Predictable Recoveries

Xiaoming Cai, Wouter J. Den Haan, and Jonathan Pinder*

August 23, 2015

Abstract

Should an unexpected change in real GNP of $x\%$ lead to an $x\%$ change in the forecasts of future GNP? The answer could be no even if GNP is a random walk. We show that US economic downturns often go together with predictable short-term recoveries and with changes in long-term GNP forecasts that are substantially smaller than the initial drop. But not always! Essential for our results is that GNP forecasts are not based on a univariate time series model, which is not uncommon. Our alternative forecasts are based on a simple multivariate representation of GNP's expenditure components.

Key Words: forecasting, unit root, business cycles

JEL Classification: C53,E32,E37

*Cai: Tongji University. E-mail: `xmingcai@gmail.com`. Den Haan: London School of Economics, CFM, and CEPR. E-mail: `wjdenhaan@gmail.com`. Pinder: London School of Economics and CFM. E-mail: `j.m.pinder@lse.ac.uk`. We would like to thank Bas Jacobs and Sweder van Wijnbergen for useful comments.

1 Introduction

Accurate forecasts of future economic growth are very valuable, for example, because they are needed for policymakers to decide on the appropriate stance of monetary and fiscal policy. Good forecasts are also important for the private sector, for example, for investment decisions or purchases of durable consumption goods. For these reasons, it is important that such forecasts are done with utmost care; forecasts that are too pessimistic or too buoyant could induce the wrong decisions and be quite harmful. Understanding what lies ahead is especially important during recessions, which explains the strong interest to understand what the short-term and long-term consequences of economic downturns are for future output levels.

Campbell and Mankiw (1987) argued that:

> "The data suggest that an unexpected change in real GNP of 1 percent should change one's forecast by over 1 percent over a long horizon."

Thus, shocks to GNP are permanent. Campbell and Mankiw (1987) base their conclusion on estimated univariate ARMA models, that is,[1]

$$\phi(L)\,\Delta y_t = a_0 + \theta(L)\,e_t, \tag{1}$$

where y_t is the log of real GNP and e_t is a serially uncorrelated shock. This class of time-series models has the following properties: (i) there is only one-type of shock, that is, the response of output to realizations of e_t is always the same, independent of why there is a shock to output, and (ii) the response of output is linear in the magnitude of the shock, that is, if the shock is twice as large then the response is also twice as large. We will refer to the time-series model of Equation (1) as the "one-type-shock" model. This name highlights the model's main deficiency, as will become clear in the next section.

This paper consists of a methodological part and an application. The first methodological point made is that *univariate* time-series models like the one given in Equation

[1] They allow for the possibility that $\theta(L)$ has a root equal to 1, which would imply that y_t is stationary around a deterministic time trend.

(1) are not well suited to make forecasts, because *any* unexpected shock *always* leads to the same forecasting pattern. The fact that a random variable is a random walk is often thought to imply that there are no forecastable changes. As discussed in Section 2, this is not true. The second methodological point made is that a very large number of AR and MA terms may be needed to describe the time series processes of variables that are the sum of random variables.

In the application, we assess the quantitative importance of these arguments for forecasts of US GNP that are made during post-war economic downturns, including the recent financial crisis, and also for forecasts of UK GNP during the recent financial crisis. We compare the univariate one-type-shock model with a very simple multivariate model, namely a VAR that predicts the expenditure components of GNP. Forecasts for GNP are obtained by explicitly aggregating the forecasts of the components. Despite its simplicity, this time-series model allows for different types of shocks and a wide variety of forecasting patterns.

It is well known that GNP is—or is close to—an $I(1)$. If a variable is an $I(1)$ process, then there must be some shocks that have permanent effects. If an $I(1)$ process is modelled with a univariate model, then *all* unexpected changes in GNP will have at least some permanent effect. Although this persistent effect could be small or large relative to the magnitude of the initial shock, we find it to be quite large, which is consistent with Campbell and Mankiw (1987). If an $I(1)$ process is modelled with a multivariate model that allows for different types of shocks, then it is no longer the case that all unexpected changes in GNP must have permanent effects. In fact, we find that a simple multivariate model for GNP quite often predicts correctly that several economic downturns do not have persistent negative effects.

In Section 2, we will provide some theoretical background. In Section 3, we discuss the forecasts made by the two time-series models during economic downturns including the recent financial crisis.

2 Econometrics of univariate time-series models

In section 2.1, we illustrate why *univariate* time-series models can give misleading predictions *even* if they are correctly specified. In particular, it is possible that the variable of interest, y_t, is a random walk and (i) it is not necessarily true that all changes in this variable have a permanent effect and (ii) the model's predictions made during recessions systematically overpredict the persistence of the downturn. In section 2.2, we give reasons why it may be difficult to get a correctly specified univariate representation for aggregate variables.

2.1 Univariate models: Missing information and bias

Consider the following data generating process (*dgp*) for y_t:[2]

$$\begin{aligned} y_t &\equiv x_t + z_t, \\ (1-\rho L)\, x_t &= e_{x,t}, \\ (1-\rho L)\,(1-\rho_z L)\, z_t &= e_{z,t}, \\ \mathbb{E}_t\left[e_{x,t+1}\right] = \mathbb{E}_t\left[e_{z,t+1}\right] = \mathbb{E}_t\left[e_{x,t+1}e_{z,t+1}\right] = 0,\ \mathbb{E}_t\left[e^2_{x,t+1}\right] &= \sigma^2_x, \mathbb{E}_t\left[e^2_{z,t+1}\right] = \sigma^2_z, \end{aligned} \tag{2}$$

where $\mathbb{E}_t[\cdot]$ denotes the expectation conditional on current and lagged values of x_t and z_t. The persistence of the effects of $e_{x,t}$ on x_t is determined by the value of ρ and the persistence of the effects of $e_{z,t}$ on z_t is controlled by both ρ and ρ_z. We assume that

$$-1 < \rho < 1, \tag{3}$$

$$-1 < \rho_z \leq 1, \tag{4}$$

$$\frac{\rho_z}{\rho} > 1. \tag{5}$$

We define $e_{y,t}$ such that the following holds:[3]

$$(1-\rho_z L)\, y_t = e_{y,t}, \tag{8}$$

[2] This time-series specification is a generalization of the one studied in Blanchard, L'Huillier, and Lorenzoni (2013).

[3] It is always true that

$$(1-\rho_z L)\,(1-\rho L)\, y_t = (1-\rho_z L)\, e_{x,t} + e_{z,t}. \tag{6}$$

The unconditional autocovariance of $e_{y,t}$ and $e_{y,t-j}$, $\mathbb{E}[e_{y,t}e_{y,t-j}]$, is given by

$$\mathbb{E}[e_{y,t}e_{y,t-j}] = \frac{\rho^j}{1-\rho^2}\sigma_z^2 + \left((\rho-\rho_z)\rho^{j-1} + \frac{(\rho-\rho_z)\rho^j}{1-\rho^2}\right)\sigma_x^2. \tag{9}$$

This implies that the autocovariances of $e_{y,t}$ are equal to zero if the following equation holds:[4]

$$\sigma_z^2 = \frac{(\rho_z-\rho)(1-\rho_z\rho)}{\rho}\sigma_x^2. \tag{10}$$

If this equation is satisfied, then the correct univariate time-series specification of y_t is indeed an $AR(1)$ with coefficient ρ_z.

In this univariate representation for y_t, there is only one shock, $e_{y,t}$, ρ does not matter at all, and the persistence of the effects of this shock is solely determined by ρ_z. This is remarkable given that ρ affects the persistence of both fundamental shocks, $e_{x,t}$ and $e_{z,t}$. To understand why the univariate representation misses key aspects of the underlying system, consider the case when $\rho_z = 1$ and $\rho > 0$. The univariate representation is then given by

$$y_t = y_{t-1} + e_{y,t}. \tag{11}$$

That is, Δy_t is white noise and y_t is a random walk. Although y_t is a random walk, *all* changes in y_t imply predictable further changes. In particular, if $\Delta y_t < 0$ because $e_{x,t} < 0$, then there is a *predictable* recovery in y_t, since $x_t = \rho x_{t-1} + e_{x,t}$ and $0 < \rho < 1$. If $\Delta y_t < 0$ because $e_{z,t} < 0$, then there is a *predictable* further deterioration, since $\Delta z_t = \rho\Delta z_{t-1} + e_{z,t}$ and $\rho > 0$. If one only observes that $\Delta y_t < 0$, then one has to weigh the two possible cases and in this example the two opposing effects exactly offset each other, leading the forecaster to predict that the level of output will remain the same.

This example is special because the forecastability that is present in the underlying system completely disappears in the univariate representation. It is true more generally,

Thus, an equivalent definition of $e_{y,t}$ would be the following:

$$(1-\rho L)e_{y,t} = (1-\rho_z L)e_{x,t} + e_{z,t}. \tag{7}$$

These two equations are helpful in deriving the formulas in this section.

[4] $\sigma_z > 0$, since we assumed that $\rho_z/\rho > 1$.

however, that important information is lost in the univariate representation of the sum of variables.

Is the predicted long-run impact correct on average? The previous discussion showed that the univariate model given in equation (8) clearly misses some useful information. Next, we turn to the question whether this model generates (long-term) predictions that are *on average* correct.

