Panel Data Forecasting*

Andreas Pick[†] Allan Timmermann[‡]

April 24, 2024

Abstract

A variety of forecasting methods are available for exploiting the time-series and cross-sectional dimensions (T and N) of panel data, including fixed and random effects estimation, Bayesian methods, pooling, and forecast combination. We discuss how the predictive accuracy of these methods depends on T and N as well as the importance of the type and magnitude of heterogeneity in the model parameters that regulate the bias-variance trade-off in the forecasting problem. Finally, we discuss choices of loss function and methods for assessing the accuracy of panel data forecasts.

JEL Classification: C53

Keywords: Panel data; forecast combination; Bayesian estimation; bias-variance trade-off.

1 Introduction

Economic forecasting experiments often take off from the univariate ARIMA models developed and popularized in the influential work of Box and Jenkins (1970). This practice has brought clear advantages from a computational perspective since least squares estimation of the parameters of autoregressive models is simple and efficient algorithms exist for this purpose.[1] Moreover,

*We thank Annika Camehl, Aisha Schmidt, and the editors, Michael Clements and Ana Galvao, for comments on an earlier draft. We also thank Hashem Pesaran for numerous discussions, which have helped shape our views and understanding of panel data modeling.

[†]Erasmus University Rotterdam and Tinbergen Institute

[‡]Rady School of Management and Department of Economics, UCSD

[1]Estimation of models with moving average dynamics is more complex and requires maximum likelihood methods.

the econometric properties of such estimators are well understood in a variety of settings with data displaying even strong forms of serial persistence (autocorrelation) and time-varying volatility (heteroskedasticity), including unit root dynamics and autoregressive conditional heteroskedasticity. Seasonal patterns and trends can also be tailored to the individual time series. Yet, this flexibility can come at the cost of imprecisely estimated parameters in settings where the data have a short time-series dimension (small T) and estimation error can lead to poor out-of-sample forecast accuracy. Additionally, the classical univariate modeling approach does not attempt to exploit any linkages or constraints on the model parameters that may hold across multiple time series.

Following the influential work of Sims (1980), univariate forecasting models have been broadened to multi-variate time-series models such as vector autorregressions (VARs). Due to their flexible modeling of dynamic interaction effects, these models typically require estimating vastly more parameters than univariate models. This can again lead to deterioration in predictive accuracy in cases where the time-series dimension is short. Bayesian estimation methods have been developed to deal with this issue. These methods generally reduce the impact of estimation error through "Minnesota" priors which shrink posterior parameter estimates towards some ex-ante motivated values (Litterman, 1980). Still, these methods tend to be applied mostly to instances in which the time-series dimension of the data (T) is relatively large and the cross-sectional dimension (N) is modest.

More recently, continued progress in data recording and storage along with the arrival of increasingly powerful computational engines have fostered considerable progress in forecasting methods that apply flexible machine learning methods to high-dimensional multivariate data (large N). However, establishing the theoretical properties of the estimates of these models, let alone analytically characterizing the accuracy of the resulting forecasts is difficult in settings with nonlinear mappings from predictors to outcomes for data that are not strictly exogenous and often display strong forms of serial persistence and even non-stationarities. Convergence rates for the estimates of the parameters of these high-dimensional models are often sufficiently slow that estimation error is a major concern unless N is very large.[2]

Panel forecasting models offer a potential mid-point between the traditional univariate or low-dimensional VAR models and modern high-dimensional

[2]Machine learning methods offer the potential of flexibly capturing nonlinearities but outside settings with independently and identically distributed data it is difficult to establish analytically the properties of the resulting estimates and forecasts.

machine learning methods. Panel models are typically assumed to be linear and can handle data sets with both high time-series and cross-sectional dimensions. Linearity of the assumed forecasting model makes it feasible to establish the properties of the subsequent forecasts (e.g., bias) and to characterize measures of forecasting performance such as mean squared forecast errors analytically. Moreover, panel designs such as two-way fixed effects can be used to capture time-invariant sources of heterogeneity and incorporate the effect of common shocks. In parallel with VAR models, dynamic panels can capture autoregressive predictive effects, although they generally do not allow for the same rich lead-lag patterns across variables as VARs are designed to do. Conversely, panel models tend to be relatively parsimonious, requiring the estimation of considerably fewer parameters than VARs fitted to the same set of variables exactly because they typically do not account for an unrestricted set of dynamic lead-lag cross-effects.

Panel forecasting methods can be particularly attractive in settings with a small or medium-sized time-series (T) dimension. Indeed, dynamic panel forecasting models can be used, when implemented with empirical Bayes and shrinkage methods, in situations with a small time-series dimension, see Liu et al. (2020) and Giacomini et al. (2023). A large cross-sectional (N) dimension can facilitate more accurate parameter estimates in settings with high degrees of homogeneity in the model parameters yet enough independent variation in the predictors included in the model.

In this review, we cover a number of topics related to panel forecasting. We begin in Section 2 by examining the determinants of the accuracy of forecasts based on unit-specific regressions, pooled regressions and random effect models. Section 3 extends the analysis to cover forecast combinations based on univariate and pooled forecasts. Alternative forecasting methods such as Bayesian estimation, panel and global VAR models and machine learning models are covered in Section 4. Choice of loss function and evaluation of panel data forecasts is discussed in Section 5. Finally, Section 6 concludes.

2 Bias-Variance Trade-offs in Panel Forecasting

Due to their linear form, panel forecasting models make it easier to analyze the bias-variance trade-offs involved in exploiting cross-sectional information and commonalities across individual units to obtain more accurate forecasts. These trade-offs are determined chiefly by the degree of parameter heterogeneity across individual series: If the model parameters are identical across units, a simple pooled estimation scheme is optimal. Conversely, the presence of strong parameter heterogeneity makes it more attractive to use unit-specific estimation and proceed on a unit by unit basis, assuming that the data has a minimal T dimension that allows reasonably accurate estimation of the required parameters. If heterogeneity is largely due to unit-specific effects, a random effects or fixed effects framework can be adopted.

We begin our analysis by discussing these points and highlighting the main determinants of which panel forecasting methods can be expected to perform well in a given empirical application. This analysis offers practical advice for applied economists with an interest in choosing among different panel forecasting approaches.

2.1 Panel Regression Model

Our starting point is the linear panel regression model:

$$y_{it} = \boldsymbol{\beta}_i' \mathbf{x}_{it} + \varepsilon_{it}, \tag{1}$$

where $i = 1, 2, \ldots, N$ refers to the individual units and $t = 1, 2, \ldots, T$ refers to the time period, y_{it} is the outcome of unit i at time t, $\mathbf{x}_{it}$ is a $K \times 1$ vector of regressors—or predictors—used to forecast y_{it}, $\boldsymbol{\beta}_i$ is the associated vector of regression coefficients, and ε_{it} is the disturbances of unit i in period t. For simplicity, we suppress any references to forecast horizon in our notation and simply assume that data up to time T is used to generate a one-step-ahead forecast of $y_{i,T+1}$. However, insofar that a direct forecasting approach is being used, this setup can easily be generalized to arbitrary forecast horizons, h, by requiring that $\mathbf{x}_{it}$ contains variables that are observed with the appropriate lag, i.e., variables known at time $t-h$ for an h-period forecast horizon.

Before proceeding, we introduce some notations. Stacking the time series of outcomes, predictors and disturbances, define $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}', \mathbf{x}_{i2}', \ldots, \mathbf{x}_{iT}')'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$. Further, let $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N')'$, $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_N')'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_N')'$.

