Leibniz-Institut für
Wirtschaftsforschung
Halle

Optimizing Policymakers' Loss Functions in Crisis Prediction: Before, Within or After?

Peter Sarlin
Gregor von Schweinitz

June 2015 No. 6

IWH-DISKUSSIONSPAPIERE
IWH DISCUSSION PAPERS

Authors: *Peter Sarlin*
Department of Economics at Hanken School of Economics and RiskLab Finland at Arcada University of Applied Sciences
E-mail: peter@risklab.fi

Gregor von Schweinitz
Martin Luther University Halle-Wittenberg, Chair of Macroeconomics
Halle Institute for Economic Research (IWH)
Department of Macroeconomics
E-mail: Gregorvon.Schweinitz@iwh-halle.de
Phone: +49 345 7753 744

The responsibility for discussion papers lies solely with the individual authors. The views expressed herein do not necessarily represent those of the IWH. The papers represent preliminary work and are circulated to encourage discussion with the authors. Citation of the discussion papers should account for their provisional character; a revised version may be available directly from the authors.

Comments and suggestions on the methods and results presented are welcome.

IWH Discussion Papers are indexed in RePEc-EconPapers and in ECONIS.

Editor:
Halle Institute for Economic Research (IWH) – Member of the Leibniz Association

Address: Kleine Maerkerstrasse 8, D-06108 Halle (Saale), Germany
Postal Address: P.O. Box 11 03 61, D-06017 Halle (Saale), Germany
Phone: +49 345 7753 60
Fax: +49 345 7753 820
Internet: www.iwh-halle.de

ISSN 1860-5303 (Print)
ISSN 2194-2188 (Online)

Optimizing Policymakers' Loss Functions in Crisis Prediction: Before, Within or After?*

Abstract

Early-warning models most commonly optimize signaling thresholds on crisis probabilities. The ex-post threshold optimization is based upon a loss function accounting for preferences between forecast errors, but comes with two crucial drawbacks: unstable thresholds in recursive estimations and an in-sample overfit at the expense of out-of-sample performance. We propose two alternatives for threshold setting: (i) including preferences in the estimation itself and (ii) setting thresholds ex-ante according to preferences only. We provide simulated and real-world evidence that this simplification results in stable thresholds and improves 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

JEL Classification: C35, C53, G01

[*] 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. The authors are grateful for 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: Halle Institute for Economic Research (IWH) – Member of the Leibniz Association Seminar, Goethe University Brown Bag Seminar, European Central Bank Financial Stability Seminar, Deutsche Bundesbank Early-Warning Modeling Seminar, and the 2015 CEUS Workshop on „Recent Issues of European Integration“.

Optimierung von Verlusten in der Krisenprognose: davor, währenddessen oder danach?*

Zusammenfassung

Frühwarnmodelle setzen üblicherweise Schwellenwerte zur Klassifikation von Krisenwahrscheinlichkeiten an und optimieren diese. Diese nachgelagerte Schwellenwertoptimierung basiert auf einer Verlustfunktion, die Präferenzen bezüglich der Art von Prognosefehlern mit einbezieht. Sie hat zwei schwerwiegende Nachteile: instabile Schwellenwerte in rekursiven Schätzungen sowie eine unnötige Reduktion der Prognosegüte (out-of-sample). Es werden zwei Alternativen zur Schwellenwertoptimierung vorgeschlagen: (i) eine Inklusion von Präferenzen in der Schätzung selbst und (ii) die Kalibrierung der Schwellenwerte ausschließlich auf Basis der Präferenzen. Anhand von simulierten und realen Datensätzen wird gezeigt, dass diese Vereinfachungen zu stabilen Schwellenwerten und verbesserter Prognosegüte führen. Die Vorschläge gelten nicht nur für Binary-Choice-Modelle, sondern gleichermaßen für den Signalansatz und alle probabilistischen Frühwarnmodelle.

Schlagwörter: Frühwarnmodelle, Verlustfunktionen, Schwellenwertsetzung, Prognosegüte

JEL-Klassifikation: C35, C53, G01

* Die Forschung von Gregor von Schweinitz wurde teilweise mit Mitteln der Europäischen Union, Europäischer Strukturfonds für regionale Entwicklung „Investition in Ihre Zukunft“ und durch das IWH Speed Project 2014/02 gefördert. Für die wertvollen Kommentare sind die Autoren Bernd Amann, Carsten Detken, Makram El-Shagi, Jan-Hannes Lang, Tuomas Peltonen und Peter Welz dankbar, ebenso wie für die Diskussionen im Rahmen des Seminars am Leibniz-Institut für Wirtschaftsforschung Halle (IWH), des Brown-Bag-Seminars an der Goethe-Universität Frankfurt am Main, des Financial-Stability-Seminars an der Europäischen Zentralbank, des Early-Warning-Modeling-Seminars an der Deutschen Bundesbank sowie des CEUS-Workshops on „Recent Issues of European Integration“ 2015.

1 Introduction

In the wake of the crisis, much research has been devoted to early-warning models for signaling the vulnerability to crisis. This provides means for triggering macroprudential policy, such as countercyclical capital buffers, and for warnings of growing macroeconomic imbalances, such as the European Commission's scoreboard. The most common setup of an early-warning model is to couple a binary-choice method or a univariate indicator with a preference-weighted loss function for ex-post optimization of signaling thresholds. This paper shows that ex-post threshold optimization finds signal in noise, which leads to unnecessary variation in thresholds and produces an in-sample overfit at the expense of out-of-sample performance. We provide two simpler alternative approaches for threshold setting ex-ante or within estimations with stable thresholds and improved out-of-sample performance.

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 & Rose (1996) and Berg & Pattillo (1999) to exchange-rate pressure, and has more recently been the predominant approach (Lo Duca & Peltonen 2013, Betz, Oprică, Peltonen & Sarlin 2014). The signaling approach relates univariate indicators to crises. It descends originally from Kaminsky & Reinhart (1999), but has also been common in the past years (Alessi & Detken 2011, Knedlik & von Schweinitz 2012). The second part of an early-warning model concerns the setting of signaling thresholds based upon loss functions tailored to the preferences of a decision-maker.[1] Demirgüç-Kunt & Detragiache (2000) introduced the notion of a policymakers' loss-function, where the policymaker mainly faces costs for missing crises (type 1 errors) and issuing false alarms (type 2 errors). Later, this loss function was extended and transformed to a usefulness function (Alessi & Detken 2011, Sarlin 2013) that indicates whether and to what extent the loss of the prediction is smaller than the loss of disregarding the model. Thus, the two elements of an early-warning model consist of a probability or indicator of vulnerability and the selection of an optimal threshold with the goal of minimizing a loss function.

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 between type 1 and type 2 errors. This approach comes with the drawback of the additional second step of threshold optimization and time-varying thresholds in recursive estimations. In practice, variation in thresholds is problematic as the rationale for policy implementation needs to descend from changes in vulnerability rather than changing thresholds. From an econometric perspective, the reason for threshold variation, given constant policymakers' preferences, is due to uncertainty about the true data-generating process (DGP). The process of threshold optimization, based on in-sample data, does not take this uncertainty into account. Thus, new observations and increased knowledge about the true DGP lead to changing thresholds. Further, by not taking uncertainty into account, optimized thresholds produce an in-sample overfit and (more often than not) an out-of-sample underfit. This paper presents two alternatives to the currently used approach for threshold setting that abstain from threshold

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

optimization. The first alternative is a weighted binary-choice model, where the weights are given by the above mentioned preferences. Instead of an optimized threshold, the fixed threshold of 50% transforms probabilistic into binary forecasts. The second alternative uses the usual binary-choice model, but sets probability thresholds ex-ante according to preferences. It can be proven that this, independently of the DGP, is the long-run optimal threshold.

This paper postulates that early-warning models based upon binary-choice methods can and should account for policymakers' preferences directly as part of the maximum-likelihood estimation or even ex-ante, rather than applying an ex-post optimization of a loss function as a second step. This has three benefits. First, it assures stable thresholds for time-varying models (i.e., equal to 0.5 or equal to preferences, respectively), which is essential for policy conclusions to descend from variation in vulnerabilities rather than thresholds. Second, we show that one-step maximization or an ex-ante threshold choice improves out-of-sample predictive power of the model and reduces the positive bias of in-sample performance on average.[2] Thus, our methods provide better performing early-warning models than the traditional threshold optimization. Third, it simplifies the process, as the second optimization step of the traditional approach is left out.

Our proposals can easily be extended to more general settings. As our critique and suggested solution applies to any loss or usefulness function optimization, it also holds for every early-warning model with this feature. Within-estimation thresholds are directly transferable to methods built on an initial maximum-likelihood estimation of event probabilities, whereas the ex-ante thresholds apply to any models resulting in probabilities of vulnerabilities. Further, our critique also extends to the signaling approach, which consists solely of the optimization step. As a weighted estimation is not possible in the context of the signaling approach, we propose it to be replaced altogether by equivalent univariate weighted binary-choice estimations with thresholds of 0.5 or univariate non-weighted binary-choice estimations with thresholds equal to preferences. In line with the above discussion, this mitigates the problem of overfit due to threshold optimization in the signaling approach, avoids an often ambiguous assumption on the sign of signaling indicators, and provides standard statistical properties of the estimator.

We provide two-fold evidence for our claims concerning in-sample and out-of-sample model performance and threshold stability. First, 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. Second, 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. For the real-world exercises, we replicate the early-warning model for currency crises in Berg & Pattillo (1999) and the early-warning model for systemic financial crises in Lo Duca & Peltonen (2013).

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

[2] This is also in line with the original evidence by El-Shagi, Knedlik & von Schweinitz (2013) and further evidence by Holopainen & Sarlin (2015), which both show and account for the fact that positive usefulness can be insignificant. This paper approaches the problem of uncertainty and significance from a different angle.

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 pre-crisis variable $C(h)$. The pre-crisis 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 (see among many others Frankel & Rose 1996, Kumar, Moorthy & Perraudin 2003, Fuertes & Kalotychou 2007, Davis & Karim 2008). In a standard binary-choice model, it is assumed that the event $C_j(h)$ is driven by a latent variable

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

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.

Non-probabilistic forecasts: The model returns probability forecasts $p_j = \mathbb{P}(y_j^* > 0)$ for the occurrence of the event. An intuitive threshold for predicted probabilities triggering counteraction would be 50%. However, crises are (luckily) scarce and (sadly) often very costly. Both features of crises pose a problem in the early-warning literature, as the estimated probability of a crisis does not take its costs into account, and seldom exceeds the intuitive threshold of 50%. For this reason, ex-post threshold optimization is commonly applied to

[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 & Fratzscher 2006).

