XIX

Volume XIX, Number 2, November 2016

Journal of Applied Economics

Makram El-Shagi
Gregor von Schweinitz

Qual VAR revisited: Good forecast, bad story

UCEMA

Edited by the Universidad del CEMA
Print ISSN 1514-0326
Online ISSN 1667-6726

Journal of Applied Economics. Vol XIX, No. 2 (November 2016), 293-322

QUAL VAR REVISITED: GOOD FORECAST, BAD STORY

Makram El-Shagi[*]
Henan University; Halle Institute for Economic Research (IWH)

Gregor von Schweinitz
Halle Institute for Economic Research (IWH);
Martin-Luther-University Halle-Wittenberg

Submitted October 2014; accepted August 2015

Due to the recent financial crisis, the interest in econometric models that allow to incorporate binary variables (such as the occurrence of a crisis) experienced a huge surge. This paper evaluates the performance of the Qual VAR, originally proposed by Dueker (2005). The Qual VAR is a VAR model including a latent variable that governs the behavior of an observable binary variable. While we find that the Qual VAR performs reasonable well in forecasting (outperforming a probit benchmark), there are substantial identification problems even in a simple VAR specification. Typically, identification in economic applications is far more difficult than in our simple benchmark. Therefore, when the economic interpretation of the dynamic behavior of the latent variable and the chain of causality matter, use of the Qual VAR is inadvisable.

JEL classification codes: C15, C35, E37
Key words: binary choice model, Gibbs sampling, latent variable, MCMC, method evaluation

[*] Makram El-Shagi (corresponding author): Henan University, Kaifeng, School of Economics and Halle Institute for Economic Research, Member of the Leibniz Association; makram.elshagi@gmail.com. Gregor von Schweinitz: Halle Institute for Economic Research, Member of the Leibniz Association and Martin-Luther-University Halle-Wittenberg; gregorvon.schweinitz@iwh-halle.de. The authors are highly indebted to two anonymous referees, Michael Berlemann, Herbert Buscher, Peter Tillmann and Jonathan Wright for valuable comments and suggestions as well as participants of the INFER annual conference and the faculty seminar at the Helmut-Schmidt-University Hamburg.

I. Introduction

Due to the recent financial crisis, the interest in econometric models that allow incorporating binary variables (such as the occurrence of a crisis) experienced a huge surge. This paper evaluates the performance of the Qual VAR, a VAR (vector autoregression) model including a latent variable governing the behavior of an observable binary variable that was originally proposed by Dueker (2005).

Harding and Pagan (2011) criticize that in macroeconomics a binary variable is often constructed from observable variables that show some persistence (they call it a *secondary* binary variable). They claim that the conventional binary choice model cannot capture the time series dependence of the binary variable inherent in current macroeconomic applications. Therefore, Harding and Pagan recommend using other methods incorporating time series dependence such as the Markov Switching approach. These models have originally been developed for the dating of business cycle turning points (Hamilton 1989; Paap et al. 2009). Like binary choice models, they can also be used for crisis prediction (Fratzscher 2003; Hartmann et al. 2012). Because Markov Switching models estimate the regimes endogenously, they cannot be applied to predefined binary variables, such as the NBER recessions, IMF interventions and the like. Essentially, while being able to incorporate time series dependence in the binary variable, there is no clear economic definition of what is really meant by the different regimes (El-Shagi, Knedlik and von Schweinitz 2013).

The Qual VAR and some other recent methods may overcome the Harding and Pagan critique and the identification problem of Markov Switching models (Kauppi and Saikkonen 2008; Dueker 2005). Especially the Qual VAR proposed by Dueker (2005) is economically appealing. In the Qual VAR, a latent variable, driving an observable binary variable, and a number of other observables (jointly) follow a VAR process.[1] Estimating a VAR process instead of a single equation, thereby exploiting more information, leads to efficiency gains in the identification of the latent variable. Moreover, since the latent variable can be interpreted as a risk indicator for the event desribed by the binary variable, the VAR structure allows to capture the feedback of the corresponding risk into the economy. The

[1] Thus, the Qual VAR is essentially an extension of the dynamic ordered probit of Eichengreen et al. (1985), as observed by Marcellino (2006).

importance of such an interaction between observable and latent variable can for example be observed in the current European debt crisis, where the risk of sovereign default strongly affects government bond interest rates and vice-versa.

A number of recent papers have used the Qual VAR. Bordo et al. (2008) apply the model to bull and bear periods on the stock market, Dueker and Assenmacher-Wesche (2010) assess the recursive forecasting performance of the Qual VAR. This performance is also tested in comparison to other models by Galvão (2006) and Fornari and Lemke (2010). However, the present literature concentrates exclusively on the forecasting performance of the Qual VAR. Furthermore, this evaluation is done based on one specific economic example (i.e., forecasting the 2001 U.S. recession) rather than being performed in a more general framework. It is therefore unclear whether the results are applicable universally. Additionally, while the Qual VAR has been developed for forecasting, the estimates have also been interpreted economically, e.g., by analyzing impulse response functions derived from coefficient estimates and considering the latent variable as in sample measure of event probability (Dueker 2005). However, even if the forecasting performance was generally high, this is not sufficient to allow a structural interpretation of the results.

Our paper aims at closing these gaps in the literature by providing a range of Monte Carlo studies. This allows a more general examination of the Qual VAR by considering its performance in idealized settings where the data generating process (DGP) and the latent variable are known. We test the performance of the original Qual VAR as published in the JBES by Dueker (2005), rather than more recent extensions by Bordo et al. (2008) and Dueker and Nelson (2006), since the first is by far the most prominent paper and — to the best of our knowledge — even today the only one for which an estimation routine in a widely used econometric package is freely available.[2]

First, we assess the in-sample estimation of the latent variable. Second, we analyze whether or not the Qual VAR identifies the true Granger causality between the latent and other variables. Third, we test the forecasting performance of the estimated system. All these tests are performed for a variety of VAR-specifications covering different chains of causality and uncertainty levels. Using

[2] The available routine is in RATS. The RATS routine has not been used in this paper since it does not allow adaptation. All estimations in this paper have been performed in Matlab. The code is available on request.

different specifications ensures that the results are relevant for typical empirical applications with similar underlying DGPs. While we find that the Qual VAR performs reasonable well in forecasting (in several respects outperforming different benchmark estimations), there are substantial identification problems even in a simple VAR specification. These identification problems are found both in one lag, two variable VARs and more complex specifications including more lags and variables. Typically, identification in economic applications is far more difficult than in our settings. Therefore, when the economic interpretation of the dynamic behavior of the latent variable and the chain of causality matter, the use of the Qual VAR is inadvisable.

The remainder of the paper is organized as follows. In Section II we present the estimation technique of the Qual VAR. Section III describes the set of Monte Carlo studies used to obtain our results. In Section IV we discuss identification issues; Section V contains the results of our tests of the forecasting performance of the Qual VAR in our benchmark settings. Section VI shows that these results mostly hold in more complex settings, and Section VII concludes.

II. Estimation of a Qual VAR

The Qual VAR (Dueker 2005) has been developed as a method for forecasting qualitative variables. Originally, it has been applied to the prediction of recessions and business cycle turning points. It assumes that the present state in the qualitative (usually binary) variable y_t is the observable manifestation of a latent variable y_t^*:

$$y_t = \begin{cases} 0, if\ y_t^* \leq 0 \\ 1, if\ y_t^* > 0 \end{cases}. \tag{1}$$

The unobservable variable y^* and $k-1$ other observable variables X are said to follow a VAR(p) process:

$$Y_t = \mu + L(\Phi)Y_{t-1} + \varepsilon_t, \tag{2}$$

where $Y_t = (X_t\ y_t^*)$, $t = 1, \dots, T$ is the time index, $L(\cdot)$ is the lag operator, Φ the corresponding coefficient matrix, μ the constant vector and ε_t the error vector at t. Errors are assumed to be multivariate-normally distributed with mean zero and covariance matrix Σ. The covariance matrix, the parameters of the VAR and the unobservable variable $y^*$are jointly estimated using a Gibbs sampler.

The Gibbs sampler is used to simulate the joint distribution of $\Lambda = (\Phi, \Sigma, y^*)$. In each iteration i, a value for each element of Λ is randomly drawn from its distribution conditional on the last generated values of all other elements. Depending on the ordering of the elements in the Gibbs sampler, the last generated value can either come from the current or the previous iteration. In the first iteration, starting values for the latent variable are randomly generated (based on the knowledge of the binary variable). Φ and Σ can then be estimated by OLS and used as initial values.

As in standard ML estimation, the coefficients Φ are assumed to be multivariate-normally distributed and the inverse of the covariance matrix of errors Φ^{-1} is assumed to be Wishart distributed. Each y_t^* is drawn from a truncated normal distribution, where the truncation is determined by the observable binary variable .

