanced versus emerging market economies separately, nor do these studies examine the out-of-sample performance of EWIs while taking into account the various tradeoffs that arise from the modelling choices.

We take a step back and systematically investigate these modelling choices. This careful consideration of the modelling choices matters for multiple reasons. It leads to an improvement in the tradeoff between false positives (type-1 errors) and false negatives (type-2 errors) which, in turn, enhances the ability of EWIs to predict crises in sample and, more importantly, out of sample. In addition, ex-post optimization may not necessarily improve performance of EWIs out of sample (see, for example Sarlin and von Schweinitz (forthcoming)). We show that out-of-sample performance can improve if one explicitly considers a preference for robustness of the credit-gap estimate to the arrival of new information.

Our modelling approach offers several practical implications. It provides insights for how high the credit gap can reach before triggering financial stability concerns. For example, if the policymaker has a strong enough preference for minimizing false positives or type-1 errors, we find that the critical threshold for the credit gap is in the narrow range between 12 and 14 percentage points of GDP, regardless of the trend-cycle decomposition method used. Furthermore, our modelling approach highlights an important distinction between advanced and emerging economies. We show that credit gaps yield good EWIs for financial crises in advanced economies both in- and out-of-sample. By contrast, the predictive power of credit gaps is much weaker for emerging economies, suggesting that there might be other factors at play for these countries, such as the political environment, as documented by Herrera, Ordoñez, and Trebesch (2020).

Our study proceeds as follows. We first create optimal EWIs. Specifically, for a sample of 33 countries (both advanced and emerging economies), we decompose the credit-to-GDP ratios into a long-run trend and a cyclical component (the credit gap) using five trend-cycle decomposition methods: (i) Hodrick and Prescott (1997) filter, (ii) the Hamilton (2018) filter, (iii) a Bayesian structural time series model (STM) (Petris, 2010; West and Harrison, 1997), (iv) a simple growth rate of the credit-to-GDP ratio, and (v) deviations of the credit-to-GDP ratio from its moving average. A common feature in all of these methods is a parameter which governs the smoothness of the long-run trend. Previous studies fix the smoothing parameter to ensure a smooth trend. Instead, we consider a wide range of possible values of the smoothing parameter to construct the credit gaps. Using these gaps, we define the EWI as an indicator function that assumes the value of one when the gap exceeds a certain critical threshold, and zero otherwise. We optimize the EWIs for each of the trend-cycle decomposition methods by jointly adjusting the smoothing parameter and the critical threshold of the credit gap which triggers the EWI.

We then evaluate the performance of the EWI using standard criteria adopted from the signal extraction literature (Kaminsky and Reinhart, 1999) (hereafter, KR) for predicting crises. Among the five decomposition methods we consider, the optimized HP filter and Bayesian STM offer the best overall performance in-sample. Out of sample, however, the optimized moving average method performs best if the policymaker is more cautious about false negatives, whereas the growth rate method performs best if the policymaker is more cautious about false positives. When the policymaker is roughly indifferent between false positives and false negatives, the HP filter performs best out of sample.

The rest of the paper is organized as follows. Section 2 reviews related literature. In Section 3 we describe the data used in this study. We describe the trend-cycle decomposition methods and the signal extraction approach we use in Section 4. We discuss the policy preferences and tradeoffs in Section 5. In Section 6, we report our in-sample findings. In Section 7, we carry out an out-of-sample exercise using the GFC as a case study. Section 8 concludes.

2 Related literature

This paper bridges the academic and policy debates on the role of credit gaps for predicting financial crises, by integrating the modelling choices associated with trend-cycle decomposition methods into the design of early warning systems. Early studies include Minsky (1977) and Kindleberger (1978), followed by Kiyotaki and Moore (1997), Kaminsky and Reinhart (1999), Chang and Velasco (2001), Lowe and Borio (2002a), among others. Following the GFC, a massive literature on financial stability blossomed. A key feature of this body of work is the concept of credit booms (Reinhart and Rogoff, 2009; Schularick and Taylor, 2012; Miranda-Agrippino and Rey, 2020) and credit cycles, also referred to as financial cycles.

Borio (2014) reports salient stylized features of credit cycles. First, these cycles are most parsimoniously characterized in terms of credit (loans and bonds) extended to the nonfinancial private sector (households and corporations). Second, credit cycles have a much lower frequency than business cycles (see Borio, 2014; Drehmann et al., 2012). Borio (2014) additionally points out that credit cycle peaks are closely associated with the onset of the financial crises, and that they help detect crises with a good lead in real time. Schularick and Taylor (2012) focus on bank credit, and find a similar link between loan growth and financial crises. Credit cycles are often characterized in terms of the credit gap– the deviations of the credit-to-GDP ratio from its long-run trend. Recovering this long-run trend, of course, requires using trend-cycle decomposition techniques.

For example, when constructing the credit gap, Drehmann, Borio, and Tsatsaronis (2011) propose using Hodrick and Prescott (1997) for trend-cycle decomposition of the credit-to-GDP ratio, while fixing the smoothing parameter (λ) at 400,000 to ensure a smooth trend. There is a debate over the use of the HP filter for decomposing macroeconomic time-series. For example, Harvey and Jaeger (1993) and Hamilton (2018) argue that it introduces spurious and unstable dynamics in the decomposed series. Hamilton (2018) proposes an alternative linear projection method (henceforth, the Hamilton filter). Drehmann and Yetman (2018) defend using the HP filter at least in the context of recovering trend and credit gap components of credit-to-GDP ratios. Hamilton and Leff (2020) argue that even for this purpose, the HP filter is inadequate. Similarly, Edge and Meisenzahl (2011) argue that the end-of-sample estimates of the HP-filtered credit-to-GDP trend are unreliable in real time, making them unsuitable for stabilization policy (which, of course, must be made in real time).[1]

We take a step back, and evaluate these modeling choices of trend-cycle decomposition methods in the context of constructing EWIs based on credit gaps. That is, we embed the choice of trend-cycle decomposition method into the design of EWIs. Importantly, for each trend-cycle decomposition method we consider, we show that the choice of smoothing parameter can improve the tradeoff between true positives and false positives of EWIs. In addition, we consider the robustness (or reliability) of EWIs constructed in real time to the arrival of new information, which previous studies have largely ignored. We show how a desire for robustness could improve the out-of-sample performance of EWIs. We also consider a range of preferences over false positives and false negatives of the EWIs.

A promising area of further research is the question of which financial or economic indicator(s) to use for designing EWIs. Several studies have shown that composite EWIs using multiple time series for each country predict crises more accurately than those based on a single indicator. Aldasoro, Borio, and Drehmann (2018), show that including measures of household and interna-

[1] In the case of the United States, they show that a number of trend-cycle decomposition techniques would have produced false positives that signalled excessively high levels of credit in 2001:Q4 and in 2003:Q2 in real time. With the benefit of hindsight, however, the full-sample estimates of these gaps would not have produced these false positives. Edge and Meisenzahl (2011) further estimate that the reduction in lending that would have been obtained were countercyclical capital buffers (wrongly) deployed in these quarters would have been substantial.

tional debt can also improve the performance of EWIs. Using a more data-intensive approach, Lee, Posenau, and Stebunovs (2020) aggregate vulnerabilities across several sectors of the economy into a single indicator (the LPS index). They use 17 to 30 different time series for each country, which significantly improves their accuracy relative to the baseline credit gap. However, given the "severe" accounting, reporting, and structural differences across countries, the LPS indicators cannot be used to make cross-country comparisons. Afanasyeva (2020) develops a Bayesian vector autoregression (BVAR) method to detect credit booms using monetary aggregates, asset prices, and measures of real economic activity, which provides a useful cross-check to the credit gaps derived from trend-cycle decomposition methods. Sarlin and von Schweinitz (forthcoming) use 14 observable macroeconomic and asset price measures to predict the probability of a crisis, defined as periods when the Financial Stress Index of Lo Duca and Peltonen (2013) exceeds its 90th percentile. They simplify the analysis by defining the thresholds for the crisis probabilities ex-ante, and achieve a similar or better out-of-sample performance compared to ex-post optimization of the threshold. They argue that optimized thresholds display unwarranted time variation with the arrival of new data, producing a disparity between in-sample and out-of-sample performance. By contrast, when constructing EWIs based on credit gaps, we find that our jointly optimized thresholds and smoothing parameters are fairly stable when new data is introduced to the estimation sample.

3 Data

Since the 1980s, non-bank credit has become increasingly important as a source of financing for both households and corporations. To construct our EWI, we therefore use 'total' credit to the non-financial private sector (households and non-financial corporations) from all sources, scaled by nominal GDP. These data are available from the Bank for International Settlements, at a quarterly frequency, for a sample of 33 advanced economies (AE) and emerging market economies (EM). Our sample period is 1952:Q1-2018:Q1.[2]

[2] Drehmann (2013) show that EWIs constructed using total credit perform better than those based only on credit provided by domestic banks. Consistent with this finding, Basel III guidelines for designing countercyclical capital buffers suggest using a measure of credit-to-GDP gap that includes "all credit extended to households and other non-financial private entities in an economy independent of its form and the identity of the supplier of funds" (Basel Committee on Banking Supervision (2010), p.10). Aikman, Haldane, and Nelson (2015), Gonzalez, Sousa

Following the literature, we define the credit gap as the deviation of the credit-to-GDP ratio from its long-run trend, estimated in real-time using the five trend-cycle decomposition methods mentioned earlier. Figure 1 compares the real-time and full-sample credit gap estimates from each method for the United States. Notice that the HP filter gap looks nearly identical to the one estimated using the Bayesian method except at the endpoint. The Hamilton filter gap is less smooth, but shows smaller differences between its full-sample and real-time estimates. We refer to this similarity between the real-time and full-sample gap estimates as robustness, which we define later.

Following Laeven and Valencia (2012), a systemic banking crisis is characterized by a surge in defaults in both corporate and financial sectors, sharp rise in non-performing loans, depletion of bank capital, and sharp declines in asset prices.[3] Our crises dates are based on definitions from Laeven and Valencia (2012) and Reinhart and Rogoff (2009) of systemic banking crises, and are taken from Drehmann et al. (2011). All told, our sample includes 39 crises for 29 countries. Of these 39 crises, 11 occurred in EMs, and the remaining 28 occurred in the AEs.

Of the 28 advanced economy crises in our sample, half occurred during the global financial crisis period of 2007-2008. Given the clustering of banking crises across advanced economies during this period, the GFC provides a natural setup for an out-of-sample analysis. We conduct this analysis using a shorter sample ending in 2004 to optimize our EWIs. We drop the EMs from this shorter sample because they did not experience a crisis in the 2007-2008 period.

As a robustness check, we repeat the analysis using a more restrictive definition of banking crises based on Drehmann and Juselius (2014) which excludes cross-border crises, resulting in a sample of 28 crises. The results are qualitatively similar.

4 Implementing the Early Warning Indicator

In this section, we take the following steps to construct EWIs (based on optimized credit gaps and thresholds), and evaluate their performance:

Gomes Marinho, and Alves de Vasconcellos e Lima (2017), and Aldasoro et al. (2018) use total credit (from banks and nonbanks) at quarterly frequency. This data is only available for post-WWII period.

[3] Laeven and Valencia (2012) define systemic banking crises using both quantitative data and a subjective assessment.

1. For each country, decompose the credit-to-GDP series into trend and cyclical components (the credit gap) using five different statistical methods. These methods are:

    (a) Hodrick and Prescott (1997) (HP) filter. For this method, smoothing parameter is λ which controls the smoothness of the trend. Example: Drehmann and Yetman (2018).

