# Identifying Noise Shocks: a VAR with Data Revisions*

Riccardo M. Masolo† Alessia Paccagnini‡

February 14, 2013

**Abstract**

We set out to show how the use of different vintages of data delivers a simple identification strategy that allows to study the impact of data imperfections on the business cycle. Our findings suggest that an erroneous report of output growth numbers delivers a persistent and hump-shaped response of real output and unemployment. When we include investment in our estimation, we find that it displays a significant response to noise shocks too, while the responses of output growth and unemployment are still significant.

JEL CODES: E3, C10, D8
KEYWORDS: Noise Shocks, Data Revisions, VAR, Impulse-Response Functions

*We are particularly grateful to Larry Christiano, Giorgio Primiceri and Roberto Motto. We would also like to thank participants at the XXI International Conference on Money, Banking and Finance. Riccardo M. Masolo thanks Northwestern University as this project formed part of his dissertation and the European Central Bank as he worked on part of this project while a trainee there. The views and opinions expressed in this work are solely those of the authors and do not necessarily reflect those of the Bank of England or the European Central Bank.

†Bank of England, Email: Riccardo.Masolo@bankofengland.co.uk

‡University of Milano-Bicocca, Department of Economics, E-mail : alessia.paccagnini@unimib.it

# 1 Introduction

The growing popularity of dispersed-information DSGE's make the empirical study of noise shocks more relevant than ever as an instrument to assess the relevance of models.

While the statistical properties of data revisions have been widely documented (see Makinw and Shapiro (1986) and Arouba (2008) among others), their economic implications for the business cycle have not received much attention.

Our aim is to cast some light on this, with a simple econometric model. We assume that early vintages of data are a noisy (or at least preliminary) version of the final release, which we assume to correspond to the truth. While the ongoing data revision processes might legitimately raise doubts on the latter assumption, we nonetheless see it as an extremely useful benchmark, as basic econometric analysis would simply assume the latest release to be the true one, disregarding past as well as future revisions altogether

We do not make any model-specific assumption although one is implicit. Our analysis would not make sense in a world (or model) in which true values for the relevant economic variables are known with certainty immediately or at a very short lag. Noise shocks have no reason to produce any economic effect once all the information is revealed[1]. In this sense, the environment

[1] A recent work by Del Negro and Schorfheide (2012) discusses the news and noise assumptions on nowcast data and its impact on forecasting performance. On the one hand, it is interesting as it is an application to widely used New Keynesian models with no particular informational friction. On the other hand, the nowcast is rationalized in terms of a particular sequence of structural shocks (which do not include a noise shock). In other words, their goal is not the propagation of the noise shock to the economy but rather making the starting condition for forecasting as accurate as the econometrician information set allows her.

we have in mind is broadly consistent with dispersed information models such as Lorenzoni (2009), Mendes (2007) and Masolo (2011). As we just said, our assumptions are consistent with standard dispersed-information DSGE's and yet we do not to impose a particular structure ex ante. We see this as providing a more robust and data-driven framework that lets the data speak.

In sum, our main assumption is a timing assumption. We maintain that the noise shock only impacts true values with a lag, through the decision making of agents who respond to the noise-ridden indicators they have at their disposal, pending the publication of more accurate figures. This paper is devoted to showing how this simple identification scheme delivers interesting business cycle effects of what we call revision or noise shocks.

As it turns out, output, unemployment and investment display statistically significant responses to a noise shock which can be thought of as a release of output growth number that does not reflect actual output growth.

Our work is related to those in Rodriguez-Mora and Schulstad (2007)[2] and Oh and Waldman, 1990 and 2005). In particular, Rodriguez-Mora and Schulstad (2007) go a long way to show the effect of data revisions on output growth and investment.

For the time being we take a clear stand in assuming that early vintages of data impact the true underlying values through the decision process of agents. We are aware of the criticism raised by Clements and Galvão (2010)

[2] Kaukoranta (2009), expands Rodriguez-Mora and Schulstad (2007) on a longer series and more countries.

who argue that the casual effects of early announcement of output growth numbers on future growth might not necessarily signal a behavioral relationship as Rodriguez-Mora and Schulstad (2007) posit but might be the by product of the specific statistical process for data revisions.

While we might consider this insight for future developments of our analysis[3], we see our current work as a simple and intuitive benchmark. Clements and Galvão (2010) need make assumptions on the revision process while we do not need to model it. Moreover, their analysis does not rule out the possibility of a behavioral relationship, rather it highlights the possibility that the data revision process itself might generate a specific correlation pattern between different releases.

Our exercise draws from general equilibrium models in which the casual relationship is straightforward so we follow this strategy. Relative to Rodriguez-Mora and Schulstad (2007), we see our methodology to be more general as it presents a simple VAR identification scheme which can be extended in numerous ways. Moreover, we focus our attention on impulse responses rather than predictive power. In this respect, we see our work as some sort of a bridge between the papers we just cited, which tend to be geared towards a statistical analysis and recent works like Lorenzoni (2009) and Blanchard, L'Hullier and Lorenzoni (2012).

The former, identifies all non-technology disturbances as noise shocks, an extreme assumption which is entertained to show the ability of the model to match even the most extreme case. The latter tries to tell noise from news

[3] Clements and Galvão (2010b) develop specific techniques for VARs with data revisions, which however they use primarily for prediction purposes.

shocks without resorting to data revisions, i.e. not exploiting the informational advantage the econometrician has relative to economic agents.

Indeed, our baseline VAR is, rather similar to that in Lorenzoni (2009)[4], comprising the final releases of output growth and unemployment together with the first release of output growth, which allows us to pinpoint the impact of noise shocks on real variables.