Table 1: A contingency matrix.

		Actual class C_j	
		Pre-crisis period	Tranquil period
Predicted class P_j	Signal	Correct call *True positive (TP)*	False alarm *False positive (FP)*
	No signal	Missed crisis *False negative (FN)*	Correct silence *True negative (TN)*

binary-choice models. In order to do that, the estimated event probabilities p_j are turned into (non-probabilistic) binary point predictions P_j by assigning the value one if p_j exceeds a threshold $\lambda \in [0, 1]$ (to be optimized) and zero otherwise. The resulting predictions P_j and the true pre-crisis variable $C_j(h)$ can be summarized in a 2×2 contingency matrix, see Table 1. It should be noted that all entries of the contingency matrix depend on the threshold λ. If λ increases, the number of signals decreases, leading to both more true negatives and false negatives.

Threshold setting: Entries of the contingency matrix can be used to define a large palette of goodness-of-fit measures. If the threshold λ is optimized ex-post (like it is usually done), then λ is chosen such that a given goodness-of-fit measure is optimized.

In this paper, we use the measures defined in Sarlin (2013).[4] It uses three components to define these measures. The first component is the unconditional probability of an event $P_1 = \mathbb{P}(C_j(h) = 1) = (TP + FN)/N$ and of no event $P_2 = 1 - P_1$. The second component describes prediction errors. Type 1 errors represent the conditional probability of a missed event in case of an observed event, $T_1 = \mathbb{P}(P_j = 0|C_j = 1) = FN/(FN + TP)$, while type 2 errors represent the conditional probability of a falsely predicted event in case of no observed event, $T_2 = \mathbb{P}(P_j = 1|C_j = 0) = FP/(FP + TN)$. The third component are policymakers' preferences that are assigned to individual errors. Falsely predicted events (FP) get a weight of $\mu \in [0, 1]$, missed events (FN) a weight of $1 - \mu$. That is, the preferences should capture the relative costs of individual errors (which include economic and political costs, among others). As preferences are used to capture relative costs, they are a free parameter that should in practice be set ex-ante by the policymaker.

From these three components, three equivalent measures are derived. The first is a *loss function* calculating the frequency and preference-weighted error rates,

$$L(\mu) = (1 - \mu)P_1T_1 + \mu P_2T_2.$$

This provides nothing else than observation-spefic relative costs for type 1 and 2 errors. A policymaker could achieve a loss of $(1 - \mu)P_1$ by never issuing a signal and μP_2 by always

[4] There exists a myriad of alternative performance measures. Three other measures have been commonly applied in the early-warning literature. The noise-to-signal ratio (Kaminsky & 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 & Detragiache 2000, El-Shagi et al. 2013). Bussiere & Fratzscher (2008) and Fuertes & Kalotychou (2007) use a slightly different loss function. The usefulness measure of Alessi & Detken (2011) is conceptually close, but preferences apply to type 1 and type 2 error rates. This means, that preferences are harder to pin down, since they do not only depend on the costs of individual false predictions, but also on the frequency of these errors. Many additional measures are summarized in Wilks (2011).

issuing a signal. Thus, the second measure of *absolute usefulness* relates the (negative) loss to the alternative outcome that could be achieved by disregarding the model altogether:

$$U_a(\mu) = \min((1-\mu)P_1, \mu P_2) - L(\mu).$$

The absolute usefulness may still be hard to interpret. A scaled measure of *relative usefulness* relates absolute usefulness to the maximal achievable usefulness,

$$U_r(\mu) = \frac{U_a(\mu)}{\min((1-\mu)P_1, \mu P_2)}.$$

It should be clear that the relation between the three measures is strictly monotonic. When interpreting models, we can hence focus mainly on U_r. This entails in particular that there is one threshold optimizing the three measures (loss function, absolute and relative usefulness) simultaneously. 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 data-generating process (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 with real-time data, 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 would be produced by the threshold that maximizes usefulness out-of-sample. Optimizing usefulness based on in-sample data does not take estimation uncertainty into account and thus follows the implicit assumption that the DGP has been perfectly estimated.[5] As the estimation uncertainty is not taken into account, it can be shown that in-sample usefulness is biased upwards for optimized thresholds. As this bias is (mostly) due to estimation uncertainty, it is not likely to persist out-of-sample. In fact, if forecasting errors are not systematically related to in-sample estimation errors (in which case the model would be misspecified), the out-of-sample performance of optimized threshold will on average be biased downward.

The signaling approach: Another common approach is the signaling approach (Kaminsky & 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 and that all results are identical. Therefore, the results presented in this paper extend to the signaling approach as well.

2.2 Alternative 1: Thresholds within binary-choice models

In the current approach, a policymaker with preferences μ would transform probability forecasts into binary signals, optimizing a threshold λ while taking into account her relative

[5] It is worth noting that this uncertainty may also come from misspecified models or even simple (Excel) coding errors (Herndon, Ash & Pollin 2014).

costs of missing crises and issuing false alarms. These costs, however, can also be accounted for by weights in the log-likelihood function of the binary-choice model. For the weighted probit model, the log-likelihood function is the following:

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

Observation-specific weights have previously been used for other purposes in binary-choice models. Manski & Lerman (1977) and *Logistic Disease Incidence Models and Case-Control Studies* (n.d.) 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, Baise & Vogel 2011, Maalouf & Siddiqi 2014). Other strategies include oversampling and undersampling (King & Zeng 2001), balancing the sample directly. However, the validity of over- and undersampling hinges depends on a certain degree of homogeneity within classes, which is harder to assure for a cross-country study of economic crises. All of these strategies share the same conceptual goal with our proposal. The main difference is that the imbalance introduced in our sample is due to the differences in preferences and is thus independent of event frequencies.

For the usefulness measures, individual type 1 and type 2 errors are weighted by $(1-\mu)$ and μ, respectively. In the spirit of the above, this introduces an imbalance into the sample, as observations have now a different importance. The reason is that type 1 errors can only occur in event periods, while type 2 errors can only occur in non-event periods. As a consequence, setting observation weights $w = \mu$ accounts for this imbalance. This is possible, because both the usefulness measure and the log-likelihood are defined on an observation-specific basis.

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. Therefore, the problematic threshold-optimization step becomes unnecessary. As this holds equally well for multivariate and univariate cases, it 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 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)). To this end, it may be reasonable to be more concerned about the U.S., and less so about Finland. Similarly, in a study over a long time, the weight of countries like China should increase strongly. As this extension goes beyond the relation between type 1 and type 2 errors, it is not possible to apply in the current approach. 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 estimation results. Thus, when the early-warning model is used with a

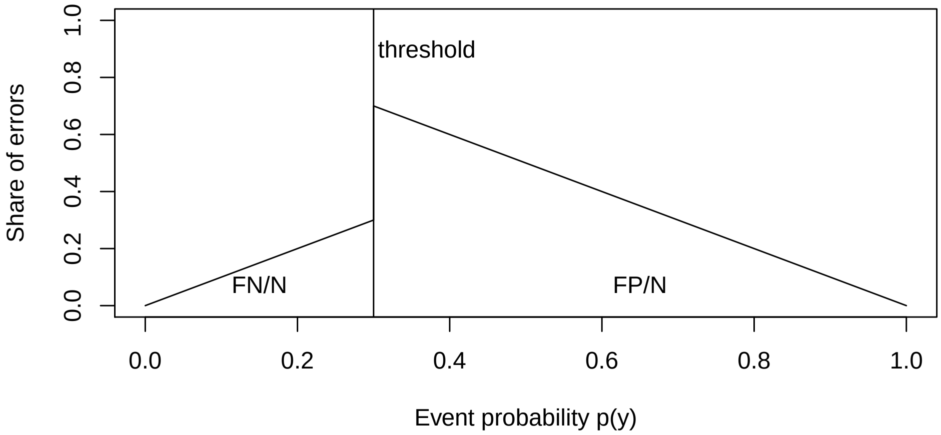

Figure 1: Type 1 and type 2 error shares at different event probabilities.
Note: The total share of errors (FN/N and FP/N) is the area under the curve, divided by the threshold λ.

set of different preferences, the outcome does not only differ in the contingency matrix, but also in different parameter estimates.

2.3 Alternative 2: Ex-ante thresholds in binary-choice models

Rather than after or within the binary-choice estimation, our final approach proposes setting the threshold before estimating the model. In the following, we will prove that the long-run optimal threshold is $\lambda^{\infty} = \mu$, independently of the DGP. That is, for infinitely many observations (when the true DGP is revealed), the optimal threshold can be written as an analytical function of the preferences. As this long-run optimal threshold is independent of the DGP, it is also optimal with a limited number of observations.

In the following, we assume that the estimated model is correctly specified.[6] This entails (a) that predicted probabilities $\widehat{p(y)}$ approach true probabilities $p(y)$ (and observed frequencies) as N increases (Hosmer & Lemeshow 1980), and (b) that out-of-sample forecasting errors are not systematically related to in-sample estimation errors. Due to this, we can work in the following with the true event probabilities p (abstracting from y). Furthermore, we observe that the probability of a missed event is just equal to the event probability (for observations with probabilities below the signaling threshold). Similarly, the probability of a false alarm is equal to one minus the event probability. This relation is shown in Figure 1. To the left of the threshold $\lambda = 0.3$, only missed events can occur (with increasing probability

[6] Note that this assumption is not only necessary for the derivation of the long-run optimal threshold, but also needs to be fulfilled by the estimation model itself, if this is to be used in early warning. Strictly speaking, we also need the assumption that the model provides some explanatory power for events. However, in the two limiting cases of no relation and perfect explanation of the latent variable, the setting of thresholds is unnecessary.

as p increases). For event probabilities $p > \lambda$, only false alarms are a concern.

The intuition for setting $\lambda^\infty = \mu$ is the following: 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 simplicity, that observations are equally distributed and that therefore every point on the curve in Figure 1 is equally important. 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 $\lambda^\infty = \mu$.

