Journal of Multivariate Analysis 136 (2015) 108–125

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Evaluating panel data forecasts under independent realization[☆]

Ryan Greenaway-McGrevy

Department of Economics, University of Auckland Business School, Owen G Glenn Building, 12 Grafton Road, Auckland 1010, New Zealand

ARTICLE INFO

Article history:
Received 6 June 2014
Available online 21 January 2015

AMS subject classifications:
62M10
62M20

Keywords:
Forecasting
Panel data
Same-sample realization
Independent realization
Vector autoregression

ABSTRACT

Independent realization is a commonly used shortcut for deriving forecast properties. It is also an unrealistic assumption in many empirical applications. In this paper we consider the effect of the assumption when deriving the properties of panel data forecasts. To do so, we derive and compare the asymptotic forecast loss of a set of vector autoregressions (VARs) under both independent and same-sample realization. We show that adopting the independent realization assumption can result in an overstatement of forecast loss when common forms of parameterized heterogeneity (such as fixed effects) are included in the VAR. Because these parameters are often used in dynamic panel forecasting applications, our results imply that the independent realization shortcut should only be used with caution when working with panel data.

1. Introduction

One of the primary obstacles to deriving the properties of out-of-sample forecasts is the potential dependence between the variable to be forecast and the sample used in model estimation. Often the dependence problem is circumvented altogether by using the so-called *independent realization* (IR) condition, whereby the forecast variable is assumed to be a statistically independent replicate of the process used for model estimation.[1] Such an assumption is unnatural in many applications. For example, in autoregressive (AR) time series forecasting, future realizations of the time series are dependent on earlier data used to estimate the AR model.

A more appropriate condition in many applications is *same-sample realization* (SSR), whereby the variable to be forecast is generated by the same process as the data used to estimate the model. However, under sufficient restrictions on the dependence in the data, the IR assumption often delivers a sufficiently accurate approximation of SSR forecast loss. For a class of short memory processes, Ing and Wei [16,17] show that the mean square forecast error (MSFE) of a least squares autoregression under IR is equivalent to the MSFE under SSR up to an $o\left(kT^{-1}\right)$ approximation. (Here k denotes the lag order of the autoregression, and T denotes the number of available time series in the estimation sample.) For stable AR(1) processes, Phillips [23] shows that the IR and SSR forecast errors are equivalent up to an $o\left(T^{-1}\right)$ approximation. Thus, although unnatural in many applications, the IR assumption is a potentially useful shortcut for deriving some of the properties of same sample forecasts.

☆ The author thanks Yong Bao, Ben Bridgman, Kyle Hood and Donggyu Sul for their comments on a previous version of the paper.
E-mail address: r.mcgrevy@auckland.ac.nz.

1 See, amongst others, Yamamoto [27], Baillie [1], Shibata [26], Reinsel [24], Lewis and Reinsel [20,21], Bhansali [3–5], Karagrigoriou [18], Findley and Wei [10] and Schorfheide [25].

http://dx.doi.org/10.1016/j.jmva.2015.01.004

Panel data models are increasingly being used in empirical forecasting applications. Panel data models that permit limited forms of cross sectional heterogeneity typically produce more accurate forecasts than the corresponding time series specification. For a detailed survey, see the "Forecasting Applications" section of [2]. In this paper we consider whether the IR assumption can also be employed when deriving the properties of these panel data forecasts. We consider forecasts generated from a set of fitted panel vector autoregressions (VARs) of the general form

$$y_{i,t+1} = \sum_{r=1}^{p} \beta_{i,r} \left(t-k+1\right)^{(r-1)} + \sum_{s=1}^{k} \alpha_s' y_{i,t-s+1} + e_{i,t}, \quad i = 1, \ldots, n;\ t = k, \ldots, T-1. \tag{1}$$

Dependence within each time series is modeled through a homogeneous vector autoregressive structure (i.e., $\{\alpha_s\}_{s=1}^{k}$ are the same for each time series in the panel). The parameter space is restricted so that each time series in the panel is a short memory process. Limited heterogeneity is permitted through a set of polynomial time trends interacting with cross section specific coefficients. Because it permits cross sectional fixed effects, lagged (in time) dependent variables and other weakly exogenous regressors, the VAR (k) given in (1) nests many of the models used in the empirical applications cited in [2].

