

von Schweinitz, Gregor; Sarlin, Peter

**Conference Paper**

## Signaling Crises: How to Get Good Out-of-Sample Performance Out of the Early Warning System

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Sovereign risk revisited: construction and use of early-warning systems, No. A06-V3

**Provided in Cooperation with:**

Verein für Socialpolitik / German Economic Association

*Suggested Citation:* von Schweinitz, Gregor; Sarlin, Peter (2015) : Signaling Crises: How to Get Good Out-of-Sample Performance Out of the Early Warning System, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Sovereign risk revisited: construction and use of early-warning systems, No. A06-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft

This Version is available at:

<https://hdl.handle.net/10419/112964>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Signaling Crises: How to Get Good Out-of-Sample Performance Out of the Early Warning System

Gregor von Schweinitz      Peter Sarlin

This version: March 1, 2015

## Abstract

In past years, the most common approaches for deriving early-warning models belong to the family of binary-choice methods, which have been coupled with a separate loss function to optimize model signals based on policymakers preferences. The evidence in this paper shows that early-warning models should not be used in this traditional way, as the optimization of thresholds produces an in-sample overfit at the expense of out-of-sample performance. Instead of ex-post threshold optimization based upon a loss function, policymakers' preferences should rather be directly included as weights in the estimation function. Doing this strongly improves the out-of-sample performance of early-warning systems.

**Keywords:** Financial Crises; Early-Warning Systems; Binary Choice Models; Out-of-Sample Performance

**JEL-Classification:** C35; C53; G01

# 1 Introduction

Early-warning models provide quantitative means for early identification of vulnerabilities preceding systemic financial crises. The most common approaches for deriving early-warning models descend from the family of binary-choice methods, which have been coupled with a separate loss function based on misclassification costs for false and missing warnings. This paper puts forward joint maximum likelihood optimization of the loss function and standard binary-choice methods by using policymakers' preferences as observation weights in the likelihood function.

The past years of financial turmoil have stimulated research on early-warning analysis, with the result of more mature models and more direct mappings to macroprudential policies. The two dominating approaches for deriving early-warning models consist in the signals approach and logit/probit analysis. The signals approach (a univariate analysis of indicators and their optimal signaling thresholds) descends originally from Kaminsky & Reinhart (1999), but has also been common in the past years (Alessi & Detken 2011, Knedlik & von Schweinitz 2012). Logit/probit analysis was already applied by Frankel & Rose (1996) and Berg & Pattillo (1999) to exchange-rate pressure. More recently, it has been the predominant approach for predicting banking and systemic financial crises (Betz, Oprică, Peltonen & Sarlin 2014, Lo Duca & Peltonen 2013).

An own strand of literature has focused on the explicit forecasting objectives of early-warning systems (EWS) and on loss functions tailored to the preferences of a political decision-maker.<sup>1</sup> Demirgüç-Kunt & Detragiache (2000) introduced the notion of a policymakers' loss-function in the context of banking crises, where the policymaker faces costs for taking unnecessary preventive actions (type 2 errors) and those of an occurring, but unpredicted, crisis (type 1 errors). Later, adaptations of this very general type of loss functions have been introduced to EWSs for other types of crises, e.g. currency crises (Bussiere & Fratzscher 2008), debt crises (Fuentes & Kalotychou 2007, Knedlik & von Schweinitz 2012), and asset price boom/bust cycles (Alessi & Detken 2011). These contributions mostly focus on the trade-off between type 1 and 2 errors, but they also provide usefulness measures that indicate whether and to what extent the loss of the prediction is smaller than the loss of disregarding the model (Sarlin 2013).

Yet, when applying binary-choice methods, common practice has been an ex-post minimization of the loss function, or, equivalently, maximization of a usefulness function which is based on the loss function. This paper postulates that early-warning models based upon binary-choice methods should account for policymakers' preferences directly as part of the maximum likelihood estimation rather than applying an ex-post optimization of a loss function as a second step. We do this by introducing observation weights (reflecting the preferences) in the likelihood function of the binary choice model.<sup>2</sup> One-step maximization as suggested by our paper strongly improves out-of-sample performance of the model and reduces the positive bias of in-sample predictive power. Therefore, our method provides a much more reliable early-warning model than the traditional estimation. Additionally, while the maxi-

---

<sup>1</sup>The literature on early-warning models has used a wide range of measures for evaluating performance. We do not herein summarize measures focusing on model robustness, such as the Receiver Operating Characteristics curve and the area below it, as they do not provide guidance on the choice of a threshold.

