Macroeconometric modelling

Estimating logs and forecasting levels

Gunnar Bårdsen

CREATES

16-17 November 2009

Overview

Joint work with Helmut Lütkepohl

- System of variables estimated in logs, but need forecasts of (functions) of levels.
- Naive forecast: Convert the log forecast by taking exponential is not optimal theoretically.
- A simple expression for the optimal forecast under normality assumptions is derived.
- Monte Carlo performance of system capturing typical applications
- Revisiting empirical study

Forecasts of VAR

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

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

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

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

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

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

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

Forecasts of VAR

continued

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,

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

Forecasts of levels of log transformed variables

Let the k-th component be 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:

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

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 $\Sigma_x(h)$.

Forecasts of levels of log transformed variables

continued

The optimal predictor for y_{t+h} is

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

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

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

Comments

Factors to consider when using $y_{t+h|t}^{opt}$ instead of $y_{t+h|t}^{nai}$

- Depends upon the normal distribution approximates the log variable well
- the parameters and adjustment factors have to be replaced by estimated quantities.
- the residual variance in estimated models is typically small relative to the level of the variable.
- the forecast error variance of the optimal forecast for a stationary variable is bounded by the variance of the log of the variable when the forecast horizon goes to infinity.
  - the adjustment factor for the optimal forecast is likely to be small.
- the forecast error variance may be unbounded when $h \longrightarrow \infty$. for integrated processes
  - The adjustment factor may have a substantial impact on the optimal forecast for large forecast horizons.

Considerations

non-stationarity

To see last point, 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.

Considerations

cointegration

- Assume that forecasting the ratio of the first two components of a vector y_t is of interest, that is, $z_t = y_{1t}/y_{2t}$. Let the log of the ratio be a cointegration relation. In that case, the adjustment factor in (6) is bounded although x_{1t} and x_{2t} are integrated processes.
- Even for long-term forcasts the adjustment factor for the optimal forecast of z_t will hence be small.

Monte Carlo Comparison of Forecasts

Linearization (continued)

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

$$x_t = \nu + A_1 x_{t-1} + u_t, \, u_t \sim \mathcal{N}(0, \Sigma_u) \tag{7}$$

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

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

x_{1t} is cointegrated with x_{2t}, which is a random walk, while x_{3t} is a stationary process. We define $y_{it} = \exp x_{it}$, $i = 1, 2, 3$, and compute RMSEs of $y^{nai}_{t+h|t}$ and $y^{opt}_{t+h|t}$, varying the forecast horizon $h = 1, \ldots, 16$.

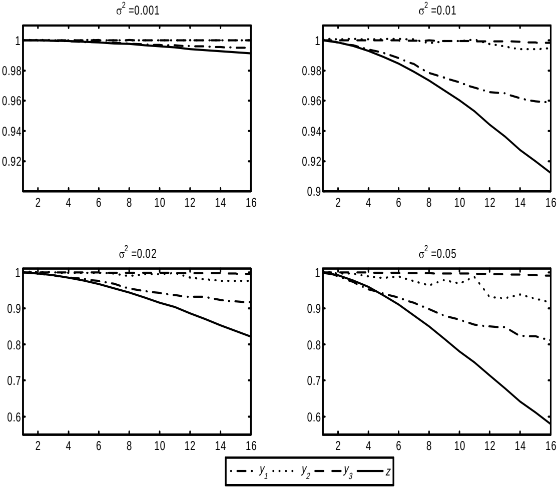
Since the adjustment factor $\frac{1}{2}\sigma^2_{kk}(h)$ involves the residual variances, the elements of Σ_u may be of importance for the relative precision of the two forecasts. To isolate this factor across variables, we let $\Sigma_u = \sigma^2 I_3$, implying

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

and vary $\sigma^2 = 0.001,\ 0.01,\ 0.02,\ 0.05$. We also investigate the impact of deterministic drifts by using $\nu_i = 0$ and 0.02. The role of cointegration is controlled by considering $\alpha_{11} = 0.1$ and 0.5, while we keep the remaining parameters fixed as $\alpha_{32} = 0.5$, $\beta_{12} = -1$, and $\beta_{13} = 0.1$. In particular, $x_{1t} - x_{2t}$ is a cointegration relation.

Monte Carlo Comparison of Forecasts

RMSEs of true optimal relative to estimated optimal h-step forecasts for y_{it} and $z_t = y_{1t}/y_{2t}$ with deterministic terms $\nu_i = 0.02$ and cointegration adjustment $\alpha_{11} = 0.1$, varying the covariance matrix of the residuals.



