Forecasting under Structural Breaks Using Improved Weighted Estimation

Tae-Hwy Lee[†] Shahnaz Parsaeian[‡*] Aman Ullah[§]

March 3, 2022

Abstract

In forecasting a time series containing a structural break, it is important to determine how much weight can be given to the observations prior to the time when the break occurred. In this context, Pesaran et al. (2013) (PPP) proposed a weighted least squares estimator by giving different weights to observations before and after a break point for forecasting out-of-sample. We revisit their approach by introducing an improved weighted generalized least squares estimator (WGLS) using a weight (kernel) function to give different weights to observations before and after a break. The kernel weight is estimated by cross-validation rather than analytically derived from a parametric model as in PPP. Therefore, the WGLS estimator facilitates implementation of the PPP method for the optimal use of the pre-break and post-break sample observations without having to derive the parametric weights which may be misspecified. We show that the kernel weight estimated by cross-validation is asymptotically optimal in the sense of Li (1987). Monte Carlo simulations and an empirical application to forecasting equity premium are provided for verification and illustration.

Keywords: Forecasting; Cross-validation; Kernel; Structural breaks; Model averaging

JEL Classifications: C14, C22, C53

[†]Department of Economics, University of California, Riverside, CA 92521. E-mail: taelee@ucr.edu
[‡]Department of Economics, University of Kansas, Lawrence, KS 66045. E-mail: sh.parsaeian@ku.edu
[*]Corresponding author: Shahnaz Parsaeian. Email: sh.parsaeian@ku.edu
[§]Department of Economics, University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu

1 Introduction

In forecasting with a time series model when there is a structural break in the conditional mean and/or in the conditional variance, the usual OLS estimator using all the observations (i.e., the full-sample, that is, observations before and after a break) is inconsistent. When the post-break sample is relatively large, a consistent estimator can be obtained from using only the post-break observations. However, as pointed out by Pesaran and Timmermann (2007), Pesaran and Pick (2011), Rossi (2013), Pesaran et al. (2013) (hereafter PPP), forecasts based on the post-break observations may not be optimal in terms of the mean squared forecast error (MSFE) as the estimation uncertainty may be large due to the relatively smaller number of observations in the post-break sample. Therefore, one can improve forecasting using the pre-break sample observations. When the break is "strong" (large break magnitude), the post-break estimator may be optimal. When there is "no" break, the full-sample estimator is clearly preferred and is optimal. However, when the break is "weak" (small break magnitude) or is of a moderate magnitude, a combination of the full-sample estimator and the post-break estimator would be optimal, where the combination weight is between 0 and 1 depending on the magnitude of the structural break and the timing of the structural break. This is because the combined estimator takes advantage of the trade-off between the bias and variance efficiency of the full-sample estimator.

An alternative approach to the fully parametric regression model is a semi-parametric kernel-based regression model. Under structural breaks, kernel smoothing of coefficients across regimes provides an appealing method of "combining information" (before the break and after the break) without forcing the choice between regression using full-sample observations and regression using only the post-break sample observations. This can also be viewed as the frequentist methods for model averaging (averaging the pre-break and post-break estimators as shown in equation (11) below, or equivalently averaging between the post-break and full-sample estimators) introduced by Hjort and Claeskens (2003), Hansen (2007, 2008), and Hansen and Racine (2012), among others. See also Clements and Hendry (2006, 2011), Pesaran and Timmermann (2002, 2005, 2007), Timmerman (2006) and the references therein, for different forecast combination methods.

The contribution of this paper is that we introduce the improved weighted estimator which facilitates implementation of using the pre-break data in addition to the post-break data, and

improves the forecasting performance under structural breaks. Motivation of the proposed estimator comes from the paper by Li et al. (2013) in which they focus on semi-parametric kernel estimation of varying-coefficient models with categorical variables.[1] In structural break models, as the slope parameters change across different regimes, they can be viewed as a function of a discrete covariate which takes different values across the regimes.[2] Inspired by the estimation method of Li et al. (2013), we develop a weighted generalized least squares (WGLS) estimator for time series structural break models. The approach is similar to semi-parametric methods since it uses kernel smoothing for the discrete covariate across regimes while the relationship between dependent variable and independent variables is parametrically specified. The kernel is based on the work of Aitchison and Aitken (1976) who offer the kernel smoothing of discrete covariates by borrowing information from neighboring subsets. For choosing the smoothing parameters we use leave-one-out cross-validation (CV) which was initially introduced by Stone (1974) and Geisser (1974). We prove theoretically that the weight estimated by CV is asymptotically optimal in the sense of Li (1987), in that the average squared error of the CV is asymptotically as small as the average squared error of the infeasible best possible estimator. We compare our WGLS estimator with the one proposed by Pesaran et al. (2013). We show that our estimator facilitates implementation of their approach in exploiting the pre-break sample observations. More specifically, the advantage of using this method is that we can find the kernel by CV without having to estimate unknown parameters such as break size in the regression coefficients or the error variance.

We conduct Monte Carlo experiments that compare the WGLS forecast with the PPP forecast, the five methods used in Pesaran and Timmermann (2007) including the post-break forecast, and the average window method proposed by Pesaran and Pick (2011). We evaluate the forecasting performance of these methods under different break size and different break points. We also apply the WGLS method to forecasting the equity premium. The results show the superior predictive ability of the WGLS forecast relative to the PPP, post-break forecast and other alternative approaches.

The outline of the paper is as follows. Section 2 sets up the structural break model and introduces the estimator proposed by Pesaran et al. (2013). Section 3 introduces our proposed

[1] Also, see the paper by Su et al. (2009) for a more general semi-parametric functional coefficient regression model with both continuous and discrete variables, and Su et al. (2013) where continuous variables are endogenous.

[2] For example, when there is a break, the discrete covariate for the pre-break observations take one value, say 0, and another value for the post-break observations, say 1.

WGLS estimator and its asymptotic properties. We also compare the WGLS estimator with the one proposed by Pesaran et al. (2013) and show their relationship. Section 4 reports Monte Carlo simulation. An empirical application is presented in Section 5. Section 6 concludes.

2 The model

Consider the linear structural break model with m breaks or $m+1$ regimes. There are T observations, and for simplicity we assume only one break, $m=1$, that occurs at T_1. The method can be easily extended to the cases with multiple breaks. Let the model have the form of $y_t = x_t'\beta_t + \sigma_t\varepsilon_t$, where x_t is a $k\times 1$ vector of stationary regressors, $\varepsilon_t \sim$ i.i.d.$(0,1)$, and the vector of coefficients, β_t, and the error variance, σ_t^2, are subject to breaks as

$$\beta_t = \begin{cases} \beta_{(1)} & \text{for} \quad 1 < t \leq T_1 \\ \beta_{(2)} & \text{for} \quad T_1 < t < T \end{cases} \tag{1}$$

and

$$\sigma_t = \begin{cases} \sigma_{(1)} & \text{for} \quad 1 < t \leq T_1 \\ \sigma_{(2)} & \text{for} \quad T_1 < t < T. \end{cases} \tag{2}$$

Let $Y = (Y_{(1)}'\ Y_{(2)}')'$ be a $T\times 1$ vector of the dependent variable where $Y_{(i)} = \big(y_{T_{i-1}+1},\ldots,y_{T_i}\big)'$, $X = (X_{(1)}'\ X_{(2)}')'$ be a $T\times k$ matrix of the independent variables where $X_{(i)} = (x_{T_{i-1}+1},\ldots,x_{T_i})'$, with $i=1,2$, and the convention that $T_0=0$, and $T_2=T$. Let $\big(\sigma_{(1)}\epsilon_{(1)}'\ \sigma_{(2)}\epsilon_{(2)}'\big)'$ represent the vector of the error terms where $\epsilon_{(i)} = (\varepsilon_{T_{i-1}+1},\ldots,\varepsilon_{T_i})'$, $i=1,2$, and the variance of the error term is denoted by $\Omega = \text{diag}\big(\sigma_{(1)}^2 I_{T_1},\ \sigma_{(2)}^2 I_{T-T_1}\big)$ which has heteroscedasticity across regimes. Also, $b_1 \equiv \frac{T_1}{T} \in (0,1)$ denotes the proportion of observations before the break. In matrix form, the model can be written as

$$\begin{cases} Y_{(1)} = X_{(1)}\beta_{(1)} + \sigma_{(1)}\epsilon_{(1)} & \text{for} \quad 1 < t \leq T_1 \\ Y_{(2)} = X_{(2)}\beta_{(2)} + \sigma_{(2)}\epsilon_{(2)} & \text{for} \quad T_1 < t < T. \end{cases} \tag{3}$$

We note that the method of break point estimation is based on the least-squares principle. That

is, for the break point T_1, the associated least-squares estimates of slope coefficients are obtained by minimizing the sum of squared residuals. Substituting these estimated coefficients in the objective function, the estimated break point is obtained. See Bai and Perron (1998, 2003).

Remark 1: In structural break models, it is often assumed that some of the slope coefficients can stay constant across the regimes; for example see Bai (1997). The current framework in (3) can also allow for partial structural change model. To see this, following Bai (1997), consider $x_t = (w_t' \; z_t')'$, where w_t is $k_1 \times 1$, z_t is $k_2 \times 1$, and $k_1 + k_2 = k$. Let $\beta_{(1)} = (\alpha' \; \delta_1')'$, $\beta_{(2)} = (\alpha' \; \delta_2')'$, $\delta = \delta_2 - \delta_1$, $\beta_{(1)} - \beta_{(2)} = -R\delta$ with $R' = \big(\mathbf{0}_{k_2 \times k_1} \; I_{k_2}\big)$, and $z_t = R'x_t$. Then, we can write Bai's equation (1) as our equation (3). We note that the case where $k_2 < k$ denotes a partial change model, while the case where $k_2 = k$ is for a full change model. □

2.1 Estimator by Pesaran et al. (2013)

Pesaran et al. (2013) propose that one can reduce MSFE under the structural breaks by using the full-sample observations instead of using only the post-break observations. Their proposed estimator is

$$\widehat{\beta}_{PPP} = (X'WX)^{-1}(X'WY). \tag{4}$$

They derive the optimal weight, W, such that MSFE of the one-step-ahead forecast based on (4) is minimized, and find that the optimal weight takes one value, $w_{(1)}$, for the pre-break observations and another value, $w_{(2)}$, for the post-break observations. In other words

$$W = \text{diag}\Big(w_{(1)}, \ldots, w_{(1)}, w_{(2)}, \ldots, w_{(2)}\Big), \tag{5}$$

where

$$\begin{cases} w_{(1)} = \frac{1}{T}\frac{1}{b_1 + (1-b_1)(q^2 + Tb_1\phi^2)}, \\ \\ w_{(2)} = \frac{1}{T}\frac{q^2 + Tb_1\phi^2}{b_1 + (1-b_1)(q^2 + Tb_1\phi^2)}, \end{cases} \tag{6}$$

$b_1 = \lim_{T\to\infty} \frac{T_1}{T}$ is the proportion of observations before the break, $q = \frac{\sigma_{(1)}}{\sigma_{(2)}}$ measures the break in the error variance, $\lambda = \beta_{(1)} - \beta_{(2)}$ is the break size in the regression coefficient, $\phi = \frac{x_{T+1}'\lambda}{\sigma_{(2)}(x_{T+1}'Q^{-1}x_{T+1})^{1/2}}$, and $Q = \mathbb{E}(x_t x_t')$. See Pesaran et al. (2013) for details.

