Macroeconomic Dynamics, **25**, 2021, 100–123. Printed in the United States of America.
doi:10.1017/S1365100518000603

OPTIMIZING POLICYMAKERS' LOSS FUNCTIONS IN CRISIS PREDICTION: BEFORE, WITHIN OR AFTER?

Peter Sarlin
RiskLab at Arcada and Hanken School of Economics, and Silo.AI

Gregor von Schweinitz
Halle Institute for Economic Research (IWH) and University of Leipzig

Recurring financial instabilities have led policymakers to rely on early-warning models to signal financial vulnerabilities. These models rely on *ex-post* optimization of signaling thresholds on crisis probabilities accounting for preferences between forecast errors, but come with the crucial drawback of unstable thresholds in recursive estimations. We propose two alternatives for threshold setting with similar or better out-of-sample performance: (i) including preferences in the estimation itself and (ii) setting thresholds *ex-ante* according to preferences only. Given probabilistic model output, it is intuitive that a decision rule is independent of the data or model specification, as thresholds on probabilities represent a willingness to issue a false alarm vis-à-vis missing a crisis. We provide real-world and simulation evidence that this simplification results in stable thresholds, while keeping or improving on out-of-sample performance. Our solution is not restricted to binary-choice models, but directly transferable to the signaling approach and all probabilistic early-warning models.

Keywords: Early-Warning Models, Loss Functions, Threshold Setting, Predictive Performance

1. INTRODUCTION

The recent financial crisis has stimulated research on early-warning models. These models signal macro-financial risks and guide macroprudential policy to mitigate real implications of an impending crisis. Early-warning models mostly

Research of Gregor von Schweinitz was partly funded by the European Regional Development Fund through the programme "Investing in your Future" and by the IWH Speed Project 2014/02. Parts of this work have been completed at the Financial Stability Surveillance Division of the ECB DG Macroprudential Policy and Financial Stability. The authors are grateful for the suggestions of two anonymous referees, useful comments from Bernd Amann, Carsten Detken, Makram El-Shagi, Jan-Hannes Lang, Tuomas Peltonen, and Peter Welz, and discussion at the following seminars and conferences: Third HenU-INFER Workshop on Applied Macroeconomics, IWH Economic Research Seminar, Goethe University Brown Bag Seminar, ECB Financial Stability Seminar, Deutsche Bundesbank Early-Warning Modeling Seminar and the 2015 CEUS Workshop. An online appendix to this paper as well as replication material are supplied at https://risklab.fi/publications/thresholdoptimization. Address correspondence to: Gregor von Schweinitz, Department of Macroeconomics, Halle Institute for Economic Research, Kleine Märkerstr. 8, 06108 Halle (Saale), Germany. e-mail: gsz@iwh-halle.de. Phone: +49 345 7753 744.

involve two parts: (i) an estimated measure of crisis vulnerability and (ii) a threshold transforming these measures into binary signals for policy recommendation. The currently predominant approach separates the two parts and optimizes thresholds *ex-post*. This ignores estimation uncertainty, provides time-varying thresholds, and results in suboptimal policy guidance out-of-sample. We propose two alternatives that avoid these problems: within-estimation and *ex-ante* threshold setting.

The first part of an early-warning model is the estimation method. The two dominating approaches for this are binary-choice methods and the signaling approach. Binary-choice analysis (like probit or logit models) was already applied by Frankel and Rose (1996) and Berg and Pattillo (1999) to exchange-rate pressure, and has more recently been the predominant approach [Lo Duca and Peltonen (2013) and Betz et al. (2014)]. The signaling approach is simpler in that it only monitors univariate indicators vis-à-vis thresholds. It originally descends from Kaminsky and Reinhart (1999), but has also been common in past years [Alessi and Detken (2011) and Knedlik and von Schweinitz (2012)]. The second part of an early-warning model concerns the setting of thresholds that transform probabilities (univariate indicators for the signaling approach) into signals. This transformation is based upon loss functions tailored to the preferences of a decision-maker.[1] These loss functions rely on the notion of a policymaker facing costs for missing crises (type-1 errors) and issuing false alarms (type-2 errors). Different versions of a loss function have, for example, been introduced by Demirgüç-Kunt and Detragiache (2000), Alessi and Detken (2011), and Sarlin (2013).

Common practice implies an estimation of a binary-choice model and an *ex-post* optimization of the threshold within a loss function given predefined preferences for type-1 and type-2 errors. This approach implies several economic and econometric drawbacks. Viewing the problem from an econometric perspective, it ignores uncertainty about the true data-generating process (DGP). Thus, optimized thresholds falsely react to and vary with probability estimates. They find signal in noise by exhibiting an in-sample overfit and (more often than not) an out-of-sample underfit. Accordingly, as optimized thresholds react to probability estimates, new observations and increased knowledge about the true DGP lead to unwarranted time variation in thresholds. For policy purposes, this is problematic as the rationale for policy implementation needs to descend from changes in vulnerability rather than changing thresholds.

This paper postulates that early-warning models should abstain from threshold optimization. Instead, we present two alternatives to the currently predominant approach for threshold setting: within-estimation and *ex-ante* threshold setting. The first alternative relies on a weighted binary-choice model, where the weights are given by the above-mentioned preferences. In the case of the loss function of Sarlin (2013) (our preferred loss function, see the next section for a more detailed description), weights are given by preferences. If a large preference is given to correctly signaling crises, these observations will receive a large weight in the

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

TABLE 1. Optimization approaches at a glance

	Current approach	Alternative 1	Alternative 2
Estimation method	Binary choice	Weighted binary choice	Binary choice
Loss function		Sarlin (2013)	
Preference parameter	μ	μ	μ
Observation weights		$\mu/1-\mu$	
Threshold	λ^* minimizes loss function in sample	0.5	$\lambda^{\text{inf}} = 1-\mu$
Loss function		Alessi and Detken (2011)	
Preference parameter	θ	θ	θ
Observation weights		$\frac{\theta}{P_1}$ / $\frac{1-\theta}{P_2}$	
Threshold	λ^* minimizes loss function in sample	0.5	$\lambda^{\text{inf}} = \frac{(1-\theta)P_1}{(1-\theta)P_1+\theta P_2}$

Note: Preference parameters $\mu, \theta \in [0, 1]$ relate to the weights of different errors. P_1 denotes the share of precrisis periods, while P_2 represents the probability of tranquil periods ($P_1 + P_2 = 1$).

estimation. The estimation shifts fitted values in a way that an invariant threshold of 50% can now be employed to transform probabilistic into binary forecasts. The second alternative is based on the usual binary-choice model, but sets probability thresholds *ex-ante* according to preferences. It can be proven that this is the long-run optimal threshold independently of the DGP. Given an unbiased probabilistic model, it is intuitive that a decision rule is independent of the exact data or model specification. By way of a simple example, the decision of signaling for probabilities above 20% indicates a willingness to issue a false alarm (80%) vis-à-vis missing a crisis (20%). In terms of preferences, this means that the *ex-ante* optimized threshold for a preference parameter μ (equal to 0.8 in the above example) is set at a value of $1 - \mu$ for the loss function of Sarlin (2013). Table 1 reports the three alternative approaches to selection for two different loss functions at a glance.

The alternative approaches have three benefits. First, even in recursive estimations they assure a stable threshold, because thresholds only depend on preferences which are exogenous to the model. With preferences and thresholds being exogenous, time-varying policy guidance only depends on time-varying macro-financial vulnerability. Second, we show that within-estimation and *ex-ante* threshold setting on average improves out-of-sample predictive power. We can show that threshold optimization does not account for estimation uncertainty. Thus, it introduces a positive bias of in-sample performance, and has on average a negative effect on out-of-sample performance.[2] Third, *ex-ante* threshold selection simplifies the process, as the second optimization step of the traditional approach is left out.

These benefits, and the underlying critique, can easily be extended to more general settings. First, the critique is not restricted to the specific loss functions

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

analyzed in this paper, but applies to any loss or usefulness function optimization that ignores estimation uncertainty. In general, using different loss functions does not alleviate the described problem. Second, the critique extends to the signaling approach that consists solely of the optimization step. However, the equivalence of the signaling approach to a univariate probit model implies that our proposed solutions equally apply. Third, the proposed alternatives extend to methods beyond binary-choice models: accounting for preferences within estimation is directly transferable to all methods used in the early-warning literature, while *ex-ante* threshold setting is valid for any model yielding unbiased crisis probabilities.

