NBER WORKING PAPER SERIES

FORECASTING WITH DYNAMIC PANEL DATA MODELS

Laura Liu
Hyungsik Roger Moon
Frank Schorfheide

Working Paper 25102
http://www.nber.org/papers/w25102

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2018

We thank Xu Cheng, Frank Diebold, Ulrich Mueller, Peter Phillips, Akhtar Siddique, and participants at various seminars and conferences for helpful comments and suggestions. Moon and Schorfheide gratefully acknowledge financial support from the National Science Foundation under Grants SES 1625586 and SES 1424843, respectively. The views expressed herein are those of the authors and do not necessarily reflect the views of the Board of Governors, the Federal Reserve System, or the National Bureau of Economic Research.

Forecasting with Dynamic Panel Data Models
Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide
NBER Working Paper No. 25102
September 2018
JEL No. C11,C14,C23,C53,G21

ABSTRACT

This paper considers the problem of forecasting a collection of short time series using cross sectional information in panel data. We construct point predictors using Tweedie's formula for the posterior mean of heterogeneous coefficients under a correlated random effects distribution. This formula utilizes cross-sectional information to transform the unit-specific (quasi) maximum likelihood estimator into an approximation of the posterior mean under a prior distribution that equals the population distribution of the random coefficients. We show that the risk of a predictor based on a non-parametric kernel estimate of the Tweedie correction is asymptotically equivalent to the risk of a predictor that treats the correlated-random-effects distribution as known (ratio-optimality). Our empirical Bayes predictor performs well compared to various competitors in a Monte Carlo study. In an empirical application we use the predictor to forecast revenues for a large panel of bank holding companies and compare forecasts that condition on actual and severely adverse macroeconomic conditions.

Laura Liu
Federal Reserve Board
20th Street and Constitution
Avenue N.W.
Washington, DC 20551
laura.liu@frb.gov

Hyungsik Roger Moon
University of Southern California
Department of Economics
KAP 300
University Park Campus
Los Angeles, CA 90089
hyungsikmoon@gmail.com

Frank Schorfheide
University of Pennsylvania
Department of Economics
The Ronald O. Perelman Center for
Political Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297
and NBER
schorf@ssc.upenn.edu

A supplemental appendix is available at: https://www.nber.org/data-appendix/w25102/

1 Introduction

The main goal of this paper is to forecast a collection of short time series. Examples are the performance of start-up companies, developmental skills of small children, and revenues and leverage of banks after significant regulatory changes. In these applications the key difficulty lies in the efficient implementation of the forecast. Due to the short time span, each time series taken by itself provides insufficient sample information to precisely estimate unit-specific parameters. We will use the cross-sectional information in the sample to make inference about the distribution of heterogeneous parameters. This distribution can then serve as a prior for the unit-specific coefficients to sharpen posterior inference based on the short time series.

More specifically, we consider a linear dynamic panel model in which the unobserved individual heterogeneity, which we denote by the vector λ_i, interacts with some observed predictors W_{it-1}:

$$Y_{it} = \lambda_i' W_{it-1} + \rho' X_{it-1} + \alpha' Z_{it-1} + U_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{1}$$

X_{it-1} is a vector of predetermined variables that may include lags of Y_{it}, Z_{it-1} is a vector of strictly exogenous covariates, and U_{it} is an unpredictable shock. Throughout this paper we adopt a correlated random effects approach in which the λ_is are treated as random variables that are possibly correlated with some of the predictors. An important special case is the linear dynamic panel data model in which $W_{it-1} = 1$, λ_i is a heterogeneous intercept, and the sole predictor is the lagged dependent variable: $X_{it-1} = Y_{it-1}$.

We develop methods to generate point forecasts of Y_{iT+1}, assuming that the time dimension T is short relative to the number of predictors (W_{iT}, X_{iT}, Z_{iT}). The forecasts are evaluated under a quadratic loss function. In this setting accurate estimates of the unit-specific parameters λ_i facilitate accurate forecasts. Our paper builds on the dynamic panel literature that has developed consistent estimators of the common parameters (α, ρ) and focuses on the estimation of λ_i, which is essential for the prediction of Y_{it}.

The benchmark for our prediction methods is the forecast that is based on the knowledge of the common coefficients (α, ρ) and the distribution $\pi(\lambda_i|\cdot)$ of the heterogeneous coefficients λ_i, but not the values λ_i themselves. This forecast is called oracle forecast. Because we are interested in forecasts for the entire cross section of N units, a natural notion of risk is that of compound risk, which is a (possibly weighted) cross-sectional average of expected losses.

In a correlated random-effects setting, this averaging is done under the distribution $\pi(\lambda_i|\cdot)$, which means that the compound risk associated with the forecasts of the N units is the same as the integrated risk for the forecast of a particular unit i. It is well known, that the integrated risk is minimized by the Bayes predictor that minimizes the posterior expected loss conditional on time T information for unit i. Thus, the oracle replaces λ_i by its posterior mean under the prior distribution $\pi(\lambda_i|\cdot)$.

The implementation of the oracle forecast is infeasible because in practice neither the common coefficients (ρ, α) nor the distribution of the unit-specific coefficients $\pi(\lambda_i|\cdot)$ are known. To approximate the oracle predictor, we first replace the unknown common parameters by a consistent ($N \longrightarrow \infty$, T is fixed) estimator. Second, rather than computing the posterior mean of λ_i based on the likelihood function and the prior $\pi(\lambda_i|\cdot)$, we use a formula – attributed to separate works of the astronomer Arthur Eddington and the statistician Maurice Tweedie – that expresses the posterior mean of λ_i as a function of the cross-sectional density of certain sufficient statistics. We estimate this density either parametrically or non-parametrically from the cross-sectional information in the panel data set and plug the density estimate into what is in the statistics literature commonly referred to as Tweedie's formula. This leads to an empirical Bayes estimate of λ_i and an empirical Bayes predictor of Y_{iT+1}.[1] The posterior mean predictor shrinks the estimates of the unit-specific coefficients toward a common prior mean, which reduces its sampling variability.

Our paper makes three contributions. First, we show in the context of the basic dynamic panel data model that an empirical Bayes predictor based on a consistent estimator of (ρ, α) and a kernel estimator of the cross-sectional density of the relevant sufficient statistics can asymptotically ($N \longrightarrow \infty$, T is fixed) achieve the same compound risk as the oracle predictor. Our main theorem provides a significant extension of the central result in Brown and Greenshtein (2009) from a vector-of-means model to a panel data model with estimated common coefficients. Importantly, the convergence result is uniform over families Π of correlated random effects distributions $\pi(\lambda_i|y_{i0})$ that contain point masses, i.e., span parameter homogeneity. As in Brown and Greenshtein (2009), we are able to show that the rate of convergence to the oracle risk accelerates in the case of homogeneous λ coefficients. Second, we provide a detailed Monte Carlo study that compares the performance of various implementations, both non-parametric and parametric, of our predictor. Third, we use our techniques to forecast pre-provision net-revenues (PPNRs) of a panel of banks.

[1] A fully Bayesian procedure would place a prior distribution on $\pi(\lambda_i|\cdot)$ and then integrate over the unknown density rather than replacing it with a plug-in estimate.

Our empirical Bayes predictor can be compared to two easily implementable benchmark predictors whose implicit assumptions about the distribution of the λ_i's correspond to special cases of our $\pi(\lambda_i|\cdot)$. The first predictor, which we call plug-in predictor, is obtained by estimating λ_i conditional on $(\hat{\rho}, \hat{\alpha})$ by maximum likelihood for each unit i. This predictor can be viewed as an approximation to the empirical Bayes predictor if $\pi(\lambda_i|\cdot)$ is very uninformative relative to the likelihood function associated with unit i. The second predictor, which we call pooled-OLS predictor, assumes homogeneity, i.e., $\lambda_i = \lambda$ for all i, and is based on a joint maximum likelihood estimate of (ρ, α, λ). It approximates the empirical Bayes estimator if $\pi(\lambda_i|\cdot)$ is very concentrated relative to unit i's likelihood.

Asymptotically, the plug-in predictor and the pooled-OLS predictor are suboptimal because they do not converge to the oracle predictor. However, in finite samples at least one of them, depending on the amount of heterogeneity in the data, may work quite well because they do not rely on (potentially noisy) density estimates. In our Monte Carlo simulations and in the empirical analysis we document that in practice the empirical Bayes predictor dominates, either weakly or strictly, both the plug-in predictor and the pooled-OLS predictor. While our theoretical results are based on a kernel implementation of the empirical Bayes predictor, in the Monte Carlo simulation and the empirical application we also consider finite mixture and nonparametric maximum likelihood estimates of the density of sufficient statistics required for the evaluation of Tweedie's formula.

In our empirical application we forecast PPNRs of bank holding companies. The stress tests that have become mandatory under the Dodd-Frank Act require banks to establish how revenues vary in stressed macroeconomic and financial scenarios. We capture the effect of macroeconomic conditions on bank performance by including the unemployment rate, an interest rate, and an interest rate spread in the vector W_{it-1} in (1). Our analysis consists of two steps. We first document the superior forecast accuracy of the empirical Bayes predictor developed in this paper under the actual economic conditions, meaning that we set the aggregate covariates to their observed values. In a second step, we replace the observed values of the macroeconomic covariates by counterfactual values that reflect severely adverse macroeconomic conditions. According to our estimates, the effect of stressed macroeconomic conditions on bank revenues is heterogeneous, but typically small relative to the cross-sectional dispersion of revenues across holding companies.

Our paper is related to several strands of the literature. For $\alpha = \rho = 0$ and $W_{it} = 1$ the problem analyzed in this paper reduces to the problem of estimating a vector of means, which is a classic problem in the statistic literature. In this context, Tweedie's formula has been used, for instance, by Robbins (1951) and more recently by Brown and Greenshtein (2009) and Efron (2011) in a "big data" application. Throughout this paper we are adopting an empirical Bayes approach, that uses cross-sectional information to estimate aspects of the prior distribution of the correlated random effects and then conditions on these estimates. Empirical Bayes methods also have a long history in the statistics literature going back to Robbins (1956) (see Robert (1994) for a textbook treatment).

We use compound decision theory as in Robbins (1964), Brown and Greenshtein (2009), Jiang and Zhang (2009) to state our optimality result. Because our setup nests the linear dynamic panel data model, we utilize results on the consistent estimation of ρ in dynamic panel data models with fixed effects when T is small, e.g., Anderson and Hsiao (1981), Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998), Alvarez and Arellano (2003). Fully Bayesian approaches to the analysis of dynamic panel data models have been developed in Chamberlain and Hirano (1999), Hirano (2002), Lancaster (2002).

The papers that are most closely related to ours are Gu and Koenker (2016, 2017). They also consider a linear panel data model and use Tweedie's formula to construct an approximation to the posterior mean of the heterogeneous regression coefficients. However, their main object of interest is the estimation of the heterogeneous coefficients, and not out-of-sample forecasting. Their papers implement the empirical Bayes predictor based on a nonparametric maximum likelihood estimator, following Kiefer and Wolfowitz (1956), of the cross-sectional distribution of the sufficient statistics, but do not provide any theoretical optimality result. A key contribution of our work is to establish a rate at which the regret associated with the empirical Bayes predictor vis-a-vis the oracle predictor vanishes uniformly across families of correlated random-effects distributions Π that include point masses. Moreover, we also provide novel Monte Carlo and empirical evidence on the performance of the empirical Bayes procedures.

Liu (2018) develops a fully Bayesian (as opposed to empirical Bayes) approach to generate panel data forecasts. She uses Dirichlet process mixtures (random effects) and mixtures of Gaussian linear regressions (correlated random effects) to construct a prior for the distribution of the heterogeneous coefficients, which then is updated in view of the observed panel data. While the fully Bayesian approach is more suitable for density forecasting and can be more easily extended to nonlinear panel data models (see Liu, Moon, and Schorfheide (2018) for an extension to a panel Tobit model), it is also a lot more computationally intensive. Moreover, it is much more difficult to establish convergence rates. Liu (2018) shows that her posterior predictive density converges strongly to the oracle's predictive density, but does not establish uniform bounds on the regret associated with the Bayes predictor.

There is an earlier panel forecast literature (e.g., see the survey article by Baltagi (2008) and its references) that is based on the best linear unbiased prediction (BLUP) proposed by Goldberger (1962). Compared to the BLUP-based forecasts, our forecasts based on Tweedie's formula have several advantages. First, it is known that the estimator of the unobserved individual heterogeneity parameter based on the BLUP method corresponds to the Bayes estimator based on a Gaussian prior (see, for example, Robinson (1991)), while our estimator based on Tweedie's formula is consistent with much more general prior distributions. Second, the BLUP method finds the forecast that minimizes the expected quadratic loss in the class of linear (in $(Y_{i0}, ..., Y_{iT})'$) and unbiased forecasts. Therefore, it is not necessarily optimal in our framework that constructs the optimal forecast without restricting the class of forecasts. Third, the existing panel forecasts based on the BLUP were developed for panel regressions with random effects and do not apply to correlated random effects settings.

Finally, there is a literature on "top-down" stress testing, which relies on publicly available bank-level income and capital data. Some authors, e.g. Covas, Rump, and Zakrajsek (2014), use time series quantile regression techniques to analyze revenue and balance sheet data for the relatively small set of bank holding companies with consolidated assets of more than 50 billion dollars. There are slightly more than 30 of these companies and they are subject to the Comprehensive Capital Analysis and Review conducted by the Federal Reserve Board of Governors. Because of mergers and acquisitions a lot of care is required to construct sufficiently long synthetic data sets that are amenable to time series analysis.

Closer to our work are the studies by Hirtle, Kovner, Vickery, and Bhanot (2016) and Kapinos and Mitnik (2016), which analyze PPNR components for a broader panel of banks. The former paper considers, among other techniques, pooled OLS estimation of models for bank income components and then uses the models to compute predictions under stressed macroeconomic conditions. The latter paper compares a standard fixed effect approach, a bank-by-bank time series approach, and fixed effects with optimal grouping in terms of out-of-sample forecasts, and finds that the bank-by-bank time-series approach does not perform well while the grouping approach provides better performance, which parallels our findings that the empirical Bayes methods usually outperform pooled OLS and plug-in predictors.

The remainder of the paper is organized as follows. Section 2 specifies the forecasting problem in the context of the basic dynamic panel data model. We introduce the compound decision problem, present the oracle forecast and its implementation through Tweedie's formula, and discuss the practical implementation. Section 3 establishes the ratio optimality of the empirical Bayes predictor and contains the main theoretical result of the paper. Monte Carlo simulation results are presented in Section 4. Section 5 discusses extensions to a more general linear panel data model with covariates and presents identification results for the homogeneous parameters and the correlated random effects distribution. The empirical application is presented in Section 6 and Section 7 concludes. Technical derivations, proofs, and the description of the data set used in the empirical analysis are relegated to the Appendix.

2 The Basic Dynamic Panel Data Model

We consider a panel with observations for cross-sectional units $i = 1, \dots, N$ in periods $t = 1, \dots, T$. The goal is to generate a point forecast for each unit i for period $t = T + h$. For now, we set $h = 1$ and assume that the observations Y_{it} are generated from a basic dynamic panel data model with homoskedastic Gaussian innovations:

$$Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}, \quad U_{it} \sim iid N(0, \sigma^2), \quad (\lambda_i, Y_{i0}) \sim iid\, \pi(\lambda, y_0). \tag{2}$$

This model is a restricted version of (1), where $W_{it-1} = 1$, $X_{it-1} = Y_{it-1}$, and $\alpha = 0$. Thus, λ_i is a unit-specific (heterogeneous) intercept. We combine the homogeneous parameters into the vector

$$\theta = (\rho, \sigma^2).$$

The purpose of investigating the simple model is to develop a full econometric theory of optimal forecasting.

In Section 2.1 we define the loss function under which the forecasts are evaluated and specify how we take expectations to construct our measure of risk. We construct an infeasible benchmark forecast in Section 2.2. This so-called oracle forecast is based on the posterior mean of λ_i under the prior $\pi(\lambda, y_0)$. The posterior mean can be conveniently evaluated using Tweedie's formula, which is discussed in Section 2.3. Finally, Section 2.4 describes how the infeasible oracle forecast can be turned into a feasible forecast by replacing unknown objects with estimates based on the cross-sectional information contained in the panel. Later in Section 5, we discuss extensions to the more general forecasting model (1), which is more relevant to empirical applications.

2.1 Compound Risk

The unit-specific forecasts $\hat{Y}_{i,T+1}$ are evaluated under the conventional quadratic loss function and the forecast error losses are summed over the units i to obtain the compound loss

$$L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1}) = \sum_{i=1}^{N} (\widehat{Y}_{iT+1} - Y_{iT+1})^2, \tag{3}$$

where $Y^N_{T+1} = (Y_{1T+1}, \dots, Y_{NT+1})$. We define the compound risk by taking expectations over (indicated by superscripts) the observed trajectories $\mathcal{Y}^N = (Y_1^{0:T}, \dots, Y_N^{0:T})$ with $Y_i^{0:T} = (Y_{i0}, Y_{i1}, \dots, Y_{iT})$, the unobserved heterogeneous coefficients $\lambda^N = (\lambda_1, \dots, \lambda_N)$, and future shocks $U^N_{T+1} = (U_{1T+1}, \dots, U_{NT+1})$:

$$R_N(\widehat{Y}^N_{T+1}) = \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N,\lambda^N,U^N_{T+1}} \left[L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1}) \right]. \tag{4}$$

We use the subscripts to indicate that the expectation is conditional on the homogeneous parameter θ and the correlated random effects distribution π. Upper case variables, e.g., Y_{it}, are generally used to denote random variables, and lower case variables, e.g., y_{it}, to denote their realizations.