To simplify the discussion, we focus on a particular version of the *dgp* given in equation (2). We assume that $\rho_z = 1$ and equation (10) is satisfied, so that the univariate representation of y_t is a random walk. Moreover, we set $\sigma_x = \sigma_z = \sigma$, which implies that $\rho = 0.381966$ according to equation (10). Finally, we assume that $e_{x,t}$ and $e_{z,t}$ can take on only two values, namely $-\sigma$ and $+\sigma$, both with equal probability. Note that the value of y_t remains unchanged if $e_{x,t}$ and $e_{z,t}$ have the opposite sign.

Although y_t has a random-walk representation, it systematically overpredicts the long-term consequences when output falls, i.e., during recessions, and it systematically underpredicts long-term consequences when output increases.

Before showing this, we first consider the case when output remains the same, which happens if $e_{x,t}$ and $e_{z,t}$ have the opposite sign. The (long-run) predictions based on the random-walk specification remain the same, since y_t remains the same. However, the true long-run predictions are affected as follows:

$$\begin{array}{l} \lim_{\tau \longrightarrow \infty} \mathbb{E}_t \left[y_{t+\tau} \right] - y_t = +\sigma / \left(1 - \rho\right) \text{ if } e_{z,t} = +\sigma \text{ and } e_{x,t} = -\sigma \text{ and} \\ \lim_{\tau \longrightarrow \infty} \mathbb{E}_t \left[y_{t+\tau} \right] - y_t = -\sigma / \left(1 - \rho\right) \text{ if } e_{z,t} = -\sigma \text{ and } e_{x,t} = +\sigma. \end{array} \tag{12}$$

Thus, when y_t remains the same, then one fails to recognize that the long-run value of y_t has gone up half of the time and fails to recognize that this long-run value has gone down the other half of the time. However, the forecasts are not systematically wrong.

Now consider the case in which output drops, which happens when $e_{x,t} = e_{z,t} = -\sigma$. The drop in output is equal to $-\sigma_x - \sigma_z = -2\sigma$. The random-walk specification implies that the long-run impact is identical to the short-term impact, that is,

$$\lim_{\tau \longrightarrow \infty} \widehat{\mathbb{E}}_t \left[y^f_{t,t+\tau} \right] - y_t = -2\sigma, \tag{13}$$

where $\widehat{\mathbb{E}}_t[\cdot]$ is the expectation according to the (correct) univariate representation. The *true* long-run impact of the shock, however, is equal to

$$\lim_{\tau \longrightarrow \infty} \mathbb{E}_t[y_{t+\tau}] - y_t = -\sigma/(1-\rho) = -1.618\sigma. \tag{14}$$

That is, in a recession, the univariate model systematically overpredicts the long-run negative impact of the economic downturn. Similarly, the univariate model systematically overpredicts the long-run positive impact of an increase in y_t. So the predictions are not biased, but one clearly is too pessimistic during recessions and too optimistic during booms if one would make predictions based on the random-walk specification.

In this stylized example in which $e_{x,t}$ and $e_{z,t}$ can take on only two values, one could drastically improve on the predictions of the univariate model even if one could not observe x_t or z_t, but knows the true *dgp*. The reason is that a drop in y_t implies that $e_{x,t}$ and $e_{z,t}$ are negative and an increase implies that both shocks are positive. The idea that the *magnitude* of the unexpected change in y_t has information about the importance of $e_{x,t}$ and $e_{z,t}$ is also true for more general specifications of $e_{x,t}$ and $e_{z,t}$, as long as one has information about the distribution of the two shocks. If one observes a very large drop in y_t, then it is typically the case that it is more likely that $e_{x,t}$ and $e_{z,t}$ are both negative than that $e_{x,t}$ is positive and $e_{z,t}$ is so negative it more than offsets the positive value of $e_{x,t}$ or vice versa. That is, the larger the economic downturn the larger the probability that a certain fraction of this downturn is driven by the transitory shock, that is, the larger the probability that a fraction of the drop in real activity will be reversed.

2.2 Aggregated variables and correctly specifying their *dgp*s

Aggregating ARMA processes. In this section, we highlight another problem with working with aggregated variables. We illustrate that the correct ARMA representation of an aggregate variable may very well be more complex than the most complex ARMA process for each of the component series. Formally, if x_t is an $ARMA(p_x, q_x)$ and z_t is an $ARMA(p_z, q_z)$, then $y_t \equiv x_t + z_t$ is an $ARMA(p, q)$ and p and q satisfy the following condition:[5]

$$p \leq p_x + p_z \text{ and } q \leq \max\{q_x + p_z, q_z + p_x\}. \tag{15}$$

These conditions give *upper* bounds for the $ARMA$ representation of the sum, y_t. Thus, the $ARMA$ representation of y_t is not necessarily of a higher order than those of x_t and z_t. In fact, in Section 2.1 we gave an example in which an $AR(1)$ variable and an $AR(2)$ variable add up to an $AR(1)$ variable.[6] But that example relies on very specific assumptions. In practice, one should not rule out the possibility that the univariate representation of a sum of several random variables could be quite complex. In fact, Granger (1980) argues that an aggregate of many components—as is the case for typical macroeconomic variables—may exhibit long memory.[7]

One might think that the solution to this dilemma is to use more complex $ARMA$ processes for aggregate variables. The problem is that the model has to be estimated with a finite amount of data, consequently the values of p and q cannot be too high. But if the values of p and/or q are too low, then the *dgp* could be misspecified.[8]

Simple example. We will now give a simple example, in which the predictions of a univariate time-series model for an aggregated variable are quite bad if that time-series model is *not* more complex than the most complex time-series representation of the components.

[5] See Granger and Morris (1976).

[6] In theory it is, of course, even possible that the sum of random variables is not random.

[7] One aspect that seems to be ignored in the econometrics literature is that the *dgp*s of the individual components may be "aligned" to the same factors. For example, if markets are complete, then market prices will align agents' marginal rates of sustitution—and, thus, their consumption growth processes—even if agents face very different income processes.

[8] The misspecification is likely to be worse than indicated in this section. Typically, log-linear processes are more suitable than linear processes. But if $y_t \equiv x_t + z_t$ and x_t and z_t are log-linear processes, then neither y_t nor $\ln(y_t)$ is a linear process and the convention of modelling $\ln(y_t)$ as a linear process is, thus, not correct. In fact, the effects of shocks on y_t would be time-varying. These issues are further discussed in Den Haan, Sumner, and Yamashiro (2011).

Consider the following *dgp*:

$$\begin{aligned} y_t &\equiv x_t + z_t, \\ x_t &= \rho_x x_{t-1} + e_{x,t}, \\ z_t &= e_{z,t}, \\ \mathbb{E}_t\left[e_{x,t+1}\right] &= \mathbb{E}_t\left[e_{z,t+1}\right] = 0, \\ \mathbb{E}_t\left[e_{x,t+1}^2\right] &= \sigma_x^2, \\ \mathbb{E}_t\left[e_{z,t+1}^2\right] &= \sigma_z^2, \end{aligned} \tag{16}$$

with $-1 < \rho_x < 1$. Thus, y_t is the sum of two stationary random variables, an AR(1) and white noise. Equation (16) implies that

$$(1 - \rho_x L)\, y_t = e_{x,t} + (1 - \rho_x L)\, e_{z,t}. \tag{17}$$

The first-order autocorrelation of the term on the right-hand side is not equal to zero unless $\rho_x = 0$, but higher-order autocorrelation coefficients of this term are equal to zero. Consequently, y_t is an $ARMA\,(1,1)$. That is, there is a value for θ such that the following is the correct univariate time-series representation of y_t:

$$(1 - \rho_x L)\, y_t = (1 + \theta L)\, e_{y,t}, \tag{18}$$

where $e_{y,t}$ is serially uncorrelated. The value of θ is given by the following expression:[9]

$$\theta = \frac{\rho_x\left(-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right]\right)}{\mathbb{E}\left[e_{y,t}^2\right]}. \tag{19}$$

The most complex component of y_t is x_t, which is an AR(1). So suppose that y_t is also modelled as an AR(1). That is,

$$y_t = \widetilde{\rho}_y y_{t-1} + \widetilde{e}_{y,t}. \tag{20}$$

[9] Since $e_{y,t}$ is white noise, it must be true that

$$\mathbb{E}\left[(1+\theta L)\, e_{y,t} \times (1+\theta L)\, e_{y,t-1}\right] = \theta \mathbb{E}\left[e_{y,t}^2\right].$$

It is also true that

$$\mathbb{E}\left[(1+\theta L)\, e_{y,t} \times (1+\theta L)\, e_{y,t-1}\right] = \rho_x\left(-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right]\right),$$

since $(1+\theta L)\, e_{y,t} = e_{x,t} + (1 - \rho_x L)\, e_{z,t}$ and both $e_{x,t}$ and $e_{z,t}$ are white noise. Combining both equations gives the expression for θ.

If we abstract from sampling uncertainty, we can pin down the value of $\widetilde{\rho}_y$ using population moments:

$$\widetilde{\rho}_y = \frac{\mathbb{E}\left[y_t y_{t-1}\right]}{\mathbb{E}\left[y_t^2\right]} = \frac{(\rho_x + \theta)(1 + \rho_x \theta)}{(1 - \rho_x^2) + (\rho_x + \theta)^2}. \tag{21}$$

We are interested in whether this AR(1) specification would tend to over- or underestimate the long term effects of shocks by comparing $|\widetilde{\rho}_y|$ with $|\widetilde{\rho}_x|$. If $|\widetilde{\rho}_y| > |\widetilde{\rho}_x|$, then the AR(1) specification would tend to overstate the true degree of persistence. It is straightforward to show that $|\widetilde{\rho}_y| > |\rho_x|$ if and only if $\theta\rho_x > 0$, that is, if ρ_x and θ have the same sign.[10] Equation (19) implies that this happens if

$$-\mathbb{E}\left[e_{x,t} e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right] > 0. \tag{23}$$

This condition is satisfied if the covariance of $e_{x,t}$ and $e_{z,t}$ is sufficiently negative. Similarly, $|\widetilde{\rho}_y| < |\rho_x|$ if and only if ρ_x and θ have the opposite sign, which happens if

$$-\mathbb{E}\left[e_{x,t} e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right] < 0. \tag{24}$$

This condition would be satisfied if the two shocks are positively correlated.