We provide two-fold evidence for our claims concerning threshold stability and model performance. First, we make use of two real-world cases to illustrate both threshold stability and in-sample versus out-of-sample performance for the three approaches. Specifically, we replicate the early-warning model for currency crises in Berg and Pattillo (1999) and the early-warning model for systemic financial crises in Lo Duca and Peltonen (2013). Second, we run simulations with different DGP to illustrate the superiority of weighted maximum-likelihood estimation and *ex-ante* thresholds vis-à-vis *ex-post* optimization of thresholds on data with known patterns. All exercises are performed for the loss functions of Alessi and Detken (2011) and Sarlin (2013).

The paper is structured as follows. The next section presents the methods, followed by a discussion of our experiments on real-world data in the third section and our exercises on simulated data in the fourth section. The last section concludes.

2. ESTIMATING AND EVALUATING EARLY-WARNING MODELS

This section presents the three methods analyzed in this paper, namely the currently used approach to derive an early-warning model as well as two alternatives. All three methods consist of two elements: the estimation of a binary-choice model and the setting of a probability threshold for the classification into signals. These two elements will be described together with the current approach in the first subsection, while the following subsections introduce the two alternatives.

In all cases, the binary event to be explained is a precrisis variable $C(h)$. The precrisis variable $C(h)$ is set to one in the h periods before a crisis, and zero in all other, so-called tranquil, periods.[3] That is, $C_j(h) = 1$ signifies that a crisis is to happen in any of the h periods after observation $j \in \{1, 2, \ldots, N\}$, while $C_j(h) = 0$ indicates that all h subsequent periods are classified as tranquil.

2.1. Binary-Choice Models and *Ex-post* Thresholds

Estimation. Binary-choice models (logit or probit models) have been the most important methods in the early-warning literature [Frankel and Rose (1996), Kumar et al. (2003), Fuertes and Kalotychou (2007), and Davis and Karim (2008),

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

TABLE 2. A contingency matrix

		Actual class C_j	
		Precrisis period	Tranquil period
Predicted class S_j	Signal	Correct call *True positive (TP)* Rel. cost: 0	False alarm *False positive (FP)* Rel. cost: $1-\mu$
	No signal	Missed crisis *False negative (FN)* Rel. cost: μ	Correct silence *True negative (TN)* Rel. cost: 0

see among many others]. In a standard binary-choice model, it is assumed that the event $C_j(h)$ is driven by a latent variable

$$y_j^* = X_j\beta + \varepsilon$$
$$C_j(h) = \begin{cases} 1, & \text{if } y_j^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Under the assumption $\varepsilon \sim \mathcal{N}(0,1)$, this leads to the probit log-likelihood function

$$LL(C(h)|\beta, X) = \sum_{j=1}^{N} 1_{C_j(h)=1} \ln(\Phi(X_j\beta)) + 1_{C_j(h)=0} \ln(1-\Phi(X_j\beta)),$$

which is maximized with respect to β. If we assume a logistic distribution of errors, the likelihood function changes only with respect to a distribution function F, which is logistic instead of normal.

Threshold setting. The model returns probability forecasts $p_j = \mathbb{P}(y_j^* > 0)$ for the occurrence of a crisis. While the level of crisis probabilities is of interest, a policymaker is mainly concerned with whether the probability ought to trigger (or signal) preventive policy measures. Thus, estimated event probabilities p_j are turned into binary point predictions S_j by assigning the value of one if p_j exceeds a threshold $\lambda \in [0,1]$ and zero otherwise. The resulting predictions S_j and the true precrisis variable $C_j(h)$ can be presented in a 2×2 contingency matrix (see Table 2). Based upon the threshold λ, the contingency matrix allows us to compute a number of common summarizing measures, such as unconditional probabilities P_1 and P_2, and type-1 and 2 error rates T_1 and T_2.[4] It should be noted that all entries of the contingency matrix, and hence all measures based upon its entries, depend on the threshold λ.

An intuitive threshold would be 50%. However, as crises are (luckily) scarce and (sadly) often very costly, one would usually choose a threshold below 50% in order to balance the frequency and costs of the two types of errors.[5] The entries

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

of the contingency matrix, as well as error rates, can be used to define a large palette of loss functions to optimize the threshold λ. Three components define these measures: unconditional probabilities, type-1 and -2 error rates, and error preferences. We mainly use the the loss and usefulness measures defined in Sarlin (2013): To set policymakers' preferences of individual errors in relative terms (including economic and political costs, among others), falsely predicted events (FP) get a weight of $\mu \in [0, 1]$, missed events (FN) a weight of $1 - \mu$. That is, we assume that the cost of falsely predicting a crisis is μ, the cost of missing a crisis is $1 - \mu$, while correct predictions incur zero costs to the policymaker (see also Table 2). Accordingly, the preference parameter μ is a free parameter that should in practice be set *ex-ante* by the policymaker. In practice, it is often chosen around the share of tranquil periods P_2 (around 80% in most samples).

From the three components (classification threshold λ and error rates, preference parameter μ, share of precrisis and tranquil periods), three equivalent measures are derived. The first is a *loss function* $L(\mu)$ of preference-weighted errors, the second is *absolute usefulness* $U_a(\mu)$ that relates the loss of the model to disregarding the model altogether, and the third is a scaled *relative usefulness* $U_r(\mu)$ that relates absolute usefulness to the maximal achievable usefulness:

$$
\begin{aligned}
L(\mu) &= \mu P_1 T_1 + (1-\mu) P_2 T_2 = \mu FN/N + (1-\mu) FP/N. \\
U_a(\mu) &= \min(\mu P_1, (1-\mu) P_2) - L(\mu). \\
U_r(\mu) &= \frac{U_a(\mu)}{\min(\mu P_1, (1-\mu) P_2)}.
\end{aligned}
$$

The relation between these three measures is strictly monotonic in thresholds: suppose the threshold λ is decreased. There will be more false positives and less false negatives. Suppose further that the change in classification errors is such that the loss function increases. Then, absolute and relative usefulness decrease, because the first summation term of the absolute loss function and the denominator of the absolute loss function ($\min(\mu P_1, (1-\mu) P_2)$) is strictly positive and independent of thresholds. When interpreting models, we can hence focus mainly on U_r. The current approach in early-warning modeling chooses the threshold that optimizes the three measures (loss function, absolute and relative usefulness) simultaneously based on the results of the probabilistic model. We call this the optimized threshold λ^*.

While the optimized threshold λ^* produces the best in-sample fit given preferences μ, it has two undesirable properties. First, it is not an analytical function of the preferences, but also depends on the realization of the DGP. Thus, if new data are added to the sample, the optimized threshold will most likely change. This is extremely relevant in practice, where the early-warning model is estimated recursively over time, re-optimizing the threshold with every new estimation. Second, good in-sample performance is not necessarily a sign of good out-of-sample performance. In principle, the best out-of-sample performance would be achieved by the threshold that maximizes usefulness out-of-sample. Thus, the optimized threshold λ^* may prove to be suboptimal out-of-sample.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

Alternative specifications: The loss function of Alessi and Detken (2011) is conceptually close, but preferences θ apply to type-1 and type-2 error rates instead of shares of all observations: $L^{AD}(\theta) = \theta T_1 + (1-\theta)T_2$.[6] In practice, values of θ around 0.5 have received most attention. That is, the loss function of Alessi and Detken (2011) measures error relative to the class in which they can occur (false positives can only occur in tranquil periods). The loss function of Sarlin (2013)—by taking errors as share of all observations—rather takes an observation-specific mindset. However, if we set $\theta = \frac{\mu P_1}{\mu P_1 + (1-\mu)P_2}$, then the loss function of Alessi and Detken (2011) becomes

$$L^{AD}(\theta) = L^{AD}\left(\frac{\mu P_1}{\mu P_1 + (1-\mu)P_2}\right) = \frac{\mu P_1 T_1 + (1-\mu)P_2 T_2}{\mu P_1 + (1-\mu)P_2}$$
$$= \frac{1}{\mu P_1 + (1-\mu)P_2} L(\mu).$$

That is, the two loss functions are equal (up to a factor). The correspondence between the preference parameters μ and θ has several consequences. First, it has to be noted that the factor $\frac{1}{\mu P_1 + (1-\mu)P_2}$ does not depend on model output and thus also not on the threshold. Thus, if θ and μ are set correspondingly, they result in an identical threshold λ (independent of the approach taken to set λ). That is, all results reported in later sections equally apply to both preference settings. Second, to assure that costs of individual (i.e., observation-specific) errors are reflected by preferences, θ should vary with the probability of the two classes P_1 and P_2. In recursive estimations, θ should thus be time-varying.