<sup>2</sup>The estimation routine (in R) can be obtained from the authors on request.

maximum likelihood estimation is nearly identical, the second optimization step of the traditional approach is left out, making our approach simpler overall.

We provide two-fold evidence for our claim concerning the quality of the early-warning system. First, we run simulations with different data generating processes to illustrate the superiority of weighted maximum likelihood estimation vis-a-vis ex-post optimization of thresholds on data with known patterns. Second, we make use of a real-world case to illustrate both in-sample and out-of-sample performance of the two approaches. We replicate the early-warning model for currency crises in Berg & Pattillo (1999).

As our critique and suggested solution applies to the failures of loss/usefulness function optimization, it holds for every early-warning system with this feature. For methods built on an initial (maximum likelihood) estimation of event probabilities, our proposed solution is directly transferable. However, our critique specifically extends to the signals approach as well, which consists solely of the optimization step. That is, in the context of the signals approach a weighted estimation is not possible. This leaves the practitioner with two different solutions to the problems presented in this paper: First, the signals approach can be replaced altogether by equivalent univariate weighted probit models (following the route taken here). Alternatively, the signals approach can be enhanced by confidence measures as proposed by El-Shagi, Knedlik & von Schweinitz (2013). While the first solution tackles the positive in-sample bias introduced by threshold optimization, the second one aims at measuring it. The paper is structured as follows. The next section presents the methods, followed by a discussion of our experiments in the third section and a conclusion.

## 2 Estimating and evaluating early-warning models

This section presents the methods analyzed in this paper. It starts with a description of the usual estimation and evaluation of (binary-choice) early-warning models, followed by an introduction to the combination of these two steps in a single, simple maximum-likelihood estimation.

### 2.1 Estimating early-warning models

The literature on early-warning models has used a range of conventional statistical methods for estimating distress probabilities. Most common approaches rely on logit/probit analyses, although model specifications and estimation strategies have varied to some extent. We follow herein the approach based on a standard pooled probit model (Kumar, Moorthy & Perraudin 2003, Fuertes & Kalotychou 2007, Davis & Karim 2008).

The occurrence of an event of interest is represented by a binary state variable  $I_j(h) \in \{0, 1\}$ , where the index  $j = 1, 2, \dots, N$  represents instances and  $h$  is a specified forecast horizon. The state variable  $I_j(h)$  is set to 1 if an event (mostly a crisis) happens sometime in the next  $h$  periods. In the standard binary choice model, it is assumed that  $I_j(h)$  is driven by a

latent variable

$$y_j^* = X_j\beta + \varepsilon$$

$$I_j(h) = \begin{cases} 1 & , \text{if } y_j^* > 0 \\ 0 & , \text{otherwise} \end{cases}.$$

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

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

which is maximized with respect to  $\beta$ . 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. This binomial model (probit or logit) returns probability forecasts  $p_j \in [0, 1]$  of the occurrence of the event.

## 2.2 Evaluating early-warning models

Luckily, crises are scarce. However, this poses a serious problem in the early-warning literature, where the estimated probability of a crisis seldom exceeds 50% (which would be an intuitive threshold for a binary choice model). Therefore, practitioners have to determine which probability should be used as a threshold above which the early warning system issues a warning, eventually leading to precautionary action to prevent a crisis from happening. Introducing a probability threshold transforms the event probabilities into binary signals. The evaluation of these signals follows the methodology of the signals approach. The framework applied here follows that in El-Shagi et al. (2013) and Sarlin (2013). As they do, we derive a loss and usefulness function for a cost-aware decision maker with class-specific misclassification costs, where the classes depend on the instances  $I_j(h)$ .

To mimic the state variable  $I_j(h)$ , the probabilities  $p_j$  need to be transformed into binary point forecasts  $P_j \in \{0, 1\}$  that equal one if  $p_j$  exceeds a specified threshold  $\lambda$  and zero otherwise. The correspondence between  $P_j$  and  $I_j$  can be summarized by a so-called contingency matrix (frequencies of prediction-realization combinations): false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN).<sup>3</sup> The sum of true positives and false negatives is the number of instances where the state variable is equal to 1 (all positives), while the sum of false positives and true negatives is just the number of instances where the state variable is equal to zero (all negatives).

