September 24, 2010

Forecasting Levels of log Variables in Vector Autoregressions

Gunnar Bårdsen[1]
Department of Economics, Dragvoll, NTNU,
N-7491 Trondheim, NORWAY
email: gunnar.bardsen@svt.ntnu.no

Helmut Lütkepohl
Department of Economics, European University Institute
Via della Piazzola 43, I-50133 Firenze, ITALY
email: helmut.luetkepohl@eui.eu

Abstract. Sometimes forecasts of the original variable are of interest although a variable appears in logarithms (logs) in a system of time series. In that case converting the forecast for the log of the variable to a naive forecast of the original variable by simply applying the exponential transformation is not optimal theoretically. A simple expression for the optimal forecast under normality assumptions is derived. Despite its theoretical advantages the optimal forecast is shown to be inferior to the naive forecast if specification and estimation uncertainty are taken into account. Hence, in practice using the exponential of the log forecast is preferable to using the optimal forecast.

Key Words: Vector autoregressive model, cointegration, forecast root mean square error

JEL classification: C32

[1]This research was undertaken while Gunnar Bårdsen was visiting the European University Institute as a Fernand Braudel Fellow. The excellent working conditions offered are gratefully acknowledged. We thank an anonymous associate editor for helpful comments on an earlier draft of the paper.

1 Introduction

It is quite common to use economic variables in logarithms (logs) in economic models. Also vector autoregressions (VARs) are often constructed for the logs of variables. There are a number of justifications for using logs rather than the original variables. For example, the statistical properties of the model fitted to the logs may be preferable to those of a model for the original variables. In particular, the residuals of a model for logs may have a more homogeneous variance or they may even be well described by a normal distribution. Furthermore, growth rates of economic variables are often of primary interest. Approximating the growth rates by changes in the logs of a variable is common practice. Hence, the log transformation is natural for many economic variables.

Since time series models are used for forecasting and sometimes forecasts of the original variables are of interest, an obvious question is how to obtain such a forecast from a forecast for the log of the variable. Although it is tempting to use the exponential of the forecast of the log variable, a classical result by Granger and Newbold (1976) for univariate models states that such a "naive" forecast is generally not optimal. This result was extended by Ariño and Franses (2000) to VAR models. In fact, these authors derive the optimal forecast for Gaussian VAR models and argue that in practice sizable gains are possible in forecast accuracy from using the optimal forecast.

In this study we reconsider this finding by first deriving a somewhat more compact expression for the optimal forecast and, second, investigating possible gains in forecast precision to be expected from using it. Having a more transparent expression of the optimal forecasting formula enables us to see more easily that for typical economic variables gains in the forecast precision from using the optimal rather than the naive forecast are not likely to be substantial. In fact, in practice the optimal forecast may well be inferior to the naive forecast. This result is fully in line with findings by Lütkepohl and Xu (2011) who compared different univariate forecasts and found that the naive forecast may be superior to the optimal forecast when specification and estimation uncertainty are taken into account. We use Monte Carlo simulations to demonstrate that for variables which have typical features of some economic variables, using the optimal forecast is likely to result in efficiency losses if the forecast precision is measured by the root mean square error (RMSE). We also reconsider the example used by Ariño and Franses (2000) and demonstrate that under our criteria gains in forecast precision may be obtained by using the naive rather than the optimal forecast. Our overall conclusion is that the common practice of forecasting the logs of a variable and getting a forecast for the original variable by applying the

exponential function is a useful strategy in practice.

Our study is structured as follows. In the next section a transparent expression of the optimal forecast of the level of a variable which enters a VAR in logs is derived. In Section 3 the results of a simulation experiment are reported which compares the naive and the optimal forecast. Empirical forecast comparisons based on economic data are discussed in Section 4 and Section 5 concludes.

2 Forecasts of Levels of log Transformed Variables

Suppose $x_t = (x_{1t}, \ldots, x_{Kt})'$ is a K-dimensional VAR process of order p (VAR(p)),

$$x_t = \nu + A_1 x_{t-1} + \cdots + A_p x_{t-p} + u_t, \tag{1}$$

where $u_t \sim \mathcal{N}(0, \Sigma_u)$ is Gaussian white noise. By successive substitution we can write

$$x_{t+h} = \nu^{(h)} + A_1^{(h)} x_t + \cdots + A_p^{(h)} x_{t+1-p} + u_{t+h} + \Phi_1 u_{t+h-1} + \cdots + \Phi_{h-1} u_{t+1},$$

where $\nu^{(h)}$ and the $A_i^{(h)}$'s are functions of the original VAR parameters and

$$\Phi_i = \sum_{j=1}^{\min(i,p)} \Phi_{i-j} A_j$$

can be computed recursively for $i = 1, 2, \ldots$, with $\Phi_0 = I_K$ (e.g., Lütkepohl (2005, Chapter 2)).