To shed some light on the possible consequences of using an $AR(1)$ as the law of motion for y_t, we consider the case when the two shocks have the following very simple relationship:

$$e_{z,t} = \alpha e_{x,t}. \tag{25}$$

Since $e_{x,t}$ and $e_{z,t}$ are perfectly correlated, there is only one type of shock and there is a univariate time-series specification of y_t that completely captures the dynamics of y_t. Now we investigate what the consequences of misspecifying the $ARMA(1,1)$ process as an $AR(1)$—as an $AR(1)$ is the most complex of the individual underlying time series processes.

[10] Equation (21) implies that $|\widetilde{\rho}_y| > |\rho_x|$ if

$$\begin{aligned} &\frac{(1-\rho_x^2)}{(1-\rho_x^2)+(\rho_x+\theta)^2}\theta > 0 \quad \text{when } \rho_x > 0, \\ &\frac{(1-\rho_x^2)}{(1-\rho_x^2)+(\rho_x+\theta)^2}\theta < 0 \quad \text{when } \rho_x < 0. \end{aligned} \tag{22}$$

Consequently, $\left|\widetilde{\rho}_y\right| > |\rho_x|$ if and only if $\theta\rho_x > 0$, that is, if ρ_x and θ have the same sign.

Figure 1 plots $\widetilde{\rho}_y$, i.e., the value of the coefficient of the $AR(1)$ representation of y_t, as a function of the true dominant root in the *dgp* of y_t, i.e., ρ_x. The top panel considers the case when the two shocks are negatively correlated ($\alpha < 0$). In this case, $\widetilde{\rho}_y$ is greater than ρ_x and so the $AR(1)$ process overstates the true amount of persistence. Conversely, if the shocks are positively correlated $\widetilde{\rho}_y$ is less than ρ_x, as shown in the lower panel.

These two panels document that long-term persistence is increased substantially for lower values of ρ_x when α is negative and that long-term persistence is decreased substantially for higher values of ρ_x when α is positive.

Figure 2 displays IRFs for three sets of parameter values. Each panel plots the true response of y_t to a one-time shock in $e_{x,t}$ and the response according to the $AR(1)$ specification for y_t. These three panels clearly document that misspecifying the aggregate variable y_t as an $AR(1)$—the correct specification of the most complex of the underlying processes—can give inaccurate impulse responses at both short and long horizons. The $AR(1)$ representation of y_t overestimates the long-term consequences of the shock when $e_{x,t}$ and $e_{z,t}$ are negatively correlated and underestimates them when the two shocks are positively correlated. The bottom two panels document that these bad long-term predictions only become apparent at forecast horizons of over 30 periods. At forecast horizons shorter than 30 periods, the $AR(1)$ representation of y_t *over*estimates the consequences of the crisis by a large margin when the shocks are *positively* correlated and vice versa. For example, when the shocks are negatively correlated, then the $AR(1)$ representation predicts that the initial reduction will be followed by an immediate but gradual recovery. By contrast, the true response is a further deterioration of almost the same magnitude followed by a somewhat faster recovery.

In this section, we focused on a case in which the most complex time-series specification of a component is an $AR(1)$, that is, a relatively simple process. Although the correct time-series specification of the aggregate is more complex, namely an $ARMA(1,1)$, it has only two parameters and one should be able to estimate this more complex time-series model with data sets of typical length. One can also improve on the $AR(1)$ specification by using higher-order AR processes, although these would—like the $AR(1)$—not be correct either, unless the number of lags is high enough to result in a sufficiently accurate approximation. However, the option to estimate a more complex representation may not always be feasible. If the two components are, for example, both an $AR(4)$, one would have to estimate an $ARMA(8,4)$, and if y_t is the sum of three $AR(4)$ processes, then one would have to estimate an $AR(12,8)$ to make sure that the univariate representation is not misspecified. In the next section, we document that a better strategy might be to estimate separate time-series models for the components and then explicitly aggregate the forecasts of the components to obtain forecasts for the aggregated variables.

3 Documenting the disadvantages of univariate models

The previous section made clear that (i) the correct univariate specification of a sum of random variables, y_t, could miss key predictable aspects of y_t and (ii) the correct univariate representation could be more complex than the most complex representation of its components. But the examples given were stylized. In this section, we discuss the empirical relevance of these claims. The purpose of this section is *not* to construct the best possible forecasting model. Instead the purpose is to document that (i) univariate time-series models that only have one type of shock—like univariate ARIMA models—are inadequate and tend to predict that shocks have very persistent effects even when the impact of the shock is short lived and (ii) a very simple multivariate model with more than one type of shock allows different types of forecasting patterns, which helps them to better capture observed recovery patterns during several recessions. We focus mainly on post-war US recessions, but also summarize the results for UK recessions and discuss in detail the UK experience during the recent financial crisis.

3.1 Empirical specifications

The specification of the multivariate model is given by the following VAR:

$$\ln(s_t) = \sum_{j=1}^{4} B_j \ln(s_{t-j}) + e_{s,t}, \tag{26}$$

where s_t is a 5×1 vector containing the expenditure components, consumption, c_t; investment, i_t; government expenditures, g_t; exports, x_t; and imports. m_t. The forecast for $y_{t+\tau}$ follows directly from

$$y_{t+\tau} \equiv e^{\ln(c_{t+\tau})} + e^{\ln(i_{t+\tau})} + e^{\ln(g_{t+\tau})} + e^{\ln(x_{t+\tau})} - e^{\ln(m_{t+\tau})}. \tag{27}$$

The estimated univariate model for aggregate output is given by:[11]

$$\ln(y_t) = \sum_{j=1}^{4} a_j \ln(y_{t-j}) + e_t. \tag{28}$$

The time series for y_t itself is also constructed using Equation (27) so that we are comparing like with like exactly. The key feature of the univariate time-series model is that there is only one type of shock. If output turns out to be unexpectedly lower than expected, i.e., $e_t < 0$, then the predicted effect on future values of y_t will always have the same pattern with the magnitude proportional to the value of e_t.

Both time-series processes are estimated with ordinary least squares (OLS). Given that the variables could very well be integrated, it is important to add enough lags to ensure that the shocks are stationary and spurious regression results are avoided. If the time series are known to be integrated, then efficiency gains are possible by imposing this. Additional restrictions can be imposed if the series are cointegrated. If these restrictions are correct, but are not imposed, then the estimated parameter values will converge towards the true parameter values at rate T, that is, there is superconsistency. If the restrictions are not correct and are nevertheless imposed, then the system is misspecified and the estimated system will not converge towards the true system. Because of superconsistency, we prefer not to impose these types of restrictions on the system.

[11] All models in this section also include a constant and a linear-quadratic deterministic trend. Campbell and Mankiw (1987) also consider $ARMA$ representations, but the results are similar to those obtained with AR represenations. The only exception is when third-order MA components are included, but the authors point out that the implied impulse response functions of this specification are estimated very imprecisely .

3.2 Impulse response functions

The impact of a negative one-standard-deviation shock to e_t on (the log of) US GNP, i.e., the impulse response function (IRF), is displayed in Figure 3.[12] Even though the specification in Equation (28) does not impose a unit root and contains a quadratic deterministic trend, the estimated specification documents that the impact to the shock e_t is very persistent. It is exactly this type of result that underlies the argument of Greg Mankiw that one should expect economic downturns to have permanent effects.

If output is generated by the multivariate model, i.e., according to equations (26) and (27), then there are five reduced-form shocks that result in a drop in output. Consequently, there are five IRFs, that is, five different ways in which output could respond. There are fierce debates in the economic literature on how to interpret shocks, but the interpretation of the shocks is not important for the point we want to make, that is, a model used to forecast GNP should allow for different forecasting patterns. For convenience, we will label the reduced-form shocks according to the regressand of the equation. For example, we will refer to $e_{c,t}$ as the consumption shock, but this is just a label and not meant to hint at a structural interpretation. The five IRFs are plotted in Figure 4. The figure makes clear that according to the multivariate model there are shocks that have an extremely persistent impact on output. The figure also makes clear, however, that there are shocks that have a transitory impact on output.

To sum up, Figures 3 and 4 illustrate two key points of this paper: (*i*) models that allow for only one type of shock will not discover that the long-term impact is not the same for each type of shock and (*ii*) that a richer model can discover this difference.

[12] See Appendix A for further details on data sources. Whereas the forecasting exercise discussed in the next subsection is based on real-time data, the results in this subsection are based on the full sample of quarterly US data from 1947Q1 to 2015Q1. The results are very similar if the sample ends in 2006Q4 and the financial crisis is, thus, excluded, except that the IRF of the "import" shock is then less persistent.