While entries of a contingency matrix can be used to define a large palette of goodness-of-fit measures, such as overall accuracy, we approach the problem from the viewpoint of a decision maker that is wary of conducting two types of errors. Type 1 errors – a missed event – represent the conditional probability  $P(p_j \leq \lambda | I_j(h) = 1)$ . This conditional probability is estimated from data as the share of false negatives to all positives ( $T_1 = FN / (FN + TP)$ ). Similarly, type 2 errors – a falsely predicted event – represent the conditional probability

---

<sup>3</sup>Kaminsky & Reinhart (1999) use a matrix of these four states, denoting them by B, A, C and D.

$P(p_j > \lambda | I_j(h) = 0)$ , and are estimated as the proportion of false positives to all negatives ( $T_2 = FP / (FP + TN)$ ). Given probabilities  $p_j$  of a model, the decision maker should focus on choosing a threshold  $\lambda$  such that her loss is minimized. To account for imbalances in class size in the loss function, the share of errors  $T_1$  and  $T_2$  has to be weighted by unconditional probabilities of positives  $P_1 = P(I_j(h) = 1)$  and negatives  $P_2 = P(I_j(h) = 0) = 1 - P_1$ . Frequency-weighted errors are then further weighted by policymakers' relative preferences between FNs ( $\mu \in [0, 1]$ ) and FPs ( $1 - \mu$ ). Finally, the loss function is as follows:

$$L(\mu) = \mu T_1 P_1 + (1 - \mu) T_2 P_2.$$

The specification of the loss function  $L(\mu)$  enables computing the usefulness (sometimes also called utility) of a model. A decision maker could achieve a loss of  $\min(P_1, P_2)$  by always issuing a signal of a crisis if  $P_1 > 0.5$  or never issuing a signal if  $P_2 > 0.5$ . When also paying regard to the policymakers' preferences between errors, the decision maker achieves a loss  $\min(\mu P_1, (1 - \mu) P_2)$  when ignoring the model. The usefulness  $U_a$  of a model is computed by subtracting the loss generated by the model from the loss of ignoring it:

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

This measure highlights the fact that achieving beneficial models on highly imbalanced data is challenging as a non-perfectly performing model is easily worse than always signaling the more frequent class. Hence, already an attempt to build a predictive model with imbalanced data implicitly requires a decision maker to be more concerned about the rare class.

As a third measure, relative usefulness computes the percentage of absolute usefulness  $U_a$  to a model's available usefulness  $\min(\mu P_1, (1 - \mu) P_2)$ :

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

The relative usefulness  $U_r$  computes absolute usefulness  $U_a$  as a share of the usefulness that a decision maker would gain with a perfectly performing model. Hence,  $U_r$  is nothing more than a rescaled measure of  $U_a$ . Yet, the  $U_r$  provides means for better assessment of usefulness by extracting a number with a meaningful interpretation; performance can be compared in terms of percentage points. When interpreting models, we can hence focus solely on  $U_r$ .

In the following, we will focus on the overfit created by usefulness maximization. This tendency (or rather, characteristic) of the EWS evaluation and optimization is independent of the exact definition of the employed loss and usefulness function. That is, our results are robust to many different specifications. However, we choose this specific formulation of the loss and usefulness function, because the preferences  $\mu$  are de facto applied to false positives and negatives as shares of the total number of instances. That is,  $\mu$  and  $1 - \mu$  are actually observation-specific weights.

### 2.3 Estimating and evaluating the early-warning system jointly

As described in the beginning of subsection 2.2, one would intuitively see an event as likely if the estimated probability exceeds 50%. Transforming probability forecasts into binary signals by the use of a threshold  $\lambda$  changes this intuitive threshold. The reason to do this is

because policy makers need to account for the different costs of crisis prevention and crisis occurrence. These costs, however, can also be introduced in the form of weights into the likelihood of the binary choice model.<sup>4</sup> For the weighted probit model, the log-likelihood function is the following:

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

If the usefulness measure is defined as in the previous section, the number of type 1 and 2 errors are weighted by  $\mu$  and  $(1 - \mu)$  respectively. That is, setting observation weights  $w = 1 - \mu$  is equivalent to the definition of usefulness given above.<sup>5</sup>

This function can be maximized just as easily as the standard binary choice model. Compared to the standard model before threshold optimization, it will result in slightly shifted probability forecasts, with the direction of the shift depending on  $\mu$ . The appealing feature of the weighted binary choice model is that finding a probability threshold that optimizes a usefulness measure is not necessary anymore. Instead, the intuitive threshold of 50% from the weighted model already accounts for all policy preferences captured in  $\mu$ . Therefore, the second and – as we shall see – extremely problematic step of the traditional construction of an EWS becomes unnecessary.