Denoting by E_t the conditional expectation operator, given information up to time t, the optimal (minimum mean square error (MSE)) h-step ahead forecast of x_t at origin t is

$$E_t(x_{t+h}) \equiv x_{t+h|t} = \nu^{(h)} + A_1^{(h)} x_t + \cdots + A_p^{(h)} x_{t+1-p}.$$

In other words, $x_{t+h} = x_{t+h|t} + u_t^{(h)}$, where $u_t^{(h)} = u_{t+h} + \Phi_1 u_{t+h-1} + \cdots + \Phi_{h-1} u_{t+1}$ is the forecast error with mean zero and covariance matrix

$$\Sigma_x(h) = \sum_{i=0}^{h-1} \Phi_i \Sigma_u \Phi_i', \tag{2}$$

that is,

$$u_t^{(h)} \sim \mathcal{N}(0, \Sigma_x(h)). \tag{3}$$

Now suppose that the k-th component is the log of a variable y_t, i.e., $x_{kt} = \log y_t$, and forecasts of y_t are desired. A naive h-step ahead forecast for y_{t+h} may be based on $x_{k,t+h|t}$, the k-th component of $x_{t+h|t}$, as follows:

$$y_{t+h|t}^{nai} = \exp(x_{k,t+h|t}). \tag{4}$$

Granger and Newbold (1976) call this forecast naive because it is biased and it is not the optimal forecast. Using that

$$E(\exp x) = \exp(\mu + \tfrac{1}{2}\sigma_x^2),$$

if $x \sim \mathcal{N}(\mu, \sigma_x^2)$, it follows from the normality of the forecast error in (3) that

$$\begin{aligned} E_t(y_{t+h}) &= E_t[\exp(x_{k,t+h|t} + u_{kt}^{(h)})] = \exp(x_{k,t+h|t}) E_t(\exp u_{kt}^{(h)}) \\ &= \exp(x_{k,t+h|t} + \tfrac{1}{2}\sigma_{kk}^2(h)), \end{aligned}$$

where $\sigma_{kk}^2(h)$ is the k-th diagonal element of $\Sigma_x(h)$. Thus, the optimal predictor for y_{t+h} is

$$y_{t+h|t}^{opt} = \exp(x_{t+h|t} + \tfrac{1}{2}\sigma_{kk}^2(h)). \tag{5}$$

Hence, the optimal forecast differs from the naive forecast by a multiplicative adjustment factor $\exp(\frac{1}{2}\sigma_{kk}^2(h))$.

More generally, if a subvector of x_t consists of variables in logs and a product or ratio of the corresponding original variables, say $z_t = \exp(c'x_t)$, is of interest, where c is a suitable $(K \times 1)$ vector, a forecast of the relevant linear combination $c'x_t$ may be obtained and transformed. In that case, the naive forecast would be $z_{t+h|t}^{nai} = \exp(c'x_{t+h|t})$ and the corresponding optimal forecast becomes

$$z_{t+h|t}^{opt} = \exp(c'x_{t+h|t} + \tfrac{1}{2}c'\Sigma_x(h)c). \tag{6}$$

In economic models the residual variance of an equation for the log of a variable is typically small relative to the level of the variable. Moreover, the forecast error variance of the optimal forecast for the log of a stationary variable is bounded by the variance of the log of the variable when the forecast horizon goes to infinity. Therefore, for stationary economic variables the adjustment factor for the optimal forecast is typically small. It is worth emphasizing, however, that the derivations of the optimal forecast do not

require stationarity of the process x_t. Hence, integrated and cointegrated VARs are also permitted. If integrated processes are involved, the forecast error variance may be unbounded when $h \to \infty$. Thus, the adjustment factor may have a substantial impact on the optimal forecast for large forecast horizons.

