Weighted-average least squares prediction[*]

October 25, 2013

Jan R. Magnus
Department of Econometrics & Operations Research,
VU University Amsterdam

Wendun Wang
Econometric Institute, Erasmus University Rotterdam

Xinyu Zhang
Academy of Mathematics & Systems Science,
Chinese Academy of Sciences, Beijing

[*] Address correspondence to Jan R. Magnus, Department of Econometrics & Operations Research, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands; E-mail: jan@janmagnus.nl

Abstract Prediction under model uncertainty is an important and difficult issue. Traditional prediction methods (such as pretesting) are based on model selection followed by prediction in the selected model, but the reported prediction and the reported prediction variance ignore the uncertainty from the selection procedure. This paper proposes a weighted-average least squares (WALS) prediction procedure that is not conditional on the selected model. Taking both model and error uncertainty into account, we also propose an appropriate estimate of the variance of the WALS predictor. Correlations among the random errors are explicitly allowed. Compared to other prediction averaging methods, the WALS predictor has important advantages both theoretically and computationally. Simulation studies show that the WALS predictor generally produces lower mean squared prediction errors than its competitors, and that the proposed estimator for the prediction variance performs particularly well when model uncertainty increases.

Keywords Model averaging; Model uncertainty; Bayesian analysis; Prediction.

JEL Classification C11; C52; C53

1 Introduction

In econometric practice one typically first selects the 'best' model based on diagnostic tests (t-ratios, R^2, information criteria) and then computes estimates within this selected model. This is called 'pretesting' (Leeb and Pötscher, 2003, 2006, 2008). There are many problems with this procedure (Magnus, 1999; Magnus and Durbin, 1999; Danilov and Magnus, 2004a,b), but the most important is that model selection and estimation are completely separated so that uncertainty in the model selection is ignored when reporting properties of the estimates. An alternative is to average the results obtained from all candidate models, but weighed to allow for prior confidence in the various models. This is called 'model averaging' and it has two major advantages. First, it avoids arbitrary thresholds (like 1.96), thus forcing continuity on a previously discontinuous estimator; second, it allows us to combine model selection and estimation into *one* procedure, thus moving from conditional to unconditional estimator characteristics.

Much of the model averaging literature has concentrated on estimation rather than on prediction. In this paper we concentrate on prediction (forecasting), which may in fact be a more appropriate application of model averaging, because the interpretation of coefficients changes with different models but the predictor always has the same interpretation. A substantial literature on the averaging of forecasts exists, going back to Bates and Granger (1969); see Granger (2003), Yang (2004), Elliott and Timmermann (2004), and Aiolfi and Timmermann (2006) for some recent contributions, and Hendry and Clements (2004) and Timmermann (2006) for recent reviews. Simulation and empirical studies indicate that predictors based on a set of models generally perform better than predictors obtained from a single model (Stock and Watson, 2004; Jackson and Karlsson, 2004; Bjørnland et al., 2012).

Our paper has two main contributions. First, we introduce the prediction counterpart to the weighted-average least squares (WALS) estimator proposed in Magnus et al. (2010) and study its properties in simulations. The WALS procedure avoids some of the problems encountered in standard Bayesian model averaging (BMA). In particular, the prior is based on a coherent notion of ignorance, thus avoiding normality of the prior and unbounded risk. Also, the computational burden increases linearly rather than exponentially with the number of regressors because of the so-called semi-orthogonalization, and is therefore trivial compared to other model averaging estimators such as standard BMA, model-selection-based weights methods (Buckland et al., 1997; Hjort and Claeskens, 2003), exponential reweighing (Yang, 2004), or Mallows model averaging (Hansen, 2007, 2008). Our proposed method explicitly allows for correlation in the observations, including possible correlation between the errors in the realized sample and the predictive sample.

The second contribution of the paper is that we propose an estimate for the prediction variance taking model uncertainty into account, and evaluate the accuracy of this estimate. The typical researcher's instinct is to favor a predictor with a small variance over one with a large variance. We argue that what we require is not a small but a 'correct' variance: in a situation with much noise a predictor with a small variance can cause much harm, while a truthfully reported large variance may lead to more prudent policy. In fact, one of the problems with the credibility of econometric predictions may be that our reported prediction variances are too small, and this is caused, at least in part, by the fact that model uncertainty is ignored.

The paper is organized as follows. Sections 2–7 develop the theory. In Section 2 we set up the model and present the traditional predictor. The commonly employed conditional predictor is presented in Section 3, and the WALS predictor in Section 4. In Section 5 we discuss the computation of the WALS predictor based on the Laplace prior. An estimator for the variance of the WALS predictor is proposed in Section 6. Finally, in Section 7, we discuss the estimation of unknown parameters in the variance matrix of the random disturbances. Then, in Sections 8–11, we compare the WALS predictor with its most important competitors: unrestricted maximum likelihood, pretesting, ridge regression, and Mallows model averaging. Our comparison is conducted through a large number of Monte Carlo simulation experiments, controlling for sample size, parameter values, and variance specifications. The simulation results show that the WALS predictor typically has the lowest mean squared prediction error among the predictors considered, and that the more uncertainty exists in the model, the better is the relative performance of WALS. Section 12 concludes.