2.2 Oracle Forecast

In order to develop an optimality theory for the panel data forecasts, we begin by characterizing an infeasible benchmark forecast, which is called the oracle forecast. In the compound decision theory it is assumed that the oracle knows the distribution of the heterogeneous coefficients $\pi(\lambda_i, h_i)$, but it does not know the specific λ_i for unit i. In addition, the oracle knows θ and has observed the trajectories $\mathcal{Y}^N$.

Conditional on θ, the compound risk takes the form of an integrated risk that can be expressed as

$$R_N(\widehat{Y}^N_{T+1}) = \mathbb{E}_{\theta,\pi}^{\mathcal{Y}^N} \left[\mathbb{E}_{\theta,\pi,\mathcal{Y}^N}^{\lambda^N,U^N_{T+1}} [L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1})] \right]. \tag{5}$$

The inner expectation can be interpreted as posterior risk, which is obtained by conditioning on the observations $\mathcal{Y}^N$ and integrating over the heterogeneous parameter λ^N and the shocks U^N_{T+1}. The outer expectation averages over the possible trajectories $\mathcal{Y}^N$. It is well known that the integrated risk is minimized by choosing the forecast that minimizes the posterior risk (with the understanding that we are conditioning on (θ, π) throughout) for each realization $\mathcal{Y}^N$.

Using the independence across i, the posterior risk can be written as follows:

$$\mathbb{E}^{\lambda^N, U^N_{T+1}}_{\theta,\pi,\mathcal{Y}^N}[L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1})] = \sum_{i=1}^{N} \left\{ \left(\widehat{Y}_{iT+1} - \mathbb{E}^{\lambda_i, U_{iT+1}}_{\theta,\pi,\mathcal{Y}_i}[Y_{iT+1}] \right)^2 + \mathbb{V}^{\lambda_i, U_{iT+1}}_{\theta,\pi,\mathcal{Y}_i}[Y_{iT+1}] \right\}, \tag{6}$$

where $\mathbb{V}^{\lambda_i, U_{iT+1}}_{\theta,\pi,\mathcal{Y}_i}[\cdot]$ is the posterior predictive variance of Y_{iT+1}. The decomposition of the risk into a squared bias term and the posterior variance of Y_{iT+1} implies that the optimal predictor is the mean of the posterior predictive distribution. Because U_{iT+1} is mean-independent of λ_i and $\mathcal{Y}_i$, we obtain

$$\widehat{Y}^{opt}_{iT+1} = \mathbb{E}^{\lambda_i, U_{iT+1}}_{\theta,\pi,\mathcal{Y}_i}[Y_{iT+1}] = \mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] + \rho Y_{iT}. \tag{7}$$

The compound risk associated with the oracle forecast is

$$R^{\text{opt}}_N = \mathbb{E}^{\mathcal{Y}^N}_{\theta} \left[\sum_{i=1}^{N} \left\{ \mathbb{V}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] + \sigma^2 \right\} \right]. \tag{8}$$

According to (8), the compound oracle risk has two components. The first component reflects uncertainty with respect to the heterogeneous coefficient λ_i and the second component captures uncertainty about the error term U_{iT+1}. Unfortunately, the direct implementation of the oracle forecast is infeasible because neither the parameter vector θ nor the correlated random effect distribution (or prior) $\pi(\cdot)$ are known. Thus, the oracle risk R^{opt}_N provides a lower bound for the risk that is attainable in practice.

2.3 Tweedie's Formula

The posterior mean $\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i]$ that appears in (7) can be evaluated using a formula which is named after the statistician Maurice Tweedie (though it had been previously derived by the astronomer Arthur Eddington). This formula is convenient for our purposes, because it expresses the posterior mean not as a function of the in practice unknown correlated random effects density $\pi(\lambda_i, y_{i0})$ but instead in terms of the marginal distribution of a sufficient statistic, which can be estimated from the cross-sectional information.

The contribution of unit i to the likelihood function associated with the basic dynamic panel data model in (2) is given by

$$p(y_i^{1:T}|y_{i0},\lambda_i,\theta) \propto \exp\left\{-\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{it}-\rho y_{it-1}-\lambda_i)^2\right\} \propto \exp\left\{-\frac{T}{\sigma^2}\big(\hat{\lambda}_i(\rho)-\lambda_i\big)^2\right\}, \tag{9}$$

where $\propto$ denotes proportionality and the sufficient statistic $\hat{\lambda}_i(\rho)$ is

$$\hat{\lambda}_i(\rho) = \frac{1}{T}\sum_{t=1}^{T}(Y_{it}-\rho Y_{it-1}). \tag{10}$$

Using Bayes Theorem, the posterior distribution of λ_i can be expressed as

$$p(\lambda_i|y_i^{0:T},\theta) = p(\lambda_i|\hat{\lambda}_i,y_{i0},\theta) = \frac{p(\hat{\lambda}_i|\lambda_i,y_{i0},\theta)\pi(\lambda_i|y_{i0})}{\exp\big\{\ln p(\hat{\lambda}_i|y_{i0})\big\}}, \tag{11}$$

where $p(\hat{\lambda}_i|\lambda_i,y_{i0},\theta)$ is proportional to the right-hand side of (9).

To obtain a representation for the posterior mean, we now differentiate the equation $\int p(\lambda_i|\hat{\lambda}_i,y_{i0},\theta)d\lambda = 1$ with respect to $\hat{\lambda}_i$. Exchanging the order of integration and differentiation and using the properties of the exponential function, we obtain

$$\begin{aligned} 0 &= \frac{1}{\sigma^2}\int(\lambda_i-\hat{\lambda}_i)p(\lambda_i|\hat{\lambda}_i,y_{i0},\theta)d\lambda_i - \frac{\partial}{\partial\hat{\lambda}_i}\ln p(\hat{\lambda}_i|y_{i0},\theta) \\ &= \frac{1}{\sigma^2}\big(\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]-\hat{\lambda}_i\big) - \frac{\partial}{\partial\hat{\lambda}_i}\ln p(\hat{\lambda}_i|y_{i0},\theta). \end{aligned}$$

Solving this equation for the posterior mean yields Tweedie's formula:[2]

$$\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] = \hat{\lambda}_i(\rho) + \frac{\sigma^2}{T}\frac{\partial}{\partial\hat{\lambda}_i(\rho)}\ln p(\hat{\lambda}_i(\rho),Y_{i0}). \tag{12}$$

Tweedie's formula was used by Robbins (1951) to estimate a vector of means λ^N for the model $Y_i|\lambda_i \sim N(\lambda_i,1)$, $\lambda_i \sim \pi(\cdot)$, $i=1,\dots,N$. Recently, it was extended by Efron (2011) to the family of exponential distributions, allowing for an unknown finite-dimensional parameter θ. The posterior mean takes the form of the sum of the sufficient statistic $\hat{\lambda}_i(\theta)$ and a correction term that captures the effect of the prior distribution of λ_i on the posterior. The correction term is expressed as a function of the marginal density of the sufficient statistic

[2] We replaced the conditional log density $\ln p(\hat{\lambda}_i(\rho),Y_{i0})$ by the joint log density $\ln p(\hat{\lambda}_i(\rho),Y_{i0})$ because the two differ only by a constant which drops out after the differentiation.

$\hat{\lambda}_i(\theta)$ conditional on Y_{i0} and θ. Thus, to evaluate the posterior mean, it is not necessary to explicitly solve a deconvolution problem that separates the prior density $\pi(\lambda_i|y_{i0})$ from the distribution of the error terms U_{it}.

2.4 Implementation

We approximate the oracle forecast using an empirical Bayes approach that replaces the unknown objects θ and $p(\hat{\lambda}_i(\rho), Y_{i0})$ in (12) by estimates that exploit the cross-sectional information. A key requirement for an estimator of the homogeneous parameter θ is that it is consistent. In our basic dynamic panel data model, consistency can be achieved by various types of generalized method of moments (GMM) estimators, e.g., Arellano and Bond (1991), Arellano and Bover (1995), or Blundell and Bond (1998), or by a quasi-maximum-likelihood estimator (QMLE) that integrates out the heterogeneous λ_i's under the misspecified correlated random effects distribution $\lambda_i|Y_{i0} \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$.[3] The density $p(\hat{\lambda}_i(\rho), Y_{i0})$ could be estimated using kernel methods, a mixture approximation, or nonparametric maximum likelihood. In the next section, we provide an optimality result for the kernel estimator and we illustrate the performance of other estimators in a Monte Carlo study in Section 4.

3 Ratio Optimality

We now will prove that the predictor $\widehat{Y}^N_{T+1}$ that is constructed by replacing θ with a consistent estimator $\hat{\theta}$ and by replacing $p(\hat{\lambda}_i(\rho), Y_{i0})$ with a kernel density estimator achieves ϵ_0-ratio optimality uniformly for priors $\pi \in \Pi$. That is, for any $\epsilon_0 > 0$

$$\limsup_{N \to \infty} \sup_{\pi \in \Pi} \frac{R_N(\widehat{Y}^N_{T+1}; \pi) - R_N^{\text{opt}}(\pi)}{N \mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\mathbb{V}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]\right] + N^{\epsilon_0}} \leq 0. \tag{13}$$

The convergence is uniform with respect to the correlated random effects distributions π in some set Π that we will characterize in more detail below. The uniformity holds in the neighborhood of point masses for which the prior and posterior variances of λ_i are zero. Thus, the convergence statement covers the case of λ_i being homogeneous across i.

First, consider the numerator in (13). The autoregressive coefficient in basic dynamic panel model can be $\sqrt{N}$-consistently estimated, which suggests that $\sum_{i=1}^N (\hat{\rho} - \rho)^2 Y_{iT}^2 =$

[3] This estimator is described in more detail in Section 4 and its consistency is proved in the Online Appendix.

$O_p(1)$. Thus, whether a predictor $\widehat{Y}_{iT+1}$ attains ratio optimality crucially depends on the rate at which the discrepancy between $\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]$ and $\widehat{\mathbb{E}}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]$ vanishes, where the latter is a function of a nonparametric estimate of $p(\hat{\lambda}_i(\rho), Y_{i0})$. Second, note that the denominator of the ratio in (13) is strictly positive due to the N^{ϵ_0} term and divergent. The rate of divergence depends on the posterior variance of λ_i. If the posterior variance is strictly greater than zero, then the denominator is of order $O(N)$. Because the posterior is based on a finite number of observations T, the posterior variance is zero only if the prior density $\pi(\lambda)$ is a point mass. In this case the definition of ratio optimality requires that the regret vanishes at a faster rate, because the rate of the numerator drops from $O(N)$ to N^{ϵ_0}.[4]

The proof of the ratio-optimality result presented below in Theorem 3.7 below is a significant generalization of the proof in Brown and Greenshtein (2009), allowing for the presence of estimated parameters in the sufficient statistic $\hat{\lambda}(\cdot)$ and uniformity with respect to the correlated random effect density $\pi(\cdot)$, which is allowed to have a unbounded support. The remainder of this section is organized as follows. The kernel estimator of $p(\hat{\lambda}_i(\rho), Y_{i0})$ and the resulting formula for the predictor $\widehat{Y}_{iT+1}$ are presented in Section 3.1. In Section 3.2 we provide high-level assumptions that lead to the main theorem. In Section 3.3 we show that the high-level conditions are satisfied if $\pi(\cdot)$ belongs to a collection of finite-mixtures of Gaussian random variables with bounded means and variances.

3.1 Kernel Estimation and Truncation

To facilitate the theoretical analysis, we use a leave-one-out kernel density estimator of the form:

$$\hat{p}^{(-i)}(\hat{\lambda}_i(\rho), y_{i0}) = \frac{1}{N-1}\sum_{j\neq i}\frac{1}{B_N}\phi\left(\frac{\hat{\lambda}_j(\rho)-\hat{\lambda}_i(\rho)}{B_N}\right)\frac{1}{B_N}\phi\left(\frac{Y_{j0}-y_{i0}}{B_N}\right), \tag{14}$$

where $\phi(\cdot)$ is the pdf of a $N(0,1)$ and B_N is the kernel bandwidth. Using the fact that the observations are cross-sectionally independent and conditionally normally distributed one

[4] If it were known that the λ_i's are in fact homogeneous and the model is estimated with a common intercept, then $\sqrt{N}(\hat{\lambda}-\lambda) = O_p(1)$ and the regret in the numerator would be $O(1)$. Thus, the convergence result could also be achieved by standardizing with sequences that diverge at a rate slower than N^{ϵ_0}.

can directly compute the expected value of the leave-one-out estimator:

$$\begin{aligned} \mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] &= \int \frac{1}{\sqrt{\sigma^2/T + B_N^2}} \phi\left(\frac{\hat{\lambda}_i - \lambda_i}{\sqrt{\sigma^2/T + B_N^2}}\right) \\ &\quad \times \left[\int \frac{1}{B_N}\phi\left(\frac{y_{i0} - \tilde{y}_{i0}}{B_N}\right) p(\tilde{y}_{i0}|\lambda_i) d\tilde{y}_{i0}\right] p(\lambda_i) d\lambda_i. \end{aligned} \tag{15}$$

Taking expectations of the kernel estimator leads to a variance adjustment for the conditional distribution of $\hat{\lambda}_i|\lambda_i$ ($\sigma^2/T + B_N^2$ instead of σ^2/T) and the density of $Y_{i0}|\lambda_i$ is replaced by a convolution. We define:

$$\begin{aligned} \pi_*(\lambda, y_0) &= \int \frac{1}{B_N}\phi\left(\frac{y_0 - \tilde{y}_0}{B_N}\right) \pi(\lambda, \tilde{y}_0) d\tilde{y}_0 \\ p_*(\hat{\lambda}, y_0; \pi) &= \int \frac{1}{\sqrt{\sigma^2/T + B_N^2}} \phi\left(\frac{\hat{\lambda} - \lambda}{\sqrt{\sigma^2/T + B_N^2}}\right) \pi_*(\lambda, y_0) d\lambda, \end{aligned} \tag{16}$$

such that we can write

$$\mathbb{E}^{\mathcal{Y}^{(-i)}}_{\theta,\pi,\mathcal{Y}_i}[\hat{p}^{(-i)}(\hat{\lambda}_i, y_{i0})] = p_*(\hat{\lambda}_i, y_{i0}; \pi). \tag{17}$$

In view of the variance adjustment for the distribution of $\hat{\lambda}_i|\lambda_i$ induced by taking expectations of the kernel estimator, we replace the scale factor σ^2/T in the Tweedie correction term in (12) by $\hat{\sigma}^2/T + B_N^2$. Moreover, we truncate the absolute value of the posterior mean approximation from above. For $C > 0$ and for any $x \in \mathbb{R}$, define $[x]^C = \mathrm{sgn}(x)\min\{|x|, C\}$. The resulting predictor is

$$\widehat{Y}_{iT+1} = \left[\hat{\lambda}_i(\rho) + \left(\frac{\hat{\sigma}^2}{T} + B_N^2\right) \frac{\partial}{\partial \hat{\lambda}_i(\rho)} \ln \hat{p}^{-i}(\hat{\lambda}_i(\rho), Y_{i0})\right]^{C_N} + \hat{\rho} Y_{iT}, \tag{18}$$

where $C_N \longrightarrow \infty$ slowly.

3.2 Main Theorem

Let Π be a collection of joint densities $\pi(\lambda, y)$. The theoretical analysis relies heavily on slowly diverging sequences. To state the assumptions and prove the main theorem, the following definitions will be convenient:

Definition 3.1

(i) $A_N(\pi) = o_{u.\pi}(N^{\epsilon})$, *for* some $\epsilon > 0$, *if there exists a sequence* $\eta_N \longrightarrow 0$ *that does not depend on* $\pi \in \Pi$ *such that* $N^{-\epsilon}A_N(\pi) \leq \eta_N$.

(ii) $A_N = o(N^+)$ *if for* every $\epsilon > 0$ *there exists a sequence* $\eta_N(\epsilon) \longrightarrow 0$ *such that* $N^{-\epsilon}A_N(\pi) \leq \eta_N(\epsilon)$.

(iii) $A_N(\pi) = o_{u.\pi}(N^+)$ *(sub-polynomial) if for* every $\epsilon > 0$ *there exists a sequence* $\eta_N(\epsilon) \longrightarrow 0$ *that does not depend on* $\pi \in \Pi$ *such that* $N^{-\epsilon}A_N(\pi) \leq \eta_N(\epsilon)$.

Our first assumption controls the tails of the marginal distributions of λ and the initial condition Y_0.[5] We essentially assume that λ and Y_0 are subexponential random variables with finite fourth moments. The assumed tail probability and moment bounds are uniform for $\pi \in \Pi$.