### 3 Comparing the two models

In this section, we compare the use of ex-post threshold optimization in early-warning models vis-a-vis direct use of a loss function when optimizing likelihoods. We find strong evidence favoring the weighted binary choice model. To illustrate differences among the approaches, we provide a large number of experiments on simulated data. Then, we compare the approaches on real-world data using the well-known early-warning model for currency crises in Berg & Pattillo (1999).

#### 3.1 Simulated data

Before testing our approach with real data, we apply it to simulated, simple datasets. We use three explanatory variables  $X = (X_1, X_2, X_3)$ , a constant and a coefficient vector  $\beta = (1, 0, 0, -1)$ . That is, in the true model only  $X_1$  contains information on the latent variable  $y^*$  and therefore the observable event. The constant (with negative coefficient) is chosen such that the probability of an event is slightly below 25%. For a normal application of an early-warning model, this would be quite a lot of events, although not unusually many.

We draw the explanatory variables independently from a standard normal distribution. We simulate  $N = 100, 1'000, 10'000$  datapoints, calculate the probability  $\Phi(X\beta)$  of an event and draw  $I(h)$  from these probabilities (abstracting from index  $j$ ).

---

<sup>4</sup>Observation-specific weights have already been introduced in binary-choice models to adjust for non-representativeness of an estimation sample in cases where an average effect for the whole population is of interest.

<sup>5</sup>In case of other definitions of usefulness, equivalent observation weights can mostly be obtained by simple algebraic transformations.

We then estimate four different econometric models from the first half of our simulated dataset ( $N = 50, 500, 5'000$ ): a standard probit, standard logit and their weighted counterparts with weights  $w = 1 - \mu$ . The logit estimations are performed as they are by far the most simplest way to test if the results are robust against an admittedly very mild form of misspecification. We use the parameter estimates from the first half of the data together with the second half of the data to construct out-of-sample event probabilities. For the standard models, we search for the optimal probability threshold  $\lambda$  given the policy preference  $\mu$ , while we use the natural threshold of 50% for weighted models. These thresholds are used to calculate in-sample and out-of-sample absolute and relative utilities.

The above steps are performed for three different preference settings for  $\mu$ : 0.2, which gives strong preference to avoiding type-2 errors. In practice, such a preference would be chosen to account for the (assumedly) higher frequency of type-2 errors, as non-events are more frequent. 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). 0.8 gives a strong preference to avoiding type-1 errors. This preference would be used in practice to account for the fact that missing a crisis may be very costly.

Simulating every model 1'000 times gives reasonable estimates of the mean and standard deviation of  $\lambda$  as well as absolute and relative utilities. Furthermore, we can calculate the probability that the in- and out-of-sample usefulness from the standard econometric model is lower than the one from the corresponding weighted model. 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 similar results.

### 3.1.1 Randomness of utility

First, let us take a look at a specification (different from above), where events have no relation to explanatory variables (that is,  $\beta = (0, 0, 0, 0)$ , and the event probability is 50% in every period). Figure 1 shows the in-sample Receiver-Operator-Characteristics (ROC) curves from a probit model for three simulations with different  $N$ . 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 (L1) distance to the red diagonal.

Ideally, the distance (and therefore absolute usefulness) should be zero, because there is no relation between  $X$  and  $I(h)$  in this specification. However, in practice this is not the case. For small  $N$ ,  $\beta$  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 stronger ROC-movements.<sup>6</sup> 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.

---

<sup>6</sup>This argument is very much related to El-Shagi et al. (2013), who argue that – in order to judge the quality of an early warning system – it is paramount to obtain a distribution of the usefulness under the null hypothesis of no relation between  $X$  and  $I(h)$ , instead of only a measure of usefulness itself.