We can rewrite the PPP estimator in equation (4) as

$$\begin{aligned}\widehat{\beta}_{PPP} &= \Big(w_{(1)}\ X'_{(1)}X_{(1)} + w_{(2)}\ X'_{(2)}X_{(2)}\Big)^{-1}\Big(w_{(1)}\ X'_{(1)}X_{(1)}\widehat{\beta}_1 + w_{(2)}\ X'_{(2)}X_{(2)}\widehat{\beta}_{(2)}\Big) \\ &= \Lambda\widehat{\beta}_{(1)} + (I-\Lambda)\widehat{\beta}_{(2)},\end{aligned} \tag{7}$$

which can be viewed as the combined estimator of the pre-break and the post-break estimators, $\widehat{\beta}_{(1)} = (X'_{(1)}X_{(1)})^{-1}(X'_{(1)}Y_{(1)})$ and $\widehat{\beta}_{(2)} = (X'_{(2)}X_{(2)})^{-1}(X'_{(2)}Y_{(2)})$, respectively, with the combination weight $\Lambda = \Big(\frac{w_{(1)}}{w_{(2)}}\ X'_{(1)}X_{(1)} +\ X'_{(2)}X_{(2)}\Big)^{-1}\Big(\frac{w_{(1)}}{w_{(2)}}\ X'_{(1)}X_{(1)}\Big)$.

3 Weighted least squared errors estimator

Li et al. (2013) introduce a semi-parametric method for estimating the parameters of varying-coefficient models with categorical covariates. Also, see Su et al. (2009) and Kiefer and Racine (2017). Inspired by these, we propose an estimator for the time series structural break model where breaks may occur in both the coefficients and the error variance. Specifically, since our interest is in forecasting, we are interested in estimating the coefficients after a break, while also exploiting the pre-break information. We estimate the post-break coefficients using the following discrete kernel weighted generalized least squares estimator (WGLS)

$$\widehat{\beta}_{WGLS}(\gamma) = \underset{\beta_{(2)}}{\operatorname{argmin}} \sum_{t=1}^{T} \Big(\frac{y_t - x'_t\beta_{(2)}}{\sigma_t}\Big)^2\ K\big(t,\gamma\big), \tag{8}$$

where

$$K\big(t,\gamma\big) = \gamma\ \mathbf{1}\big(t \le T_1\big) + \mathbf{1}\big(t > T_1\big), \quad \gamma \in (0,1], \tag{9}$$

is a discrete kernel. $\widehat{\beta}_{WGLS}(\gamma)$ is the estimator of $\beta_{(2)}$ (the post-break parameter value), which is to be used forecasting y_t at $t = T$. As more recent information is usually more relevant for forecasting, this kernel-weight estimator gives the weight 1 to the post-break observations and the weight γ to the pre-break observations. When γ is very close to zero, this estimator is heavily weighting the post-break observations. When $\gamma = 1$, all of the observations in the sample are weighted equally and the estimator is the full-sample estimator. This case is particularly useful under a small break size. As pointed out by Boot and Pick (2020), under a small break, ignoring the break and estimating

the coefficient using all the observations would result in a better forecast (lower MSFE). For other cases, $\gamma \in (0,1)$, we have the combination of the pre-break and post-break observations. We note that (8) depends on the break point and σ_t. However, given the break point, not knowing σ_t turns out not to matter in light of Remark 3 below.

3.1 The weighted squared error estimator is a model averaging estimator

The first order condition for equation (8) is

$$\sum_{t=1}^{T} K(t,\gamma)\ \frac{x_t\left(y_t - x_t'\beta_{(2)}\right)}{\sigma_t^2} = 0, \tag{10}$$

leading to

$$\begin{aligned}
\widehat{\beta}_{WGLS}(\gamma) &= \left(\sum_{t=1}^{T} K(t,\gamma)\ \left(\frac{x_t x_t'}{\sigma_t^2}\right)\right)^{-1} \sum_{t=1}^{T} K(t,\gamma)\ \left(\frac{x_t y_t}{\sigma_t^2}\right) \\
&= \left(\frac{\gamma}{\sigma_{(1)}^2} X_{(1)}'X_{(1)} + \frac{1}{\sigma_{(2)}^2} X_{(2)}'X_{(2)}\right)^{-1} \left(\frac{\gamma}{\sigma_{(1)}^2} X_{(1)}'Y_{(1)} + \frac{1}{\sigma_{(2)}^2} X_{(2)}'Y_{(2)}\right) \\
&= \left(\frac{\gamma}{q^2} X_{(1)}'X_{(1)} + X_{(2)}'X_{(2)}\right)^{-1} \left(\frac{\gamma}{q^2} X_{(1)}'X_{(1)}\ \widehat{\beta}_{(1)} + X_{(2)}'X_{(2)}\ \widehat{\beta}_{(2)}\right) \\
&= \Delta^*\ \widehat{\beta}_{(1)} + (I - \Delta^*)\ \widehat{\beta}_{(2)},
\end{aligned} \tag{11}$$

where $\Delta^* = \left(\gamma^* X_{(1)}'X_{(1)} + X_{(2)}'X_{(2)}\right)^{-1} \left(\gamma^* X_{(1)}'X_{(1)}\right)$, $\gamma^* \equiv \gamma/q^2$, $\widehat{\beta}_{(1)} = \left(X_{(1)}'X_{(1)}\right)^{-1} X_{(1)}'Y_{(1)}$ and $\widehat{\beta}_{(2)} = \left(X_{(2)}'X_{(2)}\right)^{-1} X_{(2)}'Y_{(2)}$.[3] A feasible version of this estimator is

$$\widehat{\beta}_{WGLS}(\widehat{\gamma}) = \widehat{\Delta}^*\ \widehat{\beta}_{(1)} + (I - \widehat{\Delta}^*)\ \widehat{\beta}_{(2)}, \tag{12}$$

for which we first estimate $\widehat{q} = \widehat{\sigma}_{(1)}/\widehat{\sigma}_{(2)}$ by using the least squares estimators of $\sigma_{(1)}^2$ and $\sigma_{(2)}^2$, then we estimate $\widehat{\gamma}$ by the cross-validation using equation (13), and then we let $\widehat{\gamma^*} = \widehat{\gamma}/\widehat{q}^2$.

We estimate $\widehat{\gamma}$ by minimizing the following leave-one-out CV criterion function

$$CV(\gamma) = \frac{1}{(T - T_1)} \sum_{s=T_1+1}^{T} \left(\frac{y_s - x_s'\ \widehat{\beta}_{WGLS}^{(-s)}(\gamma)}{\sigma_{(2)}}\right)^2, \tag{13}$$

[3] It is easy to see that the WGLS estimator is similar to the PPP estimator in equation (4). To show that, let $\Gamma = \text{diag}\left(\gamma I_{T_1}, I_{T-T_1}\right)$. Define $W \equiv \Gamma^{1/2}\Omega^{-1}\Gamma^{1/2}$. Therefore, $\widehat{\beta}_{WGLS}(\gamma) = \left(X'WX\right)^{-1}\left(X'WY\right)$.

where $\widehat{\beta}^{(-s)}_{WGLS}(\gamma)$ denotes the estimate of β based on (11) in which the index s goes over the post-break period, and at each time the s^{th} row of the observations is deleted. We note that the estimation of γ is cross-validated over the post-break observations. The reason is that we need to find an estimator that fits best the post-break sample observations in order to use it for forecasting. Once we have estimated $\widehat{\gamma}$, we calculate the feasible WGLS estimator in equation (12).

Remark 2: In the special case where $\sigma_{(1)} = \sigma_{(2)}$, there is no structural break in the error variance ($q = 1$), the WGLS estimator reduces to the weighted ordinary least squares (WOLS) estimator, which is

$$\widehat{\beta}_{WGLS}(\gamma) = \widehat{\beta}_{WOLS}(\gamma) = \Delta\ \widehat{\beta}_{(1)} + (I - \Delta)\ \widehat{\beta}_{(2)}, \tag{14}$$

where $\Delta = \left(\gamma X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\right)^{-1}\left(\gamma X'_{(1)}X_{(1)}\right)$. Note that $\gamma^* = \gamma$ when $q = 1$. The feasible WOLS estimator is obtained by estimating $\widehat{\gamma}$ using the cross validation in equation (13). □

Remark 3: Using the WGLS estimator, one can estimate only one unknown parameter, γ^*, by CV. In other words, even though the combination weight is in the form of the fraction, γ/q^2, one can consider it as a single unknown parameter, γ^*, and estimate it by CV without dealing with estimation of break in the error variance (q^2).[4] □

Remark 4: By comparing Δ^* in equation (11) with Λ in equation (7), we see that the PPP estimator is a WGLS estimator if

$$\gamma^* = \frac{w_{(1)}}{w_{(2)}} = \frac{1}{q^2 + T_1\phi^2}. \tag{15}$$

In the WGLS method, we only estimate the unknown parameter, γ^*, which depends on the break in the coefficients, ϕ, the break in error variance, q, and T_1. We note that for both WGLS and PPP estimators, the break point needs to be known or estimated. □

3.2 Consistency of the cross-validation

We now prove that the estimator $\widehat{\gamma}$ by CV is asymptotically optimal in the sense of Li (1987), in that the average squared error of the CV is asymptotically as small as the average squared error of

[4]Monte Carlo simulation results show that the results are very close to the two step feasible WGLS method in (12). To save space, we do not report the results in the paper.

the infeasible best possible estimator. Define $L(\gamma) = \left[\left(\widehat{\beta}_{WGLS}(\gamma) - \beta_{(2)}\right)' \mathbb{W} \left(\widehat{\beta}_{WGLS}(\gamma) - \beta_{(2)}\right)\right] = \left(\widehat{\mu}(\gamma) - \mu\right)'\left(\widehat{\mu}(\gamma) - \mu\right)$ to be the loss function where $\mathbb{W} = X'_{(2)}X_{(2)}/\sigma^2_{(2)}$, $\mu = \sigma^{-1}_{(2)}\ X_{(2)}\beta_{(2)}$ and $\widehat{\mu}(\gamma) = \widehat{\sigma}^{-1}_{(2)}\ X_{(2)}\widehat{\beta}_{WGLS}(\gamma)$, and the expected loss (risk) is $R(\gamma) = \mathbb{E}\left[L(\gamma)\right]$. Theorem 1 shows the asymptotic optimality of the weight $\widehat{\gamma}$ estimated by CV in the sense of making $L(\gamma)$ and $R(\gamma)$ as small as possible among all feasible weights γ.