    (b) Hamilton (2018) filter. Smoothing parameter here is the number of periods ahead (h). Similar to λ in HP filter, higher values for this parameter lead to smoother trends. Example: Hamilton and Leff (2020).

    (c) Bayesian structural time series model (Bayesain STM). We adjust the error variance, V, which affects the size of the deviations of credit-to-GDP from its trend. Example: Gonzalez et al. (2017).

    (d) Growth rate of the credit gaps. The smoothing parameter in this case is the number of lags, q, or the size of the differencing window. Example: Schularick and Taylor (2012).

    (e) Moving average of the credit gaps. The smoothing parameter is the length of the moving average window, q.

    Instead of fixing the smoothing parameter for each method, as is generally done in the literature, we consider a wide range of values in order to assess how the smoothing parameter affects the performance of EWIs. Details of implementation of each method are available in Appendix 1: Statistical Methods for Trend-Cycle Decomposition.

2. Using the credit gaps, define the EWI as an indicator function that assumes the value of one when the gap exceeds a certain critical threshold, and zero otherwise.

3. Evaluate the performance of the EWI using standard criteria adopted from the signal extraction literature.

4.1 Constructing the early warning indicators

We build our early warning indicators using the signal extraction approach of KR, and described in Edison (2003) and Gonzalez et al. (2017) (hereafter GML). For a given monitored variable (in this case, the credit-to-GDP gap), a crisis warning signal is generated if the deviation of this variable from its long-run trend (estimated in real-time) breaches a certain critical threshold (θ). We use a 3-year forecasting horizon, following Drehmann (2013) and Aldasoro et al. (2018).[4] We do not consider signals issued in the two-year 'grace period' after the onset of a crisis, when many economic indicators (including the credit-to-GDP ratio) tend to behave erratically (Basel Committee on Banking Supervision, 2010).[5] We also ignore all signals issued during the first 10 years of data for each country, which we treat as the 'burn-in' period.

With a smoother trend and a higher threshold, the gap is likely to cross the threshold fewer times, but may remain above the threshold for a longer period of time due to its persistence. In contrast, with a rougher trend and a lower threshold, the gap is likely to cross the threshold more often, even if briefly, and generate more false positives. To avoid favoring such a gap which produces many false positive blips, we assume that the authorities will seriously consider deploying countercyclical policies when a signal is *first* issued. They then adopt a 'wait and see' approach to evaluate the effectiveness of their policy response, and during this period they simply ignore any additional signals.[6] Specifically, we assume the following:

Assumption 1. Let $gap_{c,t}$ denote the (real-time) credit-to-GDP gap for country c, estimated at quarter t using only information available at this time . The signal $EWI_{c,t}$ is turned 'on' (takes a value of 1) when the gap breaches the threshold θ, and it stays on for two years. If the credit-to-GDP gap does not breach θ, or if the signal occurs in the two-year 'grace period' after a crisis, the signal remains off. We can thus define $EWI_{c,t}$ as follows.

$$EWI_{c,t} = \begin{cases} 1 & \text{if for any h} \in [0,7], \ \ gap_{c,t-h} \geq \theta \text{ and there was no crisis in country c} \\ 0 & \text{otherwise} \end{cases}$$

[4] These EWIs are intended to help policymakers time the deployment of macroprudential policies, which generally affect credit dynamics with substantial lags. Thus, the signals should arrive with ample warning.

[5] As a robustness check, we repeat the analysis and produce similar findings by extending the grace period to 4 years post crisis.

[6] The Appendix 2 illustrates the implications of this assumption through an example using Swedish data.

4.2 Evaluating the performance of the EWI

Once the signal is issued, we summarize its performance using the two-by-two matrix shown in Table 1, adapted from KR.

	Crisis within 12 quarters	**No crisis within 12 quarters**
Signal issued	a	b (type-1 error, false pos.)
No signal issued	c (type-2 error, false neg.)	d

Table 1: **EWI performance.** Type-1 and type-2 errors are defined under the null hypothesis that the country remains in a tranquil state (no crisis).

Let the null hypothesis be that the country is in a tranquil state (no crisis).[7] At each quarter, if the signal is followed by an actual crisis within a specified forecasting horizon (12 quarters), it is considered a true positive, and the entry a in Table 1 increases by one unit. When the signal is not followed by a crisis during that same horizon, it is considered a false positive (or a type-1 error), and b increases by one. If no signal is issued, and a crisis occurs within the forecasting horizon, c increases by one unit, indicating a false negative (or type-2 error). If no signal is issued, and no crisis occurs within the forecasting horizon, d increases by one, indicating a true negative. Thus, a perfect EWI would produce signals that fall exclusively under entries a and d.

To evaluate the performance of our EWIs, we define T_1 as the fraction of type-1 errors relative to the number of non-crises dates (false positive rate). Similarly, T_2 is the fraction of type-2 errors relative to the number of crises in our sample (false negative rate).[8] In terms of Table 1,

$$T_1 = \frac{b}{b+d} \qquad \text{and} \qquad T_2 = \frac{c}{a+c}. \tag{1}$$

In addition to the false positive and false negative rates, we also compare the performance of our EWIs using two other commonly used metrics, accuracy and predictive power, defined below:

[7] We depart from earlier studies who define the null to be the crisis state (Edison, 2003; Gonzalez et al., 2017; Borio and Drehmann, 2009). Consequently, our type-1 errors are false positives, and our type-2 errors are false negatives.

[8] When computing the false negative rate, KR define $a + c$ as as the number of quarters for which a correct signal is possible. If, for example, a country has two crises, and with a three year forecasting horizon, a perfect signal would have flashed on exactly 24 times, once for each of the 12 quarters in the forecasting horizon preceding each crisis, so $a + c = 24$. We follow Drehmann et al. (2011) and define $a + c$ as the number of crises (2 in this example).

Definition 1. *Accuracy.* The accuracy, A, of the signal is determined by the noise-to-signal ratio (NSR), which is the ratio of false positives to true positives.

$$\begin{aligned} NSR &= \frac{\#\ \textit{false positives}}{\#\ \textit{non-crisis episodes}} \bigg/ \frac{\#\ \textit{true positives}}{\#\ \textit{crisis episodes}} = \frac{T_1}{1-T_2} \qquad (2) \\ \text{A} &= 1 - NSR. \qquad (3) \end{aligned}$$

With a high value for the critical threshold (θ), the EWI will only be triggered when the credit-to-GDP gap is high, which is more likely to signal that the economy is in the midst of a significant credit boom. Following such a boom, a crisis is more likely to occur (Schularick and Taylor, 2012). Therefore, the true positive rate tends to increase with the critical threshold, boosting the accuracy of the signal.

Definition 2. *Predictive power.* Predictive power (P) is the ratio of true positives to the total number of crisis episodes.

$$P = \frac{\#\ \textit{true positives}}{\#\ \textit{crises}} = 1 - T_2$$

Predictive power decreases with the threshold because as the threshold increases, it is less likely that the gap will exceed the threshold, and thus fewer signals are issued.

We are interested in how sensitive our credit gap estimates, and as a result our EWIs, are to the arrival of new information. To that end, we estimate robustness (ex-post) as the difference between the real-time and full-sample gap estimates. Our measure of robustness, R, aggregates the difference between the real-time and full-sample gap estimates across countries and over time. It only applies to the HP Filter, Hamilton Filter, and Bayesian STM methods because the unobserved trend estimated by these methods is updated with the arrival of new information.

To estimate robustness, we first compute the absolute value of the difference between the real-time and the full-sample gaps at each time t and sum that difference over the entire sample period. We then repeat this exercise for all countries in the sample and sum the results to produce a single aggregated value. We normalize the value and take the difference from 1. Given a panel of C countries, robustness is defined as

Definition 3. *Robustness.*

$$R = 1 - \frac{\sum_{c=1}^{C}\sum_{t=1}^{N}|gap_{c,t}^{full} - gap_{c,t}^{rt}|}{2 * \sum_{c=1}^{C}\sum_{t=1}^{N}|gap_{c,t}^{full}|} \tag{4}$$

where at time t, $gap_{c,t}^{full}$ and $gap_{c,t}^{rt}$ represent the full-sample and real-time gaps for country c, respectively.

5 Policy preferences and tradeoffs

False positives (type-1 errors) are costly because they could result in the erroneous deployment of macroprudential policies that could hamper the provision of credit and adversely affect economic activity. That said, policymakers may reasonably be concerned about false negatives (type-2 errors), because failing to predict a crisis that later materializes may have dire economic consequences.[9] However, if the preference for minimizing false positives is high enough, the optimized EWI will feature a high threshold, resulting in too few signals issued, and thus a low true positive rate. This in turn leads to an unacceptably low predictive power.

We assume that the policymaker cares about the tradeoff between type-1 and type-2 errors, and robustness.[10] That is, we assume the policymaker minimizes the loss function (L) by choosing the threshold level (θ), and the smoothing parameter (ρ), subject to the constraint that at least two-thirds of crises are predicted ($P \geq \frac{2}{3}$), as in Borio and Drehmann (2009). We specify the loss function as

$$min_{\theta,\rho}[L] = w(1-R) + (1-w)(\alpha T_1 + (1-\alpha)T_2), \quad |P \geq \frac{2}{3} \tag{5}$$

where R is the measure of robustness (ex-post), and w is the preference weight on R relative to the weighted sum of type-1 and type-2 errors, and α is the weight on type-1 errors. For the in-sample analysis, we assume no preference for robustness ($w = 0$). We introduce a preference for robustness

[9] Betz et al. (2014) also explore the tradeoff between minimizing type-1 and type-2 errors.

[10] Demirgüç-Kunt and Detragiache (1998), Borio and Drehmann (2009), and Sarlin (2013) also consider minimizing a loss function that is the weighted sum of type-1 and type-2 errors, but they do not consider a preference for robustness. Borio and Drehmann (2009), Edison (2003), and Kaminsky and Reinhart (1999) minimize the noise-to-signal ratio, which is a function of type-1 and type-2 errors. However, that approach often results in an unsatisfactory low percentage of crises predicted.

when we perform an out-of-sample analysis in Section 7. Using a grid search, we optimize the values of θ and ρ under each decomposition method to minimize the loss function (5), for different values of α.[11]

We present an example of this minimization exercise for loss function (5) in Figure 2, using EWIs based on gaps generated by the HP filter for the advanced economy sample. We set $\alpha = 0.5$ and $w = 0$, implying equal weights on type-1 and type-2 errors and no weight on robustness in equation (5). The policy variables–smoothing parameter ρ and threshold θ–range between 200,000 to 600,000, and 8 to 12, respectively. The surface shows that after conducting this grid search, the loss function is minimized at $\rho^* = 221,000$ and threshold $\theta^* = 8.7$. The surface displays a number of local minima. This implies that conducting a grid search over a limited range for policy variables will likely lead to identifying only the local minima. To avoid this pitfall, we consider wide ranges of possible values for the policy variables.

5.1 The importance of the smoothing parameter

Previous attempts to design EWIs using credit gaps have fixed the smoothing parameter in the trend-cycle decomposition. For example, Borio and Drehmann (2009), Drehmann et al. (2011), and Basel Committee on Banking Supervision (2010) use the one-sided Hodrick and Prescott (1997) filter with smoothing parameter $\lambda = 400,000$. In their Bayesian framework, GML also ensure a smooth trend by setting the parameter that governs the variance of the state equation (σ^2_{w1} in equation 13) to zero. Why does the choice of smoothing parameter matter?