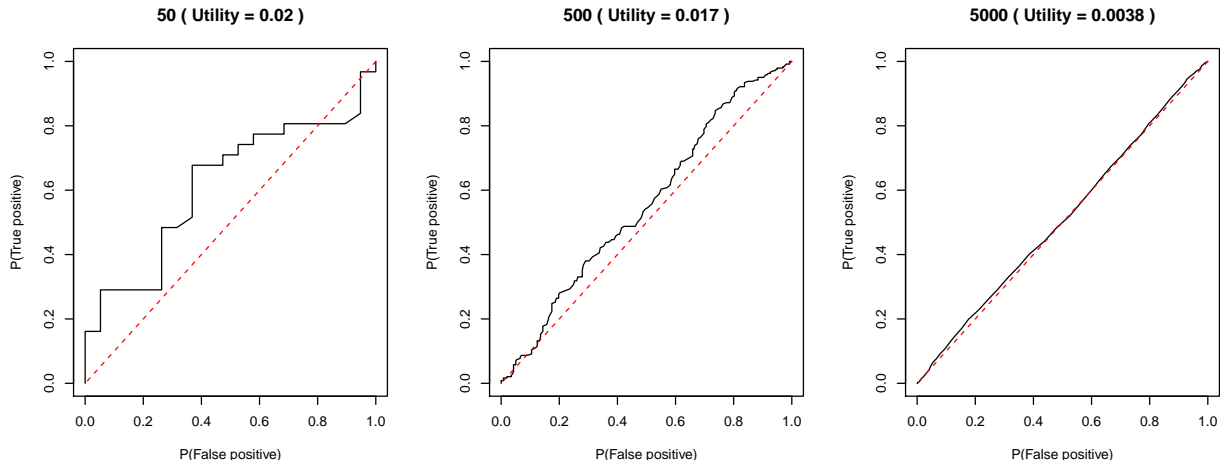


Figure 1: ROC-curve for three simulations with random events ( $N=50, 500, 5'000$ ) from the probit estimation.

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

As can be seen, this distance, and therefore usefulness of the random model, decreases strongly with increasing  $N$ . This happens because, as  $N$  increases, uncertainty on true coefficients decreases, bringing the ROC-curve closer to the diagonal and bringing usefulness closer towards its true level of zero.

### 3.1.2 Is the optimal probability threshold in the end only determined by policy preferences?

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 2 presents the mean plus/minus one standard deviation of optimal  $\lambda$  for the different policy preferences  $\mu$  and different number of observations  $N$ .

As the true data generating process is always identical, all uncertainty on  $\lambda$  comes from the number of observations. Therefore, it is quite natural that the standard deviation of  $\lambda$  does not depend on the preferences  $\mu$  and decreases with  $N$ . The much more interesting feature, however, is that  $\lambda$  approaches  $1 - \mu$  for larger  $N$ . On second thought, this is again quite logical: As  $N$  goes to infinity, the ROC-curve gets smoother and approaches the “ideal” form that is solely determined by the underlying data generating process. Under this “ideal” form, we get a uniform distribution of type-1 and 2 errors, that is,  $P(p_j \leq \lambda | I_j(h) = 1) = \lambda$ .

Figure 2 depicts another frequently found result: the difference between probit and logit estimations is marginal. If anything, the  $\lambda$  obtained from logit estimations seems to approach the true  $\lambda$  faster – even though the logit model is misspecified.