2 The traditional predictor

Our framework is the linear regression model

$$y = X\beta + u, \tag{1}$$

where y is a vector of N observations on the dependent variable, X $(N \times k)$ is a matrix of regressors, u is a random vector of N unobservable disturbances, and β is a vector of k unknown parameters. We assume throughout that $1 \leq k \leq N-1$ and that X has full column-rank k. We are interested in some specific (possibly future) values of the regressors X_f $(N_f \times k)$, and we wish to predict the value y_f $(N_f \times 1)$ likely to be associated with X_f. The regressors X and X_f are taken to be fixed. We assume that y_f is generated by

$$y_f = X_f\beta + u_f, \tag{2}$$

and our task is to find a predictor $\hat{y}_f$ of y_f.

In general the observations will be correlated, and we shall assume that

$$\begin{pmatrix} u \\ u_f \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & C_f' \\ C_f & \Omega_f \end{pmatrix}\right), \tag{3}$$

where the variance of (u, u_f) is a positive definite $(N+N_f) \times (N+N_f)$ matrix, whose component blocks Ω, C_f, and Ω_f are functions of an m-dimensional unknown parameter vector $\theta = (\theta_1, \ldots, \theta_m)'$. Normality of the errors is the basis on which we build our conditional moments and the properties of the WALS predictor. The role of the normality assumption in our theorems will be discussed in more detail at the end of Section 6. Our theory applies to both fixed and random regressors under strictly exogeneity (hence not to lagged dependent variables). To simplify notation the following derivation treats the regressors as fixed (at least for the moment); the results for random regressors can be obtained similarly if we condition appropriately.

The joint distribution of u and u_f in (3) implies that

$$\mathrm{E}(u_f|u) = C_f \Omega^{-1} u, \qquad \mathrm{var}(u_f|u) = \Omega_f - C_f \Omega^{-1} C_f', \tag{4}$$

so that

$$\mathrm{E}(y_f|y) = X_f \beta + C_f \Omega^{-1}(y - X\beta). \tag{5}$$

This leads to the traditional least squares predictor in the presence of a non-scalar variance matrix:

$$\hat{y}_f = X_f \hat{\beta} + C_f \Omega^{-1}(y - X\hat{\beta}), \tag{6}$$

where $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ denotes the generalized least squares (GLS) estimator of β, and it is assumed (for the moment) that θ is known; see Whittle (1963, p. 53, Eq. (10)) for the general formula, and Johnston and DiNardo (1997, Sec. 6.8) and Ruud (2000, Sec. 19.7) for the special case where $N_f = 1$ and the errors follow an AR(1) process. The predictor (6) is normally distributed with mean $\mathrm{E}(\hat{y}_f) = X_f\beta$ and variance

$$\mathrm{var}(\hat{y}_f) = X_f(X'\Omega^{-1}X)^{-1}X_f' + C_f(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})C_f' \tag{7}$$

from which we see *inter alia* that the presence of the covariance C_f increases the variance of the predictor, and therefore that ignoring correlation leads to misleadingly precise predictions.

The prediction error $\mathrm{PE} := \hat{y}_f - y_f$ can be conveniently written as the sum of two independent random variables:

$$\mathrm{PE} = (X_f - C_f\Omega^{-1}X)(\hat{\beta} - \beta) - (u_f - C_f\Omega^{-1}u), \tag{8}$$

and the traditional predictor $\hat{y}_f$ is a good predictor in the sense that it is unbiased and that the prediction error has minimum variance

$$\begin{aligned}\mathrm{var}(\mathrm{PE}) = {} & (X_f - C_f\Omega^{-1}X)(X'\Omega^{-1}X)^{-1}(X_f - C_f\Omega^{-1}X)' \\ & + \Omega_f - C_f\Omega^{-1}C_f'\end{aligned} \tag{9}$$

in the class of linear unbiased estimators.

3 The conditional predictor

The previous section assumes that the data-generation process (DGP) and the model coincide, which one might call the 'traditional' approach. In practice, the model is likely to be (much) smaller than the DGP. In this section we shall assume that the model is a special case of the DGP obtained by setting some of the β-parameters equal to zero. We do not know in advance which β-parameters should be set to zero and we use model selection diagnostics (such as t- and F-statistics) to arrive at a model that we like. Once we have obtained this model we derive the properties of the predictor *conditional* on the selected model and hence we ignore the noise generated by the model selection process. We call this the 'conditional' approach. This is not quite right of course, and we shall present a method which combines model selection and prediction in the next section.

We distinguish between *focus* regressors X_1 (those we want in the model on theoretical or other grounds) and *auxiliary* regressors X_2 (those we are less certain of), and write model (1) accordingly as

$$y = X_1\beta_1 + X_2\beta_2 + u, \tag{10}$$

so that $X = (X_1 : X_2)$ and $\beta = (\beta_1', \beta_2')'$. Let $k_1 \geq 0$ be the dimension of β_1 and $k_2 \geq 0$ the dimension of β_2, so that $k = k_1 + k_2$. Model selection takes place over the auxiliary regressors only. Since each of the k_2 auxiliary regressors can either be included or not, we have 2^{k_2} models to consider.