3.3 Forecasting recoveries during past recessions

The analysis above showed that the dynamics of GNP according the multivariate VAR are rich and diverse, in contrast to the univariate specification. We now investigate whether this matters in practice for the recovery of GNP in a recession, as in principle the univariate model could capture the shocks which drive economic downturns. We start with a discussion of post-war US recessions, followed by a discussion of the behavior of UK GNP during and after the recent financial crisis.

Out-of-sample forecasts. We use the univariate and the multivariate time-series models to make predictions about future output during economic recessions. Forecasts are out-of-sample forecasts, because forecasts made at t^* only use data up to date t^*.[13] We use the latest vintage of data for each forecast.[14]

Explaining the figures. The vertical line in each figure indicates the forecasting point. The thick solid line plots the actual data. Each figure also plots the predicted growth path according to the two time-series models and a deterministic time trend.[15]

1973-75 US recession. The top panel of figure 5 displays the results for the 1973-75 recession. Forecasts are made at the trough of the recession, 1975Q1. Forecasts from the univariate one-type-shock model indicate that output losses will be very persistent. Instead, there is a rapid recovery back to the long-term trend. Given that there are at times persistent changes in GNP, the univariate model will always reflect this persistence

[13] Strictly speaking, this is *pseudo* out-of-sample forecasting, since future data is available at each forecasting point.

[14] The first recession considered is the 1973-75 recession. Because we focus on out-of-sample forecasts, we have only 109 quarterly observations for forecasts at the trough of this recession, which leaves few degrees of freedom when the VAR is estimated with the default specification, that is, four lags for each of the five variables and a quadratic deterministic trend. By using a VAR with only two lags, we avoid the strong sensitivity of forecasts when the forecasting date shifts slightly.

[15] The time trend shown in the figures is a linear trend estimated on the full sample of GNP and is included as a point of reference. The linear-quadratic trends included in the univariate and multivariate models are estimated up until t^*.

to some extent.[16] By contrast, the forecast based on the multivariate model captures the fast recovery of GNP after the trough of the recession. In addition to the predicted short-term increase in growth rates, the multivariate model also captures the subsequent return to normal growth rates. Not surprisingly, the path forecasted in 1973Q2 does not predict the recessions of the early eighties.

The exercise discussed here should not be considered as a horse race of two forecasting models. What the results show is that (i) some economic downturns are followed by faster than normal growth and seem to have little or no permanent effects and (ii) this type of pattern is unlikely to be predicted by univariate models, whereas multivariate VARs do have the flexibility to capture this.

1980 US recession. The bottom panel of figure 5 displays results for the first recession of the early eighties. Forecasts are made at the trough, 1980Q3. Both models predict that the shortfall of GNP relative to its trend value observed in 1980Q3 will remain of roughly the same magnitude up till 1984. This means that both models miss the short-lived pickup in growth rates just after 1980Q3 and both miss the second recession in the early eighties. In 1984, the economy has recovered from the second recession, although GNP is still below its trend value, and GNP is in fact close to the levels predicted by both models using data up to 1980Q3.

The two 1980Q3 forecasts diverge in their predictions for the post-1984 period. The 1980Q3 forecast according to the univariate model predicts that the gap between GNP and its (ex-post) trend value will not become smaller. By contrast, the 1980Q3 forecast based on the multivariate model indicates that the gap will become smaller, which is indeed what happened. In 1986, GNP was back to its trend value, which is in line with the 1980Q3 prediction according to the multivariate model.

The recovery predicted by the multivariate model in 1980Q3 is quite different from the recovery predicted in 1973Q2. Whereas, the multivariate model predicts a quick return at the trough of the seventies recession, it predicts a much more gradual return at the trough

[16] However, since we use an $AR(4)$ to describe real output, our model does allow for a further predictable deterioration and/or for the possibility that (a large) part of the initial drop can be expected to be reversed.

of the first early eighties recession.

1981-82 US recession. The top panel of figure 6 reports the results for the forecasting exercise when forecasts are made at the end of the second early-eighties recession, 1982Q4. From this point onwards, the US economy recovers remarkably quickly. Whereas the economy is almost 9% below its (ex-post) trend level at the end of 1982, this gap is only 2.5% at the end of 1984 and only 1% at the end of 1985. The multivariate model captures this remarkable recovery very well. It does not capture, however, the fact that in subsequent years the gap gets even smaller. The univariate model completely misses the recovery and predicts, again, that ground lost during the recession is permanent.

Both the behavior of GNP during this recession and the fact that the remarkable recovery can be predicted by a simple time-series model strongly suggest that it is not always the case that an unexpected change in real output of x percent should lead to a change of the long-term forecast of x percent.

Although our multivariate model is a simple VAR, with five variables and four lags, it allows for a rich set of dynamics. It is, therefore, not always easy to understand what features of the data lead to particular predictions. For this particular period, it is possible to point at the reason why the model predicts a sharp recovery. The period just before 1982Q4 is characterized by sharp drops in investment and exports. As documented in figure 4, these correspond to *temporary* reductions in GNP. Consequently, the multivariate model predicts that these negative influences will disappear quickly. During 1982, both consumption and government expenditures have started to grow already, which according to figure 4 correspond to permanent positive changes in GNP. This is consistent with the predicted persistence of the recovery.

1990-91 US recession. The bottom panel of figure 6 displays the results for the recession of the early 1990s. The results differ from those reported above for previous recessions in that now both models predict a permanent loss in GNP. Although the loss in actual GNP is indeed very persistent and GNP does not get back to its trend level until 1997, the actual loss is not permanent.

2001 US recession. The results for the early naughties recession are displayed in the top panel of figure 7. During this recession, there is not a sharp contraction in output. It is better characterized by a period of near zero growth rates. The recovery is also very gradual. The multivariate model is wrong in predicting a short-term pick up in growth rates, but is correct in its longer-term forecast that the loss in GNP is not permanent. The univariate model predicts again that there will be no recovery, not in the short term, which in this case is indeed what happened, and also not in the long term, which is not what happened.

US financial crisis, 2008-2009 The bottom panel of figure 7 plots the results for the forecasts made in 2009Q2, when the sharp fall in GNP had come to a halt.[17] Similar to forecasts made in previous recessions, the multivariate model again predicts that part of the loss in output relative to trend will be recovered in a couple years. Different from forecasts made in previous recession is that the univariate now also predicts a recovery. In fact, at this point in time, the univariate model predicts stronger long-term growth than the multivariate model. Unfortunately, forecasts of both models were too optimistic.

Starting in 2012, the multivariate model starts to predict the future reasonably well. In particular, it correctly predicts that output loss relative to trend will not be reversed.[18] The univariate model remains more optimistic than the multivariate model until the end of the sample, sometimes marginally more optimistic, but typically substantially more optimistic. Using data up to the end of our sample, the univariate model predicts that output in 2025 will be 1% below its extrapolated trend value whereas the multivariate model predicts that the gap will be 4.5%.[19]

[17] At the beginning of the financial crisis, both time-series models wrongly predict that a substantial part of the losses will be recaptured quickly. These results are not displayed in the graphs.

[18] These results are not displayed in the figures.

[19] The economy was substantially above its trend value before the crisis, which means that these long-term predictions imply larger losses relative to the hypothetical case when there would have been no financial crisis and subsequent average real output growth would have been equal to the trend growth rate.

UK recessions before the financial crisis. Post-war UK recessions are not as interesting as US recessions. Instead of sharp contractions, like those observed for the US, UK recessions were typically prolonged periods of low growth rates. Similarly, recoveries were very gradual. Although the multivariate model has better long-term predictions than the univariate model in five of the six recessions that occurred before the financial crisis, the predictions of the two models are roughly similar. Moreover, forecasted paths are close to straight lines, which is not surprising given the shallow aspect of economic downturns in the UK. The exception to these observations is the financial crisis, which will be discussed next.

UK financial crisis, 2008-2010. Figures 8 and 9 plot the realizations of UK GNP together with forecasts made by the two models at four different forecasting points. First consider the two panels of figure 8, which plot the results when forecasts are made at the middle of the period with large negative growth rates, 2008Q4, and at the end of this period, 2009Q2.

In the middle of the period when GNP dropped sharply, the univariate model predicts an immediate and sustained return to positive growth rates. It is even somewhat more optimistic than the prediction of a random walk model with drift in that it predicts that GNP will grow faster than its trend in the next couple years, that is, it predicts that part of the reduction of the pre-crisis positive gap between GNP and its trend value will be recovered. By contrast, the multivariate model predicts that GNP will grow at rates that are somewhat lower than the trend growth rate, which is closer to the observed outcomes, although also too optimistic. In 2009Q2, the univariate model still predicts that GNP will end up substantially above its trend value. The multivariate model forecasts that growth rates would be around zero for several quarters followed by a very gradual recovery. These forecasts are slightly below the actual outcomes.

The two panels of figure 9 plot the results when forecasts are made in 2009Q3 and 2010Q1. Both of these quarters are in the period when the UK economy had just started its recovery. For both forecasting points, the univariate model's predictions indicate that the economy will start growing at rates slightly higher than those observed in the past so that it still predicts that part of the losses will be recovered. By contrast, the multivariate model—using data up to 2009Q3—predicts that there first will be a period with low growth rates, which eventually is followed by a period of faster growth rates. This is indeed what happened, although the predictions are a little bit too pessimistic. Half a year later, in 2010Q1, the forecasts of the multivariate model have improved somewhat and do a good job in predicting the subsequent development of UK GNP.