The order of drawings we use (following Dueker) is $\Phi \rightarrow \Sigma \rightarrow y^*$. That is, we draw $\Phi^{(i+1)}|Y^{(i)}, \Sigma^{(i)}$ and $\Sigma^{(i+1)}|\Phi^{(i+1)}, Y^{(i)}$. The vector y^* is sampled element by element, where the distribution of the respective element is conditional on the past of the time series in the current iteration and the future of the time series in the last iteration; i.e., we draw $\left(y_t^{*(i+1)} | \Phi^{(i+1)}, \left\{y_k^{*(i+1)}\right\}_{k<t}, \{y_k^{*(i)}\}_{k>t}, X_t, \Sigma^{(i+1)}\right)$. If $p < t < T - p$, $y_t^{*(i+1)}$, is drawn from the exact conditional distribution.[3] It is not feasible to compute these for $t \leq p$ and $t \geq T - p$. We follow Dueker (2005), who proposes a Metropolis-Hastings algorithm for the first p periods. For the last p periods, we use simple VAR forecasts to compute the mean of the distribution of $y_{T-p+1}^{*(i+1)}$ and subsequently draw errors from a truncated normal satisfying the conditions imposed by the observable binary variable. Based on $y_{T-p+1}^{*(i+1)}$, we can compute the value of the next period accordingly.

Dueker originally proposed to run the Gibbs sampler once with 10,000 iterations and discard the first 5,000. In contrast, we only use every fifth of the remaining 5,000 iterations to avoid artifacts caused by the dependencies between consecutive iterations in the sampled distributions (Casella and George 1992).[4]

From the resulting sample of 1,000 iterations, we calculate median variables Λ_{med}, confidence bands and a set of Fry-Pagan estimate variables Λ_{FP}. Median and confidence bands are calculated for every element in Φ and Σ and y^* separately. However, in the spirit of the Fry-Pagan critique,[5] the set of Fry-Pagan estimates is

[3] p is, again, the lag order of the VAR.

[4] We deviate from this rule in our forecasting tests as described in Table 2, thereby reducing otherwise exploding runtimes. Estimations show no great difference compared to the other tests.

[5] Fry and Pagan (2007) criticize the use of the pointwise median to construct impulse response functions in SVAR.

a consistent set of elements of Λ. Therefore, we select the iteration with the highest joint log likelihood as the Fry-Pagan estimate. The likelihoods are computed assuming multivariate normal distributions for Φ and normal distributions for each element of y^*, where mean and variance are drawn from the distribution obtained from the Gibbs sampler.[6]

III. Setup of the Markov Chain Monte Carlo (MCMC) simulation

Our Monte Carlo study aims to test the capability of Qual VAR to identify the latent variable, to capture the correct chain of causality and to forecast the event driven by the latent variable. To robustly do so, we must test a range of setups, covering the most important features of the data-generating process (DGP) that might affect these issues. The first important feature of the DGP is the variance of the error term in the equations governing the behavior of the observable variable(s).[7] The variance strongly affects the degree of determination in the system and, thereby, both identification and potential forecasting performances. The second feature of the DGP that we account for is the chain of causality between observable(s) and the latent variable. While this is obviously essential to answer the question whether different chains of causality can be distinguished by Qual VAR estimation, it may also affect identification. Because identification in the Qual VAR exploits both lags and leads, a causality running both directions potentially simplifies correct identification of the latent variable.

We aim to cover all of these aspects in the simplest framework possible. Therefore, the true DGP in most of our simulations uses one observable and the latent variable ($k = 2$) and one lag ($p = 1$). This strongly reduces multicollinearity issues that arise in more complex systems of interacting variables, which may cause trouble in identification. At the same time, such a simple DGP is still capable of capturing a range of potential chains of causality and different degrees of uncertainty. In our robustness section we analyze further DGPs with features frequently found in economic data that cannot be hosted by a DGP with one lag

They argue that while the median is usually the most likely outcome at any point in time, the sequence obtained from the pointwise median values is not itself necessarily a consistent impulse response. This critique also applies in the current case, if one needs to analyze the estimated latent variable itself and not only its distribution.

[6] We only use Φ and y for this calculation as the variance of the error of the latent variable is always set to one, making the calculation of a density of the inverse Wishart distribution for Σ impossible.

[7] As the scaling of the latent variable is arbitrary, the error variance in its equation is conventionally scaled to one in estimation. Therefore, we do the same in the true process in all of our MCMC experiments.

and one observable variable. The results suggest that our key findings also hold in these more complex settings.

To avoid confusing a lack of power with model uncertainty, the "true" DGP in our simulations exactly mirrors the assumptions of the Qual VAR. That is, the event occurs if and only if the latent variable is greater than zero.

Our models are simulated for 200 periods, a sample size typically found in macroeconometric time series applications (such as in the original Qual VAR paper by Dueker 2005).[8]

The typical economic event that is modeled by the Qual VAR, such as crises, recessions, and business cycle turning points, is rather rare but occurs sufficiently often in the sample period to obtain a certain idea of the underlying dynamics. That is, for our Monte Carlo study, the binary process is required to have multiple (blocks of) events while having an overall event probability clearly below 50%. Pretests show that an event probability of 20% satisfies these criteria for all coefficient matrices Φ and covariance matrices of the error terms Σ considered in this paper.

For simplicity, we set the constant term in the observable equation to zero and assume diagonal covariance matrices.[9] Therefore, as the variance of the error term in the latent variable equation, σ_y^2, is held constant at one, the volatility of the whole system is primarily driven by the variance of the error term in the observable variable equation, σ_x^2. Thus, we can scale this volatility through a single parameter of the MCMC setup. In this context, volatility may have different implications. In forecasting, volatility is a main driver of uncertainty. However, if the observable Granger causes the latent variable, a high σ_x implies that a large share of the volatility of the latent variable can be attributed to changes in the observable variable. This can greatly facilitate the identification of the latent variable. If, on the other hand, the chain of causality implies that the Qual VAR identifies the latent variable mostly through future values of the observable variable (a backward identification), a large σ_x may be an obstacle. To test the influence of σ_x on the

[8] At the same time, this is a good compromise between short samples that might cause additional small sample problems in the estimation and large samples that rapidly increase computational requirements. We add a swing in phase of 100 periods to each simulation that is dropped before estimation, thereby guaranteeing independence from the starting values.

[9] Although we denote the observable variable by x (or X in case of multiple variables), the standard deviation of the error term in the observable equation is denoted by σ_x. Correspondingly, while the latent variable is denoted consistently by y^*, the corresponding standard deviation is given by σ_y.

identification and forecasting performance of the Qual VAR, we use three different specifications of σ_x: a low volatility specification with $\sigma_x = 0.1$ (labeled *low* in the remainder of the paper), an equal variance specification with $\sigma_x = \sigma_y = 1(eq)$, and a high variance specification where $\sigma_x = 10$ $(high)$.

We test all three variance specifications in DGPs covering all potential causality chains; i.e., (1) the observable Granger causes the latent variable (labeled *ol* in the remainder of the paper), (2) the latent variable Granger causes the observable variable (*lo*), and (3) the observable and the latent variables mutually Granger cause each other (*olo*). For all model structures covered by our analysis, the matrix Φ is given in Table 1.[10]

To construct examples with strong causality chains, the parameter on the off-diagonal is set to 0.7 whenever Granger causality exists. To ensure strong intertemporal dependence and stationarity of the processes at the same time, the sum of the off-diagonal parameter and the autocorrelation term in the same column is restricted to 0.9. Depending on the setting, the latent variable may show different persistence. While some settings produce rare events of long duration, other settings result in shorter, but more frequent events. That is, our data generating process can reproduce a variety of real world examples.

Table 1. Coefficient matrix for the different causality chains

		observable → latent	
		no	yes
latent → observable	no	-	$\begin{pmatrix}0.9 & 0.7\\ 0 & 0.2\end{pmatrix}$
	yes	$\begin{pmatrix}0.2 & 0\\ 0.7 & 0.9\end{pmatrix}$	$\begin{pmatrix}0.2 & 0.7\\ 0.7 & 0.2\end{pmatrix}$

Given Φ and Σ, the probability of an event, i.e., $P(y^* > 0)$, is driven by the constant term in the latent variable equation. Assuming that the number of events in the sampled process is binomially distributed, we develop an acceptance-rejection algorithm where a VAR simulated with given parameters, covariance matrix and constants is only accepted if the simulated event probability is not statistically different from the 20%. Otherwise, the constant of the latent variable is adjusted and the process is resimulated.

[10] We do not investigate the possibility that the observable and latent variables are independent (with possible autocorrelation).

In total, we consider nine different settings of the VAR that differ along two dimensions in Sections IV and V. For robustness, four additional settings are introduced and analyzed in Section VI. In the following, we assess the quality of the in-sample estimation and the forecasting performance of the Qual VAR for the first nine settings. Due to the different requirements of the tests, the number of iterations of the Monte Carlo study and of the Gibbs sampler (applied in each Monte Carlo iteration) is set individually, as outlined in Table 2. For convenience, the table also lists the specifications of the Gibbs sampler as described in Section II.