In addition to the regressors that are always in the model (X_1) and those that are sometimes in the model (X_2), there are also regressors that are never in the model (say X_3), even though they are in the DGP. This is because the modeler is ignorant about these regressors or has no access to the necessary data. We disregard this situation for the moment, but return to it in Section 8.

We assume (at first) that θ and hence Ω is known. It is convenient to semi-orthogonalize the regression model as follows. Let

$$M_1^* := \Omega^{-1} - \Omega^{-1}X_1(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}, \tag{11}$$

where we notice that the matrix $\Omega^{1/2}M_1^*\Omega^{1/2}$ is idempotent. Let P be an orthogonal matrix and Λ a diagonal matrix with positive diagonal elements such that $P'X_2'M_1^*X_2P = \Lambda$. Next define the transformed auxiliary regressors and the transformed auxiliary parameters as

$$X_2^* := X_2P\Lambda^{-1/2}, \qquad \beta_2^* := \Lambda^{1/2}P'\beta_2. \tag{12}$$

Then $X_2^*\beta_2^* = X_2\beta_2$, so that we can write (10) equivalently as

$$y = X_1\beta_1 + X_2^*\beta_2^* + u. \tag{13}$$

The result of this transformation is that the new design matrix $(X_1 : X_2^*)$ is 'semi-orthogonal' in the sense that $X_2^{*\prime}M_1^*X_2^* = I_{k_2}$ and this has important advantages that will become clear shortly.

3.1 Estimation in model $\mathcal{M}_i$

Our strategy will be to estimate (β_1, β_2^*) rather than (β_1, β_2). Each of the k_2 components of β_2^* can either be included or not included in the model and this gives rise to 2^{k_2} models. A specific model is identified through a $k_2 \times (k_2 - k_{2i})$ selection matrix S_i of full column-rank, where $0 \leq k_{2i} \leq k_2$, so that $S_i' = (I_{k_2 - k_{2i}} : 0)$ or a column-permutation thereof. Our first interest is in the GLS estimator of (β_1, β_2^*) in the i-th model, that is, in the GLS estimator of (β_1, β_2^*) under the restriction $S_i'\beta_2^* = 0$.

Let $\mathcal{M}_i$ represent model (13) under the restriction $S_i'\beta_2^* = 0$, and let $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}^*$ denote the GLS estimators of β_1 and β_2^* under $\mathcal{M}_i$. Extending Danilov and Magnus (2004a, Lemmas A1 and A2), the GLS estimators of β_1 and β_2^* under $\mathcal{M}_i$ may be written as (see also Magnus et al., 2011):

$$\hat{\beta}_{1(i)} = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y - Q^*W_ib_2^*, \qquad \hat{\beta}_{2(i)}^* = W_ib_2^*, \tag{14}$$

respectively, where

$$b_2^* := X_2^{*\prime}M_1^*y, \quad Q^* := (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}X_2^*, \quad W_i := I_{k_2} - S_iS_i'. \tag{15}$$

Note that b_2^* is simply the GLS estimator of β_2^* in the unrestricted model, and that W_i is a diagonal $k_2 \times k_2$ matrix with k_{2i} ones and $(k_2 - k_{2i})$ zeros on the diagonal. The j-th diagonal element of W_i equals zero if β_{2j}^* (the j-th component of β_2^*) is restricted to zero, and equals one otherwise. If $k_{2i} = k_2$ then $W_i = I_{k_2}$. The diagonality of W_i is a direct consequence of the semi-orthogonal transformation.

The distributions of $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}^*$ are then

$$\hat{\beta}_{1(i)} \sim \mathrm{N}_{k_1}\left(\beta_1 + Q^*S_iS_i'\beta_2^*,\ (X_1'\Omega^{-1}X_1)^{-1} + Q^*W_iQ^{*\prime}\right), \tag{16}$$

$$\hat{\beta}_{2(i)}^* \sim \mathrm{N}_{k_2}\left(W_i\beta_2^*,\ W_i\right), \tag{17}$$

and $\mathrm{cov}(\hat{\beta}_{1(i)}, \hat{\beta}_{2(i)}^*) = -Q^*W_i$. The residual vector $e_i := y - X_1\hat{\beta}_{1(i)} - X_2^*\hat{\beta}_{2(i)}^*$ is given by $e_i = \Omega D_i^*y$, where $D_i^* := M_1^* - M_1^*X_2^*W_iX_2^{*\prime}M_1^*$ and $\Omega^{1/2}D_i^*\Omega^{1/2}$ is a symmetric idempotent matrix of rank $n - k_1 - k_{2i}$. It follows that:

- all models that include the j-th column of X_2^* as a regressor have the same estimators of β_{2j}^*, namely b_{2j}^*, irrespective of which other columns of X_2^* are included;
- the estimators $b_{21}^*, b_{22}^*, \ldots, b_{2k_2}^*$ are independent; and
- the residuals of the i-th model $\mathcal{M}_i$ depend on y only through M_1^*y.