Smoothing parameter affects tradeoff between true positives and false positives

Recall that the ideal EWI does not produce any false positives (type-1 errors) in its prediction of crises, or false negatives (type-2 errors). A common approach to visualizing this tradeoff is by comparing the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) for all thresholds. The TPR is the ratio of true positives to the number of crisis episodes (1-T_2). The FPR is the ratio of false positives to non-crisis

[11] For the grid search, θ ranges between 0.00 and 15.00, with a step size of 0.1. The ranges and step sizes for smoothing parameter depend on the decomposition method. For the HP filter, λ ranges between 1,000 and 1,100,000, with a step size of 10,000. For the Hamilton filter, the smoothing parameter (lead length) has a step size of 1 quarter. The step size for the smoothing parameter (V) in the Bayesian STM method is 100. For the growth rate and moving average we use a window step size of 2.

episodes (T_1). Figure 3 shows the ROC curves for the EWIs based on the different decomposition methods using different values for the smoothing parameter. A perfect signal would correspond to a point on the top-left corner, indicating only true positives, and no false positives. A random guess (coin toss), would result in a point along the 45-degree line. Points above the 45-degree line denote an improvement relative to the random guess.

For each plot in Figure 3, the end-points of the curves closest to the top-right corner denote the true and false positive rates when the threshold is zero. When the threshold is zero, the EWI will be triggered many times, resulting in a high false positive rate. At the same time, the true positive rate will also be high, because nearly all crises in our sample will have been preceded by a positive signal. As we raise the threshold, the signals are generated less frequently, but are more accurate, resulting in a lower false positive rate. The true positive rate also declines, because with fewer signals, some crisis are not predicted (more false negatives). As evidenced by the thicker purple ROC curves, a higher smoothing parameter generally shifts the ROC curves out because the true positive rate drops off more slowly as the threshold increases.

A number of recent studies use the Area Under Receiver Operating Characteristic curve, or AUROC, as a summary statistic for the ability of credit-based measures to predict financial crises (a better tradeoff between type-1 and type-2 errors results in a higher AUROC value).[12] By using the AUROC to compare the performance of various EWIs, the researcher avoids having to make an assumption about the policymaker's preferences over type-1 and type-2 errors. However, this simplification comes with a cost, because the AUROC embodies tradeoffs that policymakers would consider inadmissible. In other words, it embodies tradeoffs over type-1 or type-2 errors that would only arise if the policymaker had extreme preferences for one over the other (equivalent to α values close to zero or one in equation 5). For example, the left-most region of the ROC curve (produced by a high threshold) is one where the true positive rate (and the predictive power) of the signal is unacceptably low; few signals are generated, and thus the vast majority of crises are missed. Furthermore, in practice, AUROC values are computed ignoring the choice of smoothing parameter as a control variable, producing sub-optimal results. This practice, in turn, has led to some confusion about which trend-cycle decomposition method produces the best credit gaps for

[12] See, for example, Schularick and Taylor (2012), Herrera et al. (2020) Drehmann and Juselius (2014), Drehmann and Yetman (2018), and Hamilton and Leff (2020)

predicting crises.

The left panel of Figure 4 illustrates the gains from our "fully optimized" approach. For both the HP and the Hamilton filter methods, it compares the in-sample loss values from equation (5), optimized with respect to just the threshold (the "partial" optimization approach, denoted by the solid lines), with those obtained by optimizing with respect to both the threshold and smoothing parameter (the "full" optimization approach, denoted by the dashed and dotted lines). For all values of α, the HP filter produces lower in-sample loss values than their Hamilton filter-based counterparts. When fully optimized with respect to both smoothing parameter and threshold, the loss values are more similar, particularly when the preference for type-1 errors is lower ($\alpha < 0.5$). Thus, the performance differences between EWIs based on credit gaps derived using these two trend-cycle decomposition methods nearly vanish when the optimization is performed with respect to both the threshold and smoothing parameter.[13]

Smoothing parameter affects robustness

Robustness is solely affected by the choice of smoothing parameter and not by the threshold. Figure 5 illustrates how the choice of smoothing parameter affects robustness across different decomposition methods. By construction, the gaps computed using the growth rate and moving average approaches are perfectly robust ($R = 1$) because they are invariant to using real time or full sample data. For low values of the smoothing parameter, the Hamilton filter is very robust to the arrival of new information. However, the robustness of the Hamilton filter declines with the smoothing parameter. In contrast, while the Bayesian STM and HP filter methods are less robust, robustness is increasing with the smoothing parameter for both methods.

6 In-sample results

We begin by assessing the in-sample performance of our EWIs for the advanced economies in our sample. The left panels in Figure 6 show the false positive rate, false negative rate, accuracy, and loss function values of our EWIs optimized with respect to both θ and ρ, and for different preferences over α. For now, we assume the policymaker has no preference for robustness ($w = 0$).

[13] In the right panel, we relax the constraint in the loss function that the predictive power should be greater than or equal to 2/3. Relaxing this constraint allows optimized Hamilton filter-based losses to get closer to HP filter-based losses for $\alpha > 0.70$.

For comparison, we also include a horizontal line showing these same metrics using the HP Filter specification of Borio and Drehmann (2009) which assumes a smoothing parameter of $\lambda = 400,000$ and threshold of $\theta = 9$ percentage points of GDP. We refer to this specification as the 'baseline HP.'

For $\alpha \geq 0.4$, the HP Filter and Bayesian DLM optimized with respect to both θ and ρ have the lowest false positive rate compared to the other methods, resulting in a higher accuracy, and lower values for the loss function specified in Equation (5). When α is less than 0.4, all optimized methods yield a similar performance, which is notably better than that of the baseline HP filter.

6.1 Practical implications

EWIs do not accurately predict crises in EMs

Our modelling approach has an important practical implication: Our EWIs are much less accurate in predicting crises in the EMs (right panels of Figure 6). Among the decomposition methods, the moving average and growth rate deliver the highest accuracy for EMs, and have the lowest loss function values. By contrast, the baseline HP has an unacceptably high false negative rate (off the chart), which implies a predictive power of less than 50 percent (this would violate our constraint that $P \geq \frac{2}{3}$). All told, compared to the advanced economy results, the EM sample delivers much higher false positive rates, and thus poor accuracy.

Why do our EWIs perform so poorly in the EM subsample? As mentioned in GML, could it be that the shorter length of the EM data series is biasing the estimated trends? We think not. Our EM series have an average data length of 43 years (compared to 53 years for the AE series), which seems adequate. Likewise, the frequency of crises is similar across both groups: the ratio of quarters per crisis for the AE and EM subsamples are 167 and 172 respectively. As a counterfactual, we repeat the optimization exercise for the AEs using a series that have been shortened to match the average length of the EMs. We also try shortening the AE subsample series so that not only the average length but also the distribution of lengths across countries closely matches that of the EM subsample. In both exercises, the loss function values obtained using the truncated AE sample are still much lower than those of the EM subsample. This evidence suggests that the differences between the AE and EM subsample results shown in Figure 6 are not due to inadequate data length.

Financial markets in the AEs are generally better developed than those in the EMs (see Committee on the Global Financial System, 2020). The lower levels of financial market development in the EMs could result in a different underlying data generating process for the credit cycle, affecting the performance of gap-based EWIs. An alternative explanation could be that these differences reflect the importance of other vulnerabilities (besides the buildup of credit) in explaining EM crises, such as large and persistent current account or government deficits, or high levels of debt denominated in hard currencies, or even the political environment (Herrera et al., 2020). Clearly, additional research is needed to develop more accurate EWIs for EMs.

Optimized thresholds are similar across methods

Another practical implication of our modelling approach is that the optimized thresholds provide insights into how high the credit gap could reach before triggering financial stability concerns. Focusing on the AE sample, the top panel of Figure 7 shows the optimized thresholds across decomposition methods for different policy preferences over α, when $w = 0$. Interestingly, when $\alpha \geq 0.6$, the optimized threshold across all methods lies in the fairly narrow range between 12 and 14 percentage points of GDP. Thus, if the policymaker has a strong preference against false positives, the choice of critical threshold is broadly similar across all methods.

Why is this the case? A higher threshold results in a lower false positive rate. Thus, a stronger preference against false positives (greater value of α in the loss function) results in a higher optimized threshold. However, this higher threshold also produces a higher false negative rate, which in turn lowers the predictive power (recall that $P = 1 - T_2$). As a consequence, the threshold can only increase up to a certain point before the constraint that the predictive power cannot drop below two-thirds binds. Therefore, as we increase α, the thresholds eventually converge around this narrow range of 12 to 14 percentage points of GDP.

7 Out-of-sample performance: Predicting advanced economy banking crises in 2007-2008

Could our EWIs have predicted which countries experienced a banking crisis in 2007-2008? The global financial crisis of 2007-2008 was essentially an advanced economy phenomenon. None of the emerging market economies in our sample experienced a banking crisis during that period. By contrast, according to Laeven and Valencia (2012), 14 advanced economies experienced a banking crisis. Leading up to the GFC, credit grew rapidly in many advanced economies, such that on the eve of the crisis, most had elevated credit gaps.

We evaluate how well our EWIs could have predicted which of the advanced economies experienced a banking crisis in 2007-2008. To that end, we conduct both an out-of-sample and pseudo-out-of-sample exercise. In the out-of-sample exercise, we optimize the EWIs using data through 2004 (3 years before the onset of the GFC). In the pseudo-out-of-sample exercise, EWIs are optimized using data through 2018. We then use these real-time credit gaps to construct EWIs on the eve of the global financial crisis, and evaluate for which countries the EWI was triggered, and if this was a true positive or a false positive.

Table 2 presents the results assuming $\alpha = 0.5$ (equal weight for type-1 and type-2 errors in loss function) and $w = 0$ (zero weight on robustness). We report the pseudo-out-of-sample results on the left-hand side, and the out-of-sample results on the right-hand side.

The GFC looks like a "credit boom gone bust"

As shown by the first column, the gaps estimated using the baseline HP filter exceeded the threshold of 9 percent of GDP for 13 of the 21 advanced economies in our sample (highlighted in bold red font). The other trend-cycle decomposition methods also show that many of the advanced economies had elevated credit gaps just before the GFC (both in the pseudo- and out-of-sample calibrations). Thus, from a credit perspective, the GFC looks like a "credit boom gone bust." Indeed, the banking crises in countries such as the United States and the United Kingdom were followed by a prolonged period of financial deleveraging, particularly by households (Justiniano et al., 2015; Sanchez, 2014).

When optimized assuming $\alpha = 0.5$, these methods generally have a similar performance

out of sample

The set of countries for which the EWI is triggered does not change much with either pseudo-out-of-sample or out-of-sample calibrations. In the case of the HP filter, the higher smoothing parameter and threshold used in the out-of-sample exercise did not change the number of true positives and false positives. In the case of the Bayesian STM and growth rate methods, the optimized thresholds and smoothing parameters are similar using both samples, and therefore produced similar results. The Hamilton filter produces considerably fewer true positives out-of-sample than pseudo-out-of-sample.

All methods produce the same false positives in pseudo-out-of-sample exercise for Australia, New Zealand, Italy, Norway, and Finland. These countries had elevated credit gaps, but did not experience a crisis in 2007-2008.

With the preference parameters we have assumed in the loss function ($\alpha = 0.5$ and $w = 0$), the Hamilton filter performs marginally better during the 2007-2008 crisis, with an overall lower combination of type-1 and type-2 errors. The EWIs derived from the growth rate, Bayesian STM, baseline HP, and optimized HP methods perform similarly, and reasonably well. The MA method has a good performance in pseudo-out-of-sample exercise, generating 10 true positives. However, in out-of-sample calibration (using the sample ending in 2004), the optimized threshold for the MA method is very small, which results in many true positives, but also many false negatives. In sum, the EWIs were able to predict most of the countries which experienced a financial crisis during the GFC (high true positive rate), but they also generated a significant amount of false positives.