In the simulations and the example section we assume that forecasting the ratio of the first two components of a vector y_t is of interest, that is, $z_t = y_{1t}/y_{2t}$. The log of the ratio is a cointegration relation in the data generation process (DGP) of x_t used in the simulations. In that case, the adjustment factor in (6) is bounded although x_{1t} and x_{2t} are integrated processes. Even for long-term forecasts the adjustment factor for the optimal forecast of z_t will hence be small.

In practice it is not clear that generally improvements in forecast precision can be obtained by using the optimal predictor. Notice that the adjustment factor relies on the normality of the forecast error which may not be satisfied. Moreover, the parameters and forecasts have to be replaced by estimated quantities which can make a difference, in particular, because the adjustment factor also has to be estimated. Estimation errors may have a small impact if stationary variables are considered and, hence, the adjustment factor for the optimal forecast is small. The situation may be different, however, for integrated variables. For them estimation errors in the forecast error variance may in fact be substantial. To see this, consider a univariate AR(1) process, $x_t = \nu + \alpha x_{t-1} + u_t$. For this process $\Phi_i = \alpha^i$. Hence, from (2) the h-step forecast error variance is seen to be $\sigma_u^2(1+\alpha^2+\cdots+\alpha^{2(h-1)})$, where σ_u^2 is the variance of u_t. If $|\alpha| < 1$ and, hence, the process is stationary, the powers of α go to zero. However, if $\alpha = 1$ and the process is a random walk, the estimated α may well be greater than one and, hence, substantial estimation errors may accumulate in the estimated forecast error variance based on such an estimate.

In the next section the relative performance of the naive and the optimal forecasts is explored under ideal conditions in a simulation environment to obtain a better impression of the possible gains or losses in forecast precision. In the light of these results we reconsider the example system used by Ariño and Franses (2000) in Section 4.

3 Monte Carlo Comparison of Forecasts

We simulate a 3-dimensional VAR(1) process,

$$x_t = \nu + A_1 x_{t-1} + u_t, \tag{7}$$

where $u_t \sim \mathcal{N}(0, \Sigma_u)$. We define y_t to be a 3-dimensional process consisting of the exponentials of the components of x_t, that is, $y_{it} = \exp x_{it}$, $i = 1, 2, 3$, and compute RMSEs of $y^{nai}_{t+h|t}$ and $y^{opt}_{t+h|t}$, varying the forecast horizon $h = 1, \ldots, 16$.

The VAR has the vector equilibrium or error correction model (VECM) representation

$$\begin{pmatrix} \Delta x_{1t} \\ \Delta x_{2t} \\ \Delta x_{3t} \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} - \begin{pmatrix} \alpha_{11} & 0 \\ 0 & 0 \\ 0 & \alpha_{32} \end{pmatrix} \begin{pmatrix} 1 & \beta_{12} & \beta_{13} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \\ x_{3,t-1} \end{pmatrix} + u_t, \quad (8)$$

highlighting the different time series properties of the variables: x_{1t} is cointegrated with x_{2t}, which is a random walk, while x_{3t} is a stationary process. Thus, the cointegration rank is two. Since the adjustment factor $\frac{1}{2}\sigma^2_{kk}(h)$ involves the residual variances, the elements of Σ_u may be of importance for the relative precision of the two forecasts. To isolate this factor across variables, we let $\Sigma_u = \sigma^2 I_3$, implying

$$\Sigma_x(h) = \sigma^2 \sum_{i=0}^{h-1} \Phi_i \Phi_i',$$

and vary $\sigma^2 = 0.001,\ 0.01,\ 0.02,\ 0.05$. The remaining parameters are fixed as $\nu_i = 0.02$, $\alpha_{11} = 0.1$, $\alpha_{32} = 0.5$, $\beta_{12} = -1$, and $\beta_{13} = 0.1$.[2] In particular, $x_{1t} - x_{2t}$ is a cointegration relation.

We use an effective sample size of 100 observations, discarding the 50 first to reduce start-up effects, and run 10,000 replications of the experiment. In each replication the lag length p is chosen by means of Schwarz's Bayesian Information Criterion (BIC) (Schwarz (1978)), the cointegration rank is tested with Johansen's likelihood ratio trace tests (Johansen (1995)) and the VECM with the corresponding number of cointegrated vectors is estimated by Johansen's reduced rank regression. We are interested in forecasts of $y_{it} = \exp(x_{it})$, $i = 1, 2, 3$. The estimated forecasts are based on (4) and (5), where all unknown parameters are replaced by estimates.

