EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Peter Sarlin, Gregor von Schweinitz

Optimizing policymakers' loss functions in crisis prediction: before, within or after?

No 2025 / February 2017

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. 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 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

Optimizing policymakers' loss function in crisis prediction: before, within, or after?

Non-Technical Summary

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 estimated crisis probabilities with a threshold on those probabilities. The threshold is a decision rule indicating when preventive action needs to be taken. The estimation is mostly performed using binary-choice methods. Sometimes, the estimation step is circumvented by applying thresholds to univariate signaling indicators directly (the so-called signaling approach). Currently, decision thresholds are obtained from a preference-weighted loss function. That is, they are optimized on the basis of (a) costs of, or preferences for, missed crises, (b) costs of, or preferences for, falsely predicted crises, and (c) estimated error probabilities in the past.

This paper shows that such ex-post threshold optimization of the decision rule finds signal in noise, which leads to unnecessary variation in thresholds and produces an in-sample overfit of the decision rule at the expense of out-of-sample performance. We provide two simpler alternative approaches for threshold setting ex-ante or within estimations that provide stable thresholds (independent of the data and estimated crisis probabilities) as well as improved out-of-sample performance.

The current approach of optimizing thresholds, taking into account estimated error probabilities, suffers from estimation uncertainty. New observations affect estimation results and can lead to strong variation of decision thresholds over time. In practice, this variation in thresholds is problematic as the rationale for policy implementation needs to descend from changes in vulnerability rather than changing thresholds. Furthermore, the process of threshold optimization does not take estimation uncertainty into account (even though it is based on in-sample data). Thereby, optimized thresholds produce an in-sample overfit and (more often than not) an out-of-sample underfit.

The two alternatives for threshold optimization after the estimation step alleviate both problems. The first alternative is a weighted binary-choice model, where the weights are given by the above mentioned costs of, or preferences for, missed and falsely predicted crises. Instead of an optimized threshold, the fixed threshold of 50% transforms probabilistic into binary forecasts. This alternative approach combines threshold optimization and model estimation in one step. The second alternative uses the usual binary-choice model, but sets probability thresholds before the estimation, only based on preferences. It can be proven that this, independently of the data, is the long-run optimal threshold. Given probabilistic model output, it is intuitive that a decision rule is independent of the data or model specification. By way of a simple example, the decision of signaling for probabilities above 0.20 indicates a willingness to issue a false alarm (80%) vis-à-vis missing a crisis (20%).

This paper postulates that early-warning models based upon binary-choice methods can and should account for policymakers' preferences directly within the estimation or even before, rather than applying optimization of a loss function as a second step after the estimation. The

alternative approaches have three benefits. First, they assure 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 they improve out-of-sample predictive power of the model and reduce the positive bias of in-sample performance on average. Thus, our methods provide better performing early-warning models than the traditional threshold optimization. Third, the alternative approaches simplify the process, as the second optimization step of the traditional approach is left out.

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 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. 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; 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; 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. Ex-post threshold optimization has econometric and practical drawbacks. From an econometric perspective, it ignores uncertainty about the true data-generating process (DGP). Thus, optimized thresholds falsely react to (unbiased) probability estimates. They find signal in noise by exhibiting an in-sample overfit and (more often than not) an out-of-sample underfit. Further, as optimized tresholds react to probability estimates, new observations and increased knowledge about the true DGP lead to 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. The invariant threshold of 50% transforms 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%).

The alternative approaches have three benefits. First, even in recursive estimations they assure a stable threshold that only depends on preferences, which strictly relates policy guidance to macro-financial vulnerability. Second, we show that within-estimation and ex-ante threshold setting on

[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.

average improves out-of-sample predictive power and reduces the positive bias of in-sample performance.[2] Third, the alternative approaches simplify 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. The critique is not restricted to the specific loss functions 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. Further, the critique extends to the signaling approach that consists solely of the optimization step, but so do the proposed solutions via univariate binary-choice models. The proposed alternatives also 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 resulting in unbiased crisis probabilities.

We provide two-fold evidence for our claims concerning 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 and Pattillo (1999) and the early-warning model for systemic financial crises in Lo Duca and Peltonen (2013). All exercises are performed for the loss functions of Sarlin (2013) and Alessi and Detken (2011).

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. 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 (Frankel and Rose, 1996; Kumar et al., 2003; Fuertes and Kalotychou, 2007; Davis and Karim, 2008, see among many others). In a standard binary-choice model,

[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 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).

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.

Table 1: A contingency matrix.

		Actual class C_j	
		Pre-crisis 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

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 are 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 (non-probabilistic) 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 pre-crisis variable $C_j(h)$ can be presented in a 2×2 contingency matrix, see Table 1. 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. The entries 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. We mainly use the the loss and usefulness measures defined in Sarlin (2013). Three components define these measures: unconditional probabilities, type 1 and 2 error rates, and error preferences. 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 $1-\mu$. Accordingly, the preference parameter μ is a free parameter that should in practice be set ex-ante by the policymaker. From the three components, three equivalent measures are derived. The

[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 $T_2 = \mathbb{P}(P_j = 1|C_j = 0) = FP/(FP + TN)$.

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:

$$L(\mu) = \mu P_1 T_1 + (1-\mu) P_2 T_2 = \mu FN/N + (1-\mu) FP/N.$$

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

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

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. 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 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 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.

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: $L^{AD}(\theta) = \theta T_1 + (1-\theta) T_2$.[5] 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

[5] There exists a myriad of alternative performance measures with larger differences. Two other measures have been commonly applied in the early-warning literature. The noise-to-signal 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; El-Shagi et al., 2013). Bussière and Fratzscher (2008) and Fuertes and Kalotychou (2007) use a slightly different loss function. Many additional measures are summarized in Wilks (2011).

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. Therefore, the 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, pre-crisis 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:

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

For the usefulness function of Sarlin (2013), we set $w_1 = \mu$ and $w_2 = 1 - \mu$.[6] 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; 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, i.e. 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 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, the dependence of class weights on class probabilities P_1 and P_2 in the case of the loss function of Alessi and Detken (2011) may prove to be problematic as weights will in general not be constant in a real-time recursive estimation.

[6] 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 $w_2 = \frac{1-\mu}{\mu P_1 + (1-\mu)P_2}$).

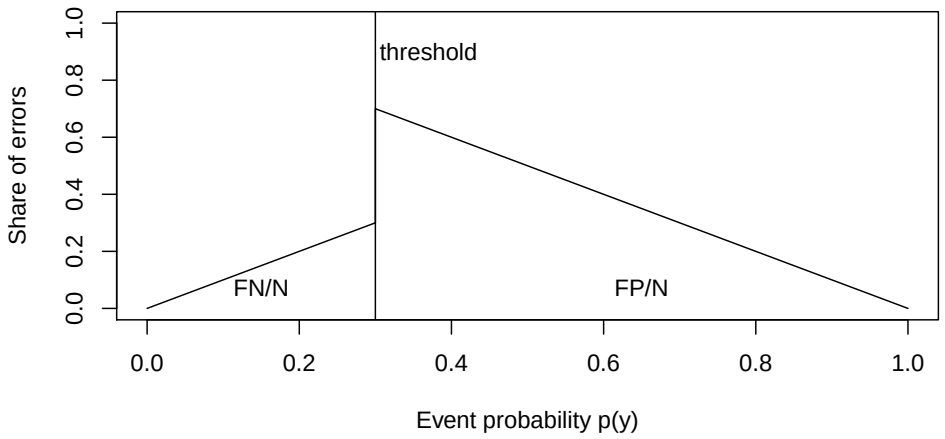

Figure 1: Type 1 and type 2 error shares at different event probabilities.
Note: The total share of errors (FN/N and FP/N) is the area under the triangle bounded by the threshold λ.