An alternative to binary-choice models in the early-warning literature is the signaling approach [Kaminsky and Reinhart (1999)]. It derives predictions from applying a threshold directly on indicator values, and proceeds with calculating the contingency matrix and a usefulness measure as described above. The large appeal it has for policymakers is due to the direct interpretability of the results and the low data requirements. It is straightforward to show that the signaling approach can be directly mapped to a univariate binary-choice model. In a univariate binary-choice model (with a positive parameter β), higher indicator values are associated with higher probabilities. Therefore, it makes conceptually no difference if a threshold is searched and set on indicator values or probabilities from the associated univariate binary-choice model. Thus, all results presented in this paper extend to the signaling approach as well.

2.2. Alternative 1: Thresholds Within Binary-Choice Models

Instead of using preferences μ to optimize thresholds, one could also include preferences as class weights in the log-likelihood function of the binary-choice model [King and Zeng (2001)]. Thus, precrisis observations in the estimation sample will receive a higher weight in the likelihood if the policymaker aims at avoiding false negatives. The log-likelihood function of the weighted probit model is the following:

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

$$LL(C(h)|\beta, X, w) = \sum_{j=1}^{N} 1_{C_j(h)=1} w_1 \ln(\Phi(X_j\beta)) + 1_{C_j(h)=0} w_2 \ln(1 - \Phi(X_j\beta)).$$

For the usefulness function of Sarlin (2013), we set $w_1 = \mu$ and $w_2 = 1 - \mu$.[7] In the case of Alessi and Detken (2011), we use weights $w_1 = \theta/P1$ and $w_2 = (1 - \theta)/P_2$.

Class-specific weights have previously been used for other purposes in binary-choice models. Manski and Lerman (1977) and Prentice and Pyke (1979) use them to adjust for non-representativeness of an estimation sample in cases where an average effect for the whole population is of interest. In other disciplines, (penalized) weights are one possibility to avoid an estimation bias in severely unbalanced samples with an absolute low number of events [Oommen et al. (2011) and Maalouf and Siddiqi (2014)]. All of these strategies share the same conceptual goal with our proposal. The imbalance introduced in our sample is due to the differences in preferences, that is, different weights of type-1 and type-2 errors in the loss function, and is thus independent of class frequencies. Setting weights according to preferences accounts for the imbalance of errors in the loss function.

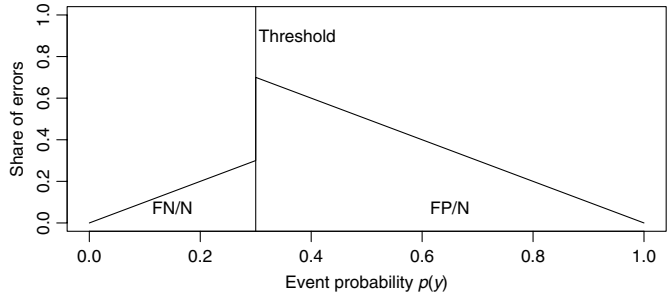
This function can be maximized just as easily as the standard binary-choice model. However, the resulting fitted values should be interpreted as preference-adjusted probabilities. The appealing feature of the weighted binary-choice model is that optimizing a probability threshold *ex-post* is not necessary anymore. Instead, the intuitive threshold of $\lambda^w = 50\%$ already accounts for all policy preferences captured in μ (or θ). This provides a means to replace *ex-post* threshold optimization in both multivariate binary-choice and univariate signaling exercises.

An advantage of this approach is the possible extension to full observation-specific weights. In a cross-country study, one could argue that the potential loss of an error depends not only on the type of error, but also on the (time-varying) size of the affected economy [see Sarlin (2013)]. A second advantage is that this extension can be applied to all methods that employ maximum-likelihood estimation. Yet, weighted binary-choice models come at the disadvantage that different preferences have a direct impact on first-stage estimation results. Thus, when the early-warning model is used with a set of different preferences, the outcome does not only differ in the contingency matrix, but also in different probability and parameter estimates. Moreover, in the case of the loss function of Alessi and Detken (2011) the dependence of class weights on class probabilities P_1 and P_2 may prove to be problematic as weights will in general not be constant in a real-time recursive estimation.

2.3. Alternative 2: *Ex-ante* Thresholds in Binary-Choice Models

Our final approach proposes setting the threshold before estimating the model. The choice of the long-run optimal threshold is based on an argument already put forward by classical decision theory in the vein of the seminal contributions by

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press



Note: The total share of errors (FN/N and FP/N) is the area under the triangle bounded by the threshold λ.

FIGURE 1. Type-1 and type-2 error shares at different event probabilities.

Wald (1950). First, we note that the selection of a threshold is a decision rule. If the (estimated) probability is above the threshold, a signal is given, guiding policy towards action. For probabilities below the threshold, no signal is given. Savage (1951) shows that the optimal decision rule only depends on the costs of different outcomes in the contingency matrix. Thus, a threshold λ can be derived independently of the DGP. Instead, λ should be set at a probability of vulnerability such that a policymaker is in expectation indifferent between a signal and no signal. A classic example is the decision whether or not to carry an umbrella: carrying an umbrella incurs a cost, as does standing in the rain unprotected. Thus, a person would always decide to take an umbrella with her if the cost of carrying one are lower than the expected disutility of being caught in the rain.

We call the threshold given by this optimal decision rule the long-run optimal threshold λ^{∞}. As correct signals have no costs, policymakers should choose a probability threshold which equalizes total costs from false negatives and false positives. The online appendix provides a mathematical derivation of λ^{∞} for the usefulness functions of Alessi and Detken (2011) and Sarlin (2013). It is shown that policymakers are indifferent between a signal and no signal at a threshold of

$$\lambda^{\infty} = \begin{cases} 1-\mu, \text{ for the loss function of Sarlin (2013)} \\ \frac{(1-\theta)P_1}{(1-\theta)P_1+\theta P_2}, \text{ for the loss function of Alessi and Detken (2011).} \end{cases} \quad (\mathbf{1})$$

In general, higher costs of missed events (i.e., a higher μ or higher θ) will lower the long-run optimal threshold, increasing the frequency of false alarms and reducing the frequency of missed events.

The intuition for setting $\lambda^{\infty} = 1-\mu$ in the case of Sarlin (2013) is the following: For every possible threshold λ, the share of false negatives and false positives is just the integral over the respective areas in Figure 1. Let's assume for the sake of the argument, that observations are equally distributed. Then the share of false negatives would be $\int_0^{\lambda} pdp = \lambda^2/2$, and the share of false positives would be $\int_{\lambda}^{1} (1-p)dp = (1-\lambda)^2/2$. Minimizing the loss function over λ now returns $\lambda^{\infty} = 1-\mu$.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

The long-run optimal thresholds λ^{∞} for the loss function of Alessi and Detken (2011) depends not only on policymakers preferences, but also on the frequency of classes P_1 and P_2. The reason is again that the loss function depends on error rates. In practice, class frequencies have to be estimated. Thus, long-run optimal thresholds in recursive estimations will vary with these estimates.

3. REAL-WORLD EVIDENCE OF THRESHOLD SETTING

This section illustrates early-warning modeling and threshold setting with two real-world examples. With these exercises, we in particular focus on illustrating challenges with threshold stability when modeling over time. We test the three different approaches for deriving early-warning models and thresholds: (i) binary-choice models with optimized thresholds, (ii) weighted binary-choice models, and (iii) binary-choice models with pre-set thresholds. We show that in addition to unstable thresholds, out-of-sample utility with optimized thresholds is on average lower than in our two alternative approaches.