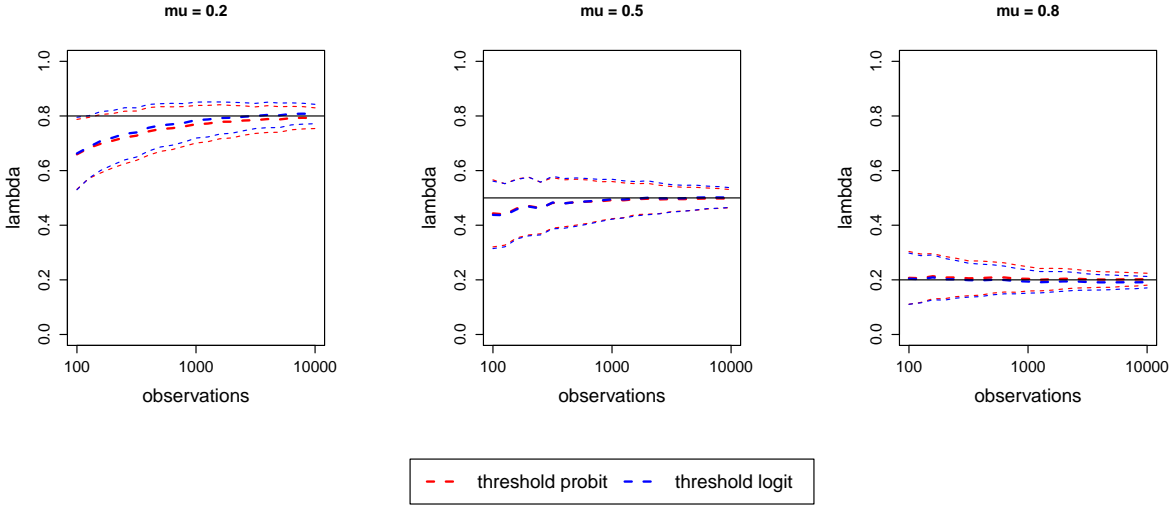


Figure 2: Development of mean  $\lambda$  plus/minus one standard deviations as  $N$  increases, for different values of  $\mu$ .

### 3.1.3 Differences in usefulness: is it really optimal to optimize usefulness?

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 data generating process, 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 3. This figure equally confirms that out-of-sample usefulness (the lower 4 curves in every plot), which essentially depend on a correct assessment of the properties of the data-generating process, improves as  $N$  goes to infinity.

In addition to the general results holding for all four estimation methods, we see that the usefulness (in- and out-of-sample) of the weighted methods is on average closer to their true value than those of the threshold methods. 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 system. If this quality is biased upwards (as it usually is), it induces an overstated sense of confidence, trust and security. This bias is much lower for weighted methods, where it only stems from estimation uncertainty. However, what really matters is out-of-sample usefulness. Here, weighted models perform better on average than their threshold peers. This holds especially in the case of the logit model: the results of the (misspecified) weighted logit are nearly identical to the ones of the weighted probit, while out-of-sample usefulness of the threshold logit is far below the one for threshold probit, when  $\mu$  is different from 0.5. That is, in addition to being on average better out-of-sample than their peers, weighted methods provide robustness against

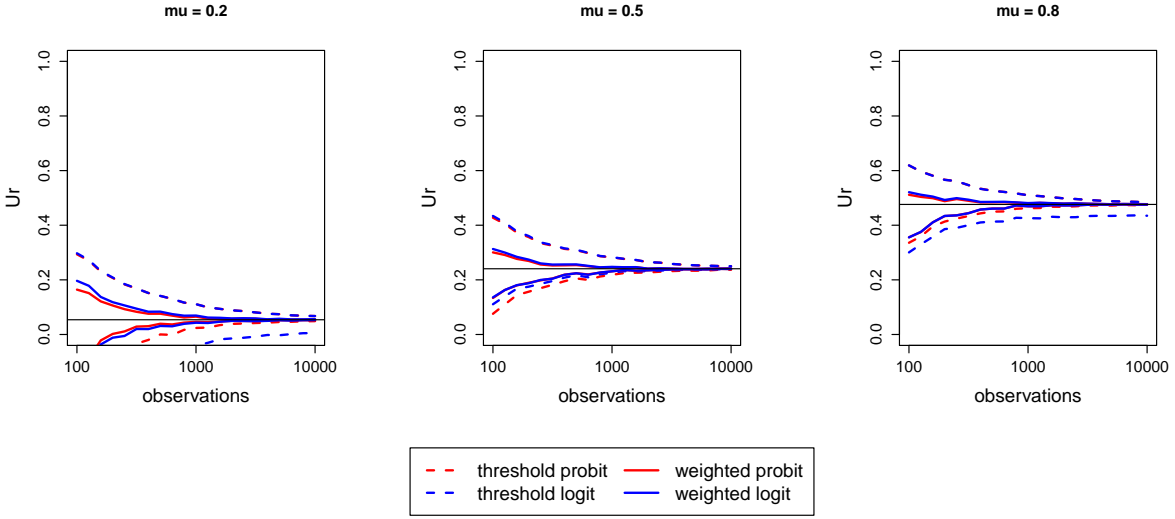


Figure 3: Mean relative usefulness for threshold estimation and weighted estimations. *Note:* In-sample usefulness is higher than out-of-sample usefulness for every number of observations  $N$ . The black line signifies the limiting usefulness as  $N$  goes to infinity. Usefulness below zero is (mostly) not displayed.

method misspecification. Altogether, the argument for weighted models is even stronger out-of-sample (which matters significantly more in practice) than in-sample.