In practice, the simplification of equally distributed observations and event probabilities if certainly not true. So let us now turn to the general case, where event probabilities p have a density $f(p)$. Note that event probabilities p and their density $f(p)$ both depend on the DGP of explanatory variables X and events $C(h)$. Therefore, p and their density $f(p)$ are unknown a priori. Furthermore, while the probabilities p themselves come from the binary-choice model, the density $f(p)$ can take arbitrary forms. If, for example, the distribution of X is bimodal, so will be $f(p)$. However, as we will see, knowledge about $f(p)$ is not required to derive the long-run optimal threshold λ for given preferences μ.

The expected value of false negatives and false positives (depending on λ) is the following:

$$\begin{aligned}\mathbb{P}(FN(\lambda)) = T_1(\lambda)P_1 &= \int_0^\lambda pf(p)dp \\ \mathbb{P}(FP(\lambda)) = T_2(\lambda)P_2 &= \int_\lambda^1 (1-p)f(p)dp.\end{aligned}$$

This gives the following loss function to be minimized

$$L(\mu) = L(\mu,\lambda) = (1-\mu)\int_0^\lambda pf(p)dp + \mu\int_\lambda^1 (1-p)f(p)dp$$

Now, we are looking for the threshold λ^∞ that minimizes $L(\mu,\lambda)$, i.e. the value λ^∞ for which $\frac{\partial}{\partial\lambda}L(\mu,\lambda) = 0$. As a derivation of an integral with respect to its boundary is just the value of the integrated function at the boundary (multiplied by -1 if the derivative is taken at the lower boundary), we get

$$\frac{\partial}{\partial\lambda}L(\mu,\lambda) = (1-\mu)\lambda f(\lambda) - \mu(1-\lambda)f(\lambda) = \lambda f(\lambda) - \mu f(\lambda).$$

The unique root solution is $\lambda^\infty = \mu$, minimizing the loss function.[7] This proves the long-run optimality of the ex-ante thresholds. Therefore, we may as well set the $\lambda^\infty = \mu$ before estimating a model and deriving estimated event probabilities.[8]

[7] In order to prove that this root indeed provides the minimum of $L(\mu,\lambda)$, it suffices to note that the second derivative of $L(\mu,\lambda)$ is

$$\frac{\partial^2}{\partial\lambda^2}L(\mu,\lambda)|_{\lambda=\lambda^\infty} = f(\lambda^\infty) + (\lambda^\infty - \mu)f'(\lambda^\infty) = f(\lambda^\infty) \geq 0.$$

This follows due to $\lambda^\infty = \mu$ and the fact that f is a density, which is by definition greater or equal to zero for all values.

[8] For the alternative loss function $L(\theta) = \theta T_1 + (1-\theta)T_2$ (Alessi & Detken 2011), the long-run optimal threshold is nearly as easy to derive. It is $\lambda = \frac{(1-\theta)P_1}{\theta P_2 + (1-\theta)P_1}$.

3 Comparing optimal thresholds with simulated data

In this section, we compare the use of ex-post threshold optimization in early-warning models vis-à-vis direct use of a loss function when optimizing likelihoods, as well as with ex-ante thresholds. The exercises on simulated data provide strong evidence favoring optimization of thresholds within binary-choice models and the use of ex-ante thresholds. To illustrate differences among the approaches, we provide a large number of experiments on a range of different simulated data.

3.1 Simulation setup

Before testing our approach with real data, we apply it to simulated, simple data. We present the setup of the baseline scenario here. The more complicated 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 range 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).[9] 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 probit with fixed thresholds $\lambda^\infty = \mu$. Third, a weighted probit with threshold $\lambda^w = 0.5$. 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 different preference settings. To start with, $\mu = 0.05$ and $\mu = 0.2$ 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). $\mu = 0.8$ gives strong preference to avoiding false alarms. which accounts for high costs related to external announcements and reputation losses.

Every simulation is performed 1'000 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

[9] 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.

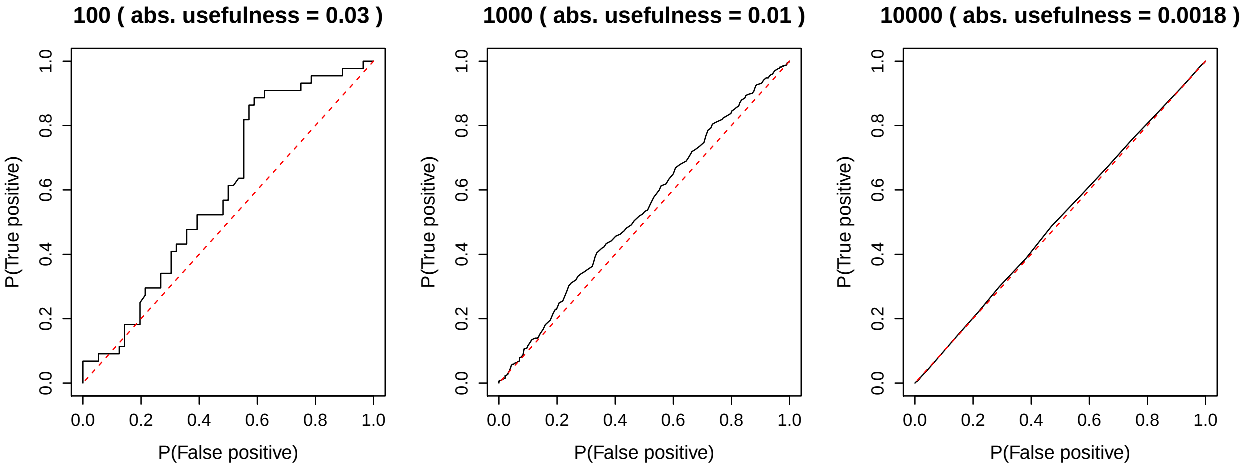

Figure 2: ROC curve for three simulations with random events (N=50, 500, 5'000) from the probit estimation.

Note: Type 2 error probability on the x-axis, (1 - type 1 error probability) on the y-axis.

with threshold optimization) is outperformed by our alternatives. In the following (with the exception of subsection 3.1.1), we will only present results from the baseline specification. Many other specifications, as described in the last subsection on robustness, yield both qualitatively and quantitatively very similar results.

3.2 Randomness of the usefulness

First, let us take a look at a specification (different from above), where events have no relation to explanatory variables, and where the event probability is 50% in every period. Figure 2 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 that one has to face 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 & Juselius (2014)).

Ideally, the distance (and therefore absolute usefulness) should be zero, because there is no relation between explanatory variables X and events $C(h)$ 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 (otherwise, the fit would be worse than for coefficients equal to zero). With less observations there is more uncertainty concerning true coefficients, resulting in a stronger upward bias of the ROC.[10] If now, in a second step, the weighted distance of the ROC curve is maximized in order to maximize usefulness, this produces an overfit. Essentially, threshold optimization chooses the best possible outcome (in-sample) instead of the most likely possible outcome.

[10] 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.

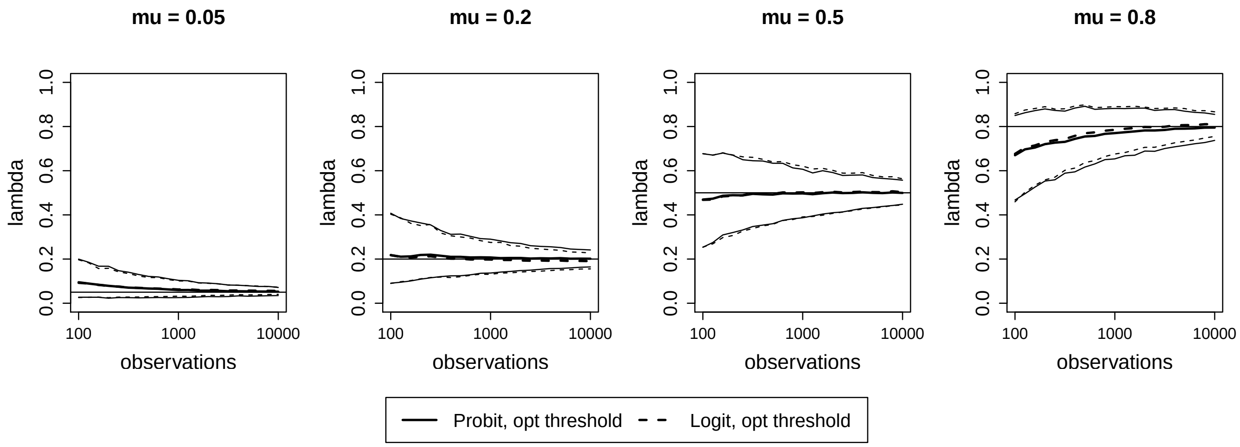

Figure 3: Mean λ^* with 90% confidence bands, for different values of μ.

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 true DGP decreases, bringing the ROC curve closer to the diagonal and bringing usefulness closer towards its true level of zero.

3.3 Variation and limit of optimized thresholds

Opposite to the previous subsection, we analyze the simple baseline specification with a true relation between the exogenous variables and the observed events (but without any additional properties that might negatively influence the estimation of the probit). Figure 3 presents the mean λ^* together with confidence bands from 1'000 replications for the different policy preferences μ and different number of observations N.

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. Therefore, in a recursive real-time context one can expect changing thresholds as new information about the true data-generating process becomes available.[11] As expected and in line with the mathematical proof of our second alternative, λ^* approaches μ as N increases. Figure 3 depicts another frequently found result: the difference between probit and logit estimations is marginal. If anything, the optimized threshold from logit estimations seems to approach μ faster – even though the logit model is misspecified.

[11] The result of Figure 3 is only indicative, as it is based on independent draws with identical numbers of observations. In a real-time context, one would only add a small number of observations (the cross-sectional dimension) to a much larger dataset, leading to a smaller change in optimized thresholds.