Assumption 3.2 (Correlated Random Effects Distribution, Part 1) *There exist positive constants* $M_1 < \infty$, $M_2 < \infty$, $M_3 < \infty$, *and* $M_4 < \infty$ *such that for every* $\pi \in \Pi$*:*

(i) $\int_{|\lambda|\geq C} \pi(\lambda)d\lambda \leq M_1 \exp(-M_2(C - M_3))$ *and* $\int \lambda^4 \pi(\lambda)d\lambda \leq M_4$.

(ii) $\int_{|y_0|\geq C} \pi(y_0)dy_0 \leq M_1 \exp(-M_2(C - M_3))$ *and* $\int y_0^4 \pi(y_0)dy_0 \leq M_4$.

The proof of the main theorem relies on truncations. Assumption 3.3 imposes a lower bound and an upper bound on the diverging truncation sequences C_N and C'_N. These bounds on the truncation sequences constrain the rate at which the bandwidth B_N of the kernel density estimator has to shrink to zero. For technical reasons, the bandwidth cannot shrink as quickly as in typical density estimation problems.[6]

Assumption 3.3 (Trimming and Bandwidth)

(i) The truncation sequence C_N *satisfies* $C_N = o(N^+)$ *and* $C_N \geq 2(\ln N)/M_2$.

(ii) The truncation sequence C'_N *satisfies* $C'_N = C_N + \sqrt{(2\sigma^2 \ln N)/T}$.

(ii) The bandwidth sequence B_N *is bounded by* $\underline{B}_N \leq B_N \leq \bar{B}_N$, *where* $1/\underline{B}_N^2 = o(N^+)$, $\bar{B}_N(C_N + C'_N) = o(1)$ *and the bounds do not depend on the observed data or* $\pi \in \Pi$.

[5] To simplify the notation we write $\pi(\lambda)$ and $\pi(y_0)$ to denote the marginals of $\pi(\lambda, y_0)$.

[6] In a nutshell, we need to control the behavior of $\hat{p}(\hat{\lambda}_i, Y_{i0})$ and its derivative uniformly, which, in certain steps of the proof, requires us to consider bounds of the form M/B_N^2, where M is a generic constant. If the bandwidth shrinks too fast, the bounds diverge too quickly to ensure that it suffices to standardize the regret in Theorem 3.7 by N^{ϵ} if the λ_i coefficients are identical for each cross-sectional unit.

We also need to impose a restriction on the conditional distribution of Y_0 given λ. First, we assume that the conditional density is bounded. Second, we regularize the shape of the conditional density $\pi(y_0|\lambda)$ by assuming that the density of as a function of y_0 should be uniformly "smooth" such that its convolution with the Kernel density remains close (in relative terms) to the original density. This assumption is required because we use a single bandwidth (independent of y_0) when estimating the density $p(\hat{\lambda}(\rho), y_0)$ for the Tweedie correction term. Thus, we rule out, for instance, that $\pi(y_0|\lambda)$ is a point mass. However, it is important to note that we do allow the marginal distribution $\pi(\lambda)$ to be a point mass (or discrete).

Assumption 3.4 (Correlated Random Effects Distribution, Part 2)

(i) There exists an $M < \infty$ such that $\pi(y_0|\lambda) \leq M$ for all $\pi \in \Pi$.

(ii) For given sequences C_N, C'_N, and B_N satisfying Assumption 3.3, the conditional distribution $\pi(y_0|\lambda)$ satisfies

$$\sup_{|y_0| \leq C'_N, |\lambda| \leq C_N, \pi \in \Pi} \left| \frac{\frac{1}{B_N} \int \phi\left(\frac{\tilde{y}_0 - y_0}{B_N}\right) \pi(\tilde{y}_0|\lambda) d\tilde{y}_0}{\pi(y_0|\lambda)} - 1 \right| = o(1).$$

Next, we impose some restrictions on the sampling distributions of the posterior means. To do so, define the posterior mean function as

$$m(\hat{\lambda}, y_0; \pi) = \hat{\lambda} + \left(\sigma^2/T\right) \frac{\partial \ln p(\hat{\lambda}, y_0; \pi)}{\partial \hat{\lambda}}, \tag{19}$$

where the joint sampling distribution of the sufficient statistic and the initial condition is given by

$$p(\hat{\lambda}, y_0; \pi) = \int \frac{1}{\sqrt{\sigma^2/T}} \phi\left(\frac{\hat{\lambda} - \lambda}{\sqrt{\sigma^2/T}}\right) \pi(y_0, \lambda) d\lambda.$$

In order to be precise about the uniformity requirements for $\pi \in \Pi$, we now included the prior density π as a conditioning argument in the functions $m(\cdot)$ and $p(\cdot)$. Moreover, we denote the posterior mean function under the $*$-distribution defined in (17) as

$$m_*(\hat{\lambda}, y_0; \pi, B_N) = \hat{\lambda} + \left(\sigma^2/T + B_N^2\right) \frac{\partial \ln p_*(\hat{\lambda}, y_0; \pi, B_N)}{\partial \hat{\lambda}}. \tag{20}$$

While the sampling distribution of $(\hat{\lambda}_i, Y_{i0})$ is tightly linked to the prior $\pi(\lambda, y_0)$ and the distribution of the sufficient statistic $\hat{\lambda}|\lambda \sim N(\lambda, \sigma^2/T)$, we find it convenient to postulate some high-level conditions, that we will verify for collections of finite mixtures of multivariate Normals (FNMN) in Section 3.3.

Assumption 3.5 (Posterior Mean Functions) *For a given sequence C_N satisfying Assumption 3.3, the posterior mean functions satisfy:*

(i) $N \int \int m(\hat{\lambda}, y_0; \pi)^2 \, \mathbb{I}\left\{ |m(\hat{\lambda}, y_0; \pi)| \geq C_N \right\} p(\hat{\lambda}, y_0; \pi) \, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+)$,

(ii) $N \int \int m_*(\hat{\lambda}, y_0; \pi, B_N)^2 \, \mathbb{I}\left\{ |m_*(\hat{\lambda}, y_0; \pi, B_N)| \geq C_N \right\} p(\hat{\lambda}, y_0; \pi) \, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+)$,

(iii) $N \int \int m(\hat{\lambda}, y_0; \pi)^2 \, \mathbb{I}\left\{ |m(\hat{\lambda}, y_0; \pi)| \geq C_N \right\} p_*(\hat{\lambda}, y_0; \pi^*, B_N) \, d\hat{\lambda} dy_0 = o_{u.\pi}(N^+)$.

Finally, we impose moment restrictions on the sampling distribution of the estimators of the homogeneous parameters $\hat{\rho}$, and $\hat{\sigma}^2$.

Assumption 3.6 (Estimators of ρ and σ^2) *The estimators $\hat{\rho}$ and $\hat{\sigma}^2$ have the following properties: (i) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[|\sqrt{N}(\hat{\rho} - \rho)|^4 \right] = o_{u,\pi}(N^+)$, (ii) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\hat{\sigma}^4 \right] = o_{u.\pi}(N^+)$, and (iii) $\mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[|\sqrt{N}(\hat{\sigma}^2 - \sigma^2)|^2 \right] = o_{u.\pi}(N^+)$.*

An example of an estimator $\hat{\rho}$ that satisfies Assumption (3.6)(i) is the truncated instrumental variable (IV) estimator

$$\hat{\rho}_{IV} = \left[\left(\sum_{i=1}^{N} \sum_{t=2}^{T} Y_{it-2} \Delta Y_{it-1}, \right)^{-1} \right]^{M_N} \left(\sum_{i=1}^{N} \sum_{t=2}^{T} Y_{it-2} \Delta Y_{it} \right),$$

where M_N is a sequence that slowly diverges to infinity. Define the residuals $\hat{U}_{it} = Y_{it} - -\hat{\lambda}(\hat{\rho}_{IV}) - \hat{\rho}_{IV} \hat{Y}_{it-1}$. Then, the sample variance of the $\hat{U}_{it}$'s is an estimator of σ^2 that satisfies Assumption (3.6).

We are now in a position to present the main result, which states that the regret associated with the vector of predictors $\widehat{Y}^N_{T+1}$, standardized by the posterior variance of the heterogeneous parameters λ_i, converges to zero as the cross-sectional dimension N of the sample becomes large.

Theorem 3.7 *Suppose that Assumptions 3.2 to 3.6 are satisfied. Then, in the basic dynamic panel model (2), the predictor $\widehat{Y}_{iT+1}$ defined in (18) achieves ϵ_0-ratio optimality uniformly in $\pi \in \Pi$, that is, for every $\epsilon_0 > 0$*

$$\limsup_{N \to \infty} \sup_{\pi \in \Pi} \frac{R_N(\widehat{Y}^N_{T+1}; \pi) - R_N^{opt}(\pi)}{N \mathbb{E}^{\mathcal{Y}^N}_{\theta,\pi}\left[\mathbb{V}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i] \right] + N^{\epsilon_0}} \leq 0. \tag{21}$$

3.3 Two Examples of Π

We now provide two specific examples of classes of distributions Π that satisfy Assumptions 3.2, 3.4, and 3.5.

Multivariate Normal Distributions. To simplify the notation we write y instead of y_0. We define the class Π of correlated random effects densities as

$$\Pi = \left\{ \pi(\lambda, y) = \pi(\lambda)\pi(y|\lambda) \;\middle|\; \pi(\lambda) \in \Pi_\lambda,\ \pi(y|\lambda) \in \Pi_{y|\lambda} \right\} \tag{22}$$

where

$$\begin{aligned} \Pi_\lambda &: \left\{ N(\mu_\lambda, \sigma^2_\lambda) \;\middle|\; |\mu_\lambda| \le M_{\mu_\lambda},\ 0 \le \sigma^2_\lambda \le M_{\sigma^2_\lambda} \right\} \\ \Pi_{y|\lambda} &: \left\{ N(\alpha_0 + \alpha_1\lambda, \sigma^2_{y|\lambda}) \;\middle|\; |\alpha_0| \le M_{\alpha_0},\ |\alpha_1| \le M_{\alpha_1},\ 0 < \delta_{\sigma^2_{y|\lambda}} \le \sigma^2_{y|\lambda} \le M_{\sigma^2_{y|\lambda}} \right\} \end{aligned}$$

We interpret $\underline{\sigma}^2_\lambda = 0$ as a point mass and impose upper bounds on the mean and the variance of λ. To obtain joint normality of (λ, y) the conditional mean function for $y|\lambda$ is linear in λ and the conditional variance is constant. We bound the absolute values of the conditional mean parameters α_0 and α_1 as well as the conditional variance from above. The lower bound $\delta_{\sigma^2_{y|\lambda}}$ rules out a point-mass prior for $y|\lambda$. Let $\Pi = \Pi_\lambda \otimes \Pi_{y|\lambda}$.

Finite Mixtures of Multivariate Normals. This class of distribution is able to approximate a wide variety of distributions with exponential tails as the number of mixture components, K, increases. Formal approximation results are provided, for instance, in Norets and Pelenis (2012). Let $K < \infty$ be the maximum number of mixture components and define:

$$\Pi^{(K)}_{mix} = \left\{ \pi_{mix}(\lambda, y) = \sum_{k=1}^{K} \omega_k \pi_k(\lambda, y) \;\middle|\; \pi_k \in \Pi\, \forall k,\ 0 \le \omega_k \le 1,\ \sum_{k=1}^{K} \omega_k = 1 \right\}, \tag{23}$$

where the class Π is defined in (22).

Theorem 3.8 *Assumptions 3.2, 3.4, and 3.5 are satisfied by (i) Π in (22) and (ii) $\Pi^{(K)}_{mix}$ in (23).*

4 Monte Carlo Simulations

We now conduct three Monte Carlo experiments to illustrate the performance of the empirical Bayes predictor. We begin by describing the various predictors that we compare in the experiments.

Oracle Forecast. The oracle knows the parameters $\theta = (\rho, \gamma)$ as well as the correlated random effects distribution $\pi(\lambda_i, Y_{i0}|\xi)$. However, the oracle does not know the specific λ_i values. Its forecast is given by (7).

Empirical Bayes Estimators. All empirical Bayes estimators used in this section are based on the QMLE of θ.[7] The QMLE is derived from the possibly misspecified distribution $\lambda_i|(Y_{i0}, \xi) \sim N(\phi_0 + \phi_1 Y_{i0}, \underline{\Omega})$, where $\xi = (\phi_0, \phi_1, \underline{\Omega})$. We denote the density as $\pi_Q(\lambda_i|Y_{i0}, \xi)$ and define

$$\big(\hat{\theta}_{QMLE}, \hat{\xi}_{QMLE}\big) = \mathrm{argmax}_{\theta,\xi} \prod_{i=1}^{N} \int p(y_i^{1:T}|y_{i0}, \lambda_i, \theta)\pi_Q(\lambda_i|Y_{i0}, \xi)d\lambda_i. \tag{24}$$

We show in the Online Appendix that the QMLE in our Monte Carlo designs is consistent under misspecification of $\pi_Q(\cdot)$. To implement Tweedie's formula (see (12) for the basic model) we consider the following three estimates of $p(\hat{\lambda}_i(\rho), Y_{i0})$:

Kernel-based Tweedie Correction. We use the kernel estimator (14) for which we developed the asymptotic theory in Section 3. In Experiment 1 we consider a random-effects design which implies that we need to estimate a one-dimensional density $p(\hat{\lambda}_i(\hat{\rho}))$. In the other experiments we consider correlated random-effects designs which require a bivariate density estimation. Prior to the density estimation we standardize the sequences $\hat{\lambda}_i(\hat{\rho})$ and Y_{i0}, $i = 1, \dots, N$, which is equivalent to scaling the bandwidths B_N used in (14) by the standard deviations of the two series. For the standardized series, we set the bandwidth to $B_N = bB_N^*$, where $b \in \mathcal{B}$ and $\mathcal{B}$ is a finite grid for the scaling constant. The baseline value for the bandwidth as well as lower and upper bounds for b are given by

$$B_N^* = \left(\frac{4}{d+2}\right)^{1/(d+4)} \max\left\{\frac{1}{N^{1/(d+4)}}, \frac{1}{(\ln N)^{1.01}}\right\}, \quad \underline{b} \leq b \leq \bar{b}. \tag{25}$$

Here d corresponds to the dimension of the density that is being estimated and is either one or two. The scaling constant in front of the max operator as well as the first rate inside the

[7] Results based on GMM estimators can be found in the working paper version Liu, Moon, and Schorfheide (2017).

max operator correspond to Silverman (1986)'s rule-of-thumb bandwidth choice. The second rate in the max operator is consistent with Assumption 3.3. We report forecast evaluation statistics for various choices of b as well as a $\hat{b}$ that is generated by minimizing the average forecast error loss for $b \in \mathcal{B}$ when predicting the last set of observations in the estimation sample, Y_{iT}, based on $Y_{i0}, \ldots, Y_{iT-1}$.[8] By setting $\underline{B}_N = \underline{b}B_N^*$ and $\bar{B}_N = \bar{b}B_N^*$ we can deduce that the kernel-estimator with data-driven bandwidth scaling still satisfies Assumption 3.3.

Finite Mixtures of Multivariate Normals. We also consider mixtures of normal distributions to approximate $p(\hat{\lambda}_i(\hat{\rho}), Y_{i0})$:

$$p_{mix}\big(\hat{\lambda}_i, Y_{i0}\big|\{\omega_k, \mu_k, \Sigma_k\}_{k=1}^K\big) = \sum_{k=1}^K \omega_k p_N(\hat{\lambda}_i, Y_{i0}|\mu_k, \Sigma_k), \quad \omega_k \geq 0, \quad \sum_{k=1}^K \omega_K = 1,$$

where $p_N(\cdot)$ is the density of a $N(\mu_k, \Sigma_k)$ random variables. The mixture probabilities as well as the means and covariance matrices are estimated by maximizing the log likelihood function using an EM algorithm. The Tweedie corrections are then computed conditional on the estimates $(\hat{\omega}_k, \hat{\mu}_k, \hat{\Sigma}_k)$. We report forecast evaluation statistics for various choices of K as well as a $\hat{K}$ that is generated by minimizing the average forecast error loss for $1 \leq K \leq \bar{K}$ when predicting the last set of observations in the estimation sample, Y_{iT}, based on $Y_{i0}, \ldots, Y_{iT-1}$.

Nonparametric MLE Tweedie Correction. Gu and Koenker (2016) proposed to estimate the density of the sufficient statistic by nonparametric MLE. We only report results for this estimator in Experiment 1 where it can be implemented using the GLmix function of the REBayes package provided by Gu and Koenker (2016). In the random-effects setup, the estimator is constructed as follows. Specify bounds for the domain of λ_i and partition it into K bins. We adopt their default setting, where the number of bins is $K = 300$. Let λ_k be the right endpoint of bin k and Δ_k its width. Moreover, let $\tilde{\omega}_k$ be the probability associated with the k'th bin and define $\omega_k = \tilde{\omega}_k \Delta_k$. Then

$$p_{NP}\big(\hat{\lambda}_i\big|\{\omega_k\}_{k=1}^K\big) = \sum_{k=1}^K \omega_k p_N(\hat{\lambda}_i|\lambda_k, \sigma^2/T), \quad \omega_k \geq 0, \quad \sum_{k=1}^K \omega_k = 1.$$

The bin probabilities are estimated by maximizing the log likelihood function. Unlike in the mixture of normals case, here the means and the variances of the normal distributions are fixed and only the probabilities ω_k are estimated. The Tweedie corrections are then

[8] To do so, we also re-estimate the homogeneous parameter θ based on the reduced sample $Y_{i0}, \ldots, Y_{iT-1}$.

computed conditional on the estimates $\hat{\omega}_k$.