Table 2. Monte Carlo and Gibbs Sampler setup

		Identification	Forecasting	Robustness
Section		4	5	6
MCMC	Simulations per model	1,000	10,000	1,000
Gibbs Sampler	Total iterations	10,000	4,000	10,000
	Swing in iterations	5,000	2,000	5,000
	Spacing	5	2	5
	Final iterations	1,000	1,000	1,000

Note: A spacing of *m* means that every *m*th iteration of the Gibbs sampling is used to compute the final distributions after discarding the first swing in iterations. Final iterations refers to the number of iterations chosen in that way.

IV. Identification problems

In this section, we analyze whether the Qual VAR is able to correctly identify the latent variable, the parameters and the covariance matrix of the VAR. As described above, the Gibbs sampler produces a distribution of the elements of Λ that we use to calculate the median results and Fry-Pagan estimates $\Lambda_{med}, \Lambda_{FP}$. Three tests are performed on the estimations of the Qual VAR. First, we determine whether the estimates of the latent variable fit the true latent. A low level of accordance would imply that conclusions drawn from the estimated values of the latent variable must be treated cautiously. We test for unbiasedness using a method from the forecasting evaluation literature (Holden and Peel 1990). However, while a perfect fit is a desirable property, economic conclusions from level differences in the latent variable can also be derived if the dynamics are correctly reproduced. Therefore, in our second series of tests, we focus on the explanatory power of the test equation rather than on the coefficient estimates that are usually considered.

Because the Gibbs sampler enforces the correct sign of the latent variable, it produces some correlation between the estimated and the true latent variable by construction, even if the economic story behind the latent variable is not correctly captured by the model. Therefore, both tests are performed using non-event periods (i.e., roughly 80% of the sample). The results of those tests are reported in subsection IV.A.

Because our model features considerable persistence, our previous tests may indicate that we correctly capture the dynamics of the system, although the turning points of the latent time series are shifted. In this case, the economic interpretation of the estimated model does not replicate the true data generating process. Therefore, in subsection IV.B, we run a series of Granger causality tests to assess whether the chain of causality implied by the true parameter matrix is correctly identified.[11] A correct estimation of the direction of causality may be enough for a qualitative, although not quantitative, economic interpretation of the results obtained by the Qual VAR.

A. Correct estimation of the latent variable

To test if the estimated latent variable is an unbiased estimate of the true latent variable, we regress the two variables on the subsample of non-event periods. That is, we exclusively consider negative values of the (true and estimated) latent. This restriction avoids the possibility that we find correlation between the two latent variables that is introduced by the sign restriction of the Qual VAR estimate.

$$y^*_{true} = \alpha + \beta\, y^* + \varepsilon,^{12} \tag{3}$$

where y^* can be both the median latent variable y^*_{med} or the Fry-Pagan latent y^*_{FP}. In Table 3, we show the estimates of regression (3) as well as the test results for the hypotheses $\alpha = 0$ and $\beta = 1$. As the Monte Carlo simulation is run 1,000 times with different true processes, we present the mean estimate of $(\alpha; \beta)$ and the share of simulations, where t-tests (individually) and F-tests (jointly) reject the two hypotheses.

[11] Strictly speaking, our test is more restrictive than a traditional Granger causality test, as our tests require the correct sign of the parameter.

[12] As the equation is estimated on the non-event periods $(y^* < 0)$, $\alpha > 0$, and $\beta < 1$ imply a negative bias of the Qual VAR estimates.

Table 3. Estimation of the latent variable and dynamics, corrected equation

	ol, high	*ol, eq*	*ol, low*	*lo, high*	*lo, eq*	*lo, low*	*olo, high*	*olo, eq*	*olo, low*
y^*_{med}	(2.74;6.07)	(-0.11;0.98)	(-0.59;0.55)	(-0.13;1.10)	(-0.01;1.10)	(0.27;1.39)	(1.33;3.71)	(-0.11;0.96)	(0.16;1.14)
% reject t	(97.3;100.0)	(21.2;55.9)	(43.4;39.0)	(32.4;52.0)	(21.8;67.2)	(94.6;99.4)	(91.1;100.0)	(16.9;37.1)	(91.2;93.9)
% reject F	100.0	73.0	39.2	82.0	81.1	99.4	100.0	55.2	90.0
R^2_{med}	0.9834	0.6523	0.0338	0.3385	0.7647	0.9717	0.9562	0.6572	0.9634
y^*_{FP}	(-0.45;5.24)	(-0.75;0.72)	(-1.16;0.04)	(-1.49;0.51)	(-0.44;0.92)	(0.07;1.30)	(-0.42;3.25)	(-0.57;0.74)	(0.04;1.07)
% reject t	(23.2;100.0)	(96.9;89.9)	(100.0;100.0)	(99.6;92.0)	(76.4;55.2)	(33.3;96.8)	(19.0;100.0)	(95.7;89.2)	(29.1;67.6)
% reject F	100.0	98.8	100.0	100.0	92.3	96.7	100.0	96.9	62.4
	0.8617	0.4823	0.0147	0.1705	0.6486	0.9027	0.8361	0.5008	0.8933

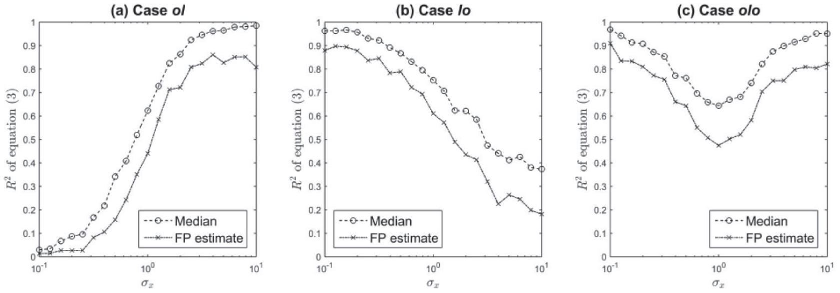
Note: Mean results ($\alpha; \beta$ of the estimation described in Equation (3) and corresponding t- and F-tests for the median latent variable and the Fry-Pagan latent variable .

We find that the null hypothesis of an unbiased estimate is rejected by the F test in the vast majority of cases in all settings except *ol,low*. However, the situation in this setting is even worse. The reason for not rejecting the rationality test is not a good fit of the estimated latent variable, but high residuals of a very bad fit. This creates high uncertainty, reducing rejection rates. Thus, in many cases we can neither reject $\beta = 1$ nor $\beta = 0$.

However, the problems are not as severe for many other settings. Although we reject $\beta = 1$ in most simulations, we often find $\beta > 0$. In all cases except *ol,low*, β is significantly greater than zero in every bootstrap iteration. That is, the Qual VAR captures at least some of the dynamics of the latent variable in most settings. This is partly reflected by the results of our second test series. However, only in four out of nine cases are the results convincing with an R^2 greater than 0.9. While some more settings (*ol,eq*; *lo,eq;* and *olo,eq*) produce at least moderate results with $0.6 < R^2 < 0.8$, it should be considered that the testing environment is rather favorable for the Qual VAR as the true structure of the model (i.e., the selection of variables and the lag order) is known and we consider particularly simple models.

When explaining the Fry-Pagan estimate of the latent with the true latent, R^2 values are even lower (see Figure 1). Again, the reason is the lower degree of noise in the median estimate (compared to the Fry-Pagan estimate). Thus, the difference in $R^2$is mostly due to differences in the variance that must be explained, rather than the variance that is explained by the true latent variable.

Figure 1. of estimation (3) for between 0.1 and 10, logarithmic equally spaced



Note: The reported results are the average of 100 MCMC iterations (instead of the normal 1,000). The reduction was necessary to reduce runtime to a tolerable level.

Whether the Qual VAR captures the dynamics of the latent variable (as described by the R^2) strongly depends on the variance of shocks in the observable equation. If the observable Granger causes the latent variable, estimation is simplified by a high σ_x. If, however, the latent Granger causes the observable variable, a high σ_x strongly decreases the R^2 of our test equation. The reason is that in both cases, the latent variable is mostly identified using information from the observable variable. In the first case (*ol*), the latent variable is strongly correlated to past values of the observable variable. Because the entire variation of the observable variable affects the latent variable, more variance represents information that can be exploited in the estimation. On the contrary, in the second case (*lo*), the latent variable is correlated to future values of the observable variable. However, only the predetermined part of the observable variable contains information on the latent variable. A high σ_x obfuscates the view on the predetermined part of the observable, thus impeding the estimation of y. This difficulty is slightly alleviated (compared to the case *ol*) by the high autocorrelation of the latent variable.

Figure 1, panels (a) and (b) show the corresponding results for a larger set of different levels of σ_x based on MCMC simulations with fewer iterations (100 instead of 1,000). In the *olo* settings (see Figure 1, panel (c)), we find a non-monotonic impact of σ_x on R^2. In this case, we can draw information on the latent variable from both past and future values of the observable variable. Initially, when σ_x increases, the loss of information drawn from the future outweighs the gains of information from the past. However, when the variance increases further, the benefit of more information from past observations dominates the impact of σ_x. In our case, with symmetric mutual causality between the latent and the observable variables, the turning point coincides with the equal variance setting (*olo,eq*).