The specification in (1) includes as a special case the canonical dynamic panel model studied, in some form, by Baltagi (2005), Trapani and Urga (2009), Liu et al. (2020) and other authors:

$$y_{it} = \beta_{0i} + \beta_{1i} y_{i,t-1} + \beta_{2i} x_{it} + \epsilon_{it}. \tag{2}$$

It is common in the panel forecasting literature to consider three broad classes of panel estimators, namely homogeneous (pooled), heterogeneous, and shrinkage estimators, the latter often implemented in a Bayesian setting. Homogeneous estimators reduce the impact of estimation error by pooling information across units in the cross-section. The benefit from this tends to be strongest when the time-series dimension, T, is small and the cross-sectional dimension, N is large. When the underlying regression parameters display considerable heterogeneity, however, the benefits from pooling come at the cost of biasing individual estimates which will lead to a deterioration in forecasting performance.[3]

Early empirical evidence indicated that pooled estimators often produce better forecasts than heterogeneous estimates (Baltagi and Griffin, 1997; Baltagi et al., 2002; Rapach and Wohar, 2004). Other studies have found that the ranking of the performance of different panel forecasting approaches varies across applications. Trapani and Urga (2009) use Monte Carlo simulations to show that the ranking of forecasts based on heterogeneous versus homogeneous estimators depends on the level of parameter heterogeneity. In particular, homogeneous estimators perform best if the degree of parameter heterogeneity and dependence across units are both low. If the degree of heterogeneity is high, the pooled estimator ceases to produce accurate forecasts and the relative performance of heterogeneous and Bayesian estimators tends to improve, with the latter performing particularly well in the presence of strong cross-sectional dependencies. Similarly, Brucker and Siliverstovs (2006) find that a conventional fixed effect estimator along with a hierarchical Bayes estimator produce more accurate out-of-sample forecasts of international migration than the pooled OLS estimator. In turn, these estimators tend to produce better forecasts of international migration than heterogeneous estimators. Finally, Garcia-Ferrer et al. (1987) empirically investigate the bias-variance trade-off in the context of predicting growth rates for the US and EU countries. Their findings suggest that for some series individual forecast were more precise whereas others benefited from pooling.

To examine the trade-off between pooled and individual estimation, we will make a number of simplifying assumptions designed to make the analysis particularly transparent. First, we will assume that the error term $\boldsymbol{\varepsilon}_i \sim \text{iid}(\mathbf{0}, \sigma_i^2 \mathbf{I}_T)$ with a finite variance and that the error term is (conditionally) uncorrelated across individual units. Additionally, we assume that the regressors are exogenous with finite and positive definite sample covariance matrices, $\mathbf{Q}_{iT} = T^{-1}\mathbf{X}_i'\mathbf{X}_i$ and $\mathbf{Q}_{NT} = T^{-1}N^{-1}\mathbf{X}'\mathbf{X}$.

[3]Notice, however, that the individual estimates will also be biased in the presence of lagged dependent variables, see Pesaran and Smith (1995).

We can then quantify the forecast accuracy in the square error sense. We start by considering two opposite approaches to generating forecasts with panel data, namely zero pooling and full pooling. This corresponds to cases with very high and very low levels of parameter heterogeneity, respectively.

Pesaran et al. (2022) discuss the case where the coefficients are random and independent of the regressors[4]

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim (\mathbf{0}, \boldsymbol{\Omega}).$$

An alternative assumption is non-random, fixed coefficients $\boldsymbol{\beta}_i$ and we will comment on this case below, too.

2.2 Individual Estimation

The first set of forecasts we consider are based on individual estimates and ignore information from other units:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i.$$

Using these estimates, we can generate one-step-ahead forecasts of the individual units, $y_{i,T+1}$, as

$$\hat{y}_{i,T+1} = \hat{\boldsymbol{\beta}}_i'\mathbf{x}_{i,T+1}. \tag{3}$$

The associated forecast error is

$$\hat{e}_{i,T+1} = y_{i,T+1} - \hat{y}_{i,T+1} = \varepsilon_{i,T+1} - (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\mathbf{x}_{i,T+1}.$$

The mean square forecast error (MSFE) associated with the individual forecasts is

$$\begin{aligned} \mathrm{E}(\hat{e}_{i,T+1}^2) &= \sigma_i^2 + \mathbf{x}_{i,T+1}'\mathrm{E}\left[(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\right]\mathbf{x}_{i,T+1} \\ &= \sigma_i^2 + \frac{1}{T}\sigma_i^2\mathbf{x}_{i,T+1}'\left(\frac{\mathbf{X}_i'\mathbf{X}_i}{T}\right)^{-1}\mathbf{x}_{i,T+1} = \sigma_i^2 + O_p\left(\frac{1}{T}\right) \end{aligned} \tag{4}$$

see Pesaran et al. (2022). In many economic panels, T is often quite small such that the $O_p\left(\frac{1}{T}\right)$ component of the MSFE can be substantial. Note also that the MSFE depends on $\mathbf{x}_{i,T+1}$. The extent to which the uncertainty around a given parameter impacts the MSFE therefore depends on the value of the predictor vector in the forecast period, $\mathbf{x}_{i,T+1}$.

[4] Pesaran et al. (2024) consider the case where coefficient vectors and regressors are correlated. However, for simplicity of exposition, we maintain the assumption of uncorrelated random coefficients here.

2.3 Pooled Estimation

Alternatively, forecasts can be based on pooled estimation of the parameters

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The forecasts associated with this pooled estimator are given by

$$\tilde{y}_{i,T+1} = \tilde{\boldsymbol{\beta}}'\mathbf{x}_{i,T+1}, \tag{5}$$

while the associated MSFE becomes

$$\begin{aligned}
\mathrm{E}(\tilde{e}^2_{i,T+1}) &= \sigma_i^2 + \mathbf{x}'_{i,T+1}\mathrm{E}\left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\right]\mathbf{x}_{i,T+1} \\
&= \sigma_i^2 + \mathrm{E}\left[\mathbf{x}'_{i,T+1}\left(\frac{\mathbf{X}'\mathbf{X}}{NT}\right)^{-1}\frac{1}{N}\sum_{j=1,i\neq i}^{N}\sigma_j^2\left(\frac{\mathbf{X}_j'\mathbf{X}_j}{T}\right)^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\right]^2 \\
&\quad + \frac{1}{NT}\mathbf{x}'_{i,T+1}\left(\frac{\mathbf{X}'\mathbf{X}}{NT}\right)^{-1}\frac{1}{N}\sum_{j=1}^{N}\sigma_j^2\frac{\mathbf{X}_j'\mathbf{X}_j}{T}\left(\frac{\mathbf{X}'\mathbf{X}}{NT}\right)^{-1}\mathbf{x}_{i,T+1} \\
&= \sigma_i^2 + \mathrm{E}\left[\mathbf{x}'_{i,T+1}\left(\frac{\mathbf{X}'\mathbf{X}}{NT}\right)^{-1}\frac{1}{N}\sum_{j=1,i\neq i}^{N}\sigma_j^2\left(\frac{\mathbf{X}_j'\mathbf{X}_j}{T}\right)^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\right]^2 \\
&\quad + O_p\left(\frac{1}{N}\right).
\end{aligned} \tag{6}$$

Under the assumption of random coefficients, the MSFE under pooled estimation simplifies to

$$\mathrm{E}(\tilde{e}^2_{i,T+1}) = \sigma_i^2 + \mathbf{x}'_{i,T+1}\boldsymbol{\Omega}\mathbf{x}_{i,T+1} + O_p\left(\frac{1}{N}\right), \tag{7}$$

see Pesaran et al. (2022). Under the assumption of fixed coefficients, the MSFE remains as listed in the last line of (6).