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. 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 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 data-generating process. Instead, λ should be set at a probability of vulnerability such that a policymaker is in expectation indifferent between a signal and no signal.

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. Appendix A provides a mathematical derivation of λ^{∞} for the usefulness functions of Sarlin (2013) and Alessi and Detken (2011). It is shown that policymakers are indifferent between a signal and no signal at a threshold of

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

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 $\lambda^{\infty} = 1-\mu$.

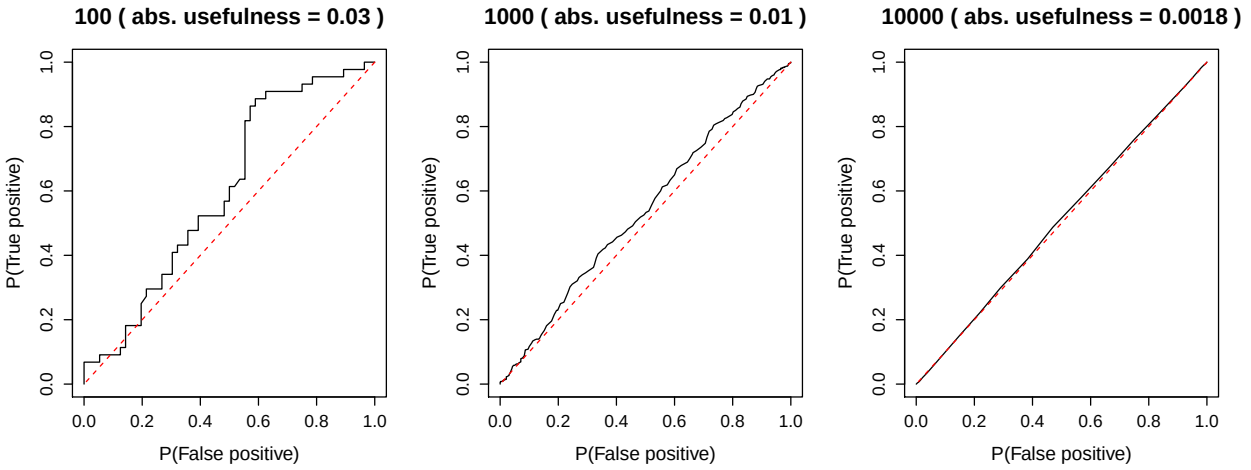

Figure 2: ROC curve for three simulations with random events (N=100, 1000, 10'000) from the probit estimation. *Note:* The type 2 error probability is given on the x-axis, (1 - type 1 error probability) on the y-axis. The absolute usefulness of the model for the drawn number of observations is given in the title.

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. 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. To illustrate differences among the approaches, 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.

3.1. Randomness of the usefulness

Before testing our approach with real data, we apply it to simulated, simple data. First, let us take a look at a data generating process, where explanatory variables and events are unrelated, 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 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 (otherwise, the fit would be worse than for coefficients equal to

zero). In fact, the area under the ROC curve (that is, 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 stronger upward bias of the ROC.[7] 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.

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.

3.2. Simulation setup

In this section, we compare our approaches in a simulation setup where explanatory variables and events are related, i.e., 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).[8] 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 four different preference settings. To start with, $\mu = 0.95$ and $\mu = 0.8$ 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.

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

[7] 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.

[8] 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.

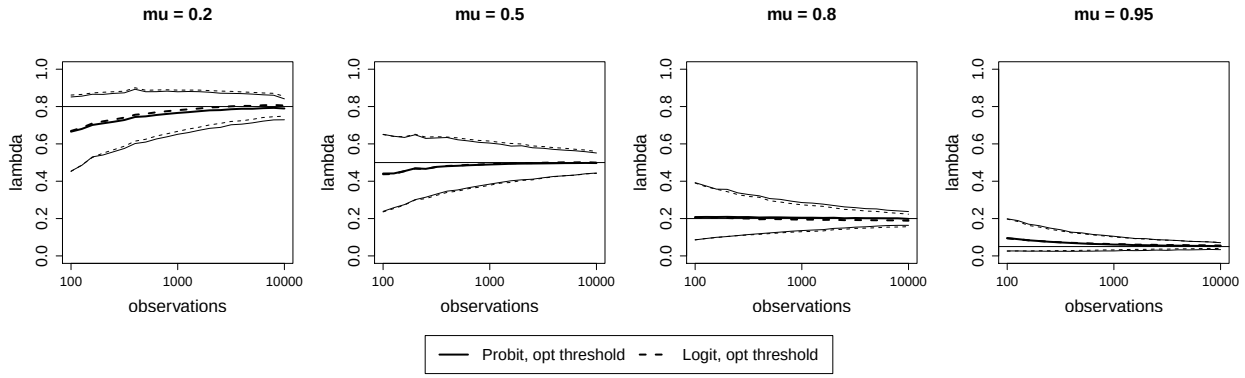

Figure 3: Mean λ^* with 90% confidence bands, for different values of μ.

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 endogeneously. We follow the latter approach.

In the following, 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.

3.3. 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 3 presents the mean λ^* together with confidence bands from R 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. 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.

3.4. 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 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

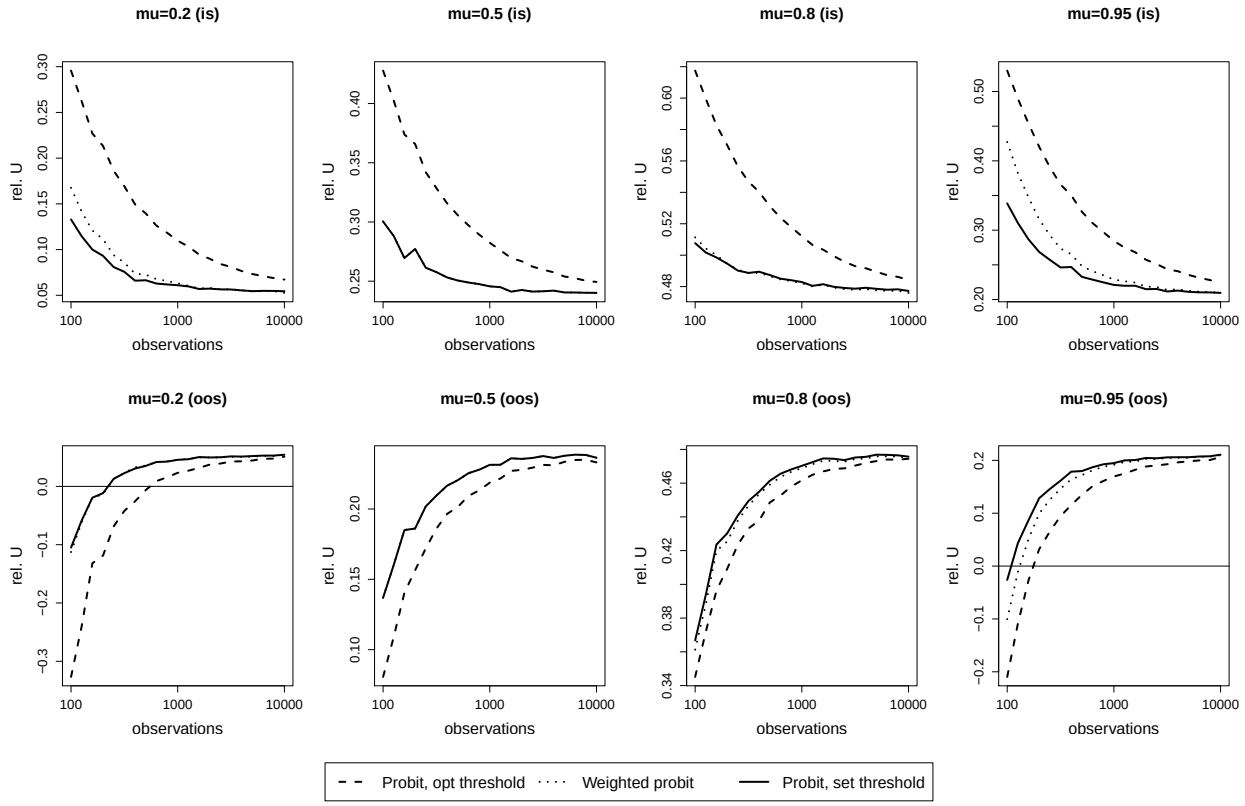

Figure 4: Mean relative usefulness of the three probit models.
Note: In-sample (*is*) usefulness is higher than out-of-sample (*oos*) usefulness for every number of observations N. The black line at zero signifies the boundary below which it is optimal not to use the model.

