Synthetic prediction methods for minimizing post shock forecasting error

Daniel J. Eck and Soheil Eshghi

May 2019

Abstract

We seek to develop a forecasting methodology for time series data that is thought to have undergone a shock which has origins that have not been previously observed. We still can provide credible forecasts for a time series in the presence of such systematic shocks by drawing from disparate time series that have undergone similar shocks for which post-shock outcome data is recorded. These disparate time series are assumed to have mechanistic similarities to the time series under study but are otherwise independent (Granger noncausal). The inferential goal of our forecasting methodology is to supplement observed time series data with post-shock data from the disparate time series in order to minimize average forecast risk.

1 Setting

Suppose that an analyst is interested in forecasting a real-valued time series $y_1, y_2, \ldots$. Given each time point $t \geq 1$, let $\mathbf{x}_t$ be the (possibly multivariate) information variable vector revealed prior to the observation of y_t. To gauge the performance of a procedure that produces forecasts $\{\hat{y}_t, t = 1, 2, \ldots\}$ given time horizon T, we consider the average forecast risk

$$R_T = \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}(y_t - \hat{y}_t)^2$$

in our analyses.

2 Dynamic panel model

We consider a dynamic panel data model similar with autoregressive structure. For simplicity, we will consider the following simplified version with one covariate:

$$y_{i,t} = \eta_i + \alpha_i D_{i,t} + \phi_i y_{i,t-1} + \theta_i \mathbf{x}_{i,t} + \varepsilon_{i,t}, \tag{1}$$

where $i = 1, \ldots, J+1$ and $t = 1, \ldots, T$ and $D_{i,t} = 1(t > T_i^*)$, $T_i^* < T$. For simplicity we will assume that $T_i^* = T$ for all $i = 1, \ldots, J$. We will consider a simple random effects structure where

$$
\begin{aligned}
\eta_i &\overset{iid}{\sim} N(0, \sigma_\eta^2), \qquad i = 1, \ldots J+1, \\
\alpha_i &\overset{iid}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \qquad i = 1, \ldots J+1, \\
\phi_i &\overset{iid}{\sim} U(-1, 1), \qquad i = 1, \ldots J+1, \\
\theta_i &\overset{iid}{\sim} U(-1, 1), \qquad i = 1, \ldots J+1, \\
\varepsilon_{i,t} &\overset{iid}{\sim} N(0, \sigma^2), \qquad i = 1, \ldots J+1;\ t = 1, \ldots T, \\
\eta \perp\!\!\!\perp \alpha &\perp\!\!\!\perp \phi \perp\!\!\!\perp \theta \perp\!\!\!\perp \varepsilon;
\end{aligned}
$$

Without loss of generality, we will suppose that we want to forecast for y_{1,T^*+1} where we have no observations for $y_{1,t}, t > T^*$. Inference about the random effects parameters is not of interest in this forecasting context. Conditional on all regression parameters (except α?), previous responses, and covariates, model (1) has distribution

$$y_{i,t} \sim N(\eta_i + \alpha_i D_{i,t} + \phi_i y_{i,t-1} + \theta_i \mathbf{x}_{i,t}, \sigma^2).$$

All parameters will be estimated with OLS. In particular, let $\hat{\alpha}_i$, $i = 2, \ldots, J+1$ be the OLS estimate of α_i and define the adjusted α plugin estimator for time series $i = 1$ by,

$$\hat{\alpha}_{\text{adj}} = \frac{1}{J} \sum_{i=2}^{J+1} \hat{\alpha}_i \tag{2}$$

where the $\hat{\alpha}_i$s are MLEs of all of the α_is. We can use $\hat{\alpha}_{\text{adj}}$ as an estimator for the unknown α_1 term for which no meaningful estimation information otherwise exists.

Consider the candidate forecasts:

$$
\begin{aligned}
\text{Forecast 1} : \hat{y}^1_{1,T^*+1} &= \hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1}, \\
\text{Forecast 2} : \hat{y}^2_{1,T^*+1} &= \hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} + \hat{\alpha}_{\text{adj}},
\end{aligned}
$$

where $\hat{\eta}_1$, $\hat{\phi}_1$, and $\hat{\theta}_1$ are MLEs of η_1, ϕ_1, and θ_1 respectively. The two forecasts do not differ in their predictions for y_t, $t = 1, \ldots T^*$. They only differ in predicting y_{T^*+1}.

We want to determine when either $\hat{y}^1_{1,T^*+1}$ or $\hat{y}^2_{1,T^*+1}$ minimizes the average forecast risk.