Alternative Predictors. In addition to the empirical Bayes predictors we consider several alternative procedures.

QMLE Plug-In Predictor. This predictor takes the form $\widehat{Y}_{iT+1} = \hat{\lambda}_i(\hat{\rho}_{QMLE}) + \hat{\rho}_{QMLE} Y_{iT}$ and does not use the Tweedie correction.

Pooled-OLS Predictor. Ignoring the heterogeneity in the λ_i's and imposing that $\lambda_i = \lambda$ for all i, we can define

$$(\hat{\rho}_P, \hat{\lambda}_P) = \mathrm{argmin}_{\rho,\lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(Y_{it} - \rho Y_{it-1} - \lambda\big)^2. \tag{26}$$

The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_P + \hat{\rho}_P Y_{iT}$.

Loss-Function-Based Predictor. We construct an estimator of (ρ, λ^N) based on the objective function:

$$\hat{\rho}_L = \mathrm{argmin}_{\rho} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(Y_{it} - \rho Y_{it-1} - \hat{\lambda}_i(\rho)\big)^2, \quad \hat{\lambda}_i(\rho) = \frac{1}{T} \sum_{t=1}^{T} \big(Y_{it} - \rho Y_{it-1}\big). \tag{27}$$

This estimator minimizes the loss function under which the forecasts are evaluated in sample. It is well-known that due to the incidental parameter problem, the estimator $\hat{\rho}_L$ is inconsistent under large-N and fixed-T asymptotics. The resulting predictor is $\widehat{Y}_{iT+1} = \hat{\lambda}_i(\hat{\rho}_L) + \hat{\rho}_L Y_{iT}$.

First-Difference Predictor. In the panel data literature it is common to difference-out idiosyncratic intercepts, which suggests predicting ΔY_{iT+1} based on ΔY_{iT}. We evaluate the first-difference predictor at the QMLE estimator of ρ to obtain $\widehat{Y}^{FD}_{iT+1}(\hat{\rho}_{QMLE})$.

Compound Risk. We generalize our definition of compound risk by introducing a selection rule that lets us focus on forecasts for a specific group of individuals. Let

$$D_i = D_i(\mathcal{Y}^N) \in \{0, 1\}, \quad i = 1, \ldots, N, \tag{28}$$

where $D_i(\mathcal{Y}^N)$ is a measurable function of the observations $\mathcal{Y}^N$. For instance, suppose that $D_i(\mathcal{Y}^N) = \mathbb{I}\{Y_{iT} \in A\}$ for $A \subset \mathbb{R}$. In this case, the selection is homogeneous across i and, for individual i, depends only on its own sample. Alternatively, suppose that units are selected based on the ranking of an index, e.g., the empirical quantile of Y_{iT}. In this case, the selection dummy D_i depends on $(Y_{1T}, ..., Y_{NT})$ and thereby also on the data for the other

Table 1: Monte Carlo Design 1

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0, \gamma^2)$. $\rho = 0.8$, $\gamma = 1$
Initial Observations: $Y_{i0} \sim N(0, 1)$
Random effects: $\lambda_i \vert Y_{i0} \sim \text{Gamma}(2, b)$, various choices of b
Sample Size: $N = 1,000$, $T = 4$
Number of Monte Carlo Repetitions: $N_{sim} = 1,000$

$N - 1$ individuals. The compound loss of interest is the average of the individual losses weighted by the selection dummies:

$$L_N(\widehat{Y}^N_{T+1}, Y^N_{T+1}) = \sum_{i=1}^{N} D_i(\mathcal{Y}^N)(\widehat{Y}_{iT+1} - Y_{iT+1})^2.$$

Because the selection rule is a function of the observed trajectories $\mathcal{Y}^N$ and the optimal predictor (7) is a Bayes predictor that conditions on $\mathcal{Y}^N$, the presence of the selection rule does not change the optimal forecast. Thus, the theory developed in Section 3 has a straightforward extension to the case $D_i(\mathcal{Y}^N) \neq 1$.

4.1 Experiment 1: Gamma-Distributed Random Effects

The first Monte Carlo experiment is based on the basic dynamic panel data model in (2). The design of the experiment is summarized in Table 1. We assume that the λ_i's follow a Gamma$(2, b)$ distribution and are uncorrelated with the initial condition Y_{i0} (random effects). The Gamma distribution has exponential tails and satisfies the tail bound condition in Assumption 3.2. We set $\rho = 0.8$, $\sigma^2 = 1$, and choose the parameter b to generate various values for $\mathbb{V}[\lambda_i]$. Each panel consists of $N = 1,000$ cross-sectional units and the number of time periods is $T = 4$. Generally, the smaller T relative to the number of right-hand-side variables with heterogeneous coefficients, the larger the gain from using a prior distribution to compute posterior mean estimates of the λ_i's. The subsequent results are based on $N_{sim} = 1,000$ Monte Carlo repetitions.

In Table 2 we report the regret associated with each predictor relative to the posterior variance of λ_i, averaged over all trajectories $\mathcal{Y}^N$, as specified in Theorem 3.7 (setting $\epsilon_0 = 0.1$ which leads to $N^{\epsilon_0} \approx 2$). For the oracle predictor the regret is by definition zero and we tabulate the risk R^{opt}_N instead (in parentheses). The columns titled "All" correspond to $D_i(\mathcal{Y}^N) = 1$. Using the 95% quantile of the population distribution of Y_{iT}, we define the

Table 2: Experiment 1: Relative Regrets under Gamma Random Effects

	$\mathbb{V}[\lambda_i] = 1$		$\mathbb{V}[\lambda_i] = 0.1$		$\mathbb{V}[\lambda_i] = .002$		$\mathbb{V}[\lambda_i] = 0$	
Predictor	All	Top	All	Top	All	Top	All	Top
Oracle Predictor	(1177)	(57.6)	(1067)	(57.7)	(1002)	(49.3)	(1000)	(49.1)
Empirical Bayes, $\hat{\theta}_{QMLE}$								
Kernel $b = \hat{b}$	0.04	0.20	0.12	0.42	2.30	1.37	4.61	1.48
Kernel $b = 1.0$	0.09	1.34	0.38	3.14	5.87	5.09	11.8	5.61
Kernel $b = 1.5$	0.05	0.25	0.12	0.54	2.08	1.24	4.21	1.35
Kernel $b = 2.0$	0.07	0.08	0.14	0.25	2.93	1.46	5.91	1.57
Kernel $b = 3.0$	0.13	0.03	0.38	0.51	8.30	4.11	16.7	4.41
Mixture $K = \hat{K}$	0.04	0.12	0.05	0.25	0.32	0.18	0.64	0.20
Mixture $K = 1$	0.13	0.53	0.07	0.41	0.18	0.10	0.37	0.10
Mixture $K = 3$	0.04	0.11	0.05	0.16	0.59	0.32	1.03	0.32
Mixture $K = 5$	0.03	0.09	0.05	0.15	0.68	0.38	1.33	0.39
NP MLE	0.03	0.16	0.04	0.16	0.36	0.30	0.70	0.31
Plug-in Predictor, $\hat{\theta}_{QMLE}$	0.42	0.45	2.73	6.14	61.6	30.7	123	32.9
Loss-Function Based	0.53	1.21	3.18	3.18	64.9	8.47	129	8.68
Pooled OLS	1.84	0.73	0.17	0.84	0.14	0.07	0.27	0.07
First Differences	4.67	4.46	13.8	18.4	249	74.6	496	79.6

Notes: The design of the experiment is summarized in Table 1. The regret is standardized by the average posterior variance of λ_i and we set $\epsilon_0 = 0.1$; see Theorem 3.7. For the oracle predictor we report the compound risk (in parentheses) instead of the regret.

cut-off the top 5% group. Because the cut-offs are computed from the population distribution of Y_{iT}, for unit i the selection rules only depends on Y_{iT} and not on Y_{jT} with $j \neq i$. The corresponding columns are labeled "Top."

We report results for different choices of $\mathbb{V}[\lambda_i]$, starting from $\mathbb{V}[\lambda_i] = 1$ on the left and ending with $\mathbb{V}[\lambda_i] = 0$ ($\lambda_i = \lambda$ for all i). When the variance of λ_i is reduced to zero, the oracle risk decreases because there is less uncertainty. At the same time, the relative regret increases because the numerator in the relative regret in Theorem 3.7 drops from approximately 200 for $\mathbb{V}[\lambda_i] = 1$ to approximately 2 for $\mathbb{V}[\lambda_i] = 0$. As expected from the theoretical analysis, if the λ_i's exhibit (strong) heterogeneity, the empirical Bayes predictors attain a lower regret than the alternative predictors. The plug-in predictor, the loss-function based predictor, and the first-differences predictor are consistently dominated by the best empirical Bayes predictors. Not surprisingly, the pooled OLS predictor beats the best empirical Bayes predictor when the λ_i's are essentially homogeneous, i.e., $\mathbb{V}[\lambda_i] = .002$ and $\mathbb{V}[\lambda_i] = 0$. The estimate of ρ obtained under pooled OLS is generally larger (closer to unity) than $\hat{\rho}_{QMLE}$. If we condition the pooled OLS estimate of λ on $\hat{\rho}_{QMLE}$, the regret increases noticeably.[9]

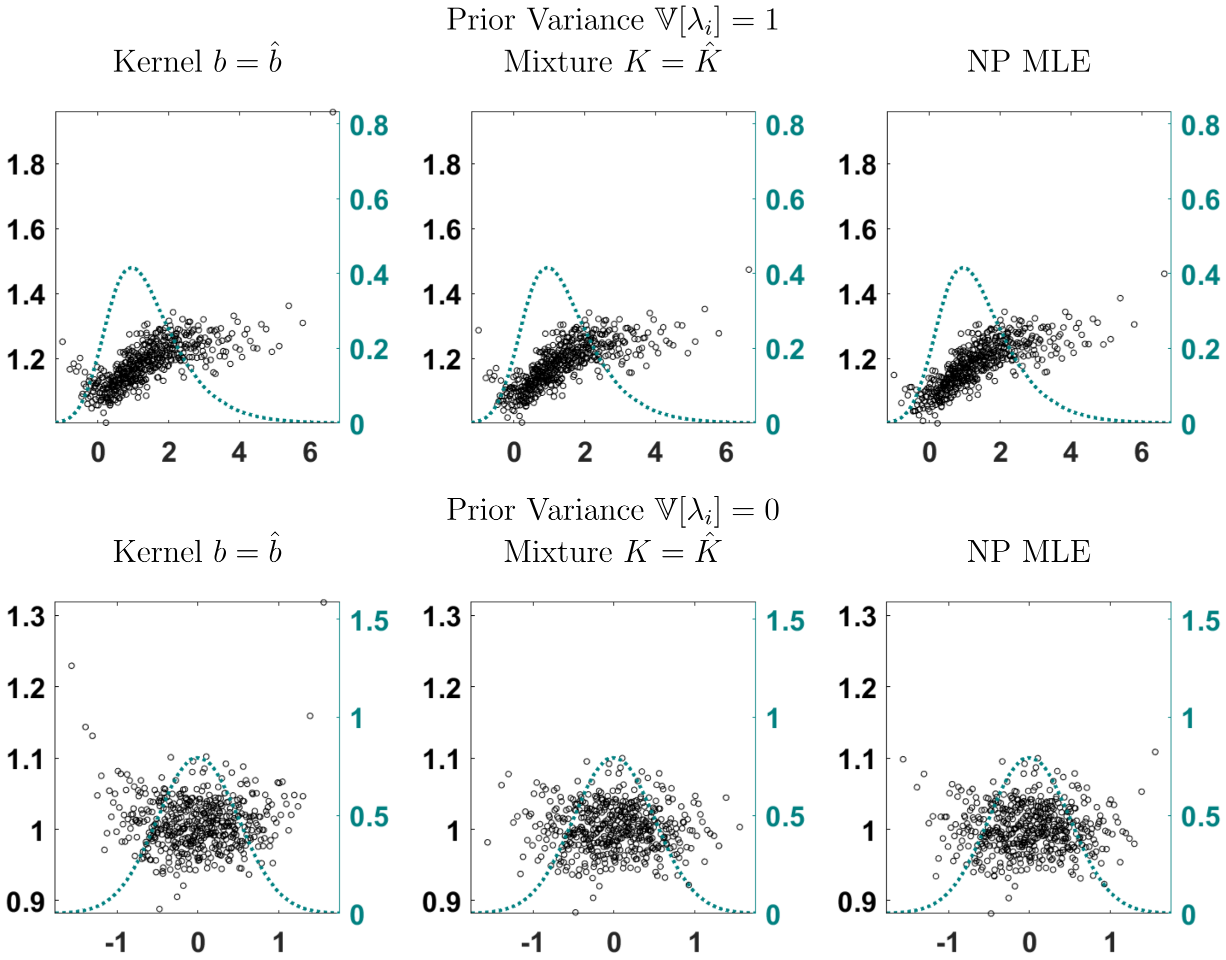
The empirical Bayes estimators in Table 2 differ with respect to the estimation of the density required to calculate the Tweedie correction. We begin with the fixed-bandwidth kernel estimators, which are covered by our theoretical results. If the squared forecast error losses are averaged across all units, then a scaling of B_N^* defined in (25) by approximately 1.5 minimizes the relative regret *ex post*. The *ex ante* selection of this scaling constant based on the pseudo-out-of-sample forecast of Y_{iT} works very well. We determine $\hat{b}$ by minimizing the regret over the interval $[1,3]$ using a grid-spacing of 0.1. The regret obtained by setting $b = \hat{b}$ is very close to the regret with the *ex-post* optimal b. If the expected loss is computed for the top-5% group, then it is preferable to use a larger bandwidth scaling than $b = 1.5$, because it provides a better estimate of the thin right tail of the density $p(\hat{\lambda}_i)$. Note that we could have re-computed $\hat{b}$ by minimizing the time T prediction risk for the top-5% group, but we did not do so for this table.

The mixture approximation of $p(\hat{\lambda}_i)$ leads to a Tweedie correction that tends to generate more accurate forecasts than the kernel-based correction, in particular for small $\mathbb{V}[\lambda_i]$. The selection of the number of mixture components, K, based on the period T forecast errors also works well. For large values of $\mathbb{V}[\lambda_i]$ when the distribution of $\hat{\lambda}_i$ has a long right tail, we select a large value of K, whereas for small values $\mathbb{V}[\lambda_i]$ when the distribution of $\hat{\lambda}_i$ is approximately normal, we select a small value of K. Finally, the nonparametric MLE of $p(\hat{\lambda}_i)$ proposed by Gu and Koenker (2016) performs about as well as the mixture estimator in this experiment.

To shed some more light on the performance differentials among the various Tweedie corrections, we plot forecast errors as a function of $\hat{\lambda}_i$ in Figure 1 and overlay the density of $\hat{\lambda}_i$. The plots are generated as follows. We have a total of 10^6 forecasts across the 1,000 repetitions of the Monte Carlo experiment. We group the $\hat{\lambda}_i$'s into 500 bins such that each bin contains 2,000 observations. For each bin, we compute the average squared forecast error, which leads to 500 pairs of bin location and forecast performance. Unlike Table 2 which focuses on regrets, the figure reports mean-squared errors. While the regret differentials, say, between the kernel-based correction versus the mixture-based on NP-MLE correction appear to be large, overall, in terms of MSE, the forecast performance of the various estimators is very similar. For instance, for $\mathbb{V}[\lambda_i] = 1$ the average MSEs for kernel $b = \hat{b}$, mixture $K = \hat{K}$, and NP MLE are 1.185, 1.184, and 1.183, respectively.

[9] For the values of $\mathbb{V}[\lambda_i]$ reported in Table 2 the relative regrets for "All" units are 4.6, 0.49, 0.24, and 0.42, respectively.