drops towards a boundary value. This view is confirmed by Figure 4 for probit models (and for logit models Figure B.2 in the appendix).[9] 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 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 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.

[9] 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 B.1 and B.3 in the Appendix.

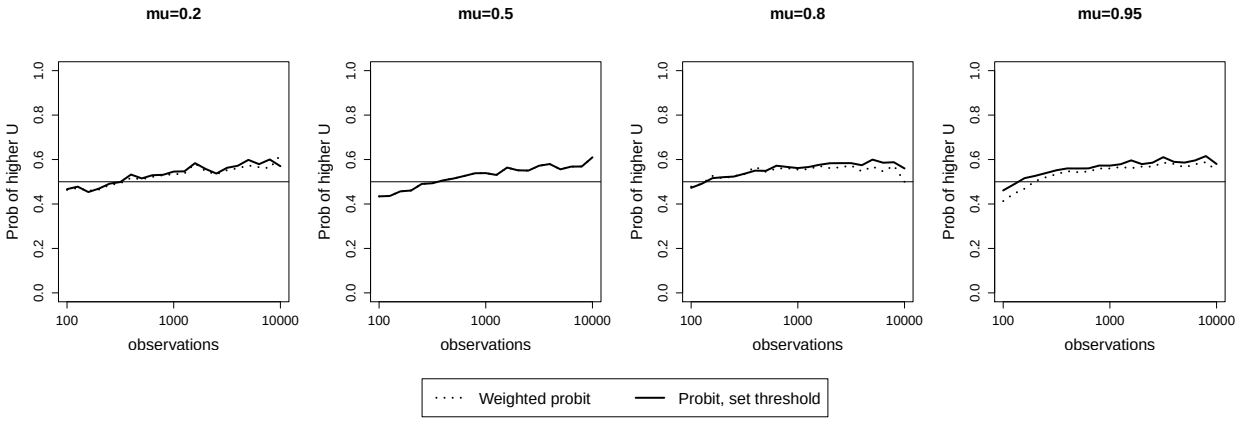

Figure 5: Probability of outperformance of alternative approaches out-of-sample (probit estimations)

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 5 (and Figure B.4 in Appendix B for the logit model). 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 in 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 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 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.
5. 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.
6. 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^*.
7. 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. For example, extreme events (outliers) that affect coefficient estimates disproportionately will be more likely if the distribution of X has heavy tails. We test the robustness of our results 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.[10]

4. Real-world evidence of threshold setting

To compare both threshold stability and in-sample versus out-of-sample performance in a real-world setting, this section provides empirical evidence on threshold setting based upon policymakers' preferences. 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. This section provides two types of evidence for the three different approaches: in-sample versus out-of-sample performance for a one-off split of the data in Subsection 4.2, and in-sample versus out-of-sample performance and threshold stability in real-time recursive estimations in Subsection 4.3.

4.1. 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 to 2010Q4 (a total of 1,729 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 90^{th} 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 pre-crisis periods

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

with a forecast horizon of two years. 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 US. The dataset is divided into two partitions: in-sample data (1990Q4 to 2005Q1) and out-of-sample data (2005Q2 to 2009Q2, out of which LDP use only data until 2007Q2 for analysis). It should be noted that the out-of-sample data contain the run-up to the great financial crisis, increasing the unconditional probability of being in an pre-crisis window from 22% in-sample to 33% out-of-sample.

The second model is the probit model for currency crises by Berg and 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, BP uses 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 defines an observation to be in a vulnerable state, or pre-crisis period, if it experienced a crisis within the following 24 months. To replicate the set-up in BP, the data is 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). Despite the short period of the test sample, nearly 25% of all events happen in that window due to the Asian crisis.

4.2. In-sample versus out-of-sample performance

In this subsection, we test in-sample and out-of-sample performance for a one-off split of the data for μ ranging between 0 and 1. That is, we test over all potential preferences that a policymaker may have.

To start with, we report results for LDP. Figure 6 shows in-sample performance in the top and out-of-sample performance in the middle for all the approaches for different μ. The bottom panel shows the probability that the two alternatives (weighted logit and logit with ex-ante set thresholds) are better out-of-sample than the current approach. This probability is derived from 1'000 draws of a panel block bootstrap over in-sample data with a block-length of 12 quarters.[11]

In-sample usefulness is by definition always equal to or above 0 (for optimized thresholds). The figure shows that out-of-sample relative usefulness is mostly negative for $\mu < 0.5$, i.e., for preferences that pay comparatively little regard to correctly predicting crises. Across different threshold setting approaches, the figure provides evidence of generally similar performance on in-sample data, with slightly higher performance of ex-post threshold optimization. The picture reverses for out-of-sample usefulness. The optimized threshold λ^* leads to inferior out-of-sample performance. For preferences $\mu \geq 0.7$, out-of-sample relative usefulness of the weighted logit is on average 0.1 higher, while the average gain is 0.03 for ex-ante threshold setting. This result is further confirmed by the bottom panel of Figure 6, which shows that our two alternatives outperform optimized tresholds in more than 50%, independently of preferences. As in our simulation studies, we thus find that our alternatives are better than the current approach in the majority of the cases, and that their average out-of-sample performance is higher. Moreover, the weighted logit is slightly better than threshold

[11] 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.