We believe that one of the main benefits of VAR analysis is its ability to isolate the part of the revision which is orthogonal to the variables of interest[5]. In this sense, VARs are a natural econometric counterpart to dispersed information DSGE's but their flexible setup might actually help correct for the fact that real-world data revisions do not necessarily correspond to noise shocks in models[6].

The rest of the paper presents an overview of the state-space representation of dispersed information models, a discussion of our econometric setup and our identification strategies. We then discuss results from the estimation of our baseline VAR as well as an alternative specification which includes a measure of investment on top of output growth and unemployment.

## 2 Noise Shocks in Dispersed Information Models

A recent series of dispersed information general equilibrium models, e.g. Lorenzoni (2009), Mendes (2007) and Masolo (2011), provide the ideal the-

[4] We use unemployment instead of hours and consider output growth as opposed to levels. Obviously we also include the first release of output growth which does not enter Lorenzoni (2009) VAR specification.

[5] Rodriguez-Mora and Schulstad (2007) have an ad-hoc equation to isolate what they call surprises.

[6] We expand on this ides when we discuss our identification strategy.

oretical foundation to study noise shocks.

In standard full information DSGE's, information about the past is irrelevant: the agents know the current state of the economy, hence will not respond to any noisy information about the past[7].

In reality, however, people do not perfectly know the state of the economy so they take advantage of published data about, say, GDP growth in recent quarters. The very fact that such series get revised shows that those numbers are not fully accurate (especially for the most recent periods), yet they contain useful information for the economic agents.

Dispersed information models capture exactly this. Because agents do not fully know the state of the economy, they will respond to an informative, albeit noisy, signal about the state of the economy improves the precision of their predictions.

The precision of the signal will impact the size of the response but will not prevent agents from responding to noise. The impact of the noise embedded in the signal will only die out as agents learn about the true fundamentals[8].

Models such as those sketched above can typically be readily cast in a state-space[9] form in which at least the observation equation is household specific

[7] This is somewhat symmetric to the news literature which is concerned with information about future states of the world.

[8] The speed of learning is obviously a matter of one's preferred calibration in a model. One of the benefits of our analysis is to cast light on the time span over which these effects are statistically significant.

[9] It is usually the case that more lags of the state variables are needed to solve a dispersed information model. Typically they are stacked to form a first-order system.

(as denoted by the h subscript):

$$\mathbf{Z}_t = \Psi_1 \mathbf{Z}_{t-1} + \Psi_0 \underline{u}_t \quad (1)$$

$$\underline{\mathbf{s}}_{ht} = \Gamma_1 \mathbf{Z}_t + \Gamma_0 \underline{\zeta}_{ht} \quad (2)$$

Usually the noise shock will be a component of the vector $\underline{u}_t$[10]. A key characteristics is that the noise shock will impact a signal observable by all the agents in the economy[11].

This is realistic if one has in mind something like early releases of output growth figures which are available to everyone. Moreover, it is convenient because it does not integrate out across households but has an impact on economy-wide variables; as a result it lends itself to simple time-series analysis.

It is realistic to assume that agents will only receive signals about aggregate variables at the end of each period[12], i.e. once the aggregate variable of interest has materialized.

As a result the following timing pattern arises naturally:

1. The noise shock hits the economy.

2. At the end of the period the economy-wide signal, e.g. early release of output growth, is affected by the noise shock.

3. At the beginning of the following period the noise will impact the economic decisions of the agents.

[10] In principle the noise component might be itslef autocorrelated, in which case it will enter the state vector $\mathbf{Z}_t$, while $\underline{u}_t$ will include the innovation to that same process.

[11] The row of $\Gamma_0$ corresponding to the early-data release will be all zero.

[12] Or equivalently at the beginning of the following period

4. Over time agents will gradually learn that the signal was driven by a noise shock as opposed to a fundamental and revise the expectations and decisions accordingly.

This results in set of zero-restrictions on the matrix $\Psi_0$. In particular if the economy-wide signal we are concerned with (early data release) is the j-th entry in $\mathbf{Z}_t$ and the noise shock is the i-th entry in $\underline{u}_t$ than the i-th column of $\Psi_0$ will comprise all zeros except on row j.

A crucial aspect of this class of models, though, is that the characteristics of the noise embedded in those signals has to be specifically defined and calibrated. In particular, it is usually the case that the noise shock is the same as the revision implied by the model.

On the other hand, estimating a VAR we do not we do not need to make ad-hoc assumptions on the noise shock. We only use the timing restriction highlighted above and we show in the next two sections which assumptions on data revisions are consistent with our identification scheme. As it turns out, our strategy is robust to some deviations from the basic assumption that data revisions and noise shocks are the same.

# 3 Setup

We now illustrate how our assumption fares when faced with the most commonly held assumptions about data revisions. In the next paragraph we illustrate the case in which the early vintage of the date is the sum of true value plus a noise component, which the assumption usually maintained in theoretical models.

The following section shows how our estimation is also consistent with the case in which the economic agency releasing the data tries to optimally forecast the actual series, so that the revision is no longer orthogonal to the fundamental value of the variable at hand.

### 3.1 Classical Noise

For the sake of the discussion let us assume that only two vintages of data are available. We will refer to the first release as $x_t^0$ and to the latest as $x_0^f$explain the different cases for the latest.

For the moment let us entertain the assumption that the early vintage of data equals the latest (i.e. true or fundamental) plus a noise shock:

$$x_t^0 = x_t^f + v_t \tag{3}$$

A number of empirical papers discuss and test the validity of this assumption on data revisions (e.g. Arouba (2008)). We start off with it because it is the one commonly used in modeling[13] and leave an alternative scenario to the next paragraph.