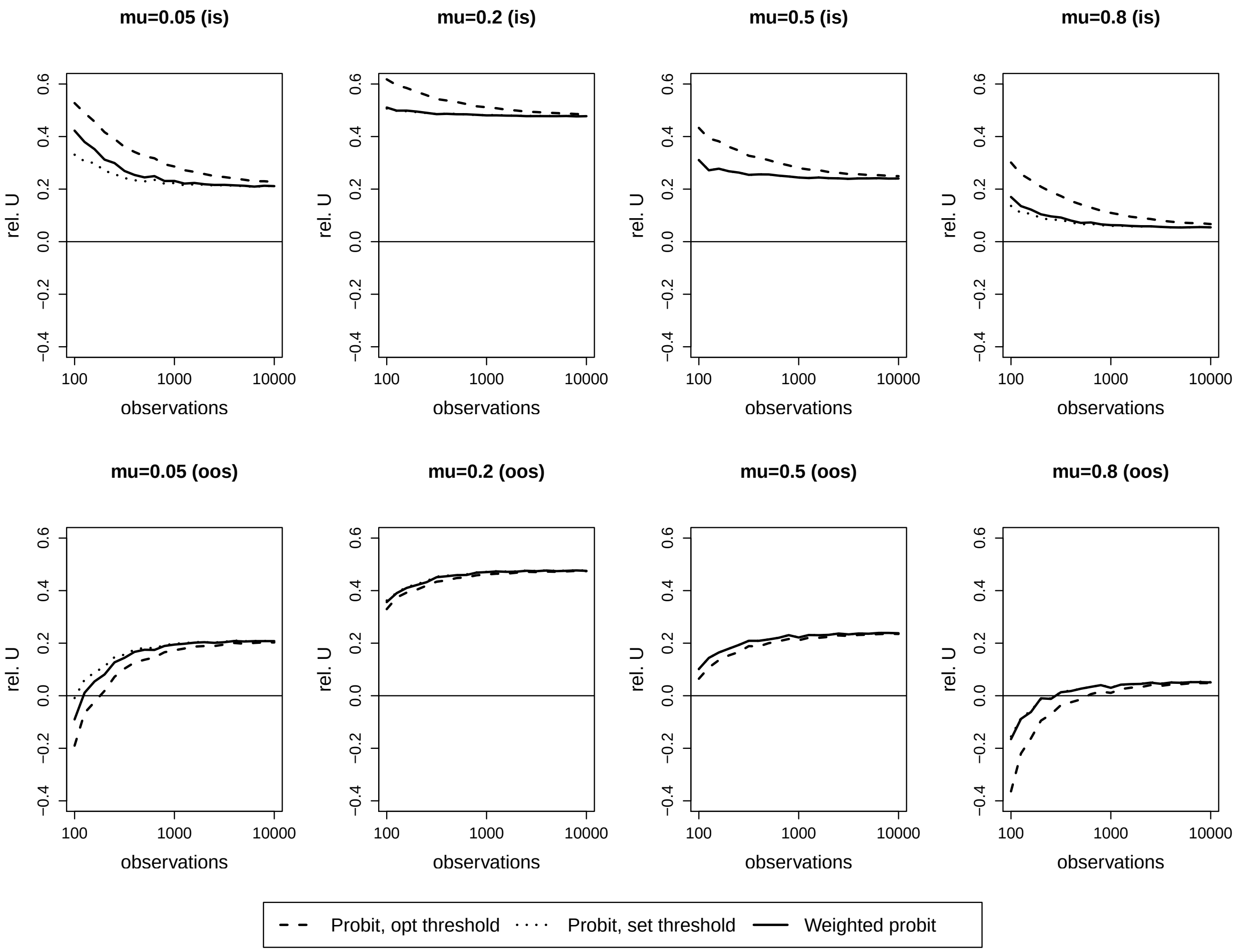

Figure 4: Mean relative usefulness of the three probit models.

Note: In-sample usefulness is higher than out-of-sample 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.

3.4 Comparison of out-of-sample performance

Differences in usefulness among different models are probably the most important aspect for practitioners, as this is the main quality measure of an early-warning model.

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 drops towards a boundary value. This view is confirmed by Figure 4 for probit models (and for logit models Figure A.1 in the appendix).[12] These figures show the mean relative usefulness from simulations with different numbers of observations for the three

[12] An alternative way to look at this would be the difference of relative usefulness between the benchmark model and our two proposals. This is shown in Figures A.2 and A.3 in the Appendix.

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 strengthens the 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 value than those of the benchmark models. Concerning in-sample usefulness (which is higher than the true value from the DGP), 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 this quality is biased upwards, it induces an overstated sense of confidence, trust and security. This bias is much lower for our proposals, where it only stems from estimation uncertainty. However, what really matters in the early-warning practice is out-of-sample usefulness. Here, our proposals perform on average better. This holds especially for the weighted logit model for all $\mu \neq 0.5$: the results of the (misspecified) weighted logit are nearly identical to the ones of the weighted probit, while out-of-sample usefulness of the standard logit (with or without threshold maximization) is far below the one for threshold probit, when μ is different from 0.5. That is, in addition to being on average better out-of-sample than their peers, weighted methods may provide robustness against method misspecification.

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 only in slightly more than 50% of the cases, see Figure 5. The exception to this is, again, the misspecified weighted logit model (Figure A.4 in the appendix). 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.

3.5 Robustness to other specifications

Above, we reported only results for a very simple specification where no estimation problems are to be expected. This may change if the complexity of the DGP is increased. For example, it could well be that estimation suffers disproportionately from slightly more complicated weighted models. Therefore, we test many different specifications. The only unchanged properties in these robustness tests are that we keep the number of exogenous variables at three, and that we keep the constant at -1. The following adjustments were tested:

1. Correlation of 50% among all exogenous variables. Multicollinearity is known to be a

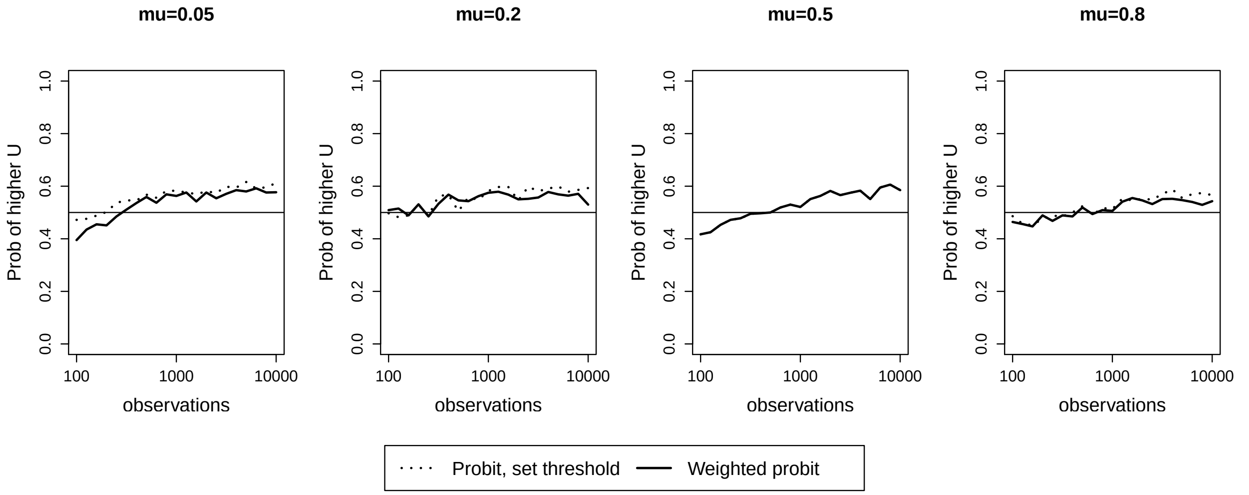

Figure 5: Probability that the weighted model has a higher usefulness than the threshold model (probit estimations).

bigger problem for binary-choice models than it is for OLS. Thus, it could potentially affect the weighted estimations strongly. The relevance in practice is evident, where an early warning model with non-correlated exogenous variables is virtually non-existent.

2. Autocorrelation of all exogenous variables with lag coefficients 0.7 (first lag) and -0.3 (second lag) in order to allow for cyclical behavior of X. Autocorrelation is highly relevant for macroeconomic variables that are usually used in early-warning models.

3. Combination of correlated and autocorrelated exogenous variables.

4. Testing omitted variables, excluding X_1 in the baseline model. As X_2 and X_3 do not provide any information on y^*, the results should be very similar to a purely random model as presented in subsection 3.2.

5. Testing omitted variables, excluding X_1 in the correlated model. Now, X_1 is correlated with X_2 and X_3. Thus, y^* given X_2 andd X_3 is not completely random. We would therefore expect results close to the correlated model.

6. Having multiple exogenous variables explaining the latent variable. We change the coefficient vector to $\beta = (1, 1, 0)$, allowing X_2 to influence y^* as well.

7. Varying the explained variance of the model. We use different coefficient vectors ($\beta_1 = (10, 0, 0)$, $\beta_2 = (0.1, 0, 0)$, $\beta_3 = (10, 10, 0)$, $\beta_4 = (0.1, 0.1, 0)$, $\beta_5 = (10, 0.1, 0)$) that increase or decrease the influence of the exogenous variables. As they are drawn from a standard normal distribution, this changes both the total variance of y^* as well as the share of (potentially) explained variance in y^*.

8. Changing the DGP of exogenous variables. It may be that different underlying distributions of X influence both the inference on y^* and the speed with wich optimized thresholds approach the long-run optimal threshold. We test this by changing the distribution of X_1, X_2, X_3 to a Cauchy distribution and a shifted exponential distribution. Both distributions are calibrated to have mean zero, and are tested with different standard deviations.

In short, the results are nearly identical for different models. That is, our baseline results are representative for the full battery of different model specifications (with the exception of test number 4).[13]

4 Real-world evidence of threshold setting

This section provides empirical evidence on threshold setting based upon policymakers' preferences for two real-world cases. We again 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. To compare both threshold stability and in-sample versus out-of-sample performance in a real-world setting, we replicate the early-warning model for currency crises by Berg & Pattillo (1999) and the early-warning model for systemic financial crises by Lo Duca & Peltonen (2013). This provides empirical evidence for a probit model and a logit model, and follows the cases used to introduce the usefulness measure in Sarlin (2013).

4.1 Currency crisis model by Berg and Pattillo (1999)

This section reproduces the probit model for currency crises by Berg & Pattillo (1999) (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 2,916 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, we use the exchange market pressure index defined by BP. A crisis occurs if the weighted average of monthly percentage currency depreciation and monthly percentage declines in reserves exceeds its mean by more than three standard deviations. Using the resulting crisis occurrences, we define an observation to be in a vulnerable state, or pre-crisis period, if it experienced a crisis within the following 24 months. This subsection provides two types of evidence: (i) in-sample versus out-of-sample performance for a one-off split of the data, and (ii) in-sample versus out-of-sample performance and threshold stability in recursive real-time estimations.

To replicate the set-up in BP, the data is divided in 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). Despite the short period of the test sample, nearly 25% of all events happen in that window. Large differences in unconditional event probabilities

[13] Detailed results can be obtained from the authors on request.