Since in a forecasting situation we are often also interested in functions of the modelled variables, we also compute forecasts of the ratio $z_t \equiv y_{1t}/y_{2t}$, which in our case corresponds to the cointegrated stationary combination of x_t variables. We consider forecasts $z^{nai}_{t+h|t} = y^{nai}_{1,t+h|t}/y^{nai}_{2,t+h|t}$ and $z^{opt}_{t+h|t}$ obtained from (6). Finally, to investigate the impact of specification and estimation on the performance of optimal forecasts, we also compute forecasts based on the true parameters.

[2] We also investigated the case of $\nu_i = 0$, but the results were qualitatively unaltered.

We first investigate the impact of specification and estimation variability on the optimal forecast. A comparison of optimal forecasts based on true and estimated parameters is summarized in Figure 1 which plots the RMSEs of the optimal forecasts based on true parameters relative to estimated optimal forecasts as a function of the forecast horizon with four different variances of the shocks to the processes. Notice that also specification uncertainty enters the estimated forecasts because the model order and the cointegration rank are data based while the true order and cointegration rank are used when true parameters are considered. Four conclusions emerge. First, the loss of forecasting precision due to estimation is negligible for the stationary variable y_{3t}. Second, for the nonstationary, integrated variables y_{1t} and y_{2t}, the negative effects of estimation have an increasingly negative impact on the forecast precision with increasing forecast horizon. Third, the larger the error variance (σ^2), the more negative is the effect of estimation on the performance of optimal forecasts. Finally, even though cointegration is imposed, estimation has an increasingly negative effect with growing horizon on the forecasts of the ratio, as shown for $z^{opt}_{t+h|t}$, denoted by z in Figure 1.

[Figure 1 about here.]

All these results are fully in line with what was to be expected by evaluating the optimal forecasting formula with the possible exception of the fourth observation. As mentioned in the previous section, the forecast error variance of a stationary variable is bounded and small relative to the level of the variable for our DGP. Therefore the estimation errors are also relatively small. This feature is in line with properties of economic variables in logs. In contrast, the forecast error variances for the integrated variables grow with the forecast horizon and are unbounded. Here estimation errors may grow with the forecast horizon, as mentioned in Section 2. Given that the cointegration relation is stationary, one may wonder why estimation errors become so important for the optimal forecast of z_t. These results reflect the fact that we use an estimated rather than true cointegration rank. In our DGP the cointegration relation is not very strong. The loading coefficient $\alpha_{11} = 0.1$ which is a speed of adjustment coefficient often found in empirical studies. On the other hand, the cointegration rank tests are known to have low power. Note that the implied A_1 matrix of our DGP with $\alpha_{11} = 0.1$ is

$$A_1 = \begin{pmatrix} 0.9 & 0.1 & -0.01 \\ 0 & 1.0 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}.$$

Its characteristic roots are 1.0, 0.9 and 0.5 and the true cointegration rank is two (one genuine cointegration relation and one stationary variable). Given

the low power of the cointegration rank tests, underestimation of the cointegration rank is quite likely which may lead to substantial estimation errors in the forecast error variance of the genuine cointegration relation. These estimation and specification errors are reflected in Figure 1.

[Figure 2 about here.]

With these conclusions in mind, we next turn to the relative performance of estimated naive to estimated optimal forecasts. The results are summarized in Figure 2 which plots the RMSEs of estimated naive forecasts relative to estimated optimal forecasts as a function of the forecast horizon with the same four different variances of the shocks to the processes as before. Five conclusions emerge. First, for the stationary variable there are no gains from using optimal forecasts at any forecast horizon. Second, for integrated variables, the naive forecasts generally perform better than optimal forecasts, the relative gains increasing with the forecast horizon. Third, in general, the worse the fit of the equations, the better are the naive relative to the estimated optimal forecasts. Fourth, the performance of the forecasts of the random walk y_{2t} have some benefits of optimal forecasting for shorter horizons. Fifth, for the stationary function of integrated variables, z_t, the naive forecasts are clearly superior to the optimal forecasts and the gain in forecast precision increases with the forecast horizon and the residual variance.