Here, $x_t^f$ is understood as including variables set by the economic agents given their information set[14]. Their dynamics reflect the fact that agents respond to both fundamentals and (past) noise in the way highlighted above. Implicitly we are assuming that the revision process eventually converges to

[13] Consistent with our assumption, the first vintage of data we consider is the one published in the quarter following the one to which it refers. That way computation is made when the quarter has ended so that remaining imperfection should be more easily attributable to noisy data as opposed to a forecasting error.

[14] We do not tackle the issue of invertibility in this context, assuming it throughout, the aim of our work being primarily empirical.

the true underlying value, which the econometrician can observe while the agent in the model cannot (at least timely and perfectly).

In particular, appendix A shows that given the above mentioned state-space representation, the process governing $x_t^f$ can be expressed as:

$$x_t^f = A(L)x_{t-1}^f + B(L)v_{t-1} + \varepsilon_t \tag{4}$$

Where all the elements of equation (4) can be vectors and $A(L)$ and $B(L)$ are finite-order polynomials in the lag operator and $\varepsilon_t$ are the other shocks hitting the economy (which we are not trying to identify here).

Equations (3) and (4) define the evolution of the two set of variables we are interested in, namely the early and the latest vintages of data.

Combining the two defines our model for the early releases of economic data:

$$x_t^0 = A(L)x_{t-1}^f + B(L)v_{t-1} + \varepsilon_t + v_t \tag{5}$$

which shows how the revision component affects directly the early vintage of data and indirectly, i.e. through the decision-making process of the economic agents, the future values of fundamental variables.

That is consistent with the idea that when early vintages of data are released, decisions for the period at hand are sunk[15] but future decisions will respond to this noisy indicator of economic activity.

[15] This is certainly true in our case as we consider as first release the one published in the quarter following the one of interest.

### 3.2 Identification

The identification strategy hinges on the fact that the information set of the econometrician is richer than that of the economic agent who made the decision, because the econometric analysis is carried out at a later timeexpand on why we need to have some time after the end of the sample.
Rearranging equation (3) and substituting it into equation (5) yields:

$$x_t^0 = (A(L) - B(L))\, x_{t-1}^f + B(L) x_{t-1}^0 + \varepsilon_t + v_t \tag{6}$$

Using equation (6) to substitute for $x_t^0$ in equation (3) delivers a VAR representation of the final observations:

$$x_t^f = (A(L) - B(L))\, x_{t-1}^f + B(L) x_{t-1}^0 + \varepsilon_t \tag{7}$$

Stacking up the early and the latest vintages produces the following VAR:

$$\begin{bmatrix} x_t^f \\ x_t^0 \end{bmatrix} = \begin{bmatrix} A(L) - B(L) & B(L) \\ A(L) - B(L) & B(L) \end{bmatrix} \begin{bmatrix} x_{t-1}^f \\ x_{t-1}^0 \end{bmatrix} + C \begin{bmatrix} \varepsilon_t \\ v_t \end{bmatrix}$$

As usual matrix, C maps fundamental shocks into observed residuals so it has to satisfy:

$$CC' = E\left[\underline{w}_t \underline{w}_t'\right]$$

Where $w_t = [\varepsilon_t \; v_t]'$. What we call noise shocks are identified by the fact that they contemporaneously affect only the early release of data. As described in the previous section, noisy components enter the decision making process of the imperfectly informed agents but they do so with a lag because when early numbers concerning period $t$ are made public, decision for that period have been already made.

Suppose that the n-th element of our VAR is the early release of output growth. Our identifying assumption implies that the n-th column of C will have zero entries corresponding to final releases and a non-zero entry in row n[16].

### 3.3 Prediction Error

A vast empirical literature shows that revisions for some series are better characterized as resulting from forecasting errors made by the agency which publishes early releases, e.g. Mankiw and Shapiro (1986).

The key difference with respect to the case illustrated above is that the revision is not orthogonal to the final release, which makes the revision not a suitable candidate for a noise shock.

In this paragraph we illustrate how our VAR procedure actually mitigates this problem as what we call noise shock is *not* the revision of data vintages, but that part of the revision which is orthogonal to the final release.

An exhaustive discussion of this issue would require the knowledge of the prediction models used by the agencies which publish early vintages of data. Since that is not the case, we will proceed with an example and discuss how our procedure is robust to a simple statistical model. In particular, this example will illustrate to which extent our procedure is robust to different models used by the data-producing agency.

[16] This is consistent with our assumption on $\Psi_0$.

Let us assume that a statistical agency receives a noisy signal on the true underlying economic variable which takes on the following form:

$$x_t^{00} = x_t^f + v_t$$

The key difference with respect to the case above is that now the agency correctly anticipates that the data they collect are noise ridden (e.g. because they only collect a sample of the data of interest) and so perform a filtering procedure before making them public. In particular, it is reasonable to assume that they will consider the linear projection of the true underlying variable onto the known signal so that the early release would take on the following form:

$$x_t^0 = P[x_t^f | x_t^{00}] = \phi x_t^{00} = \phi x_t^f + \phi v_t$$

Where the projection coefficient $\phi$ depends on the relative variance of noise in the signal $x_t^{00}$.

The key difference, relative to the case above is that now the data revision is not orthogonal to the final release, in fact:

$$x_t^f - x_t^0 = (1 - \phi) x_t^f + \phi v_t$$

As a consequence, one would be incorrect in taking the revision as an indicator of the noise shock. However, our VAR strategy provides a simple fix to this.

Under the maintained assumption that economic agents in the model know the data-generating process, the newly defined early release would simply result in a different observation equation but would otherwise not change the model which could be summed up as:

$$x_t^f = \tilde{A}(L)x_{t-1}^f + \tilde{B}(L)v_{t-1} + \varepsilon_t$$

Where different filters $\tilde{A}(L)$ and $\tilde{B}(L)$ reflect the different observation equation.