Comparing the out-of-sample optimized threshold (θ^*) and smoothing parameter (ρ^*) in the bottom portion of Table 2 with their pseudo-out-of-sample counterparts allows us to assess how stable they are to the arrival of new information. The optimized thresholds and smoothing parameters are quite similar across the out-of-sample and pseudo-out-of-sample columns for the HP, Bayesian STM, and growth rate methods. By contrast, they change notably for the Hamilton and MA methods. Therefore, it appears that the optimization results for these two methods are more sensitive to the arrival of new information. Policymakers interested in using EWIs in real time may want to take this sensitivity into account.

Fully optimizing EWIs generally improves out-of-sample performance

We have shown that fully optimizing EWIs with respect to both the smoothing parameter (ρ) and the threshold (θ) improves their *in-sample* performance, at the least for the HP filter and the Hamilton filter (Figure 4). Even so, there is a debate in the literature about whether or not this type of ex-post optimization of EWIs improves *out-of-sample* performance (Sarlin and von Schweinitz, forthcoming). Figure 8 compares the out-of-sample loss function values of the fully-optimized EWIs (i.e. the dashed blue lines, optimized with respect to both smoothing parameter and threshold), with their partially-optimized counterparts (the solid black line, optimized with respect to the threshold only), and for different preferences over type-1 versus type-2 errors (α). As we did for Table 2, we optimize the EWIs using data through 2004 (3 years before the onset of the GFC), and use them to predict which of the advanced economies experienced a banking crisis in 2007-2008.

Fully optimizing the EWIs clearly improves out-of-sample performance for the Hamilton filter, for most values of α, as evidenced by the lower level of the dashed blue line relative to the solid black line. For the growth rate method, the out-of-sample improvement occurs for α greater than 0.5. For the HP filter method, such improvement only occurs when $\alpha = 0.7$. For the Bayesian STM and moving average approaches, fully optimizing EWIs does not improve out-of-sample performance.

A preference for ex-ante robustness can improve out-of-sample performance

Earlier we introduced the concept of robustness, which measures how sensitive our credit gap estimates, and as a result our EWIs, are to the arrival of new information (see Definition 4). We are interested in knowing whether an ex-ante preference for robustness can improve out-of-sample performance. To that end, we re-optimize the EWIs using a weight of 0.5 on robustness ($w = 0.5$) in the loss function equation (5). For a consistent comparison between the out-of-sample performance of the fully-optimized EWIs derived assuming a preference for robustness with that of the partially- and fully-optimized EWIs assuming no preference for robustness, we evaluate out-of-sample performance using only the average (weighted by α) of the type-1 and type-2 error rates. Recall that the growth and moving average methods are perfectly robust by construction, making their out-of-sample performance invariant to a preference for robustness.

A preference for robustness clearly improves out-of-sample performance for the Bayesian STM method relative to the fully-optimized version with no preference for robustness, regardless of α, as evidenced by the lower level of the dotted red line relative to the dashed blue line in Figure 8. It also improves out-of-sample performance relative to the partially-optimized version when α is less or equal to 0.6. A preference for robustness also improves out-of-sample performance for the HP filter, under most values of α. For the Hamilton filter, a preference for robustness improves out-of-sample performance only when α is less than 0.4. All told, taking into account our ex-ante measure of robustness in the optimization helps improve the out-of-sample performance of EWIs, making them more useful for predicting crises in real time.

Table 3 compares the levels of the loss function values (i.e. the α-weighted average of tpye-1 and type-2 error rates) shown in Figure 8. The EWIs derived using the moving average method perform best when $\alpha < 0.5$ (i.e. the policymaker is more cautious against false negatives). When the preference over false positives and false negatives is the same, both the moving average and the HP filter (fully optimized assuming a preference for robustness) produce the lowest loss values. The HP continues to dominate for α between 0.5 and 0.6. When $\alpha > 0.6$ (i.e. the policymaker is more cautious against false positives), the growth rate method performs best out of sample.

8 Conclusion

The credit gap is commonly used to measure the buildup of vulnerabilities stemming from excessive credit growth, and to characterize the credit cycle. Accordingly, the existing literature has proposed several methods to construct credit gap measures, primarily through a trend-cycle decomposition. These estimated gaps, in turn, have been used to design EWIs for predicting financial crises. In this paper, we carefully examine the tradeoffs that result from the modelling choices inherent in the design of gap-based EWIs. We show that optimizing EWIs with respect to both the threshold and the smoothing parameter, and allowing for different preferences over type-1 errors, type-2 errors, and robustness improves the out-of-sample performance of EWIs.

Moreover, we show that differences in performance between EWIs based on various partially optimized trend-cycle decomposition methods, which has been widely debated in the literature, largely disappear with full optimization. Among the five decomposition methods we consider, the HP filter and Bayesian STM offer the best overall performance in-sample when α is greater than 0.4, and perform similarly to the other methods when α is less than 0.4. The moving average method, however, performs best out-of-sample, when the policymaker is more cautious about false negatives. The fully optimized HP filter performs best out of sample when the policymaker is indifferent between false positives and false negatives, whereas the growth method performs best when the policymaker is more cautious against false positives. When we incorporate a preference for ex-ante robustness in the loss function, out-of-sample performance generally improves for the three parametric methods (HP, Hamilton, and Bayesian). Our optimized credit-gaps could easily be included in composite EWIs for predicting crises.

We uncover a number of practical implications. Among them, we find that if the policymaker has a strong enough preference for minimizing false positives, then regardless of the decomposition method used, the critical threshold for the credit gap is in the narrow range between 12 and 14 percentage points of GDP. In addition, we find that gap-based EWIs do not accurately predict crises in the emerging market economies. Thus, we recommend being cautious when using credit gaps to inform macroprudential policies in EMs. The stark differences between the EM and AE subsample results suggest that further research is needed on the behavior of credit cycles between these groups, and in understanding the role of other vulnerabilities in explaining EM crises. Finally, our results suggest that incorporating a preference for robustness of the estimated credit gaps to the arrival of new information in the optimization process can help improve their out of sample predictive power.

References

Afanasyeva, E., 2020. Can Forecast Errors Predict Financial Crises? Exploring the Properties of a New Multivariate Credit Gap. Finance and Economics Discussion Series 2020-045, Board of Governors of the Federal Reserve System, Washington, D.C., U.S.A.

Aikman, D., Haldane, A. G., Nelson, B. D., 2015. Curbing the credit cycle. the Economic Journal 125 (585), 1072–1109.

Aldasoro, I., Borio, C., Drehmann, M., 2018. Early warning indicators of banking crises: expanding the family. BIS Quarterly Review.

Basel Committee on Banking Supervision, 12 2010. Guidance for national authorities operating the countercyclical capital buffer. Tech. rep., Bank for International Settlements.

Betz, F., Opricǎ, S., Peltonen, T. A., Sarlin, P., 2014. Predicting distress in European banks. Journal of Banking & Finance 45, 225 – 241.

Borio, C., 2014. The financial cycle and macroeconomics: What have we learnt? Journal of Banking & Finance 45, 182 – 198.

Borio, C., Drehmann, M., March 2009. Assessing the risk of banking crises - revisited. BIS Quarterly Review.

Chang, R., Velasco, A., 2001. A model of financial crises in emerging markets. The Quarterly Journal of Economics 116 (2), 489–517.

Committee on the Global Financial System, 2020. Establishing viable capital markets. CGFS Papers No. 62, the Bank for International Settlements, Basel, Switzerland.

Demirgüç-Kunt, A., Detragiache, E., March 1998. The Determinants of Banking Crises in Developing and Developed Countries. IMF Staff Papers 45 (1), 81–109.

Drehmann, M., June 2013. Total credit as an early warning indicator for systemic banking crises. BIS Quarterly Review.

Drehmann, M., Borio, C., Tsatsaronis, K., 2011. Anchoring countercyclical capital buffers: The role of credit aggregates. International Journal of Central Banking 7 (4), 189–240.

Drehmann, M., Borio, C., Tsatsaronis, K., 2012. Characterising the financial cycle: don't lose sight of the medium term! BIS Working Papers 380, Bank for International Settlements.

Drehmann, M., Juselius, M., 2014. Evaluating early warning indicators of banking crises: Satisfying policy requirements. International Journal of Forecasting 30 (3), 759–780.

Drehmann, M., Yetman, J., 2018. Why you should use the Hodrick-Prescott filter – at least to generate credit gaps. BIS Working Paper Series 744, 1–21.

Edge, R. M., Meisenzahl, R. R., 2011. The unreliability of Credit-to-GDP Ratio Gaps in real time: Implications for countercyclical capital buffers. International Journal of Central Banking 7 (4), 261–298.

Edison, H. J., 2003. Do indicators of financial crises work? an evaluation of an early warning system. International Journal of Finance & Economics 8 (1), 11–53.

Gonzalez, R. B., Sousa Gomes Marinho, L., Alves de Vasconcellos e Lima, J. I., 2017. Re-anchoring countercyclical capital buffers: Bayesian estimates and alternatives focusing on credit growth. International Journal of Forecasting 33 (4), 1007 – 1024.

Hamilton, J. D., 2018. Why you should never use the Hodrick-Prescott filter. Review of Economics and Statistics 100 (5), 831–843.

Hamilton, J. D., Leff, D., 2020. Measuring the credit gap. Working Paper, University of California San Diego (), .

Harvey, A. C., 1985. Trends and cycles in macroeconomic time series. Journal of Business and Economic Statistics 3 (3), 216–227.

Harvey, A. C., Jaeger, A., 1993. Detrending, stylized facts and the business cycle. Journal of Applied Econometrics 8 (3), 231–247.

Herrera, H., Ordoñez, G., Trebesch, C., 2020. Political booms, financial crises. Journal of Political Economy 128 (2), 507–543.

Hodrick, R. J., Prescott, E. C., 1997. Postwar U.S. business cycles: An empirical investigation. Journal of Money, Credit and Banking 29 (1), 1–16.

Justiniano, A., Primiceri, G. E., Tambalotti, A., 2015. Household leveraging and deleveraging. Review of Economic Dynamics 18 (1), 3 – 20.

Kaminsky, G. L., Reinhart, C. M., 1999. The twin crises: The causes of banking and balance-of-payments problems. The American Economic Review 89 (3), 473–500.

Kindleberger, C. P., 1978. Manias, Panics, and Crashes: A History of Financial Crises, 1st Edition. Basic Books, New York, NY.

Kiyotaki, N., Moore, J., 1997. Credit cycles. Journal of Political Economy 105 (2), 211–248.

Laeven, L., Valencia, F., 2012. Systemic banking crises database: An update. IMF Working Paper (WP/12/163).

Lee, S. J., Posenau, K. E., Stebunovs, V., 2020. The anatomy of financial vulnerabilities and banking crises. Journal of Banking & Finance 112, 105334.

Lo Duca, M., Peltonen, T. A., 2013. Assessing systemic risks and predicting systemic events. Journal of Banking & Finance 37 (7), 2183 – 2195.

Lowe, P., Borio, C., December 2002a. Assessing the risk of banking crises. BIS Quarterly Review.

Lowe, P., Borio, C., 2002b. Asset prices, financial and monetary stability: Exploring the nexus. BIS Working Papers (114).

Minsky, H. P., 1977. The Financial Instability Hypothesis: An Interpretation of Keynes and an Alternative to “Standard” Theory. Challenge 20 (1), 20–27.

Miranda-Agrippino, S., Rey, H., 2020. U.S. Monetary Policy and the Global Financial Cycle. The Review of Economic Studies.

Petris, G., 2010. An R package for dynamic linear models. Journal of Statistical Software 36 (12).