Conclusions:

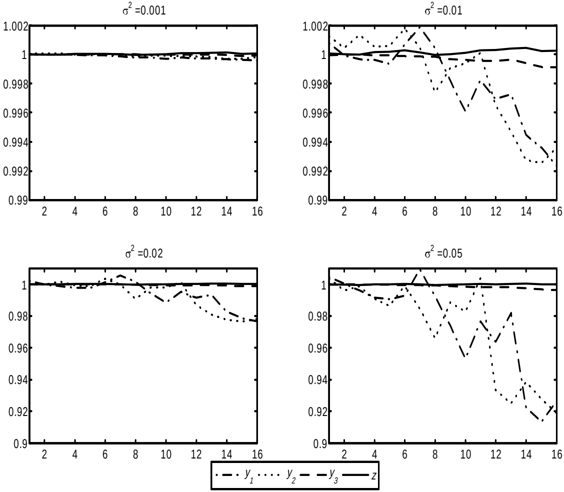
1. the loss of precision is negligible for y_{3t}.
2. for y_{1t} and y_{2t} effects are noticable.
3. increasing with error variance (σ^2).
4. for $z^{opt}_{t+h|t}$, effects are noticeable, despite cointegration imposed

Monte Carlo Comparison of Forecasts

- forecast error variance of a stationary variable is bounded and small relative to the level of the variable
  - Therefore the estimation errors are also relatively small.
  - This feature is in line with properties of economic variables in logs.
- forecast error variances for the integrated variables grow with the forecast horizon and are unbounded.
- z_t is stationary, why estimation errors so important for optimal forecasts
  - estimated rather than true cointegration rank.
  - cointegration relation is not very strong. The loading coefficient $\alpha_{11} = 0.1$: characteristic roots are 1.0, 0.9 and 0.5.
  - underestimation of rank is quite likely may lead to estimation errors in the forecast error variance of z_t
- increase this coefficient to the unrealistically high value of $\alpha_{11} = 0.5$, implying characteristic roots of 1.0, 0.5 and 0.5

Monte Carlo Comparison of Forecasts

RMSEs of true optimal relative to estimated optimal h-step forecasts for y_{it} and $z_t = y_{1t}/y_{2t}$ with deterministic terms $\nu_i = 0.02$ and cointegration adjustment $\alpha_{11} = 0.5$, varying the covariance matrix of the residuals.

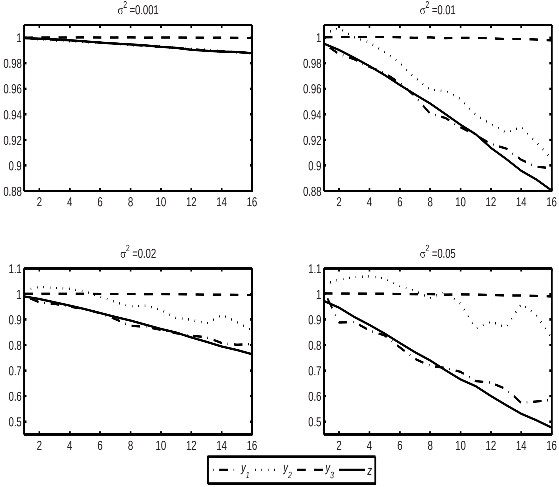


Conclusions:

1. the nonstationary variable y_{1t} now behaves like its common trend y_{2t}
2. the ratio z_t behaves as y_{3t}.
3. shows importance of correct choice of cointegration rank in longer horizon forecasting.

Monte Carlo Comparison of Forecasts

RMSEs of naive relative to estimated optimal h-step forecasts for y_{it} and $z_t = y_{1t}/y_{2t}$ with deterministic terms $\nu_i = 0.02$ and cointegration adjustment $\alpha_{11} = 0.1$, varying the covariance matrix of the residuals.

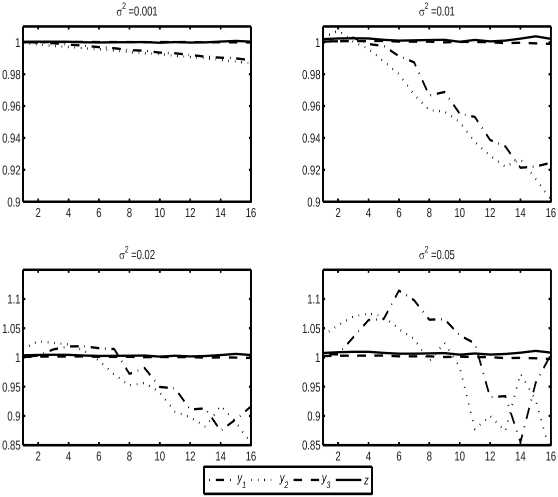


Conclusions:

1. there is no gain for y_{3t}.
2. for y_{1t} and y_{2t} naive forecasts perform better.
3. naive performance increasing with error variance (σ^2).
4. for y_{2t} optimal forecasts perform better for short horizons.
5. for $z^{opt}_{t+h|t}$, naive forecasts relatively

Monte Carlo Comparison of Forecasts

RMSEs of naive relative to estimated optimal h-step forecasts for y_{it} and $z_t = y_{1t}/y_{2t}$ with deterministic terms $\nu_i = 0.02$ and cointegration adjustment $\alpha_{11} = 0.5$, varying the covariance matrix of the residuals.