3.2 Prediction in model $\mathcal{M}_i$

Next we wish to predict N_f (possibly future) values y_f, based on values of the regressors X_{1f} $(N_f \times k_1)$ and X_{2f} $(N_f \times k_2)$. Corresponding to X_2^* we define $X_{2f}^* := X_{2f}P\Lambda^{-1/2}$, so that

$$\begin{pmatrix} y \\ y_f \end{pmatrix} = \begin{pmatrix} X_1 & X_2^* \\ X_{1f} & X_{2f}^* \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2^* \end{pmatrix} + \begin{pmatrix} u \\ u_f \end{pmatrix}, \tag{18}$$

where the errors (u, u_f) are distributed as in (3). From (5) we obtain

$$\mathrm{E}(y_f|y) = X_{1f}\beta_1 + X_{2f}^*\beta_2^* + C_f\Omega^{-1}(y - X_1\beta_1 - X_2^*\beta_2^*), \tag{19}$$

leading to the predictor in model $\mathcal{M}_i$, using (14),

$$\begin{aligned}\hat{y}_f^{(i)} &= X_{1f}\hat{\beta}_{1(i)} + X_{2f}^*\hat{\beta}_{2(i)}^* + C_f\Omega^{-1}(y - X_1\hat{\beta}_{1(i)} - X_2^*\hat{\beta}_{2(i)}^*)\\ &= X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y + C_fM_1^*y + Z_fW_ib_2^*, \end{aligned}\tag{20}$$

where

$$Z_f := (X_{2f}^* - X_{1f}Q^*) - C_f\Omega^{-1}(X_2^* - X_1Q^*). \tag{21}$$

The prediction error $\mathrm{PE}^{(i)} := \hat{y}_f^{(i)} - y_f$ can now be written as

$$\mathrm{PE}^{(i)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}u + Z_f(W_ib_2^* - \beta_2^*) - v_f, \tag{22}$$

where

$$Z_{1f} := X_{1f} - C_f\Omega^{-1}X_1, \qquad v_f := u_f - C_f\Omega^{-1}u. \tag{23}$$

Since v_f and u are uncorrelated, and $X_1'\Omega^{-1}u$ and b_2^* are also uncorrelated, we find that $\mathrm{PE}^{(i)}$ is the sum of three *independent* random variables.

Theorem 1: The prediction error $\mathrm{PE}^{(i)}$ follows a normal distribution with

$$\mathrm{E}(\mathrm{PE}^{(i)}) = -Z_f(I - W_i)\beta_2^*$$

and

$$\mathrm{var}(\mathrm{PE}^{(i)}) = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_fW_iZ_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

and hence the mean squared prediction error $\mathrm{MSPE}^{(i)} := \mathrm{MSE}(\mathrm{PE}^{(i)})$ is

$$\mathrm{MSPE}^{(i)} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f\Delta_iZ_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

where

$$\Delta_i := W_i + (I - W_i)\beta_2^*\beta_2^{*\prime}(I - W_i).$$

Proof: The results follow directly from (22). ‖

The best model is therefore the one where the matrix Δ_i is as 'small' as possible. Since W_i is a diagonal matrix with only zeros and ones on the diagonal, Δ_i is 'small' if the selected model $\mathcal{M}_i$ includes precisely those regressors x_{2j}^* of X_2^* whose corresponding parameter β_{2j}^* is larger than one in absolute value. Since the β_{2j}^* are 'theoretical' t-ratios, this result corresponds exactly to econometric intuition.

4 The WALS predictor

The problem, of course, is that we don't know which model to choose. Given estimates $\hat{\beta}^*_{2j}$ of the k_2 components β^*_{2j} of β^*_2, we could include the regressor x^*_{2j} if $|\hat{\beta}^*_{2j}| > 1$, and exclude it otherwise. This would lead to a *pretest* estimator with well-established poor properties. These poor properties stem primarily from the fact that the pretest estimator is 'kinked'; it has a discontinuity at one. This is not only mathematically undesirable but also intuitively: If $\hat{\beta}^*_{2j} = 0.99$ we exclude x^*_{2j}; if $\hat{\beta}^*_{2j} = 1.01$ we include it. It would seem better to include x^*_{2j} 'continuously' in such a way that the higher is $|\hat{\beta}^*_{2j}|$, the more of x^*_{2j} is included in our model. This is precisely the idea behind model averaging. The additional benefit of model averaging is that we develop the theory taking into account both model uncertainty and parameter uncertainty. In other words, we think of model selection and parameter estimation as *one* combined procedure, so that the reported standard errors reflect both types of uncertainty.