3.1. Two Datasets

We replicate the (logit) early-warning model for systemic financial crises by Lo Duca and Peltonen (2013) and the (probit) early-warning model for currency crises by Berg and Pattillo (1999).

The first model is the logit model of systemic financial crises of Lo Duca and Peltonen (2013) (referred to as LDP). The dataset includes quarterly data for 28 countries, 18 emerging market and 10 advanced economies, for the period 1990Q1–2010Q4 (a total of 1729 observations). The crisis definition uses a Financial Stress Index (FSI) with five components: the spread of the 3-month interbank rate over the 3-month government bill rate, quarterly equity returns, equity index volatility, exchange-rate volatility, and volatility of the yield on the 3-month government bill. Following LDP, a crisis is defined to occur if the FSI of an economy exceeds its country-specific 90th percentile. That threshold on the FSI defines 10% of the quarters to be systemic events. It is derived such that the events led, on average, to negative consequences for the real economy. To enable policy actions for avoiding a further build-up of vulnerabilities, the focus is on identifying precrisis periods with a forecast horizon of six quarters. This goal is achieved by employing 14 macro-financial indicators that proxy for a large variety of sources of vulnerability, such as asset price developments, asset valuations, credit developments and leverage, as well as traditional macroeconomic measures, such as GDP growth and current account imbalances. The variables are used both on a domestic and a global level, where the latter is an average of data for the Euro area, Japan, UK, and USA. The dataset is divided into two partitions: in-sample data (1990Q4–2005Q1) and out-of-sample data (2005Q2–2009Q2, out of which LDP use only data until 2007Q2 for analysis). Figure 2 shows the share of precrisis observations at every point in time. It should be

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

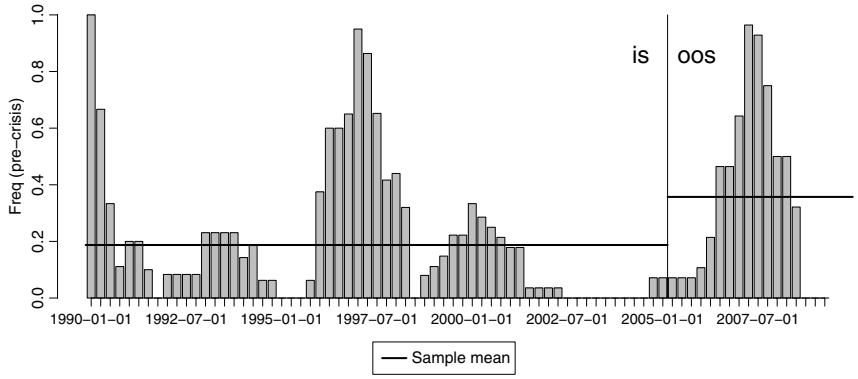


FIGURE 2. Frequency of precrisis periods, full sample, LDP model.

noted that the out-of-sample data contain the run-up to the great financial crisis, increasing the unconditional probability of being in an precrisis window from 19% in-sample to 36% out-of-sample, which we also indicate in the figure.

The second model is the probit model for currency crises by Berg and Pattillo (1999) (hereafter referred to as BP). The dataset consists of five monthly indicators for 23 emerging market economies from 1986:1 to 1996:12 with a total of 2916 country-month observations: foreign reserve loss, export loss, real exchange-rate overvaluation relative to trend, current account deficit relative to GDP, and short-term debt to reserves. To control for cross-country differences, each indicator is transformed into its country-specific percentile distribution. In order to date crises, BP use an exchange market pressure index. A crisis occurs if the weighted average of monthly currency depreciation and monthly declines in reserves exceeds its mean by more than three standard deviations. BP define an observation to be in a vulnerable state, or precrisis period, if it experienced a crisis within the following 24 months. To replicate the setup in BP, the data are divided into an estimation sample for in-sample fitting from 1986:1 to 1995:4, and a test dataset for out-of-sample analysis from 1995:5 to 1996:12 (around 15% of the sample). Figure 3 shows again the share of precrisis observations over time, together with the in-sample and out-of-sample mean. Despite the short period of the test sample, nearly 25% of all events happen in that window due to the Asian crisis.

One obvious difference between the two models is that currency crises and the preceding early-warning windows (BP) are much more equally distributed over time than systemic financial crises. The strong clustering of financial crises, in turn, could lead to imprecise estimates of the true DGP. In the LDP case, the estimated unconditional probability of being in an early-warning window will fluctuate around the true probability, with the fluctuations being large and persistent. Thus, new crisis observations in a recursive analysis may affect in-sample probability estimates and potentially also thresholds. In principle, this could warrant a certain variation of optimized thresholds λ^* around the long-run

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

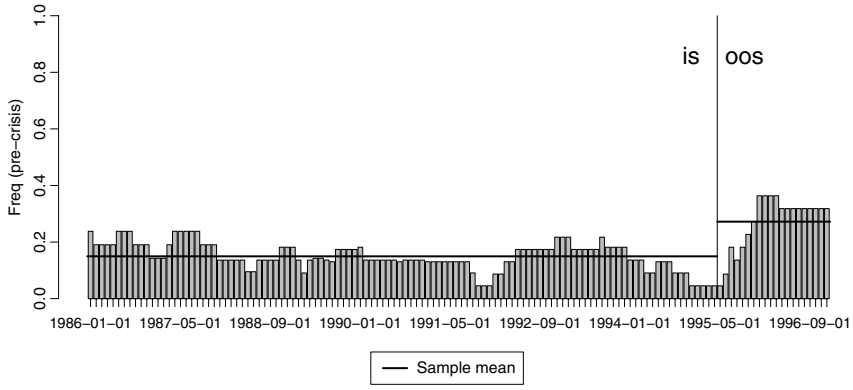


FIGURE 3. Frequency of precrisis periods, full sample, BP model.

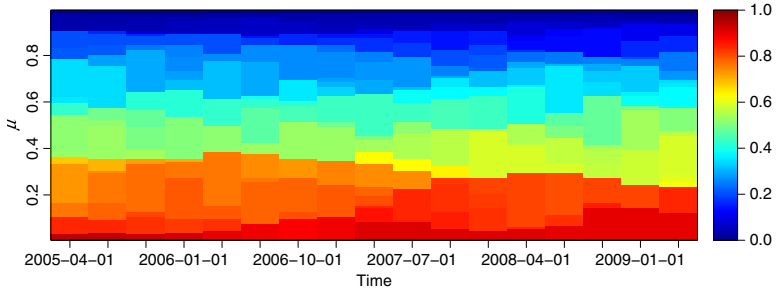
threshold λ^{∞}.[8] However, this variation would still be problematic in real-time analysis, as full real-time knowledge of the "status" of current observations (if they are in an early-warning window before a crisis or not) becomes only available in the future, when the full early-warning horizon has passed.

3.2. Real-Time Thresholding

The main proposition of this paper is that *ex-post* threshold optimization leads to unwarranted variation in thresholds, and that this is problematic for policy. New observations increase knowledge on the true DGP and should thus affect estimates of crisis probabilities. However, to the extent that these estimates are unbiased, new observations should not have an effect on thresholds. Put differently, the rational of recommendations for policy action (i.e., if estimated probabilities are above or below the threshold) should descend from changes in crisis vulnerability rather than changing thresholds. Especially for policymakers, it should be problematic to take different actions based on identical probability estimates, only because of small variations in thresholds, everything else equal. Indeed, we would argue that only policy preferences should affect thresholds [which is our reason to prefer the usefulness function of Sarlin (2013)]. However, in reality we observe sometimes strong variations in optimized thresholds. Moreover, we find that these time-varying thresholds are only optimal for in-sample data, but generate on average suboptimal signals out-of-sample.

The first line of evidence that we put forward is based upon recursive real-time estimations. With the same division of data as in the two original papers, we explore the characteristics and performance of the three approaches when applying them recursively over the out-of-sample part of the data. The recursive analysis implies that we use information available at each period t to derive model output and set optimal thresholds for the same period in question.[9] This mimics a real-time setting when applying early-warning models.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press



Note: The scale refers to λ^* values for each μ and quarter. The models are estimated in a recursive manner by using only information available up to each quarter between 2005Q2 and 2009Q2. Even though LDP only uses data up to 2007Q2, we extend the analysis to the longest available time-series.