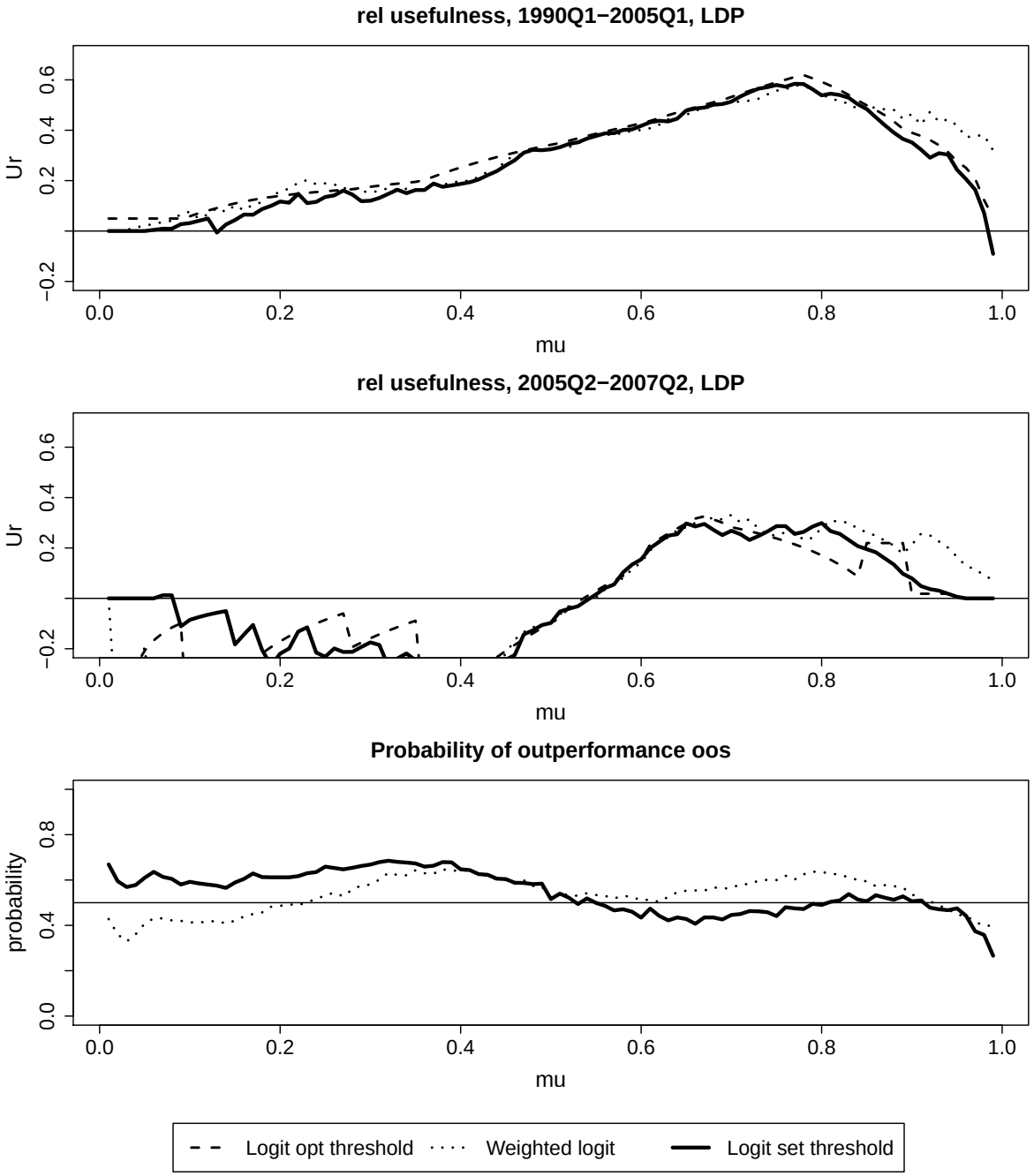

Figure 6: In-sample and out-of-sample analysis with the LDP model.
Note: The LDP model is estimated on in-sample data (top panel) and applied to out-of-sample data (mid and bottom panel).

setting ex-ante for $\mu \geq 0.7$, both in terms of mean usefulness and the probability of outperformance. As expected due to the relationship $\theta = \frac{\mu P_1}{\mu P_1 + (1-\mu)P_2}$, we obtain similar results for the loss function of Alessi and Detken (2011), as reported in Figure 7.

Turning to the BP model, Figure 8 shows out-of-sample performance for all three approaches for the loss function of Sarlin (2013) in the top and Alessi and Detken (2011) in the bottom panel. The results are qualitatively very comparable the ones in the LDP case. Especially the logit with ex-ante set thresholds strongly outperforms ex-post threshold optimization for the majority of μ (and θ).

4.3. Real-time recursive behavior

The second 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 performance of the three approaches when applying them recursively over the out-of-sample part of the data. This mimics a real-time setting when applying early-warning models. The recursive analysis implies that we use information available at each period t to derive model output for the same period in question. Due to the panel dimension of the dataset, we have enough observations in every period to calculate period-specific performance measures. Another aspect that recursive models allow to explicitly illustrate is the (in)stability of thresholds λ^*, while ex-ante and within estimation setting of thresholds assure stability by definition.

For the LDP paper, the recursive tests run from 2005Q2 to 2007Q2. Figure 9 shows the resulting absolute usefulness in each quarter in the out-of-sample data. We can see that the the two alternatives often outperform the current approach of ex-post threshold optimization. Weighted logit is again the better of the two approaches. In line with Figure 6, we find negative usefulness for μ values below 0.5, for which the difference among approaches is smaller. However, usefulness is calculated on a much smaller number of observations. Therefore, the results are much less stable than in the previous subsection, which can be seen by the large fluctuations of period-specific usefulness.

In the case of BP, our recursive tests run from 1995:5 to 1996:12.[12] The results are stronger in favour of our two proposed alternatives than in the LDP case. For most values of μ (especially for values above 0.5) and for most of the out-of-sample periods, the alternative approaches outperform the benchmark in regions of positive usefulness where the model provides added value.

A major source of uncertainty (and potentially confusion) is the variability of thresholds in ex-post optimization. We illustrate this by showing threshold variation for the LDP model with ex-post optimization. Figure 11 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 $\mu = 0.8$ 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 B.5 in Appendix B for the BP model.

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 potentially reduced out-of-sample usefulness and unstable probability thresholds. Accordingly, we show that the traditional approach is exposed to