Thus motivated, we define the WALS predictor of y_f as

$$\hat{y}_f = \sum_{i=1}^{2^{k_2}} \lambda_i \hat{y}_f^{(i)}, \tag{24}$$

where the sum is taken over all 2^{k_2} different models obtained by setting a subset of the β^*_2's equal to zero, and the λ_i's are weight-functions satisfying certain minimal regularity conditions, namely

$$\lambda_i \geq 0, \quad \sum_{i=1}^{2^{k_2}} \lambda_i = 1, \quad \lambda_i = \lambda_i(M^*_1 y). \tag{25}$$

The first two conditions define the λ_i as proper weights, lying between zero and one and adding up to one. The third condition says that each of the λ_i can only depend on $M^*_1 y$. This is motivated by tho facts. First, we observe from (14) that the estimators of (β_1, β^*_2) differ over models by a linear transformation of $M^*_1 y$. Second, it follows from the discussion below (17) that the residual vector in each model is also a function of $M^*_1 y$, and diagnostics typically are functions of the residuals. Our assumption that the weights depend only on $M^*_1 y$ is in line with the commonly used model selection criteria, such as t- and F-tests but also AIC and BIC, which also depend on y only through $M^*_1 y$. For connections between our weights $\lambda_i(M^*_1 y)$ with model selection criteria and other weight functions, see Liang et al. (2011).

The assumption $\lambda_i = \lambda_i(M^*_1 y)$ significantly alleviates the computational burden, because the WALS procedure does not require 2^{k_2} λ_i's but only the k_2 diagonal elements of $W := \sum_i \lambda_i W_i$. The definition (24) now specializes as follows.

Definition 1 (WALS predictor): The WALS predictor of y_f is given by

$$\hat{y}_f := X_{1f}(X_1' \Omega^{-1} X_1)^{-1} X_1' \Omega^{-1} y + C_f M^*_1 y + Z_f \hat{\beta}^*_2,$$

where $\hat{\beta}_2^* := Wb_2^*$.

Note that, while the W_i's are non-random diagonal matrices, the matrix W is random (but still diagonal) because it depends on the random λ_i's. The prediction error $\mathrm{PE} := \hat{y}_f - y_f$ now takes the form

$$\mathrm{PE} = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}u + Z_f(\hat{\beta}_2^* - \beta_2^*) - v_f, \tag{26}$$

and we present its moments in the following 'equivalence' theorem.

Theorem 2 (Equivalence theorem): If the weights λ_i satisfy condition (25), then the WALS prediction error PE has the following expectation, variance and mean squared error:

$$\mathrm{E}(\mathrm{PE}) = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*),$$

$$\mathrm{var}(\mathrm{PE}) = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f\,\mathrm{var}(\hat{\beta}_2^*)Z_f' + \Omega_f - C_f\Omega^{-1}C_f',$$

and hence

$$\mathrm{MSE}(\mathrm{PE}) = Z_{1f}(X_1'\Omega^{-1}X_1)^{-1}Z_{1f}' + Z_f\,\mathrm{MSE}(\hat{\beta}_2^*)Z_f' + \Omega_f - C_f\Omega^{-1}C_f'.$$

Proof: The key ingredient is that $\mathrm{cov}(M_1^*u, X_1'\Omega^{-1}u)$ and $\mathrm{cov}(u, v_f)$ are both zero. In addition, the λ_i (and hence W) depend only on M_1^*y so that $\hat{\beta}_2^* = Wb_2^*$ also depends only on M_1^*y. Hence, the three random variables $X_1'\Omega^{-1}u$, $\hat{\beta}_2^*$, and v_f are all independent of each other. The results follow. $\|$

The equivalence theorem tells us that the WALS predictor $\hat{y}_f$ will be a 'good' predictor of y_f in the mean squared error sense if and only if $\hat{\beta}_2^*$ is a 'good' estimator of β_2^*. That is, if we can find λ_i's such that $\hat{\beta}_2^*$ is an 'optimal' estimator of β_2^*, then *the same* λ_i's will provide an 'optimal' predictor of y_f.

Next we obtain expressions for the bias and variance of the predictor itself, under the assumption that the diagonal elements of W depend only on $b_2^* = X_2^{*\prime}M_1^*y$ rather than only on M_1^*y.

Theorem 3: If the diagonal elements w_j of W depend only on b_2^*, then the WALS predictor $\hat{y}_f$ has the following bias and variance:

$$\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}\beta_2) = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*)$$

and

$$\begin{aligned}\mathrm{var}(\hat{y}_f) = {} & X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_fM_1^*C_f' + Z_f\,\mathrm{var}(\hat{\beta}_2^*)Z_f' \\ & + C_fM_1^*X_2^*\,\mathrm{cov}(b_2^*, \hat{\beta}_2^*)Z_f' + Z_f\,\mathrm{cov}(\hat{\beta}_2^*, b_2^*)X_2^{*\prime}M_1^*C_f'.\end{aligned}$$

Under the stronger assumption that w_j depends only on b_{2j}^*, the $k_2 \times k_2$ matrices $\mathrm{var}(\hat{\beta}_2^*)$ and $\mathrm{cov}(b_2^*, \hat{\beta}_2^*)$ are both diagonal.