B. Correct identification of Granger causality

When testing the correct identification of Granger causality, we report two results. First, for each parameter in Φ, we report the share of iterations where causality is correctly identified (see Table 4). To allow sound economic interpretation, the Qual VAR must capture existing causalities while avoiding the erroneous identification of causalities where none exist. Therefore, if positive true parameters are considered, we report the share of Monte Carlo iterations producing significantly positive estimates of the respective parameter. If parameters that are set to zero are considered, we report the share of iterations producing insignificant results. Second, as a summary of those results, we report the share of iterations where each of the four entries of Φ indicates the correct causality.

Table 4. Causality and parameter estimates

	ol, high	*ol, eq*	*ol, low*	*lo, high*	*lo, eq*	*lo, low*	*olo, high*	*olo, eq*	*olo, low*
Φ	$\begin{pmatrix}0.90 & 0.70\\ 0.00 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.90 & 0.70\\ 0.00 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.90 & 0.70\\ 0.00 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.00\\ 0.70 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.00\\ 0.70 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.00\\ 0.70 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.70\\ 0.70 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.70\\ 0.70 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.70\\ 0.70 & 0.20\end{pmatrix}$
Φ_{med}	$\begin{pmatrix}0.89 & 0.12\\ -0.03 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.89 & 0.73\\ -0.01 & 0.14\end{pmatrix}$	$\begin{pmatrix}0.89 & 0.69\\ -0.00 & 0.15\end{pmatrix}$	$\begin{pmatrix}0.19 & 0.00\\ 0.75 & 0.86\end{pmatrix}$	$\begin{pmatrix}0.21 & 0.05\\ 0.77 & 0.82\end{pmatrix}$	$\begin{pmatrix}0.41 & 0.13\\ 0.71 & 0.68\end{pmatrix}$	$\begin{pmatrix}0.18 & 0.19\\ 2.42 & 0.19\end{pmatrix}$	$\begin{pmatrix}0.18 & 0.67\\ 0.70 & 0.18\end{pmatrix}$	$\begin{pmatrix}0.26 & 0.62\\ 0.73 & 0.13\end{pmatrix}$
Φ_{FP}	$\begin{pmatrix}0.89 & 0.13\\ -0.02 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.89 & 0.75\\ -0.01 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.88 & 0.70\\ 0.00 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.18 & -0.00\\ 0.97 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.11 & 0.00\\ 0.91 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.38 & 0.12\\ 0.77 & 0.90\end{pmatrix}$	$\begin{pmatrix}0.17 & 0.19\\ 2.49 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.14 & 0.64\\ 0.80 & 0.20\end{pmatrix}$	$\begin{pmatrix}0.25 & 0.60\\ 0.76 & 0.20\end{pmatrix}$
$\%\,t_{\Phi}$	$\begin{pmatrix}91.1 & 0.0\\ 90.5 & 98.7\end{pmatrix}$	$\begin{pmatrix}90.2 & 88.8\\ 89.3 & 86.9\end{pmatrix}$	$\begin{pmatrix}87.3 & 89.9\\ 89.4 & 88.9\end{pmatrix}$	$\begin{pmatrix}90.8 & 90.0\\ 87.4 & 83.9\end{pmatrix}$	$\begin{pmatrix}87.8 & 87.0\\ 82.9 & 82.2\end{pmatrix}$	$\begin{pmatrix}9.1 & 56.3\\ 91.1 & 37.2\end{pmatrix}$	$\begin{pmatrix}88.6 & 0.0\\ 0.2 & 96.0\end{pmatrix}$	$\begin{pmatrix}87.9 & 88.5\\ 86.6 & 91.1\end{pmatrix}$	$\begin{pmatrix}62.8 & 82.8\\ 83.4 & 84.3\end{pmatrix}$
$\%F_{\Phi}$	0.0	69.0	61.6	61.4	62.8	6.9	0.0	66.3	43.4
G%Gr(ind)	$\begin{pmatrix}100.0 & 100.0\\ 90.5 & 50.4\end{pmatrix}$	$\begin{pmatrix}100.0 & 100.0\\ 89.3 & 29.3\end{pmatrix}$	$\begin{pmatrix}100.0 & 47.1\\ 89.4 & 30.8\end{pmatrix}$	$\begin{pmatrix}80.8 & 90.0\\ 49.1 & 100.0\end{pmatrix}$	$\begin{pmatrix}71.9 & 87.0\\ 100.0 & 100.0\end{pmatrix}$	$\begin{pmatrix}99.7 & 56.3\\ 100.0 & 100.0\end{pmatrix}$	$\begin{pmatrix}76.6 & 100.0\\ 100.0 & 92.6\end{pmatrix}$	$\begin{pmatrix}74.5 & 100.0\\ 100.0 & 62.6\end{pmatrix}$	$\begin{pmatrix}99.2 & 100.0\\ 100.0 & 40.9\end{pmatrix}$
G%Gr(all)	45.3	26.3	11.4	33.6	61.1	56.3	69.6	38.9	40.2
σ_x	10.00	1.00	0.10	10.00	1.00	0.10	10.00	1.00	0.10
$\sigma_{x,\text{med}}$	10.02	1.00	0.10	9.97	0.99	0.32	9.78	1.00	0.25
$\sigma_{x,\text{FP}}$	9.99	0.99	0.10	9.89	0.92	0.29	9.67	0.94	0.22
$\%\,t_{\sigma_x}$	89.6	90.4	90.5	89.6	88.7	0.0	88.2	89.7	0.0

Note: This table contains the true matrix Φ, the median and Fry-Pagan parameter matrix Φ_{med}, Φ_{FP}. Furthermore, it contains the percentage of Monte Carlo iterations, in which the parameter matrix Φ was inside the confidence interval given by the Gibbs sampler (element-by-element in row t_{Φ} and as a whole in row F_{Φ}). The same applies to the standard observation of the observable, σ_x.Furthermore, we report the share of iterations in which estimated individual parameters captured the true chain of causality individually (Gr(ind)) and jointly (Gr(all)).

We find a substantial share of iterations, where at least one parameter produces an incorrect estimate. Even in the setting where the Qual VAR performs best in this respect, the correct chain of causality is merely identified in less than 70% of the MCMC iterations (see Gr(all) in Table 4). The most frequent reason to reject the joint test is the inability of the Qual VAR to identify weak autoregressive behavior in the latent variable ($\Phi_{1,2} = 0.2$). In three additional settings, there are severe identification problems concerning causality ($\Phi_{1,2}$and $\Phi_{2,1}$).

First, if the observable Granger causes the latent variable and σ_x is low (*ol,low*), the Qual VAR does not capture this causality in approximately 50% of all iterations. Second, a similarly high rejection rate is found for the opposite case *lo,high*, where we are unable to detect the causality from the latent to the observable. This corresponds to the two settings where the identification of the latent variable is most difficult. Third, with causality running from the latent to the observable variable with low σ_x (*lo,low*), the Qual VAR incorrectly finds a significant effect from the observable on the latent variable in roughly 40% of the Monte Carlo iterations. This is because the actual correlation between past and future values of the observable is much higher in this setting than indicated by the autoregressive parameter of the observable ($\Phi_{1,1} = 0.2$). The high persistence of the observable variable is mostly due to the high persistence of the latent variable ($\Phi_{2,2} = 0.9$) that is the main driving force of the observable ($\Phi_{2,1} = 0.7$combined with $\sigma_x = 0.1$). Therefore, y^* and correspondingly the observable occurrence of the modeled event (y) is correlated to both past and future values of the observable variable. Accordingly, the autoregressive coefficient of the observable variable is overestimated, as implicit autocorrelation (via the latent) is mistaken for true autoregressive behavior. The autoregressive behavior of the latent variable is underestimated. This is reflected in the parameter estimates. To a lesser extent, the same problem is found with higher values of σ_x for the same causality setting.[13]

[13] As reference, Table 4 also lists the parameter estimates and the share of iterations where the true parameter values are within the estimated confidence bounds. However, the high shares we find for most parameters are often due to extremely broad confidence bounds around the parameters. These broad confidence bounds are caused by the uncertainty concerning the latent variable that is usually poorly identified.

V. Forecasting performance

The Qual VAR has mostly been used for forecasting binary events, which may still work despite the identification problems documented in the previous section. Therefore, we assess the forecasting performance of the Qual VAR over different horizons in comparison to three econometric benchmark models (and the unconditional event probability). We perform both a probabilistic assessment, using root mean squared forecast errors, and a non-probabilistic assessment based on hit rates and false alarm rates. Before presenting these results, we compare predicted event probabilities of the Qual VAR estimates to the true model directly.