[12] The original authors do not perform this type of a test

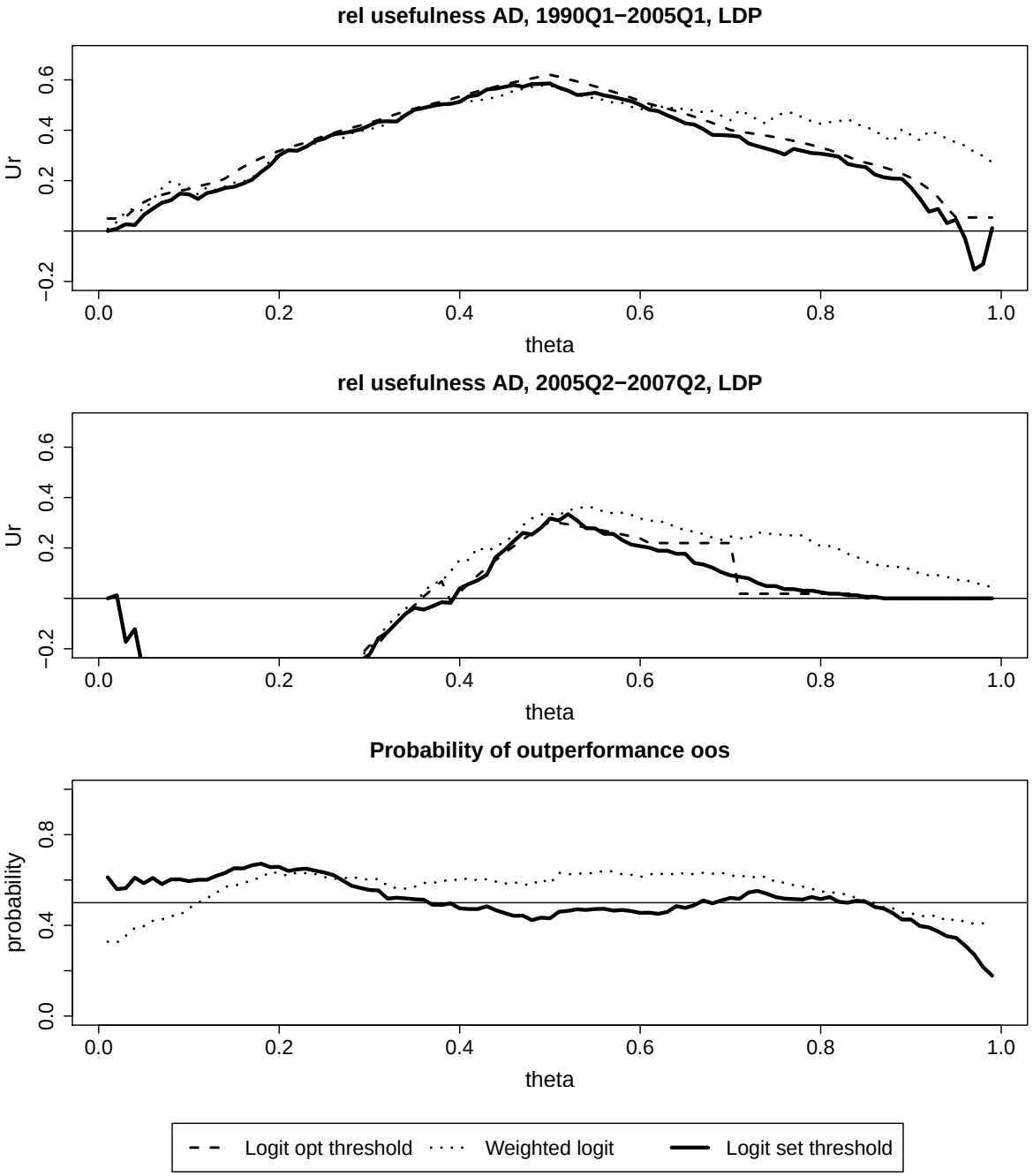

Figure 7: In-sample and out-of-sample analysis with the LDP model for Alessi-Detken preferences.
Note: The LDP model is estimated on in-sample data (top panel) and applied to out-of-sample data (mid and bottom panel).

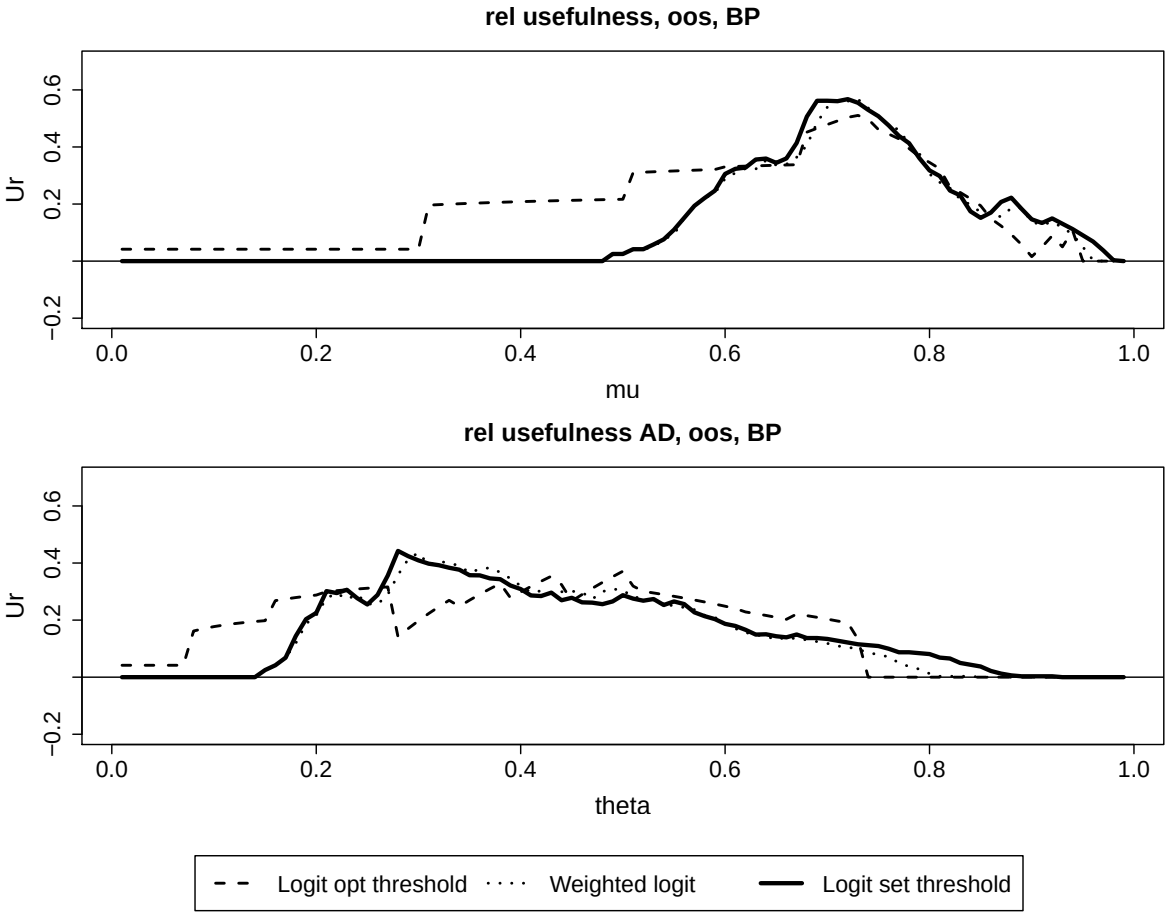

Figure 8: Out-of-sample analysis with the BP model, Sarlin and Alessi-Detken preferences.
Note: The BP models is estimated on in-sample data and applied to out-of-sample data, with different results for different usefulness functions.

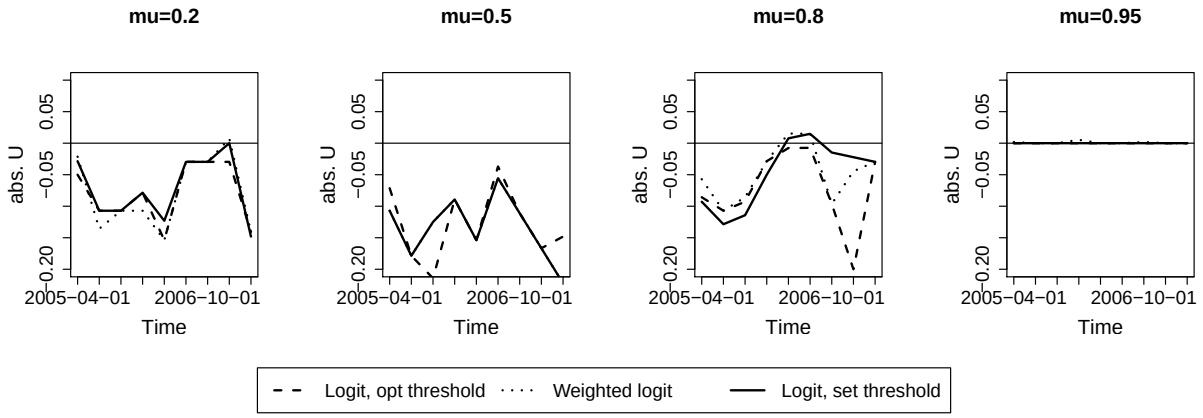

Figure 9: Recursive real-time analysis with the LDP model.
Note: The models are estimated recursively by using only information available up to each quarter between 2005Q2 and 2007Q2. We display absolute instead of relative usefulness due to the negative parts for $\mu \leq 0.5$.