Proof: The bias follows directly from Theorem 2. Noting that

$$\mathrm{cov}(X_1'\Omega^{-1}y,\ M_1^*y) = X_1'M_1^* = 0, \qquad \mathrm{cov}(X_1'\Omega^{-1}y,\ \hat{\beta}_2^*) = 0,$$

Definition 1 implies that

$$\begin{aligned}\operatorname{var}(\hat{y}_f) &= X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_f M_1^* C_f' + Z_f \operatorname{var}(\hat{\beta}_2^*)Z_f' \\ &\quad + C_f \operatorname{cov}(M_1^* y, \hat{\beta}_2^*)Z_f' + Z_f \operatorname{cov}(\hat{\beta}_2^*, M_1^* y)C_f'.\end{aligned}$$

Since $\Omega^{1/2}M_1^*\Omega^{1/2}$ is idempotent, we can write

$$\Omega^{1/2}M_1^*\Omega^{1/2} = AA', \qquad A'A = I_{n-k_1}.$$

Define $y^* := A'\Omega^{-1/2}y$ and $B_1 := A'\Omega^{-1/2}X_2^*$, so that $y^* \sim \mathrm{N}(B_1\beta_2^*, I_{n-k_1})$. Since $B_1'B_1 = I_{k_2}$ there exists an $(n-k_1)\times(n-k)$ matrix B_2, such that $B := (B_1 : B_2)$ is orthogonal. This allows us to write

$$M_1^* y = \Omega^{-1/2}A(B_1B_1' + B_2B_2')y^*, \qquad \hat{\beta}_2^* = WB_1'y^*,$$

so that

$$\begin{aligned}\operatorname{cov}(M_1^* y, \hat{\beta}_2^*) &= \operatorname{cov}(\Omega^{-1/2}AB_1B_1'y^*, WB_1'y^*) + \operatorname{cov}(\Omega^{-1/2}AB_2B_2'y^*, WB_1'y^*) \\ &= M_1^*X_2^* \operatorname{cov}(b_2^*, \hat{\beta}_2^*) + \Omega^{-1/2}AB_2 \operatorname{cov}(B_2'y^*, WB_1'y^*) \\ &= M_1^*X_2^* \operatorname{cov}(b_2^*, \hat{\beta}_2^*),\end{aligned}$$

because $B_1'y^*$ and $B_2'y^*$ are independent, and the diagonal elements w_j of W depend only on $X_2^{*\prime}M_1^*y = B_1'y^*$.

Finally, if w_j depends only on b_{2j}^*, then

$$\operatorname{cov}(b_{2i}^*, w_j b_{2j}^*) = 0, \qquad \operatorname{cov}(w_i b_{2i}^*, w_j b_{2j}^*) = 0 \qquad (i \neq j),$$

because b_{2i}^* and b_{2j}^* are independent. In that case both $\operatorname{cov}(b_2^*, \hat{\beta}_2^*)$ and $\operatorname{cov}(\hat{\beta}_2^*, b_2^*)$ are diagonal. This completes the proof. ‖

Note that we write Theorem 3 in terms of $\operatorname{cov}(\hat{\beta}_2^*, b_2^*)$ and not in terms of $\operatorname{cov}(\hat{\beta}_2^*, M_1^*y)$, which would have been much easier. The reason is that the latter is difficult to compute, while the former is easier because it allows us to make use of the relation between prior and posterior as we shall see in Section 6.

5 Computation of the WALS predictor based on prior ignorance

The WALS predictor proposed in Definition 1 cannot be computed unless we know $W = \sum_i \lambda_i W_i$. Because of the semi-orthogonal transformation, we do know that W is diagonal, say $W = \operatorname{diag}(w_1, \dots, w_{k_2})$. There are 2^{k_2} λ_i's, but there are only k_2 w_j's. These are functions of the λ_i's, but we cannot identify the λ_i's from the w_j's. This does not matter because we are not interested in the λ_i's as we are not interested in selecting the 'best' model. We are only interested in the 'best' predictor.

The k_2 components b^*_{2j} of b^*_2 are independent with $\operatorname{var}(b^*_{2j}) = 1$. Therefore, if we choose w_j to be a function of b^*_{2j} only, then the components $\hat{\beta}^*_{2j} = w_j b^*_{2j}$ of $\hat{\beta}^*_2$ will also be independent, and our k_2-dimensional problem reduces to k_2 one-dimensional problems. The one-dimensional problem is simply how to estimate β^*_{2j} using only the information that $b^*_{2j} \sim \mathrm{N}(\beta^*_{2j}, 1)$.