A. Forecast performance compared to the true model

In Table 5 we start with a descriptive analysis of the predicted probability of the binary event $P(\hat{y}^*_{T+h} > 0)$ in periods where the binary event occurs at the forecast horizon h (i.e. $y_{T+h} = 1$), compared to situations with no event at the forecast horizon (i.e. $y_{T+h} = 0$).[14] The probability of an event in the far future can be highly uncertain even for the true DGP. Therefore, we contrast the performance of forecasts obtained from Qual VAR estimates with forecasts obtained from the true DGP. This concept that we employ for a first glance at model performance combines benefits of probabilistic prediction (i.e. the prediction of event probabilities) and non-probabilistic forecasts (where the prediction itself is binary like the event). The approach is related to the (non-probabilistic) concept of hit rates and false alarm rates that is more commonly found in the literature Ratcliff (2013). In the non-probabilistic evaluation of binary forecast performance, the sample is split into event and non-event periods. These periods are then compared to binary signals obtained from predicted event probabilities that are separated using a calibrated probability threshold. Our descriptive approach borrows from this evaluation method by separating the sample in event and non-event periods, but still retains the key information of the probabilistic prediction, i.e. the predicted probability. This allows a similar comparison of performance in event and non-event periods as hit rates do, but does not require the additional technical layer of threshold calibration (which we will introduce in Subsection V.B).

[14] The probabilities are determined through a series of simulations rather than using a deterministic forecast, to account for the covariance of shocks that is considered in the Qual VAR.

Table 5. Forecasting event probabilities with the Qual VAR

	ol, high	*ol, eq*	*ol, low*	*lo, high*	*lo, eq*	*lo, low*	*olo, high*	*olo, eq*	*olo, low*
true forecast:	1.00 / 0.00	0.69 / 0.01	0.21 / 0.18	0.70 / 0.01	0.71 / 0.01	0.71 / 0.01	1.00 / 0.00	0.60 / 0.03	0.40 / 0.08
event vs	0.79 / 0.00	0.56 / 0.03	0.20 / 0.18	0.53 / 0.04	0.53 / 0.04	0.53 / 0.04	0.49 / 0.05	0.42 / 0.07	0.39 / 0.09
no event	0.38 / 0.09	0.34 / 0.10	0.20 / 0.19	0.34 / 0.12	0.32 / 0.12	0.32 / 0.12	0.32 / 0.13	0.30 / 0.14	0.27 / 0.15
1,2,5,10 periods	0.26 / 0.17	0.23 / 0.16	0.19 / 0.19	0.25 / 0.18	0.24 / 0.17	0.25 / 0.18	0.25 / 0.19	0.23 / 0.18	0.22 / 0.19
med forecast:	0.89 / 0.00	0.66 / 0.01	0.20 / 0.18	0.65 / 0.01	0.64 / 0.02	0.56 / 0.03	0.84 / 0.00	0.57 / 0.03	0.35 / 0.09
event vs	0.65 / 0.01	0.53 / 0.03	0.20 / 0.19	0.46 / 0.05	0.48 / 0.06	0.44 / 0.07	0.42 / 0.06	0.38 / 0.08	0.33 / 0.10
no event	0.32 / 0.09	0.31 / 0.11	0.19 / 0.19	0.23 / 0.13	0.24 / 0.13	0.24 / 0.14	0.26 / 0.12	0.26 / 0.14	0.23 / 0.16
1,2,5,10 periods	0.20 / 0.15	0.21 / 0.17	0.19 / 0.19	0.18 / 0.17	0.19 / 0.17	0.19 / 0.17	0.19 / 0.17	0.20 / 0.18	0.19 / 0.18
FP forecast:	0.88 / 0.00	0.66 / 0.01	0.21 / 0.19	0.57 / 0.02	0.60 / 0.02	0.53 / 0.04	0.85 / 0.00	0.55 / 0.04	0.34 / 0.09
event vs	0.63 / 0.01	0.53 / 0.04	0.21 / 0.20	0.42 / 0.07	0.44 / 0.06	0.41 / 0.08	0.42 / 0.06	0.39 / 0.08	0.34 / 0.11
no event,	0.31 / 0.09	0.30 / 0.12	0.20 / 0.20	0.23 / 0.14	0.25 / 0.14	0.25 / 0.14	0.26 / 0.12	0.26 / 0.14	0.22 / 0.16
1,2,5,10 periods	0.20 / 0.16	0.21 / 0.17	0.20 / 0.20	0.18 / 0.18	0.20 / 0.18	0.20 / 0.18	0.19 / 0.17	0.20 / 0.18	0.19 / 0.18

Note: The blocks contain the average forecasted probability of an event in event periods vs. non-event periods at forecast horizons of one, two, five, and ten periods.

In all cases except *ol,low* (where conditional probabilities are always close to the unconditional probability of approximately 20%), the one period ahead conditional probability in event (non-event) periods is above (below) 40% (10%).

In all cases where the true model contains substantial information about the risk of an event in the future, a large share of this is captured by the estimates. Generally, we find that the estimated probabilities are much closer to the probabilities implied by the true model than they are to the unconditional probability. That is, more than half of the risk explained by the true model is also explained by the Qual VAR estimates.

B. Forecast performance relative to benchmark models

Selection of benchmark models

We turn to a more formal analysis of forecast performance, matching the performance of the Qual VAR with alternative models using both probabilistic and non-probabilistic forecast evaluation tools. We consider four different benchmarks as challengers for the Qual VAR. First, we use an uninformed unconditional probability forecasts (labeled “uncond” in Tables 6 and 7). Second, we use a simple probit model that only considers the exogenous variables (labeled “probit” in the tables). Third, we use a dynamic probit that uses the lagged observable binary variable as additional explanatory variable following Moneta (2005) (labeled “M probit”). Forth, we use an approximation of the dynamic probit proposed by Kauppi and Saikkonen (2008). Kauppi and Saikkonen reframe the probit model slightly, defining the latent variable to be strictly deterministic on the observables and its own lags. That is, shocks do not affect the latent itself, but only the realization of the binary variable. Since the latent variable is deterministic (given some initial values) this model can be estimated by a straightforward, but computationally intensive, maximum likelihood. However, Ratcliff (2013) argues that the forecasts of a Kauppi and Saikkonen (2008) model are equivalent to a simple probit that considers higher lag orders of the observables. Thus, we estimate a probit augmenting the model with three additional lags of the observables to approximate the Kauppi and Saikkonen (2008) model. The corresponding rows are labeled “KS probit’.

Since only the Qual VAR models the entire system, it is also the only one that allows for indirect forecasts (obtained from a simulation of the system). All benchmark models need to rely on direct forecasts.

Table 6. Root Mean Squared Forecast Errors, the Qual VAR in comparison

		ol, high	*ol, eq*	*ol, low*	*lo, high*	*lo, eq*	*lo, low*	*olo, high*	*olo, eq*	*olo, low*
RMSFE(1)	med	0.13	0.06	0.06	0.12	0.12	0.12	0.11	0.05	0.06
	FP	0.13	0.07	0.08	0.16	0.15	0.14	0.12	0.07	0.07
	M probit	$0.10^{--/--}$	$0.07^{***/\dagger\dagger\dagger}$	$0.07^{***/--}$	$0.21^{***/\dagger\dagger\dagger}$	$0.19^{***/\dagger\dagger\dagger}$	$0.17^{***/\dagger\dagger\dagger}$	$0.13^{***/--}$	$0.08^{***/\dagger\dagger\dagger}$	$0.07^{***/--}$
	KS probit	$0.14^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.21^{***/\dagger\dagger\dagger}$	$0.20^{***/\dagger\dagger\dagger}$	$0.18^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$
	probit	$0.10^{--/--}$	$0.06^{***/--}$	$0.07^{***/--}$	$0.31^{***/\dagger\dagger\dagger}$	$0.21^{***/\dagger\dagger\dagger}$	$0.17^{***/\dagger\dagger\dagger}$	$0.13^{***/\dagger\dagger}$	$0.07^{***/--}$	$0.07^{***/--}$
	uncond	$0.40^{***/\dagger\dagger\dagger}$	$0.30^{***/\dagger\dagger\dagger}$	$0.08^{***/--}$	$0.31^{***/\dagger\dagger\dagger}$	$0.31^{***/\dagger\dagger\dagger}$	$0.31^{***/\dagger\dagger\dagger}$	$0.39^{***/\dagger\dagger\dagger}$	$0.28^{***/\dagger\dagger\dagger}$	$0.22^{***/\dagger\dagger\dagger}$
RMSFE(2)	med	0.06	0.04	0.04	0.13	0.11	0.11	0.05	0.08	0.09
	FP	0.07	0.05	0.06	0.16	0.14	0.13	0.07	0.11	0.11
	M probit	$0.05^{--/--}$	$0.05^{***/\dagger\dagger\dagger}$	$0.06^{***/--}$	$0.19^{***/\dagger\dagger\dagger}$	$0.17^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$
	KS probit	$0.08^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.19^{***/\dagger\dagger\dagger}$	$0.18^{***/\dagger\dagger\dagger}$	$0.16^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$
	probit	$0.03^{--/--}$	$0.04^{--/--}$	$0.05^{***/--}$	$0.27^{***/\dagger\dagger\dagger}$	$0.18^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$
	uncond	$0.33^{***/\dagger\dagger\dagger}$	$0.27^{***/\dagger\dagger\dagger}$	$0.06^{***/--}$	$0.27^{***/\dagger\dagger\dagger}$	$0.27^{***/\dagger\dagger\dagger}$	$0.26^{***/\dagger\dagger\dagger}$	$0.25^{***/\dagger\dagger\dagger}$	$0.23^{***/\dagger\dagger\dagger}$	$0.22^{***/\dagger\dagger\dagger}$
RMSFE(5)	med	0.05	0.04	0.04	0.12	0.10	0.09	0.05	0.06	0.07
	FP	0.07	0.06	0.05	0.14	0.12	0.11	0.07	0.08	0.08
	M probit	$0.08^{***/\dagger\dagger\dagger}$	$0.07^{***/\dagger\dagger\dagger}$	$0.05^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.13^{***/\dagger\dagger\dagger}$	$0.12^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$
	KS probit	$0.10^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.16^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$
	probit	$0.06^{***/--}$	$0.05^{***/--}$	$0.05^{***/--}$	$0.19^{***/\dagger\dagger\dagger}$	$0.13^{***/\dagger\dagger\dagger}$	$0.11^{***/--}$	$0.09^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$
	uncond	$0.22^{***/\dagger\dagger\dagger}$	$0.19^{***/\dagger\dagger\dagger}$	$0.05^{***/--}$	$0.19^{***/\dagger\dagger\dagger}$	$0.19^{***/\dagger\dagger\dagger}$	$0.18^{***/\dagger\dagger\dagger}$	$0.19^{***/\dagger\dagger\dagger}$	$0.16^{***/\dagger\dagger\dagger}$	$0.15^{***/\dagger\dagger\dagger}$
RMSFE(10)	med	0.05	0.04	0.03	0.09	0.08	0.08	0.05	0.06	0.06
	FP	0.08	0.07	0.05	0.12	0.10	0.10	0.07	0.08	0.08
	M probit	$0.10^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.06^{***/\dagger\dagger\dagger}$	$0.13^{***/\dagger\dagger\dagger}$	$0.12^{***/\dagger\dagger\dagger}$	$0.12^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$
	KS probit	$0.12^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.14^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$
	probit	$0.08^{***/\dagger\dagger\dagger}$	$0.07^{***/--}$	$0.05^{***/--}$	$0.12^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.08^{***/\dagger\dagger\dagger}$	$0.08^{***/--}$	$0.08^{***/\dagger\dagger\dagger}$
	uncond	$0.13^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.04^{***/--}$	$0.12^{***/\dagger\dagger\dagger}$	$0.12^{***/\dagger\dagger\dagger}$	$0.12^{***/\dagger\dagger\dagger}$	$0.11^{***/\dagger\dagger\dagger}$	$0.10^{***/\dagger\dagger\dagger}$	$0.09^{***/\dagger\dagger\dagger}$