Following the same steps as above, while using the new definition of early release, one gets the following formulas for the early and the latest vintages of data:

$$x_t^0 = \phi(\tilde{A}(L) - \tilde{B}(L))x_{t-1}^f + \tilde{B}(L)x_{t-1}^0 + \phi\varepsilon_t + \phi v_t \tag{8}$$

$$x_t^f = (\tilde{A}(L) - \tilde{B}(L))x_{t-1}^f + \frac{1}{\phi}\tilde{B}(L)x_{t-1}^0 + \varepsilon_t \tag{9}$$

Despite the scaling factor $\phi$ showing up in the equations and different lag polynomials reflecting the fact that equilibrium responses will in general be different under this alternative scheme, it is still the case that the noisy component $v_t$ contemporaneously only the early release while not the final, thus being consistent with the identification strategy laid down above. Not only that, but this analysis suggests that the resulting noise shock is the share of noise $\phi$ which is not filtered out by the statistical agency. In other words, it is the chunk of noise that enters the economy through the informational set of agents[17].

[17] It does not appear that $\phi$ should drive the roots of the polynomial over one, because it shows up in the term in $x_{t-1}^f$ in the equation for $x_t^0$ and vice versa, in the equation for

The example above illustrates a situation in which taking the data revision naively would lead to an incorrect assessment of the noise shock because the revision incorporates a component which is not orthogonal to the true value $x_t^f$. The VAR however cleanses the revision of the component that depends on $x_t^f$.

While the example assumes a very simple information set of the statistical agency it casts light on the benefits of our strategy, which whitens revisions so that we can call noise shock the component of the revision which is orthogonal to the variables included in the VAR.

In fact, the only possible problem with this strategy appears to be in the number of variables and lags included in the VAR. In abstract, since the agents in the model know the data generating process, any variable, or lag thereof, used by the agency would be included in the state equation. In practice, since we do not know the information set and the procedures of the statistical agency, we rely on the standard lag-selection tests to gauge whether our statistical model appears to be correctly specified.

So, while the limited number of data points curtails the number of series and lags we can realistically include in our estimation, we find that using a VAR is the correct approach to isolate noise shocks as it orthogonalizes data revisions relative to the variables included in the estimation (this is somewhat related to Rodriguez-Mora and Schulstad (2007)).

$x_t^f$. In any case, in the limit as $\phi \to 1$ which occurs when the agency has more precise information, the effect of $\phi$ becomes smaller and smaller.

# 4 VAR

## 4.1 Baseline

When it comes to deciding which variables should enter our VAR specification we start off with a parsimonious set, in part motivated by the fact that using different vintages of data increases the number of parameters to estimate on a relative small time series and in part because, as we said above, we would like to contribute to the evidence provided in Lorenzoni (2009). Different vintages of data are available only from the mid 1960's and we also want to leave a sufficiently long period after the end of the sample so that we can be reasonably confident that the bulk of the revisions has ended by the time we carry out are analysis. For this reason we limit our sample to 2006.

In a similar fashion to Lorenzoni (2009) we include the final releases of output growth and unemployment (as part of $x_t^f$), adding the early release of output growth numbers as part of $x_t^0$.
This is a simple enough setting and yet allows to study the impact of noise shocks on two crucial variables without assuming that all non-permanent effects are noise related as in Lorenzoni (2009).
Our estimation equation then reads:

$$\begin{bmatrix} \Delta y_t^f \\ u_t^f \\ \Delta y_t^0 \end{bmatrix} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \begin{bmatrix} \Delta y_{t-1}^f \\ u_{t-1}^f \\ \Delta y_{t-1}^0 \end{bmatrix} + C \begin{bmatrix} \nu_t^1 \\ \nu_t^2 \\ \nu_t^3 \end{bmatrix}$$

What we are interested is the identification of the third column of matrix C which, consistent with the discussion above, will read:

$$C_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Which implies that the our empirical noise shock $\nu_t^3$ will be orthogonal to all the variables included in the VAR except the current value of the first release (see the Appendix B for details).

#### 4.1.1 Data Description

We consider the real GDP and the unemployment rate from the Historical Data Files for the Real-Time Data Set provided by the Federal Reserve Bank of Philadelphia. The quarterly vintages and quarterly observations of the Real GNP/GDP (ROUTPUT) is in Billions of real dollars, and seasonally adjusted. We take the first difference logarithmic transformation, so we consider it as a quarterly (annualized) growth rate[18]. Instead, the quarterly vintages and monthly observations of the Unemployment Rate (RUC) is in percentage points, seasonally adjusted. We transform our data from monthly to quarterly frequency considering the first observation of the quarter.

We consider the quarterly sample from 1966:1 to 2006:4. This is so that we allow for a sufficiently long window for revisions even for the end-of-sample

[18] Using growth rates is motivated not simply by non-stationarity consideration but also by the fact that, as Rodriguez-Mora and Schulstad (2007) point out, it is easy to account for big long-term data revisions in growth rates (because typically affect one value which we substitute with the average of the previous and the following quarter) than in level, because in this case the effect of the revision is essentially permanent.

observations. In fact, our final releases are those published in the third quarter of 2011[19] so we allow for about five years worth of revisions even for the data at the end of the sample. For the first release, on the other hand, we considered that derived from output level numbers published one quarter after the period of interest[20]vertical difference of diagonal difference?. That way it seems safe to consider that the agency had some time to collect data, reducing the forecasting component in the release and it also guarantees that such a number cannot affect the decisions of agents in the current period.

#### 4.1.2 Results

Figures 1 and 2 report the responses of final output (in log-levels) and unemployment to a revision shock.

First, both of them are significant. Output appears to be higher than it would otherwise be for several quarters, while unemployment is significantly below its long-run level for around three years.