Petris, G., Petrone, S., Campagnoli, P., 2009. Dynamic Linear Models with R, 1st Edition. Use R! Springer-Verlag, New York, NY.

Reinhart, C. M., Rogoff, K. S., May 2009. The aftermath of financial crises. American Economic Review 99 (2), 466–72.

Sanchez, J. M., 2014. The Deleveraging of U.S. Households Since the Financial Crisis. Economic Synopses, Federal Reserve Bank of St. Louis (5).

Sarlin, P., 2013. On policymakers loss functions and the evaluation of early warning systems. Economics Letters 119 (1), 1 – 7.

Sarlin, P., von Schweinitz, G., forthcoming. Optimizing policymakers loss function in crisis prediction: Before, within or after? Macroeconomic Dynamics, 1–24.

Schularick, M., Taylor, A. M., 2012. Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008. American Economic Review 102 (2), 1029–1061.

West, M., Harrison, P. J., 1997. Bayesian Forecasting & Dynamic Models, 2nd Edition. Springer Series in Statistics. Springer Verlag, New York, NY.

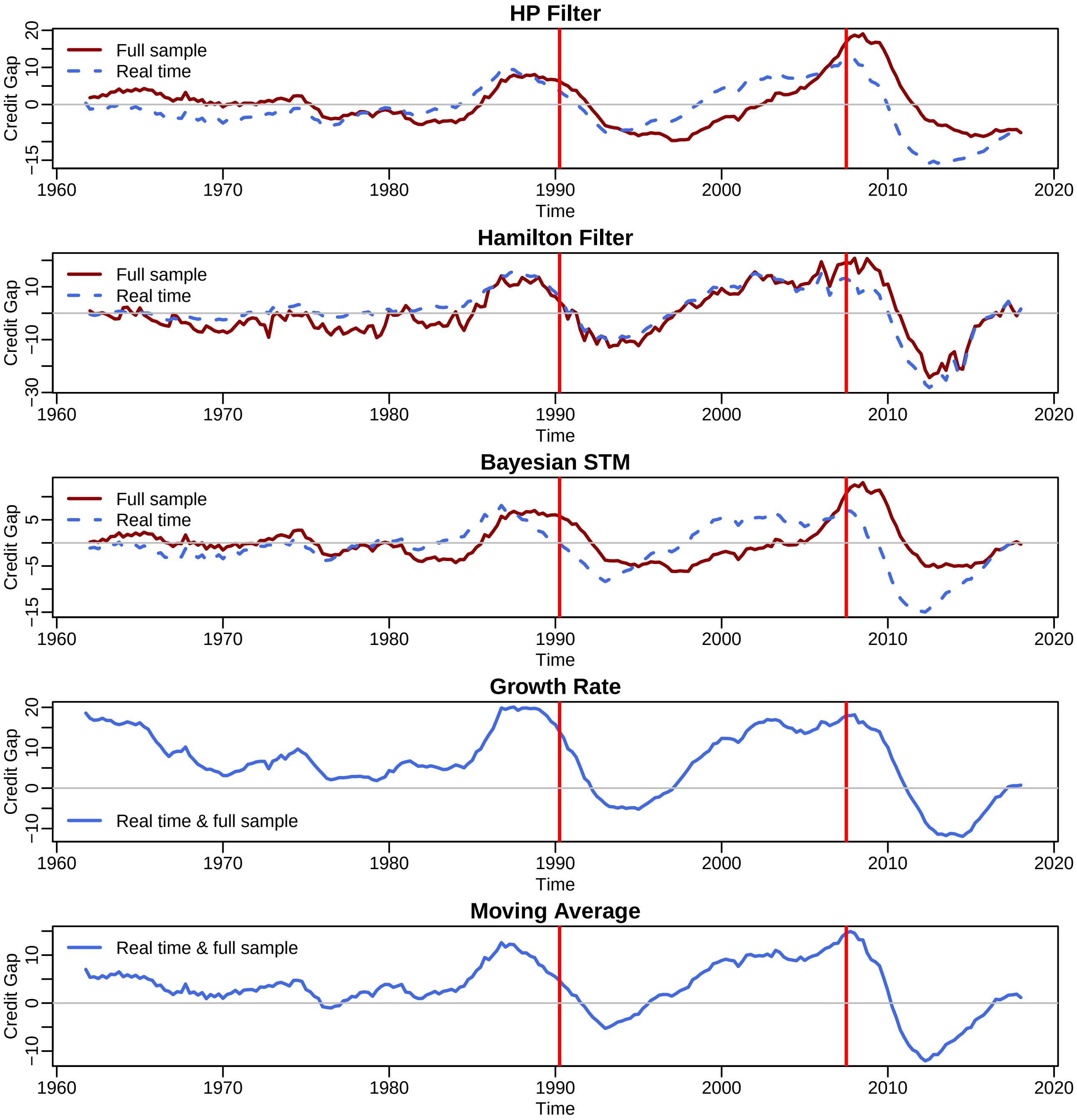

Figure 1: US credit cycle by trend-cycle decomposition method using literature-recommended policy values. Credit gap is expressed as percent of GDP. HP filter $\lambda = 400{,}000$. Hamilton filter $h = 20$. Bayesian STM $V = 600$. Growth rate window $= 21$. Moving average window $= 21$. Red lines denote the onset of a financial crisis.

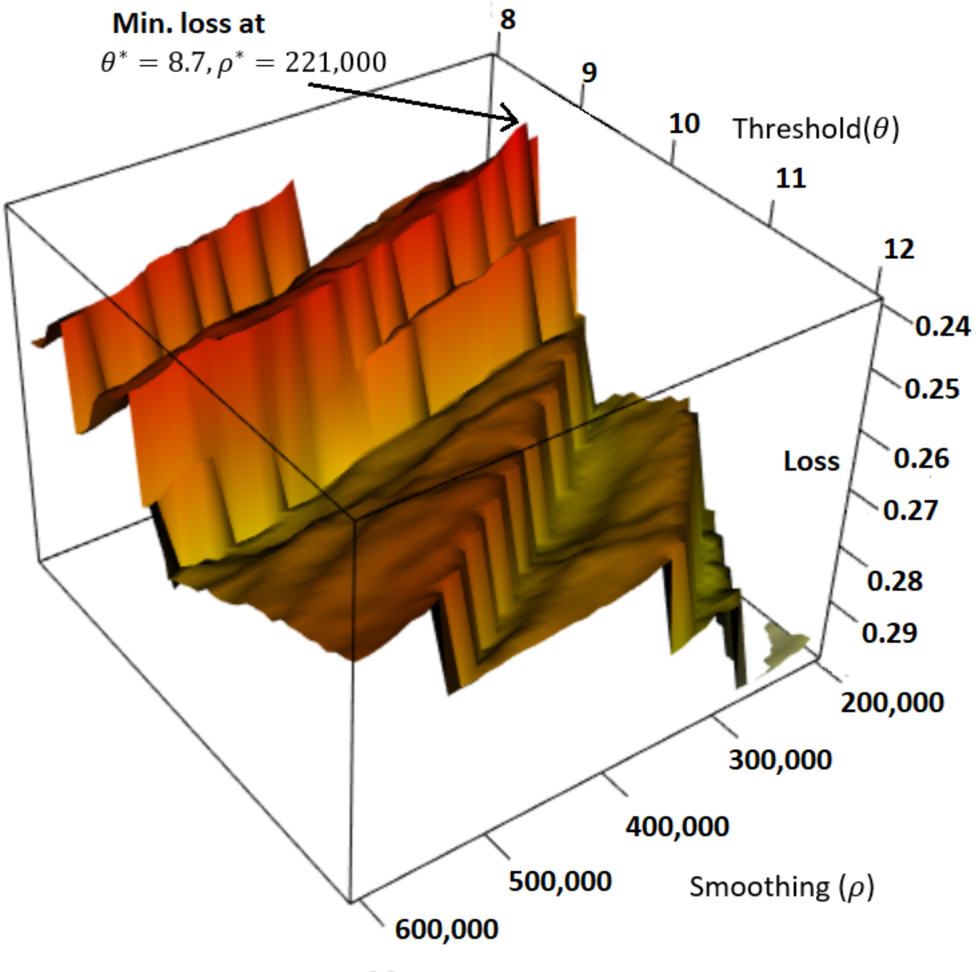

Figure 2: Loss function values for different combinations of the threshold and smoothing parameter, based on EWIs derived using HP filter, assuming $\alpha = 0.5$ and $w = 0$, and using the advanced economy sample.

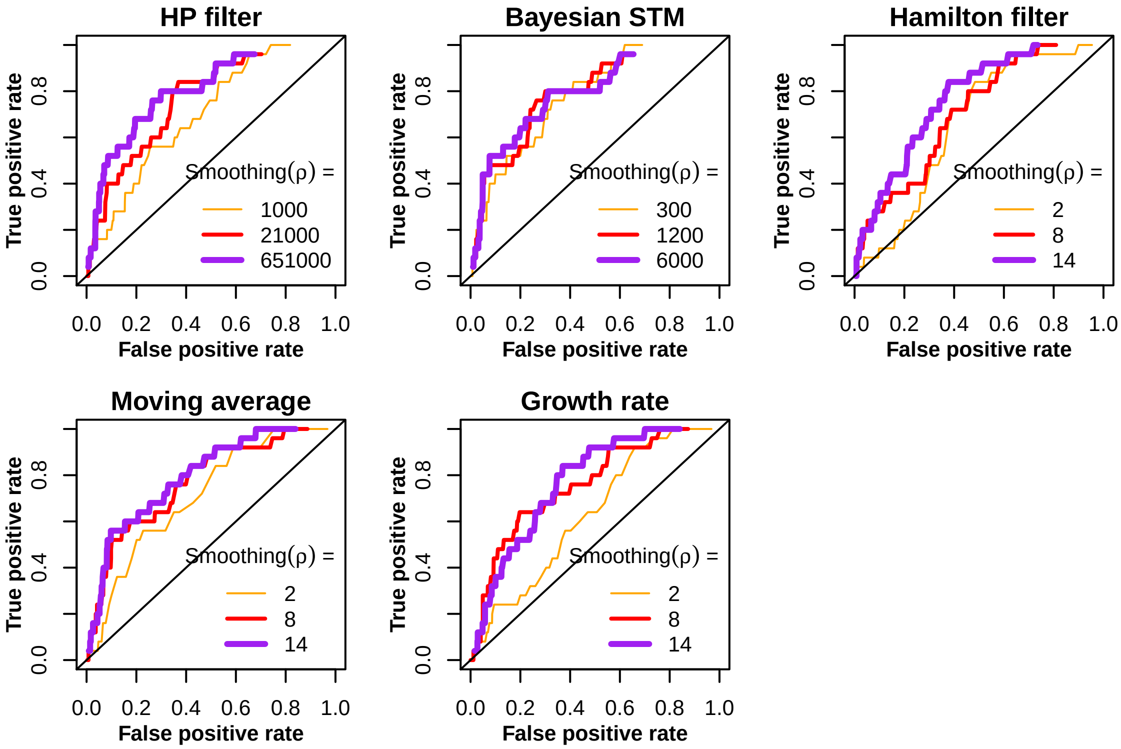

Figure 3: ROC curves for different values of the smoothing parameter, as the thresholds go from 15 (left-most point of each curve) to 0 (right-most point of each curve). A higher degree of smoothing results in a slower drop off in the true positive rate as the threshold increases.