Note: The next four blocks contain, for each of those four forecast horizons, the RMSFE of the forecast based on the median latent variable, the Fry-Pagan latent variable, a probit using only the observable variable and that of the unconditional probability. For the first two RMSFEs, significant outperformance of the median Qual-VAR prediction at the 1% (5% / 10%) level is indicated by *** (** / *), while significant outperformance of the Fry-Pagan prediction is indicated by ††† (†† / †).

Table 7. Hit Rates (HR) and False Alarm Rates (FAR), the Qual VAR in comparison

		ol, high	*ol, eq*	*ol, low*	*lo, high*	*lo, eq*	*lo, low*	*olo, high*	*olo, eq*	*olo, low*
HR / FAR(1)	true	0.99 / 0.04	0.89 / 0.16	0.45 / 0.29	0.75 / 0.07	0.74 / 0.07	0.75 / 0.07	0.99 / 0.06	0.88 / 0.23	0.77 / 0.25
	med	1.00 / 0.11	0.89 / 0.18	0.48 / 0.35	0.74 / 0.07	0.74 / 0.07	0.75 / 0.07	1.00 / 0.17	0.87 / 0.23	0.76 / 0.26
	FP	0.99 / 0.11	0.89 / 0.18	0.51 / 0.40	0.75 / 0.08	0.74 / 0.08	0.75 / 0.08	0.99 / 0.17	0.86 / 0.23	0.75 / 0.26
	M probit	0.97 / 0.03	0.87 / 0.16	0.51 / 0.44	0.65 / 0.11	0.76 / 0.19	0.82 / 0.21	0.95 / 0.06	0.84 / 0.20	0.77 / 0.27
	KS probit	0.94 / 0.03	0.85 / 0.15	0.47 / 0.43	0.68 / 0.14	0.74 / 0.19	0.81 / 0.21	0.93 / 0.05	0.82 / 0.20	0.74 / 0.26
	probit	0.97 / 0.04	0.87 / 0.16	0.55 / 0.45	0.54 / 0.47	0.79 / 0.25	0.83 / 0.22	0.95 / 0.06	0.84 / 0.20	0.78 / 0.27
TP / FP(2)	true	0.94 / 0.28	0.89 / 0.26	0.55 / 0.45	0.69 / 0.11	0.69 / 0.11	0.69 / 0.12	0.98 / 0.66	0.74 / 0.21	0.66 / 0.17
	med	0.96 / 0.26	0.89 / 0.27	0.53 / 0.44	0.67 / 0.11	0.67 / 0.11	0.66 / 0.12	0.93 / 0.43	0.69 / 0.19	0.59 / 0.15
	FP	0.96 / 0.27	0.89 / 0.27	0.51 / 0.44	0.68 / 0.15	0.68 / 0.13	0.69 / 0.14	0.92 / 0.42	0.67 / 0.19	0.62 / 0.17
	M probit	0.90 / 0.13	0.83 / 0.19	0.53 / 0.45	0.59 / 0.15	0.72 / 0.24	0.79 / 0.26	0.74 / 0.27	0.71 / 0.30	0.69 / 0.32
	KS probit	0.88 / 0.13	0.82 / 0.19	0.50 / 0.44	0.62 / 0.17	0.69 / 0.23	0.75 / 0.24	0.72 / 0.24	0.67 / 0.27	0.64 / 0.30
	probit	0.90 / 0.13	0.83 / 0.19	0.55 / 0.46	0.54 / 0.48	0.75 / 0.29	0.80 / 0.26	0.76 / 0.29	0.72 / 0.31	0.70 / 0.34
TP / FP(5)	true	0.97 / 0.80	0.89 / 0.53	0.61 / 0.52	0.71 / 0.29	0.71 / 0.29	0.71 / 0.30	0.99 / 0.92	0.81 / 0.46	0.68 / 0.34
	med	0.98 / 0.73	0.90 / 0.56	0.60 / 0.56	0.73 / 0.40	0.74 / 0.37	0.72 / 0.33	0.99 / 0.86	0.82 / 0.51	0.67 / 0.36
	FP	0.98 / 0.73	0.91 / 0.56	0.63 / 0.57	0.73 / 0.45	0.74 / 0.41	0.71 / 0.36	0.98 / 0.83	0.79 / 0.47	0.67 / 0.38
	M probit	0.76 / 0.28	0.72 / 0.30	0.50 / 0.46	0.48 / 0.23	0.64 / 0.33	0.68 / 0.35	0.69 / 0.32	0.66 / 0.34	0.62 / 0.37
	KS probit	0.73 / 0.27	0.68 / 0.29	0.50 / 0.44	0.48 / 0.25	0.57 / 0.31	0.60 / 0.32	0.65 / 0.31	0.62 / 0.33	0.57 / 0.35
	probit	0.77 / 0.29	0.74 / 0.30	0.52 / 0.49	0.53 / 0.49	0.69 / 0.37	0.70 / 0.35	0.71 / 0.33	0.69 / 0.36	0.67 / 0.39
TP / FP(10)	true	1.00 / 1.00	0.98 / 0.92	0.63 / 0.59	0.92 / 0.79	0.91 / 0.81	0.92 / 0.79	1.00 / 1.00	0.93 / 0.81	0.75 / 0.59
	med	1.00 / 0.97	0.98 / 0.92	0.67 / 0.64	0.98 / 0.97	0.97 / 0.89	0.92 / 0.78	1.00 / 0.99	0.95 / 0.87	0.79 / 0.67
	FP	0.99 / 0.96	0.97 / 0.89	0.69 / 0.66	0.93 / 0.89	0.93 / 0.84	0.90 / 0.75	0.99 / 0.97	0.89 / 0.79	0.79 / 0.67
	M probit	0.57 / 0.38	0.59 / 0.40	0.48 / 0.47	0.44 / 0.36	0.50 / 0.42	0.55 / 0.43	0.55 / 0.40	0.52 / 0.41	0.51 / 0.44
	KS probit	0.54 / 0.38	0.55 / 0.38	0.46 / 0.46	0.43 / 0.38	0.49 / 0.40	0.51 / 0.41	0.51 / 0.39	0.48 / 0.40	0.48 / 0.42
	probit	0.65 / 0.42	0.65 / 0.42	0.53 / 0.51	0.54 / 0.50	0.59 / 0.47	0.61 / 0.47	0.61 / 0.44	0.57 / 0.46	0.56 / 0.49

Note: Hit Rates are the number correctly predicted events as a share of total events. False Alarm Rates are the number of falsely predicted events as a share of non-event periods. The probability threshold used to separate predicted probabilities into events and non-events is (separately for every forecasting horizon and estimation method) chosen such that the difference between hit rate and false alarm rate is maximized.