Interestingly, both responses build over time in a hump-shaped fashion which is consistent with learning in models with dispersed information.think about the scale

Moreover it should be noticed that, not only the growth-rate of output converges back to zero but its log-level does as well, which is consistent with

[19] Clements and Galvão (2010) entertain both the definition of final release as the latest available or that occurring a fixed number of quarters after the end of the period of interest (in their case 14 quarters, seeming to favor the latter because it is less affected by long-term revisions. On the other hand, Rodriguez-Mora and Schulstad (2007) seem to favor our approach. In any event, we find our approach a sensible benchmark because the standard counterpart of our VAR would be one in which the latest releases available are used, not those published a certain fixed number of quarters after the end of the period of interest.

[20] The computer code we used to elaborate raw data can be requested to the authors.

the idea that while noise shocks can be expected to produce variability at business cycle frequency, no long-run effects on output seem likely.

Finally, the sign of the responses deserves some attention as well. Firstly, because the shock at hand is not technological in nature - it is a demand shock in the spirit of Lorenzoni (2009) - it is natural to expect that output and unemployment be negatively correlated. And so they turn out to be. [inline]discuss the sheer size of the shock and responses if possible

## 4.2 Alternative VAR Setup

Rodriguez-Mora and Schulstad (2007) suggest that investment is a crucial variable when considering the impact of data revisions, which is reasonable given the forward-looking nature of investment decisions.

Long-term projects, such as investment plans tend to be, are more susceptible to data imperfections as they necessarily have to rely on forecasts of future conditions. On top of that, investment decisions are costly to reverse, once undertaken.

Adding a measure of investment in our VAR we want to address two main points. First, we are interested in verifying if investment exhibits a significant response, somewhat along the lines of Rodriguez-Mora and Schulstad (2007). Secondly, introducing investment we can verify if the response of output to a noise shock is significant even when a measure of investment is included in the VAR.

Finally, adding an extra regressor further "cleanses" our definition of noise shock for potential correlations with variables which could enter the datapublishing agency's information set. In particular, we make our noise shock orthogonal to our measure of lagged and current investment as well.

Our definition of investment is similar to that in Altig, Christiano, Eichenbaum, Linde (2004). It is the log of the ratio of investment to GDP (in this case the final value of GDP). In other words, it is a measure of the investment-rate as a share of output.

In particular, the investment is given by the logarithmic transformation of the ratio between the sum of Gross Private Domestic Investment (GPDI) and Personal Consumption Expenditures in Durable Goods (PCDG) and the Gross Domestic Product, 1 Decimal (GDP). All the quarterly-observation variables used to build the series for investment were taken from the FRED Database of the Federal Reserve Bank of St. Louis.

### 4.2.1 Results

Results if Figure 3, 4 and 5 show that, once more, responses to revision shocks appear to be significant at business cycle frequencies.

Interestingly, the size of the responses of output and unemployment is similar to that in the baseline setup we considered above, which appears to rule out the possibility that the response of output had to do with the omission of investment from the setup.

At the same time, the investment rate is higher than average for about two years following an "overly optimistic" early release of output growth numbers.

In this respect, it should be noticed that it is not simply investment per se that goes up, but investment as a share of output. Because output itself grows after a positive noise shock, this suggests that the level of investment increases in response to a noise shock more than output, consistent with basic business cycle facts that show how investment is positively correlated but more volatile than output (see King and Rebelo (2000)).

Finally, the sheer size of the responses appears to support the idea that noise shocks produce real effects even when cleansed from any linear correlation of the revision with lagged and current investment rates.

### 4.3 Variance Decomposition

While the significant responses of output, unemployment and investment to a noise shock suggest that revisions can play an important role in the behavior of macro variables, the share of the variance explained by noise shocks helps assess the relative magnitude of revision shocks.

Figures 6, 7 and 8 display a dynamic variance-decomposition exercise. The charts show the share of the variance explained by revision shocks at different points into the futurecheck that the description makes sense. In other words, they report the share of forecast variance which can be traced back to revision shocks.

The share is necessarily zero in the first period because of our identification restriction and then tends to grow leveling off just below 5 percent for what concerns output growth, around 7 percent for unemployment and 6 percent for investment.

These numbers help define the scope of noise shocks in a more precise manner compared to Lorenzoni (2009), who took a very conservative stance in assuming that all shocks that are not technological (and hence do not have a long-run effect in his setting) are to be regarded as potential noise shocks. Here, on the other hand, we can more precisely assess the relative importance of revision shocks and we find the numbers sensible for two reasons. Firstly, they noise shock appear to represent a non-negligible share of the variability of key business cycle variables. Secondly, the share is small enough to be reconciled with common wisdom. Indeed, if noise shocks would explain, say, a third of the variation in output growth one could doubt the soundness of the analysis. It seems obvious that only a relatively small share of the variance of economic variables can be explained by noise shocks. At the same time, a shock explaining 7 percent of the variance of unemployment over the business cycle deserves attention.

Furthermore, Figures 6 and 7 also confirm that including investment share in the specification of the VAR does not dramatically impacts the ability of the VAR to explain the behavior of unemployment and output growth.

## 5 Conclusion

We set out to show how the use of different vintages of data delivers a simple identification strategy that allows to study the impact of data imperfections on the business cycle. Our findings suggest that an erroneous report of output growth numbers delivers a persistent and hump-shaped response of real output and unemployment. When we include investment in our estimation, we find that it displays a significant response to noise shocks too, while the responses of output growth and unemployment are still significant.

Our analysis does not rely on a specific economic model although it is consistent with standard assumptions made in dispersed information models. At the same time, it acknowledges the fact that in the real world the revision of data need not be orthogonal to state variables as is usually the case in models (e.g. Mendes (2007)). That is why what we call noise shock is not the revision per se but a shock made orthogonal to the final-release variables included in the VAR.