Since N is typically large in panel data, estimation uncertainty either vanishes or is greatly reduced under pooled estimation. Conversely, pooled estimation gives rise to the term capturing the squared bias, which is caused by parameter heterogeneity.

The bias-variance trade-off in forecasting performance is clear from a comparison of equations (4) and (6). Individual forecasts will be more precise if the estimation uncertainty, weighted by the regressors in the forecast period, is smaller than the squared bias introduced by parameter heterogeneity. Otherwise pooled forecasts will be more precise. Note, that the effects of parameter heterogeneity and estimation uncertainty alone do not determine forecast uncertainty. What matters is instead the interaction of these two elements with the values of the regressors in the forecast period.

2.4 Panel Data with Short T

A sizeable part of the literature on panel data is concerned with models that allow for T to be very small. In this setting, most attention is on random and fixed effects models and forecasting has been extensively reviewed by Baltagi (2013). We merely highlight the main issues here and refer to Baltagi's survey for a detailed discussion.

With short T, estimating heterogeneous slope coefficients may not be possible or desirable. The literature therefore focused on the random effects specification, which takes the form

$$y_{it} = \alpha + \boldsymbol{\theta}'\mathbf{x}_{it} + u_{it}, \qquad u_{it} = \eta_i + \varepsilon_{it},$$

where $\eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$ and $\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. Forecasts are based on the Best Linear Unbiased Predictor (BLUP) of Goldberger (1962)

$$\hat{y}_{i,T+1} = \hat{\alpha}_{\mathrm{GLS}} + \hat{\boldsymbol{\theta}}'_{\mathrm{GLS}}\mathbf{x}_{i,T+1} + \frac{\hat{\sigma}_\eta^2}{T\hat{\sigma}_\eta^2 + \hat{\sigma}_\varepsilon^2}\,(\mathbf{l}_i' \otimes \boldsymbol{\iota}_T)\hat{\mathbf{u}}_{\mathrm{GLS}} \tag{8}$$

and $\mathbf{l}_i$ is the ith column of $\mathbf{I}_N$, $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones, $\hat{\alpha}_{\mathrm{GLS}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{GLS}}$ are estimated by GLS with covariance matrix

$$\boldsymbol{\Sigma} = T\sigma_\eta^2\mathbf{P} + \sigma_\varepsilon^2\mathbf{I}$$

$$\mathbf{P} = \mathbf{M}_\mu(\mathbf{M}_\mu'\mathbf{M}_\mu)^{-1}\mathbf{M}_\mu', \ \mathbf{M}_\mu = \mathbf{I}_N \otimes \boldsymbol{\iota}_T, \ \hat{\mathbf{u}}_{\mathrm{GLS}} = \mathbf{Y} - \hat{\alpha}_{\mathrm{GLS}} - \mathbf{x}\hat{\boldsymbol{\theta}}_{\mathrm{GLS}},$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_i)^2,$$

and $\hat{\sigma}_\eta^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{\eta}_i^2$ with $\hat{\eta}_i$ obtained from the fixed effects estimation.[5] The last term in (8) is an estimate of the individual specific effect.

Baltagi and Li (1992) derive the BLUP when the error term is autocorrelated and Baltagi and Liu (2020) extend the BLUP of the random effects model to unbalanced panels.

[5] See Pesaran (2015) for details of the random effects estimation.

Liu et al. (2020) consider a correlated random effects approach that allows the parameters that are heterogeneous in the cross-section to be correlated with the predictors. They develop an empirical Bayes estimator of the parameter that captures unobserved individual heterogeneity. They also consider predictors based on plug-in and pooled-OLS empirical Bayes estimators and find in Monte Carlo simulations and in an empirical application that their empirical Bayes predictor dominates these two alternatives.

Giacomini et al. (2023) also discuss shrinkage methods, including the James and Stein (1961) shrinkage estimator, which pull individual estimates towards a common mean. Their frequentist random effects approach aims to obtain accuracy of the individual forecasts as opposed to forecasts that are accurate "on average". Effectively, their individual weighting approach uses past time-series data to estimate the weights of a combination of time-series and pooled forecasts. They show that their individual weighting approach avoids poor forecasting performance even in areas of the parameter space where the performance of the underlying forecasts can differ significantly, i.e., it is minimax-regret optimal relative to an approach that uses *either* individual time-series forecasts or pooled forecasts.[6]

A fixed effects specification may be preferred in some settings. Any correlation between the intercept and the regressors, which are the concern of Liu et al. (2020), are then of no importance. The fixed effects model is

$$y_{it} = \alpha_i + \boldsymbol{\theta}'\mathbf{x}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim (0, \sigma^2).$$

Unbiased estimates of α_i and $\boldsymbol{\theta}$ are

$$\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} - \hat{\boldsymbol{\theta}}'_{FE}\mathbf{x}_{it},$$

and

$$\hat{\boldsymbol{\theta}}_{FE} = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}\mathbf{y}_i, \quad \mathbf{M} = \mathbf{I}_T - \boldsymbol{\iota}_T\boldsymbol{\iota}_T'/T.$$

However, the estimator for α_i does not benefit from the cross-section dimension and relies on large T, so for panels with few time-series observations the fixed effects estimator and, consequently, the forecasts will suffer from low precision.

[6]Focusing on estimating the intercept of a forecasting model, Giacomini et al. (2023) consider three different weighting schemes, including minimax regret optimal weights, inverse MSFE weights, and estimated oracle weights.

Forecasts of time-series with a very short T-dimension is related to the so-called "cold-start" forecasting problem which arises when no data points are available. Examples include forecasting the initial stock market price of a company undertaking an IPO or predicting the number of users of a new app. Without any historical data on the variable, one will have to draw heavily on estimates from existing data and make assumptions on how strong the similarities are between existing series with longer time records and the new one. One approach is to identify clusters of existing variables for which parameter estimates are available and then construct forecasts of the new variable as a weighted average of forecasts from those clusters. Heterogeneity among the units within each cluster can then be used to compute confidence bands on the forecasts of the new unit. Bayesian hierarchical approaches (discussed below) is another possibility.

3 Forecast Combinations

The trade-offs between individual and pooled forecasts suggests that combinations of individual and pooled forecasts in (3) and (5) may increase forecast accuracy. For the random coefficients case, Pesaran et al. (2022) consider the following combined forecast:

$$y_{i,T+1}^{(c)} = w_i \hat{y}_{i,T+1} + (1 - w_i)\tilde{y}_{i,T+1},$$

where w_i is the combination weight. The corresponding forecast error is

$$e_{i,T+1}^{(c)} = w_i \hat{e}_{i,T+1} + (1 - w_i)\tilde{e}_{i,T+1},$$

and the MSFE of the combination forecast becomes

$$\begin{aligned} \text{MSFE}(e_{i,T+1}^{(c)}) &= w_i^2 \text{Var}(\hat{e}_{i,T+1}) + (1 - w_i)^2 \text{Var}(\tilde{e}_{i,T+1}) \\ &\quad +2w_i(1 - w_i)\text{Cov}(\tilde{e}_{i,T+1}\hat{e}_{i,T+1}), \end{aligned} \tag{9}$$

since both forecasts are unbiased in the random coefficients model. The optimal weights are then

$$w_i^* = \frac{\text{Var}(\tilde{e}_{i,T+h}) - \text{Cov}(\tilde{e}_{i,T+h}, \hat{e}_{i,T+h})}{\text{Var}(\tilde{e}_{i,T+h}) + \text{Var}(\hat{e}_{i,T+h}) - 2\text{Cov}(\tilde{e}_{i,T+h}, \hat{e}_{i,T+h}).} \tag{10}$$

Using the expressions in (4) and (7), and the fact that $\text{Cov}(\tilde{e}_{i,T+h}, \hat{e}_{i,T+h}) = \sigma_i^2 + O_p(1/N)$, yields the optimal weights

$$w_i^* = \frac{\mathbf{x}_{i,T+1}' \boldsymbol{\Omega} \mathbf{x}_{i,T+1}}{\mathbf{x}_{i,T+1}' \left[T^{-1} \sigma_i^2 \left(\frac{\mathbf{X}_i' \mathbf{X}_i}{T} \right)^{-1} + \boldsymbol{\Omega} \right] \mathbf{x}_{i,T+1}} + O_p\left(\frac{1}{N}\right),$$

see Pesaran et al. (2022). Under a fixed coefficients specification, a similar expression can be derived using the expressions in (4) and (6).