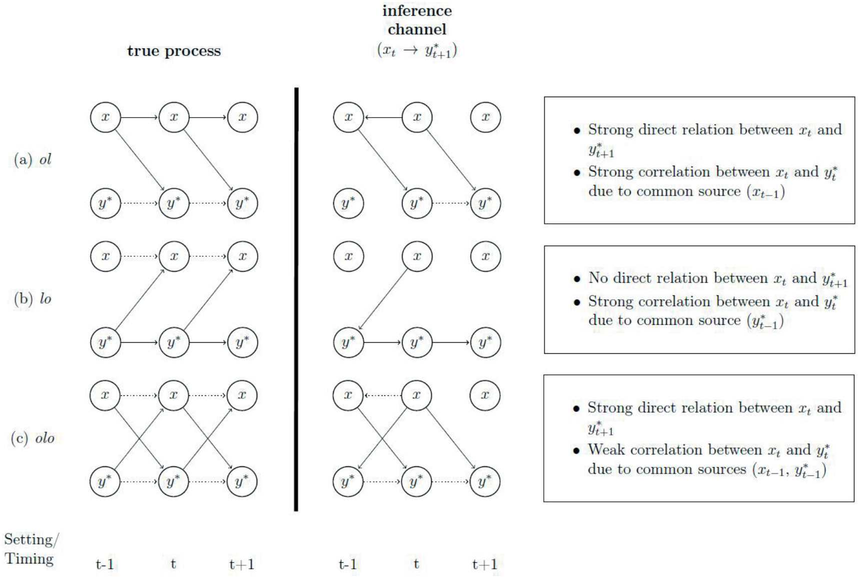
Probabilistic evaluation of forecast performance

Traditionally, the probabilistic evaluation of forecasts is plagued by the lack of information of true probabilities. Therefore, in analyses relying on actual data the accuracy of predictions cannot be calculated precisely. The next best thing commonly done is looking at entire intervals of probability forecasts, and comparing the average prediction in this interval to the share of events in periods where the prediction falls into that interval (Lemeshow and Hosmer 1982). In other words, we can for example assess whether the relative frequency of recessions in periods where the predicted probability is between 30% and 40% is in the same order of magnitude.

Since the data generating process is actually known in our simulation, we can refrain from such approximations and use methods commonly employed in evaluating forecasts of continuous observable variables, such as root mean squared forecast errors (RMSFE). However, since the scaling of the latent variable in probabilistic models is arbitrary, our RMSFE are based on probabilities rather than the prediction of the latent variable itself. More precisely, in Table 6 we compare the predicted probability of an event at $t + h$ given the information set at time with the true conditional probability of an event at $t + h$ given the true model and the state of the world in t.

Considering the probability rather than the actual values has some implications on the interpretation. Commonly, RMSFEs are computed as the difference between the forecast and the true realization of the variable of interest. Thus, they usually increase over the forecast horizon. On the contrary, this is not necessarily true for our application, where the RMSFE is defined as the root mean squared difference between the predicted probability and the true conditional probability of an event. There are two opposing effects on the magnitude of the RMSFE when the forecast horizon increases. First, the forecast uncertainty increases (as with the standard definition of the RMSFE). Second, both the true conditional probability and the forecasted probability converge to the unconditional probability. This causes a reduction of the RMSFE over the forecast horizon. The dominant effect varies both between different forecast horizons and between settings. Therefore, we find n-shaped, u-shaped and decreasing developments of the RMSFE when the forecast horizon increases from one to ten.

Figure 2. Direct causalities in the different settings and inferences (both direct and indirect) from the current observable on the forecasted latent variable



- Strong direct relation between x_t and y^*_{t+1}
- Strong correlation between x_t and y^*_t due to common source (x_{t-1})

- No direct relation between x_t and y^*_{t+1}
- Strong correlation between x_t and y^*_t due to common source (y^*_{t-1})

- Strong direct relation between x_t and y^*_{t+1}
- Weak correlation between x_t and y^*_t due to common sources (x_{t-1}, y^*_{t-1})

Note: The direction of arrows shows the direct causalities on the left hand side. On the right hand side, arrows point in the opposite direction, if the inference channel from to uses that detour. Strong links are given by a solid line, weak links are dotted.

In all cases, the median forecast has a lower RMSFE than the unconditional probability. The Fry-Pagan estimate performs similarly well, outperforming the unconditional probability in all settings except *ol,low*. In this setting, the noise included in the Fry-Pagan estimate makes it impossible to outperform the unconditional probability, which is — in this case — a quite accurate approximation of the true event probability.

However, the more appropriate benchmark is given by alternative models that are used in the prediction of binary events. For most true DGPs and for almost all forecast horizons the median Qual VAR forecast significantly outperforms all benchmarks.

Contrary to the benchmarks, including the single equation dynamic forecasts, only the Qual VAR replicates the full system dynamics, which are particularly important with respect to forecasting over a long horizon. Therefore, the median estimate significantly outperforms all competitors over the five to ten period horizons in all settings. The Fry-Pagan estimate performs slightly worse, but still good.

Whenever the latent variable affects the observable variable (i.e., all *lo* and *olo* settings), these results also hold for the shorter forecast horizons. Only in the *ol* setting with *high* variance in the observable equation can the probit play its strengths and significantly outperform the Qual VAR over short forecast horizons. Since there is only little autoregressive behavior of the latent variable and no feedback from the latent to observables, there is little to be gained from the Qual VAR setup. Yet, the high uncertainty induced by the complex model essentially worsens forecasting performance. In *ol,eq*, the Qual VAR is outperformed by a simple probit for a two period ahead forecast because the benefit of explicitly modeling the time series behavior of the latent variable (as accomplished by the Qual VAR) is very limited in the *ol* settings. The latent variable (and thus the event probability) in $t+1$ depends mostly on the observable in t and – to a lesser extent – on the latent variable in t. However, the impact of y_t^* on y_{t+1}^* is strongly reflected in the correlation between x_t and y_{t+1}^* as both x_t and y_t^* are primarily driven by x_{t-1}. Because the correlation exploited in the probit captures most of the impact of the lagged latent, the value added of the Qual VAR is generally small. If this is combined with situations where the importance of the lagged observable variable (that is included in the probit) is particularly high (e.g., if σ_x is high), the uncertainty carried into the model by trying to identify the dynamic behavior over time overcompensates for the benefits of the identification. This argument is visualized in Figure 2, panel (a).

On the contrary, the probit forecast performs extraordinarily poorly in the *lo* settings, where it is usually outperformed by all of the dynamic approaches, most strongly by the Qual VAR. In *lo,high*, probit cannot even outperform an unconditional forecast. Because a causal link between the lagged observable and the contemporary latent – as modeled by the probit – does not even exist in this case, the probit must entirely rely on the correlation between x_t and y_t^* caused by the common origin y_{t-1}^* (see Figure 2, panel (b)). Especially if σ_x is high, the correlation between x_t and y_t^* is low, thereby further obfuscating the dynamics.

The results for the Fry-Pagan estimate are not quite as good as the results for the median Qual VAR forecast, but still rather convincing. Far more often than not, the Fry-Pagan estimate outperforms all competitors.

The good fit in terms of probability, indicates that the dynamics of the system are captured reasonably well by a Qual VAR despite its flawed identification. This implies that impulse response functions (IRFs) can be derived reasonably well at least in terms of the dynamics. The impulse response functions implied by our coefficient estimates confirm this.[15] However, while the dynamics of the system are generally well captured, the magnitude is substantially misestimated in 4 out of the 9 scenarios for at least one IRF. Also, in particular in scenarios where the estimation of the variances is distorted, the actual identification of structural shocks can be problematic.

Non-probabilistic forecast evaluation

For the non-probabilistic forecast evaluation, we transform the probabilistic predictions into binary predictions, using a probability threshold that is calibrated to maximize the difference between hit rate and false alarm rate within the sample (Ratcliff 2013).[16] In the case of the Qual VAR the in-sample distribution of predicted probabilities at horizon h is based on h period ahead forecasts that are calculated from coefficients using the entire in-sample data. This is necessary to allow comparing the indirect forecasts of the Qual VAR with the direct forecasts provided by the alternative models.

While the performance between the benchmark models and the Qual VAR is similar for short horizons, the Qual VAR tends to overpredict at longer horizons. While this implies a higher hit rate, the benefit in terms of additional hits is bought at the expense of an extremely high rate of false alarms, see Table 7.

VI. Robustness

Previous applications of the Qual VAR are based on much more complex models than the one outlined in this analysis, which might add to the identification problems. While we have very limited information on the true interaction of the latent variables discussed in those models and observable variables, there are some well known features of macroeconomic models that might cause further problems when one of the variables considered is non-observable.

While our benchmark model is fairly flexible, it cannot capture some of those features. One of those features is cyclical behavior that requires a lag order of two

[15] The IRF analysis is not part of this paper. The IRFs are available from the authors on request.