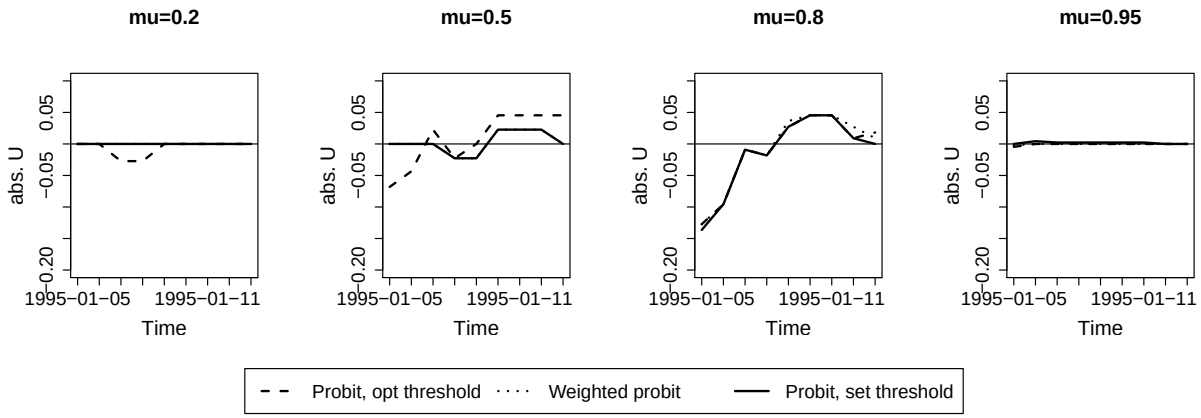

Figure 10: Recursive real-time analysis with the BP model.
Note: The models are estimated in a recursive manner by using only information available up to each month between 1995:5 and 1996:12. We display absolute instead of relative usefulness due to the negative parts for $\mu = 0.5$ and $\mu = 0.8$.

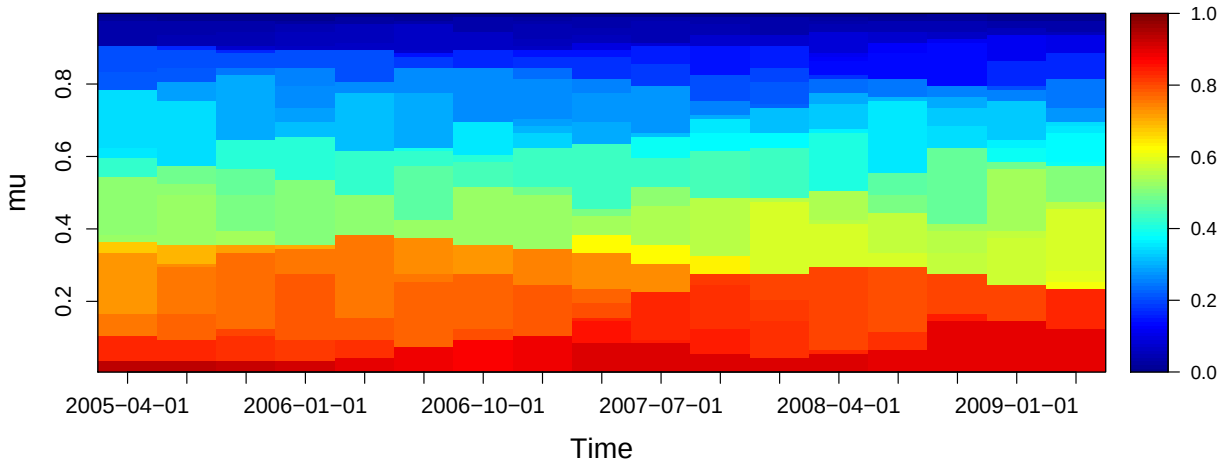

Figure 11: Variation of λ^* in recursive analysis with the LDP model.
Note: The color scale refers to λ^* values for each μ and quarter. The models are estimated in a recursive manner by using only information available up to each quarter between 2005Q2 and 2009Q2. Even though LDP only uses data up to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

identifying positive usefulness even in random data (see Figure2). 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 ($\lambda^\infty = 1 - \mu$).

Including preferences as estimation weights (resulting in $\lambda^w = 0.5$) in the early-warning model outperforms optimized thresholds out-of-sample in 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.[13]

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 3. That is, in 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 show our results in simulation studies with simple and known data-generating processes. These examples already show that all our results are robust to small degrees of misspecification (if a logit model is applied to data generated by a probit process). In a second part, we apply our approaches to real-world examples, strengthening 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. 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 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.

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

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

improved by our approaches, while the bias on in-sample usefulness is reduced. Third, our proposals are simpler.

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.

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

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

Chawla, N., Japkowicz, N., Kotcz, A., 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explorations 6 1, 1–6.

Davidson, R. and MacKinnon, J.G., 2000. Bootstrap Tests: How Many Bootstraps?. Econometric Reviews 19 (1), 55–68.

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.

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 45 (8), 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.

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

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. Wiley.
Wilks, D. S., 2011. Statistical Methods in the Atmospheric Sciences, 3rd Edition. Vol. 100 of International Geophysics Series. Academic Press.

Appendix A: Mathematical derivation of the long-run optimal threshold

This appendix provides a mathematical derivation of the long-run optimal threshold given in equation (1).

In the following, we assume that the estimated model is correctly specified.[14] This entails (a) that predicted probabilities $\widehat{p(y)}$ approach true probabilities $p(y)$ (and observed frequencies) as N increases (Hosmer and 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 as p increases). For event probabilities $p > \lambda$, only false alarms are a concern.

In general, 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:

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

Using these in the loss function of Sarlin (2013) results in

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

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

$$\frac{\partial}{\partial \lambda} L(\mu, \lambda) = \mu\lambda f(\lambda) - (1-\mu)(1-\lambda)f(\lambda) = \lambda f(\lambda) - (1-\mu)f(\lambda) \stackrel{!}{=} 0.$$

The unique solution is $\lambda^{\infty} = (1-\mu)$, minimizing the loss function.[15] This proves the long-run

[14] 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. 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.

[15] In order to prove that $\lambda^{\infty} = (1-\mu)$ indeed provides the minimum of $L(\mu, \lambda)$, it suffices to note that the second derivative of $L(\mu, \lambda)$ is

$$\frac{\partial^2}{\partial \lambda^2} L(\mu, \lambda)|_{\lambda=\lambda^{\infty}} = f(\lambda^{\infty}) + (\lambda^{\infty} - (1-\mu))f'(\lambda^{\infty}) = f(\lambda^{\infty}) \geq 0.$$

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.

optimality of the ex-ante thresholds.

For the loss function of Alessi and Detken (2011), the solution is nearly as easy to derive. The loss function to be minimized is

$$L(\theta, \lambda) = \theta T_1 + (1-\theta)T_2 = \theta \frac{1}{P_1} \int_0^{\lambda} pf(p)dp + (1-\theta)\frac{1}{P_2}\int_{\lambda}^{1}(1-p)f(p)dp.$$

Setting the partial derivative of $L(\theta, \lambda)$ with respect to λ to zero results in the long-run optimal threshold of $\lambda^{\infty} = \frac{(1-\theta)P_1}{(1-\theta)P_1+\theta P_2}$.

This shows that we may as well set the threshold to λ^{∞} as in equation (1) before estimating a model.