FIGURE 4. Variation of λ^* in recursive analysis with the LDP model, continuous preferences.

The variability of thresholds in *ex-post* optimization (λ^*) is a major source of uncertainty (and potentially confusion). We illustrate this by showing threshold variation for the LDP model with *ex-post* optimization, where recursive tests run from 2005Q2 to 2009Q2. Figure 4 shows a heatmap of thresholds λ^* for every quarter in that time and for different preferences μ. For a given μ value (horizontal row), a model with stable thresholds would have a constant color over time. We can observe that this is not the case. For instance, for $\mu = 0.8$ the thresholds seem to vary between 13% and 28%. This points to significant uncertainty that would have serious implications in policy use. A similar result can be seen in the corresponding Figure A.1 in Appendix for the BP model, where recursive tests run from 1995:5 to 1996:12.[10]

As discussed in the description of datasets, fluctuations could in principle be warranted by the clustering of crises. We report the development of optimized thresholds λ^* in the recursive estimation for four different preference parameters $\mu = \{0.2, 0.5, 0.8, 0.95\}$ together with the frequency of crises in Figure 5.[11] That is, we test over different potential preferences that a policymaker may have. High values of $\mu = \{0.8, 0.95\}$ give a strong preference to avoiding crises, which accounts for the fact that missing a crisis may be very costly. $\mu = 0.5$ gives equal weights to both errors and is a setting, where the weighted models boil down to standard binary-choice estimation (without threshold optimization). $\mu = 0.2$ gives strong preference to avoiding false alarms, which accounts for high costs related to external announcements and reputation losses. Overall, $\mu = 0.8$ is probably the most realistic choice of preference parameter. Crucially, there seems to be no systematic link between the occurrence of crises and threshold variation in the dataset. That is, thresholds, independent of the preferences, do not vary systematically with clustered financial crises. Instead, the variation of thresholds seems to be largely driven by noise.

In the case of the BP model, we see in general less threshold fluctuation (see Figure A.2 in the Appendix). This is consistent with the observation that precrisis

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

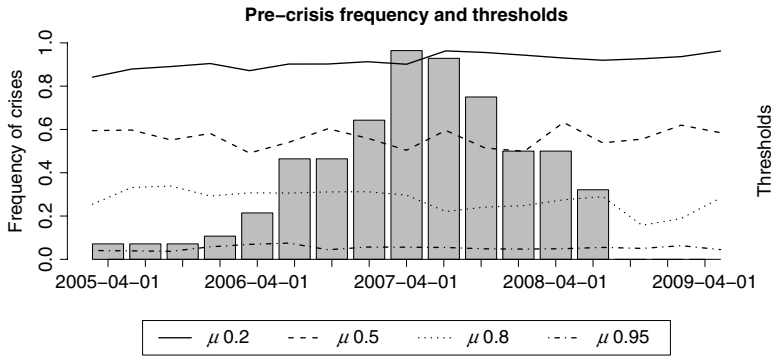


FIGURE 5. LDP model, frequency of precrisis periods, and variation of λ^*, selected preferences μ.

windows are more or less equally distributed across the sample. However, another point of criticism with respect to *ex-post* threshold optimization can be raised: when using the usefulness function of Alessi and Detken (2011), the thresholds for $\theta = 0.5$ and $\theta = 0.7$ nearly overlap. Thus, the model seems to have difficulties finding truly different thresholds for very different preference parameters.

3.3. Performance Comparison

In this subsection, we show that the variation of thresholds can lead to worse out-of-sample performance. Table 3 reports the usefulness for the above-mentioned preference parameters and the three different logit models (LDP), together with the probability that our proposed alternatives outperform the model with optimized thresholds. This probability is derived from 1000 draws of a panel block bootstrap over (recursive) in-sample data with a block-length of 12 quarters.[12]

We can first observe that absolute and relative usefulness is always negative, because the frequency of crises out-of-sample is much higher than in-sample. However, even though usefulness is negative, the models with *ex-ante* or within threshold selection are nearly always on average better than their counterparts with *ex-post* threshold optimization. This holds especially for preference parameters that put a higher weight on the less frequent and more costly type-1 errors (false negatives). For the most realistic preference parameter $\mu = 0.8$, we see for example that our two proposals signal much more often (correctly) than an early-warning model with optimized thresholds would. If we examine the data themselves (where signals should ideally be sent throughout the full early-warning window of one to six quarters before the crisis), we see that the additional correct warnings are mostly located at the beginning of the early-warning windows. This makes sense as signals should "get stronger" once a crisis becomes more and more imminent. However, in one example (Hong Kong during the great financial crisis), optimized thresholds would have only picked up on the upcoming

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

TABLE 3. Performance for LDP, recursive oos estimation 2005Q2–2009Q2

Method	TP	FP	FN	TN	L	U_a	U_r	Probability
				$\mu = 0.2$				
Logit opt threshold	7	21	157	166	0.137	−0.044	−0.470	
Weighted logit	8	21	156	166	0.137	−0.043	−0.463	0.516
Logit set threshold	6	18	158	169	0.131	−0.038	−0.402	0.819
				$\mu = 0.5$				
Logit opt threshold	51	71	113	116	0.262	−0.028	−0.122	
Weighted logit	40	66	124	121	0.271	−0.037	−0.159	0.466
Logit set threshold	40	66	124	121	0.271	−0.037	−0.159	0.466
				$\mu = 0.8$				
Logit opt threshold	104	132	60	55	0.212	−0.105	−0.989	
Weighted logit	113	128	51	59	0.189	−0.083	−0.775	0.810
Logit set threshold	117	146	47	41	0.190	−0.084	−0.786	0.706
				$\mu = 0.95$				
Logit opt threshold	155	167	9	20	0.048	−0.022	−0.807	
Weighted logit	152	160	12	27	0.055	−0.029	−1.075	0.642
Logit set threshold	152	167	12	20	0.056	−0.030	−1.112	0.586

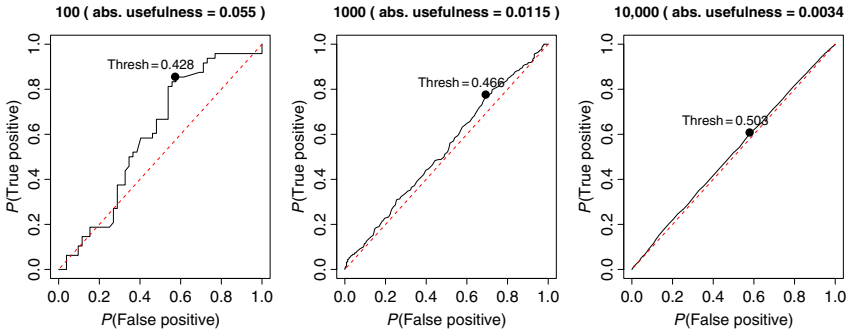
Note: The table reports performances [Sarlin (2013)] of recursive estimations in the LDP model over an out-of-sample period from 2005Q2 to 2009Q2 for the three different methods and four preference choices.

crisis half a year before it actually happened, while our two approaches would have been able to send signals four quarters earlier.

A similar result can be derived (i) for all possible policy preferences μ and (ii) for the usefulness function of Alessi and Detken (2011). For computational reasons, we perform a one-off split of the data instead of a recursive out-of-sample analysis. That is, we derive probabilities and signals for observations between 2005Q2 and 2009Q2 based on estimates on data prior to 2005Q2. As above, we use a panel-block bootstrap to derive the probabilities that our two proposals outperform *ex-post* threshold optimization. Our two alternatives outperform optimized thresholds in more than 50%, independently of the employed usefulness function and nearly independently of preferences.[13] We thus find that our alternatives are better than the current approach in the majority of cases, and that their average out-of-sample performance is higher. Moreover, the weighted logit is slightly better than threshold setting *ex-ante* for $\mu \geq 0.7$ (or $\theta > 0.3$), both in terms of mean usefulness and the probability of outperformance.