The two forecasts do not differ in their predictions for y_t, $t = 1, \ldots T^*$. repetition. They only differ in predicting y_{T^*+1}. The forecast risk for $y_{T+1,2}$ is:

$$
\begin{aligned}
&\mathrm{E}(\hat{y}^2_{1,T^*+1} - y^2_{1,T^*+1})^2 \\
&\quad = \mathrm{E}\left\{\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} + \hat{\alpha}_{\text{adj}} - (\eta_1 + \phi_1 y_{1,T^*} + \theta_1 \mathbf{x}_{1,T^*+1} + \alpha_1 + \varepsilon_{1,T^*+1})\right\}^2 \\
&\quad = \mathrm{E}\left\{(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1 \mathbf{x}_{1,T^*+1} - \alpha_1 - \varepsilon_{1,T^*+1}) + \hat{\alpha}_{\text{adj}}\right\}^2 \\
&\quad = R_{T,1} + \mathrm{E}(\hat{\alpha}^2_{\text{adj}}) - 2\,\mathrm{E}\left\{(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1 \mathbf{x}_{1,T^*+1} - \alpha_1 - \varepsilon_{1,T^*+1})\hat{\alpha}_{\text{adj}}\right\} \\
&\quad = R_{T,1} + \mathrm{E}(\hat{\alpha}^2_{\text{adj}}) - 2\,\mathrm{E}(\hat{\alpha}_{\text{adj}})\,\mathrm{E}(\alpha_1) - 2\,\mathrm{E}\left\{(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1 \mathbf{x}_{1,T^*+1})\hat{\alpha}_{\text{adj}}\right\} \\
&\quad = R_{T,1} + \mathrm{Var}(\hat{\alpha}_{\text{adj}}) - \mu_\alpha^2 - 2\,\mathrm{E}\left(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1 \mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1 \mathbf{x}_{1,T^*+1}\right)\mu_\alpha,
\end{aligned}
$$

where

$$\mathrm{E}(\hat{\alpha}_{\mathrm{adj}}) = \mathrm{E}\left\{\mathrm{E}(\hat{\alpha}_{\mathrm{adj}}|\mathcal{H})\right\} = \mathrm{E}\left(\frac{1}{J}\sum_{i=2}^{J+1}\alpha_i\right) = \mu_\alpha,$$

and

$$\mathcal{H} = \{(\eta_i, \phi_i, \theta_i, \alpha_i, \mathbf{x}_{i,t}, y_{i,t}); i = 2, \ldots, J+1, t = 1, \ldots, T\}.$$

Define

$$\mathcal{H}_1 = \{\eta_1, \phi_1, \theta_1, \alpha_1, \mathbf{x}_{1,T^*+1}, (\mathbf{x}_{1,t}, y_{1,t}); t = 1, \ldots, T^*\}.$$

Notice that

$$\begin{aligned}
&\mathrm{E}\left(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1\mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1\mathbf{x}_{1,T^*+1}\right)\\
&\quad = \mathrm{E}\left\{\mathrm{E}\left(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1\mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1\mathbf{x}_{1,T^*+1}\right)|\mathcal{H}\cup\mathcal{H}_1\right\} = 0,
\end{aligned}$$

(**is the above true?**) and that

$$\begin{aligned}
\mathrm{Var}(\hat{\alpha}_{\mathrm{adj}}) &= \mathrm{E}\left\{\mathrm{Var}(\hat{\alpha}_{\mathrm{adj}}|\mathcal{H})\right\} + \mathrm{Var}\left\{\mathrm{E}(\hat{\alpha}_{\mathrm{adj}}|\mathcal{H})\right\}\\
&= \mathrm{E}\left\{\mathrm{Var}(\hat{\alpha}_{\mathrm{adj}}|\mathcal{H})\right\} + \frac{\sigma^2_\alpha}{J}\\
&= \frac{1}{J^2}\sum_{i=2}^{J+1}\mathrm{E}\left\{\mathrm{Var}(\hat{\alpha}_j|\mathcal{H})\right\} + \frac{\sigma^2_\alpha}{J}\\
&= \frac{1}{J^2(T-1)}\sum_{i=2}^{J+1}\mathrm{E}\left\{\frac{s^2_{e,j}}{(1-R^2_j)s^2_{\alpha_j}}\right\} + \frac{\sigma^2_\alpha}{J},
\end{aligned}$$

where $s^2_{e,j}$ is the estimated variance of model j, $R^2_{\alpha_j}$ is the multiple R^2 obtained from regressing the shock indicator D on the other regressors, and s^2_j is... .

Forecast 2 has lower forecast risk than Forecast 1 when,

$$\mathrm{Var}(\hat{\alpha}_{\mathrm{adj}}) - \mu^2_\alpha - 2\,\mathrm{E}\left(\hat{\eta}_1 + \hat{\phi}_1 y_{1,T^*} + \hat{\theta}_1\mathbf{x}_{1,T^*+1} - \eta_1 - \phi_1 y_{1,T^*} - \theta_1\mathbf{x}_{1,T^*+1}\right)\mu_\alpha < 0. \tag{3}$$

Putting everything together, we see that (3) is equivalent to

$$\frac{1}{J^2(T-1)}\sum_{i=2}^{J+1}\mathrm{E}\left\{\frac{s^2_{e,j}}{(1-R^2_j)s^2_{\alpha_j}}\right\} + \frac{\sigma^2_\alpha}{J} < \mu^2_\alpha.$$

Forecast 2 is preferable to Forecast 1 asymptotically in both T and J whenever $\mu_\alpha \neq 0$. In finite samples, Forecast 2 is preferable to Forecast 1 when the μ_α is large relative to its variability and overall regression variability.