Note that the weights are based on the square error loss function of the forecast rather than the mean square error of the parameter estimate. This results in the presence of the regressors in the forecast period in the expression of the weights, which determine the importance of estimation uncertainty and heterogeneity.

As T increases, one of the weights will converge to unity and the combined forecast will equal the forecast from this individual model. However, when T is small or of modest size, the weights can differ substantially from unity if the parameter heterogeneity is sufficiently large.[7]

Inserting equation (10) in (9) and noting that the two forecasts are asymptotically uncorrelated implies that the MSFE of the combination forecast using the optimal weights can be written as

$$\text{MSFE}(w_i^*) = w_i^* \text{MSFE}(\hat{y}_{i,T+1}) = (1 - w_i^*) \text{MSFE}(\tilde{y}_{i,T+1})$$

Hence, for finite T and abstracting from estimation error in determining w_i^*, we have that $0 < w_i^* < 1$ and the combination MSFE will always be smaller than both the individual and the pooled forecast's MSFEs.

In practice, $\mathbf{\Omega}$ and σ_i^2 need to be replaced by estimated counterparts, which yields

$$\hat{w}_i^* = \frac{\mathbf{x}'_{i,T+1} \hat{\mathbf{\Omega}} \mathbf{x}_{i,T+1}}{\mathbf{x}'_{i,T+1} \left[T^{-1} \hat{\sigma}_i^2 \left(\frac{\mathbf{X}_i' \mathbf{X}_i}{T} \right)^{-1} + \hat{\mathbf{\Omega}} \right] \mathbf{x}_{i,T+1}}, \tag{11}$$

where one could use the following plug-in estimates

$$\begin{aligned} \hat{\mathbf{\Omega}} &= \frac{1}{N-1} \sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}} \right) \left(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}} \right)', \quad \bar{\hat{\boldsymbol{\beta}}} = N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_i, \\ \hat{\sigma}_i^2 &= (T-K)^{-1} (\mathbf{Y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_i)' (\mathbf{Y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_i). \end{aligned}$$

Yet, $\text{E}(\hat{\mathbf{\Omega}}) = \mathbf{\Omega} + \frac{1}{NT} \sum_{i=1}^{N} \sigma_i^2 \mathbf{Q}_{iT}^{-1}$ and an unbiased estimator of $\mathbf{\Omega}$ is given by $\tilde{\mathbf{\Omega}} = \hat{\mathbf{\Omega}} - \frac{1}{NT} \sum_{i=1}^{N} \hat{\sigma}_i^2 \mathbf{Q}_{iT}^{-1}$. Using this, we have

$$\tilde{w}_i^* = \frac{\mathbf{x}'_{i,T+1} \left[\hat{\mathbf{\Omega}} - \frac{1}{NT} \sum_{i=1}^{N} \hat{\sigma}_i^2 \mathbf{Q}_{iT}^{-1} \right] \mathbf{x}_{i,T+1}}{\mathbf{x}'_{i,T+1} \left[\frac{1}{T} \hat{\sigma}_i^2 \mathbf{Q}_{iT}^{-1} + \hat{\mathbf{\Omega}} - \frac{1}{NT} \sum_{i=1}^{N} \hat{\sigma}_i^2 \mathbf{Q}_{iT}^{-1} \right] \mathbf{x}_{i,T+1}}, \tag{12}$$

[7] For the fixed coefficients case, (4) and (6) could be used in (10) to obtain combination weights.

which Pesaran et al. (2022) refer to as bias-corrected weights. While $\tilde{\mathbf{\Omega}}$ is an unbiased estimate, it could lead to negative weights in small samples. When using $\tilde{\mathbf{\Omega}}$, Pesaran et al. (2022) therefore restrict the weights to lie between 0 and 1.

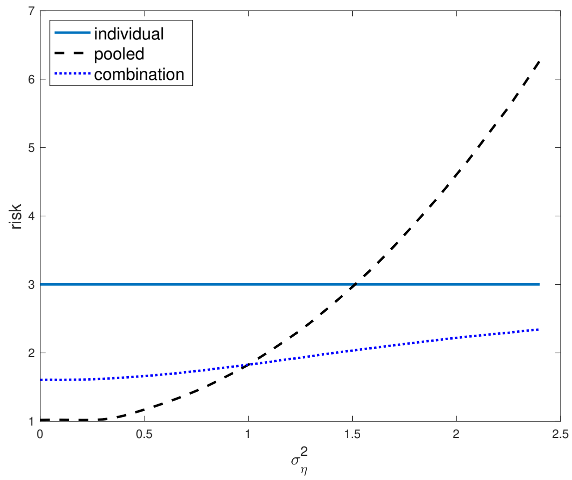
Figure 1, taken from Pesaran et al. (2022), plots the MSFE of the individual, pooled and combination forecasts with estimated weights. The MSFE is calculated by simulation using the model $y_{it} = \beta_i x_{it} + \varepsilon_{it}$ with $\text{Var}(\beta_i) = \sigma_\eta^2$, which is the variable shown on the horizontal axis. On the vertical axis is the resulting risk in terms of MSFE.

The risk of the individual forecast does not depend on the degree of parameter heterogeneity and therefore shows as a horizontal line. When the parameters are homogeneous ($\sigma_\eta^2 = 0$), the risk of the pooled forecast is much smaller than that of the individual forecast due to the increased estimation efficiency. As the parameter heterogeneity increases, the risk of the pooled forecast increases and eventually exceeds that of the individual forecast. The third line in the plot is the risk of the combination forecast. At low levels of parameter heterogeneity, the risk from the combination forecast exceeds that of the pooled forecast but falls below that of the individual forecast as estimation uncertainty yields a small positive weight on the individual forecasts. As the level of parameter heterogeneity increases, the relative performance of the combination forecast improves and eventually it becomes more precise than either of the two forecasts that it combines.

Pesaran et al. (2024) extend the analysis to the case where the regressors are weakly exogenous and correlated with the parameters. They derive weights designed to be optimal for the average forecast in the panel. Compared to the simple case discussed here, these optimal weights are somewhat more complicated and involve expressions that reflect the less restrictive assumptions. Additionally, they introduce optimal weights for the combination of forecasts based on individual and fixed effects estimation. In addition to the average MSFE, Pesaran et al. (2024) examine the entire distribution of MSFE-values across the individual units. It emerges in two applications that, among the range of forecasting methods they examine, the combination forecasts are the only methods not to provide the worst forecast for any unit in any period, thus making them attractive from a minmax regret perspective.