Appendix B: Additional figures

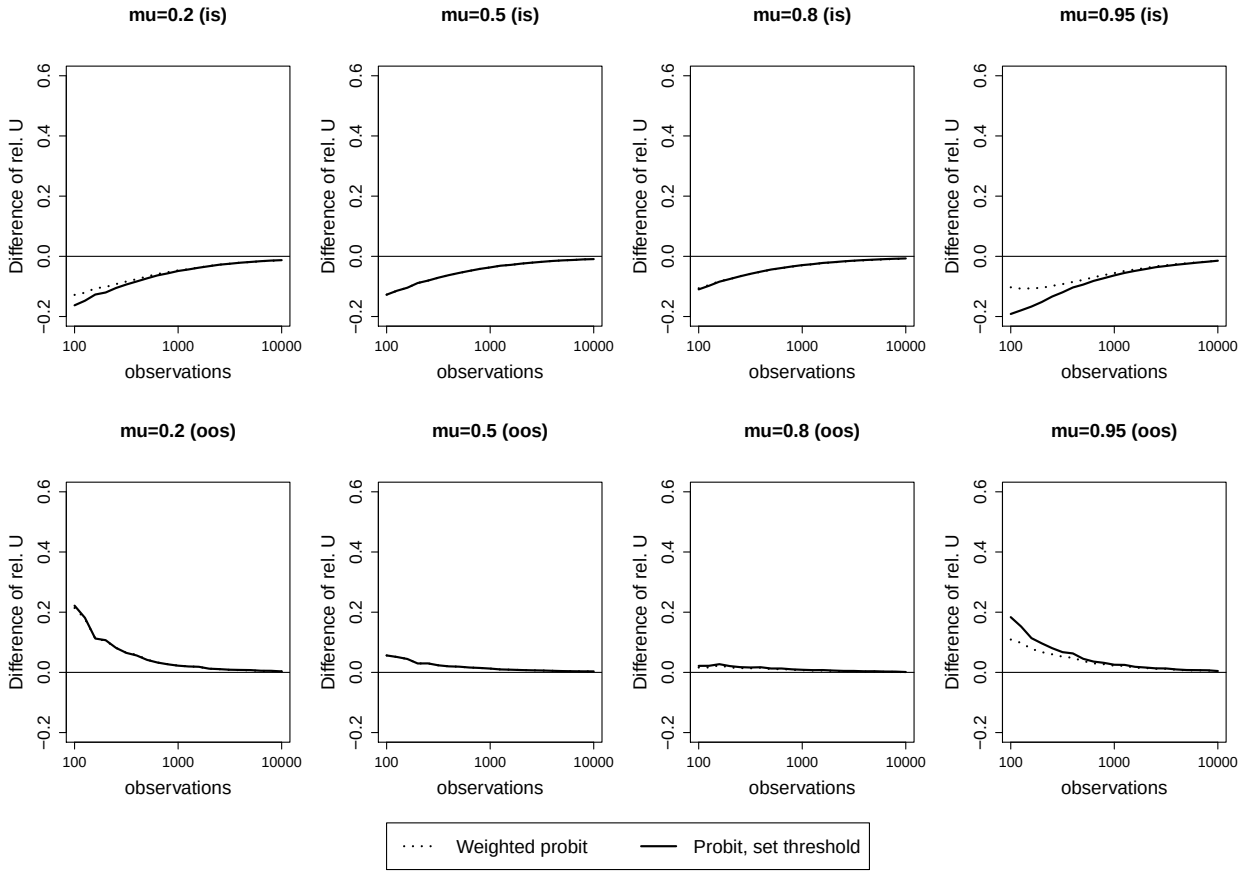

Figure B.1: Mean difference of relative usefulness of alternative probit methods to a probit with optimized λ^*. *Note:* The estimation with optimized λ^* outperforms the two alternative approaches in-sample (*is*, negative difference), but provides lower usefulness out-of-sample (*oos*, positive difference).

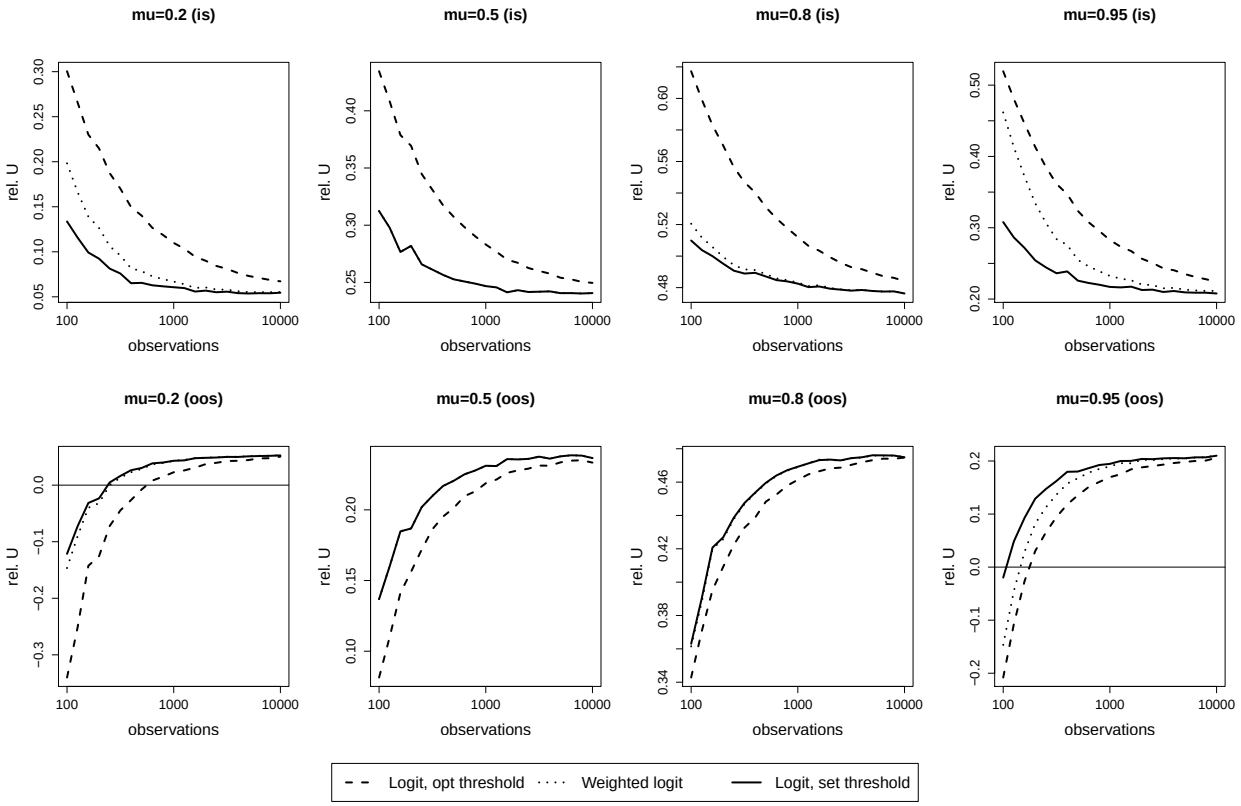

Figure B.2: Mean relative usefulness of the three logit models.
Note: In-sample usefulness is higher than out-of-sample usefulness for every number of observations N. The black line at zero signifies the boundary below which it is optimal not to use the model.