Conclusions:

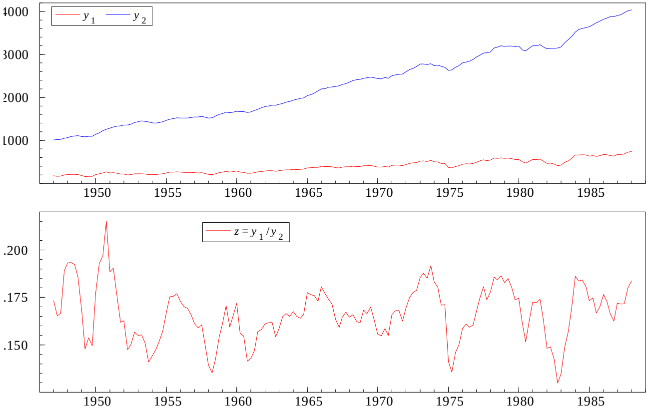
1. the nonstationary variable y_{1t} now behaves like its common trend y_{2t}
2. the ratio z_t behaves as y_{3t}.
3. correct choice of cointegration rank important.

- Ariño and Franses (2000) emphasize the forecast efficiency gains obtained from using $y^{opt}_{t+h|t}$ instead of $y^{nai}_{t+h|t}$.
  1. their results are based on one dataset only and only one forecast period is used
  2. they compute one set of forecasts only and compute the RMSE across forecasts of different horizons which may not be the best way to average out estimation and specification errors.
- results obtained by Ariño and Franses (2000) are very special and not representative.

Reexamination

Frame Subtitle

Quarterly U.S. series of real investment (y_{1t}) and real gross national product (GNP) (y_{2t}) for the period 1947(1)–1988(1). Using data until 1980(4), estimate VAR(3) in logs with one cointegration relation.



Reexamination

Computing of RMSE

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 $y^{opt}_{t+h|t}$.

Reexamination

Computing of RMSE

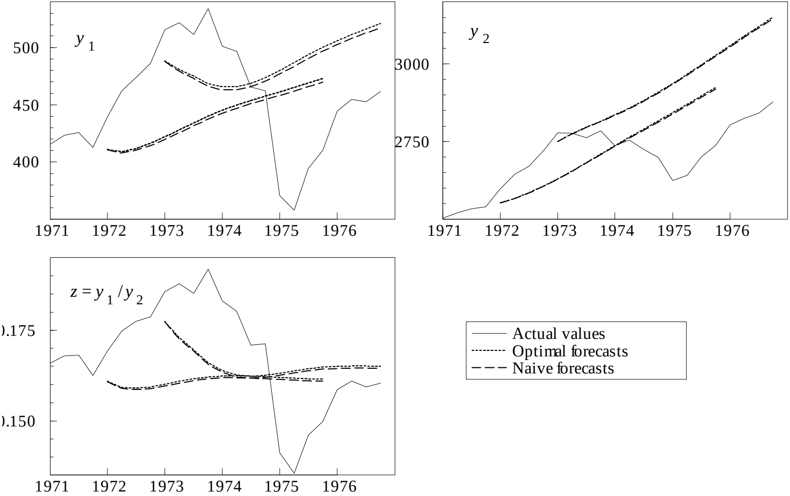
Starting with a sample of 100 observations, the forecasts $f_{t+h|t}$, $h = 1, \ldots, 16$, are computed recursively, increasing the sample by one period and redoing the estimation and forecasting over an evaluation period of 65 quarters at the end of the sample. The RMSE at forecast horizon h is then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

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

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, $z_t \equiv y_{1t}/y_{2t}$.

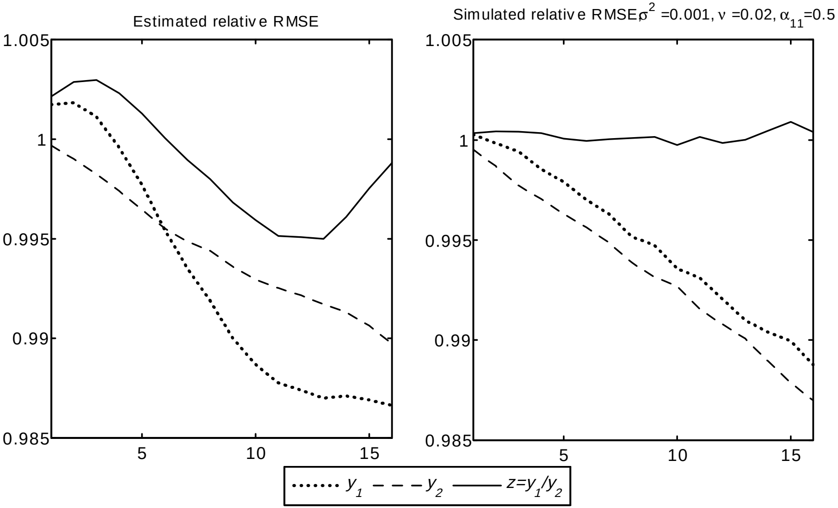
Examples of forecasts

Naive and optimal forecasts starting in 1972 and 1973



Comparing Monte Carlo with empirical performance

Estimated and simulated naive relative to optimal RMSE



Conclusions

- The optimal forecast is inferior (in most cases) to the naive forecast in Monte Carlo
- In practice using the naive forecast is preferable to using the optimal forecast.

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.

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