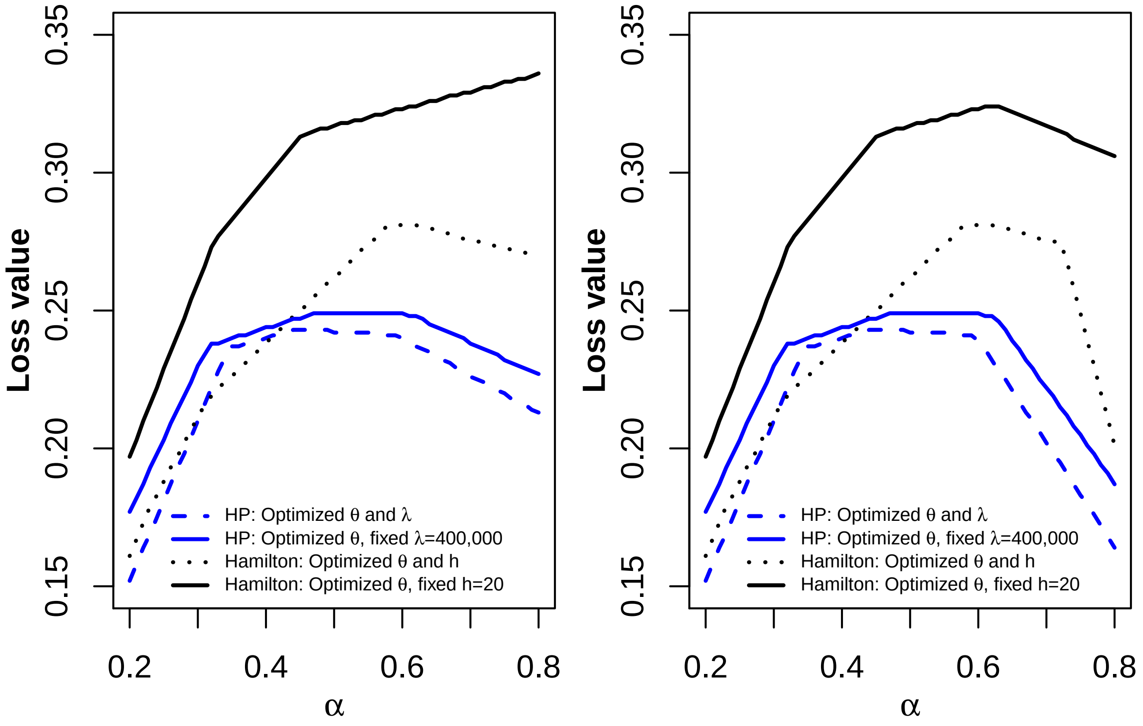

Figure 4: In-sample loss function values for partial and fully optimized HP and Hamilton filters. This chart compares fully optimized vs partially optimized HP and Hamilton Left panel: Predictive power constrained to be greater than 2/3. Right panel: No constraint on predictive power.

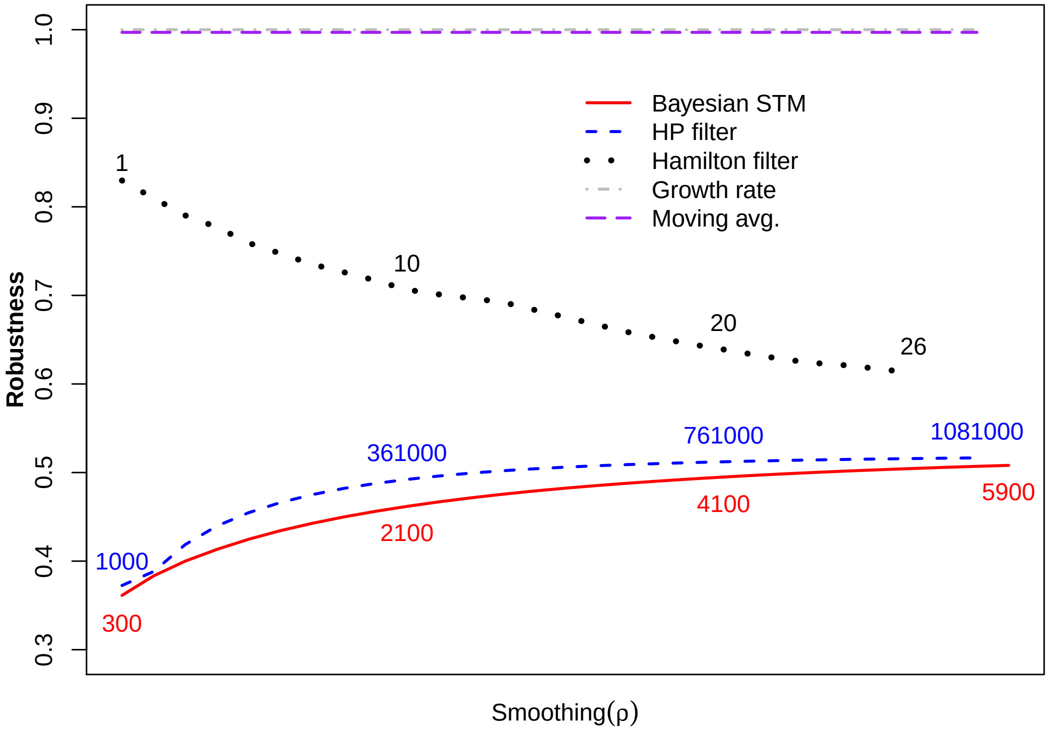

Figure 5: How smoothing affects robustness. Labels indicate value of the smoothing parameter. By construction, robustness for Bayesian STM and Growth rate methods are 1.

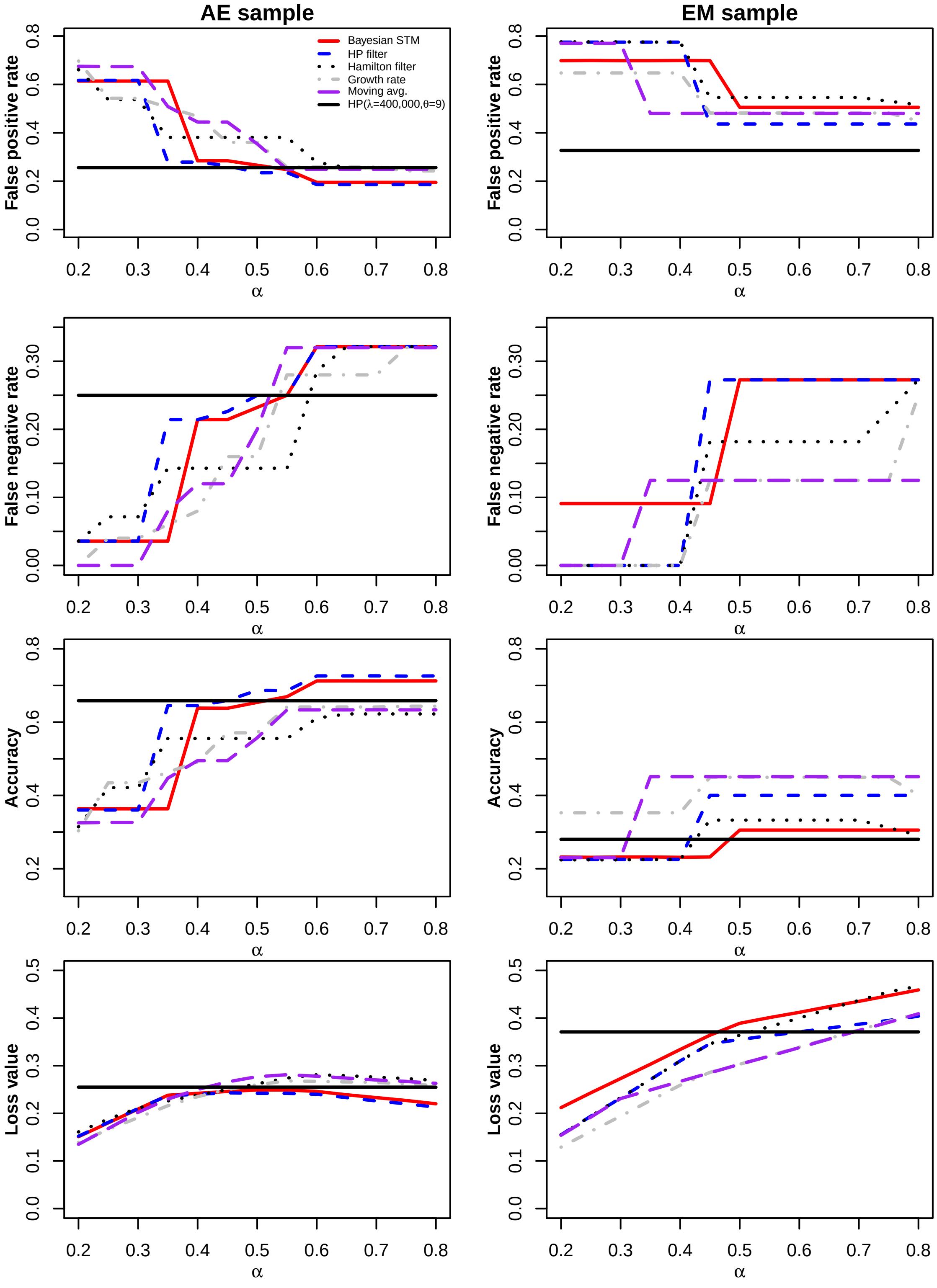

Figure 6: Tradeoffs between type-1, type-2, and accuracy for AE and EM samples. Left panels: AE sample. Right panels: EM sample.

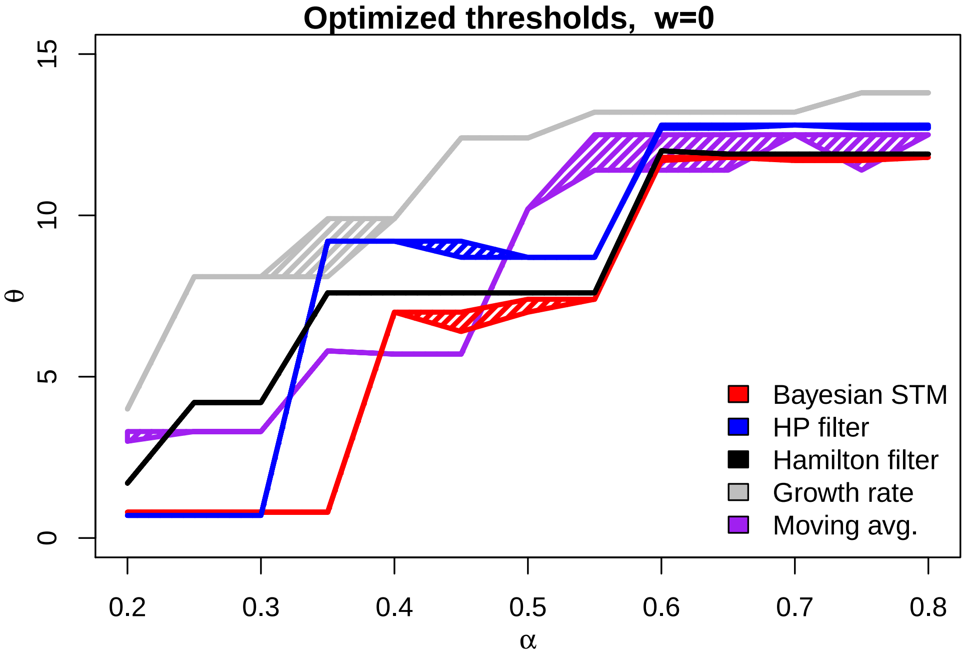

Figure 7: Optimized thresholds for different preferences over Type-1 errors using AE sample.