We do not want to argue that the multivariate model is a remarkably good forecasting model. Neither model does very well in predicting subsequent output growth during this period, although it is worth noting that the multivariate model realizes quickly that output losses will be very persistent. The point that we want to make is that multivariate models have the flexibility to predict different types of forecasting patterns. By contrast, univariate models are quite restrictive and may miss both predictable recoveries and—as is shown here—a predictable deterioration during a downturn. The main reason why the univariate model is restrictive is that it has only one type of shock. Since the GNP data used to estimate the univariate model contains a persistent component, changes in GNP will *always* lead to changes in the long-term forecasts of the univariate model. Although, univariate forecasts always have a permanent component, we allow for the possibility that short-term forecasts are different from long-term forecasts, since our empirical univariate model has four lags. But all of our estimated univariate models imply predictions that are quite close to those of a random walk with drift.

4 Concluding comments

In this paper, we documented that a correctly specified univariate representation of a sum of random variables could miss key predictable aspects of this random variables. In fact, even if a random variable is a random walk, then that does not mean that there are no forecastable changes. Moreover, the correct specification of an aggregate of random variables could be quite complex. We argued that it might be better to estimate time-series models for the components and obtain forecasts for the aggregate by explicitly aggregating the forecasts of the components. We demonstrate this is indeed a better strategy when forecasting GNP by comparing the forecasts of a univariate representation of GNP and the forecast implied by a VAR of the expenditure components of GNP for post-war recessions in the UK and the US. One point that we do not address is the correct level of (dis)aggregation. Consumption is the sum of non-durable and durable consumption and both are sums of individual expenditures. So further disaggregation may lead to further improvements. It is not clear, however, whether one should disaggregate to the lowest possible level, since sampling variation typically increases when one considers disaggregated variables.

A Data sources

US data. Data are downloaded from the web site of the Federal Reserve Bank of St. Louis. They are (i) Consumption: real personal consumption expenditures; (FRED code: PCECC96)(ii) Investment: real gross private domestic investment (GPDIC1); (iii) Government expenditures: real government consumption expenditures & gross investment (GCEC1); (iv) Exports: real exports of goods & services (EXPGSC1); and (v) Imports: real imports of goods & services (IMPGSC1). All time series are seasonally adjusted quarterly data measured in billions of chained 2009 dollars. The data were last updated May 29, 2015.

The GNP data used is the sum of the consumption, investment, government expenditures, and exports minus imports. Adding up these real time series generates a time series that is extremely close, but not exactly identical to the actual GNP data. Our approach ensures that the components used in the multivariate model add up *exactly* to the data used in the univariate model. This way, we avoid clutter in the paper by describing small differences in the GNP data used in the two types of time-series models.

UK data. Data are from the Office of National Statistics. They are (i) household final consumption expenditures (ONS code: ABJR) plus final consumption expenditure of non-profit institutions serving households (HAYO); (ii) total gross fixed capital formation (NPQT); (iii) general government: Final consumption expenditures (NMRY); (iv) balance of payments: Trade in goods and services: Total exports (IKBK); (v) Balance of payments: Imports: Trade in Goods and services (YBIM). All data are seasonally adjusted quarterly data and the base period is 2011. The GNP data used is the sum of these five components. Investment in inventories are excluded, since they contain some very volatile high frequency movements.

References

Blanchard, O. J., J.-P. L'Huillier, and G. Lorenzoni (2013): "News, Noise, and Fluctuations: An Empirical Exploration," *American Economic Review*, 103, 3045–3070.

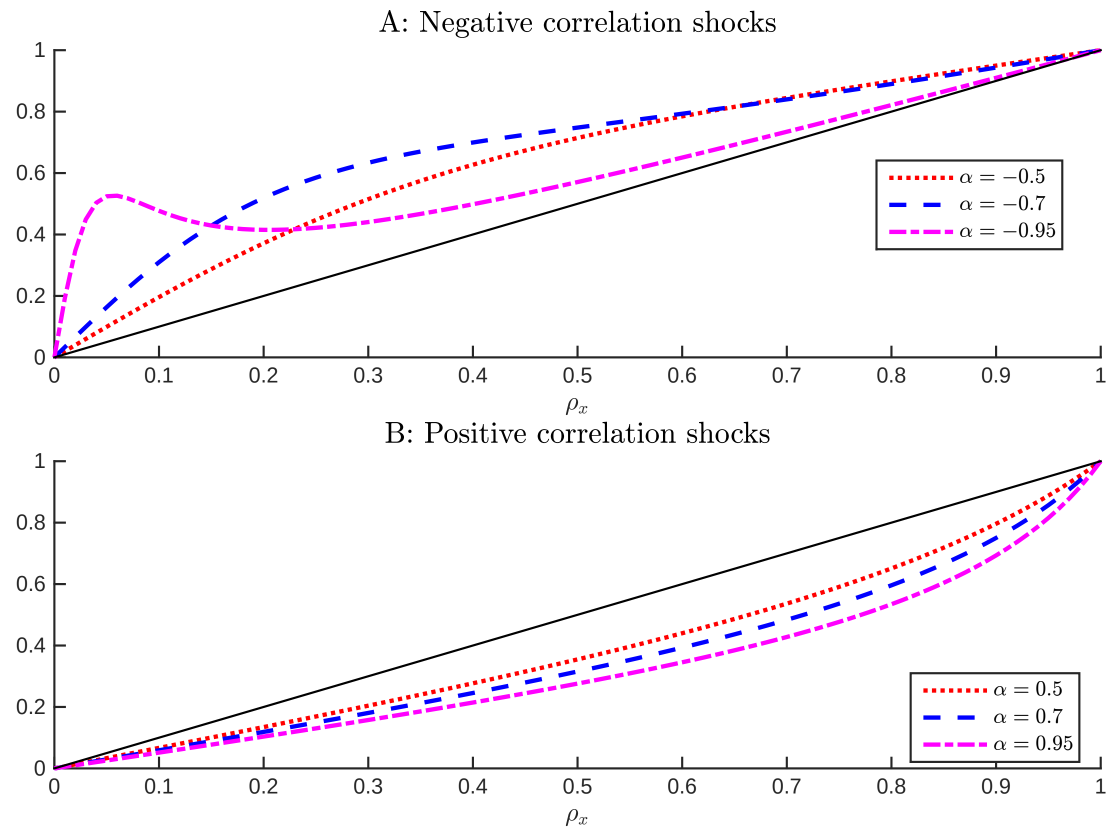
Campbell, J. Y., and N. G. Mankiw (1987): "Are Output Fluctuations Transitory," *Quarterly Journal of Economics*, 102, 857–880.

Den Haan, W. J., S. W. Sumner, and G. Yamashiro (2011): "Bank Loan Components and the Time-Varying Effects of Monetary Policy Shocks," *Economica*, 78, 593,617.

Granger, C., and M. Morris (1976): "Time Series Modelling and Interpretation," *Journal of the Royal Statistics Society A*, 139, 246–257.

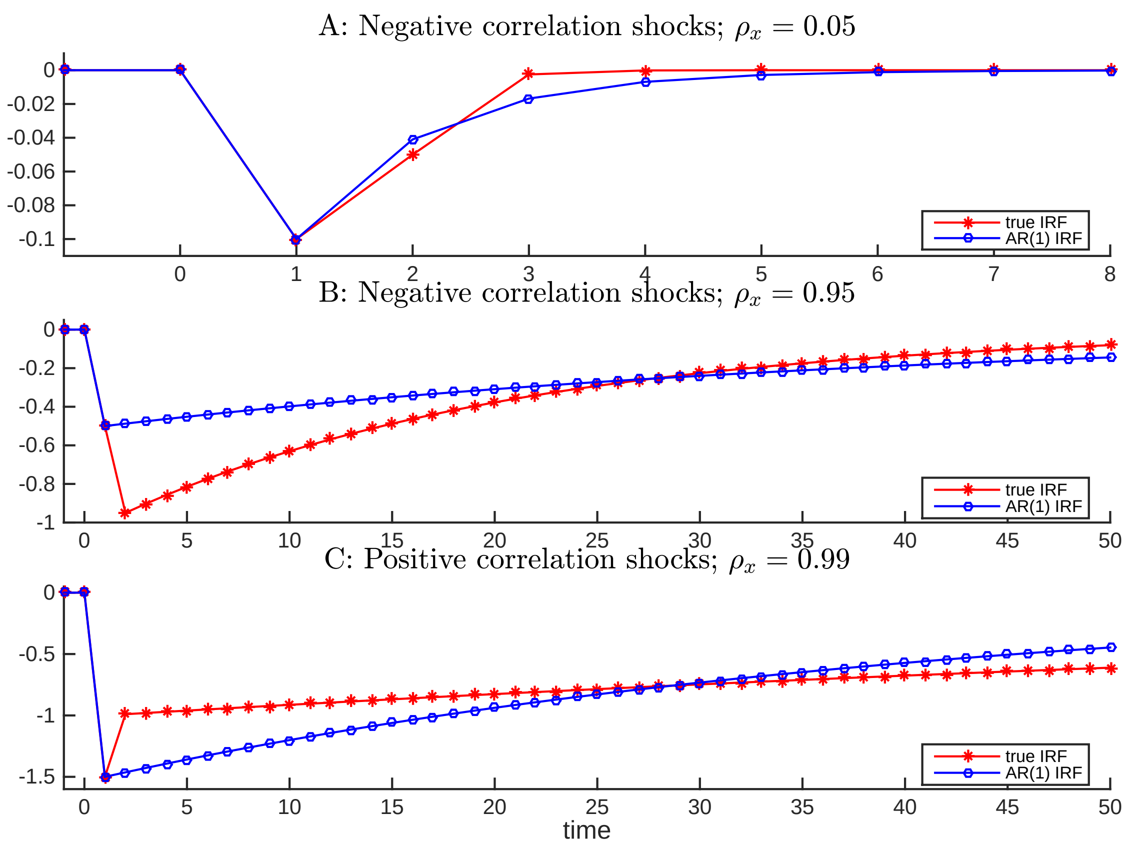
Granger, C. W. (1980): "Long Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227–238.