Theorem 1: *As* $T \to \infty$,

$$\frac{L(\hat{\gamma})}{\inf\limits_{\gamma \in \mathcal{H}} L(\gamma)} \xrightarrow{p} 1, \tag{16}$$

$$\frac{R(\hat{\gamma})}{\inf\limits_{\gamma \in \mathcal{H}} R(\gamma)} \xrightarrow{p} 1. \tag{17}$$

■

Theorem 1 shows that conditional on knowing the break date, the average squared error of the WGLS estimator obtained by CV is asymptotically as small as the average squared error of the infeasible best possible estimator. Online Appendix has the full proof of this theorem.

3.3 Efficiency of the WGLS estimator

In this subsection, we compare the bias and variance of the WGLS estimator with the PPP estimator and the post-break estimator. Denote $V_{(1)} = T_1\ Var(\widehat{\beta}_{(1)})$, $V_{(2)} = (T - T_1)\ Var(\widehat{\beta}_{(2)})$, $V_{WGLS} = T\ Var(\widehat{\beta}_{WGLS})$ and $V_{PPP} = T\ Var(\widehat{\beta}_{PPP})$ to be the covariance matrices of the pre-break estimator, the post-break estimator, the WGLS estimator and the PPP estimator, respectively.

The bias of the WGLS estimator and the PPP estimator are

$$Bias(\widehat{\beta}_{WGLS}) = \left(X'_{(1)}X_{(1)} + \frac{1}{\gamma^*}X'_{(2)}X_{(2)}\right)^{-1} X'_{(1)}X_{(1)}\ (\beta_{(1)} - \beta_{(2)}) \tag{18}$$

and

$$Bias(\widehat{\beta}_{PPP}) = \left(X'_{(1)}X_{(1)} + \frac{w_{(2)}}{w_{(1)}}X'_{(2)}X_{(2)}\right)^{-1} X'_{(1)}X_{(1)}\ (\beta_{(1)} - \beta_{(2)}). \tag{19}$$

Thus, the bias of the WGLS estimator is less than that of the PPP estimator if $\gamma^* < \frac{w_{(1)}}{w_{(2)}}$.

Besides, the variance of the WGLS estimator is

$$
\begin{aligned}
Var(\widehat{\beta}_{WGLS}) &= \frac{1}{T}\, V_{WGLS} \\
&= \frac{1}{T}\Big(\gamma b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)^{-1}\Big(\gamma^2 b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)\Big(\gamma b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)^{-1},
\end{aligned} \tag{20}
$$

and the variance of the PPP estimator is

$$
\begin{aligned}
Var(\widehat{\beta}_{PPP}) &= \frac{1}{T}\, V_{PPP} \\
&= \frac{1}{T}\Big(\frac{w_{(1)}}{w_{(2)}} q^2 b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)^{-1}\Big((\frac{w_{(1)}}{w_{(2)}})^2 q^4 b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big) \\
&\times \Big(\frac{w_{(1)}}{w_{(2)}} q^2 b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)^{-1}.
\end{aligned} \tag{21}
$$

Thus, the variance of the WGLS estimator is less than that of the PPP estimator if $\gamma^* > \frac{w_{(1)}}{w_{(2)}}$. This highlights the bias-variance tradeoff feature of the WGLS estimator. When the WGLS estimator assigns a higher weight to the pre-break sample observations, it increases the bias while reducing the variance. The goal of our WGLS method is to obtain the weighted estimator and forecast where the weight is to use the pre-break observations and is determined by cross-validation. When $\gamma^* = \frac{w_{(1)}}{w_{(2)}}$, both estimators are equal as noted in Remark 4.

To compare the variance of the WGLS estimator with the post-break estimator, we note that since $\gamma \in (0, 1]$, using (20), we have

$$
\begin{aligned}
Var(\widehat{\beta}_{WGLS}) &\leq \frac{1}{T}\Big(\gamma b_1 V_{(1)}^{-1} + (1-b_1)V_{(2)}^{-1}\Big)^{-1} \\
&= \frac{1}{T(1-b_1)} V_{(2)}^{1/2}\Big(\frac{\gamma b_1}{1-b_1} V_{(2)}^{1/2} V_{(1)}^{-1} V_{(2)}^{1/2} + I_k\Big)^{-1} V_{(2)}^{1/2}.
\end{aligned} \tag{22}
$$

Also, the variance of the post-break estimator is

$$
Var(\widehat{\beta}_{(2)}) = (\frac{1}{T-T_1})V_{(2)}. \tag{23}
$$

Therefore, using (22) and (23),

$$T\Big(Var(\widehat{\beta}_{(2)}) - Var(\widehat{\beta}_{WGLS})\Big) = \frac{1}{1-b_1} V_{(2)}^{1/2}\Big[I_k - \Big(\frac{\gamma b_1}{1-b_1} V_{(2)}^{1/2} V_{(1)}^{-1} V_{(2)}^{1/2} + I_k\Big)^{-1}\Big] V_{(2)}^{1/2} \\ = \frac{1}{1-b_1} V_{(2)} A^* V_{(2)}, \tag{24}$$

where $A^* \equiv \left(\frac{1-b_1}{\gamma b_1} V_{(1)} + V_{(2)}\right)^{-1}$, and the second equality holds using the Woodbury matrix identity. Thus the WGLS estimator is more efficient than the post-break estimator. As to be expected, the superiority in the efficiency of the WGLS estimator compared to the post-break estimator diminishes as $T(1-b_1) \to \infty$. Besides, the WGLS estimator is biased under a break. This highlights the fact that the WGLS method trade-offs between the bias and variance efficiency. Based on the simulation study and the empirical example results presented in Sections 4 and 5, respectively, the WGLS estimator outperforms the post-break estimator in the sense of mean squared forecast errors, for any break size and break points. This means that the bias created by the WGLS method would be offset by its variance efficiency, and therefore it results in a better forecast.

4 Monte Carlo evidence on forecasting performance

In this section, we use Monte Carlo simulation to compare forecasting performance of the WGLS estimator in equation (11), in comparison with the PPP estimator in equation (7) (labeled as "PPP" in tables), the five methods used in Pesaran and Timmermann (2007), namely the post-break method ("Postbk"); the trade-off method ("Troff"); weighted average of forecasts ("WA"); the pooled forecast combination ("Pooled"); cross validation ("CV"), the full-sample forecast ("Full"), and the average window forecast proposed by Pesaran and Pick (2011) ("AveW"). We also report the results with the infeasible PPP method ("$\text{PPP}_{\text{inf.}}$") in which it uses the true break point and true break size in the coefficient and error variance. We generate data from the following equation,

$$y_t = \begin{cases} x_t'\beta_{(1)} + \sigma_{(1)}\varepsilon_t & \text{if} \quad 1 < t \leq T_1 \\ x_t'\beta_{(2)} + \sigma_{(2)}\varepsilon_t & \text{if} \quad T_1 < t \leq T, \end{cases} \tag{25}$$

where x_t is $k \times 1$, $x_t \sim \text{N}(0, I_k)$, $\varepsilon_t \sim i.i.d.$ $\text{N}(0,1)$, $T = 100$, $q_1 = \frac{\sigma_{(1)}}{\sigma_{(2)}} \in \{0.5, 1, 2\}$ and $k \in \{5, 10\}$. We consider different values for T_1 which are proportional to the pre-break sample observations,

$b_1 \equiv \frac{T_1}{T} \in \{0.2, 0.5, 0.8\}$. Let $\beta_{(2)}$ be a $k \times 1$ vector of ones, and the break size in the coefficients, $\lambda = \beta_{(1)} - \beta_{(2)} \in \{0.1, 0.3, 0.6, 1\}$.[5] To incorporate the uncertainty regarding the unknown parameters (b_1, λ, q), we estimate the break point, break size in the coefficient and the ratio of break in error variance. For the estimation of the break point, we use Bai and Perron (1998, 2003) approach and set the trimming parameter at 0.2.[6] The number of replications is 1000.

Tables 1-3 report the Monte Carlo results. For comparison, we report the one-period ahead forecasts in MSFE for different methods relative to that of the forecast using post-break observations, i.e., for method i we have $RMSFE_i = MSFE_i / MSFE_{\text{Postbk}}$. The MSFE for each method is calculated as $MSFE_i = \frac{1}{1000} \sum_{m=1}^{1000} \left(y_{i,T+1}^{(m)} - \widehat{y}_{i,T+1}^{(m)}\right)^2$, where $\widehat{y}_{i,T+1}^{(m)}$ is the forecast computed using method i for the mth replication. Thus, an RMSFE less than one has a lower MSFE than the post-break forecast, and an RMSFE exceeding one has a higher MSFE than the benchmark. We report the relative MSFE for different methods with estimated break dates in columns 2-8 of Table 1.[7] Columns 9-10 show the results for the average window forecast and the infeasible PPP estimator, respectively.[8]

As it is clear from the tables, the WGLS estimator outperforms the post-break estimator. Specifically, for the small to medium break sizes in the coefficients (approximately $\lambda < 0.6$), there is a huge gain from using the WGLS estimator rather than the post-break estimator. This gain increases even more, as we increase the number of regressors, k. For the large break size in the coefficient, the WGLS estimator becomes close to the post-break estimator, as CV chooses the weight close to zero.

When the break point is close to the end of the sample (b_1 close to 1), the out-performance of the WGLS estimator is much larger than when the break happens close to the beginning of the sample (b_1 close to 0). This is due to the small number of observations in the post-break sample. For example, with $T = 100$, if the break happens at $T_1 = 80$, then the post-break sample has only

[5] We have tried different break size, ranging from 0 to 1 in increments of 0.1. The pattern of results are similar to those that we present in the paper.

[6] The results are similar when trimming of 0.15 is used.

[7] We have also implemented these methods imposing the true break point and the results are qualitatively similar to those with the estimated break point. Because imposing the true break date is infeasible in practice, we only report the results with estimated break point here to save space.

[8] See Sections 3.2-3.5 in Pesaran and Timmermann (2007) for details. For the CV, WA and Pooled methods, as used in Pesaran and Timmermann (2007), we set the size of the forecast evaluation window to $0.25T$ and size of the minimum estimation window to $0.1T$, in the simulation study in this section and the empirical study in the next section.

20 observations. This is when exploiting the pre-break observations can help, and therefore the WGLS estimator performs much better than the post-break estimator. But when break happens at $T_1 = 20$, it does not significantly improve the forecast. This is expected as discussed above based on equation (24), i.e., the gain decreases as b_1 gets smaller.