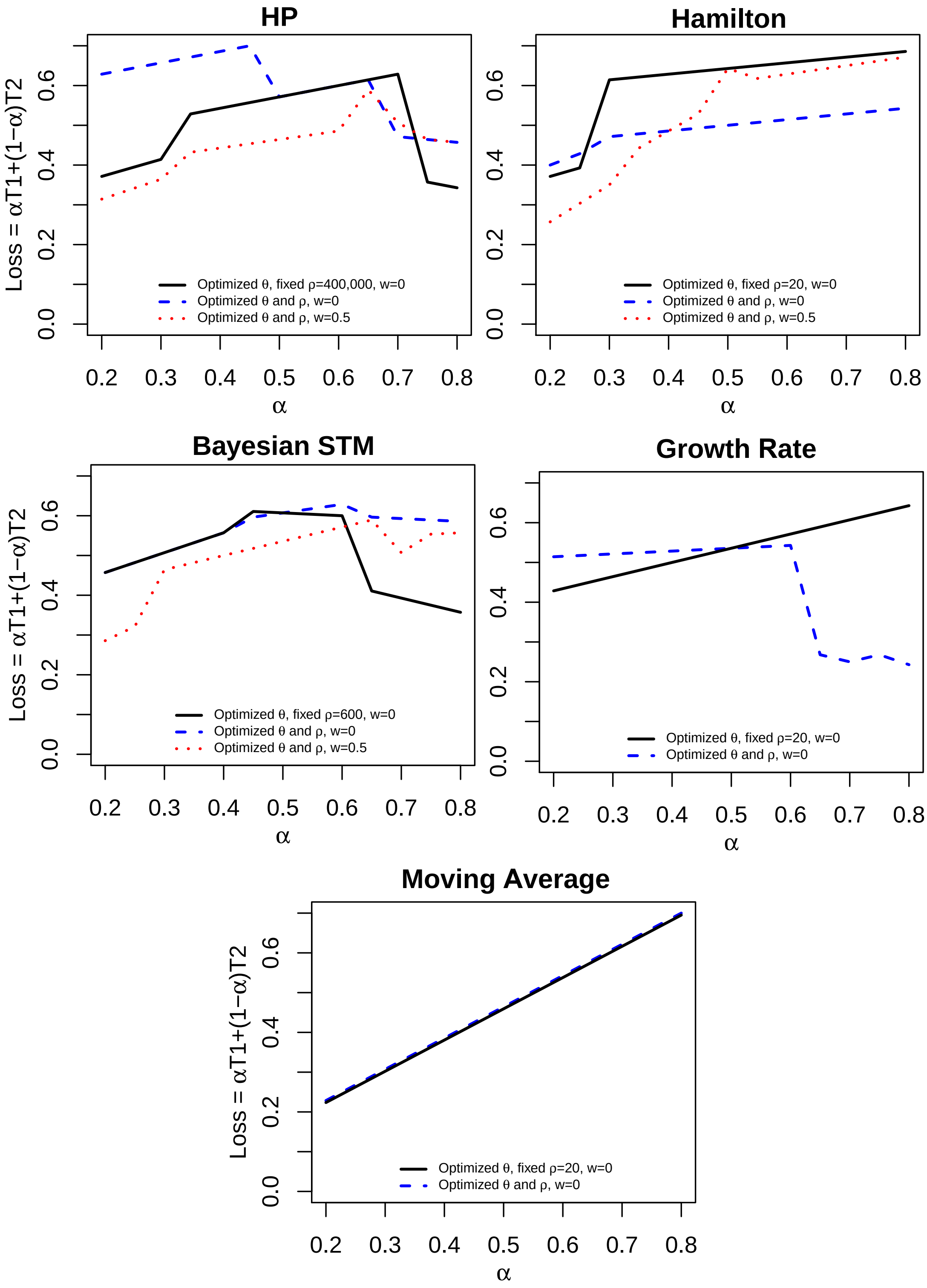

Figure 8: Out-of-sample values of the loss function, specified in equation (5), for partially and fully optimized smoothing parameters (ρ) and threshold levels (θ). Parameter α represents the weight assigned to type-1 errors in the loss function. When the policymaker values robustness, $w = 0.5$. If the policymaker does not value robustness, $w = 0$. For growth rate and moving average methods, changing w does not affect robustness. Values of fully optimized loss functions (dashed and dotted lines) below the partially optimized counterparts (solid lines) represent out-of-sample gains.

Table 2: Estimated credit gaps on eve of 2007-2008 global financial crisis (as percent of GDP) and EWI performance, $\alpha = 0.5, w = 0.0$.

Estimated credit gaps for advanced economies on eve of the Global Financial Crisis (Percent of GDP)											
		Pseudo-out-of-sample					Out-of-sample				
		$\alpha = 0.5, w = 0.0$					$\alpha = 0.5, w = 0.0$				
	Baseline HP	HP	Bayesian STM	Hamilton	Growth	MA	HP	Bayesian STM	Hamilton	Growth	MA
Crisis Countries											
Ireland	**57.2**	**50.0**	**43.0**	**39.8**	**58.2**	**52.4**	**60.0**	**41.9**	**21.8**	**51.4**	**46.4**
Portugal	**25.6**	**18.7**	**7.9**	**37.7**	**17.7**	**16.5**	**28.0**	6.9	**50.5**	12.6	**16.0**
Spain	**42.1**	**35.3**	**27.2**	**15.8**	**40.0**	**35.4**	**44.7**	**26.3**	**15.8**	**34.7**	**31.6**
Denmark	**34.5**	**30.4**	**25.3**	**35.5**	**28.0**	**28.5**	**36.0**	**24.7**	**54.4**	**23.9**	**25.8**
Greece	**24.4**	**20.3**	**15.8**	**17.5**	**42.4**	**19.2**	**26.0**	**15.2**	**36.5**	**35.7**	**16.9**
U.K.	**11.0**	**9.5**	**8.0**	**8.3**	**14.6**	**15.2**	**11.7**	**7.8**	**12.4**	**14.1**	**13.9**
U.S.	**10.4**	**9.1**	**8.1**	**12.0**	12.3	**11.3**	**10.9**	**7.9**	**18.8**	11.1	**10.3**
Belgium	7.8	5.5	**7.3**	**7.7**	**13.8**	**17.1**	8.9	7.1	-2.4	**14.2**	**16.4**
Sweden	**21.9**	**21.2**	**23.5**	**37.4**	**26.4**	**28.2**	**22.3**	**23.3**	**38.7**	**21.9**	**26.1**
France	5.6	5.1	5.3	**8.7**	10.5	8.3	5.8	5.2	9.0	8.5	**7.4**
Netherlands	3.2	2.9	2.5	**7.9**	11.2	**13.3**	3.3	2.5	-3.2	10.0	**12.0**
Austria	1.4	1.2	0.8	3.0	8.0	6.1	1.5	0.8	2.6	8.0	**5.6**
Germany	-1.7	-2.6	-4.7	-6.5	-1.7	-2.5	-1.4	-4.9	-1.1	-1.7	-2.6
Switzerland	-4.9	-2.6	-0.2	**7.8**	4.4	3.3	-5.8	0.0	-2.4	4.0	**2.8**
Non-Crisis Countries											
Australia	**17.3**	**15.9**	**13.4**	**17.4**	**20.0**	**18.1**	**17.8**	**13.1**	**22.0**	**17.3**	**16.2**
N.Z.	**13.6**	**13.2**	**12.5**	**20.2**	**25.0**	**21.6**	**13.8**	**12.4**	8.3	**22.5**	**19.3**
Italy	**12.7**	**10.5**	**8.8**	**17.3**	**21.7**	**11.5**	**13.7**	**8.6**	**29.2**	**18.8**	**10.2**
Norway	**12.6**	**13.1**	**12.5**	**22.5**	**15.6**	**13.2**	**12.3**	**12.3**	**30.1**	12.5	**12.5**
Finland	**10.7**	**12.2**	**12.6**	**22.2**	**17.2**	**12.4**	**10.1**	**12.6**	**21.5**	**13.4**	**11.1**
Canada	1.6	2.7	4.8	6.1	5.8	8.6	1.3	5.0	4.6	5.7	**8.3**
Japan	-17.7	-11.4	-4.4	0.4	-6.1	-4.3	-20.2	-3.7	-6.3	-3.2	-3.1
Optimization results											
Threshold (θ^*)	9	8.7	7	7.6	13	8.7	9.4	7.3	12	12.8	1.7
Smoothing par. (ρ^*)	400000	221000	1100	13	14	16	501000	1000	24	12	14
True Positive	8	8	9	12	8	10	8	7	8	7	13
False Positive	5	5	5	5	5	5	5	5	4	4	6
True Negative	2	2	2	2	2	2	2	2	3	3	1
False Negative	6	6	5	2	6	4	6	7	6	7	1
Type 1 Err. Rate	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.57	0.57	0.86
Type 2 Err. Rate	0.43	0.43	0.36	0.14	0.43	0.29	0.43	0.50	0.43	0.50	0.07

Notes: Colored numbers denote gaps that exceed the threshold optimized for each decomposition method. Crisis countries are those that experienced a financial crisis in 2007-2008. For each of these countries, the gap shown is the maximum of the credit gaps during the 3-year forecasting window just prior to the onset of the crisis. For countries that did not experience a crisis, the gap shown is the maximum of the credit gaps between 2004:Q3 and 2007:Q2. Type-1 error rate is the number of false positives divided by 7 (the number of countries that did not experience a crisis in 2007-2008). Type 2 error rate is the number of false negatives divided by 14 (number of countries that experienced a crisis in 2007-2008).

Table 3: Out-of-Sample Loss Function Values

	α	0.2	0.3	0.4	0.5	0.6	0.7	0.8	Average	Average ($0.3 < \alpha < 0.7$)
DLM	Full ($w = 0$)	0.46	0.51	0.56	0.61	0.63	0.59	0.59	0.57	0.58
	Partial	0.46	0.51	0.56	0.61	0.60	0.39	0.36	0.50	0.54
	Full ($w = 0.5$)	0.29	0.46	0.50	0.54	0.57	0.51	0.56	0.50	0.52
HP	Full ($w = 0$)	0.63	0.66	0.69	0.57	0.60	0.47	0.46	0.60	0.62
	Partial	0.37	0.41	0.54	0.57	0.60	0.63	0.34	0.50	0.56
	Full ($w = 0.5$)	0.31	0.36	0.44	**0.46**	**0.49**	0.51	0.46	**0.45**	**0.47**
Hamilton	Full ($w = 0$)	0.40	0.47	0.49	0.50	0.51	0.53	0.54	**0.49**	**0.50**
	Partial	0.37	0.61	0.63	0.64	0.66	0.67	0.69	0.61	0.64
	Full ($w = 0.5$)	0.26	0.35	0.49	0.64	0.63	0.65	0.67	0.53	0.55
Growth Rate	Full ($w = 0$)	0.51	0.52	0.53	0.54	0.54	**0.25**	**0.24**	**0.45**	**0.47**
	Partial	0.43	0.46	0.50	0.54	0.57	0.61	0.64	0.54	0.54
Moving Avg.	Full ($w = 0$)	**0.23**	**0.31**	**0.39**	**0.46**	0.54	0.62	0.70	**0.46**	**0.46**
	Partial	**0.23**	**0.31**	**0.39**	**0.46**	0.54	0.62	0.70	**0.46**	**0.46**

The table reports out-of-sample loss function values, defined as the weighted average of type-1 and type-2 errors (weighted by α values) for each trend-cycle decomposition method used to generate credit gaps. We use the GFC as our out of sample case study. Numbers in bold indicate the lowest loss value (the best out-of-sample performance) for a given α value.

Appendix

Appendix 1: Statistical Methods for Trend-Cycle Decomposition

The five trend-cycle decomposition methods we consider are discussed below.