As discussed in the introduction, we see our work as bridging a gap between more statistically-oriented papers and model-based analysis as that in Lorenzoni (2009).

In this respect, it is particularly useful to note that our estimation exercise delivers an estimate of the share of the variance of the economic variables that is explained by revisions, thus improving on the identification strategy employed in Lorenzoni (2009), which considered all non-permanent effects as noise-related. Our VAR suggests that between 4 and 5 percent of the variance of output growth can be attributed to noise shocks, with slightly higher shares of the variances of unemployment and investment being due to revision shocks. We find these shares reasonable, as revision shocks are certainly not the main driving force of the business cycle but, at the same time, turn out to be non-negligible.

In future developments we might consider refining the statistical methods used to treat different data vintages, possibly adopting those proposed by Clements and Galvão (2010, 2010b). In any event, we believe that using the latest revision is an important benchmark in that it is the series that one would normally use if he was not concerned with data revisions.

## 6 References

[1] Altig, D., Christiano, L., Eichenbaum, M., and Linde, J. Firm-specific capital, nominal rigidities and the business cycle. Working paper (Dec 2004), 1(49).

[2] Arouba, S. B. Data revisions are not well behaved. Journal of Money, Credit and Banking (Mar 2008), 22.

[3 Barsky, R., and Sims, E. News shocks and business cycles. Journal of Monetary Economics (May 2011), 1(34).

[4] Barsky, R. B., and Sims, E. R. Information, animal spirits, and the meaning of innovations in consumer confidence. American Economic Review 102, 4 (Jun 2012),1343(1377).

[5] Blanchard, O., L'Hullier, J., and Lorenzoni, G. News, noise and fluctuations:An empirical exploration. Revise and Resubmit, American Economic Review (May 2011), 1(40).

[6] Clements, M. P., and Galvão, A. B. First announcements and real economic activity. European Economic Review 54, 6 (Jul 2010), 803(817).

[7] Clements, M. P., and Galvão, A. B. Vector autoregressive models of data vintages for us output growth and inflation. Working paper (Oct 2010), 1(47).

[8] Del Negro, M., and Schorfheide, F. DSGE Model-Based Forecasting, prepared for Handbook of Economic Forecasting, Volume 2 (2012).

[9] King, R., and Rebelo, S. Resuscitating real business cycles. NBER Working Paper 7534 (Jan 2000).

[10] Kaukoranta, I. Do Measurement Errors in GDP Announcements

Cause Output Fluctuations? University of Helsinki, mimeo (2009).

[11] Lorenzoni, G. Optimal monetary policy with uncertain fundamentals and dispersed information. Review of Economic Studies (Feb 2009), 1(49).

[12] Lorenzoni, G. A theory of demand shocks. The American Economic Review (Jan2009).

[13] Mankiw, G., and Shapiro, M. News or noise? an analysis of gnp revisions. NBER Working Paper (Oct 1986), 24.

[14] Masolo, R. M. Wage Setting in a Dispersed Information DSGE. Northwestern University, mimeo (2011).

[15] Mendes, R. Information expectations and the business cycle. Mimeo (Dec 2007),33.

[16] Mora, J. R., and Shulstad, P. The effect of gnp announcements on fluctuations of gnp growth. European Economic Review (Nov 2007), 1(19).

[17] Oh, S., and Waldman, M. The macroeconomic effects of false announcements. The Quarterly Journal of Economics (May 1990), 1(19).

[18] Oh, S., and Waldman, M. The index of leading economic indicators as a source of expectational shocks. Eastern Economic Journal (Mar 2005), 1(21).

# 7 Figures

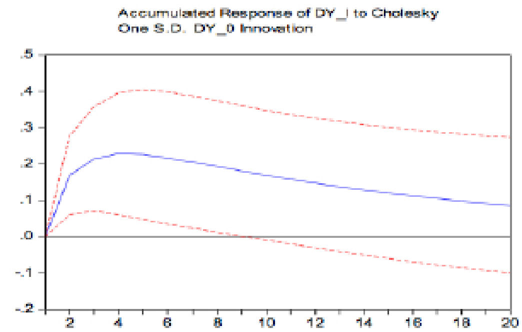


Figure 1.1. Impulse response (with confidence bands) of output to a one-stdev revision shock

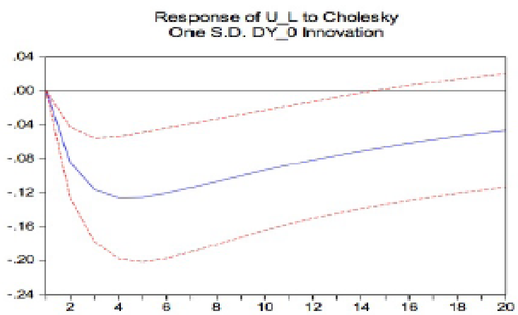


Figure 1.2. Impulse response (with confidence bands) of unemployment to a one-stdev revision shock

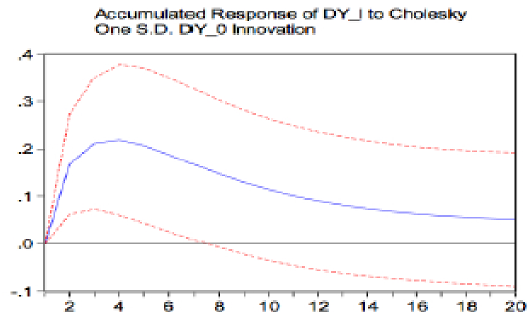


Figure 1.3. Impulse response (with confidence bands) of output to a one-stdev revision shock when investment is included in the VAR