We also compare the performance of the proposed estimator under different break ratio in the error variance, $q = \sigma_{(1)}/\sigma_{(2)}$, before and after the break. For $q < 1$, the pre-break observations are less volatile than the post-break observations. Thus, including the pre-break observations can benefit the estimator as we are using more information. When $q > 1$, the pre-break observations are more volatile compare to the post-break observations. But as we can see from the tables, even under this case, we still can gain by using the WGLS estimator over the post-break estimator but the gain is bigger when $q < 1$.

Furthermore, we compare the performance of the WGLS estimator with the PPP estimator of Pesaran et al. (2013). Results in Tables 1-3 show that, the WGLS estimator performs better than the PPP estimator for any break points and break size. Besides WGLS estimator overall outperform other methods. The performance of the cross validation (CV) approach deteriorates when T_1 gets close to T. The trade-off (Troff) method performs almost in line with the post-break method. The weighted average (WA) and pooled forecast combination (Pooled) methods perform better than the post-break estimator only for small break sizes. However, they perform poorly for large breaks. Similarly, the AveW forecast perform better than the post-break forecast for small break size. The full-sample method performs good under small break sizes, but as the break size increases, it performs much worse than the post-break estimator. We note that the WGLS, PPP, AveW, CV, Troff, WA, and Pooled methods generally perform better than the full-sample estimator, except for a small break size as expected. This mirrors the findings in Pesaran and Timmermann (2007), Pesaran and Pick (2011) and Pesaran et al. (2013).

We have also included the MSFE for the infeasible PPP method in Tables 1-3, which as expected, it almost always outperforms all other methods. To see how far our estimated optimal CV weight is from the PPP infeasible optimal weights, we plot the frequency of the pre-break and post-break sample observations weights for the PPP infeasible estimator and compare them with that of the normalized WGLS weights.[9] The simulation results are for $T = 100, T_1 = 80, k = 5, q = 1$ and

[9]The normalized pre-break and post-break observations weights for the WGLS estimator are $\gamma/(T_1\gamma + T - T_1)$

presented in Figure 1. Based on the results, the frequency of the WGLS optimal weights are close to that of the infeasible optimal weights. However, since the infeasible version uses the true parameters instead of estimates, and estimation of weights using CV introduces additional noise for the WGLS estimator, the pattern of weights are not exact between the two methods. We observed similar pattern for frequency of weights under other specifications.

5 Empirical analysis

In this section, we present an empirical application of our method. We consider forecasting the equity premium using the six predictors from Welch and Goyal (2008), which are dividend yield, earning-price ratio, stock variance, long term return, inflation and treasury bill rate.[10] The data are quarterly which we examine from 1950:Q1-2018:Q4 to assess the performance of our model. This provides the time-series dimension of 276 observations. The equity premium is the return on the stock market minus the return on a short-term risk-free treasury bill. We use the return on S&P 500 index as the proxy of the stock market return.

We recursively compute one-step-ahead forecasts using the WGLS estimator, and the same alternative methods used in the simulation study in Section 4. We divide the sample of observations into two parts. The first T observations are used as the initial in-sample estimation period, and the remaining observations are the pseudo out-of-sample evaluation period. Forecasts are recursively generated at each point in the out-of-sample period using only the information available at the time the forecast is made. As we expand the estimation window, we re-estimate break points by the sequential procedure introduced by Bai and Perron (1998, 2003), where we search for up to eight breaks and set the trimming parameter to 0.1 (i.e., the minimal permissible length of a regime is $0.1T$) and the significance level at 5%.

We consider two different estimation window sizes so the beginning of the forecast evaluation periods start from 1995:Q1 and 2005:Q1, respectively. Table 4 shows the out-of-sample forecasting results. We evaluate the forecasts for horizons $h = 1, 2, 3, 4$ quarters. The WGLS estimator has lower MSFE than the post-break estimator and the PPP estimator for all forecast horizons. Besides, the WGLS estimator overall performs better than other forecasting methods. However, there are

and $1/(T_1\gamma + T - T_1)$, respectively.

[10] We use the first difference transformation for the dividend yield, earning price ratio and treasury bill rate.

some cases that other forecasting methods underperform the post-break forecast.

Finally, we evaluate the statistical significance of superior predictive ability of method i by testing for the null hypothesis $H_0 : MSFE_{\text{Postbk}} = MSFE_i$ against the alternative hypothesis $H_1 : MSFE_{\text{Postbk}} > MSFE_i$. The Diebold and Mariano (1995) statistics reported in Table 4, show that the WGLS estimator produces forecasts with significantly smaller MSFEs than the post-break estimator does for all forecast horizons $h = 1, 2, 3, 4$. MSFEs from using the PPP estimator are not significantly smaller than those of the post-break estimator.

6 Conclusion

In this paper, we introduce the novel WGLS method for forecasting under structural breaks. As it uses the discrete kernel function instead of the parametrically derived weights as in the PPP method, WGLS facilitates implementation of the PPP method. The discrete kernel weight is estimated by cross-validation, which we show is asymptotically optimal in the sense of Li (1987), i.e., the average squared error of the loss or risk estimated by cross-validation is asymptotically as small as that of the oracle estimator. Monte Carlo simulation illustrates the superiority of the WGLS forecast relative to the post-break forecast, the PPP forecast, and other methods. When the breaks happen closer to end of the sample or when there are larger number of regressors, the WGLS method outperforms them even more. The empirical application of predicting equity premium is presented which shows the superiority of the WGLS method over other forecasting methods.

References

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*. New York: Cambridge University Press.

Aitchison, J. and Aitken, C.G.G. (1976). "Multivariate binary discrimination by the kernel method." *Biometrika* 63, 413–420.

Bai, J. (1997a). "Estimation of a change point in multiple regression models." Review of Economics and Statistics 79, 551-563.

Bai, J. and Perron, P. (1998). "Estimating and testing linear models with multiple structural changes." *Econometrica* 66, 47-78.

Bai, J. and Perron, P. (2003). "Computation and analysis of multiple structural change models." *Journal of Applied Econometrics* 18, 1-22.

Boot, T. and Pick, A. (2020). "Does modeling a structural break improve forecast accuracy?" *Journal of Econometrics* 215, 35-59.

Clements, M.P. and Hendry, D.F. (2006). "Forecasting with breaks." In Elliott, G., C.W.J. Granger and A. Timmermann (Eds.), *Handbook of Economic Forecasting* vol. 1, 605-658. North-Holland.

Clements, M. P. and Hendry, D.F. (2011). *The Oxford Handbook of Economic Forecasting.* Oxford University Press.

Diebold, F. X. and Mariano, R. S. (1995). "Comparing predictive accuracy." *Journal of Business & Economic Statistics* 13, 253-263.

Geisser, S. (1974). "The predictive sample reuse method with applications." *Journal of the American Statistical Association* 70, 320–328.

Hansen, B. (2007). "Least squares model averaging." *Econometrica* 75, 1175–1189.

Hansen, B. (2008). "Least squares forecast averaging." *Journal of Econometrics* 146, 342–350.

Hansen, B. and Racine, J. (2012). "Jackknife model averaging." *Journal of Econometrics* 167, 38–46.

Hjort, N. and Claeskens, G. (2003). "Frequentist model average estimators." *Journal of the American Statistical Association* 98, 879–899.

Kiefer, N.M. and Racine, J.S. (2017). "The smooth colonel and the reverend find common ground." *Econometric Reviews* 36, 241-256.

Li, K.-C. (1987). "Asymptotic optimality for C_P, C_L, cross-validation and generalized cross-validation: discrete index set." *The Annals of Statistics* 15, 958-975.

Li, Q., Ouyang, D. and Racine, J.S. (2013). "Categorical semiparametric varying coefficient models." *Journal of Applied Econometrics* 28, 551–579.

Pesaran, M.H. and Pick, A. (2011). "Forecast combination across estimation windows." *Journal of Business & Economic Statistics* 29, 307–318.

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

Pesaran, M.H. and Timmermann, A. (2002). "Market timing and return prediction under model instability." *Journal of Empirical Finance* 9, 495–510.

Pesaran, M.H. and Timmermann, A. (2005). "Small sample properties of forecasts from autoregressive models under structural breaks." *Journal of Econometrics* 129, 183-217.

Pesaran, M.H. and Timmermann, A. (2007). "Selection of estimation window in the presence of breaks." *Journal of Econometrics* 137, 134-161.

Rossi, B. (2013). "Advances in forecasting under instability." In G. Elliott, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, volume 2, Part B, 1203 - 1324. Elsevier.

Timmerman, Allan (2006). "Forecast combinations." In G. Elliott, C. W. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, vol. 1, 135-196. Elsevier.

Stone, C. J. (1974). "Cross-validatory choice and assessment of statistical predictions (with discussion)." *Journal of the Royal Statistical Society* Series B 36, 111–147.

Su, L., Chen, Y. and Ullah, A. (2009). "Functional coefficient estimation with both categorical and continuous data." *Advances in Econometrics* 25, 131-167.

Su, L., Murtazashvili, I. and Ullah, A. (2013). "Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling." *Journal of Business & Economic Statistics* 31, 184-207.

Welch, I. and Goyal, A. (2008). "A Comprehensive look at the empirical performance of equity premium prediction." *Review of Financial Studies* 21, 1455-508.

Zhang, X., Wan, A. T. K. and Zou, G. (2013). "Model averaging by Jackknife criterion in models with dependent data." *Journal of Econometrics* 174, 82–94.