We derive asymptotic expressions for the quadratic forecast loss of an ordinary least squares (OLS) fitting of (1) under both independent and same sample realization. Asymptotics are derived as both n (number of cross sections) and T (number of time series) jointly tend to infinity (denoted as $n, T \to \infty$). The fitted models are used to predict the realization of each cross sectional unit in the next time period, and forecast risk is evaluated by an average out-of-sample quadratic forecast error loss function (with the average taken over the cross sections in the panel), henceforth referred to as mean square forecast error (MSFE). Because expressions of out-of-sample loss are often used to analyze the effect of model specification on forecast accuracy, we derive the asymptotic expressions up to the largest terms that are a function of k (i.e., the lag order of the fitted VAR).

We show that the asymptotic MSFE under SSR is smaller than the asymptotic MSFE under IR across the permissible parameter space, provided that n grows at a rate no slower than T. Because SSR is a more realistic assumption in most empirical applications, this means that derivations based on the IR assumption will tend to over-state forecast loss in practice. This has severe practical implications. For example, model selection based on forecast loss expressions derived under IR would tend to underfit the model, leading to suboptimal forecasting. When n grows at a rate slower than T, the asymptotic MSFEs are shown to be equivalent.

The difference in MSFEs arises because the *within group transformation* used in OLS to partial out the cross-section specific parameters $\left\{\beta_{i,r}\right\}_{r=1}^{p}$ induces correlation between the transformed regressors and regression errors within each time series comprising the panel. The dependence is not weak in a conventional mixing sense: The correlation does not get smaller as the distance between time series observations grows larger. The correlation does however approach zero (in a certain sense) as T grows large. Under asymptotic sequences in which n is at least of the same order of magnitude as T, the correlation is strong enough to have an effect on the relevant approximation of the asymptotic MSFEs.

These asymptotic expressions naturally depend to some extent on the assumptions imposed on the panel process and the set of models to be fitted. In this regard we follow [26,4,16,17], who consider fitting an *AR* (k) model to an infinite order autoregressive process, permitting the set of fitted lag orders to grow with the sample size T at a $o(T^{1/2})$ rate. Similarly, our data generating process is an infinite order VAR, and we allow the maximum lag order (denoted $k_{n,T}$) to grow in the asymptotics at a restricted rate satisfying $k_{n,T}(T^{-1} + n^{-1/2}) \to 0$. Under this framework we derive an analytic expression that uniformly (in $k \leq k_{n,T}$) approximates the MSFEs of the LS fittings.

The remainder of the paper is organized as follows. In the following section we outline the data generating process underlying the panels. In Section 3 we derive asymptotic expressions for the MSFE of the SSR and IR forecasts, and discuss the difference between the SSR and IR MSFEs. In Section 4 we conduct a small Monte Carlo study in order to validate the asymptotic theory. Section 5 concludes. Throughout, "$:=$" is used as the definitional equality; C denotes an arbitrary finite constant that may take on different values in different places; $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix; $\|\cdot\|$ denotes the spectral norm; and 1_h denotes a $h \times 1$ vector of ones, and I_h denotes a $h \times h$ identity matrix, for some arbitrary integer $h \geq 1$. Proofs are contained in the Appendix.

2. Assumptions and preliminaries

The set of candidate forecasting models under consideration is given by (1), where $k = 1, \ldots, k_{n,T}$ indexes the different models. We outline the conditions imposed of the panel process $y_{i,t}$ in the following subsection before introducing the LS estimator and the conditions imposed on the maximum lag order $k_{n,T}$.

2.1. Assumptions

The data generating process for the $m \times 1$ vector $y_{i,t}$ is

$$y_{i,t} = x_{i,t} + \sum_{r=1}^{p} \beta_{i,r}^{*} \left(t-1\right)^{(r-1)}, \quad x_{i,t} = \sum_{s=1}^{\infty} \alpha_s' x_{i,t-s} + e_{i,t}. \tag{2}$$