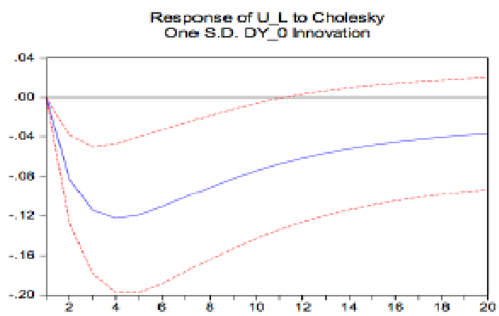


Figure 1.4. Impulse response (with confidence bands) of unemployment to a one-stdev revision shock when investment is included in the VAR

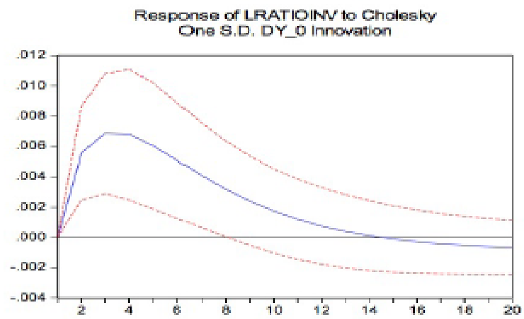


Figure 1.5. Impulse response (with confidence bands) of the investment ratio to a one-stdev revision shock

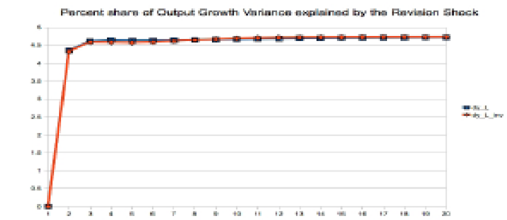


Figure 1.6. Variance share of Output Growth due to the revision shock in both the Baseline (squares) and the Alternative (triangles) setups.

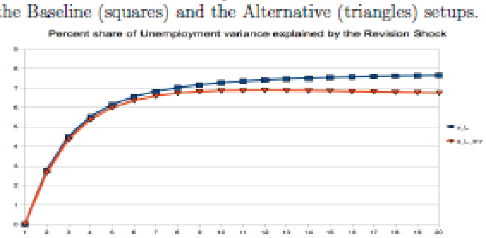


Figure 1.7. Variance share of Unemployment due to the revision shock in both the Baseline (squares) and the Alternative (triangles) setups.

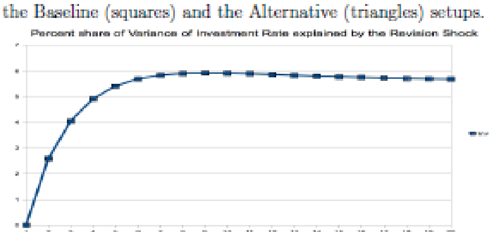


Figure 1.8. Variance share of the Investment Ratio due to the revision shock

# A Derivation of the VAR from the state-space representation

We now show how the VAR specification we employ relates to the state-space representation in which dispersed-information models are usually cast in.

We will try to keep it general, although obviously there tend to be multiple ways to write the state-space representation of a model which might change the algebra, although the substance of the model would be the same.

Throughout the derivation I will maintain the assumption that $\mathbf{Z}_t$ is defined by stacking up multiple lags of period-by-period state which is assumed to include $x_t^f$ and $x_t^0$.

## A.1 Derivation of the Law of Motion for the final release

First define $\Xi^F$ and $\Xi^u$ such that:

$$x_t^f = \Xi^F \mathbf{Z}_t \tag{10}$$

$$\underline{u}_t = \begin{bmatrix} \Xi^u \underline{u}_t \\ v_t \end{bmatrix} \tag{11}$$

Given the state-space representation we then have:

$$\begin{aligned}
x_t^f &= \Xi^F \Psi_1 \mathbf{Z}_{t-1} + \Xi^F \Psi_0 \underline{u}_t && (12)\\
&= \Xi^F \left( \sum_{l=1}^{s} \Psi_1^{F,l} x_{t-l}^f + \Psi_1^{0,l} x_{t-l}^0 \right) + \Xi^F \Psi_0 \underline{u}_t && (13)\\
&= \Xi^F \left( \sum_{l=1}^{s} \Psi_1^{F,l} x_{t-l}^f + \Psi_1^{0,l} x_{t-l}^0 \right) + \Xi^F \Psi_0 \underline{u}_t && (14)\\
&= \Xi^F \left( \sum_{l=1}^{s} \Psi_1^{F,l} x_{t-l}^f + \Psi_1^{0,l} x_{t-l}^0 \right) + [\Xi^F \Psi_0]_{\forall m \neq i} \Xi^u \underline{u}_t + [\Xi^F \Psi_0]_i u_t && (15)
\end{aligned}$$

Where $s$ is the number of lags stacked in the state vector, $\Psi_1^{F,l}$ and $\Psi_1^{0,l}$ refer to the column of $\Psi_1$ multiplying the l-th lag of the $x^f$ and $x^0$ respectively and $[.]_i$ refers to the i-th column of such matrix.
Given our zero-restriction assumption on $\Psi_0$:

$$[\Xi^F \Psi_0]_i = \Xi_j^F$$

The j-th column in $\Xi^F$. However, the j-th element in the state vector is, by our assumption, $x_t^0$ which will not be selected by $\Xi^F$ so $\Xi_j^F = \underline{0}$.
Using that and defining $\varepsilon_t \equiv [\Xi^F \Psi_0]_{\forall m \neq i} \Xi^u \underline{u}_t$, i.e. the rotation of fundamental shocks we are not trying to identify in our VAR, delivers:

$$\begin{aligned}
x_t^f &= \Xi^F \left( \sum_{l=1}^{s} \Psi_1^{F,l} x_{t-l}^f + \Psi_1^{0,l} x_{t-l}^0 \right) + \varepsilon_t && (16)\\
&= \Xi^F \left( \sum_{l=0}^{s} \Psi_1^{F,l} x_{t-1-l}^f + \Psi_1^{0,l} x_{t-1-l}^0 \right) + \varepsilon_t && (17)\\
&= \mathcal{R}(L) x_{t-1}^f + \mathcal{S}(L) x_{t-1}^0 + \varepsilon_t && (18)
\end{aligned}$$

Which corresponds to equation (7) given the appropriate matrix definitions.