A Appendix: Mathematical details

A.1 Proof of Theorem 1

To derive $L(\gamma)$ and $R(\gamma)$, rewrite (14) as

$$\widehat{\beta}_{WGLS}(\gamma) - \beta_{(2)} = \Big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\Big)^{-1}\Big(\gamma^*\ X'_{(1)}X_{(1)}\lambda + \gamma^*\ X'_{(1)}\sigma_{(1)}\epsilon_{(1)} + X'_{(2)}\sigma_{(2)}\epsilon_{(2)}\Big), \tag{A.1}$$

where $\lambda = \beta_{(1)} - \beta_{(2)}$, $\epsilon_{(1)} = (\varepsilon_1, \ldots, \varepsilon_{T_1})'$, $\epsilon_{(2)} = (\varepsilon_{T_1+1}, \ldots, \varepsilon_T)'$. Then,

$$\begin{aligned}
R(\gamma) &= \mathbb{E}\left[\Big(\widehat{\beta}_{WGLS}(\gamma) - \beta_{(2)}\Big)'\ \mathbb{W}\ \Big(\widehat{\beta}_{WGLS}(\gamma) - \beta_{(2)}\Big)\right] \\
&= \mathbb{E}\Bigg[\Big(\gamma^*\ \lambda' X'_{(1)}X_{(1)} + \gamma^*\ \sigma_{(1)}\epsilon'_{(1)}X_{(1)} + \sigma_{(2)}\epsilon'_{(2)}X_{(2)}\Big)\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W} \\
&\times \big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\Big(\gamma^*\ X'_{(1)}X_{(1)}\lambda + \gamma^*\ X'_{(1)}\sigma_{(1)}\epsilon_{(1)} + X'_{(2)}\sigma_{(2)}\epsilon_{(2)}\Big)\Bigg] \\
&= (\gamma^*)^2\ \lambda' X'_{(1)}X_{(1)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}X_{(1)}\lambda \\
&+ (\gamma^*)^2\ \mathbb{E}\left[\sigma^2_{(1)}\ \epsilon'_{(1)}X_{(1)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}\epsilon_{(1)}\right] \\
&+ \mathbb{E}\left[\sigma^2_{(2)}\ \epsilon'_{(2)}X_{(2)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(2)}\epsilon_{(2)}\right] \\
&= \lambda' A_1\lambda + (\gamma^*)^2\ \sigma^2_{(1)}\ \mathrm{tr}(A_2) + \sigma^2_{(2)}\ \mathrm{tr}(A_3),
\end{aligned} \tag{A.2}$$

where $\gamma^* = \gamma/q^2$, $A_1 = (\gamma^*)^2\ X'_{(1)}X_{(1)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\ \big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}X_{(1)}$, $A_2 = X'_{(1)}X_{(1)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}$ and $A_3 = X'_{(2)}X_{(2)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\mathbb{W}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}$. Since λ is the break size which is not known, we can plug the unbiased estimator for this term instead. The unbiased estimator for $\lambda' A_1\lambda$ is equal to

$$\widehat{\lambda}' A_1\widehat{\lambda} - \widehat{\sigma}^2_{(1)}\ \mathrm{tr}\Big(A_1(X'_{(1)}X_{(1)})^{-1}\Big) - \widehat{\sigma}^2_{(2)}\ \mathrm{tr}\Big(A_1(X'_{(2)}X_{(2)})^{-1}\Big), \tag{A.3}$$

where $\widehat{\sigma}^2_{(1)}$ and $\widehat{\sigma}^2_{(2)}$ are an unbiased estimator for the variance of the pre-break and post-break observations, respectively. By plugging (A.3) back into (A.2) we have:

$$\begin{aligned}\widehat{R}(\gamma) &= \widehat{\lambda}' A_1 \widehat{\lambda} + \widehat{\sigma}^2_{(1)}\ \mathrm{tr}\Big((\gamma^*)^2 A_2 - A_1 (X'_{(1)} X_{(1)})^{-1}\Big) + \widehat{\sigma}^2_{(2)}\ \mathrm{tr}\Big(A_3 - A_1 (X'_{(2)} X_{(2)})^{-1}\Big) \\ &= \widehat{\lambda}' A_1 \widehat{\lambda} + 2\widehat{\sigma}^2_{(2)}\ \mathrm{tr}\Big(\mathbb{W}\left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1}\Big) - k,\end{aligned} \tag{A.4}$$

where $\mathbb{W} = X'_{(2)} X_{(2)}/\sigma^2_{(2)}$ and $(\gamma^*)^2 A_2 - A_1 (X'_{(1)} X_{(1)})^{-1} = 0$. Also, $A_3 - A_1 (X'_{(2)} X_{(2)})^{-1} = 2\ \mathrm{tr}\Big(\mathbb{W}\left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1}\Big) - k/\sigma^2_{(2)}$. Besides, we can rewrite $\widehat{\lambda}' A_1 \widehat{\lambda}$ as

$$\begin{aligned}\widehat{\lambda}' A_1 \widehat{\lambda} &= \big(\widehat{\beta}_{(1)} - \widehat{\beta}_{(2)}\big)' A_1 \big(\widehat{\beta}_{(1)} - \widehat{\beta}_{(2)}\big) \\ &= \widehat{\beta}'_{(1)} A_1 \widehat{\beta}_{(1)} + \widehat{\beta}'_{(2)} A_1 \widehat{\beta}_{(2)} - 2\widehat{\beta}'_{(2)} A_1 \widehat{\beta}_{(1)} \\ &= (\gamma^*)^2\ Y'_{(1)} X_{(1)} \big(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\big)^{-1} \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} X'_{(1)} Y_{(1)} + \widehat{\beta}'_{(2)}\, \mathbb{W}\, \widehat{\beta}_{(2)} \\ &- 2\widehat{\beta}'_{(2)}\, \mathbb{W}\, \widehat{\beta}_{WGLS}(\gamma) + 2\widehat{\beta}'_{(2)}\, \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} \gamma^* X'_{(1)} Y_{(1)} + \widehat{\beta}'_{WGLS}(\gamma)\, \mathbb{W}\, \widehat{\beta}_{WGLS}(\gamma) \\ &- 2\widehat{\beta}'_{WGLS}(\gamma)\, \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} \gamma^* X'_{(1)} Y_{(1)} \\ &+ (\gamma^*)^2\ Y'_{(1)} X_{(1)} \big(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\big)^{-1} \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} X'_{(1)} Y_{(1)} \\ &- 2\widehat{\beta}'_{(2)}\, \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} \gamma^* X'_{(1)} Y_{(1)} + 2\widehat{\beta}'_{WGLS}(\gamma)\, \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} \gamma^* X'_{(1)} Y_{(1)} \\ &- 2(\gamma^*)^2\ Y'_{(1)} X_{(1)} \big(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\big)^{-1} \mathbb{W} \left(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\right)^{-1} X_{(1)} Y_{(1)} \\ &= \widehat{\beta}'_{(2)}\, \mathbb{W}\, \widehat{\beta}_{(2)} - 2\widehat{\beta}'_{(2)}\, \mathbb{W}\, \widehat{\beta}_{WGLS}(\gamma) + \widehat{\beta}'_{WGLS}(\gamma)\, \mathbb{W}\, \widehat{\beta}_{WGLS}(\gamma) \\ &= \sigma^{-2}_{(2)}\ \widehat{\beta}'_{(2)}\ X'_{(2)} X_{(2)}\ \widehat{\beta}_{(2)} - 2\ \sigma^{-2}_{(2)}\ Y'_{(2)} X_{(2)}\ \widehat{\beta}_{WGLS}(\gamma) + \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)} + \sigma^{-1}_{(2)}\ Y_{(2)}\Big)' \\ &\times \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)} + \sigma^{-1}_{(2)}\ Y_{(2)}\Big) \\ &= \sigma^{-2}_{(2)}\ \widehat{\beta}'_{(2)}\ X'_{(2)} X_{(2)}\ \widehat{\beta}_{(2)} + \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)}\Big)' \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)}\Big) - \sigma^{-2}_{(2)}\ Y'_{(2)} Y_{(2)},\end{aligned} \tag{A.5}$$

where we plug $\mathbb{W} = X'_{(2)} X_{(2)}/\sigma^2_{(2)}$ to derive the MSFE.

Thus, based on equation (A.4), the MSFE for the WGLS estimator up to the relevant terms to γ is

$$\widehat{R}(\gamma) = \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)}\Big)' \Big(\widehat{\mu}(\gamma) - \sigma^{-1}_{(2)}\ Y_{(2)}\Big) + 2\ \mathrm{tr}\Big(X'_{(2)} X_{(2)} \big(\gamma^* X'_{(1)} X_{(1)} + X'_{(2)} X_{(2)}\big)^{-1}\Big). \tag{A.6}$$

Define $P(\gamma) \equiv X_{(2)}\Big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\Big)^{-1}X'_{(2)}$. Thus, we can rewrite $\widehat{\mu}(\gamma)$ as

$$\begin{aligned}\widehat{\mu}(\gamma) &= \sigma_{(2)}^{-1}\ X_{(2)}\ \widehat{\beta}_{WGLS}(\gamma) \\ &= \sigma_{(2)}^{-1}\ X_{(2)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\big(\gamma^* X'_{(1)}Y_{(1)} + X'_{(2)}Y_{(2)}\big) \\ &= \sigma_{(2)}^{-1}\Big(\Phi + P(\gamma)Y_{(2)}\Big),\end{aligned} \tag{A.7}$$

where $\Phi \equiv X_{(2)}\big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\big(\gamma^* X'_{(1)}Y_{(1)}\big)$. Based on this definition, we can rewrite the asymptotic risk in (A.6) as

$$\begin{aligned}\widehat{R}(\gamma) &= \Big(\widehat{\mu}(\gamma) - \mu - \epsilon_{(2)}\Big)'\Big(\widehat{\mu}(\gamma) - \mu - \epsilon_{(2)}\Big) + 2\ \mathrm{tr}\big(P(\gamma)\big) \\ &= L(\gamma) + \epsilon'_{(2)}\epsilon_{(2)} - 2\ \epsilon'_{(2)}\Big(\widehat{\mu}(\gamma) - \mu\Big) + 2\ \mathrm{tr}\big(P(\gamma)\big) \\ &= L(\gamma) + \epsilon'_{(2)}\epsilon_{(2)} - 2\ \epsilon'_{(2)}\ \widehat{\mu}(\gamma) + 2\ \epsilon'_{(2)}\mu + 2\ \mathrm{tr}\big(P(\gamma)\big) \\ &= L(\gamma) + \epsilon'_{(2)}\epsilon_{(2)} - 2\sigma_{(2)}^{-1}\ \epsilon'_{(2)}\ \big(\Phi + P(\gamma)Y_{(2)}\big) + 2\ \epsilon'_{(2)}\mu + 2\ \mathrm{tr}\big(P(\gamma)\big) \\ &= L(\gamma) + \epsilon'_{(2)}\epsilon_{(2)} - 2\sigma_{(2)}^{-1}\ \epsilon'_{(2)}\ \Phi - 2\ \epsilon'_{(2)}\ P(\gamma)\mu - 2\ \epsilon'_{(2)}P(\gamma)\ \epsilon_{(2)} + 2\ \epsilon'_{(2)}\ \mu + 2\ \mathrm{tr}\big(P(\gamma)\big).\end{aligned} \tag{A.8}$$

We consider the choice of γ from minimizing (A.8), and by restricting γ to the set $\mathcal{H}$. Thus, the selected γ is

$$\widehat{\gamma} = \underset{\gamma\in\mathcal{H}}{\mathrm{argmin}}\ \widehat{R}(\gamma). \tag{A.9}$$