These results are not surprising, given the impact of the estimation error on the optimal forecast. Since the correction factor used in the optimal forecast is small, as usual for economic variables, the naive and optimal forecasts based on true parameters do not differ much. Hence, the rather substantial specification and estimation error in the optimal forecast for the integrated variables y_{1t} and y_{2t} as well as for the ratio z_t becomes important and affects the optimal forecasts negatively in particular for large forecast horizons.

The conclusions that emerge from this Monte Carlo study are that, in general, there are no gains from optimal forecasts relative to naive forecasts at any horizon—with the possible exception when forecasting integrated variables at short horizons. The next question is therefore whether these results from stylized data generating processes carry over when applied to real data and how they can be aligned with the substantial improvements obtained by Ariño and Franses (2000) from using the optimal forecast.

4 Empirical Example

To investigate the importance of optimal forecasts compared to naive forecasts for real data, we reexamine the example used by Ariño and Franses (2000). The data are quarterly U.S. series of real investment (y_{1t}) and real gross national product (GNP) (y_{2t}) for the period 1947(1)–1988(1).[3] Using data until 1980(4), Ariño and Franses (2000) find that the data are well represented by a VAR(3) in logs with one cointegration relation. The data are shown in Figure 3. Ariño and Franses (2000) estimate one pair of naive and optimal forecasts for each $h = 1, \ldots, 29$ and evaluate them by taking averages of various error measures over h horizons, so, for example, the RMSE is computed as

$$RMSE^{AF} = \sqrt{\frac{1}{29}\sum_{h=1}^{29}\left(y_{t+h} - f_{t+h|t}\right)^2},$$

where $f_{t+h|t} = y^{nai}_{t+h|t}$ or $y^{opt}_{t+h|t}$.

[Figure 3 about here.]

To investigate the forecasting properties as a function of the forecasting horizon, we choose a different strategy. Starting with a sample of 100 observations, the forecasts $f_{t+h|t}$, $h = 1, \ldots, 16$, are computed recursively, increasing the sample by one period and redoing the estimation and forecasting over an evaluation period of 65 quarters at the end of the sample. The RMSE at forecast horizon h is then computed as

$$RMSE\,(h) = \sqrt{\frac{1}{66-h}\sum_{i=1}^{66-h}\left(y_{t+i+h} - f_{t+i+h|t+i}\right)^2},\ h = 1, \ldots, 16, \quad (9)$$

with $f_{t+h|t} = y^{nai}_{t+h|t}$ or $y^{opt}_{t+h|t}$, as before. The system is reestimated for each sample size and, as in the Monte Carlo, the lag length p is chosen by means of the BIC, the cointegration rank is tested and the system is estimated by reduced rank regression with the corresponding number of cointegration vectors. The estimated forecasts are based on (4) and (5), replacing unknown parameters by estimates. We also compute forecasts of the investment-GNP ratio, $z_t \equiv y_{1t}/y_{2t}$.

[3]The data set corresponds to Table 13.5 in Pindyck and Rubinfeld (1998), but with the series starting in 1947. The data are available at `http://www.estima.com/textbookindex.shtml`.

[Figure 4 about here.]

We present the ratio of RMSEs of naive to optimal forecasts for the estimated models as a function of the forecast horizon in Figure 4. The following results are apparent: First, for the seemingly stationary variable z_t, there are no or at best very small gains of using optimal forecasts at any forecast horizon. Second, for the integrated variables y_{1t} and y_{2t}, the relative performance of the naive forecasts improves with the forecast horizon. They generally perform better than optimal forecasts except for short horizons where both have very similar RMSEs. Hence, the results of the example model mimic those from the Monte Carlo study. Thus, the gains from using the optimal forecast reported by Ariño and Franses (2000) are an artefact of their specific way to compute RMSEs.

5 Conclusions

In this study we have considered forecasting levels variables which appear in logs in a VAR or VECM. Theory asserts that forecasting the log variable and then converting it to a 'naive' forecast of the original variable by applying the exponential function is not optimal. We have derived a simple expression for the optimal forecast which has enabled us to investigate possible factors which may lead to gains from using the optimal forecast. We have found that for typical economic variables substantial RMSE gains cannot be expected even theoretically from using the optimal forecast.