Figure 1: $AR(1)$ coefficient of $y_t = x_t + z_t$ according to incorrect univariate representation



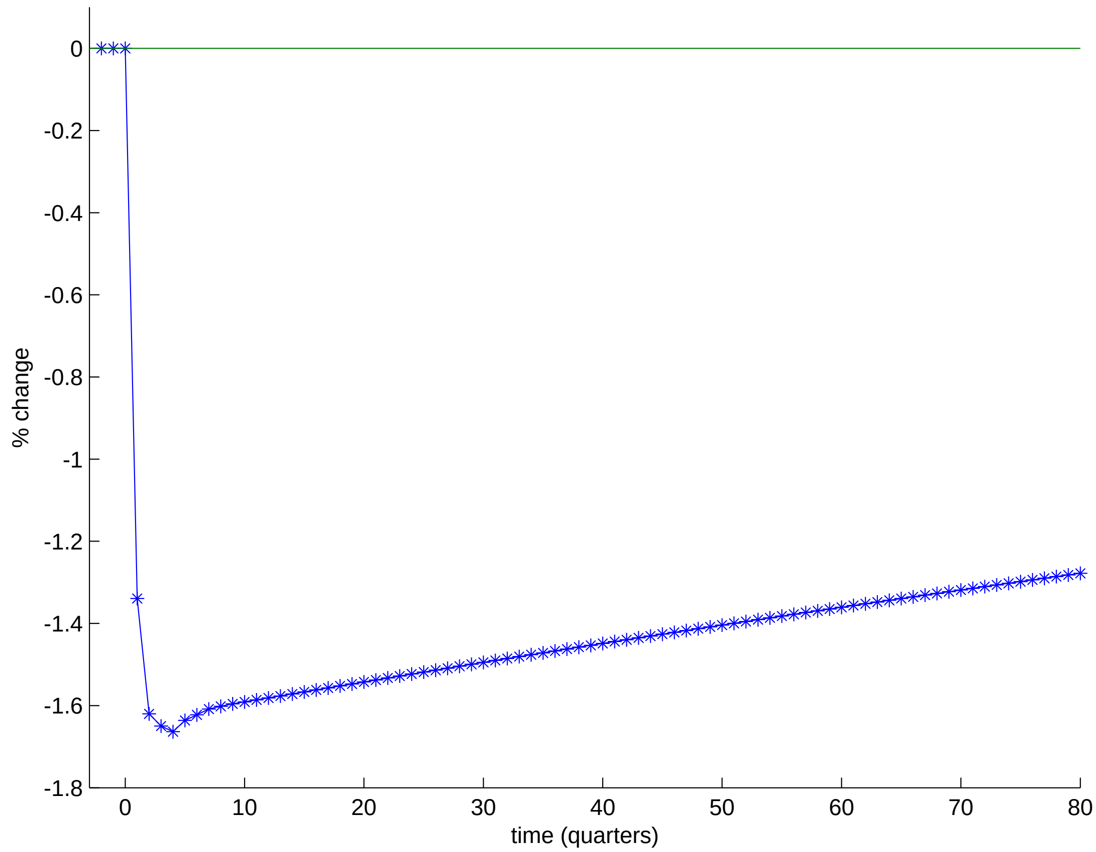
Notes: The graph displays the root of the $AR(1)$ representation of $y_t = x_t + z_t$ as a function of the AR root in the true time-series representation of y_t when $e_{z,t} = \alpha e_{x,t}$.

Figure 2: IRFs of $y_t = x_t + z_t$ according to correct and incorrect univariate representation



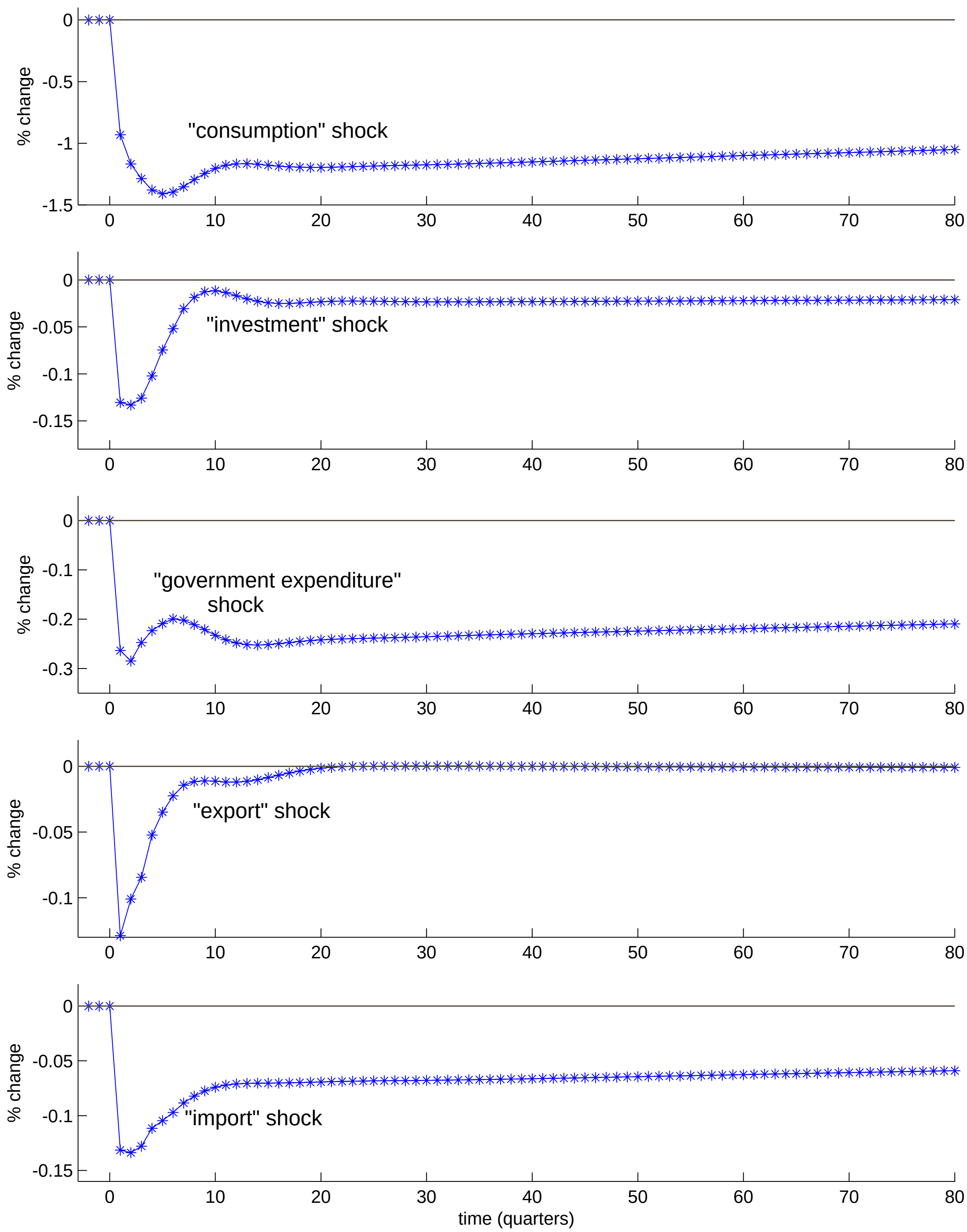
Notes: The graph plots the true responses of $y_t = x_t + z_t$ to a one-time shock in $e_{x,t}$ and the response according to the $AR(1)$ representation, which is the time-series representation of the most complex of the y_t components. In panel A, $e_{z,t} = -0.9e_{x,t}$; in panel B, $e_{z,t} = -0.5e_{x,t}$; and in panel C, $e_{z,t} = 0.9e_{x,t}$.