Let $\xi = \inf_{\gamma\in\mathcal{H}} R(\gamma) \to \infty$ as $T \to \infty$. This specifies that the bias of the WGLS estimator is not zero, and thus the trade-off between bias and variance is present. This condition is similar to that of Li (1987), Hansen (2007) for Mallows weight selection, Hansen and Racine (2012) and Zhang et al. (2013) for Jackknife model averaging. We assume that

$$\mu'\mu = O(T), \tag{A.10}$$

$$X'_{(2)}\sigma_{(2)}\epsilon_{(2)} = O_p(\sqrt{T}). \tag{A.11}$$

As shown by Theorem 2.1 of Li (1987), the estimated weight, $\widehat{\gamma}$, obtained from CV is asymptotically

optimal for the WGLS risk $\widehat{R}(\gamma)$, that is

$$\frac{\widehat{R}(\widehat{\gamma})}{\inf\limits_{\gamma\in\mathcal{H}}\widehat{R}(\gamma)} \xrightarrow{p} 1, \tag{A.12}$$

under $\lambda_{\max}(\Omega) < \infty$ (moment bound condition), and $\lim\limits_{T\to\infty} \operatorname{Sup}\limits_{\gamma\in\mathcal{H}} \lambda_{\max}\big(P(\gamma)\big) < \infty$, where we define $\lambda_{\max}(B)$ to be the largest eigenvalue of a matrix B. We note that

$$\operatorname*{Sup}_{\gamma\in\mathcal{H}} \lambda_{\max}\big(P(\gamma)\big) = \operatorname*{Sup}_{\gamma\in\mathcal{H}} \lambda_{\max}\Big(X_{(2)}\big(\gamma X'_{(1)}X_{(1)}+X'_{(2)}X_{(2)}\big)^{-1}X'_{(2)}\Big) \leq \lambda_{\max}\Big(X_{(2)}(X'_{(2)}X_{(2)})^{-1}X'_{(2)}\Big) = 1. \tag{A.13}$$

Also notice that

$$\begin{aligned}
R(\gamma) &= \mathbb{E}\left[\Big(\widehat{\mu}(\gamma)-\mu\Big)'\Big(\widehat{\mu}(\gamma)-\mu\Big)\right] \\
&= \mathbb{E}\left[\Big(\sigma_{(2)}^{-1}\ \Phi+\sigma_{(2)}^{-1}\ P(\gamma)Y_{(2)}-\mu\Big)'\Big(\sigma_{(2)}^{-1}\ \Phi+\sigma_{(2)}^{-1}\ P(\gamma)Y_{(2)}-\mu\Big)\right] \\
&= \mathbb{E}\left[\Big(\sigma_{(2)}^{-1}\ \Phi+P(\gamma)\mu+P(\gamma)\ \epsilon_{(2)}-\mu\Big)'\Big(\sigma_{(2)}^{-1}\ \Phi+P(\gamma)\mu+P(\gamma)\ \epsilon_{(2)}-\mu\Big)\right] \\
&= \mathbb{E}\left[\sigma_{(2)}^{-2}\ \Phi'\Phi+\epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)}+\big(P(\gamma)\ \mu-\mu\big)'\big(P(\gamma)\ \mu-\mu\big)+2\sigma_{(2)}^{-1}\ \Phi'\big(P(\gamma)\ \mu-\mu\big)\right] \\
&= \mathbb{E}\left[\Phi'\sigma_{(2)}^{-2}\ \Phi\right]+\operatorname{tr}\Big(P^2(\gamma)\Big)+\big(P(\gamma)\ \mu-\mu\big)'\big(P(\gamma)\ \mu-\mu\big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= \Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)-P^2(\gamma)\big)+\operatorname{tr}\big(P^2(\gamma)\big)+\big(P(\gamma)\mu-\mu\big)'\big(P(\gamma)\mu-\mu\big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= \Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)\big)+\big(1-q^2\big)\ \operatorname{tr}\big(P^2(\gamma)\big)+\Big(P(\gamma)\ \big(\mu+\epsilon_{(2)}-\epsilon_{(2)}\big)-\mu\Big)' \\
&\quad \Big(P(\gamma)\ \big(\mu+\epsilon_{(2)}-\epsilon_{(2)}\big)-\mu\Big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= \Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)\big)+\big(1-q^2\big)\ \operatorname{tr}\big(P^2(\gamma)\big)+\Big(\sigma_{(2)}^{-1}\ P(\gamma)Y_{(2)}+\sigma_{(2)}^{-1}\ \Phi-\mu-\sigma_{(2)}^{-1}\ \Phi-P(\gamma)\epsilon_{(2)}\Big)' \\
&\quad \Big(\sigma_{(2)}^{-1}\ P(\gamma)Y_{(2)}+\sigma_{(2)}^{-1}\ \Phi-\mu-\sigma_{(2)}^{-1}\ \Phi-P(\gamma)\epsilon_{(2)}\Big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= \Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)\big)+\big(1-q^2\big)\ \operatorname{tr}\big(P^2(\gamma)\big)+\Big(\widehat{\mu}(\gamma)-\mu-\sigma_{(2)}^{-1}\ \Phi-P(\gamma)\epsilon_{(2)}\Big)' \\
&\quad \Big(\widehat{\mu}(\gamma)-\mu-\sigma_{(2)}^{-1}\ \Phi-P(\gamma)\epsilon_{(2)}\Big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= L(\gamma)+\Phi'\sigma_{(2)}^{-2}\ \Phi+\epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)}-2\big(\widehat{\mu}(\gamma)-\mu\big)'\big(\sigma_{(2)}^{-1}\ \Phi+P(\gamma)\epsilon_{(2)}\big)+2\epsilon'_{(2)}P(\gamma)\sigma_{(2)}^{-1}\Phi \\
&\quad +\Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)\big)+\big(1-q^2\big)\ \operatorname{tr}\big(P^2(\gamma)\big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi \\
&= L(\gamma)+\Phi'\sigma_{(2)}^{-2}\ \Phi+\epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)}-2\Big(\sigma_{(2)}^{-1}\ \Phi+P(\gamma)\big(\mu+\epsilon_{(2)}\big)-\mu\Big)'\Big(\sigma_{(2)}^{-1}\ \Phi+P(\gamma)\epsilon_{(2)}\Big) \\
&\quad +2\epsilon'_{(2)}P(\gamma)\sigma_{(2)}^{-1}\Phi+\Upsilon+q^2\ \operatorname{tr}\big(P(\gamma)\big)+\big(1-q^2\big)\ \operatorname{tr}\big(P^2(\gamma)\big)+2\big(P(\gamma)\ \mu-\mu\big)'\ \Pi
\end{aligned}$$

$$= L(\gamma) - \Psi - \epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)} - 2\sigma_{(2)}^{-1}\ \Phi' P(\gamma)\epsilon_{(2)} - 2\Big(P(\gamma)\mu - \mu\Big)'\Big(\sigma_{(2)}^{-1}\ \Phi + P(\gamma)\epsilon_{(2)} - \Pi\Big)$$
$$+ q^2\ \mathrm{tr}\big(P(\gamma)\big) + \big(1-q^2\big)\ \mathrm{tr}\big(P^2(\gamma)\big)$$
$$= L(\gamma) - \Psi - \epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)} - 2\sigma_{(2)}^{-1}\ \Phi' P(\gamma)\epsilon_{(2)} - 2\Big(P(\gamma)\mu - \mu\Big)'\Big(\Theta + P(\gamma)\epsilon_{(2)}\Big)$$
$$+ q^2\ \mathrm{tr}\big(P(\gamma)\big) + \big(1-q^2\big)\ \mathrm{tr}\big(P^2(\gamma)\big), \tag{A.14}$$

where $\Phi'\sigma_{(2)}^{-2}\ \Phi = \Upsilon + \Psi$, $\Upsilon \equiv (\gamma^*)^2\sigma_{(2)}^{-2}\ \beta'_{(1)}X'_{(1)}X_{(1)}\big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(2)}X_{(2)}\ \big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}X_{(1)}\beta_{(1)}$, $\Psi \equiv (\gamma^*)^2q^2\ \epsilon'_{(1)}X_{(1)}\big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(2)}X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}\epsilon_{(1)}$, $\Pi = \gamma^*\sigma_{(2)}^{-1}X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\ X'_{(1)}X_{(1)}\beta_{(1)}$, and
$\Theta = \sigma_{(2)}^{-1}X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\big)^{-1}\gamma^*X'_{(1)}\sigma_{(1)}\epsilon_{(1)}$.

By using the conditions in (A.10) and (A.11), and the proof steps of Theorem 2.2 of Zhang et al. (2013), in order to establish (A.12) and therefore the optimality condition in Theorem 1, it suffices to prove that the following conditions hold.

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\epsilon'_{(2)}\Phi|}{R(\gamma)} = o_p(1), \tag{A.15}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\epsilon'_{(2)}P(\gamma)\epsilon_{(2)}|}{R(\gamma)} = o_p(1), \tag{A.16}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\ \mu' P(\gamma)\epsilon_{(2)}|}{R(\gamma)} = o_p(1), \tag{A.17}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\mathrm{tr}(P(\gamma))|}{R(\gamma)} = o_p(1), \tag{A.18}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\Psi|}{R(\gamma)} = o_p(1), \tag{A.19}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)}|}{R(\gamma)} = o_p(1), \tag{A.20}$$

$$\underset{\gamma\in\mathcal{H}}{\mathrm{Sup}}\ \frac{|\Phi' P(\gamma)\epsilon_{(2)}|}{R(\gamma)} = o_p(1), \tag{A.21}$$

$$\underset{\gamma\in\mathcal{H}}{\text{Sup}}\ \frac{|(P(\gamma)\mu-\mu)'\ \Theta|}{R(\gamma)}=o_p(1), \tag{A.22}$$

$$\underset{\gamma\in\mathcal{H}}{\text{Sup}}\ \frac{|(P(\gamma)\mu-\mu)'\ P(\gamma)\epsilon_{(2)}|}{R(\gamma)}=o_p(1), \tag{A.23}$$

$$\underset{\gamma\in\mathcal{H}}{\text{Sup}}\ \frac{|\text{tr}(P^2(\gamma))|}{R(\gamma)}=o_p(1). \tag{A.24}$$

Proof of equation (A.15):

$$\begin{aligned}\epsilon'_{(2)}\Phi &= \gamma^*\ \epsilon'_{(2)}X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)}+X'_{(2)}X_{(2)}\big)^{-1}X'_{(1)}Y_{(1)}\\ &\leq \gamma^*\ \epsilon'_{(2)}X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)}\big)^{-1}X'_{(1)}Y_{(1)}\\ &= \epsilon'_{(2)}X_{(2)}\big(X'_{(1)}X_{(1)}\big)^{-1}X'_{(1)}Y_{(1)}\\ &= \epsilon'_{(2)}X_{(2)}\ \widehat{\beta}_{(1)}=O_p(\sqrt{T}).\end{aligned} \tag{A.25}$$

Proof of equation (A.16):

$$\underset{\gamma\in\mathcal{H}}{\text{Sup}}\ \epsilon'_{(2)}P(\gamma)\epsilon_{(2)}\leq \epsilon'_{(2)}\ X_{(2)}(X'_{(2)}X_{(2)})^{-1}X'_{(2)}\ \epsilon_{(2)}=O_p(1). \tag{A.26}$$