Figure 1: Squared Forecast Error Loss as a Function of $\hat{\lambda}_i$



Prior Variance $\mathbb{V}[\lambda_i] = 1$

Kernel $b = \hat{b}$ Mixture $K = \hat{K}$ NP MLE

Prior Variance $\mathbb{V}[\lambda_i] = 0$

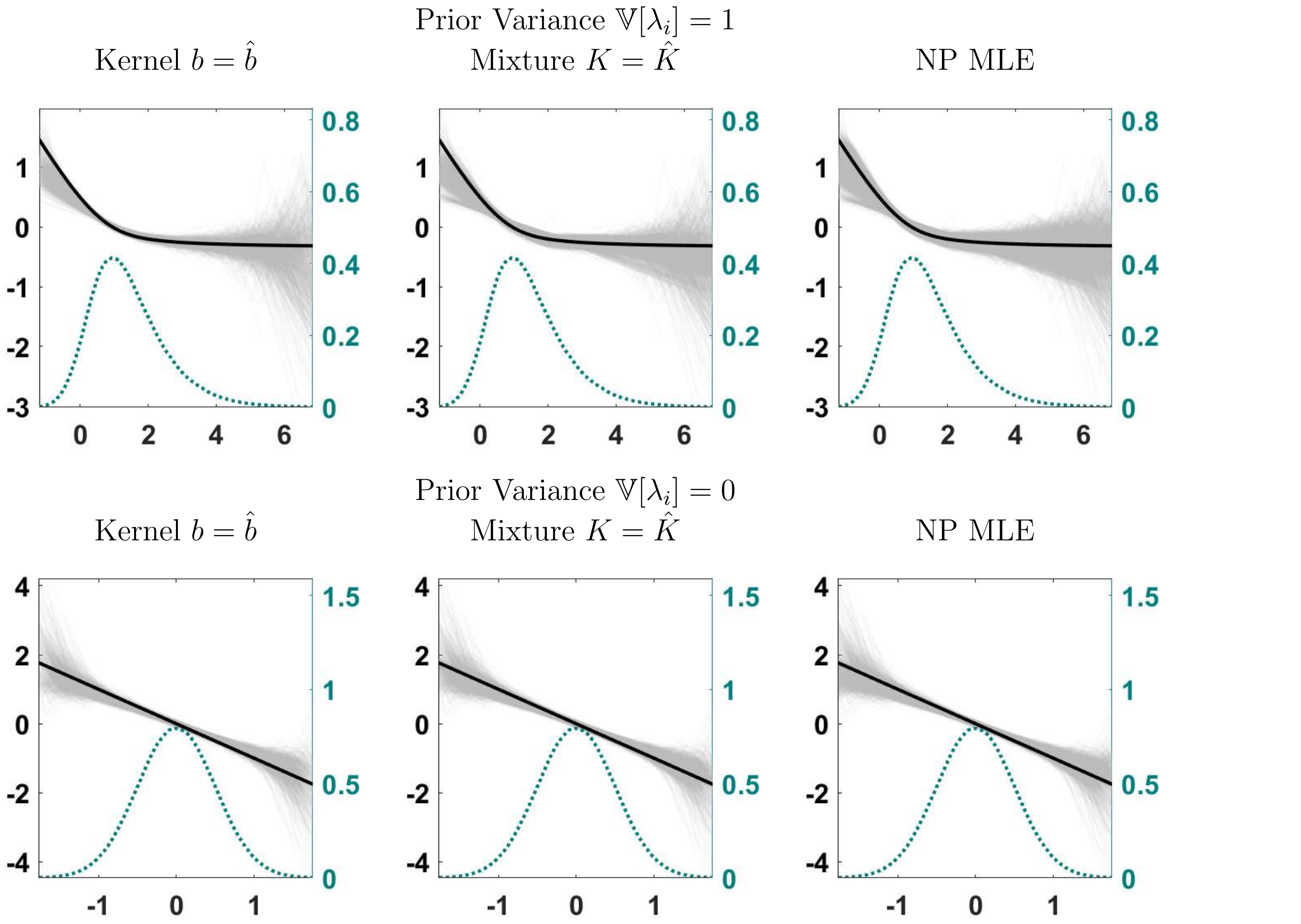
Kernel $b = \hat{b}$ Mixture $K = \hat{K}$ NP MLE

Notes: Each dot corresponds to a $\hat{\lambda}_i$ bin and the average squared-forecast error for observations assigned to that bin (scale on the left). The panels also show density estimates of the empirical distribution of $\hat{\lambda}_i(\rho)$ based on $N_{sim} \cdot N = 10^6$ simulations of $\hat{\lambda}_i(\rho)$ (scale on the right). The Monte Carlo design is described in Table 1.

For $\mathbb{V}[\lambda_i] = 1$, the further $\hat{\lambda}$ is in the tails of its distribution, the larger the MSEs. This is the result of two effects. First, the Tweedie correction is less precisely estimated in the tails, because there are fewer realizations of $\hat{\lambda}_i$. Second, as we will show below, the population Tweedie correction in the right tail of the $\hat{\lambda}_i$ distribution is essentially flat. Thus, there is less shrinkage and the posterior mean has a higher sampling variance. The visual inspection of the plots indicates that differences between MSEs arise mostly in the tails of the $\hat{\lambda}_i$ distribution. Here, the kernel-based methods produce less accurate density estimates, which translate into a few MSE outliers and thereby slightly larger average prediction errors.

In Figure 2 we plot the Tweedie corrections $(\sigma^2/T)\partial \ln p(\hat{\lambda}_i|\theta)/\partial\hat{\lambda}_i$ for the oracle forecast

Figure 2: Tweedie Corrections



Prior Variance $\mathbb{V}[\lambda_i] = 1$

Kernel $b = \hat{b}$ Mixture $K = \hat{K}$ NP MLE

Prior Variance $\mathbb{V}[\lambda_i] = 0$

Kernel $b = \hat{b}$ Mixture $K = \hat{K}$ NP MLE

Notes: Solid (black) lines depict oracle Tweedie correction based on $p(\hat{\lambda}_i|y_{i0}, \theta)$ (scale on the left). Grey "hairs" depict estimates from the Monte Carlo repetitions. The panels also show density estimates of the empirical distribution of $\hat{\lambda}_i(\rho)$ based on $N_{sim} \cdot N = 10^6$ simulations of $\hat{\lambda}_i(\rho)$ (scale on the right). The Monte Carlo design is described in Table 1.

and the various empirical Bayes predictors. Each hairline corresponds to an estimate from a particular Monte Carlo repetition. We again overlay the density of $\hat{\lambda}_i$ to indicate the likelihood of the various $\hat{\lambda}_i$ values on the x-axis. For $\mathbb{V}[\lambda_i] = 1$ the oracle correction is L-shaped. In the left tail of the $\hat{\lambda}_i$ distribution there is a lot of shrinkage to the prior mean and the correction is approximately linear with a large slop. In the right tail, the correction is essentially flat, meaning that for large values of $\hat{\lambda}_i$ (outliers) the optimal shrinkage is small in relative terms. For $\mathbb{V}[\lambda_i] = 0$ the $p(\hat{\lambda}_i|\theta)$ density is Normal and the oracle Tweedie correction is linear.

For $\mathbb{V}[\lambda_i] = 1$ all estimators do a fairly good job in approximating the optimal correction for values $\hat{\lambda}_i < 5$, i.e., in the center of the $\hat{\lambda}_i$ distribution. In the right tail for $\hat{\lambda}_i > 5$, however, the kernel-based correction appears to be fairly unstable and highly variable across Monte Carlo repetitions. If the λ_i's are homogeneous and the distribution of $\hat{\lambda}_i$ is Normal, then all procedures generate a good approximation of the optimal correction for $-1 \leq \hat{\lambda}_i \leq 1$. Unfortunately, outside of this range the estimates of the Tweedie correction become less accurate. Our simulations suggest that forecasting "winners" and "losers" remains difficult. While the Bayes correction induces shrinkage that off-sets the selection bias inherent in $\hat{\lambda}_i$, i.e., $\hat{\lambda}_i$ over-estimates (under-estimates) the "true" λ_i for λ_i's in the top (bottom) quantile of the cross-sectional distribution, estimating the optimal amount of this shrinkage is difficult because the density estimates may be based on a small number of tail observations.

4.2 Experiment 2: Non-Gaussian Correlated Random Effects

We now change the Monte Carlo design by replacing the Gaussian random effects specification with a non-Gaussian specification in which the heterogeneous coefficient λ_i is correlated with the initial condition Y_{i0}. The Monte Carlo design is summarized in Table 3. Starting point is a joint normal distribution for (λ_i, Y_{i0}), factorized into a marginal distribution $\pi_*(\lambda_i)$ and a conditional distribution $\pi_*(Y_{i0}|\lambda_i)$. According to this joint normal distribution $\lambda_i \sim N(\underline{\mu}_\lambda, \underline{V}_\lambda)$ and $Y_{i0}|\lambda_i$ corresponds to the stationary distribution of Y_{it} associated with its autoregressive law of motion. The implied marginal distribution for Y_{i0} is used as $\pi(Y_{i0})$ in the Monte Carlo design, whereas we replace $\pi_*(\lambda_i|Y_{i0})$ by a mixture $\pi(\lambda_i|Y_{i0})$, indexed by a parameter δ. For $\delta = 0$ the mixture reduces to $\pi_*(\lambda_i|Y_{i0})$, whereas for large values of δ the density becomes multi-modal and looks like a pair of scissors.

The risks associated with the oracle predictions and the relative regrets of the feasible predictors are summarized in Table 4, which has the same format as Table 2 with two exceptions. First, we dropped the nonparametric MLE, because its performance was similar to the mixture estimator and in the software provided by Gu and Koenker (2016) it is implemented for a random-effects but not for a correlated random-effects second. Second, we changed the domain of the scaling constant b of the kernel bandwidth to the interval [0.1,1.9] with a spacing of 0.1.

The results are qualitatively similar to Experiment 1. The plug-in predictors, the predictors obtained from the loss-function-based estimator, and the first difference predictors are clearly dominated by the empirical Bayes predictors. The empirical Bayes predictors also beat the pooled OLS predictor by a significant margin for $\delta = 0.1$ and $\delta = 0.3$. At first glance

Table 3: Monte Carlo Design 2

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$ where $U_{it} \sim iidN(0, \gamma^2)$; $\rho = 0.8$, $\gamma = 1$

Initial Observation: $Y_{i0} \sim N\left(\frac{\mu_\lambda}{1-\rho}, V_Y + \frac{V_\lambda}{(1-\rho)^2}\right)$, $V_Y = \gamma^2/(1-\rho^2)$; $\underline{\mu}_\lambda = 1$, $\underline{V}_\lambda = 1$

Correlated Random Effects:

$$\lambda_i | Y_{i0} \sim \left\{ \begin{array}{ll} N\big(\phi_+(Y_{i0}), \underline{\Omega}\big) & \text{with probability } p_\lambda \\ N\big(\phi_-(Y_{i0}), \underline{\Omega}\big) & \text{with probability } 1 - p_\lambda \end{array} \right. ,$$

$$\phi_+(Y_{i0}) = \phi_0 + \delta + (\phi_1 + \delta) Y_{i0},$$

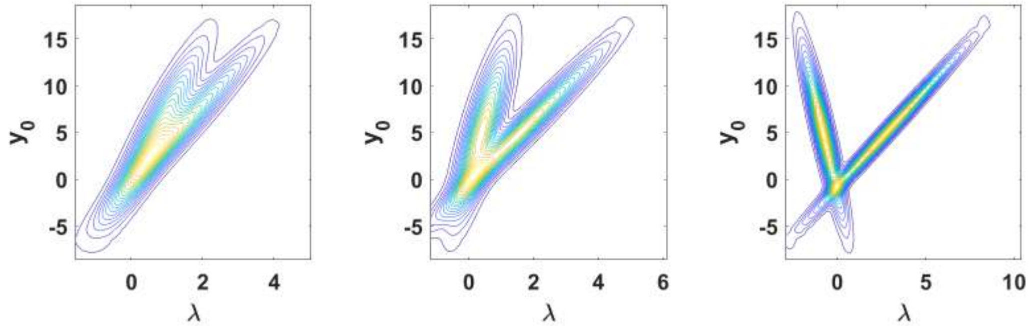
$$\phi_-(Y_{i0}) = \phi_0 - \delta + (\phi_1 - \delta) Y_{i0},$$

$$\underline{\Omega} = \left[\frac{1}{(1-\rho)^2} V_Y^{-1} + \underline{V}_\lambda^{-1}\right]^{-1}, \ \phi_0 = \underline{\Omega}\, \underline{V}_\lambda^{-1} \underline{\mu}_\lambda, \ \phi_1 = \frac{1}{1-\rho} \underline{\Omega} V_Y^{-1},$$

$$p_\lambda = 1/2, \ \delta \in \{0.05, \, 0.1, \, 0.3\}$$

Sample Size: $N = 1,000$, $T = 4$

Number of Monte Carlo Repetitions: $N_{sim} = 1,000$



The plots depict contours of the density $\pi(\lambda_i, Y_{i0})$ for $\delta = 0.05$, $\delta = 0.1$, and $\delta = 0.3$.

the good performance of pooled OLS for $\delta = 0.05$ is surprising in view of the heterogeneity in λ_i implied by the Monte Carlo design. It turns out that under pooled OLS $\hat{\lambda}$ is close to zero and $\hat{\rho}$ is approximately one. Thus, the predictor essentially generates no-change forecasts that perform quite well.

For $\delta = 0.05$ and $\delta = 0.1$ the optimal bandwidth scaling b is close to one, which is picked up by our bandwidth selection procedure based on pseudo-out-of-sample forecasts for period $t = T$. For $\delta = 0.3$ the optimal scaling is approximately 0.5. This is qualitatively plausible, because the scissor-shape density of $(\hat{\lambda}_i, Y_{i0})$ has a relatively large overall variance, which translates into a large baseline bandwidth, but at the same time it has sharp peaks in the modal regions which require a small bandwidth. As in Experiment 1, kernel estimation of the Tweedie correction does generally not work as well as the mixture estimation, in part because the latter is more stable in the tails of the $(\hat{\lambda}_i, Y_{i0})$ distribution and in part because the density of $p(\hat{\lambda}_i, Y_{i0})$ is in fact a mixture of normal distributions.

Table 4: Experiment 2: Relative Regrets under Correlated Random Effects

	$\delta = 0.05$		$\delta = 0.1$		$\delta = 0.3$	
Predictor	All	Top	All	Top	All	Top
Oracle Predictor	(1110)	(55.6)	(1122)	(53.1)	(1093)	(52.9)
Empirical Bayes, $\hat{\theta}_{QMLE}$						
Kernel $b = \hat{b}$	0.22	0.44	0.22	0.78	0.43	0.94
Kernel $b = 0.5$	1.46	6.94	0.78	5.41	0.42	1.10
Kernel $b = 1.0$	0.22	0.42	0.25	0.74	0.93	1.15
Kernel $b = 1.5$	0.37	0.35	0.43	0.97	1.25	1.44
Mixture $K = \hat{K}$	0.06	0.16	0.05	0.16	0.06	0.10
Mixture $K = 1$	0.15	0.56	0.52	1.52	1.53	1.64
Mixture $K = 3$	0.06	0.15	0.06	0.18	0.40	0.08
Mixture $K = 5$	0.25	0.39	0.54	1.66	0.46	0.15
Plug-in Predictor, $\hat{\theta}_{QMLE}$	1.28	1.13	1.03	2.05	1.63	1.91
Loss-Function Based	1.50	1.93	1.33	4.00	1.78	2.48
Pooled OLS	0.20	0.59	0.78	2.73	4.98	3.37
First Differences	7.92	5.99	7.04	9.47	9.48	9.39

Notes: The design of the experiment is summarized in Table 3. The regret is standardized by the average posterior variance of λ_i and we set $\epsilon_0 = 0.1$; see Theorem 3.7. For the oracle predictor we report the compound risk (in parentheses) instead of the regret.

4.3 Experiment 3: Misspecified Likelihood Function

In the third experiment, summarized in Table 5, we consider a misspecification of the Gaussian likelihood function by replacing the Normal distribution in the DGP with two mixtures. We consider a scale mixture that generates excess kurtosis and a location mixture that generates skewness. The innovation distributions are normalized such that $\mathbb{E}[U_{it}] = 0$ and $\mathbb{V}[U_{it}] = 1$. For (λ_i, Y_{i0}) we use the correlated random effects distribution of Experiment 2 with $\delta = 0.1$.

The oracle risks and the relative regrets are summarized in Table 6. Columns 2 and 3 refer to the case of normally distributed innovations and reproduce columns 4 and 5 of Table 4. The remaining columns contain the results for scale and location mixture innovations. The QMLE estimator of θ remains consistent under the likelihood misspecification. However, the (non-parametric) Tweedie correction no longer delivers a valid approximation of the posterior mean. Accordingly, the regrets under mixture innovations are generally higher than under the normal innovations. However, in comparison to the other four predictors (plug-in, loss-function based, pooled OLS, and first differences) the empirical Bayes predictors continue to perform very well and attain relative regrets that are more than 50% smaller than the regrets

Table 5: Monte Carlo Design 3

Law of Motion: $Y_{it} = \lambda_i + \rho Y_{it-1} + U_{it}$, $\rho = 0.8$, $\mathbb{E}[U_{it}] = 0$, $\mathbb{V}[U_{it}] = 1$

Scale Mixture: $U_{it} \sim iid \begin{cases} N(0, \gamma_+^2) & \text{with probability } p_u \\ N(0, \gamma_-^2) & \text{with probability } 1 - p_u \end{cases},$

$\gamma_+^2 = 4$, $\gamma_-^2 = 1/4$, $p_u = (1 - \gamma_-^2)/(\gamma_+^2 - \gamma_-^2) = 1/5$

Location Mixture: $U_{it} \sim iid \begin{cases} N(\mu_+, \gamma^2) & \text{with probability } p_u \\ N(\mu_-, \gamma^2) & \text{with probability } 1 - p_u \end{cases},$

$\mu_- = -1/4$, $\mu_+ = 2$, $p_u = |\mu_-|/(|\mu_-| + \mu_u^+) = 1/9$,

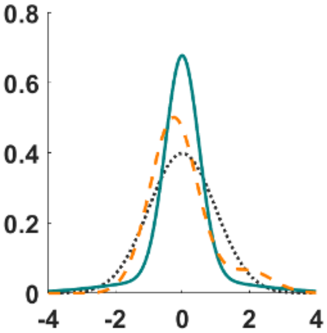
$\gamma^2 = 1 - p_u \mu_+^2 - (1 - p_u)\mu_-^2 = 1/2$

Initial Observations: same as Design 2

Correlated Random Effects: same as Design 2 ($\delta = 0.1$)

Sample Size: $N = 1,000$, $T = 4$

Number of Monte Carlo Repetitions: $N_{sim} = 1,000$



The plot overlays a $N(0,1)$ density (black, dotted), the scale mixture (teal, solid), and the location mixture (orange, dashed).

associated with the best competitors. As in Experiments 1 and 2, the kernel approximation of the Tweedie correction tends to perform not quite as well as the mixture approximation.