Now, using the definition of early release[21] we get:

$$
\begin{aligned}
x_t^f &= \Xi^F \left( \sum_{l=0}^{s} \Psi_1^{F,l} x_{t-1-l}^f + \Psi_1^{0,l} (x_{t-1-l}^f + v_{t-1-l}) \right) + \varepsilon_t && (19) \\
&= \Xi^F \left( \sum_{l=0}^{s} (\Psi_1^{F,l} + \Psi_1^{0,l}) x_{t-1-l}^f + \Psi_1^{0,l} v_{t-1-l} \right) + \varepsilon_t && (20) \\
&= (\mathcal{R}(L) + \mathcal{S}(L)) x_{t-1}^f + \mathcal{S}(L) v_{t-1} + \varepsilon_t && (21)
\end{aligned}
$$

Which is the same as equation (4) when $A(L)$ and $B(L)$ are defined accordingly.

# B Orthogonality of the revision shock

The following paragraph illustrates the benefits of using a VAR procedure to study revision shocks. In particular it will show that the revision shock resulting from our analysis is a reasonably close proxy to the classical noise shock employed in models.

We will illustrate the point for our baseline specification, the extension to the alternative specification being essentially identical.

[21] At the modeling stage it does not qualitatively matter whether $x_t^0 = x_t^f + v_t$ or $\phi x_t^f + \phi v_t$ as it would just rescale the matrices so the derivation would be the same.

In particular, let us start with the VAR specification:

$$\begin{bmatrix} \Delta y_t^f \\ u_t^f \\ \Delta Y_t^0 \end{bmatrix} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \begin{bmatrix} \Delta y_{t-1}^f \\ u_{t-1}^f \\ \Delta y_{t-1}^0 \end{bmatrix} + C\underline{w}_t$$

Where the identification assumption delivers the following definition of fundamental shocks:

$$\begin{bmatrix} c_{11} & 0 & 0 \\ c_{21} & c_{22} & 0 \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} \nu_t^1 \\ \nu_t^2 \\ \nu_t^3 \end{bmatrix} = \begin{bmatrix} w_t^1 \\ w_t^2 \\ w_t^3 \end{bmatrix}$$

First note that basic projection theory implies that $w_t^1,\ w_t^2,\ w_t^3$ are orthogonal to any of the RHS variables, namely $\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0$, and in fact all other lags that could be included as well as the constant when variables are not demeaned.

Then obviously[22]:

$$Cov(x, w_t^1) = 0 \ \Rightarrow \ Cov(x, \nu_t^1) = 0 \quad x = \{\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0\}$$

[22] The maintained assumption is that all the $c_{ij}$'s are nonzero, which is realized in standard VARs because the covariance matrix of residuals will not have any zeros.

Moreover:

$$Cov(x, w_t^2) = 0,\ Cov(x, \nu_t^1) = 0 \quad and \quad Cov(x, w_t^2) = c_{21}Cov(x, \nu_t^1) + c_{22}Cov(x, \nu_t^2)$$

$$\Downarrow$$

$$Cov(x, \nu_t^2) \quad = \quad 0 \quad x = \{\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0\} \tag{22}$$

And:

$$\begin{aligned} Cov(x, w_t^3) = 0,\ Cov(x, \nu_t^1) &= 0,\ Cov(x, \nu_t^2) = 0 \\ and\ Cov(x, w_t^3) = c_{31}Cov(x, \nu_t^1) &+ c_{32}Cov(x, \nu_t^2) + c_{33}Cov(x, \nu_t^3) \\ &\Downarrow \\ Cov(x, \nu_t^3) &= 0 \quad x = \{\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0\} \end{aligned} \tag{23}$$

So far we have showed that the noise shock as we defined it ($\nu_t^3$) is orthogonal to all the variables included in the RHS of the VAR.
Indeed it is orthogonal also to $\Delta y_t^f$. In fact, writing out the the VAR equation for $\Delta y_t^f$, it is immediate to notice that it includes the constant, the terms in $\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0$ and finally the residual which can be written out as $c_{11}\nu_t^1$. We have shown that $\nu_t^3$ is orthogonal to $\Delta y_{t-1}^f,\ u_{t-1}^f,\ \Delta y_{t-1}^0$. Moreover, the fundamental shocks, $\nu_t^3$ and $\nu_t^1$ in this case, are orthogonal to each other by construction. Hence it is indeed the case that $Cov(\Delta y_t^f, \nu_t^3) = 0$. By the same argument it is immediate to show that $Cov(u_t^f, \nu_t^3) = 0$ as well. Obviously this is not true for $\Delta y_t^0$, in fact $\nu_t^3$ enters the equation which defines it. And so it should be, because the whole idea is that the noise shock affect contemporaneously only the early vintage of data.

Hence we have shown how our noise shock is indeed orthogonal to all the lags of the variables included in the VAR as well as to all contemporaneous values of final releases. This makes our definition of noise shock much more robust than that one would obtain focusing simply on the revision which indeed could very well correlate with some of those variables.

As discussed in the main body of the paper, the main limitation to this procedure lies in the limited number of lags and variables which can be included. Given the limited number of data points we do not see a simple way out except looking at standard statistical tests for the specifications of VARs.

In particular the inclusion of unemployment appears to be important because it is likely to be a key "ingredient" in any forecast of output growth, since it is relatively easier to measure and is obviously correlated with output.