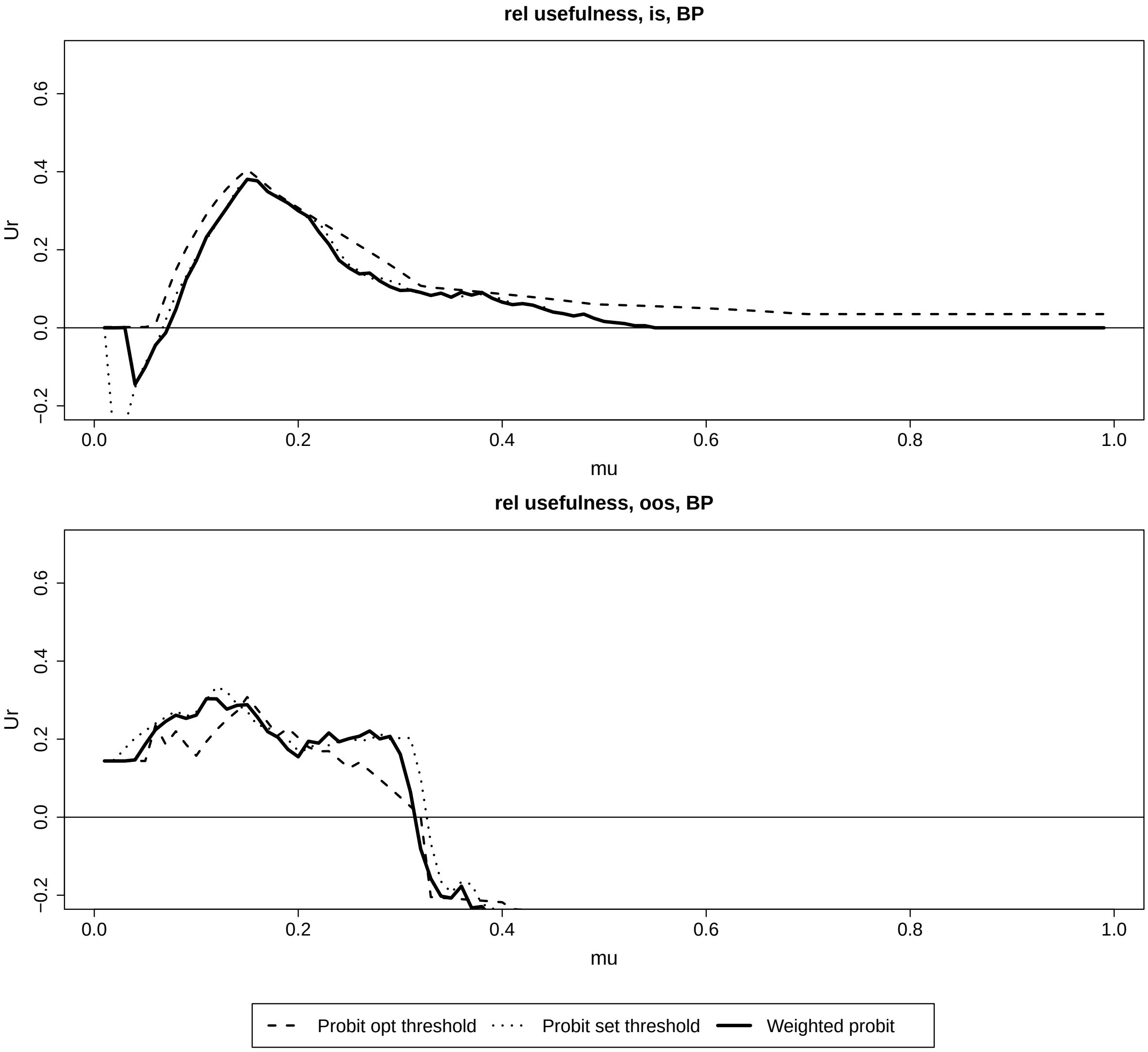

Figure 6: In-sample and out-of-sample analysis with the BP model.
Note: The models are estimated on in-sample data and applied to out-of-sample data.

point to higher uncertainty on the true DGP. This in turn should be especially problematic for the benchmark approach with optimized thresholds. To test the relative performance of before, within and after threshold setting, we test in-sample and out-of-sample performance for μ ranging between 0 and 1. That is, we test over all potential preferences that a policymaker may have.

To start with, we estimate models with all three approaches using data separated into in-sample and out-of-sample datasets. Figure 6 shows in-sample and out-of-sample performance for all the approaches over different μ values. Even though in-sample usefulness is by definition always equal to or above 0 (for optimized thresholds), the figure shows that μ values above 0.3 exhibit out-of-sample a negative usefulness, which is intuitive and in line with previous results in the early-warning literature. Across different threshold setting approaches, the figure provides evidence of generally similar performance on in-sample data, with a slight outperformance of ex-post threshold optimization. The higher usefulness of optimized thresholds may be explained by increased uncertainty and therefore "room to optimize". The picture reverses for out-of-sample usefulness. Out-of-sample performance is most often inferior for the probit model with ex-post threshold optimization. This can be also seen from table 2, which displays the mean gain (or loss) in relative usefulness from using one of our two proposed approaches over different ranges of preferences. The mean gain is calculated for μ between 0 and different maxima, to account for the fact that some extreme preferences would in practice not be chosen. In addition, we always exclude those preferences where the in-sample estimation results in negative usefulness, as in these cases the model should be disregarded altogether. For $\mu \leq 0.3$, the probit model with ex-ante threshold has on average an out-of-sample relative usefulness which is 4.5 percentage points above the benchmark model, while the weighted probit provides a gain in relative usefulness of 3.8 percentage points. Given that relative usefulness hovers around 25% for the BP model, these are sizable gains.

Max μ	Probit, set threshold	Weighted probit
0.200	0.030	0.017
0.300	0.045	0.038
0.400	0.046	0.024
0.500	-0.041	-0.057
0.600	-0.103	-0.115
0.700	-0.140	-0.150
0.800	-0.130	-0.139

Table 2: Mean difference of relative usefulness compared to the benchmark for the BP model.
Note: In every row, we report the difference of relative usefulness for μ below the maximum μ, excluding those observations where the in-sample usefulness was negative.

The second line of evidence that we put forward with the BP model is based upon recursive real-time estimations. With the same division of data, we explore the performance of the three approaches when applying them recursively. Even though the original authors do not perform this type of a test, this mimics a real-time setting when applying early-warning models. The recursive analysis implies that we only use data up to each specific month to derive model output for the same quarter in question, which is done from 1995:5

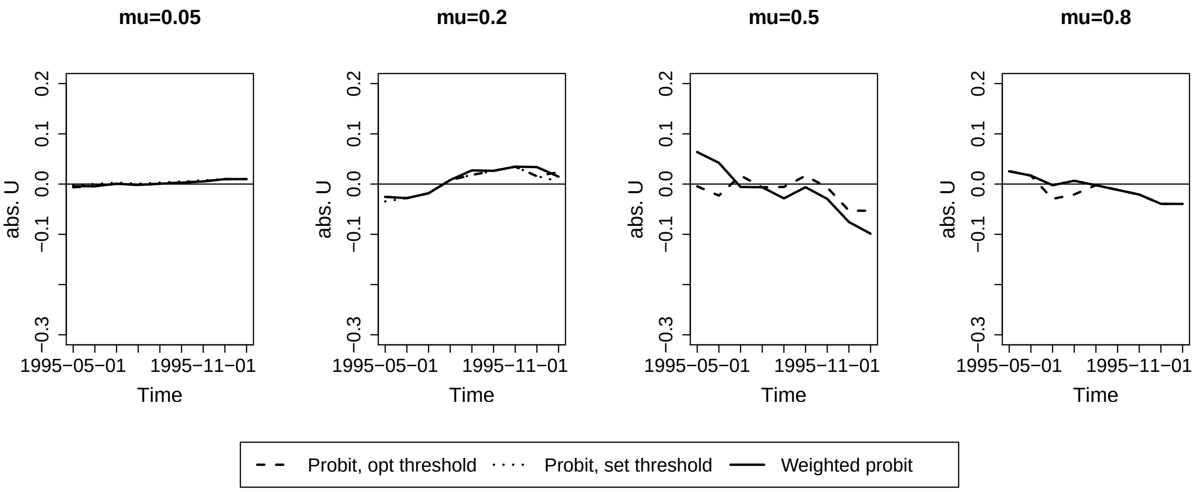

Figure 7: Recursive real-time analysis with the BP model.

Note: The models are estimated in a recursive manner by using only information available up to each month between 1995:5 and 1996:12. We display absolute instead of relative usefulness due to the negative parts for $\mu = 0.5$ and $\mu = 0.8$.

to 1996:12. We can see in Figure 7 that the three approaches generally perform equally well. However, the alternative approaches tend to outperform the benchmark in regions of positive usefulness where the model provides added value.

Another aspect that recursive models allow to explicitly illustrate is the stability of thresholds λ^*. While ex-ante and within estimation setting of thresholds assure stability by definition, a major source of uncertainty (and potentially confusion) is the variability of thresholds in ex-post optimization. Herein, we illustrate this by showing threshold variation for the BP model with ex-post optimization. Figure 8 shows a heatmap coloring of thresholds λ^* for different preferences μ. For a given μ value (horizontal row), a model with stable thresholds would also have a constant color over time. We can observe that this is not the case. For instance, for $\mu = 0.2$ the thresholds seem to vary between 18% and 24%. This illustration points out potential problems for policy.

This real-world example further strengthens our simulation results. Compared to the simulations, the mismatch between in- and out-of-sample fit may be further enhanced by the possibility that the importance of explanatory variables changes over time. Although this may not necessarily be due to a change in the DGP, it will make an estimation of the true process harder with limited number of observations. The resulting uncertainty, in turn, influences threshold optimization more negatively than the alternative approaches. In practice, it is very likely that different crises have slightly different origins.[14] That is, the importance of explanatory variables will most definitely change over time. Therefore, our example with real data provides evidence that early-warning models relying on within or ex-ante setting of thresholds are more robust to these changes than their traditional

[14] If different crises had identical origins, this would indicate strongly that economists, policymakers and market participants would be unable to learn from the past.