3 Synthetic prediction via synthetic control method

Taken from Soheil's notes (begin). Assume we have covariate-observation pairs $(y_i, \mathbf{x}_i)$ for n points, and that $\mathbf{x}_1$, the covariates of the intervened unit, lies within the convex hull of $(\mathbf{x}_i)_{i=1}^n$. One way to match covariates is to pick a vector $\mathbf{w}$ such that $\mathbf{w}'1 = 1$ and $\|\mathbf{x}_1 - \mathbf{X}\mathbf{w}\|$, where $\mathbf{X}$ is the matrix of available covariates (arranged in column form), is minimized.

One can incorporate a notion of locality by penalizing the use of covariates far from the covariates of interest, through incorporating the following term in either a constraint or as part of the objective: $\sum_{i=1}^{n} \mathbf{w}_i \|\mathbf{x}_1 - \mathbf{x}_i\|$.

This leads to the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{w}} \quad & \|\mathbf{x}_1 - \mathbf{X}\mathbf{w}\| \\
\text{s.t.} \quad & \mathbf{w}'1 = 1, \\
& \sum_{i=1}^{n} \mathbf{w}_i \|\mathbf{x}_1 - \mathbf{x}_i\| \leq \delta, \\
& \mathrm{Var}(\mathbf{Y}\mathbf{w}) \leq \gamma,
\end{aligned}$$

with δ and γ being rightsizing parameters.

We can simplify the above using the following assumptions:

...Once I have time I will write these out so that e can extract the fixed values mostly to one side.

The rest is kind of bad as is, to be honest, and hard to justify. I still think we should weigh down the ones that are pretty far from the expected value, but it's not clear why the variance condition above will not be enough and we have to do a prediction model for all $(n-1)-$ tuples. Any ideas?

The use of the transformed linear model (linking covariates to observations of outcomes) to create the counterfactual show our belief in the model specification. Under these conditions, the prediction of the model for the outcome at covariates $\mathbf{x}_i$ based on all other observed covariates $\mathbf{x}_{-i}$ (if it is within the convex hull) should closely match its observed outcome. Any discrepancy should make us less confident in the observation of the outcome (if we hold the model specification to be correct), so we would want y_i to play less of a role in the creation of the counterfactual in that scenario. Thus, we will put an upper limit on $\mathbf{w}_i$, the weight assigned to the covariate-observation pair by a function of the discrepancy between y_1 and the created synthetic prediction for it $\hat{y}_i$ from $\mathbf{x}_{-i}$, i.e.,

Taken from Soheil's notes (end).

Prediction via this synthetic control method can be applied to the dynamic panel models in Section 2. In this case, the individual level shock effects α_i are correlated with the covariate information $\mathbf{x}_i$. We consider the model

$$\alpha_i = \mu_{\alpha_i} + \beta \mathbf{x}_i + \gamma_i \tag{4}$$

for the shock effects, where γ_i are iid with $\mathrm{E}(\gamma_i) = 0$ and $\mathrm{Var}(\gamma_i) = \sigma_\gamma^2$ and are independent of all other random effects. The specification in (4) can be substitutes into the dynamic panel model (1) which yields,

$$\begin{aligned}
y_{i,t} &= \eta_i + \alpha_i D_{i,t} + \phi_i y_{i,t-1} + \theta_i \mathbf{x}_{i,t} + \varepsilon_{i,t} \\
&= \eta_i + (\mu_{\alpha_i} + \beta \mathbf{x}_i + \gamma_i) D_{i,t} + \phi_i y_{i,t-1} + \theta_i \mathbf{x}_{i,t} + \varepsilon_{i,t} \\
&= (\eta_i + \mu_{\alpha_i} D_{i,t}) + \phi_i y_{i,t-1} + (\theta_i + \beta D_{i,t}) \mathbf{x}_{i,t} + (\gamma_i D_{i,t} + \varepsilon_{i,t}) \\
&= \tilde{\eta}_i + \phi_i y_{i,t-1} + \tilde{\beta}_{i,t} \mathbf{x}_{i,t} + \tilde{\varepsilon}_{i,t},
\end{aligned}$$

where $\tilde{\eta}_i \sim N(\mu_{\alpha_i}, \sigma_\eta^2)$ and $\tilde{\varepsilon}_{i,t} \sim N(0, \sigma^2 + D_{i,t}\sigma_\gamma^2)$.