Issler and Lima (2009) propose ways to combine bias-corrected panel forecasts. Their approach assumes that forecast errors can be decomposed into a constant forecaster-specific bias term and aggregate and idiosyncratic shocks, both of which have zero means unconditionally. The forecaster-specific bias term is assumed to be identically distributed but need not be

Figure 1: Risk versus parameter heterogeneity



Note: The plot displays the risk in terms of the expected MSFE of forecasts based on pooled and individual estimates and of combination forecasts. The horizontal axis measures the degree of parameter heterogeneity in the simple panel regression model. The expected MSFE is calculated via 10,000 simulations with the DGP $y_{it} = \beta_i x_{it} + \sigma_i \varepsilon_{it}$, where $x_{it} = \mu_{xi} + \sigma_{xi} v_{it}$, $\sigma_i^2 \sim \text{iid}\left(1+\chi_1^2\right)/2$, $\sigma_{xi}^2 \sim \text{iid}\left(1+\chi_1^2\right)/2$, $\beta_i = 1 + \sigma_\eta \eta_i$, $\varepsilon_{it}, v_{it}, \eta_i, \mu_{xi} \sim \text{iidN}(0,1)$. Since there is only one regressor, forecast performance is a scaled version of parameter heterogeneity. Source: Pesaran et al. (2022).

independent across forecasters. Using this setup, they show that a simple pooling of forecasts is MSFE optimal in the limit $N, T \to \infty$. They propose a feasible bias-corrected average forecast that subtracts an estimated bias term from the equal-weighted average of the individual forecasts and show that this is optimal and equivalent to the conditional expectation. This procedure is appealingly simple, requiring only the estimation of a single parameter as opposed to a large set of combination weights.[8] Monte Carlo simulations suggest that this simple bias-adjustment performs well compared to an equal-weighted forecast that does not bias-adjust the underlying fore-

[8]Specifically, their estimator first computes the time-series averages of individual biases from the mean of individual forecasters' prediction errors. It then computes the equal-weighted cross-sectional average of these forecaster-specific time-series averages. Finally, the feasible bias-corrected average forecast is computed as the cross-sectional average forecast adjusted for the average bias.

casts or a forecast combination scheme that estimates the individual weights through a simple time-series projection of outcomes on forecasts. Even in situations where the simple bias-adjusted average forecast is not optimal, it typically only performs marginally worse than the best of these alternative forecasts.

Wang et al. (2019) propose using in-sample statistics, such as an F-test statistic for parameter homogeneity or Mallows criterion to construct weights for the combination of different panel data models. Their Monte Carlo simulations suggest that in panel data models with high R^2, forecast combinations using the Mallows criterion can work well. In medium-to-low R^2 settings, an FGLS estimator of Swamy (1970) performs best.

Huang et al. (2019) use combinations of random and fixed effects forecasts to address the issue discussed in Section 2.4 that the fixed effects forecast will be inefficient and the random effects forecast will be biased in the presence of correlation between α_i and the regressors. They use Stein-like combination of the two estimators in their forecast combination. Their Monte Carlo results suggest that the combined forecast does well in delivering a precise forecast in the mean square error sense.

4 Alternative Forecasting Approaches

Multivariate forecasting approaches that incorporate dynamics into panel data models have been developed in both the frequentist and the Bayesian context. This section discusses alternative forecasting methods that can be used for at least some types of panel data such as Bayesian hierarchical forecasts, BVARs with large numbers of variables, panel and global VARs, and machine learning approaches.

4.1 Bayesian Panel Forecasts

A range of Bayesian methods is available to use for forecasting with panel data. Lindley and Smith (1972) discuss the hierarchical approach to the linear regression model and Gelfand et al. (1990) provide a Gibbs sampler for this model.[9] Consider the linear regression model in (1) with the added assumption of normality,

$$y_{it} = \boldsymbol{\beta}_i' \mathbf{x}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathrm{N}(0, \sigma^2).$$

[9] A review can be found in the book by Greenberg (2008).

Suppose we introduce a prior that assumes $\boldsymbol{\beta}_i$ is normally distributed with mean $\bar{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Completion of the set up requires priors for $\bar{\boldsymbol{\beta}}$, which is again normal with mean $\tilde{\boldsymbol{\beta}}$ and precision matrix $\tilde{\mathbf{S}}_\beta$ and a prior for $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, which is an inverse Wishart with scale matrix $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\nu}_\Sigma$ degrees of freedom. Finally, the error variance σ_i^2 is assumed to have an inverse Gamma prior with scale parameter $\tilde{\sigma}^2$ and $\tilde{\nu}_\sigma$ degrees of freedom.[10]

Lee and Griffiths (1979) propose several different approaches to estimate the parameters of such a model, including an empirical Bayes approach and a Bayesian approach, without the use of the Gibbs sampler. Their estimators were used in an application to electricity consumption by Maddala et al. (1997). Pesaran et al. (2022) evaluated the forecast performance of these methods and found this to be quite similar to that of the combined forecast of Section 3.

The model parameters can also be estimated using the Gibbs sampler, which draws iteratively from a set of conditional posteriors (Gelfand et al., 1990):

- $\boldsymbol{\beta}_i|\cdot \sim \mathrm{N}(\mathbf{b}_\mathrm{i}, \mathbf{S}_\mathrm{i})$, where $\mathbf{b}_i = \mathbf{S}_i \left(\sigma^{-2}\mathbf{X}_i'\mathbf{y}_i + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\bar{\boldsymbol{\beta}}\right)$ and $\mathbf{S}_i = \left(\sigma^{-2}\mathbf{X}_i'\mathbf{X}_i + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right)^{-1}$
- $\sigma^2|\cdot \sim \mathrm{iG}\left([NT + \tilde{\nu}_\sigma]/2, \frac{1}{2}\left[\frac{1}{N}\sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i) + \tilde{\nu}_\sigma\tilde{\sigma}^2\right]\right)$
- $\bar{\boldsymbol{\beta}}|\cdot \sim \mathrm{N}(\mathbf{h}, \mathbf{S}_\mathrm{h})$, where $\mathbf{h} = \mathbf{S}_h \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\sum_{i=1}^N \boldsymbol{\beta}_i + \tilde{\mathbf{S}}_\beta\tilde{\boldsymbol{\beta}}\right)$ and $\mathbf{S}_h = \left(N\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \tilde{\mathbf{S}}_\beta\right)^{-1}$
- $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\cdot \sim \mathrm{W}\left(N + \tilde{\nu}_\Sigma, \left[\sum_{i=1}^N \left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}\right)' + \tilde{\nu}_\Sigma\tilde{\boldsymbol{\Sigma}}\right]^{-1}\right)$.

The Gibbs sampling approach is the most computationally expensive of these methods. Consider the example of a simple panel AR model with an intercept and one lagged dependent variable. For the case of $N = 500$, $T = 20$, we calculated the ratio of computation time different methods take to compute the forecasts for each of the 500 units relative to the individual forecast using Matlab. The pooled forecast uses about 20% of the computing time of the individual forecasts, the combination forecast takes about 3.5 times longer, the Bayesian approach of Lee and Griffiths (1979) about 6.5

[10] Normality of the parameters and error terms can be relaxed using scale mixtures of normals as suggested by Geweke (1993). Then $\boldsymbol{\beta}_i \sim \mathrm{N}(\bar{\boldsymbol{\beta}}, \eta_{\beta,i}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ with $\eta_{\beta,i} \sim \mathrm{G}(\nu_\beta/2, \nu_\beta/2)$ and $\varepsilon_{it} \sim \mathrm{N}(0, \eta_{\varepsilon,i}^{-1}\sigma^2)$ with $\eta_{\varepsilon,i} \sim \mathrm{G}(\nu_\varepsilon/2, \nu_\varepsilon/2)$. Further extensions can be found in Greenberg (2008). Group structure in the individual units can be accommodated in the model via a Dirichlet process prior (Escobar, 1994; Escobar and West, 1995).

times longer, but a Gibbs sampler with 1500 iterations takes over 1000 times longer.