### 3.1.4 Probability of out-performance by weighted models

Figure 3 shows that out-of-sample usefulness of weighted model is on average be better than out-of-sample usefulness of threshold models. The question now is: are weighted models so much and so often better that we should abandon threshold optimization and use weighted binary choice models instead? Here, the result presented in figure 4 is not as clear-cut as above. In-sample, weighted models produce nearly always worse usefulness than their threshold peers. This should be desired if in-sample usefulness is biased upwards. Out-of-sample, weighted models have a slight advantage. For the probit model (which is the true econometric model in this case), the advantage of the weighted model is minor, although it seems to be slightly growing as  $N$  increases. However, as seen in subsection 3.1.3, the precision of estimates and the fit of the model increases for growing  $N$ , which may make such small difference mostly irrelevant in practice. For the logit model and  $\mu \neq 0.5$  (i.e., the misspecified model where weights actually play a role), the probability of a better early-warning system approaches 100%. This result reflects again the result on average usefulness from subsection 3.1.3.

### 3.1.5 Robustness to other specification

Above, we reported only results for a very simple specification where no estimation problems are to be expected. This may change if the data generating process gets more complicated. Specifically, it could well be that estimation of the slightly more complicated weighted mod-

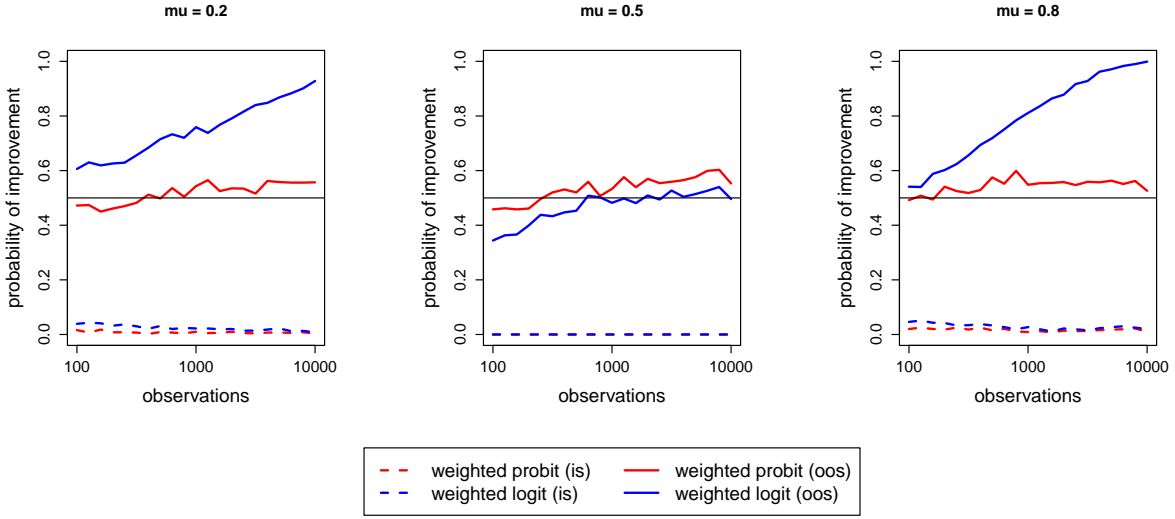


Figure 4: Probability that the weighted model has a higher usefulness than the threshold model.

*Note:* In-sample usefulness is higher than out-of-sample usefulness for every number of observations  $N$ . The black line signifies the limiting usefulness as  $N$  goes to infinity. Usefulness below zero is (mostly) not displayed.

els suffers, if the complexity of the data generating process is increased. Therefore, we tested many different specifications. The only constant in this was that we kept the number of exogenous variables at three, and that the true  $\beta$  remained  $(1, 0, 0, -1)$  (including the constant). 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. Furthermore, an early warning system with non-correlated exogenous variables is virtually non-existent in practice.
2. Autocorrelation of all exogenous variables with lag coefficients 0.7 (first lag) and  $-0.3$  (second lag). Autocorrelation is highly relevant for macroeconomic variables that are usually used in early-warning systems.
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 3.1.1.
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  $X_2$  and  $X_3$  is not completely random. We would therefore expect results close to the correlated model.