Proof of equation (A.17):

$$\begin{aligned}\mu'P(\gamma)\epsilon_{(2)} &= \mu'X_{(2)}\big(\gamma^*X'_{(1)}X_{(1)}+X'_{(2)}X_{(2)}\big)^{-1}X'_{(2)}\epsilon_{(2)}\\ &\leq \mu'X_{(2)}(X'_{(2)}X_{(2)})^{-1}X'_{(2)}\epsilon_{(2)}\\ &= \sigma^{-1}_{(2)}\ \beta'_{(2)}X'_{(2)}X_{(2)}(X'_{(2)}X_{(2)})^{-1}X'_{(2)}\epsilon_{(2)}\\ &= \sigma^{-1}_{(2)}\ \beta'_{(2)}X'_{(2)}\epsilon_{(2)}=O_p(\sqrt{T}).\end{aligned} \tag{A.27}$$

Proof of equation (A.18):

$$\underset{\gamma\in\mathcal{H}}{\text{Sup}}\ \text{tr}\big(P(\gamma)\big)\leq \text{tr}\big(X'_{(2)}(X'_{(2)}X_{(2)})^{-1}X_{(2)}\big)=k=O(1). \tag{A.28}$$

Proof of equation(A.19):

$$\begin{aligned}\Psi &= (\gamma^*)^2 q^2\ \epsilon'_{(1)}X_{(1)}\Big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\Big)^{-1} X'_{(2)}X_{(2)}\ \Big(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)}\Big)^{-1} X'_{(1)}\epsilon_{(1)} \\ &\leq (\gamma^*)^2 q^2\ \epsilon'_{(1)}X_{(1)}\Big(\gamma^* X'_{(1)}X_{(1)}\Big)^{-1} X'_{(2)}X_{(2)}\ \Big(\gamma^* X'_{(1)}X_{(1)}\Big)^{-1} X'_{(1)}\epsilon_{(1)} \\ &= \epsilon'_{(1)}\sigma_{(1)}X_{(1)}\ \big(X'_{(1)}X_{(1)}\big)^{-1}\ X'_{(2)}\sigma^{-2}_{(2)}X_{(2)}\ \big(X'_{(1)}X_{(1)}\big)^{-1} X'_{(1)}\sigma_{(1)}\epsilon_{(1)} \\ &= \Big(\widehat{\beta}_{(1)} - \beta_{(1)}\Big)' X'_{(2)}\sigma^{-2}_{(2)}X_{(2)}\ \Big(\widehat{\beta}_{(1)} - \beta_{(1)}\Big) = O_p(1).\end{aligned} \tag{A.29}$$

Proof of equation (A.20):

$$\underset{\gamma\in H}{\text{Sup}}\ \epsilon'_{(2)}P^2(\gamma)\epsilon_{(2)} \leq\ \underset{\gamma\in H}{\text{Sup}}\ \lambda_{\max}\big(P(\gamma)\big)\ \epsilon'_{(2)}P(\gamma)\epsilon_{(2)} \leq \epsilon'_{(2)}X_{(2)}(X'_{(2)}X_{(2)})^{-1}X'_{(2)}\epsilon_{(2)} = O_p(1), \tag{A.30}$$

where the first inequality holds because for any $n\times 1$ vector of $\mathbf{x}$, and any $n\times n$ symmetric matrix A with maximum eigenvalue of $\lambda_{\max}(A)$, we have $\mathbf{x}'A\mathbf{x} \leq \mathbf{x}'\mathbf{x}\ \lambda_{\max}(A)$, see Abadir and Magnus (2005), page 343.

Proof of equation (A.21):

$$\begin{aligned}\Phi' P(\gamma)\epsilon_{(2)} &= \gamma^*\ Y'_{(1)}X_{(1)}(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)})^{-1}\ X'_{(2)}X_{(2)}(\gamma^* X'_{(1)}X_{(1)} + X'_{(2)}X_{(2)})^{-1}X'_{(2)}\epsilon_{(2)} \\ &\leq \gamma^*\ Y'_{(1)}X_{(1)}(X'_{(2)}X_{(2)})^{-1}\ X'_{(2)}X_{(2)}\ \Big(\gamma^* X'_{(1)}X_{(1)}\Big)^{-1} X'_{(2)}\epsilon_{(2)} \\ &= Y'_{(1)}X_{(1)}(X'_{(1)}X_{(1)})^{-1}X'_{(2)}\epsilon_{(2)} \\ &= \widehat{\beta}_1\ X'_{(2)}\epsilon_{(2)} = O_p(\sqrt{T}).\end{aligned} \tag{A.31}$$

Proof of equation (A.22):

Using the Cauchy-Schwarz inequality we have:

$$\begin{aligned}\Big[\Big(P(\gamma)\mu - \mu\Big)'\Theta\Big]^2 &\leq \Big(P(\gamma)\mu - \mu\Big)'\Big(P(\gamma)\mu - \mu\Big)\Theta'\Theta \\ &\leq R(\gamma)\ q^2\epsilon'_{(1)}X_{(1)}\big(X'_{(1)}X_{(1)}\big)^{-1}\big(X'_{(2)}X_{(2)}\big)\big(X'_{(1)}X_{(1)}\big)^{-1}X'_{(1)}\epsilon_{(1)} \\ &= O_p(T).\end{aligned} \tag{A.32}$$

Therefore, $\Big(P(\gamma)\mu - \mu\Big)'\Theta = O_p(\sqrt{T})$.

Proof of equation (A.23):

Using the Cauchy-Schwarz inequality we have:

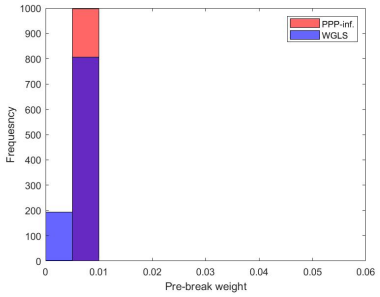
$$\begin{aligned}\left[\Big(P(\gamma)\mu-\mu\Big)'P(\gamma)\epsilon_{(2)}\right]^2 &\leq \Big(P(\gamma)\mu-\mu\Big)'\Big(P(\gamma)\mu-\mu\Big)\epsilon_{(2)}'P(\gamma)^2\epsilon_{(2)} \\ &\leq R(\gamma)\ \lambda_{\max}\big(P(\gamma)\big)\ \epsilon_{(2)}'P(\gamma)\epsilon_{(2)} = O_p(T).\end{aligned} \tag{A.33}$$

Therefore, $\Big(P(\gamma)\mu-\mu\Big)'P(\gamma)\epsilon_{(2)} = O_p(\sqrt{T})$.

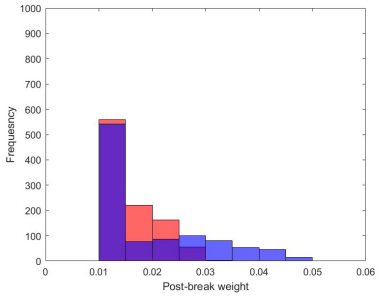
Proof of equation (A.24):

$$\begin{aligned}\mathrm{tr}\Big(P^2(\gamma)\Big) &= \mathrm{tr}\Big[X_{(2)}\big(\gamma^*X_{(1)}'X_{(1)}+X_{(2)}'X_{(2)}\big)^{-1}X_{(2)}'X_{(2)}\ \big(\gamma^*X_{(1)}'X_{(1)}+X_{(2)}'X_{(2)}\big)^{-1}X_{(2)}'\Big] \\ &\leq \mathrm{tr}\Big[X_{(2)}(X_{(2)}'X_{(2)})^{-1}X_{(2)}'X_{(2)}\ (X_{(2)}'X_{(2)})^{-1}X_{(2)}'\Big] = k = O(1).\end{aligned} \tag{A.34}$$

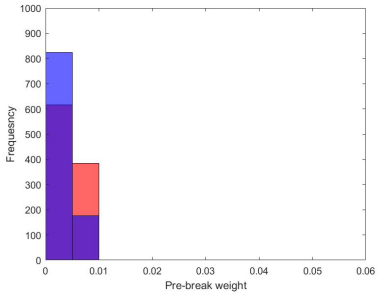
Based on the above conditions, the proof of Theorem 1 thus follows. Therefore, the weight derives by CV is optimal. ■