[16] This maximization is identical to the usefulness maximization of the signals approach El-Shagi, Knedlik and von Schweinitz (2013).

or more. Another feature that makes VAR models so popular is the possibility to trace the chain of events, that is, when a shock to one variable eventually affects another variable only through a third factor. Inference might be more difficult when it is the latent variable that links two observable variables, i.e., is caused by one and causes the other. The weak performance of the Qual VAR in our simple benchmark frameworks casts doubt on the ability of the Qual VAR in those more complex settings. However, to demonstrate that our results are robust, we also assess the performance of the Qual VAR in some settings with those features. In this analysis we focus on the area where the Qual VAR had most problems according to our results, i.e., identification.

A. Cyclical behavior

We consider cyclical behavior in the observable and the latent variable, both individually and simultaneously. This is most easily modeled by a DGP with two lags. We choose coefficients in the order of magnitude of the cyclical component of quarterly GDP. The correlation between variables is chosen to guarantee stationarity, albeit allowing some persistence. Setting the variance of the observable to one, this leaves us with the three settings given in Table 8.

Table 8. Specifications with cyclical behavior and results

	Specification		
	cycle o	*cycle l*	*cycle all*
Φ	$\begin{pmatrix} 1.3 & 0.4 \\ 0.3 & 0 \\ -0.8 & 0.4 \\ 0.1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.3 \\ 0.4 & 1.3 \\ 0 & 0.1 \\ 0.4 & -0.8 \end{pmatrix}$	$\begin{pmatrix} 1.3 & 0.3 \\ 0.3 & 1.3 \\ -0.8 & 0.1 \\ 0.1 & -0.8 \end{pmatrix}$
σ_x	1	1	1
%Gr(ind)	$\begin{pmatrix} 100.0 & 99.9 \\ 69.4 & 95.1 \end{pmatrix}$	$\begin{pmatrix} 93.9 & 6.2 \\ 97.5 & 100.0 \end{pmatrix}$	$\begin{pmatrix} 100.0 & 95.6 \\ 98.5 & 100.0 \end{pmatrix}$
%Gr(all)	65.5	5.2	94.4

Note: is the parameter matrix of the VAR $Y_t = \mu + (Y_{t-1} \quad Y_{t-2})\Phi + \varepsilon_t$, see equation (2). As in Table 4, the rows *Gr(ind)* and *Gr(all)* contain the share of Monte Carlo simulations with correctly identified Granger causalities (individual and combined), where the tested causalities are $\begin{pmatrix} x \to x & x \to y^* \\ y^* \to x & y^* \to y^* \end{pmatrix}$.

As in the benchmark scenarios, the test for unbiasedness of y^* the estimate of rejects the null hypothesis in the clear majority of our simulations. The average of R^2 the corresponding auxiliary regression (3) ranges from 0.61 to 0.75 when considering the median estimate, and from 0.47 to 0.62 for the Fry-Pagan estimate. This does more or less match the results obtained from the benchmark scenarios with σ_x=1.

In the *cycle o* and *cycle l* settings, causalities from one variable to the other are only weakly identified if the parameters are low. This problem is particularly severe in the *cycle l* setting, where the impact of the observable on the latent is merely detected in 5% of the simulations.[17] This confirms our findings from the benchmark scenarios.

The identification of the Granger causality by the Qual VAR is surprisingly strong in the case where both latent and observable variable exhibit (individual) cyclical behavior (*cycle all*). In this scenario the true Granger causality is correctly identified (at the 5% significance level) in almost 95% of our simulations. However, this is mostly due to the high variance of both the observable and the latent variable that is caused by the interaction of the cyclical behavior. Therefore, the relation of the variance of shocks to the total variance of the model variables (about 1:10) is much lower than in all our other simulations (about 1:4), and lower than in many empirical models. That is, this setting is particularly simple to identify.

B. Latent link variable

Our second robustness test applies the Qual VAR to a model where the latent variable links two observable variables:

$$Y_t = \mu + Y_{t-1}\begin{pmatrix} 0.2 & 0.5 & 0.7 \\ 0.5 & 0.2 & 0 \\ 0 & 0.7 & 0.2 \end{pmatrix} + \varepsilon_t. \tag{4}$$

This setting allows inference about the latent variable both from lead and lag observations of the observable variables, thereby closely resembling our original *olo* settings. However, there is no need to extract the lead and lag information about the latent variable from a single observable variable. This might simplify identification.

[17] Due to the multiple lag structure, we cannot apply our adjusted causality test that accounts for the sign of the effect. The values reported in this section are obtained from conventional Granger causality tests using a Wald type test.

Nevertheless, the test for unbiasedness of the estimate of the latent again rejects in the clear majority of simulations. The R^2 of the test regression (3) is slightly better than in the *olo* setting but in the same order of magnitude (0.67 for the median estimate and 0.53 for the Fry-Pagan estimate).

The share of simulations where all causalities are correctly identified is around 18%. While that is worse than the performance in the (*olo,eq*) setting, it has to be considered that a larger number of individual tests is aggregated, increasing the chance of missing once, even if the individual tests perform equally well. As in the (*olo,eq*) setting we find that most problems occur in the estimation of the autoregressive parameters with a value of 0.2. In particular, the autocorrelation of the latent variable is merely identified in one of three simulations.

VII. Conclusions

Our results on the performance on the Qual VAR are mixed. The forecasting performance is fairly good. Most notably, compared to a standard procedure in binary forecasting such as a probit, the Qual VAR generally adds substantially. Especially if the dynamic behavior of the latent variable is relevant, the Qual VAR is strong in that respect. Even in situations where the Qual VAR cannot play its strength (such as in short horizon forecasts in our *ol* settings), the loss compared to different benchmark models is moderate and the absolute forecast errors are minimal.

However, Qual VAR has severe problems in the identification of the economic story. Even when the Qual VAR is only confronted with rather simple models in our MCMC framework, it produces substantial errors when estimating the dynamics of the latent variable. Moreover — and at least as problematic from an economic perspective — is the Qual VAR's failure to capture the correct Granger causality in approximately 50% of our simulations. As the Granger causality essentially tells which setting prevails, it is difficult to identify the true setting (i.e., the general economic story). These results hold in more complex settings using more lags and more variables. Because the quality of the identification of the latent variable strongly depends on the setting, it is basically impossible to determine whether the Qual VAR results are reliable.

Thus, while providing a good forecasting tool, using the Qual VAR is inadvisable with respect to economic analysis.

References

Bordo, Michael D., Michael J. Dueker, and David C. Wheelock (2008). Inflation, monetary policy and stock market conditions. Working Paper 14019, NBER.

Casella, George and Edward I. George (1992). Explaining the Gibbs Sampler. *American Statistician* 46: 167–174.

Dueker, Michael J. (2005). Dynamic Forecasts of Qualitative Variables: A Qual VAR model of U.S. recessions. *Journal of Business and Economic Statistics* 23: 96–104.

Dueker, Michael J. and Katrin Assenmacher-Wesche (2010). Forecasting macro variables with a Qual VAR business cycle turning point index. *Applied Economics* 42: 2909–2920.

Dueker, Michael J. and Charles R. Nelson (2006). Business-cycle filtering of macroeconomic data via a latent business-cycle index. *Macroeconomic Dynamics* 10: 573.

Eichengreen, Barry, Mark W. Watson, and Richard S. Grossman (1985). Bank rate policy under the interwar Gold Standard: a dynamic probit model. *Economic Journal* 95: 725–745.

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

Fornari, Fabio and Wolfgang Lemke (2010). Predicting recession probabilities with financial variables over multiple horizons. ECB Working Paper 1255.

Fratzscher Marcel (2003). On currency crises and contagion. *International Journal of Finance & Economics* 8: 109–129.

Fry, Renee and Adrian Pagan (2007). Some issues in using sign restrictions for identifying structural VARs. NCER Working Paper 14.

Galvão, Ana Beatriz C. (2006). Structural break threshold VARs for predicting US recessions using the spread. *Journal of Applied Econometrics* 21: 463–487.

Hamilton, James D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.

Harding, Don and Adrian Pagan (2011). An econometric analysis of some models for constructed binary time series. *Journal of Business and Economic Statistics* 29: 86–95.

Hartmann, Philipp, Kirstin Hubrich, Martin Kremer, and Robert J. Tetlow (2012). Widespread financial instabilities and the macroeconomy-regime switching in the euro area? Mimeo.

Holden, K. and David A. Peel (1990). On testing for unbiasedness and efficiency of forecasts. *Manchester School* 58 120–127.

Kauppi, Heikki and Pentti Saikkonen (2008). Predicting US recessions with dynamic binary response models. *Review of Economics and Statistics* 90: 777–791.

Lemeshow, Stanley and David W Hosmer (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* 115: 92–106.

Marcellino, Massimiliano (2006). Leading indicators. *Handbook of Economic Forecasting* 1: 879–960.

Moneta, Fabio (2005). Does the yield spread predict recessions in the euro area? *International Finance* 8: 263–301.

Paap, Richard, Rene Segers, and Dick van Dijk (2009). Do leading indicators lead peaks more than troughs? *Journal of Business & Economic Statistics* 27: 528–543.

Ratcliff, Ryan (2013). The "probability of recession": Evaluating probabilistic and non-probabilistic forecasts from probit models of US recessions. *Economics Letters* 121: 311–315.