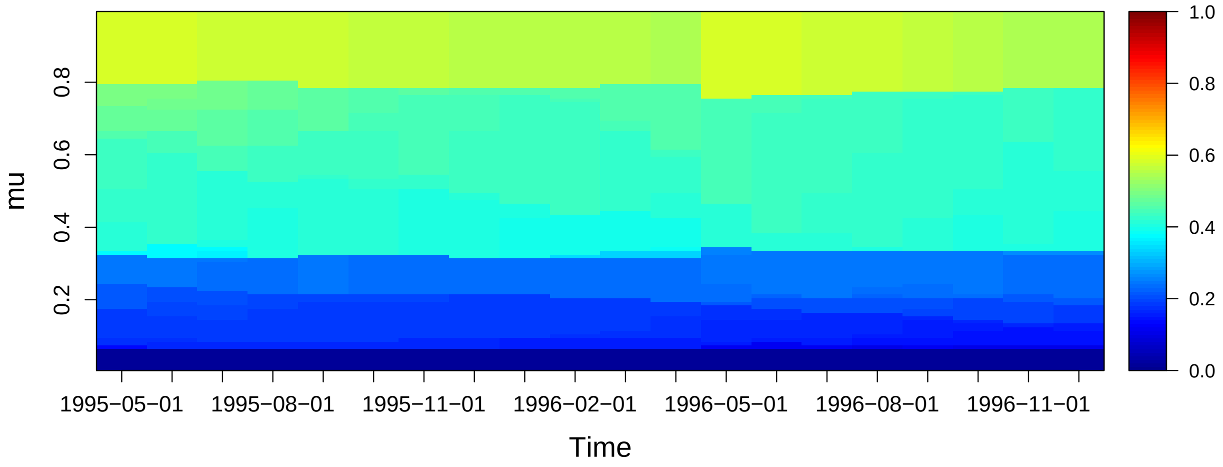

Figure 8: λ variation in recursive analysis with the BP model.
Note: The color scale refers to λ values for each μ and month. The models are estimated in a recursive manner by using only information available up to each month between 1995:5 and 1996:12.

counterparts. It is central to note that beyond evidence on out-of-sample outperformance, the most valuable merits of the two approaches relate to the stability of thresholds. In the vein of real-world cases, this is a key concern for policy as variations in thresholds due to uncertainty might be challenging to communicate. How could a policymaker be convinced to implement policies in a country with unchanged macro-financial conditions only due to a shift in "optimal" λ? Signals should depend on changes in the vulnerability indicators, not on unjustified (random) variation in thresholds. Accordingly, thresholds equaling 0.5 or μ allow by definition for constant thresholds.

4.2 Model of systemic financial crises by Lo Duca and Peltonen (2013)

This section reproduces the logit model of systemic financial crises of Lo Duca & 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 to 2010Q4 (a total of 1,729 observations). The crisis definition uses a Financial Stress Index (FSI) of 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 have 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 pre-crisis periods with a forecast horizon of two years. The dataset also consists of 14 macro-financial indicators that proxy for a large variety of sources of vulnerability, such as asset price developments, asset valuations, credit

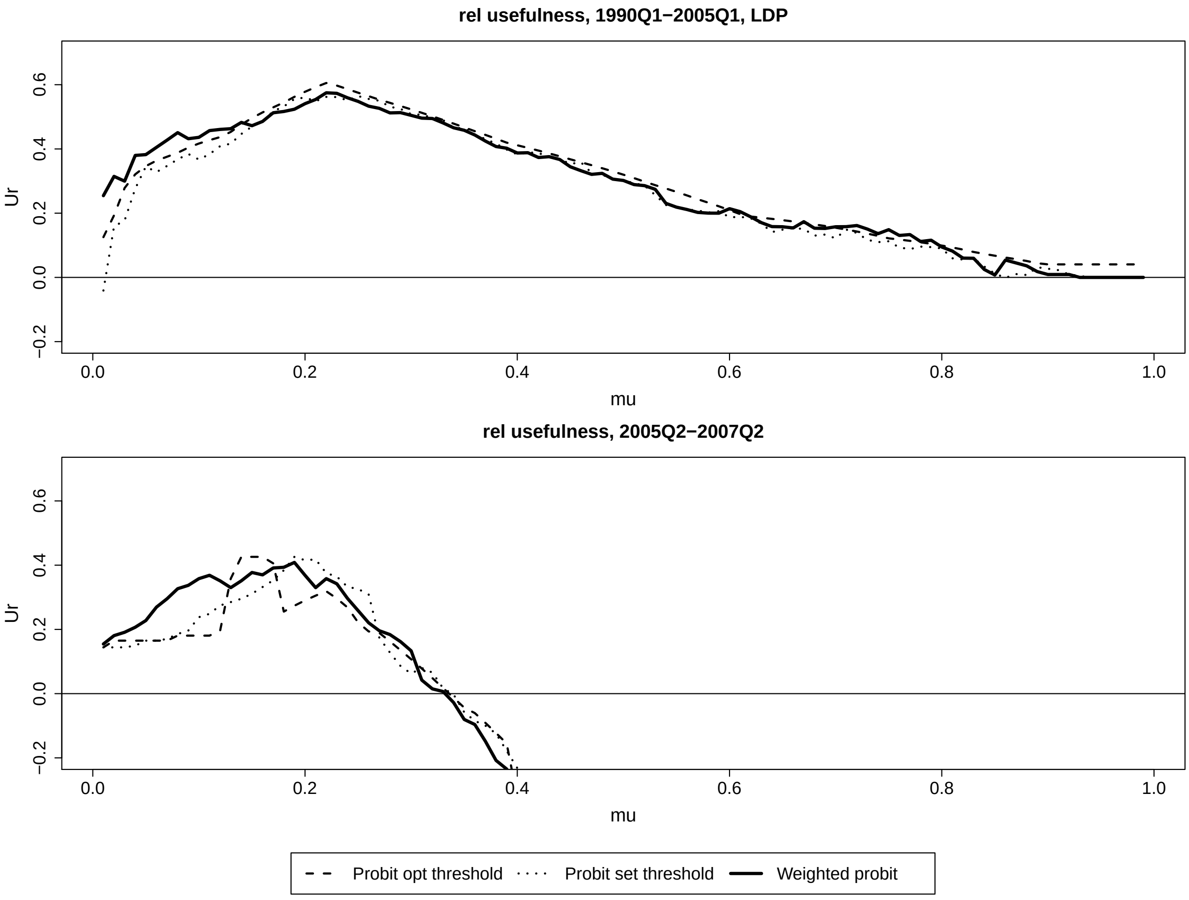

Figure 9: In-sample and out-of-sample analysis with the LDP model.
Note: The models are estimated on in-sample data and applied to out-of-sample data.

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 US. The dataset is divided into two partitions: the in-sample data (1990Q4 to 2005Q1) and out-of-sample data (2005Q2 to 2009Q2). As in the previous subsection, we analyze both the one-off split and the performance in a recursive real-time exercise.

In the vein of the above, we estimate models with all three approaches using in-sample and out-of-sample data. Figure 9 shows in-sample and out-of-sample performance for all the approaches over different μ values. This evidence again confirms that while in-sample performance is similar in nature, out-of-sample performance is most often inferior for logit models with ex-post threshold optimization. As in the case of the currency crisis model of BP, Table 3 presents the average gain over different preference ranges for our two alternative approaches. Again, the gains for reasonable ranges of μ are positive and even more sizable than in the BP case. For μ below 0.4, the weighted logit provides on average nearly 10 percentage points higher relative usefulness out-of-sample than the benchmark approach with optimized thresholds. This evidence even holds for higher preference settings that

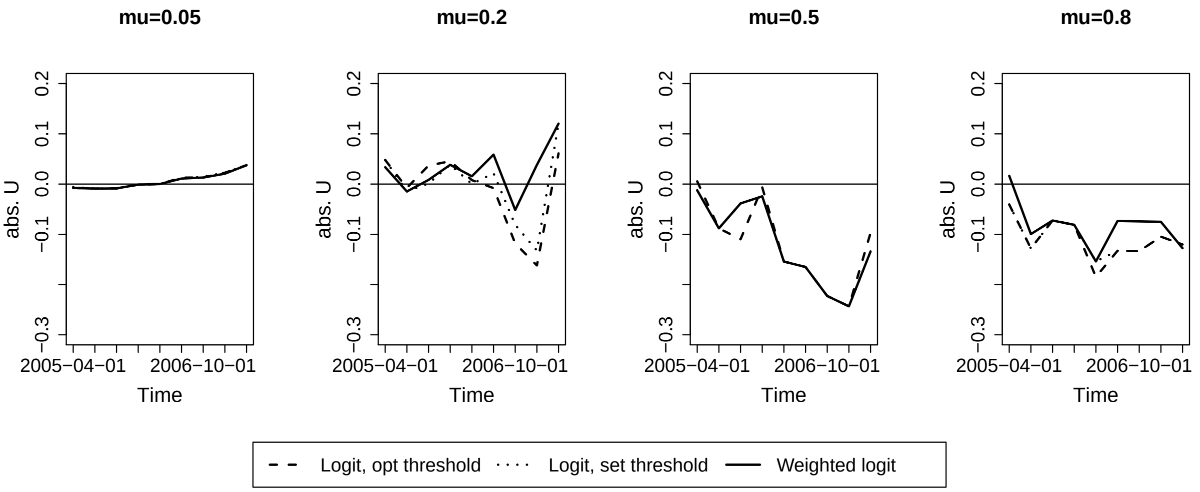

Figure 10: Recursive real-time analysis with the LDP model.

Note: The models are estimated in a recursive manner by using only information available up to each quarter between 2005Q2 and 2007Q2. We display absolute instead of relative usefulness due to the negative parts for $\mu = 0.5$ and $\mu = 0.8$.

would usually not be employed in practice.

Max μ	Logit, set threshold	Weighted logit
0.200	0.059	0.126
0.300	0.033	0.106
0.400	0.024	0.093
0.500	0.023	0.078
0.600	0.015	0.062
0.700	-0.003	0.042
0.800	-0.001	0.057

Table 3: Mean difference of relative usefulness compared to the benchmark for the LDP model.

Note: In every row, we report the difference of relative usefulness for μ below the maximum μ, excluding those observations where the in-sample usefulness was negative.

Again, a second line of evidence relies on real-time analysis by exploring the performance of different approaches when applying them in recursively during the recent global financial crisis. As performed by the original authors, and common in the literature, the recursive tests run from 2005Q2 to 2007Q2. We can see in Figure 10 that the weighted models perform better than both the ex-post and ex-ante threshold setting (out of which ex-ante is, in turn, slightly better). In line with Figure 9, we can also see that μ values above 0.5 exhibit a negative usefulness, for which the difference among approaches is smaller.