5 Extensions

In this section we discuss three extensions: multi-step forecasts (Section 5.1), Tweedie's formula for the general linear model in (1) (Section 5.2), and the identifiability of the correlated random effects distribution $\pi(\cdot)$ in the general linear model.

5.1 Multi-Step Forecasting

While this paper focuses on single-step forecasting, we briefly discuss in the context of the basic dynamic panel data model how the framework can be extended to multi-step forecasts.

Table 6: Experiment 3: Relative Regrets under Misspecified Likelihood Function

	Normal		Scale Mixture		Location Mixture	
Predictor	All	Top	All	Top	All	Top
Oracle Predictor	(1121)	(53.1)	(1075.6)	(48.9)	(1101)	(45.6)
Empirical Bayes, $\hat{\theta}_{QMLE}$						
Kernel $b = \hat{b}$	0.22	0.78	1.34	1.43	0.67	1.22
Kernel $b = 0.5$	0.78	5.41	2.76	8.47	1.52	7.04
Kernel $b = 1.0$	0.25	0.74	1.37	1.37	0.71	1.15
Kernel $b = 1.5$	0.43	0.97	1.77	1.72	0.98	1.45
Mixture $K = \hat{K}$	0.05	0.16	1.00	0.50	0.40	0.30
Mixture $K = 1$	0.52	1.52	1.99	2.43	1.10	1.96
Mixture $K = 3$	0.06	0.18	1.15	0.72	0.42	0.31
Mixture $K = 5$	0.54	1.66	2.46	1.81	0.59	0.72
Plug-in Predictor, $\hat{\theta}_{QMLE}$	1.03	2.05	3.16	3.35	1.92	2.87
Loss-Function Based	1.33	4.00	3.79	5.67	2.36	4.69
Pooled OLS	0.78	2.73	2.56	4.02	1.50	3.24
First Differences	7.04	9.47	16.8	15.6	11.4	13.7

Notes: The design of the experiment is summarized in Table 5. The regret is standardized by the average posterior variance of λ_i and we set $\epsilon_0 = 0.1$; Theorem 3.7. For the oracle predictor we report the compound risk (in parentheses) instead of the regret.

We can express

$$Y_{iT+h} = \left(\sum_{s=0}^{h-1} \rho^s \right) \lambda_i + \rho^h Y_{iT} + \sum_{s=0}^{h-1} \rho^s U_{iT+h-s}.$$

Under the assumption that the oracle knows ρ and $\pi(\lambda_i, Y_{i0})$ we can express the oracle forecast as

$$\widehat{Y}^{opt}_{iT+h} = \left(\sum_{s=0}^{h-1} \rho^s \right) \mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i] + \rho^h Y_{iT}.$$

As in the case of the one-step-ahead forecasts, the posterior mean $\mathbb{E}^{\lambda_i}_{\theta,\mathcal{Y}_i}[\lambda_i]$ can be replaced by an approximation based on Tweedie's formula and the ρ's can be replaced by consistent estimates. A model with additional covariates would require external multi-step forecasts of the covariates, or the specification in (1) would have to be modified such that all exogenous regressors appear with an h-period lag.

5.2 Tweedie's Formula (Generalization)

The general model (1) distinguishes three types of regressors. First, the $k_w \times 1$ vector W_{it} interacts with the heterogeneous coefficients λ_i. In addition to a constant, we allow W_{it}

to also include deterministic time effects such as seasonality, time trends and/or strictly exogenous variables observed at time t. To distinguish deterministic time effects $w_{1,t+1}$ from cross-sectionally varying and strictly exogenous variables $W_{2,it}$, we partition the vector into $W_{it} = (w_{1,t+1}, W_{2,it})$.[10] The dimensions of the two components are k_{w_1} and k_{w_2}, respectively. Second, X_{it} is a $k_x \times 1$ vector of predetermined predictors with homogeneous coefficients. Because the predictors X_{it} may include lags of Y_{it+1}, we collect all the predetermined variables other than the lagged dependent variable into the subvector $X_{2,it}$. Third, Z_{it} is a k_z-vector of strictly exogenous regressors, also with common coefficients.

Collect the exogenous conditioning variables in $H_i = (X_{i0}, W_{2,i}^{0:T}, Z_i^{0:T})$. To introduce heteroskedasticity, we allow the error terms to conditionally heteroskedastic in the cross section and across time:

$$U_{it} = \sigma_t V_{it} = \varsigma(H_i, \gamma_t) V_{it}, \quad V_{it} \mid (Y_i^{1:t-1}, X_i^{1:t-1}, H_i, \lambda_i) \sim N(0,1), \tag{29}$$

where $\varsigma(\cdot)$ is a parametric function indexed by the (time-varying) finite-dimensional parameter γ_t. We allow $\varsigma(\cdot)$ to be dependent on the initial condition of the predetermined predictors, X_{i0}, and other exogenous variables. Because the time dimension T is assumed to be small, the dependence through X_{i0} can generate a persistent ARCH effect. We stack the γ_t's into the vector $\gamma = [\gamma_1', \ldots, \gamma_T']$. Note that even in the homoskedastic case $\sigma_t = \sigma$, the distribution of Y_{it} given the regressors is non-normal because the distribution of the λ_i parameters is fully flexible.

Let $\theta = [\alpha', \rho', \gamma']'$, $\tilde{y}_t(\theta) = y_t - \rho' x_{t-1} - \alpha' z_{t-1}$, and $\Sigma(\theta, h) = \mathrm{diag}\big(\sigma_1^2, \ldots, \sigma_T^2\big)$ (the i subscripts are dropped to simplify the notation). Moreover, let $\tilde{y}(\theta)$ and w be the matrices with rows $\tilde{y}_t(\theta)$ and w_{t-1}', $t = 1, ..., T$. To generalize Tweedie's formula, we re-define the sufficient statistic $\hat{\lambda}$ as follows:

$$\hat{\lambda}(\theta) = \big(w'\Sigma^{-1}(\theta, h) w\big)^{-1} w'\Sigma^{-1}(\theta, h)\tilde{y}(\theta). \tag{30}$$

Using the same calculations as in Section 2.3, it can be shown that the posterior mean of λ_i has the representation

$$\mathbb{E}^{\lambda}_{\theta,\pi,\mathcal{Y}}[\lambda] = \hat{\lambda}(\theta) + \left(W^{0:T-1'}\Sigma^{-1}(\theta, H) W^{0:T-1}\right)^{-1} \frac{\partial}{\partial \hat{\lambda}(\cdot)} \ln p\big(\hat{\lambda}(\cdot), H|\theta\big). \tag{31}$$

[10] Because W_{it} is a predictor for Y_{it+1} we use a $t+1$ subscript for the deterministic trend component w_1.

The optimal forecast is given by

$$\widehat{Y}^{opt}_{iT+1}(\theta) = \left(\mathbb{E}^{\lambda_i}_{\theta,\pi,\mathcal{Y}_i}[\lambda_i]\right)' W_{iT+1} + \rho' X_{iT} + \alpha' Z_{iT}. \tag{32}$$

The existing panel data model literature has developed numerous estimators for the homogeneous parameters that could be used to generate a $\hat{\theta}$ for the general model (1). In principle one can proceed with the estimation of the Tweedie correction as for the basic model. However, the larger the set of conditioning variables H_i the more difficult it becomes to estimate $p(\hat{\lambda}_i(\theta, H_i), H_i|\theta)$ precisely.

5.3 Identification

Our forecasts rely on an implicit estimation of the correlated random effects distribution $\pi(\cdot)$ to obtain approximations of the unit-specific posterior distributions. Thus, an interesting and important question is whether this distribution is in fact identifiable based on the information contained in a panel data set with fixed time dimension T and increasing cross-sectional dimension $N \longrightarrow \infty$. For the special case of point forecasting under a quadratic loss function considered in this paper, there is a short-cut to the identification argument. Provided that the vector of homogeneous parameters θ is identifiable, it is possible to identify the cross-sectional distribution of the sufficient statistic $\hat{\lambda}(\theta)$. This, through Tweedie's formula, also identifies the posterior mean $\mathbb{E}^{\lambda}_{\theta,\pi,\mathcal{Y}}[\lambda]$ and leads to a unique Bayes predictor. In the remainder of this section, show that under the subsequent assumptions both θ and $\pi(\cdot)$ are identifiable from the panel data set. As a consequence, not just the posterior mean, but the entire posterior distribution of λ is identifiable. This result is useful, for instance, for density forecasting applications as in Liu (2018).

Assumption 5.1

(i) $(Y_i^{1:T+1}, X_i^{1:T}, H_i, \lambda_i)$ *are independent across* i.

(ii) (λ_i, H_i) *are iid with joint density*

$$\pi(\lambda, h) = \pi(\lambda|h)\pi(h).$$

(iii) For $t = 1, \ldots, T$, *the distribution of* $X_{2,it}$ *conditional on* $(Y_i^{1:t}, X_i^{1:t-1}, H_i)$ *does not depend on the heterogeneous parameters* λ_i *and the homogeneous parameter* θ.

(iv) The marginal distribution of $(W_{2,i}^{0:T}, Z_i^{0:T})$ *does not depend on* θ.

(v) $U_{it} = \varsigma(H_i, \gamma_t)V_{it}$*, where* V_{it} *has density* $\varphi(v)$ *and is iid across* i *and* t *with* $\mathbb{E}[V_{it}] = 0$ *and* $\mathbb{V}[V_{it}] = 1$ *for* $t = 1, \ldots, T+1$*. The vector* $(V_{i1}, \ldots, V_{iT})$ *is independent of* H_i*.* γ_t *is an unknown finite-dimensional parameter vector* γ_t*.*

We dropped the deterministic trend regressors $w_{1,t}$ from the notation for now. Assumption 5.1(i) states that conditionally on the predictors, the Y_{it+1}s are cross-sectionally independent. Thus, we assume that all the spatial correlation in the dependent variables is due to the observed predictors. Assumption 5.1(ii) formalizes the correlated random effects assumption. The subsequent Assumptions 5.1(iii) and (iv) imply that λ_i may affect X_{it} only indirectly through $Y_i^{1:t}$ – an assumption that is clearly satisfied in the dynamic panel data model (2) – and that the strictly exogenous predictors do not depend on θ. Assumption 5.1(v) reproduces (29), but without restricting V_{it} to be normally distributed which is not necessary to obtain the identification result.

The identification can be established in three steps. First, the identification of the homogeneous regression coefficients (ρ, α) follows from a standard argument used in the instrumental variable (IV) estimation of dynamic panel data models, e.g., Arellano and Bover (1995). Second, the variance parameters γ can be identified from a moment condition that is obtained after projecting $Y_i - X_i\rho - Z_i\alpha$ onto W_i. The identification of $\pi(\lambda_i|h_i)$ can be established using a characteristic function argument similar to that in Arellano and Bonhomme (2012). For the general model (1) we make the following additional assumptions:

Assumption 5.2

(i) The parameter vectors α *and* ρ *are identified for fixed* T *from the cross-sectional distribution of the observables.*

(ii) For each $t = 1, \ldots, T$ *and almost all* h_i $\varsigma^2(h_i, \tilde{\gamma}_t) = \varsigma^2(h_i, \gamma_t)$ *implies* $\tilde{\gamma}_t = \gamma_t$*. Moreover,* $\varsigma^2(h_i, \gamma_t) > 0$*.*

(iii) The characteristic functions for $\lambda_i|(H_i = h_i)$ *and* V_i *are non-vanishing almost everywhere.*

(iv) $W_i = [W_{i0}, ..., W_{iT-1}]'$ *has full rank* k_w*.*

Because the identification of α and ρ in panel data models with fixed or random effects is well established, we make the high-level Assumption 5.2(i) that the homogeneous parameters are identifiable.[11] Assumption 5.2(ii) enables us to identify the volatility parameters γ, and (iii) and (iv) deliver the identifiability of the distribution of heterogeneous coefficients. The following theorem proved in the Online Appendix summarizes the identification result.

[11] Textbook / handbook chapter treatments can be found in, for instance, Baltagi (1995), Arellano and Honoré (2001), Arellano (2003) and Hsiao (2014).

Theorem 5.3 *Suppose that Assumptions 5.1 and 5.2 are satisfied. Then the parameters α, ρ, and γ as well as the correlated random effects distribution $\pi(\lambda_i|h_i)$ and the distribution of V_{it} in model (1) are identified for fixed T from the cross-sectional distribution of the observables.*

6 Empirical Application

We will now use the previously-developed predictors to forecast PPNRs of bank holding companies (BHC). The stress tests that have become mandatory under the 2010 Dodd-Frank Act require banks to establish how PPNRs vary in stressed macroeconomic and financial scenarios. A first step toward building and estimating models that provide trustworthy projections of PPNRs and other bank-balance-sheet variables under hypothetical stress scenarios, is to develop models that generate reliable forecasts under the observed macroeconomic and financial conditions. Because of changes in the regulatory environment in the aftermath of the financial crisis as well as frequent mergers in the banking industry our large N small T panel-data-forecasting framework is particularly attractive for stress-test applications.

We generate a collection of panel data sets in which PPNR as a fraction of consolidated assets (the ratio is scaled by 400 to obtain annualized percentages) is the dependent variable. The data sets are based on the FR Y-9C consolidated financial statements for bank holding companies for the years 2002 to 2014, which are available through the website of the Federal Reserve Bank of Chicago. The time period t in our analysis is one quarter.

We construct rolling samples that consist of $T + 2$ observations, where T is the size of the estimation sample and varies between $T = 4$ and $T = 10$ quarters. The additional two observations in each rolling sample are used, respectively, to initialize the lag in the first period of the estimation sample and to compute the error of the one-step-ahead forecast. For instance, with data from 2002:Q1 to 2014:Q4 we can construct $M = 45$ samples of size $T = 6$ with forecast origins running from $\tau =$ 2003:Q3 to $\tau =$ 2014:Q3. Each rolling sample is indexed by the pair (τ, T). The cross-sectional dimension N varies from sample to sample and ranges from 613 to 920. Further details about the data as well as a description of our procedure to create balanced panels and eliminate outliers are provided in the Appendix. We discuss the accuracy of baseline forecasts for various model specifications and predictors in Section 6.1 and compare the baseline predictions to predictions under stressed macroeconomic and financial conditions in Section 6.2.

6.1 Baseline Forecast Results

The forecast evaluation criterion is the mean-squared error (MSE) computed across institutions:

$$\mathrm{MSE}(\widehat{Y}^N_{\tau+1}) = \frac{\frac{1}{N_\tau}\sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau})\big(Y_{i\tau+1} - \widehat{Y}_{i\tau+1}\big)^2}{\frac{1}{N_\tau}\sum_{i=1}^{N_\tau} D_i(\mathcal{Y}_{i\tau})}. \tag{33}$$

For the empirical analysis we consider four predictors in total. The first two predictors are empirical Bayes predictors based on $\hat{\theta}_{QMLE}$. The Tweedie corrections are generated either using a kernel estimator with $b = \hat{b}$ or a mixture estimator with $K = \hat{K}$. The third predictor is the plug-in predictor that, conditional on $\hat{\theta}_{QMLE}$ estimates the heterogeneous coefficients for each unit separately, without using prior information. The fourth predictor assumes that all coefficients are homogeneous and is based on pooled OLS. The predictors were described in detail in Section 4.

We consider three model specifications. The first model is the basic dynamic panel data model in (2) that was also used in the Monte Carlo experiments in Section 4. The second and third specification include additional covariates that reflect aggregate macroeconomic and financial conditions, which we will use subsequently to generate counterfactual forecasts under stress scenarios. We assume that the banks' exposure to the aggregate condition is heterogeneous and include these predictors into the vector W_{it-1}, using the notation in (1). When analyzing stress scenarios, one is typically interested in the effect of stressed economic conditions on the current performance of the banking sector. For this reason, we are changing the timing convention slightly and include the time t macroeconomic and financial variables into the vector W_{it-1}.