Hodrick and Prescott (HP) Filter

The Hodrick and Prescott (1997) filter decomposes a time series y_t into a trend (g_t) and a cyclical component (c_t). Thus, we have $y_t = g_t + c_t$. The two-sided (full sample) HP filter solves

$$\arg\min \sum_{t=1}^{T} (y_t - g_t)^2 + \lambda \sum_{t=2}^{T-1} (g_{t+1} - 2g_t + g_{t-1})^2. \tag{6}$$

The smoothness of g_t is determined by the λ parameter; a larger λ yields a smoother trend. The one-sided (real time) version of the filter (using data up to time $T' \leq T$) solves a similar equation

$$\arg\min \sum_{t=1}^{T'} (y_t - g_t)^2 + \lambda \sum_{t=2}^{T'} (g_t - 2g_{t-1} + g_{t-2})^2. \tag{7}$$

which we recast in Harvey (1985) state-space representation:

$$\begin{aligned} y_t &= g_t + e_t && (8) \\ \Delta g_t &= \Delta g_{t-1} + v_t^2 && (9) \\ & \quad e_t \sim N(0,1), v_t \sim N(0, 1/\lambda) \end{aligned}$$

where Equation (8) is the measurement equation and Equation (9) is the transition equation. We estimate the state-space equations by applying the Kalman filter.

Hamilton Filter

The Hamilton (2018) filter uses ordinary least squares regression on four lags of the observed quarterly values of a data series plus a constant to predict the series h-steps ahead.[14] In this approach, we fit the following model to the data

$$y_{t+h} = \nu + \sum_{j=0}^{J} \beta_j y_{t-j} + u_{t+h}, \tag{10}$$

where $J = 4$. Thus, the residuals from the above regression constitute the cyclical component (or credit gap). That is,

$$\hat{u}_{t+h} = y_{t+h} - \hat{\nu} - \hat{\beta}_0 y_t - \hat{\beta}_1 y_{t-1} - \hat{\beta}_2 y_{t-2} - \hat{\beta}_3 y_{t-3}. \tag{11}$$

Similar to λ in the HP filter, higher values of h will result in a smoother trend. With quarterly data, Hamilton (2018) recommends $h = 8$ quarters for decomposing variables that co-move with the business cycle, and $h = 20$ quarters for variables that co-move with the credit cycle. We treat h as the smoothing parameter of choice. As with λ, the value of h will be guided by the performance of the EWI, and not by our prior on the duration of the credit-cycle. The full sample credit gap is

[14] See Hamilton (2018) and Drehmann and Yetman (2018) for details.

estimated using the entire time series to recover the regression coefficients ($\hat{\beta}_j$'s). For the real time credit gap, measured at time t, only data up until t is used to estimate the regression coefficients.

Bayesian STM

In contrast to the frequentist approaches discussed earlier, we also consider the Bayesian structural time series model (STM), discussed in Petris et al. (2009) and Petris (2010). The implementation requires a state vector $\zeta_t = (u_t, \beta_t)'$ where u_t is a time-varying local level (the trend) and β_t is the time-varying local growth rate. The cyclical component is $Y_t - u_t$. The dynamics are given by the equations

$$\begin{aligned} y_t &= u_t + v_t \quad v_t \sim N(0, V) &(12)\\ u_t &= u_{t-1} + \beta_{t-1} + w_{1,t} \quad w_{1,t} \sim N(0, \sigma^2_{w1}) &(13)\\ \beta_t &= \beta_{t-1} + w_{2,t} \quad w_{2,t} \sim N(0, \sigma^2_{w2}). &(14)\end{aligned}$$

Notice that the equation for y_t mirrors the HP filter specification for trend and cycle, $y_t = g_t + c_t$. We adjust the error variance V which affects the size of the deviations of credit-to-GDP from its trend.[15] In other words, V is the smoothing parameter for this trend-cycle decomposition method, analogous to λ in the HP filter.

We choose the system variances $\sigma^2_{w1} = 1$ and $\sigma^2_{w2} = 0.01$ where $W = diag(\sigma^2_{w1}, \sigma^2_{w2})$. However, this does not introduce assumptions into the trend because the behavior of Y_t is dependent on the ratio of W/V. We are simply limiting the dimensions of our optimization problem by fixing the numerator (W) and then varying the denominator (V).

We recover the trend u_t using the Kalman filter, a backward looking Bayesian updating algorithm. The outcome is qualitatively comparable to the one-sided HP filter. For a full sample (two-sided) trend-cycle decomposition, we use a Kalman smoothing algorithm on the Kalman filter output. These computations are performed using the functions `dlmFilter` and `dlmSmooth` in the `R` package `dlm` by Petris (2010).

Growth Rate

The HP Filter, Bayesian STM, and Hamilton Filter methods rely on estimating the long-run trend in the ratio of credit-to-GDP, which is an unobserved latent variable. We also consider a purely backward looking growth rate based only on observable data, obtained by differencing the credit-to-GDP ratio. While strictly speaking this is not a trend-cycle decomposition, this method removes time trends and yields a stationary cyclical component which resembles a credit gap. The growth rate will not change with the arrival of new information, making it perfectly robust (i.e. the real time and full-sample versions of the growth rate are identical). The 'smoothing parameter' in this case is the number of lags, q, or the size of the differencing window. Thus, for a series y_t in real time, a q-quarter differencing would yield:

$$GR_t = \frac{y_t - y_{t-q+1}}{y_{t-q+1}}. \tag{15}$$

[15] Alternatively, Petris et al. (2009) state that "the relative magnitude of V is an important factor that enters the gain matrix of the Kalman filter which, in turn, determines how sensitive the state prior-to-posterior updating is to unexpected observations."

Moving Average

We also detrend the credit-to-GDP series by differencing its level at time t from its q-quarter moving average. Similar to the growth rate method, the moving average approach removes time trends producing a stationary gap. Like the growth rate method, the moving average method is perfectly robust because it relies only on known observable information available at time t. At time t, the real-time difference of a series y_t from its q-quarter moving average is:

$$MA_t = y_t - \frac{\sum_{i=t-q+1}^{t} y_i}{q}. \tag{16}$$

The smoothing parameter is the length of the moving average window, q.

Appendix 2: Example

Example 1. Consider two different gap estimates for Sweden. The first (EWI-1) is based on the HP filter with a high degree of smoothing ($\lambda = 400,000$) and threshold of 5 percent of GDP (Figure 9, left panel). The other (EWI-2) also uses the HP filter, but with a lower smoothing parameter ($\lambda = 100$) and threshold of 3 percent of GDP (Figure 9, right panel). The threshold is marked by the dotted yellow line. The red shaded regions at the bottom of each panel denote quarters when the gaps exceed the threshold. When this signal arrives inside the 3-year forecasting horizon (the grey shaded region that precedes the crisis dates denoted by the blue line), it is considered a true positive. EWI-1 correctly signals the crisis in 1992 and the one in 2008. But it also issues false positives over several quarters on two separate occasions (in 2002 and in 2012). EWI-2 also correctly flashes on before the 2008 crisis, but fails to predict the 1992 crisis. Furthermore, EWI-2 generates false positives on three separate occasions. On the face of it, it would seem that EWI-1 is more accurate. Roughly speaking, it made 2 good crisis calls and 2 bad crisis calls (compared just 1 good call and 3 bad calls for EWI-2). But because EWI-2 produces fewer false positives measured in quarters, it has a *lower* noise-to-signal ratio (higher accuracy), which seems counterintuitive.

EWI-2 is inadequate for anchoring countercylical capital buffers, because it generated false positives on three separate occasions (even if briefly), which in turn could have resulted in the erroneous deployment of countercyclical policies.

Returning to Example 1, the purple shaded regions at the bottom of each panel in Figure 9 indicate periods when the signal is on based on Assumption 1. With this assumption, EWI-2 is now less accurate than EWI-1. As a robustness check we repeat the optimization exercise and find similar results for the scenario where the signal only stays on for 1 year as opposed to 2 years.

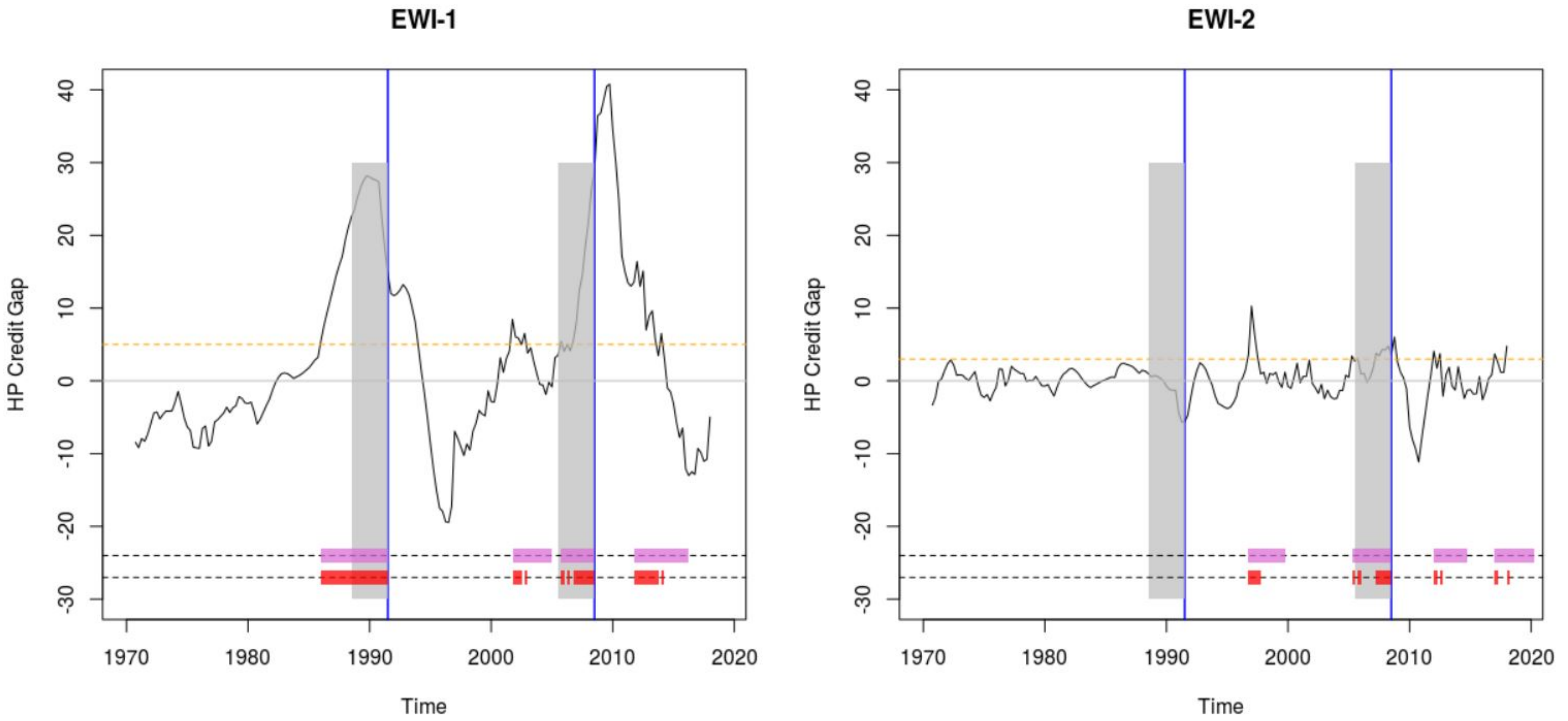

Figure 9: Sweden credit gap from Example 1. Left panel calculated with $\lambda = 400000$ and threshold of 5. Right panel calculated with $\lambda = 100$ and threshold of 3