In short, the results are nearly identical for different models. That is, our baseline results hold fully for the full battery of different model specifications.<sup>7</sup>

### 3.2 The currency crisis model of Berg and Pattillo (1999)

This section turns the attention to indicators commonly used to describe and explain the vulnerabilities to a currency crisis. These indicators were originally introduced to crisis monitoring in a predictive model by Berg & Pattillo (1999). 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 an exchange market pressure index as defined by Berg & Pattillo (1999). A crisis occurs if the sum of a weighted average of monthly percentage depreciation in the currency and monthly percentage declines in reserves exceeds its mean by more than three standard deviations. Using the crisis occurrences, we define an observation to be in a vulnerability state, or pre-crisis period, if it experienced a crisis within the following 24 months.

The data is divided in a training sample for in-sample fitting from 1986:1 to 1995:4, and a test sample 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. This shows that there may have been a structural difference between the training and the test sample which could be reflected in in- and out-of-sample usefulness.

In our estimations, we set  $\mu$  to the share of instances that are not vulnerable to a crisis (0.832). That is, we choose preferences that give higher weight to the more frequent calm periods. If we optimize the early warning system first on the full sample (instead of differing between training and test sample), relative usefulness is for all models between 40.1% and 42.5%, with threshold models having a small advantage. The difference is minor and occurs mainly, because weighted models issue fewer false negatives at the expense of many more false positives.

Dividing the sample in training and test part, the difference between models is dramatic. In-sample usefulness of threshold models is above 80%, while weighted models achieve only around 35%. The higher usefulness of threshold models may be explained by increased uncertainty and therefore “room to optimize”, while the slightly lower usefulness of weighted models might just be random. The picture reverses completely for out-of-sample usefulness. Here, the absolute usefulness of threshold models is strongly negative. That is, as a policymaker it would have been better to disregard the model altogether. This result is shocking insofar as the in-sample fit provides overwhelming support for the fitted early warning models in the threshold case. For weighted models, however, relative usefulness remains around 20%. Despite dropping from 35%, this is still relatively good. That is, weighted early-warning models may be suitable – to a certain extent – to counter the very general concern on out-of-sample performance voiced by Rose & Spiegel (2012).

The real dataset strengthens our simulation results. The traditional way of estimating and

---

<sup>7</sup>Detailed results can be obtained from the authors on request.

evaluating binary choice models by applying a threshold to probability forecasts introduces an overfit with strongly negative consequences for out-of-sample capabilities. As hinted above, the difference 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 data-generating process, it will make an estimation of the true process harder with limited datasets. The resulting uncertainty, in turn, influences threshold models much more negatively than weighted models. In practice, it is very likely that different crises have different origins<sup>8</sup>. That is, the importance of explanatory variables will most definitely change over time. Therefore, our example with real data provides suggestive evidence that early warning models relying on weighted binary choice models may be far more robust to these changes than their traditional counterparts.

## 4 Conclusion

To subsume, we find that early-warning systems where preferences for certain types of errors are included as weights in the estimation outperform their traditional peers (which account for preferences only in a second optimization step) in two ways. First, the bias of in-sample utility is much lower, reducing the false degree of confidence in out-of-sample capabilities of the early-warning system. Second, out-of-sample performance is higher, especially in cases where the econometric model is misspecified or where estimation uncertainty is high. This is because weighted models estimate the *most likely* instead of the *implausible best possible* outcome, and do therefore properly account for estimation uncertainty. We think therefore that weighted models are preferable.

As our results hold not only for the simple binary choice models tested in this paper, but for every early-warning system using threshold optimization (including the much-used signals approach), we strongly recommend to include policymakers' preferences as weights in the estimated likelihood and 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.

---

<sup>8</sup>If different crises had identical origins, this would indicate strongly that economists, policymakers and market participants would be unable to learn from the past.

- 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.
- 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.
- 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.
- 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.
- Rose, A. K. & Spiegel, M. M. (2012). Cross-Country Causes and Consequences of the 2008 Crisis: Early Warning, *Japan and the World Economy* **24**(1): 1–16.
- Sarlin, P. (2013). On Policymakers' Loss Functions and the Evaluation of Early Warning Systems, *Economics Letters* **119**(1): 1–7.