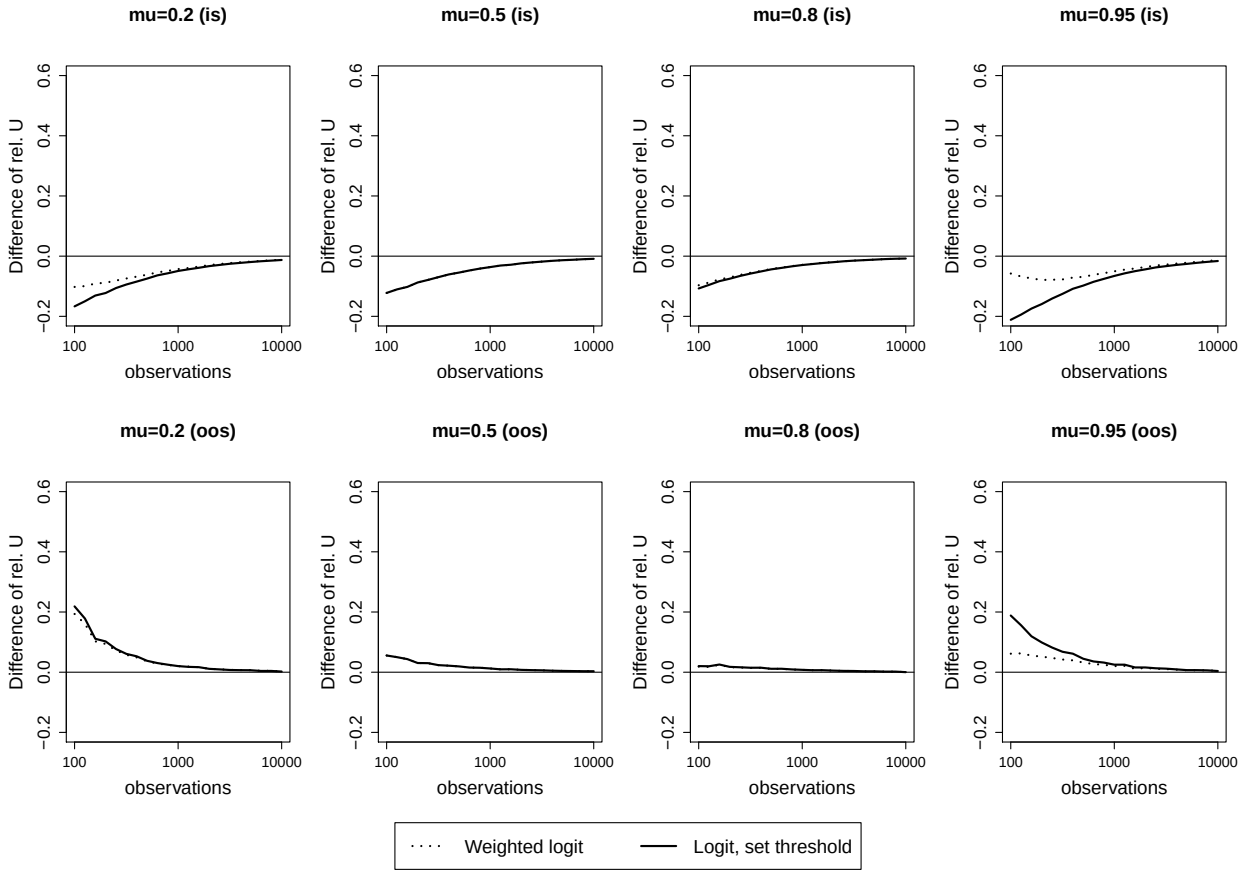

Figure B.3: Mean difference of relative usefulness of alternative logit methods to a logit with optimized λ^*. *Note:* The estimation with optimized λ^* outperforms the two alternative approaches in-sample (*is*, negative difference), but provides lower usefulness out-of-sample (*oos*, positive difference).

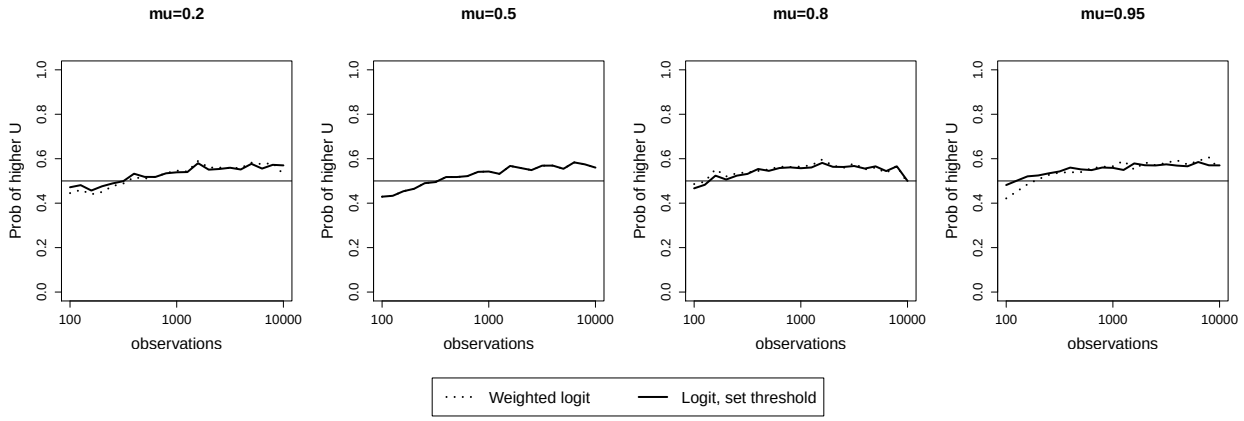

Figure B.4: Probability of outperformance of alternative approaches out-of-sample (logit estimations)

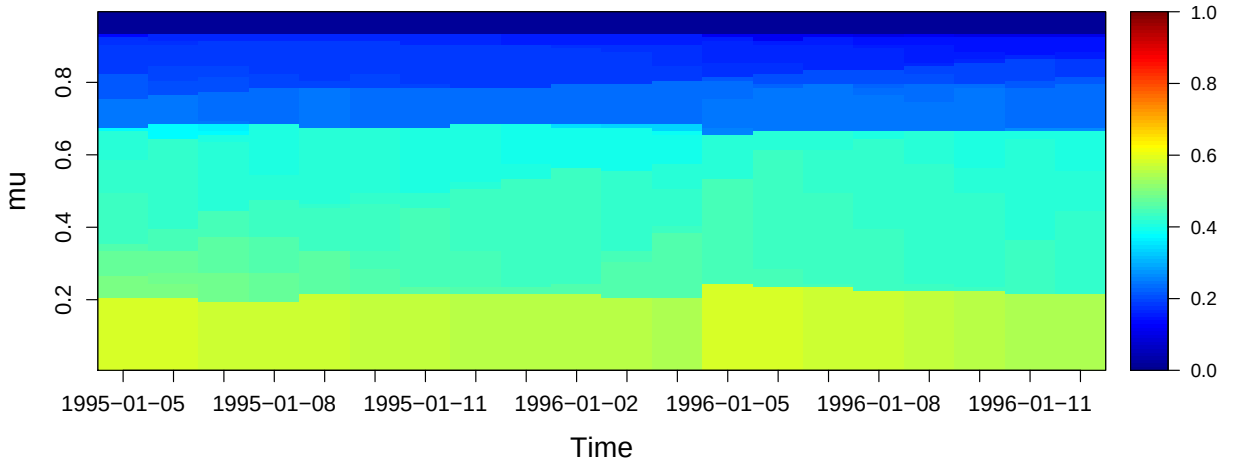

Figure B.5: λ variation in recursive analysis with the BP model.
Note: The color scale refers to λ values for each μ and month. The models are estimated in a recursive manner by using only information available up to each month between 1995:5 and 1996:12.

Acknowledgements

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 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 Seminar, Goethe University Brown Bag Seminar, European Central Bank Financial Stability Seminar, Deutsche Bundesbank Early-Warning Modeling Seminar, and the 2015 CEUSWorkshop on “Recent Issues of European Integration”. This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Peter Sarlin
Department of Economics at Hanken School of Economics; RiskLab Finland at Arcada University of Applied Sciences

Gregor von Schweinitz
Halle Institute for Economic Research (IWH), Department of Macroeconomics; Martin-Luther University Halle-Wittenberg, Chair of Macroeconomics; Deutsche Bundesbank; email: gsz@iwh-halle.de

Postal address	60640 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Website	www.ecb.europa.eu

ISSN	1725-2806 (pdf)	DOI	10.2866/48672 (pdf)
ISBN	978-92-899-2747-5 (pdf)	EU catalogue No	QB-AR-17-037-EN-N (pdf)