The instability of thresholds is again assessed by showing the extent to which they vary for the LDP model with ex-post optimization. Figure 11 shows a similar heatmap coloring

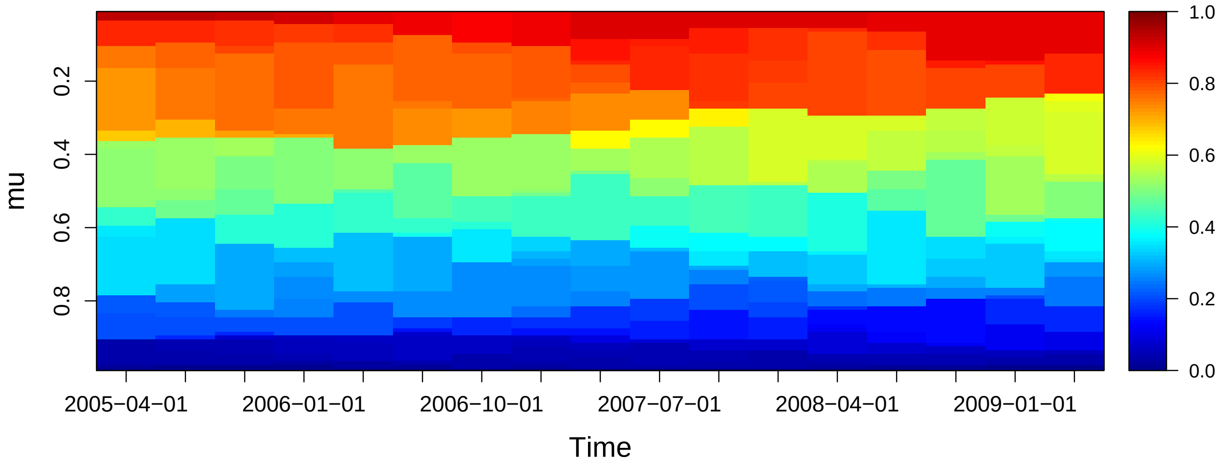

Figure 11: λ variation in recursive analysis with the LDP model.
Note: The color scale refers to λ values for each μ and quarter. The models are estimated in a recursive manner by using only information available up to each quarter between 2005Q2 and 2009Q2. Even though LDP only uses up to 2007Q2, we extend the analysis to the longest available time-series.

of thresholds λ for different preferences μ as for the BP case. We can observe that thresholds exhibit even more variation for the LDP model than for the BP model. Looking again at $\mu = 0.2$, the thresholds vary between values of below 13% and close to 28%. This points to significant uncertainty that would have serious implications in policy use.

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. Accordingly, we show that the traditional approach is exposed to identifying positive usefulness even in random data. Rather than looking for signals in noise, this paper provides simple means for noise reduction.

We propose two alternative approaches for threshold setting in early-warning models, where preferences for forecast errors are accounted for by setting thresholds within (weighted models with $\lambda^w = 0.5$) or even before ($\lambda^\infty = \mu$). To subsume, we find that these two 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, our proposals are simpler.

We think therefore that weighted models and ex-ante threshold setting are preferable approaches. Out of these two approaches, the ex-ante threshold setting is deemed to be more appealing for two reasons: (i) it is simpler than the weighted approach, and (ii) it does not require models (including coefficients, their standard errors, etc) to be re-estimated for different preference parameters. However, it comes at the disadvantage that accounting for observation-specific benefits and costs is not possible, to which weighted models can be easily extended.

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

References

Alessi, L. & Detken, C. (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. & Pattillo, C. (1999). What Caused the Asian Crises: An Early Warning System Approach, *Economic Notes* **28**(3): 285–334.

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

Bussiere, M. & Fratzscher, M. (2008). Low Probability, High Impact: Policy Making and Extreme Events, *Journal of Policy Modeling* **30**(1): 111–121.

Bussière, M. & Fratzscher, M. (2006). Towards a New Early Warning System of Financial Crises, *Journal of International Money and Finance* **25(6)**: 953–973.

Davis, E. P. & Karim, D. (2008). Comparing Early Warning Systems for Banking Crises, *Journal of Financial Stability* **4**(2): 89–120.

Demirgüç-Kunt, A. & Detragiache, E. (2000). Monitoring Banking Sector Fragility: a Multivariate Logit Approach, *The World Bank Economic Review* **14**(2): 287–307.

Drehmann, M. & Juselius, M. (2014). Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements, *International Journal of Forecasting* **30**(3): 759–780.

El-Shagi, M., Knedlik, T. & von Schweinitz, G. (2013). Predicting Financial Crises: The (Statistical) Significance of the Signals Approach, *Journal of International Money and Finance* **35**: 76–103.

Frankel, J. A. & Rose, A. K. (1996). Currency Crashes in Emerging Markets: An Empirical Treatment, *Journal of International Economics* **41**(3): 351–366.

Fuertes, A.-M. & Kalotychou, E. (2007). Optimal Design of Early Warning Systems for Sovereign Debt Crises, *International Journal of Forecasting* **23**(1): 85–100.

Herndon, T., Ash, M. & Pollin, R. (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. & Sarlin, P. (2015). Toward Robust Early-Warning Models: A Horse Race, Ensembles and Model Uncertainty, *Bank of Finland Discussion Paper 06/2015*.

Hosmer, D. & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model, *Communications in Statistics* **10**: 1043–1069.

Kaminsky, G. L. & Reinhart, C. M. (1999). The Twin Crises: the Causes of Banking and Balance-of-Payments Problems, *American Economic Review* **89**(3): 473–500.

King, G. & Zeng, L. (2001). Logistic Regression in Rare Events Data, *Political Analysis* **9**(2): 137–163.

Knedlik, T. & von Schweinitz, G. (2012). Macroeconomic Imbalances as Indicators for Debt Crises in Europe, *JCMS: Journal of Common Market Studies* **50**(5): 726–745.

Kumar, M., Moorthy, U. & Perraudin, W. (2003). Predicting Emerging Market Currency Crashes, *Journal of Empirical Finance* **10**(4): 427–454.

Lo Duca, M. & Peltonen, T. A. (2013). Assessing Systemic Risks and Predicting Systemic Events, *Journal of Banking & Finance* **37**(7): 2183–2195.

Logistic Disease Incidence Models and Case-Control Studies (n.d.).

Maalouf, M. & Siddiqi, M. (2014). Weighted Logistic Regression for Large-Scale Imbalanced and Rare Events Data, *Knowledge-Based Systems* **59**: 142–148.

Manski, C. F. & Lerman, S. R. (1977). The Estimation of Choice Probabilities from Choice Based Samples, *Econometrica* pp. 1977–1988.

Oommen, T., Baise, L. G. & Vogel, R. M. (2011). Sampling Bias and Class Imbalance in Maximum-Likelihood Logistic Regression, *Mathematical Geosciences* **43**(1): 99–120.

Sarlin, P. (2013). On Policymakers' Loss Functions and the Evaluation of Early Warning Systems, *Economics Letters* **119**(1): 1–7.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, Vol. 100 of *International Geophysics Series*, 3rd edn, Academic Press.

Appendix A

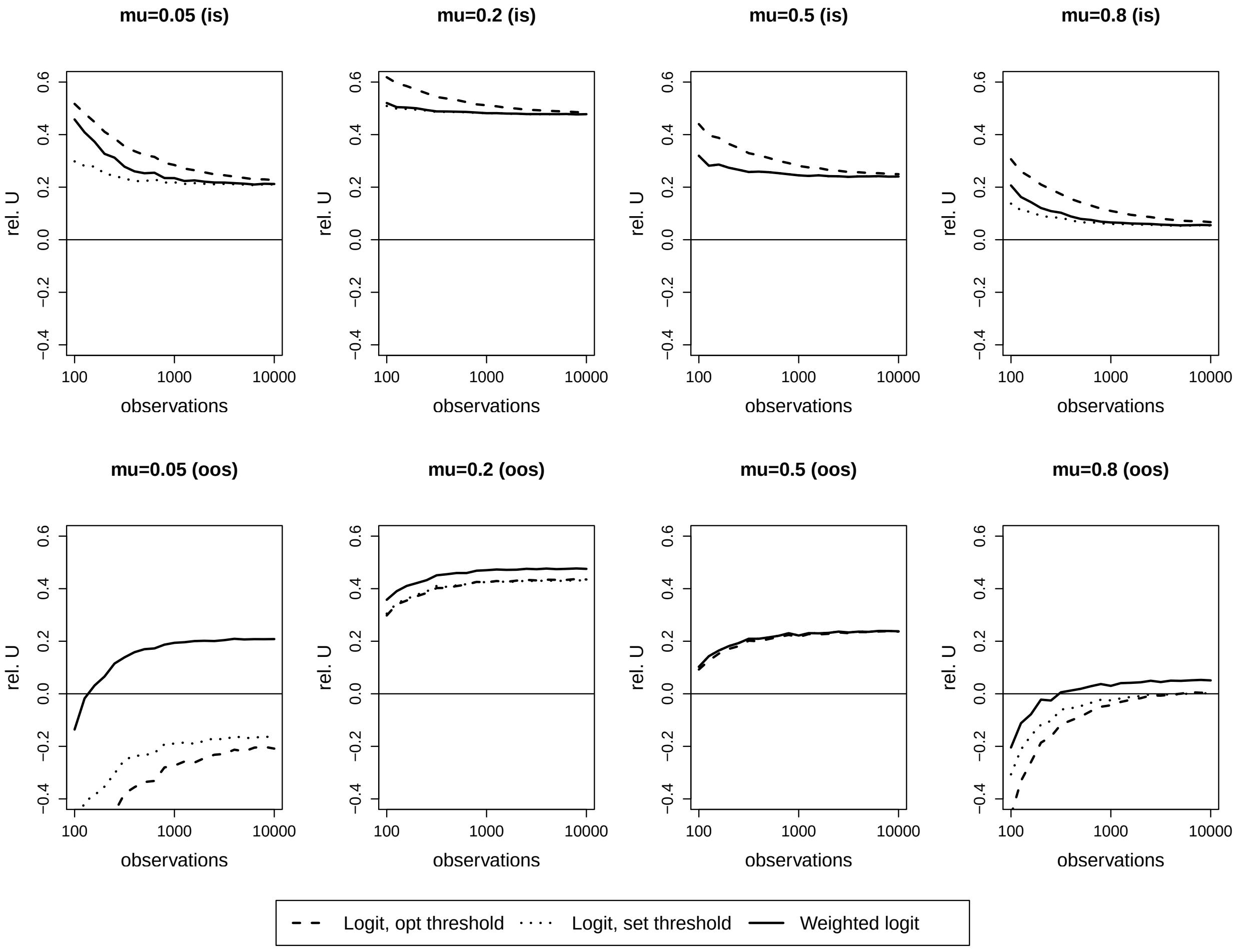

Figure A.1: Mean relative usefulness of the three (misspecified) logit models.
Note: In-sample usefulness is higher than out-of-sample usefulness for every number of observations N. The black line at zero signifies the boundary below which it is optimal not to use the model.