Zellner and Hong (1989) use Bayesian shrinkage in the context of forecasting international growth rates for 18 countries. They find that using the Bayesian approach improves forecast accuracy over alternatives used by Garcia-Ferrer et al. (1987).

Hsiao et al. (1999) consider empirical Bayes estimation of the hierarchical model for a dynamic panel data model with short T. Under the empirical Bayes approach, the variances are estimated from the data without a prior distribution. In general, however, their results suggest that the full Bayesian estimation of the hierarchical model is preferable.

4.2 VARs with Many Variables

VAR models provide a flexible way to handle (linear) dynamics of quite general form. Typically, VARs include variables with little in common such as GDP growth, interest rates, and unemployment rate. It would therefore not be meaningful to impose that the parameters are identical across individual variables. For example, the variables are unlikely to have the same mean or degree of persistence. In contrast, panel data sets typically consists of variables that are broadly comparable and measured in the same units such as inflation in different regions or stock returns across different firms. It can then make sense to assume common parameters and impose stronger homogeneity on the estimates. Hence, panel data often provide natural shrinkage targets for the parameters which are different from the targets assumed by common priors for VARs such as the Minnesota prior.

VARs are commonly used when N is small and T is large. However, Bańbura et al. (2010) show that Bayesian VAR methods can now be extended to handle hundreds of variables with appropriately chosen priors. This development is promising and offers different ways to handle variable selection and implement shrinkage.[11] Chapters 2 and 3 of this Handbook deal with Bayesian VARs in macroeconomic forecasting. Generally, however, the specification of the model and priors does not necessarily account for the panel nature of the data. This contrasts with the panel and global VAR methods discussed in the next section.

Koop and Korobilis (2019) forecast inflation in the eurozone using a large dimensional panel VAR. They develop shrinkage priors that account for the

[11] Bańbura et al. (2010) show that Laplace priors over the parameters achieve Lasso variable selection while Bayesian priors correspond to standard shrinkage of the parameters.

panel nature of the data. See also Korobilis (2016) for a discussion of priors for panel VAR models.

4.3 Panel and Global VARs

International macroeconomic data sets often contain observations on a moderate number of countries, each with a long history of data. This is thus an environment with T large and N small to moderate, which contrasts with the settings considered so far. Panel and global VAR models offer a natural setting for these types of data.[12]

Panel VAR models allow the $K \times 1$ outcome vector, $\mathbf{y}_{it}$, to depend on lags of $\mathbf{y}_{jt}$ for all $j = 1, 2, \ldots, N$ in addition to a set of exogenous regressors, $\mathbf{x}_{it}$:

$$\mathbf{y}_{it} = \boldsymbol{\alpha}_{it} + \mathbf{A}_i(L)\mathbf{y}_t + \mathbf{B}_i\mathbf{x}_{it} + \mathbf{u}_{it}, \tag{13}$$

where $\mathbf{y}_t = (\mathbf{y}'_{1t}, \mathbf{y}'_{2t}, \ldots, \mathbf{y}'_{Nt})'$, and $\mathbf{A}_i(L)$ is a lag polynomial, $\boldsymbol{\alpha}_{it}$ is a vector of constants, $\boldsymbol{x}_{it}$ a vector of exogenous variables with associated parameter matrix $\mathbf{B}_i$, and $\mathbf{u}_t = (\mathbf{u}'_{1t}, \mathbf{u}'_{2t}, \ldots, \mathbf{u}'_{Nt})' \sim iid(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

The general version of the panel VAR in (13) suffers from a proliferation of parameters and different approaches have been taken to either reduce the number of parameters or estimate them via shrinkage. For example, in the context of a model for microeconomic panel data, Holtz-Eakin et al. (1998) restrict heterogeneity to the constant intercept but pool the time-varying slope coefficients assuming that only own lags have non-zero coefficients. This contrasts with studies that focus on macroeconomic data, where dynamic interdependencies are typically included.

Canova and Ciccarelli (2004) estimate the coefficients of (13) using Bayesian methods with two kinds of priors. First, they use hierarchical priors similar to the ones discussed in Section 4.1 but decompose the vector of autoregressive parameters into constant, unit-specific parameters and common and time-varying coefficients. Second, they use Minnesota-type shrinkage priors (Litterman, 1980). They apply the Bayesian panel VAR models to forecast growth rates in G7 economies and find that the hierarchical models often provide particularly accurate forecasts.

Camehl (2023) suggests using LASSO penalization in the estimation of the parameters in (13). She applies this approach to a set of two and four macroeconomics variables over a set of five and 20 countries, respectively.

[12] The panel VAR literature has been reviewed by Canova and Cicarelli (2013) and the global VAR literature by Chudik and Pesaran (2014).

Her analysis suggests that the LASSO approach performs well without the need for numerical integration.

Global VAR models (Pesaran et al., 2004; Dées et al., 2007), or GVARs, provide an alternative to panel VARs. GVARs establish separate VAR models for each country and then connect the different VAR models via global variables. Pesaran et al. (2009) use GVAR models to forecast five macro and financial variables. Their results suggest that model averaging offers a simple but effective tool for handling model uncertainty in a forecasting context. Additional forecast applications of GVARs are reviewed by Chudik and Pesaran (2014). A Bayesian approach to generate forecasts with the GVAR model has been suggested by Crespo Cuaresma et al. (2016).

4.4 Machine Learning Methods

What are often referred to as machine learning (ML) methods have become widely used in the economic forecasting profession. For example, Chapter 10 of this Handbook discusses the use of the LASSO in fixed effects models to select among a large number of possible predictors when the time series dimension is large.

Many ML methods can be thought of as flexible semi-parametric estimation techniques. When applied to suitable data structures, these methods offer several advantages such as (i) a flexible mapping from input variables to forecasts that can accommodate both nonlinear effects and interaction terms; (ii) the capacity to handle high-dimensional data sets; (iii) the ability to simultaneously provide parameter estimation and variable selection.

While these can be important advantages, there are also some real limitations to the application of machine learning methods in economic forecasting due to the nature of the data sets encountered: (i) Theoretical results on the properties of forecasts generated from ML methods typically assume data that are independently and identically distributed. In practice, economic data typically deviate strongly from this assumption with time-varying and persistent heteroskedasticity (ARCH effects) and serially correlated (persistent) outcomes. (ii) Often, the training data used to estimate economic forecasting model is not very long (small T-dimension) or wide (small N dimension). In these cases, approximations to underlying non-linear functional forms for the forecasting model are likely to be inaccurately estimated and dominated by estimation error. (iii) Low-frequency components such as recessions are often important in economic data. By definition these components do not show up very often in the data, making it difficult even for supervised learning algorithms to uncover stable and reliable predictive patterns, let alone approximate them with a high degree of accuracy. Non-stationarities induced by structural breaks to the data generating process pose another challenge for economic forecasting models and further limit the effective T dimension of the estimation sample.

5 Forecast Evaluation

While statistical methods for evaluating univariate forecasts are well developed, how to evaluate panel forecasts poses additional challenges that have not been addressed to the same extent. For example, one forecaster may be interested in the average MSFE, i.e., the MSFE averaged cross-sectionally. Another forecaster could be interested in avoiding very poor forecasts for any individual unit, which suggests more of a minmax loss function defined on the individual forecast errors. A third forecaster may be interested in the joint distribution of forecast errors, suggesting more of a portfolio approach.

In this section we discuss various formulations of the loss function before turning to the question of how to statistically evaluate panels of forecasts.