Figure 3: Effect of the shock in univariate model on US GDP



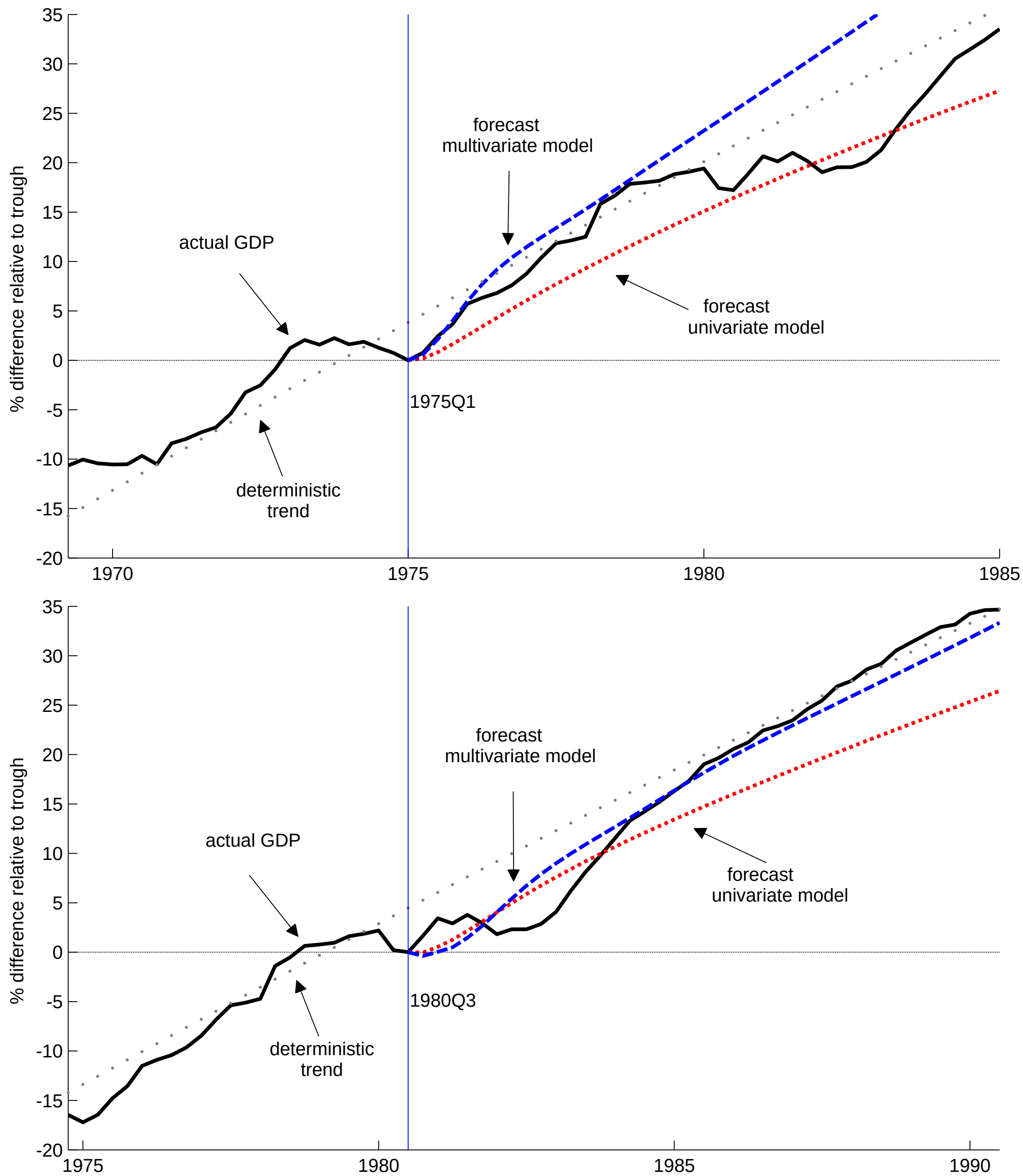
Notes:The graph plots the response of output following a one-standard-deviation negative shock according to the univariate, one-type-shock, model.

Figure 4: Effect of reduced-form VAR shocks on US GDP



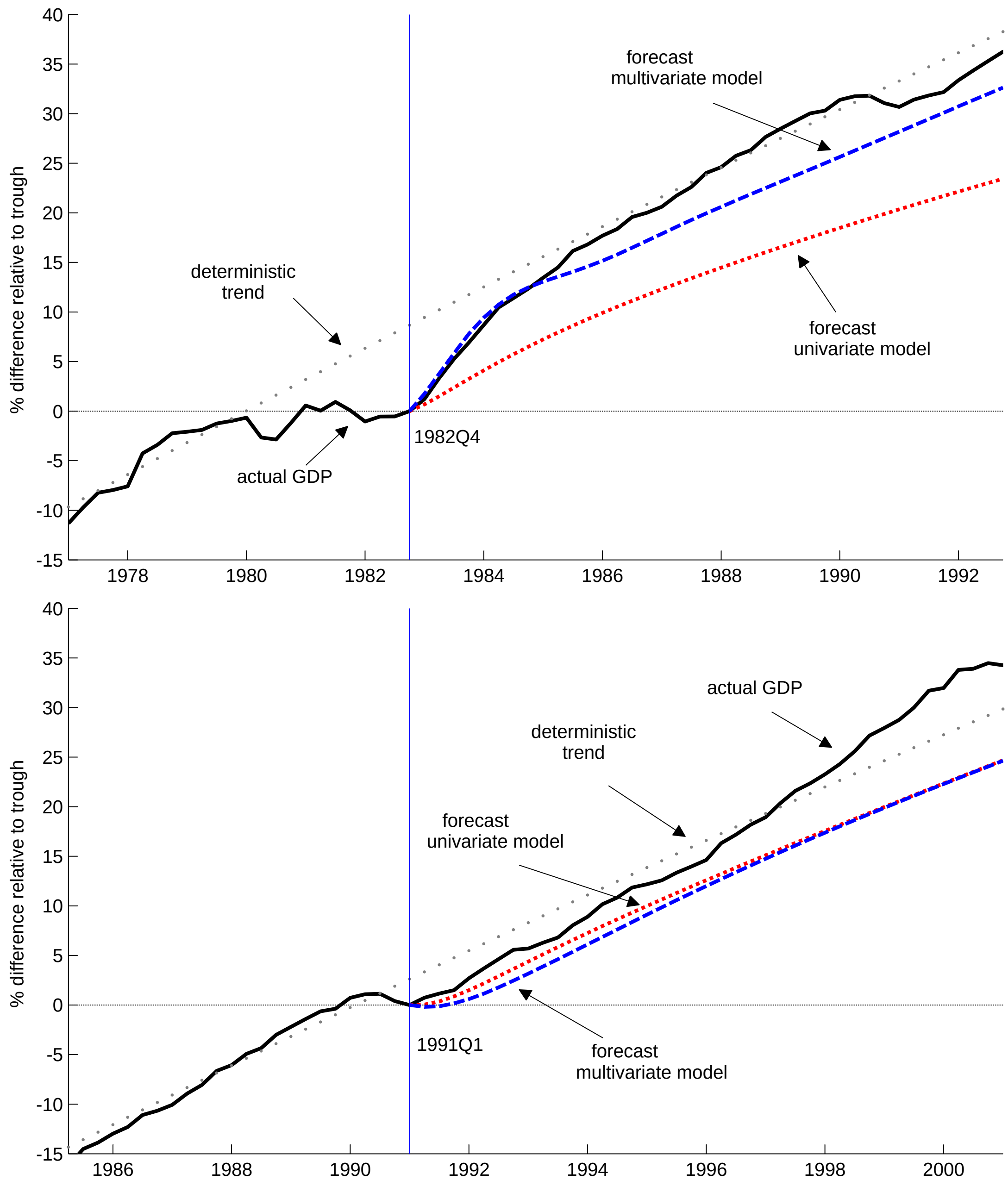
Notes: The graphs plots the predicted responses of output following a one-standard-deviation shock in the indicated reduced-form VAR shock that leads to a reduction in GNP.

Figure 5: The 1973-75 and the 1980 US recessions



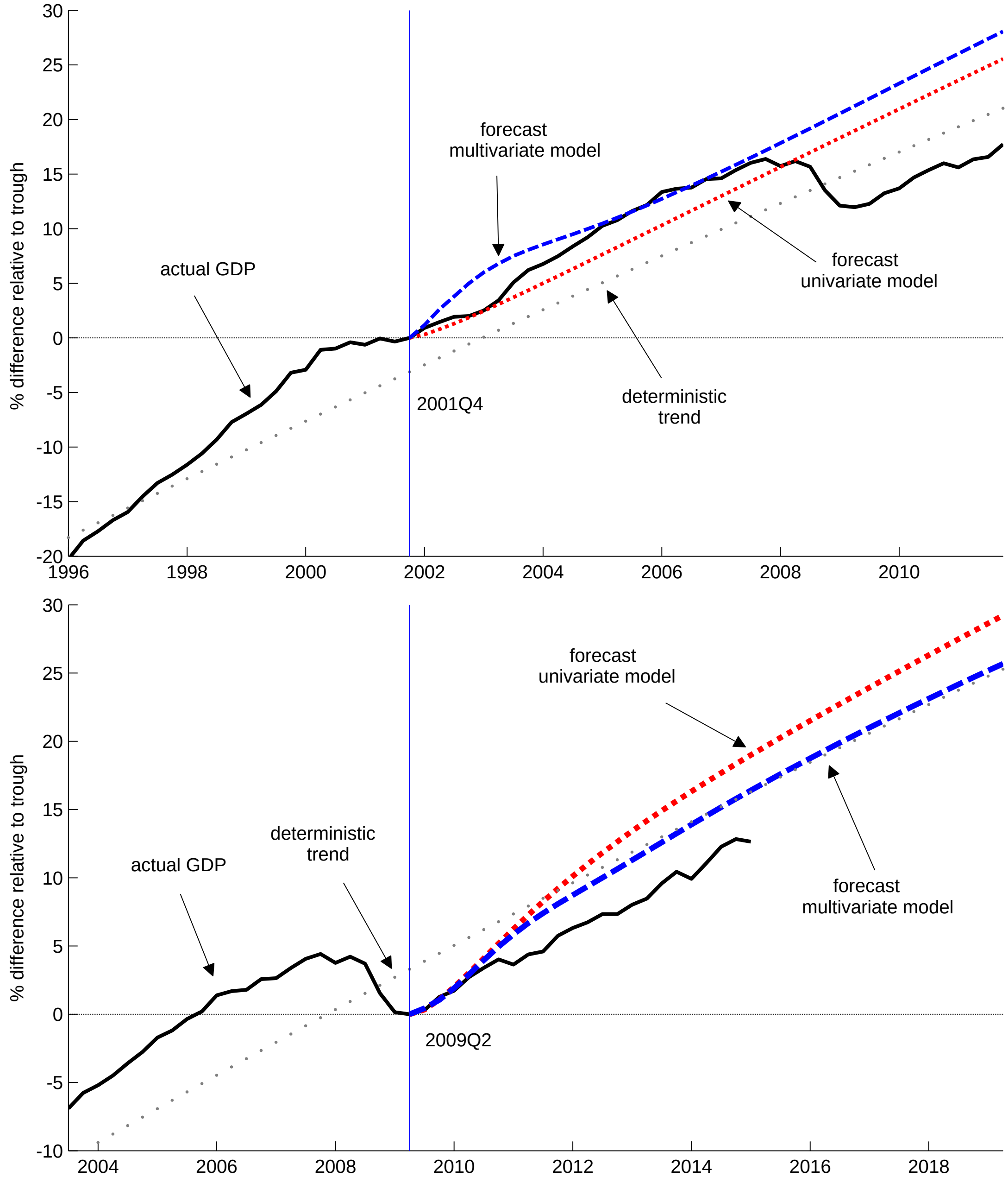
Notes: This figure plots the two forecasted time paths for US GDP together with the realized values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