The findings above are largely corroborated by the BP model. However, in this case we find larger areas where threshold optimization seems on average to be better than our two proposals. A possible reason for this is that the uncertainty regarding (short-run) optimal thresholds is lower than in the LDP case, as indicated by the lower degree of threshold variability during the out-of-sample period (see Figure A.1 in the Appendix).

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press



Note: Type-2 error probability is given on the *x*-axis, (1—type-1 error probability) on the *y*-axis. The absolute usefulness of the model for the drawn number of observations is given in the title.

FIGURE 6. ROC curve for three simulations with random events (N = 100, 1000, and 10,000) from the probit estimation.

4. COMPARING OPTIMAL THRESHOLDS WITH SIMULATED DATA

The real-world experiments allowed us to observe obvious differences in threshold setting. However, it did not allow us to show variation in thresholds and performance differences on a scale beyond single cases. In this section, we rely on simulated data to further strengthen the evidence in this paper.

To illustrate differences among the three approaches to threshold selection, we provide a large number of experiments on a range of different simulated data. Given that λ^* is selected to optimize the loss function on in-sample data, we expect λ^* to perform best on that part of the data. However, we are mainly interested in the out-of-sample performance of the the three approaches to threshold selection. There, we expect the optimized threshold to fare much worse, possibly to be outperformed by our proposed alternatives.

4.1. Randomness of Thresholds

As an illustrative example, let us take a look at a DGP, where explanatory variables and events are unrelated, and where the event probability is 50% in every period. Figure 6 shows the in-sample receiver operator characteristics (ROC) curves from a probit model for three simulations with different numbers of observations N. An ROC curve shows the trade-off between type-1 errors and type-2 errors at different thresholds. Usefulness optimization basically chooses the combination of type-1 and -2 errors on the black curve that maximizes the weighted distance to the red diagonal [for a discussion of the ROC curve, see Drehmann and Juselius (2014)].

Ideally, the distance (and therefore absolute usefulness) should be zero, because explanatory variables X and events $C(h)$ are unrelated in this specification. However, in practice this is not the case. For small N, β is estimated to produce an optimal fit. This means that the ROC curve will be above the diagonal on average

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

(otherwise, the fit would be worse than for coefficients equal to zero). In fact, the area under the ROC curve (i.e., the AUC) is significantly above 0.5 at the 10% level for the three simulations.

With less observations there is more uncertainty concerning true coefficients, resulting in a larger upward bias of the ROC.[14] If now, in a second step, the weighted distance of the ROC curve is maximized in order to maximize usefulness, this introduces randomness in thresholds and creates an overfit. Essentially, threshold optimization chooses the best possible outcome (in-sample) instead of the most likely possible outcome, which leads to threshold instability, as indicated by the three substantially different threshold values in the plot.

The distance of the ROC curve to the diagonal, and therefore usefulness of the random model, decreases strongly with increasing N. This happens because, as N increases, uncertainty on the true DGP decreases, bringing the ROC curve closer to the diagonal and bringing usefulness closer towards its true level of zero.

4.2. Simulation Setup

Now, let us compare our approaches in a simulation setup where explanatory variables and events are related, that is, where the estimation of event probabilities is actually meaningful. We present the setup of the baseline scenario here. A number of robustness checks are introduced in a later subsection. In our (simple) simulated data, we use three explanatory variables $X = (X_1, X_2, X_3)$, a coefficient vector $\beta = (1, 0, 0)$ and a negative constant of -1. That is, only X_1 contains information on the latent variable y^* and therefore the observable event. The constant is chosen such that the probability of an event is slightly below 25%, in-line with usual event frequencies in early-warning models.

We draw the explanatory variables independently from a standard normal distribution. Every simulation study is performed with 21 logarithmic-spaced number of observations between $N = 100$ and $N = 10,000$. For every N, we draw X, calculate the event probabilities $\Phi(X\beta)$ and draw $C(h)$ from these probabilities (abstracting from index j).[15] Drawing events from a normal distribution means that we simulate data from a probit model. Every simulated dataset is split evenly into an in-sample and an out-of-sample part.

We then apply the three approaches presented in Section 2 to the in-sample part of the data, using both probit and logit estimations. That is, for every dataset and policy preference μ, we construct six different early-warning models. First, a probit with optimized thresholds λ^*. Second, a weighted probit with threshold $\lambda^w = 0.5$. Third, a probit with fixed thresholds $\lambda^\infty = 1 - \mu$. The fourth, fifth and sixth model are equal to the first three, replacing the probit estimation by a logit estimation. Logit estimations are a simple way to test if the results are robust against an admittedly very mild form of misspecification. For all models, we calculate the in-sample and out-of-sample measures of goodness-of-fit defined in the previous section. The above steps are performed for the four different preferences $\mu = \{0.2, 0.5, 0.8, 0.95\}$ already employed in the real-world examples.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

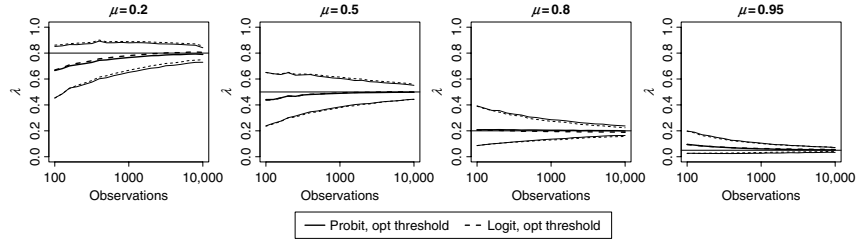


FIGURE 7. Mean λ^* with 90% confidence bands, for different values of μ.

Every simulation is performed R times to get a clear picture of the influence of sampling uncertainty. This allows us to provide a measure for the uncertainty of optimized thresholds λ^*, as well as the size of the in- and out-of-sample bias of usefulness. Furthermore, we can calculate the probability that the current early-warning model (probit/logit with threshold optimization) is outperformed by our alternatives. The probabilities of outperformance are bootstrap estimates. That is, they vary slightly with the number of replications R. To be sure that probabilities of outperformance are truly larger than 50% (and not only by chance), one can either choose a very large number of replications R, or adopt the approach of Davidson and MacKinnon (2000) to select R endogenously. We follow the latter approach.

In the following, we will only present results from the baseline specification. Many other specifications, as described in in the online appendix, yield both qualitatively and quantitatively very similar results.

4.3. Variation and Limit of Optimized Thresholds

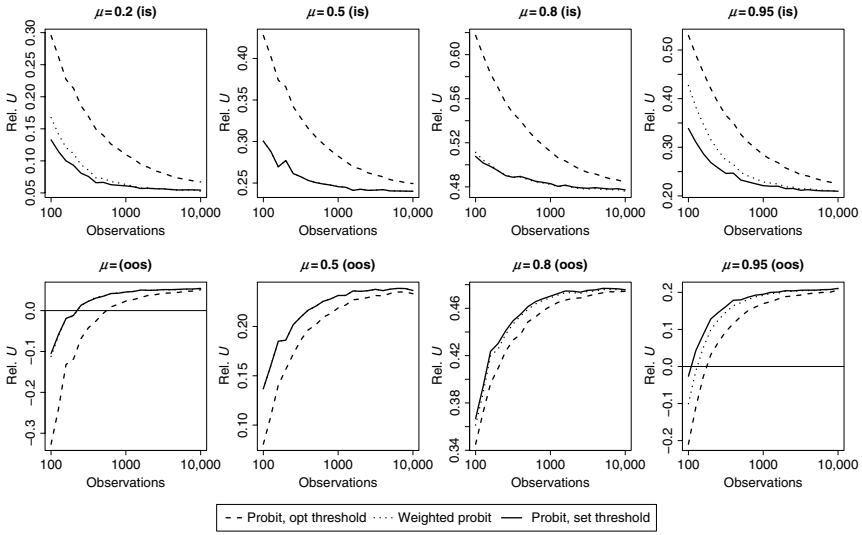
In this subsection, we analyze the behavior of the optimized threshold λ^* in our simulation setup. We are specifically interested in the question if λ^* approaches the long-run optimal threshold λ^∞ as $N \to \infty$. Figure 7 presents the mean λ^* together with confidence bands from R replications for the different policy preferences μ and different number of observations N. We first see that there is basically no difference between probit and logit estimations.