We estimate the following two models with covariates: (i) a model that only includes the unemployment rate as an additional predictor; (ii) a model that includes the unemployment rate, the federal funds rate, and an interest rate spread.[12] Because these predictors are not bank-specific, the effect of the predictors on PPNRs has to be identified from time-series variation, which is challenging given the short time-dimension of our panels. In this subsection, we generate forecasts using the actual values of the aggregate predictors (which we can evaluate based on the actual PPNR realizations for the forecast period). In Section 6.2, we

[12] All three series are obtained from the FRED database maintained by the Federal Reserve Bank of St. Louis: Unemployment is UNRATE, the effective federal funds rate is EFFR, and the spread between the federal funds rate and the 10-year treasury bill is T10YFF. We use temporal averaging to convert high-frequency observations into quarterly observations.

compare these forecasts to predictions under a stressed scenario, in which we use hypothetical values for the covariates.

Figure 3 depicts MSE differentials relative to the MSE of the plug-in predictor:

$$\Delta(\widehat{Y}^N_{\tau+1}) = \frac{\text{MSE}(\widehat{Y}^N_{\tau+1}) - \text{MSE}(\text{plug-in})}{\text{MSE}(\text{plug-in})},$$

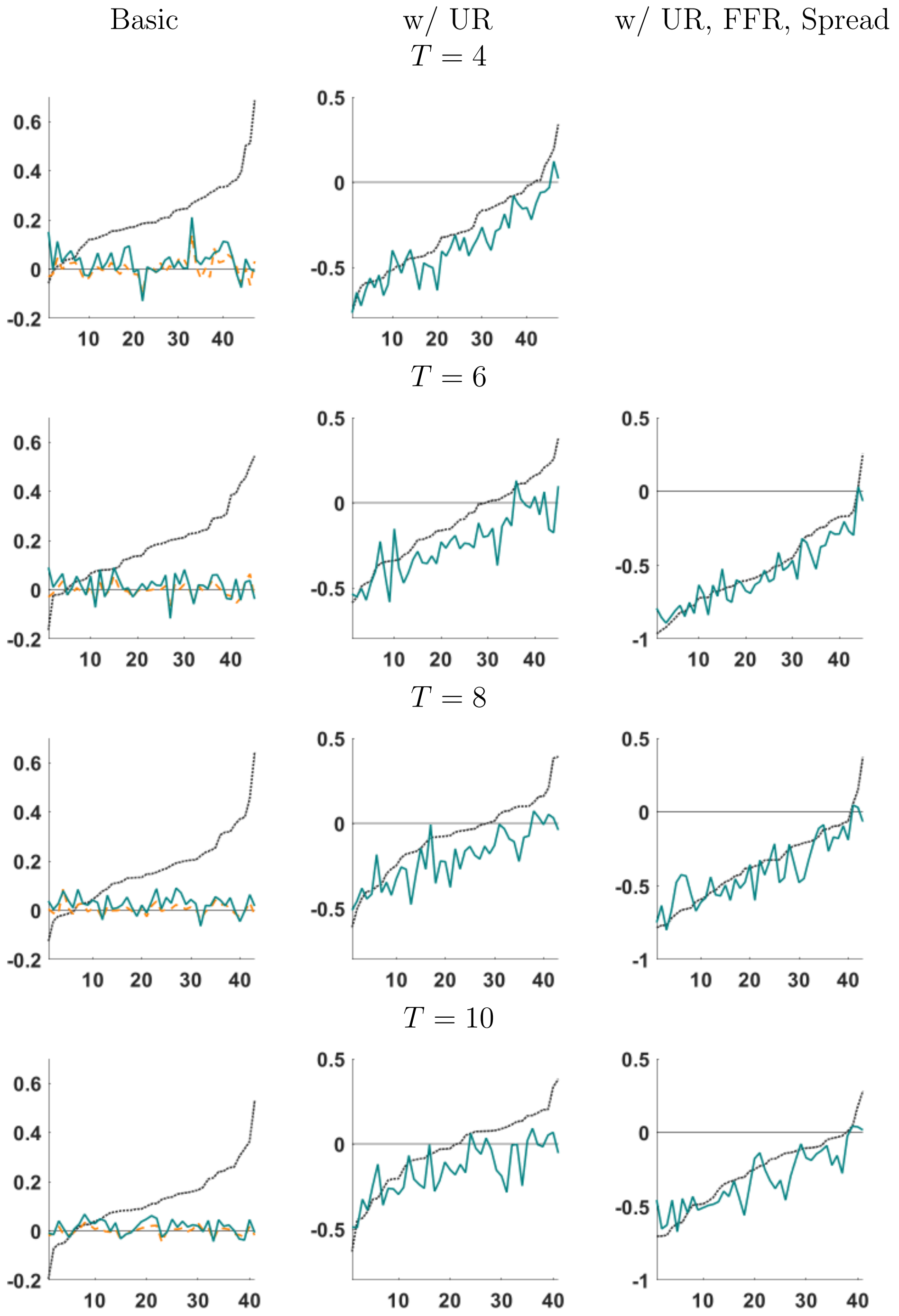
where $\text{MSE}(\widehat{Y}^N_{\tau+1})$ is defined in (33) and for now we set the selection operator $D_i(\mathcal{Y}_{i\tau}) = 1$, meaning we are averaging over all banks. If $\Delta(\widehat{Y}^N_{\tau+1}) < 0$, then the predictor $\widehat{Y}^N_{\tau+1}$ is more accurate than the plug-in predictor. The three columns correspond to the three different forecast models under consideration and the four rows correspond to the sample sizes $T = 4$, $T = 6$, $T = 8$, and $T = 10$, respectively. In the x-dimension, the MSE differentials are not arranged in chronological order by τ. Instead, we sort the samples based on the magnitude of the MSE differential for the pooled OLS predictor.

The plug-in predictor (the zero lines in Figure 3) and the pooled-OLS predictor (black dotted lines) provide natural benchmarks for the assessment of the empirical Bayes predictors. The former is optimal if the heterogeneous coefficients are essentially “uniformly” distributed in $\mathbb{R}^{k_w}$, whereas the latter is optimal in the absence of heterogeneity. The plug-in predictor dominates the pooled-OLS predictor whenever the number of heterogeneous coefficients is small relative to the time series dimension, which is the case for the basic model. For the unemployment model with $T = 4$ and the model with three covariates the pooled OLS predictor is more accurate than the plug-in predictor. For the model with unemployment only, the ranking is sample dependent.

The empirical Bayes procedure works generally well, in that it is adaptive: for most samples the empirical Bayes predictor is at least as accurate as the better of the plug-in and the pooled-OLS predictor. The unemployment-rate model provides a nice illustration of this adaptivity. In panels (2,2), (3,2), and (4,2) the fraction of samples in which the plug-in predictor dominates the pooled-OLS predictor ranges from 1/3 to 1/2. In all of these samples the MSE differential for the empirical Bayes predictor is close to zero or below zero. In the remaining samples the MSE differential of the empirical Bayes predictor tends to be smaller than the MSE differential associated with pooled OLS, highlighting that the shrinkage induced by the estimated correlated random effects distribution improves on the two benchmark procedures.

For the basic panel data model, we report results for two versions of the empirical Bayes

Figure 3: Percentage Change in MSE Relative to Plug-in Predictor, All Banks



Basic

w/ UR

w/ UR, FFR, Spread

$T = 4$

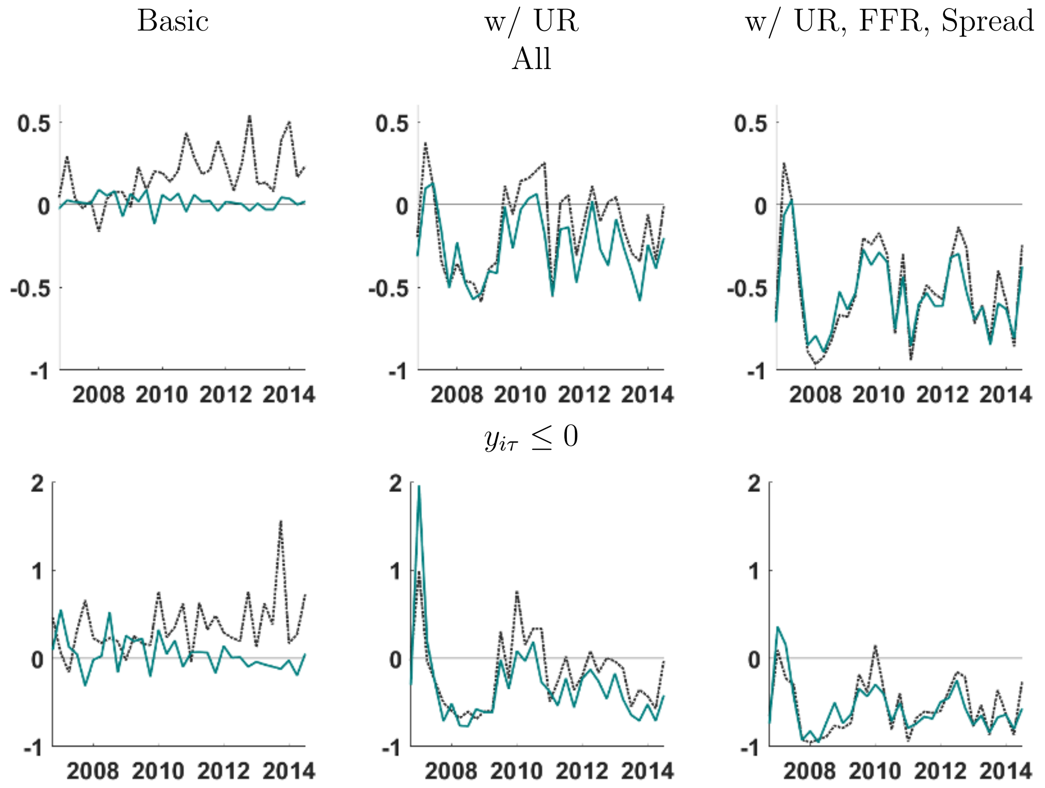
$T = 6$

$T = 8$

$T = 10$

Notes: Benchmark is the plug-in predictor. y-axis shows percentage changes in MSE, whereby a negative value is an improvement compared to the plug-in predictor. Time periods are sorted such that the relative MSE of pooled OLS is monotonically increasing. Comparison to: (i) empirical Bayes predictor with kernel estimator ($b = \hat{b}$), dashed orange; (ii) empirical Bayes predictor with mixture estimator ($K = \hat{K}$), solid teal; (iii) pooled OLS, dotted black.

Figure 4: Percentage Change in MSE Relative to Plug-in Predictor, $T = 6$



Basic

w/ UR

w/ UR, FFR, Spread

All

$y_{iT} \leq 0$

Notes: Benchmark is the plug-in predictor. y-axis shows percentage changes in MSE, whereby a negative value is an improvement compared to the plug-in predictor. x-axis shows forecast origin. Comparison to: (i) empirical Bayes predictor with mixture estimator ($K = \hat{K}$), solid teal; (iii) pooled OLS, dotted black.

predictor: one is based on the kernel estimation of the Tweedie correction term with $b = \hat{b}$ and the other one is based on the mixture estimation with $K = \hat{K}$. Here the dimension of the density that needs to be estimated to construct the correction is equal to two and the kernel and mixture estimation perform approximately equally well. For the models with covariates, a higher-dimensional density needs to be estimated, and the mixture estimation approach works generally better. In fact, for some of the samples the kernel estimates were quite erratic, which is why in columns 2 and 3 of Figure 3 we only report results for the mixture-based predictor.

In Figure 4 we focus on results for the samples of length $T = 6$. We now arrange the samples in chronological order. The first row of Figure 4 contains the same MSE differentials as the second row of Figure 3. For the basic panel data model (see Panel (1,1)), the relative performance of the pooled OLS predictor deteriorated over time, whereas the empirical Bayes predictor mimicked the performance of the plug-in predictor over time. For forecast origins from 2007:Q3 to 2009:Q2 shrinkage toward a common coefficient improved the forecasts quite substantially compared to the plug-in predictor. In subsequent periods the relative gain from shrinkage toward homogeneity oscillates over time.

For the models with covariates (see Panels (1,2) and (1,3)) there is no gain from shrinkage in 2007:Q3, but strong gains around 2008. In subsequent periods the relative performance of the empirical Bayes and pooled-OLS predictors oscillate, but MSE differentials remain negative. The second row of Figure 4 shows MSE differentials for banks that were generating losses at the forecast origin. The pattern for the relative performance of the predictors appears to be quite similar to the full sample.

Figure 5 examines the bank-specific squared forecast error loss differentials (subtracting the squared forecast error associated with the plug-in predictor) for the unemployment model. Each dot corresponds to a bank. We standardize the squared forecast error loss differentials by the MSE of the plug-in predictor, i.e., we are plotting

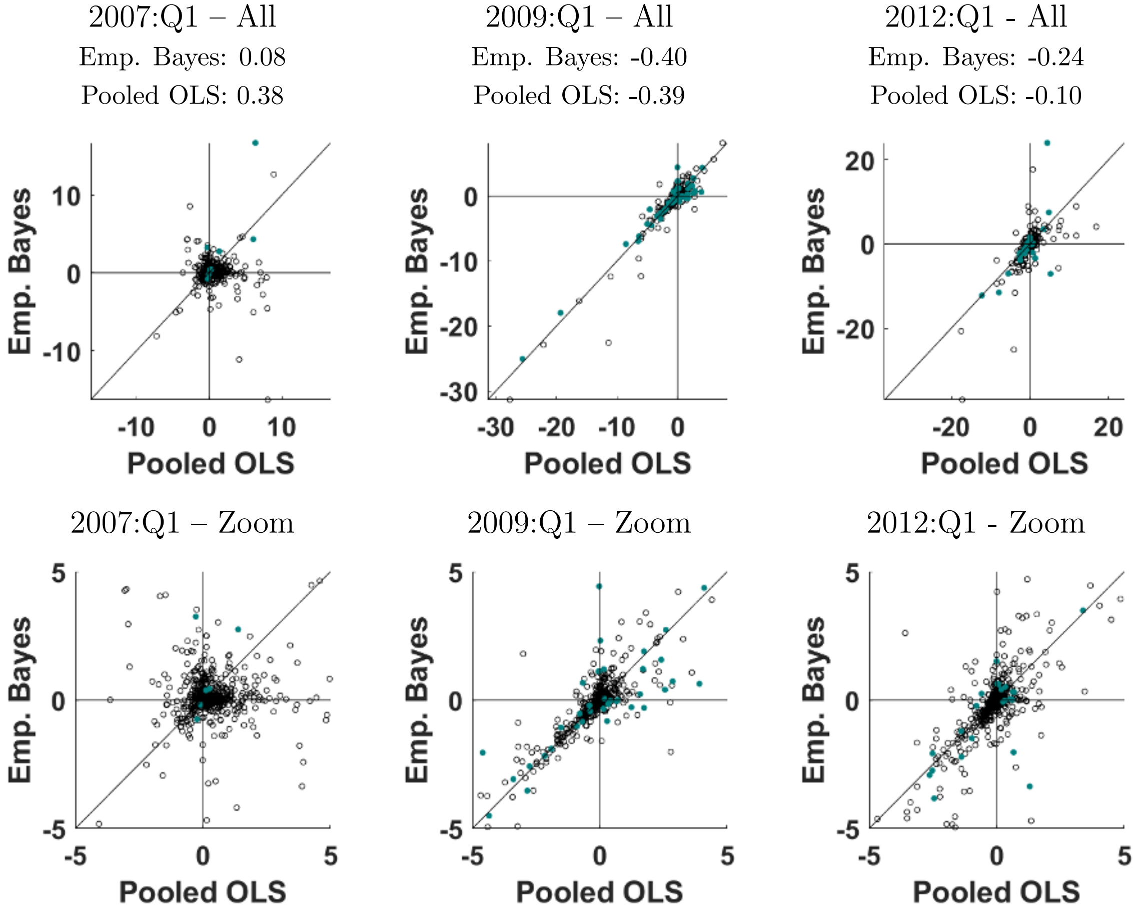
$$\frac{1}{\text{MSE(plug-in)}}\left[\big(Y_{iT+1}-\widehat{Y}_{iT+1}\big)^2-\big(Y_{iT+1}-\widehat{Y}_{iT+1}(\text{plug-in})\big)^2\right].$$

Thus, averaging the dots leads to the MSE differentials in Figure 4. The two zero lines and the 45-degree line partition each panel into six segments. Dots to the left of the vertical zero line correspond to banks for which the pooled OLS forecast is more accurate (lower squared forecast error) than the plug-in predictor. Dots below the horizontal zero line are associated with banks for which the empirical Bayes forecast is more accurate than the forecast from the plug-in predictor. Finally, dots below the 45-degree line correspond to institutions for which the empirical Bayes forecast is more accurate than the pooled OLS forecast.

We focus on three different time periods. In 2007:Q1 the MSE of the pooled OLS predictor is 38% larger than that of the plug-in predictor, whereas the empirical Bayes predictor is only slightly worse, an 8% MSE increase, than the plug-in predictor. In 2009:Q1, the pooled OLS and empirical Bayes predictors perform equally well, and generate a 40% MSE reduction relative to the plug-in predictor. Finally, in 2012:Q1, the empirical Bayes predictor performs better than both the pooled OLS and the plug-in predictor. The top row of Figure 5 shows squared forecast error differentials for all banks, whereas the bottom figure zooms in on differentials between -5 and 5.