Figure 6: The 1981-82 and the 1990-91 US recessions



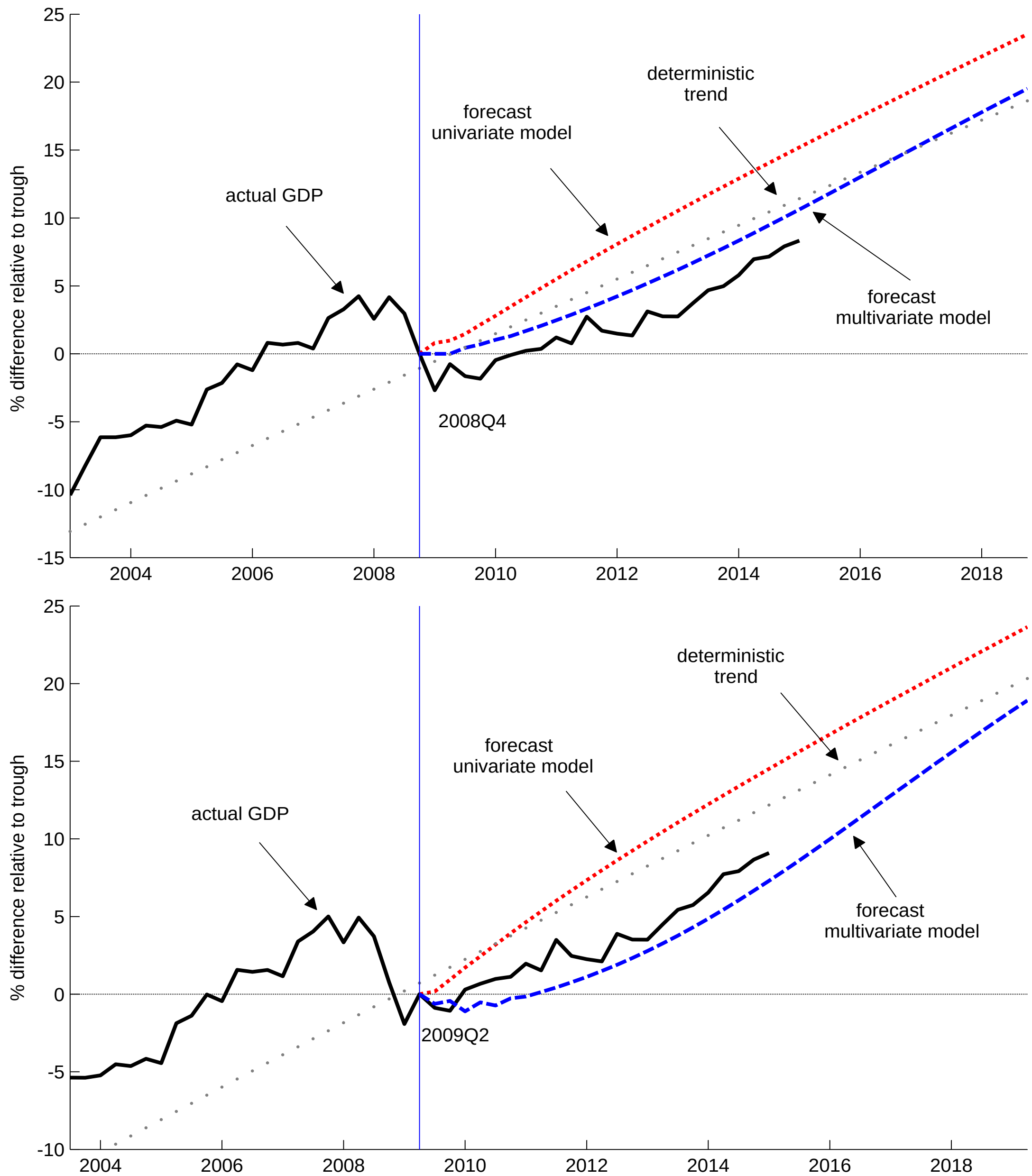
Notes: This figure plots the two forecasted time paths for US GDP together with the realized values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

Figure 7: The 2001 and great US recession



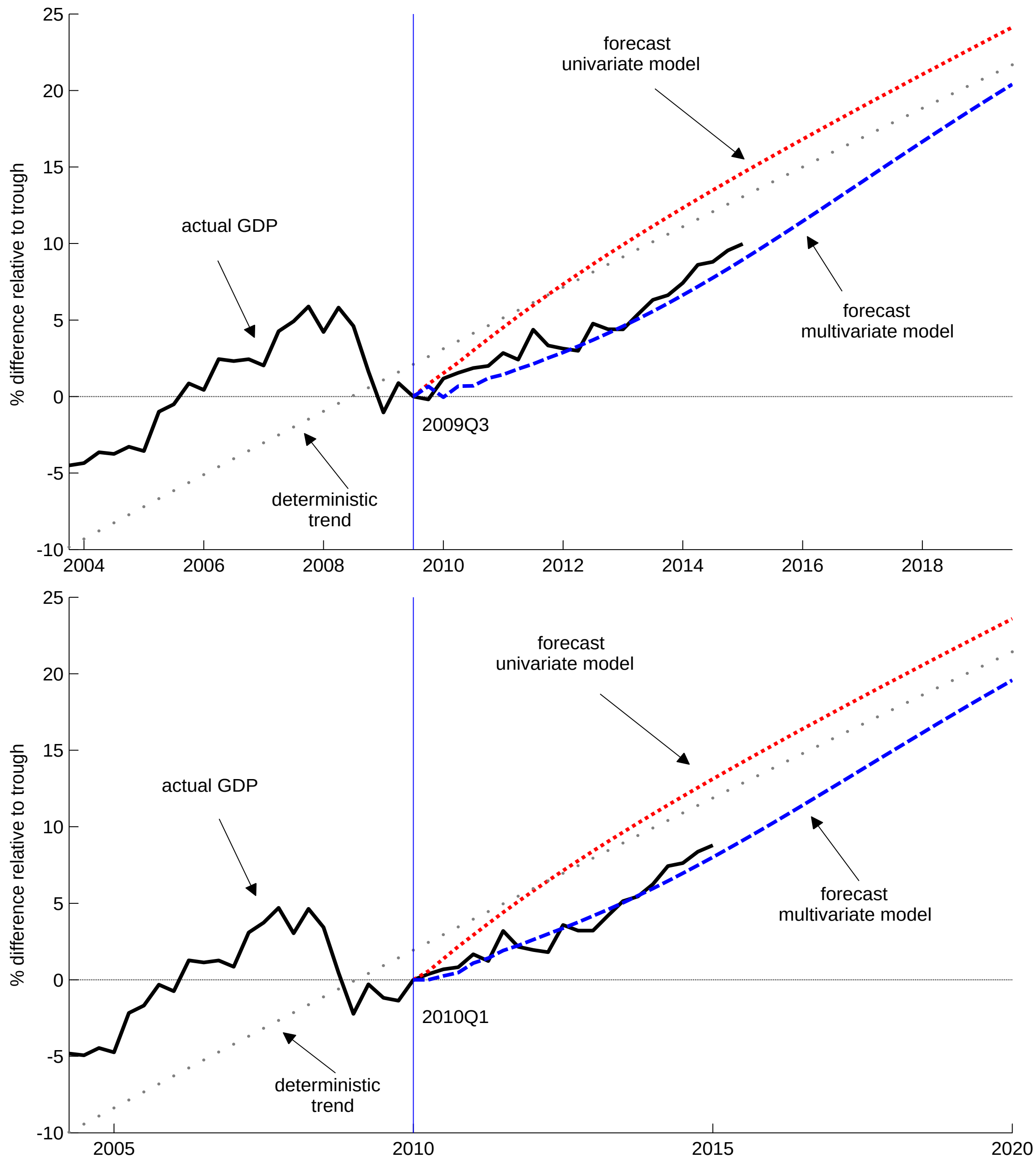
Notes: This figure plots the two forecasted time paths for US GDP together with the realized values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

Figure 8: The start and trough of the great UK recession



Notes: This figure plots the two forecasted time paths for UK GDP together with the realized values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

Figure 9: The initial recovery of the great UK recession



Notes: This figure plots the two forecasted time paths for UK GDP together with the realized values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.