As the true DGP is always identical, all uncertainty on λ^* comes from the estimation uncertainty, which depends mainly on the number of observations. Therefore, the width of the confidence bands of λ^* does not depend on preferences μ and decreases with N. However, even for a large number of observations there remains considerable uncertainty. As expected and in line with the mathematical proof of our second alternative, λ^* approaches $1 - \mu$ as N increases.

4.4. Comparison of Out-of-Sample Performance

This subsection analyzes the out-of-sample performance of the three approaches to threshold setting. We are particularly interested in the question if the in-sample

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press



Note: In-sample (*is*) usefulness is higher than out-of-sample (oos) usefulness for every number of observations *N*. The black line at zero signifies the boundary below which it is optimal not to use the model.

FIGURE 8. Mean relative usefulness of the three probit models.

superiority of the current approach has negative effects on its out-of-sample performance or not.

Under the assumption that data are created by a constant DGP, and that this process can be captured by the estimated model, in-sample and out-of-sample usefulness should both converge to the true long-run usefulness of that process. As in-sample models are fitted to the data, we would expect that in-sample usefulness is higher for a lower number of observations and that it drops towards a boundary value. This view is confirmed by Figure 8 for probit models.[16] These figures show the mean relative usefulness from simulations with different numbers of observations for the three different approaches. In-sample results are presented in the first row of plots, out-of-sample results in the second row, differentiating for different preferences μ. Contrary to in-sample usefulness, the out-of-sample usefulness improves as N goes to infinity. The reason is the slow uncovering of the true DGP, which improves inference from in- to out-of-sample data.

In addition to these general results holding for all estimation methods, we see that the usefulness (in- and out-of-sample) of our proposals is on average closer to their true limiting value than those of the benchmark models. Concerning in-sample usefulness, this seems to be bad at first sight. However, it has to be acknowledged that one of the main reasons for calculating in-sample usefulness is an evaluation of the quality of the early-warning model. If there is an upward bias, it induces an overstated sense of confidence, trust and security. This bias is

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

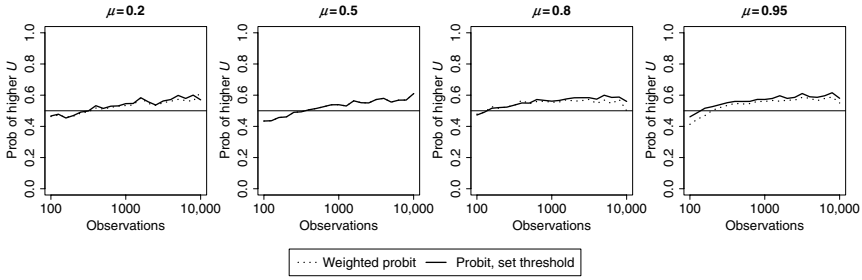


FIGURE 9. Probability of outperformance of alternative approaches out-of-sample (probit estimations).

much lower for our proposals. However, what really matters in the early-warning practice is out-of-sample usefulness. Here, our proposals perform on average better. This holds both for the weighted model and for the *ex-ante* threshold setting.

Even though out-of-sample usefulness of our proposals is on average better than that of threshold optimization, this difference is not statistically significant in most cases. By construction, our proposals produce nearly always worse in-sample usefulness than their threshold peer. Out-of-sample, our proposals outperform the benchmark in slightly more than 50% of the cases (see Figure 9). Why do our alternatives often outperform the benchmark model only in slightly more than 50% of the cases, while still providing (on average) sizable higher out-of-sample relative usefulness? The reason for this is the uncertainty in the DGP that makes threshold optimization prone to variation. As the innovations in- and out-of-sample are uncorrelated, there is a (roughly) 50% chance that the out-of-sample innovations would push the optimized threshold in a similar direction as the in-sample innovations. Therefore, there is a 50% chance that thresholds optimized based on in-sample data perform (slightly) better for out-of-sample data than the fixed thresholds of our two alternatives. However, in the other 50% the performance losses are much higher.

5. CONCLUSION

The traditional approach for deriving early-warning models relies on a separate *ex-post* threshold optimization step. We show in this paper that this *ex-post* optimization of thresholds is prone to suffer from estimation uncertainty, resulting in unstable probability thresholds and potentially reduced out-of-sample usefulness.

We propose two alternative approaches for threshold setting in early-warning models, where preferences for forecast errors are accounted for by setting thresholds not after, but within or even before the estimation of early warning probabilities.

Including preferences as estimation weights (resulting in a threshold $\lambda^w = 0.5$) in the early-warning model outperforms optimized thresholds out-of-sample in

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

the large majority of the cases. Thus, weighted binary-choice models are a valid alternative to the current approach of threshold optimization. Moreover, the idea of weighting classes according to preferences is not restricted to binary-choice or even maximum-likelihood methods. As weighting can be implemented by resampling data, our approach can be extended to any classification method employed in the early-warning literature [Chawla et al. (2004)]. However, weighting comes with two drawbacks: First, fitted values can only be interpreted as weighted probabilities. Second, introducing weights into an estimation requires moving away from standard statistical packages.[17]

Contrary to the two other approaches, the long-run optimal threshold $\lambda^{\infty} = 1 - \mu$ is independent of estimated vulnerabilities and the DGP as a whole. Moreover, λ^{*} will approach λ^{∞} as the true DGP is uncovered over time (see Figure 7). That is, in the case of a correctly specified model, the long-run optimal threshold will alleviate all challenges to optimized thresholds. However, in comparison to the two other approaches, the performance of long-run optimal thresholds depends more on the correct estimation of the true DGP. For example, a DGP with clustered events could easily lead to biased probability estimates in-sample, which affects the performance of long-run optimal thresholds both in- and out-of-sample.

We first compare our two approaches to the current standard of threshold selection *ex-post* by looking at two real-world examples. In both these models, we can document a strong variability of optimized thresholds which is not warranted by the data. For policymakers, variations in thresholds due to uncertainty might be challenging to communicate. How can policies in a country with unchanged macro-financial conditions be implemented only due to a shift in "optimal" λ? Signals should depend on changes in the vulnerability indicators, not on unjustified (random) variation in thresholds. But our two proposals do not only imply stable thresholds. A bootstrap analysis shows us that at least in the case of the model of systemic financial crises of Lo Duca and Peltonen (2013), our approaches on average outperform threshold optimization out-of-sample for nearly all preferences. Both results are confirmed by a range of simulation studies, where we sample explanatory variables and crises from a simple and known DGP.

To subsume, we find that our two alternative proposals outperform their traditional counterpart in three ways. First, we eliminate unjustified (random) variation in thresholds and allow hence all signals to descend purely from variation in probabilities. This supports policy implementation and communication based upon these models. Second, out-of-sample performance can on average be improved by our approaches, while the bias on in-sample usefulness is reduced. Third, at least *ex-ante* threshold setting is simpler than *ex-post* threshold optimization, as it forgoes the second optimization step.

As our results hold not only for the simple binary-choice models tested in this paper, but for every early-warning model using threshold optimization (including the much-used signaling approach), we strongly recommend to include

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

policymakers' preferences as weights in the estimated likelihood or specifying thresholds *ex-ante*, and thus to move away from threshold optimization in general.

NOTES

1. We do not herein summarize measures used for assessing model robustness that do not explicitly provide guidance on optimal thresholds, such as the Receiver Operating Characteristics curve and the area below it.

2. This was also indicated by El-Shagi et al. (2013) and later by Holopainen and Sarlin (2015), which both show and account for the fact that a positive usefulness can be insignificant. We approach the problem of uncertainty and significance from a different angle.

3. In most applications, one would exclude actual crisis periods and possibly even some periods after a crisis from the estimation altogether, as they may not be tranquil, and should therefore not be used for early-warning purposes [Bussière and Fratzscher (2006)].