The visual impression from the panels is consistent with the MSE ranking of the predictors. For instance, in the left panels there are more banks above the horizontal zero line (410 vs. 317) and to the right of the vertical zero line (470 vs. 257). Moreover, there are more

Figure 5: Squared Forecast Error Differentials Relative To Plug-in Predictor, Model w/ UR, $T = 6$



Notes: Figure depicts scatter plots of scaled squared forecast error differentials for pooled OLS and empirical Bayes with mixture estimator ($K = \hat{K}$) relative to the plug-in predictor. The differentials are divided by the average MSE (across all units) of the plug-in predictor. Cross-sectional averaging of the dots yields the values (%) that are listed below the plots and depicted in panels (1,2) and (2,2) of Figure 4 for the corresponding time periods. Negative values are improvements compared to the plug-in predictor. Teal dots indicate banks for which $y_{i\tau} \leq 0$. The thin solid lines correspond to zero lines and the 45-degree line, respectively.

banks below the 45-degree line than above (440 vs. 287). The panels in the center column of the figure indicate that the good performance of the empirical Bayes and pooled OLS predictors is driven in part by some banks for which the plug-in predictor performs very poorly. The corresponding squared forecast error differentials line up along the 45-degree line. It is important to note that the empirical Bayes predictor and the pooled-OLS predictor, despite the similarity in average performance, are not based on the same prediction function. The estimated autoregressive coefficient for the pooled-OLS predictor is much larger than the QMLE estimate of ρ that is used for the empirical Bayes predictor.

6.2 Forecasts Under Stressed Macroeconomic Conditions

We proceed by comparing the baseline forecasts from the previous subsection to predictions under a stressed scenario, in which we use hypothetical values for the predictors. When analyzing stress scenarios, one is typically interested in the effect of stressed economic conditions on the current performance of the banking sector. This is why we used the convention that the vector W_{it-1} includes time t macroeconomic and financial variables. Our subsequent analysis assumes that in the short run there is no feedback from disaggregate BCH revenues to aggregate conditions. While this assumption is inconsistent with the notion that the performance of the banking sector affects macroeconomic outcomes, elements of the Comprehensive Capital Analysis and Review (CCAR) conducted by the Federal Reserve Board of Governors have this partial equilibrium flavor.

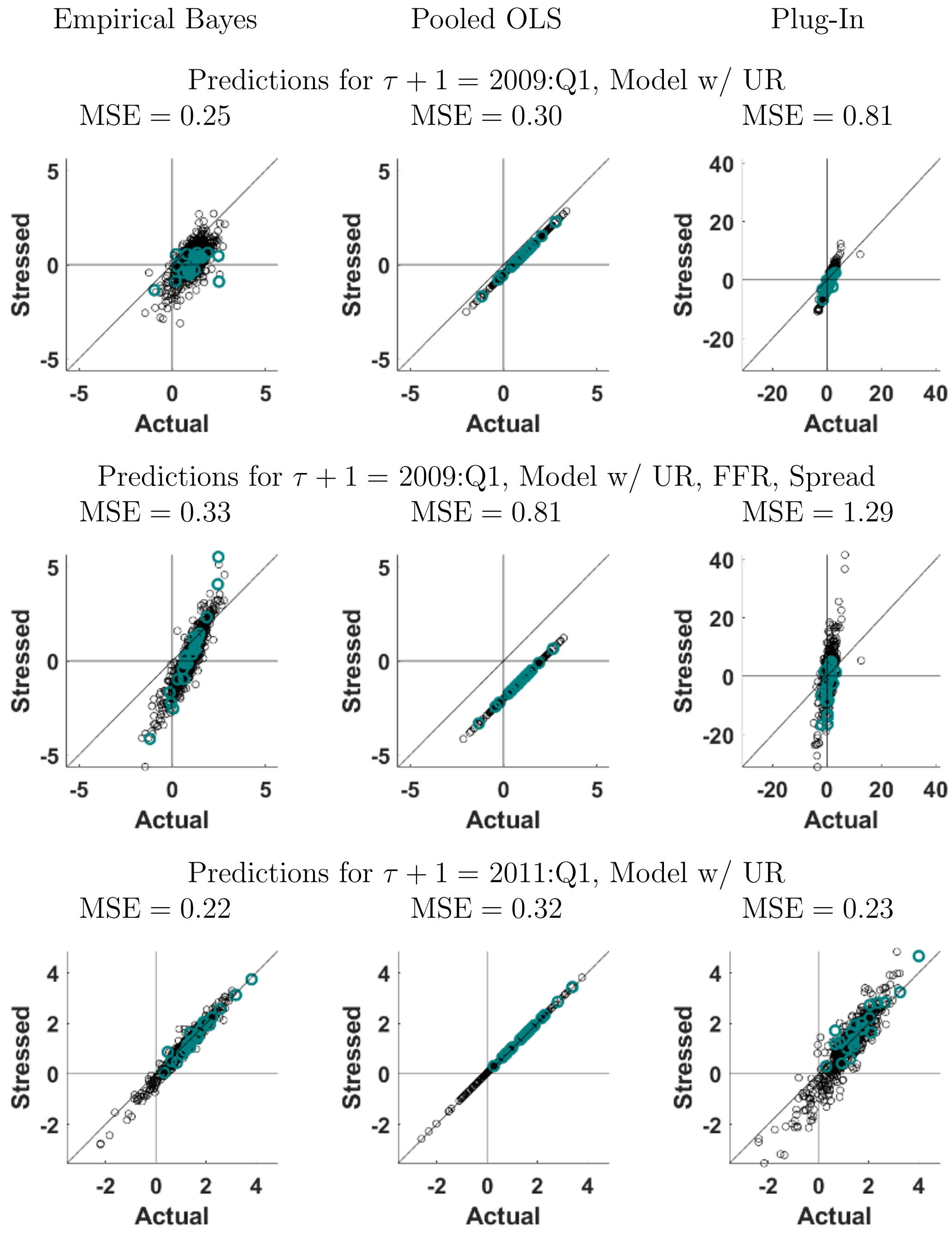
Results for one-quarter-ahead predictions of PPNRs are depicted in Figure 6. Each circle in the graphs corresponds to a particular BHC. We indicate institutions with assets greater than 50 billion dollars[13] by teal circles, while the other BHCs appear as black circles. The x-dimension is the forecast under actual macroeconomic conditions and the y-dimension indicates the forecast under the stressed scenario. For the model with unemployment as covariate we impose stress by increasing the unemployment rate by 5%. This corresponds to the unemployment movement in the *severely adverse* macroeconomic scenario in the Federal Reserve's CCAR 2016. For the model with three covariates the stressed scenario consists of an increase in the unemployment rate by 5% (as before) and an increase in nominal interest rates and spreads by 5%. This scenario could be interpreted as an aggressive monetary tightening that induced a sharp drop in macroeconomic activity. In each figure we also report the MSE associated with the various forecasts conditional on the actual macroeconomic conditions.

The two zero lines and the 45-degree line partition each panel into six segments. Dots to the right of the vertical zero line correspond to banks for which the predicted profits under the actual macroeconomic conditions are positive. Dots above the horizontal zero line are associated with banks for which predicted profits are positive under stressed macroeconomic conditions. Finally, dots below the 45-degree line correspond to institutions that are adversely affected by the stressed macroeconomic conditions: predicted revenues are smaller than under the actual conditions.

The graphs in the top two rows of Figure 6 depict forecasts for 2009:Q1 made in the midst of the Great Recession. For the majority of banks – 90% or more based on the empirical

[13] These are the BHCs that are subject to the CCAR requirements.

Figure 6: Predictions under Actual and Stressed Scenarios, $T = 10$



Empirical Bayes Pooled OLS Plug-In

Predictions for $\tau + 1 =$ 2009:Q1, Model w/ UR

MSE = 0.25 MSE = 0.30 MSE = 0.81

Predictions for $\tau + 1 =$ 2009:Q1, Model w/ UR, FFR, Spread

MSE = 0.33 MSE = 0.81 MSE = 1.29

Predictions for $\tau + 1 =$ 2011:Q1, Model w/ UR

MSE = 0.22 MSE = 0.32 MSE = 0.23

Notes: Forecast origins are $\tau =$ 2008:Q4 (panels in rows 1 and 2) and $\tau =$ 2010:Q4 (panels in row 3). Each dot corresponds to a BHC in our dataset. We indicate institutions with assets greater than 50 billion dollars by teal circles. We plot point predictions of PPNR under the actual macroeconomic conditions and a stressed scenario. Model w/ UR: unemployment rate is 5% higher than its actual level. Model w/ UR, FFR, Spread: the unemployment rate, the federal funds rate, and spread are 5% higher than their actual level. We also report actual MSEs.

Bayes and pooled OLS predictors, and between 80% and 85% under the plug-in predictor – the predicted PPNRs remain positive. The MSEs reported in the figure imply that the predictions from the model with one covariate are more accurate than the prediction for the model with three covariates. This result is not surprising, because in our sample we only have 10 time periods to disentangle the marginal effects of unemployment, federal funds rate, and spreads on bank revenues. For each of the two model specifications, empirical Bayes predictor dominates the pooled-OLS predictor, which in turn attains a lower MSE than the plug-in predictor.[14] Thus, overall, the lowest MSE among the six predictors depicted in the top two rows of the figure is attained by the empirical Bayes predictor based on the model with unemployment.

The predictions for 2009:Q1 from the model with one covariate under stressed macroeconomic conditions obtained from pooled OLS and the plug-in predictor look markedly different. The former essentially predicts no effects on bank revenues because the dots line up along the 45-degree line. The latter predicts a response that is very heterogeneous across institutions: 33% of the banks are predicted to be able to raise their revenues, whereas for 67% of the institutions the revenues are expected to fall relative to the baseline scenario. Predicted losses are as large as 10% of the bank assets. According to the preferred (based on the MSE under the baseline scenario) empirical Bayes predictor, 93% of the institutions are expected to experience a drop in PPNRs by 1 to 2 percent of their assets.

The last row of Figure 6 shows predictions for 2011:Q1 made during the recovery, based on the one-variable model. Unlike in the earlier sample, now the plug-in predictor generates more accurate forecasts (lower MSE) than the pooled-OLS predictor. As before, the empirical Bayes predictor beats both alternatives, albeit the plug-in predictor only by a small margin. The actual-versus-stressed predictions from the empirical Bayes procedure line up along the 45-degree line. For 68% of the institutions predicted profits are lower under the stressed scenario than under the benchmark scenario, but the drop in revenues is very small. Under the plug-in predictor, there is more heterogeneity in the response of banks' PPNRs, with some banks revenues dropping by 1.5 percentage points, whereas other banks are predicted to observe a modest increase in revenues. However, the baseline forecasts of this predictor

[14] The computation of the empirical Bayes predictor in this section is slightly different. After fitting mixture models to $p(\hat{\lambda}|y_0)$ we discovered that our data driven selection typically generates $\hat{K} = 1$, which means that $\hat{\lambda}|y_0$ is multivariate normal. Rather than directly estimating a normal distribution with an unrestricted variance-covariance matrix, we parameterize $p(\hat{\lambda}|y_0)$ in terms of the coefficients of a Gaussian prior $\pi(\lambda|y_0)$, imposing that the prior covariance matrix is diagonal. While in most periods the two approaches lead to the same results, there are some periods in which the latter approach is numerically more stable.

are less accurate than those from the empirical Bayes predictor, lending more credibility to the latter.

Discussion. We view this analysis as a first step toward applying state-of-the-art panel data forecasting techniques to stress tests. First, it is important to ensure that the empirical model is able to accurately predict bank revenues and balance sheet characteristics under observed macroeconomic conditions. Our analysis suggests that there are substantial performance differences among various plausible estimators and predictors. Second, a key challenge is to cope with the complexity of models that allow for heterogeneous coefficients in view of the limited information in the sample. There is a strong temptation to over-parameterize models that are used for stress tests. We use prior information to discipline the inference. In our empirical Bayes procedure, this prior information is essentially extracted from the cross-sectional variation in the data set. While we *a priori* allowed for heterogeneous responses, it turned out *a posteriori*, trading-off model complexity and fit, that the estimated coefficients exhibited very little heterogeneity. Third, our empirical results indicate that relative to the cross-sectional dispersion of PPNRs, the effect of severely adverse scenarios on revenue point predictions are very small. We leave it future research to explore richer empirical models that focus on specific revenue and accounting components and consider a broader set of covariates. Finally, it would be desirable to allow for a feedback from the performance of the banking sector into the aggregate conditions.

7 Conclusion

The literature on panel data forecasting in settings in which the cross-sectional dimension is large and the time-series dimension is small is very sparse. Our paper contributes to this literature by developing an empirical Bayes predictor that uses the cross-sectional information in the panel to construct a prior distribution that can be used to form a posterior mean predictor for each cross-sectional unit. The shorter the time-series dimension and the smaller the parameter heterogeneity, the more important this prior becomes for forecasting and the larger the gains from using the posterior mean predictor instead of a plug-in predictor. We consider a particular implementation of this idea for linear models with Gaussian innovations that is based on Tweedie's posterior mean formula. It can be implemented by estimating the cross-sectional distribution of sufficient statistics for the heterogeneous coefficients in the forecast model. We provide a theorem that establishes a ratio-optimality property for a nonparametric kernel estimator of the Tweedie correction and consider implementations based on the estimation of mixtures of normals and nonparametric MLE in Monte Carlo simulations. We illustrate in an application that our forecasting techniques work well in practice and may be useful to execute bank stress tests.

References

Alvarez, J., and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71(4), 1121–1159.

Anderson, T. W., and C. Hsiao (1981): "Estimation of dynamic models with error components," *Journal of the American statistical Association*, 76(375), 598–606.

Arellano, M. (2003): *Panel Data Econometrics*. Oxford University Press.

Arellano, M., and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, 58(2), 277–297.

Arellano, M., and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 79(3), 987–1020.

Arellano, M., and O. Bover (1995): "Another look at the instrumental variable estimation of error-components models," *Journal of econometrics*, 68(1), 29–51.

Arellano, M., and B. Honoré (2001): "Panel data models: some recent developments," *Handbook of econometrics*, 5, 3229–3296.

Baltagi, B. (1995): *Econometric Analysis of Panel Data*. John Wiley & Sons, New York.

Baltagi, B. H. (2008): "Forecasting with panel data," *Journal of Forecasting*, 27(2), 153–173.

Blundell, R., and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of econometrics*, 87(1), 115–143.

Brown, L. D., and E. Greenshtein (2009): "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means," *The Annals of Statistics*, pp. 1685–1704.

Chamberlain, G., and K. Hirano (1999): "Predictive distributions based on longitudinal earnings data," *Annales d'Economie et de Statistique*, pp. 211–242.

Covas, F. B., B. Rump, and E. Zakrajsek (2014): "Stress-Testing U.S. Bank Holding Companies: A Dynamic Panel Quantile Regression Approach," *International Journal of Forecasting*, 30(3), 691–713.

Efron, B. (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106(496), 1602–1614.

Goldberger, A. S. (1962): "Best linear unbiased prediction in the generalized linear regression model," *Journal of the American Statistical Association*, 57(298), 369–375.

Gu, J., and R. Koenker (2016): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Economics (Forthcoming).*

——— (2017): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Business & Economic Statistics*, 35(1), 1–16.

Hirano, K. (2002): "Semiparametric Bayesian inference in autoregressive panel data models," *Econometrica*, 70(2), 781–799.

Hirtle, B., A. Kovner, J. Vickery, and M. Bhanot (2016): "Assessing Financial Stability: The Capital and Loss Assessment under Stress Scenarios (CLASS) Model," *Journal of Banking & Finance*, 69, S35–S55.

Hsiao, C. (2014): *Analysis of Panel Data.* Cambridge University Press, Cambridge.

Jiang, W., and C.-H. Zhang (2009): "General Maximum Likelihood Empirical Bayes Estimation of Normal Means," *The Annals of Statistics*, 37(4), 1647–1684.

Kapinos, P., and O. A. Mitnik (2016): "A Top-down Approach to Stress-testing Banks," *Journal of Financial Services Research*, 49(2), 229–264.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27(4), 887–906.

Lancaster, T. (2002): "Orthogonal parameters and panel data," *The Review of Economic Studies*, 69(3), 647–666.

Liu, L. (2018): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *arXiv preprint arXiv:1805.04178.*

Liu, L., H. R. Moon, and F. Schorfheide (2017): "Forecasting With Dynamic Panel Data Models," *Manuscript, arXiv*, 1709.10193.

——— (2018): "Forecasting With a Panel Tobit Model," *Manuscript, University of Pennsylvania.*

Norets, A., and J. Pelenis (2012): "Bayesian Modeling of Joint and Conditional Distributions," *Journal of Econometrics*, 168, 332–346.

Robbins, H. (1951): "Asymptocially Subminimax Solutions of Compound Decision Problems," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press, Berkeley and Los Angeles.

——— (1956): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, Berkeley and Los Angeles.

——— (1964): "The empirical Bayes approach to statistical decision problems," *The Annals of Mathematical Statistics*, pp. 1–20.

Robert, C. (1994): *The Bayesian Choice*. Springer Verlag, New York.

Robinson, G. K. (1991): "That BLUP is a good thing: the estimation of random effects," *Statistical science*, pp. 15–32.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.