The situation is even worse in practice where forecasts have to be based on processes which are specified and estimated from data. In a controlled simulation experiment we have shown that in this case the optimal forecast will rarely result in RMSE reductions relative to the naive forecast. In fact, for stationary variables, including transformations based on cointegration relations no gains can be expected from using the optimal forecast when specification and estimation errors are accounted for. For integrated variables we found small improvements from using the optimal forecast for short horizons whereas substantial losses may occur at longer horizons. These features are also obtained for an example based on quarterly U.S. investment and GNP data. Our results suggest that in applied work using the naive forecast is the preferred option.

References

Ariño, M. A. and P. H. Franses (2000). Forecasting the levels of vector autoregressive log-transformed time series. *International Journal of Forecasting 16*, 111–116.

Granger, C. W. J. and P. Newbold (1976). Forecasting transformed series. *Journal of the Royal Statistical Society B 38*, 189–203.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Lütkepohl, H. and F. Xu (2011). The role of the log transformation in forecasting economic variables. *Empirical Economics forthcoming*.

Pindyck, R. S. and D. L. Rubinfeld (1998). *Econometric Models and Economic Forecasts* (4 ed.). New York: McGraw-Hill, Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

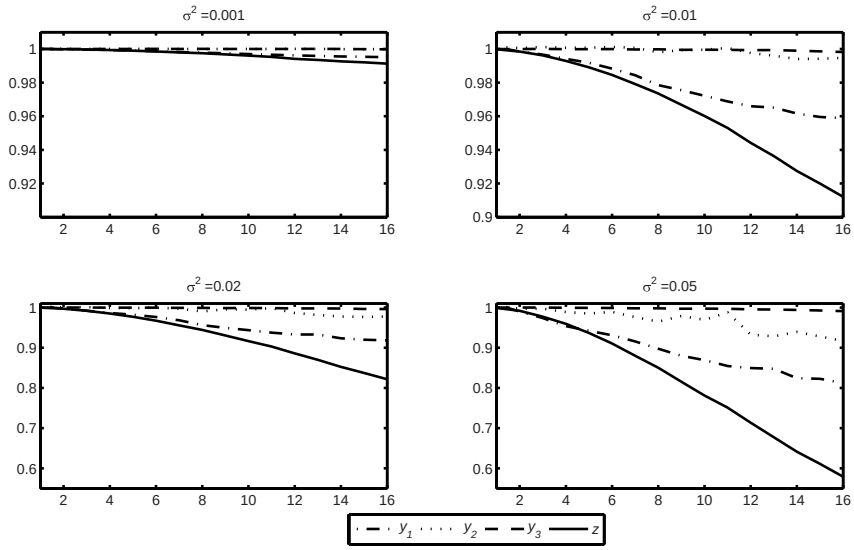

Figure 1: RMSEs of true optimal relative to estimated optimal h-step forecasts for y_t and $z_t = y_{1t}/y_{2t}$ with deterministic terms $\nu_i = 0.02$, varying the covariance matrix of the residuals.

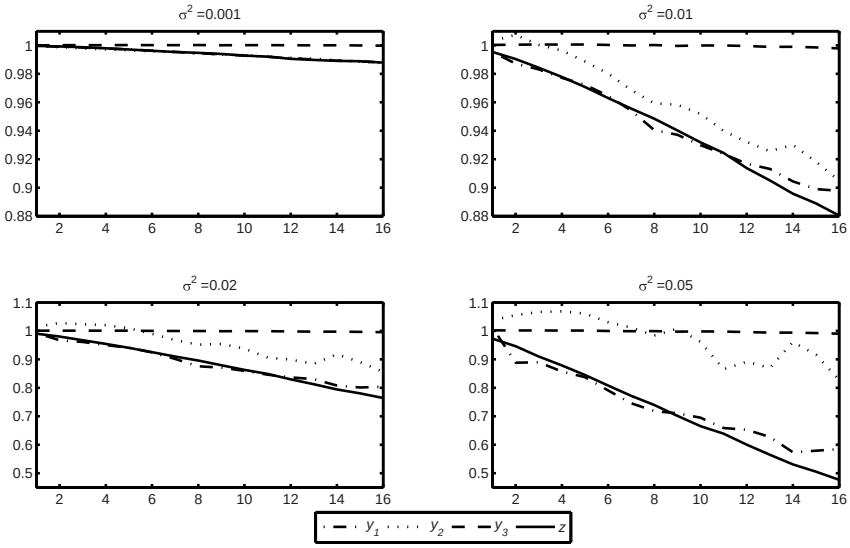

Figure 2: RMSEs of estimated naive relative to estimated optimal h-step forecasts for y_t and $z_t = y_{1t}/y_{2t}$ with $\nu_i = 0.02$, varying the covariance matrix of the residuals.