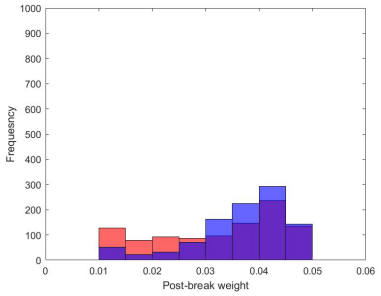
(a) $\lambda = 0.1$



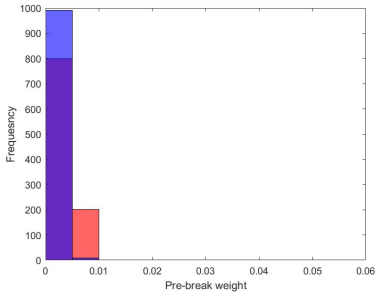
(b) $\lambda = 0.1$



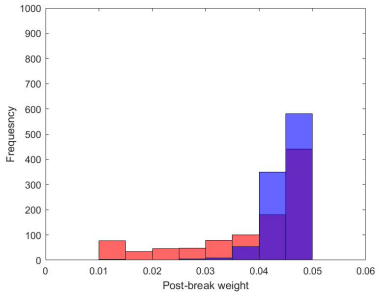
(c) $\lambda = 0.4$



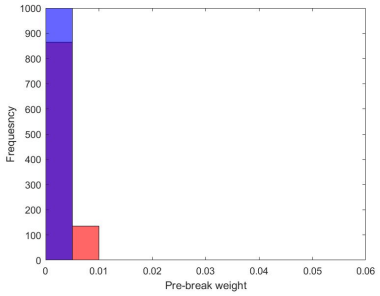
(d) $\lambda = 0.4$



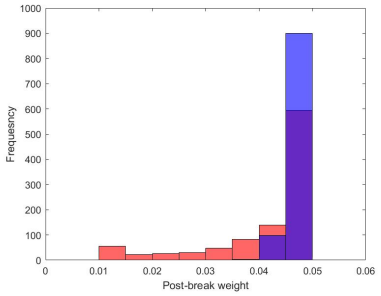
(e) $\lambda = 0.7$



(f) $\lambda = 0.7$



(g) $\lambda = 1$



(h) $\lambda = 1$

Figure 1: Frequency of weights for $T = 100$, $T_1 = 80, k = 5$ and different λ

Table 1: Simulation results with $q = 0.5$

λ	WGLS	PPP	CV	Troff	WA	Pooled	Full	AveW	$\text{PPP}_{\text{inf.}}$
$T = 100, T_1 = 20, k = 5$									
0.1	0.964	0.979	0.955	0.988	0.941	0.941	0.935	0.983	0.936
0.3	0.966	0.985	0.968	0.983	0.942	0.943	0.978	0.970	0.939
0.6	0.993	0.997	0.995	0.993	0.996	0.995	1.358	1.019	0.996
1.0	0.999	1.011	1.001	0.996	1.055	1.049	2.154	1.033	1.013
$T = 100, T_1 = 50, k = 5$									
0.1	0.957	0.980	0.960	0.978	0.942	0.941	0.940	0.968	0.934
0.3	0.974	0.983	0.980	0.974	0.960	0.958	1.112	0.953	0.934
0.6	0.997	1.002	0.990	1.002	1.119	1.107	1.919	1.031	0.982
1.0	1.000	1.003	0.994	0.998	1.462	1.404	3.678	1.141	0.994
$T = 100, T_1 = 80, k = 5$									
0.1	0.931	0.962	0.921	0.972	0.908	0.911	0.906	0.931	0.902
0.3	0.921	0.935	0.902	0.960	0.919	0.904	1.001	0.878	0.835
0.6	0.962	0.966	1.100	1.002	1.355	1.226	1.825	1.168	0.877
1.0	0.984	1.005	1.688	1.029	2.500	2.093	3.855	1.938	0.928
$T = 100, T_1 = 20, k = 10$									
0.1	0.802	0.893	0.791	0.973	0.765	0.768	0.750	0.816	0.746
0.3	0.870	0.927	0.876	0.974	0.840	0.843	0.906	0.881	0.826
0.6	0.998	1.011	1.001	1.004	1.028	1.026	1.665	1.050	1.000
1.0	1.001	1.030	1.009	1.021	1.151	1.138	3.107	1.058	1.022
$T = 100, T_1 = 50, k = 10$									
0.1	0.718	0.853	0.729	0.954	0.692	0.699	0.685	0.746	0.669
0.3	0.891	0.926	0.893	0.966	0.862	0.860	1.088	0.879	0.807
0.6	0.990	0.994	0.997	0.995	1.227	1.205	2.447	1.075	0.948
1.0	1.000	1.018	1.002	1.028	1.897	1.795	5.411	1.298	1.006
$T = 100, T_1 = 80, k = 10$									
0.1	0.607	0.776	0.627	0.925	0.593	0.602	0.592	0.626	0.579
0.3	0.712	0.784	0.702	0.949	0.713	0.688	0.836	0.675	0.601
0.6	0.856	0.873	1.118	1.027	1.358	1.192	1.921	1.118	0.720
1.0	0.937	0.975	2.074	1.145	2.888	2.383	4.472	2.169	0.833

Note: This table reports the results of the relative MSFE for different methods. All MSFEs are reported relative to the associated MSFE based on the post-break sample. λ in the first column shows different break sizes in the mean. In the heading of the table, WGLS shows the results for our proposed estimator, PPP is the one proposed by Pesaran et al. (2013), Postbk, Troff, CV, WA, and Pooled are the five methods used in Pesaran and Timmermann (2007), Full is forecast based on the full-sample observations, AveW is the method proposed by Pesaran and Pick (2011) with $T(1 - w_{\min}) + 1$ windows and $w_{\min} = 0.15$, and $\text{PPP}_{\text{inf.}}$ is the infeasible PPP method.

Table 2: Simulation results with $q = 1$

λ	WGLS	PPP	CV	Troff	WA	Pooled	Full	AveW	$\text{PPP}_{\text{inf.}}$
$T = 100, T_1 = 20, k = 5$									
0.1	0.973	0.985	0.966	0.988	0.950	0.952	0.952	0.990	0.951
0.3	0.986	0.992	0.986	0.993	0.973	0.973	0.986	1.002	0.970
0.6	0.993	0.998	0.997	0.991	1.001	0.999	1.084	1.019	0.992
1.0	0.999	1.003	1.000	0.999	1.060	1.054	1.269	1.034	1.009
$T = 100, T_1 = 50, k = 5$									
0.1	0.978	0.987	0.967	0.992	0.964	0.964	0.968	0.976	0.959
0.3	0.984	0.988	0.984	0.991	0.982	0.981	1.038	0.957	0.943
0.6	0.999	1.001	0.991	1.002	1.137	1.124	1.397	1.034	0.983
1.0	1.001	1.005	0.994	0.999	1.482	1.424	2.149	1.143	0.993
$T = 100, T_1 = 80, k = 5$									
0.1	0.968	0.983	0.957	0.986	0.945	0.945	0.942	0.968	0.938
0.3	0.946	0.965	0.962	0.991	0.969	0.954	0.987	0.912	0.860
0.6	0.954	0.967	1.119	1.008	1.334	1.224	1.496	1.140	0.849
1.0	0.981	1.006	1.703	1.018	2.489	2.099	2.991	1.927	0.919
$T = 100, T_1 = 20, k = 10$									
0.1	0.853	0.921	0.846	0.975	0.822	0.823	0.817	0.871	0.813
0.3	0.904	0.953	0.905	0.991	0.882	0.885	0.918	0.927	0.874
0.6	0.998	1.006	1.003	1.011	1.036	1.033	1.198	1.049	0.995
1.0	1.001	1.015	1.008	1.021	1.158	1.144	1.589	1.060	1.009
$T = 100, T_1 = 50, k = 10$									
0.1	0.829	0.917	0.830	0.983	0.811	0.813	0.818	0.861	0.802
0.3	0.923	0.952	0.934	0.981	0.932	0.928	1.019	0.929	0.867
0.6	0.991	0.998	0.999	1.000	1.257	1.234	1.654	1.084	0.963
1.0	1.001	1.016	1.003	1.012	1.926	1.822	3.013	1.304	1.013
$T = 100, T_1 = 80, k = 10$									
0.1	0.824	0.912	0.817	0.980	0.780	0.780	0.777	0.822	0.767
0.3	0.784	0.866	0.812	0.994	0.797	0.771	0.826	0.745	0.659
0.6	0.856	0.885	1.159	1.052	1.357	1.207	1.543	1.115	0.716
1.0	0.947	0.987	2.133	1.116	2.897	2.409	3.446	2.176	0.837

Note: See the notes to Table 1.

Table 3: Simulation results with $q = 2$

λ	WGLS	PPP	CV	Troff	WA	Pooled	Full	AveW	$\text{PPP}_{\text{inf.}}$
$T = 100, T_1 = 20, k = 5$									
0.1	1.000	1.000	1.003	0.999	0.999	0.999	1.002	1.014	0.978
0.3	0.997	0.998	1.006	1.001	1.013	1.012	1.007	1.000	0.968
0.6	0.999	1.002	1.004	1.000	1.105	1.098	1.049	1.030	0.997
1.0	1.000	1.003	1.004	1.000	1.317	1.288	1.132	1.048	1.007
$T = 100, T_1 = 50, k = 5$									
0.1	0.999	1.000	0.998	1.000	1.005	1.004	1.011	0.941	0.930
0.3	1.000	0.997	0.992	1.002	1.141	1.135	1.089	0.968	0.910
0.6	1.000	1.002	0.999	0.997	1.747	1.690	1.377	1.225	0.988
1.0	1.001	1.007	1.004	1.004	3.099	2.848	1.986	1.639	0.997
$T = 100, T_1 = 80, k = 5$									
0.1	0.964	0.985	0.955	0.993	0.941	0.938	0.934	0.912	0.845
0.3	0.972	0.991	1.108	1.000	1.220	1.168	1.128	0.999	0.713
0.6	0.993	1.012	1.981	0.996	2.866	2.443	2.245	2.143	0.811
1.0	0.992	1.032	4.514	0.994	7.753	6.175	5.330	5.469	0.984
$T = 100, T_1 = 20, k = 10$									
0.1	0.985	0.994	0.982	1.004	0.986	0.986	0.995	1.026	0.971
0.3	0.998	1.002	1.005	1.005	1.052	1.050	1.033	1.041	0.985
0.6	1.000	1.003	1.020	1.008	1.247	1.235	1.125	1.065	1.000
1.0	1.001	1.002	1.028	0.998	1.716	1.661	1.336	1.093	1.005
$T = 100, T_1 = 50, k = 10$									
0.1	0.975	0.990	0.998	1.001	0.996	0.996	0.988	0.918	0.872
0.3	0.999	1.006	0.993	1.000	1.310	1.298	1.183	1.051	0.917
0.6	1.001	1.007	1.007	1.002	2.353	2.258	1.738	1.424	0.982
1.0	1.001	1.025	1.042	1.000	4.990	4.552	3.094	2.263	1.051
$T = 100, T_1 = 80, k = 10$									
0.1	0.923	0.966	0.946	0.990	0.894	0.892	0.891	0.863	0.759
0.3	0.890	0.931	1.057	1.023	1.092	1.039	1.021	0.863	0.540
0.6	0.965	1.008	2.580	1.072	3.401	2.862	2.653	2.504	0.784
1.0	0.995	1.155	6.546	1.081	9.624	7.711	6.767	6.811	1.059

Note: See the notes to Table 1.

Table 4: Empirical results for forecasting equity premium

h	Postbk	WGLS	PPP	CV	Troff	WA	Pooled	Full	AveW
	Panel A: 1995:Q1-2018:Q1								
1	0.779	0.721**	0.756	0.713**	0.830	0.780	0.779	1.048	0.729*
2	0.833	0.794*	0.840	0.818	0.846	0.775*	0.769*	0.684*	0.841
3	0.783	0.742*	0.854	0.920	0.810	0.834	0.805	0.795	0.824
4	1.355	1.342*	1.014	0.798	1.505	0.768	0.794	0.724	0.803
	Panel B: 2005:Q1-2018:Q1								
1	0.675	0.627*	0.696	0.633	0.759	0.717	0.714	1.102	0.645
2	0.836	0.749*	0.842	0.805	0.843	0.713*	0.709*	0.554*	0.841
3	0.654	0.641	0.864	1.033	0.694	0.896	0.840	0.792	0.827
4	1.819	1.796**	1.221	0.836	2.079	0.785	0.830	0.696	0.840

Note: This table reports the $100 \times$ MSFE for different forecasting methods. The out-of-sample forecast periods are 1995:Q1-2018:Q1 and 2005:Q1-2018:Q1 in Panel A and Panel B, respectively. h in the first column shows the forecast horizon. In the heading of the table, WGLS shows the results for our proposed estimator, PPP is the one proposed by Pesaran et al. (2013), Postbk, Troff, CV, WA, and Pooled are the five methods used in Pesaran and Timmermann (2007), Full is forecast based on the full-sample observations, and AveW is the method proposed by Pesaran and Pick (2011) with $T(1-w_{\min})+1$ windows and $w_{\min}=0.1$. An asterisk denote forecast that is significantly better than that obtained from the post-break forecasts according to the Diebold and Mariano (1995) statistic. ** and * indicate significance at 5% and 10% levels, respectively.