5.1 Loss Function

What is considered a "good" forecast needs to be specified in the context of the forecaster's objective function, which is often referred to as the loss function, $L(\cdot)$. It is common to assume that this only depends on the distance between the realized, $y_{i,T+1}$, and predicted values, $\hat{y}_{i,T+1}$, of the outcome, i.e., the forecast error, $e_{i,T+1}$. Further, it is common to assume squared error loss for the individual units,

$$L_{i,T+1} \equiv L(y_{i,T+1}, \hat{y}_{i,T+1}) = \hat{e}^2_{i,T+1}. \tag{14}$$

This loss function is regularly applied to the individual units and it is common to estimate the associated loss by computing a simple time-series average over a range of pseudo-out-of-sample forecasts from, say, $T_0 + 1$ to T, which yields the MSFE:

$$\overline{L}_i \equiv \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \hat{e}^2_{it}. \tag{15}$$

It is less clear how to evaluate an $N \times (T - T_0)$ dimensional matrix of forecast errors. With both a cross-sectional and a time-series dimension, several loss functions become possible.

A natural starting point to evaluate the N-dimensional vector of MSFE losses $(\overline{L}_1, \overline{L}_2, \ldots, \overline{L}_N)$ is the grand average loss, i.e., the squared-error loss averaged across both, the $T - T_0$ dimensional time-series of individual forecasts and N cross-sectional units:

$$\overline{L} \equiv \frac{1}{N} \sum_{i=1}^{N} \overline{L}_i. \tag{16}$$

Instead of the grand average in (16), one could consider minimizing for a given time-period, $T + 1$, the average cross-sectional loss:

$$\overline{L}^c_{T+1} \equiv \frac{1}{N} \sum_{i=1}^{N} \hat{e}^2_{i,T+1} \tag{17}$$

This loss function makes most sense if each unit has the same weight in the forecaster's objectives; alternatively, a weighted average can be used. Forecast evaluation on a period-by-period basis for a cross-section of outcomes is still a relatively unexplored area, see Qu et al. (2023).

When the units of the variables being predicted are not directly comparable, it is less appealing to compute a simple average of MSFE-values across all variables. Alternatives based on percentage loss such as the mean absolute percentage error (MAPE) or mean squared percentage error (MSPE) can be used in this case although the scaling of forecast errors by either the outcome or by the forecast introduces new issues, particularly when these values are close to zero or switch signs.

One can also consider loss functions defined over the distribution or quantiles of the cross-section of MSFE-values, $F_q(\overline{L}_i)$, where $F(.)$ is the cumulative distribution function defined on the cross-section of expected squared error loss and $q \in [0, 1]$ is the particular quantile. Forecasting methods with a low probability of generating large losses for any individual unit might be desirable. A minmax loss function would attempt to minimize $\max_i \overline{L}_i$ for $i = 1, 2, \ldots, N$.

Any ranking of forecasting approaches is likely to depend on which of these loss functions is chosen. Forecasting methods differ in their sensitivity to sample information; approaches such as forecast combination often do not provide the single most accurate forecasts with the lowest MSFE values but possess desirable robustness properties by offering relative safety as they rarely generate the least accurate forecasts with the highest MSFE values for individual units.

5.2 Tests of Equal Predictive Accuracy

Time-series of cross-sectional averages of losses such as those in (16) can be used to set up panel versions of the Diebold and Mariano (1995) test which is commonly used to compare the accuracy of alternative forecasts of individual time series. Papers that pursue this route include Pesaran et al. (2013), Akgun et al. (2024), and Qu et al. (2024).[13]

Suppose we are interested in testing the null that two forecasts, A and B, are expected to be equally accurate across all time periods and cross-sectional units:[14]

$$H_0 : \mathrm{E}[L_{it,A} - L_{it,B}] = \mathrm{E}[\Delta L_{it}] = 0.$$

This null is naturally tested on $N(T - T_0)$ pseudo-out-of-sample forecasts using a test statistic that extends the Diebold and Mariano (1995) test to the panel domain:

$$\mathrm{PDM} = \frac{\frac{1}{\sqrt{N(T-T_0)}} \sum_{t=T_0+1}^{T} \sum_{i=1}^{N} \Delta L_{it}}{\hat{\sigma}(\Delta L_{it})}, \tag{18}$$

where $\hat{\sigma}(\Delta L_{it})$ is a consistent estimator for $\sqrt{\mathrm{Var}\left(\frac{1}{\sqrt{N(T-T_0)}} \sum_{t=T_0+1}^{T} \sum_{i=1}^{N} \Delta L_{it}\right)}$.

One can be compute $\hat{\sigma}(\Delta L_t)$ in different ways. Let $R_t = N^{-1/2} \sum_{i=1}^{N} \Delta L_{it}$ be the scaled cross-sectional average loss differential at time t. Under standard assumptions of weak serial dependence in the sequence of forecast losses, the standard error in the denominator of (18) can be computed using a Newey-West type estimator (see, e.g., Qu et al. (2024)):

$$\hat{\sigma}(\Delta L_t) = \sqrt{\sum_{j=-J}^{J} (1 - j/J)\hat{\gamma}(j)},$$

where $J > 0$ is the maximum lag length and $\hat{\gamma}(j) = (T - T_0 - j)^{-1} \sum_{t=T_0+j+1}^{T} \tilde{R}_{t-j}\tilde{R}_t$ with $\tilde{R}_t = R_t - \bar{R}$ and $\bar{R} = (T - T_0)^{-1} \sum_{st=T_0+1}^{T} R_t$.[15] Under standard regularity conditions, PDM $\xrightarrow{d} N(0, 1)$.

In practice, we may not only be interested in testing whether two sets of forecasts are equally accurate for the grand mean average loss in (16) but

[13] For a review of testing of equal predictive accuracy see Chapter 11 of this Handbook.

[14] For simplicity, we suppress any dependence on forecast horizon, but the null can be tested separately for a given forecast horizon or jointly across multiple horizons.

[15] For $j < 0$, $\hat{\gamma}(j) = \hat{\gamma}(-j)$.

also whether this holds for clusters (subgroups) of the data. In this context, Qu et al. (2024) consider testing predictive ability for time-series clusters or cross-sectional clusters of variables. For the latter case, suppose that the individual units have been categorized into K cross-sectional clusters, denoted by $H_1, H_2, \ldots, H_K$, with $|H_j|$ denoting the number of elements in the jth cluster, with $\sum_{j=1}^{K} |H_j| = N$. Define

$$D_j = |H_j|^{-1/2}(T - T_0)^{-1/2} \sum_{i \in H_j} \sum_{t=T_0+1}^{T} \Delta L_{it}.$$

The null hypothesis of equal predictive accuracy within each cross-sectional cluster can then be written as

$$H_0 : \ \mathrm{E}D_1 = \mathrm{E}D_2 = \cdots = \mathrm{E}D_K = 0. \tag{19}$$

Defining the average of the loss differences across the K cross-sectional clusters as $\bar{D} = K^{-1} \sum_{j=1}^{K} D_j$, Qu et al. (2024) apply the test statistic[16]

$$J_n^D = \frac{\sqrt{K}\bar{D}}{\sqrt{(K-1)^{-1} \sum_{j=1}^{K} (D_j - \bar{D})^2}}.$$

6 Conclusion

This chapter reviewed different approaches for forecasting with panel data. Key to the success or failure of a given forecasting approach is how it handles parameter heterogeneity and estimation error. Strategies such as pooling, individual unit-specific estimation, fixed or random effect estimation, forecast combination, and Bayesian modeling exploit the associated bias-variance trade-off in different ways. Panel data sets often provide natural shrinkage targets which can be exploited to reduce the impact of estimation error - without introducing too sizeable a bias - assuming that some level of commonality exists in the model parameters pertaining to individual variables.