This seemingly trivial question was addressed in Magnus (2002), who proposed the 'Laplace' estimator. This estimator is obtained by combining the normal likelihood with the Laplace prior,

$$b^*_{2j} \,|\, \beta^*_{2j} \sim \mathrm{N}(\beta^*_{2j}, 1), \qquad \pi(\beta^*_{2j}) = (c/2)\exp(-c|\beta^*_{2j}|), \tag{27}$$

where c is a positive constant. The Laplace estimator is now defined as the resulting posterior expectation $\hat{\beta}^*_{2j} := \mathrm{E}(\beta^*_{2j}|b^*_{2j})$. It is admissible, has bounded risk, has good properties around $|\beta^*_{2j}| = 1$, and is near-optimal in terms of minimax regret. It is also easily computable. The mean and variance of $\beta^*_{2j}|b^*_{2j}$ are given in Theorem 1 of Magnus et al. (2010). The mean is

$$\hat{\beta}^*_{2j} = \mathrm{E}(\beta^*_{2j} \,|\, b^*_{2j}) = b^*_{2j} - c \cdot h(b^*_{2j}) \tag{28}$$

with

$$h(x) := \frac{e^{-cx}\Phi(x-c) - e^{cx}\Phi(-x-c)}{e^{-cx}\Phi(x-c) + e^{cx}\Phi(-x-c)}, \tag{29}$$

and the variance $v_j := \operatorname{var}(\beta^*_{2j}|b^*_{2j})$ is

$$v_j = v(b^*_{2j}) = 1 + c^2(1 - h^2(b^*_{2j})) - \frac{c(1 + h(b^*_{2j}))\phi(b^*_{2j} - c)}{\Phi(b^*_{2j} - c)}, \tag{30}$$

where ϕ and Φ denote the density function and the cumulative distribution function of the standard-normal distribution, respectively.

The weights w_j are defined implicitly by $\hat{\beta}^*_{2j} = w_j b^*_{2j}$ and are thus given by

$$w_j = w(b^*_{2j}) = 1 - \frac{c \cdot h(b^*_{2j})}{b^*_{2j}}. \tag{31}$$

Each w_j satisfies $w(-b^*_{2j}) = w(b^*_{2j})$ and increases monotonically between $w(0)$ and $w(\infty) = 1$. Hence, $\hat{\beta}^*_{2j}$ is a shrinkage estimator, and we have

$$w(0)|b^*_{2j}| < |\hat{\beta}^*_{2j}| < |b^*_{2j}|. \tag{32}$$

In particular, when $c = \log 2$, we find that $w(0) = 0.5896$ which defines the maximum allowable shrinkage.

The hyperparameter c is chosen as $c = \log 2$, because this implies

$$\Pr(\beta^*_{2j} > 0) = \Pr(\beta^*_{2j} < 0), \qquad \Pr(|\beta^*_{2j}| > 1) = \Pr(|\beta^*_{2j}| < 1). \tag{33}$$

What this means is that we assume a priori ignorance about whether β^*_{2j} is positive or negative, and also about whether $|\beta^*_{2j}|$ is larger or smaller than one. These seem natural properties for a prior in our context, because we don't know a priori whether the β_2^* coefficients are positive or negative, and we don't know either whether adding a specific column of X_2^* to the model will increase or decrease the mean squared error of the predictors. Such a prior thus captures prior ignorance in a natural way. Given the choice of the weights w_j and hence of the estimator $\hat{\beta}_2^*$, the WALS predictor $\hat{y}_f$ can be computed.

6 Moments of the WALS predictor

The moments of the WALS predictor are given in Theorem 3, but the expressions provided there depend on unknown quantities. Under the assumption that the weights w_j are specified as in (31), and hence depend on b_{2j}^* only, we estimate these unknown quantities as follows.

Theorem 4: If the diagonal elements w_j of W depend only on b_{2j}^* as specified in (31), then the expected bias of the WALS predictor $\hat{y}_f$, based on prior densities $\pi(\beta_{2j}^*)$, is zero:

$$\mathrm{E}\left(\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}\beta_2)|\beta_2^*\right) = 0.$$

Proof: According to Theorem 3, the prediction bias, conditional on β_2^*, is

$$\mathrm{E}(\hat{y}_f - X_{1f}\beta_1 - X_{2f}^*\beta_2^*|\beta_2^*) = Z_f\,\mathrm{E}(\hat{\beta}_2^* - \beta_2^*|\beta_2^*).$$

Further,

$$\begin{aligned}\mathrm{E}(\hat{\beta}_{2j}^* - \beta_{2j}^*) &= \mathrm{E}\left(\mathrm{E}(\hat{\beta}_{2j}^* - \beta_{2j}^*|\beta_{2j}^*)\right)\\ &= \mathrm{E}\left(\mathrm{E}(b_{2j}^* - \beta_{2j}^*|\beta_{2j}^*)\right) - c\cdot\mathrm{E}\left(\mathrm{E}(h(b_{2j}^*)|\beta_{2j}^*)\right) = 0,\end{aligned}$$

because $\mathrm{E}(h(b_{2j}^*)|\beta_{2j}^*)$ is antisymmetric in β_{2j}^* and $\pi(\beta_{2j}^*)$ is symmetric in β_{2j}^*. Hence the expected bias of $\hat{y}_f$ vanishes. $\|$