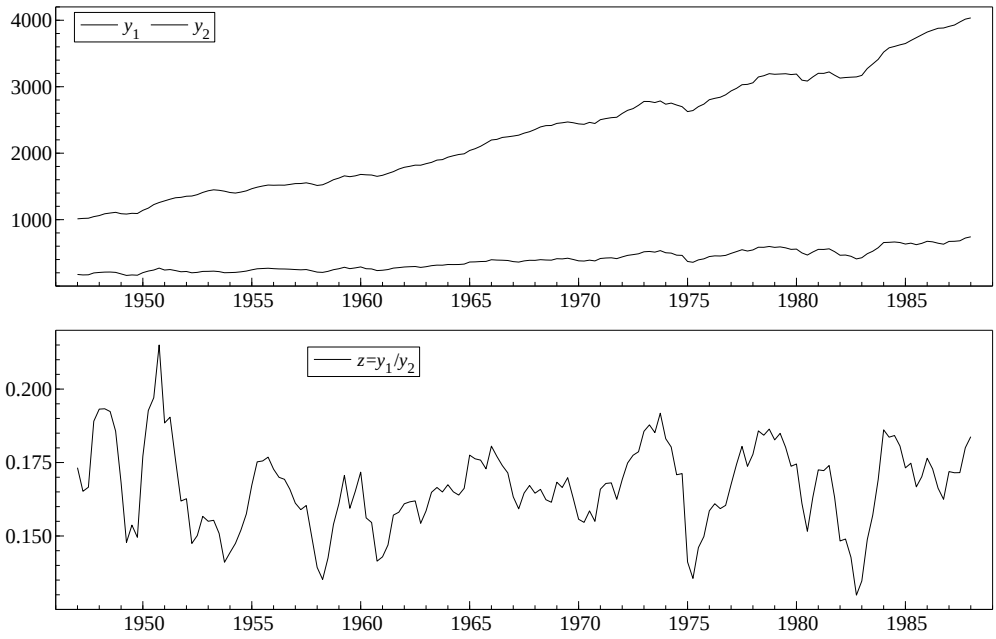

Figure 3: Levels (y_i) of U.S. real investment ($i = 1$) and real GNP ($i = 2$) and the investment/GDP ratio (z).

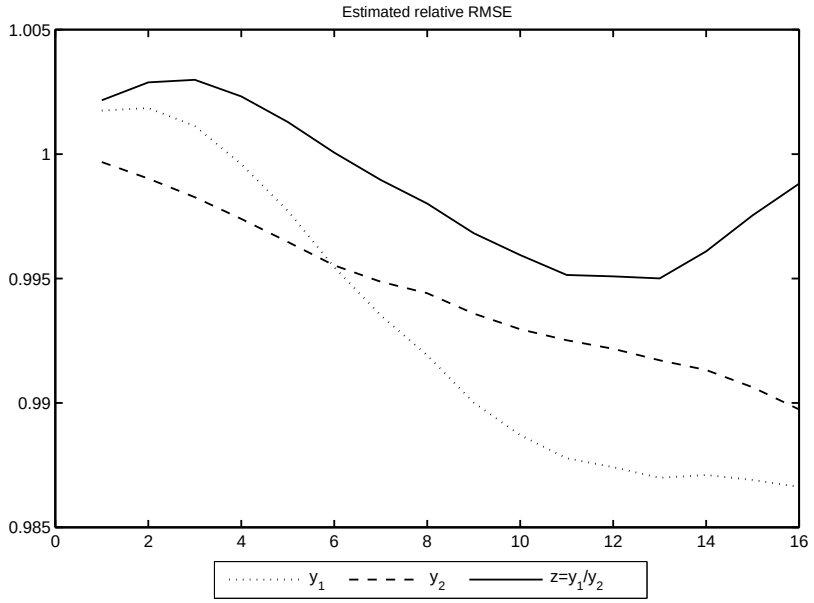

Figure 4: RMSEs of naive relative to estimated optimal h-step forecasts for y_1, y_2 and $z \equiv y_1/y_2$, $h = 1, \ldots, 16$ for the U.S. data.