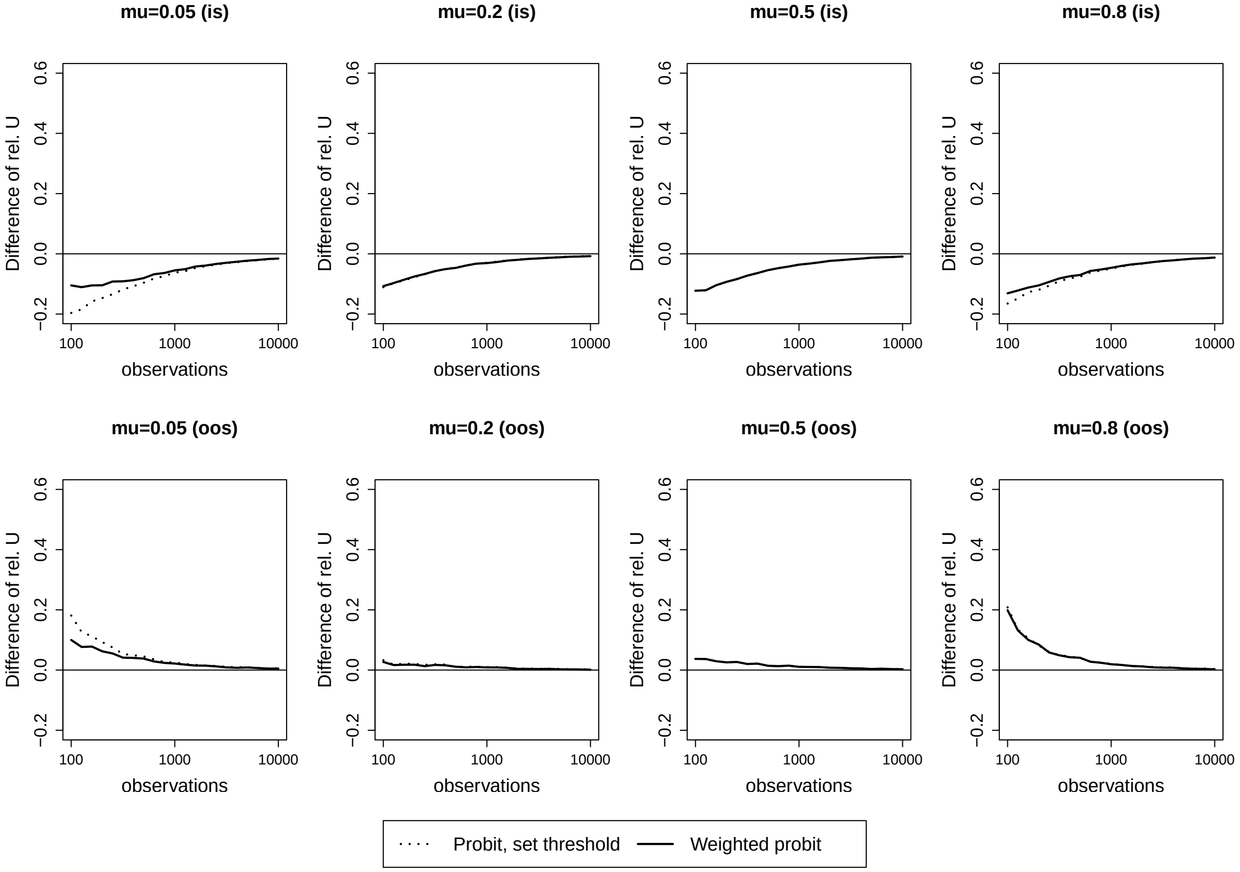

Figure A.2: Mean difference of relative usefulness of alternative probit methods to probit estimation with optimized λ.

Note: The estimation with optimized λ outperforms the two alternative approaches in-sample (negative difference), but provides lower usefulness out-of-sample (positive difference).

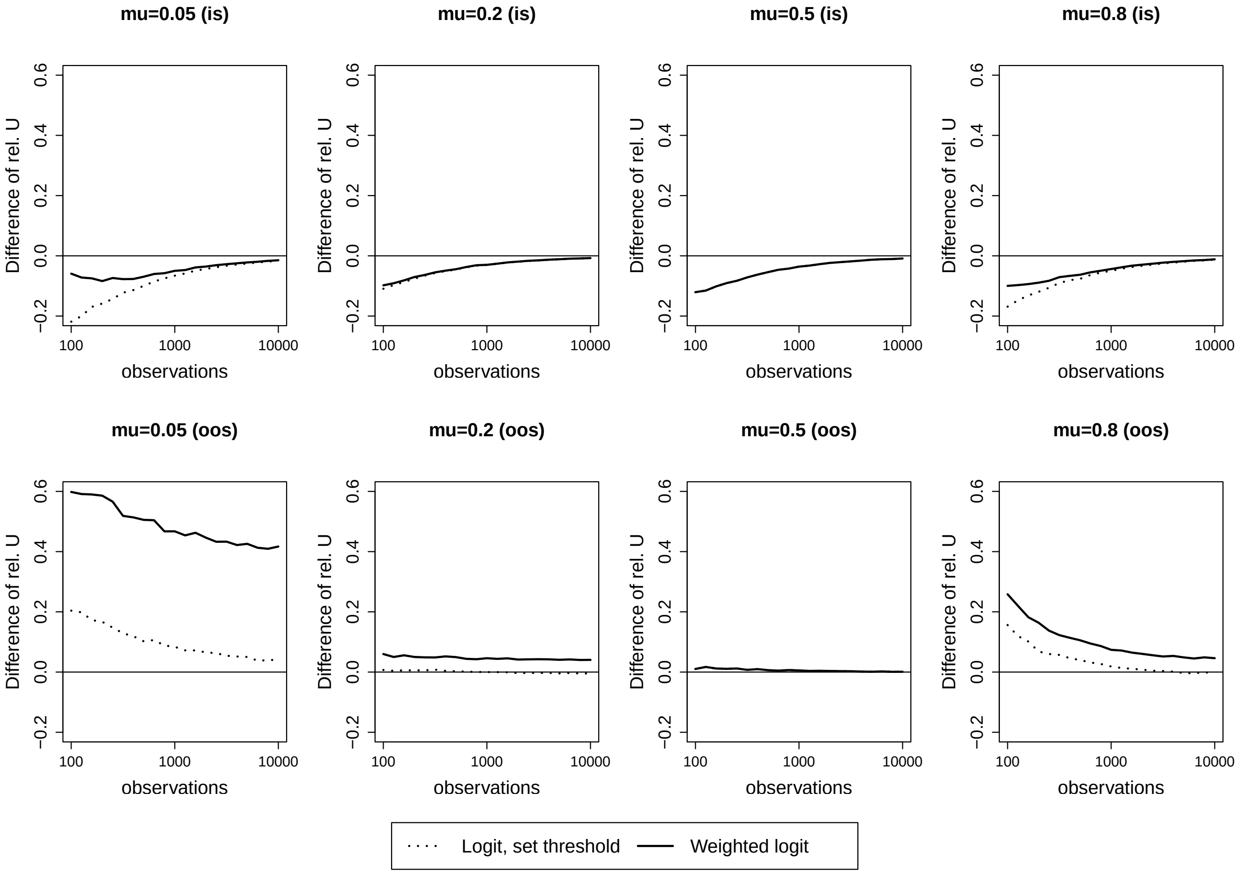

Figure A.3: Mean difference of relative usefulness of alternative logit methods to logit estimation with optimized λ.

Note: The estimation with optimized λ outperforms the two alternative approaches in-sample (negative difference), but provides lower usefulness out-of-sample (positive difference).

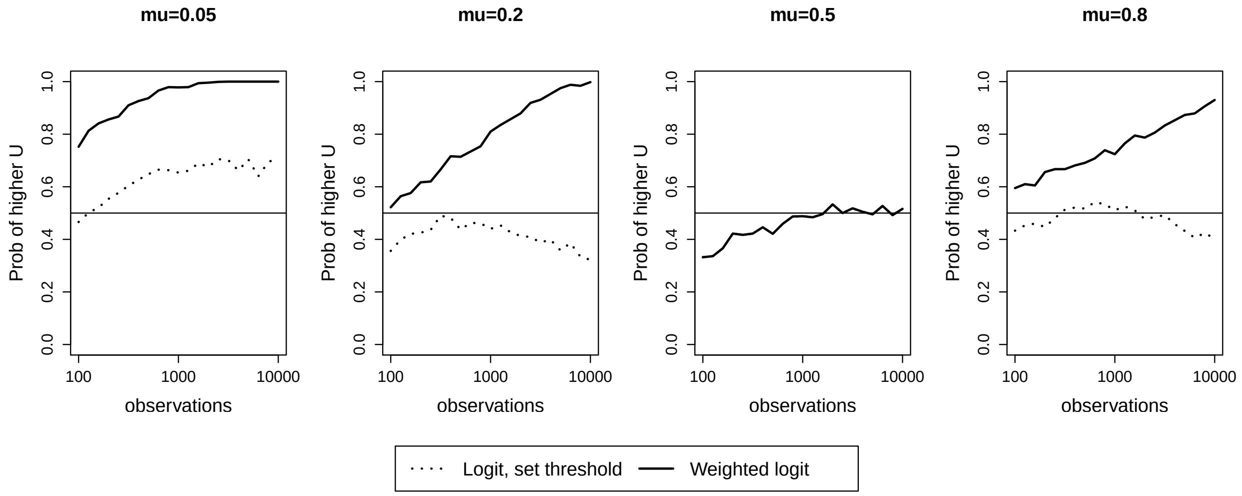

Figure A.4: Probability that the weighted model has a higher usefulness than the threshold model (logit estimations).