4. Following the literature, the measures are defined as follows: $P_1 = \mathbb{P}(C_j(h) = 1) = (TP + FN)/N$, $P_2 = 1 - P_1$, $T_1 = \mathbb{P}(P_j = 0|C_j = 1) = FN/(FN + TP)$, and $T_2 = \mathbb{P}(P_j = 1|C_j = 0) = FP/(FP + TN)$.

5. In spirit, this is very similar to the finding of Riccetti et al. (2018), that overly tight regulation puts too much of a burden on credit availability, while overly loose regulation increases the probability of financial crises.

6. There exists a myriad of alternative performance measures with larger differences. Two other measures have been commonly applied in the early-warning literature. The signal-to-noise ratio [Kaminsky and Reinhart (1999)] has been shown to lead to corner solutions, resulting in a high share of missed crisis episodes if crises are rare [Demirgüç-Kunt and Detragiache (2000) and El-Shagi et al. (2013)]. Fuertes and Kalotychou (2007) and Bussière and Fratzscher (2008) use a slightly different loss function. Many additional measures are summarized in Wilks (2011).

7. This is in principle equivalent to the approach of King and Zeng (2001), where weights are normalized to have a sample mean of unity (i.e., $w_1 = \frac{\mu}{\mu P_1 + (1-\mu)P_2}$ and $w_2 = \frac{1-\mu}{\mu P_1 + (1-\mu)P_2}$).

8. If crises are equally costly and if probability estimates do not increase, a higher frequency of crises could imply lower thresholds.

9. Thus, we recalculate the precrisis variable in every recursive step given available information on the crisis variable.

10. The original authors do not perform recursive out-of-sample analysis.

11. Results regarding the performance under the usefulness function of Alessi and Detken (2011) are very similar and reported in the online appendix.

12. We combine the two approaches by El-Shagi et al. (2013) and Holopainen and Sarlin (2015). To allow measuring uncertainty around usefulness (taking countries as given) we use a simple panel block bootstrap that accounts for cross-sectional and autocorrelation of both right and left-hand side variables and pairs events and indicators.

13. For a visual result, we refer to Figure B.3 in the online appendix, which also reports corresponding results for the BP model.

14. El-Shagi et al. (2013) therefore argue that—in order to judge the quality of an early-warning model—it is paramount to obtain a distribution of the usefulness under the null hypothesis of no relation between X and $C(h)$, instead of only a measure of usefulness itself.

15. This procedure introduces one difference to usual early-warning models: there is no continuous chain of events in an early-warning window of predefined length. However, this difference is irrelevant from an econometric perspective.

16. An alternative way to look at this would be the difference of relative usefulness between the benchmark model and our two proposals, see the online appendix. Similar results for the logit models can also be found there.

17. An R-package for weighted binary-choice models can be obtained from the authors.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press

REFERENCES

Alessi, L. and C. Detken (2011) Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity. *European Journal of Political Economy* 27(3), 520–533.

Berg, A. and C. Pattillo (1999) What caused the Asian crises: An early warning system approach. *Economic Notes* 28(3), 285–334.

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

Bussière, M. and M. Fratzscher (2006) Towards a new early warning system of financial crises. *Journal of International Money and Finance* 25(6), 953–973.

Bussière, M. and M. Fratzscher (2008) Low probability, high impact: Policy making and extreme events. *Journal of Policy Modeling* 30(1), 111–121.

Chawla, N., N. Japkowicz and A. Kotcz (2004) Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations* 6(1), 1–6.

Davidson, R. and J. G. MacKinnon (2000) Bootstrap tests: How many bootstraps? *Econometric Reviews* 19(1), 55–68.

Davis, E. P. and D. Karim (2008) Comparing early warning systems for banking crises. *Journal of Financial Stability* 4(2), 89–120.

Demirgüç-Kunt, A. and E. Detragiache (2000) Monitoring banking sector fragility: A multivariate logit approach. *The World Bank Economic Review* 14(2), 287–307.

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

El-Shagi, M., T. Knedlik and G. von Schweinitz (2013) Predicting financial crises: The (statistical) significance of the signals approach. *Journal of International Money and Finance* 35, 76–103.

Frankel, J. A. and A. K. Rose (1996) Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics* 41(3), 351–366.

Fuertes, A.-M. and E. Kalotychou (2007) Optimal design of early warning systems for sovereign debt crises. *International Journal of Forecasting* 23(1), 85–100.

Herndon, T., M. Ash and R. Pollin (2014) Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38(2), 257–279.

Holopainen, M. and P. Sarlin (2015) Toward Robust Early-Warning Models: A Horse Race, Ensembles and Model Uncertainty. Bank of Finland Discussion Paper 06/2015.

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

King, G. and L. Zeng (2001) Logistic regression in rare events data. *Political Analysis* 9(2), 137–163.

Knedlik, T. and G. von Schweinitz (2012) Macroeconomic imbalances as indicators for debt crises in Europe. *JCMS: Journal of Common Market Studies* 50(5), 726–745.

Kumar, M., U. Moorthy and W. Perraudin (2003) Predicting emerging market currency crashes. *Journal of Empirical Finance* 10(4), 427–454.

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

Maalouf, M. and M. Siddiqi (2014) Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems* 59, 142–148.

Manski, C. F. and S. R. Lerman (1977) The estimation of choice probabilities from choice based samples. *Econometrica* 45(8), 1977–1988.

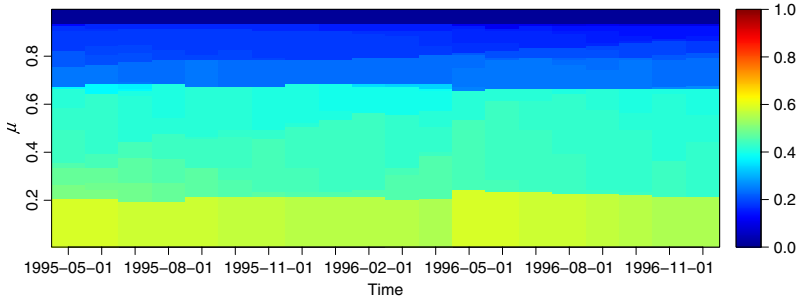
Oommen, T., L. G. Baise and R. M. Vogel (2011) Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences* 43(1), 99–120.

Prentice, R. L. and R. Pyke (1979) Logistic disease incidence models and case-control studies. *Biometrika* 66(3), 403–411.

Riccetti, L., A. Russo and M. Gallegati (2018) Financial regulation and endogenous macroeconomic crises. *Macroeconomic Dynamics* 22(4), 896–930.

Sarlin, P. (2013) On policymakers' loss functions and the evaluation of early warning systems. *Economics Letters* 119(1), 1–7.
Savage, L. J. (1951) The theory of statistical decision. *Journal of the American Statistical Association* 46(253), 55–67.
Wald, A. (1950) *Statistical Decision Functions*. New York: Wiley.
Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd ed. International Geophysics Series, Vol. 100. Academic Press.

APPENDIX: ADDITIONAL TABLES AND FIGURES



Note: The scale refers to λ values for each μ and month. The models are estimated in a recursive manner by using only information available up to each month between 1995:5 and 1996:12.

FIGURE A.1. λ variation in recursive analysis with the BP model.

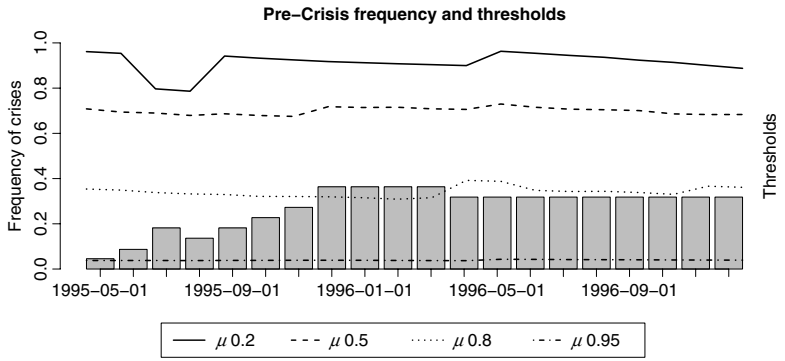


FIGURE A.2. BP model, frequency of precrisis periods, and variation of λ^*, selected preferences μ.

https://doi.org/10.1017/S1365100518000603 Published online by Cambridge University Press