An advantage of panel forecasting methods is that analytical results are available for understanding the effect of (correlated) parameter heterogeneity and dynamics on the performance of different forecasting approaches. Since these are defining features of many economic data sets, it is important to have theoretical results available for a class of models that is sufficiently

[16] Qu et al. (2024) discuss how to compute critical values for this test statistic.

simple and transparent to lend itself to analysis. Insights from this class of models can then be used as building blocks for more sophisticated and flexible forecasting methods, e.g., from machine learning.

We would not, in general, expect a single panel estimation approach to be uniformly dominant but, rather, that different approaches should be chosen for data sets with different T and N dimensions and different degrees of parameter heterogeneity. A broad reading of the literature does suggest, however, that forecast combination and Bayesian methods can be used to significantly reduce the likelihood of producing poor forecasts for individual variables. In contrast, while they may perform well for specific data sets, individual forecasts and pooled forecasts can produce poor forecasts for data sets that do not satisfy their underlying assumptions.

References

Akgun, O., Pirotte, A., Urga, G., and Yang, Z. (2024). Equal predicitive ability tests based on panel data with applications to OECD and IMF forecasts. *International Journal of Forecasting*, 40:202–228.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*. Wiley, 3rd edition.

Baltagi, B. H. (2013). Panel data forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2B, chapter 18. Elsevier.

Baltagi, B. H., Bresson, G., and Pirotte, A. (2002). Comparison of forecast performance for homogeneous, heterogeneous and shrinkage estimators: Some empirical evidence from US electricity and natural-gas consumption. *Economics Letters*, 76:375–382.

Baltagi, B. H. and Griffin, J. M. (1997). Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics*, 77:303–327.

Baltagi, B. H. and Li, Q. (1992). Prediction in the one-way error component model with serial correlation. *Journal of Forecasting*, 11:561–567.

Baltagi, B. H. and Liu, L. (2020). Forecasting with unbalanced panel data. *Journal of Forecasting*, 39:709–724.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25:71–92.

Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Brucker, H. and Siliverstovs, B. (2006). On the estimation and forecasting of international migration: How relevant is heterogeneity across countries? *Empirical Economics*, 31:735–754.

Camehl, A. (2023). Penalized estimation of panel vector autoregressive models: A panel LASSO approach. *International Journal of Forecasting*, 39:1185–1204.

Canova, F. and Cicarelli, M. (2013). Panel vector autoregressive models: A survey. In Fomby, T. B., Kilian, L., and Murphy, A., editors, *Advances in Econometrics (VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims)*, volume 32. Emerald Group Publishing.

Canova, F. and Ciccarelli, M. (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics*, 120:327–359.

Chudik, A. and Pesaran, M. H. (2014). The theory and practice of GVAR modelling. *Journal of Economic Surveys*, 30:165–197.

Crespo Cuaresma, J., Feldkircher, M., and Huber, F. (2016). Forecasting with global vector autoregressive models: a Bayesian approach. *Journal of Applied Econometrics*, 31:1197–1483.

Dées, S., Di Mauro, F., Pesaran, M. H., and Smith, L. V. (2007). Exploring the international linkages of the euro area: A globar VAR analysis. *Journal of Applied Econometrics*, 22:1–38.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 12:253–263.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89:268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Garcia-Ferrer, A., Highfield, R. A., Palm, F., and Zellner, A. (1987). Macroeconomic forecasting using pooled international data. *Journal of Businss and Economic Statistics*, 5:53–67.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of American Statistical Association*, 85:972–985.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8:S19–S40.

Giacomini, R., Lee, S., and Sarpietro, S. (2023). A robust method for microforecasting and estimation of random effects. Unpublished working paper, University College London.

Goldberger, A. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57:369–375.

Greenberg, E. (2008). *Introduction to Bayesian Econometrics*. Cambridge University Press.

Holtz-Eakin, D., Newey, W., and Rosen, H. S. (1998). Estimating vector autoregressions with panel data. *Econometrica*, 56:1371–1395.

Hsiao, C., Pesaran, M. H., and Tahmiscioglu, A. K. (1999). Bayes estimation of short-run coefficients in dynamic panel data models. In Hsiao, C., Lahiri, K., Lee, L.-F., and Pesaran, M. H., editors, *Analysis of Panels and Limited-Dependent Variable Models: In Honour of G.S. Maddala*, chapter 11. Cambridge University Press.

Huang, B., Lee, T.-H., and Ullah, A. (2019). A combined random effect and fixed effect forecast for panel data models. *Journal of Management Science and Engineering*, 4:28–44.

Issler, J. V. and Lima, L. R. (2009). A panel data approach to economic forecasting: The bias-corrected average forecast. *Journal of Econometrics*, 152:153–164.

Koop, G. and Korobilis, D. (2019). Forecasting with high-dimensional panel VARs. *Oxford Bulletin of Economics and Statistics*, 81:937–959.

Korobilis, D. (2016). Prior selection for panel vector autoregressions. *Computational Statistics & Data Analysis*, 101:110–120.

Lee, L.-F. and Griffiths, W. E. (1979). The prior likelihood and best linear unbiased prediction in stochastic coefficient linear models. *Center for Economic Research, University of Minnesota, Discussion paper 79–107*.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of Royal Statistical Society, Series B*, 34:1–41.

Litterman, R. (1980). A Bayesian procedure for forecasting with vector autoregression. Working paper, Massachusetts Institute of Technology, Department of Economics.

Liu, L., Moon, H. R., and Schorfheide, F. (2020). Forecasting with dynamic panel data models. *Econometrica*, 88:171–201.

Maddala, G., Trost, R., Li, H., and Joutz, F. (1997). Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage priors. *Journal of Business & Economic Statistics*, 15:90–100.

Pesaran, M. H. (2015). *Time Series and Panel Data Econometrics*. Oxford University Press.

Pesaran, M. H., Pick, A., and Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177:134–152.

Pesaran, M. H., Pick, A., and Timmermann, A. (2022). Forecasting with panel data: estimation uncertainty versus parameter heterogeneity. *CEPR Discussion Paper DP17123*.

Pesaran, M. H., Pick, A., and Timmermann, A. (2024). Forecasting with panel data: estimation uncertainty versus parameter heterogeneity. *(in preparation)*.

Pesaran, M. H., Schuermann, T., and Smith, L. V. (2009). Forecasting economic and financial variables with global VARs. *International Journal of Forecasting*, 25:642–675.

Pesaran, M. H., Schuermann, T., and Weiner, S. M. (2004). Modeling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business & Economics Statistics*, 22:129–162.

Pesaran, M. H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68:79–113.

Qu, R., Timmermann, A., and Zhu, Y. (2023). Comparing forecasting performance in cross-sections. *Journal of Econometrics*, 237:105186.

Qu, R., Timmermann, A., and Zhu, Y. (2024). Comparing forecasting performance in panels. *International Journal of Forecasting*, forthcoming.

Rapach, D. and Wohar, M. (2004). Testing the monetary model of exchange rate determination: A closer look at panels. *Journal of International Money and Finance*, 23:867–895.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica*, 38:311–323.

Trapani, L. and Urga, G. (2009). Optimal forecasting with heterogeneous panels: A Monte Carlo study. *International Journal of Forecasting*, 25:567–586.

Wang, W., Zhang, X., and Paap, R. (2019). To pool of not to pool: What is a good strategy for parameter estimation and forecasting in panels. *Journal of Applied Econometrics*, 34:724–745.

Zellner, A. and Hong, C. (1989). Forecasting international growth rates using Bayesian shrinkage and other procedures. *Journal of Econometrics*, 40:183–202.