The variance of $\hat{y}_f$ is given in Theorem 3. Under the assumption that the weights w_j depend only on b_{2j}^*, the matrices $\mathrm{var}(\hat{\beta}_2^*)$ and $\mathrm{cov}(b_2^*, \hat{\beta}_2^*)$ are both diagonal. Hence it suffices to discuss the estimation of $\mathrm{var}(\hat{\beta}_{2j}^*)$ and $\mathrm{cov}(b_{2j}^*, \hat{\beta}_{2j}^*)$. The variance in the posterior distribution of $\beta_{2j}^*|b_{2j}^*$ is given by v_j in (30), and hence provides the obvious estimate of $\mathrm{var}(\hat{\beta}_{2j}^*)$. It is less obvious how to find an appropriate estimate of $\mathrm{cov}(b_{2j}^*, \hat{\beta}_{2j}^*)$. We propose to use the weight as the estimator of the covariance, i.e.

$$w_j = \widehat{\mathrm{cov}}(b_{2j}^*, \hat{\beta}_{2j}^*) = \widehat{\mathrm{cov}}(b_{2j}^*, w(b_{2j}^*)b_{2j}^*). \tag{34}$$

Since $\mathrm{var}(b_{2j}^*) = 1$, this would be a perfect estimate if w_j were a constant. Now, w_j depends on b_{2j}^* and is therefore not a constant. Still, its variation is very small compared to the variation in b_{2j}^*. The correlation associated with the covariance is

$$\widehat{\mathrm{corr}}(b_{2j}^*, \hat{\beta}_{2j}^*) = \frac{\widehat{\mathrm{cov}}(b_{2j}^*, \hat{\beta}_{2j}^*)}{\sqrt{\widehat{\mathrm{var}}(b_{2j}^*)\widehat{\mathrm{var}}(\hat{\beta}_{2j}^*)}} = \frac{w(b_{2j}^*)}{\sqrt{v(b_{2j}^*)}}, \tag{35}$$

since we estimate $\operatorname{var}(\hat{\beta}^*_{2j})$ by $v_j = v(b^*_{2j})$. The estimated correlation is therefore always positive (in fact, larger than 0.7452) and smaller than one, such that when b^*_{2j} approaches $\pm\infty$ the correlation approaches one.

We conclude that a suitable estimator for the variance of the WALS predictor is given by

$$\begin{aligned}\widehat{\operatorname{var}}(\hat{y}_f) = X_{1f}(X_1'\Omega^{-1}X_1)^{-1}X_{1f}' + C_f M_1^* C_f' + Z_f V Z_f' \\ + C_f M_1^* X_2^* W Z_f' + Z_f W X_2^{*\prime} M_1^* C_f',\end{aligned} \tag{36}$$

where V and W are diagonal $k_2 \times k_2$ matrices whose j-th diagonal elements v_j and w_j are given in (30) and (31), respectively. Having thus obtained estimators for all unknown quantities, the prediction variance can be computed.

A few words on the impact and limitations of the normality assumption are in order. Our derivations are based on the normality of the error term, and this allows us to incorporate correlations between contemporary and future errors, and obtain the conditional expectation of future observations given contemporary observations, as in equations (4) and (5). These conditional expectations are our starting point. Without the normality assumption, we can still obtain the conditional expectation $\mathrm{E}(y_f|y)$ (of course in a different form), but the WALS procedure would not apply directly, because it depends on the validity of the equivalence theorem. At the moment, we don't yet have a version of the equivalence theorem under non-normality.

7 Unknown variance matrix

We have thus far assumed that Ω and C_f are known, whereas in practice they are of course unknown. If the structure of the variance matrix is known, we can estimate Ω and C_f once we have an estimate of unknown parameter θ. The parameter θ can be estimated based on the unrestricted model by minimizing

$$\varphi(\theta) := \log|\Omega| + y'(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})y \tag{37}$$

with respect to θ.

This leads to the maximum likelihood estimator $\hat{\theta}$ of θ, and hence to the estimators $\hat{\Omega} = \Omega(\hat{\theta})$ and $\hat{C}_f = C_f(\hat{\theta})$. Note that the gradient of φ is the $m \times 1$ vector whose i-th component is given by

$$\frac{\partial\varphi(\theta)}{\partial\theta_i} = \operatorname{tr}\left(\Omega^{-1}\frac{\partial\Omega}{\partial\theta_i}\right) - (M^*y)'\frac{\partial\Omega}{\partial\theta_i}(M^*y), \tag{38}$$

where

$$M^* = M_1^*(\Omega - X_2^*X_2^{*\prime})M_1^*. \tag{39}$$

Therefore, $\hat{\theta}$ depends on y only through M_1^*y and the same holds for $\hat{\Omega}$ and $\hat{C}_f$.

Replacing the unknown variance matrix with its estimator can have an effect on the property of the WALS predictor. However, Danilov (2005) showed that plugging in the unknown variance has a marginal effect on the WALS estimator, at least in terms of the risk. We shall study the effect of plugging in an *inconsistent* variance estimator in the simulation.

8 Simulation setup

Sections 2–7 contain the theoretical framework. Our next task is to evaluate the performance of the WALS predictor in a number of common situations and in comparison with other often-used predictors. In the current section we describe the setup of our simulation experiment. The simulation results are presented in Section 9. Many extensions of the benchmark setup were considered and some of these are summarized in Sections 10 and 11.

8.1 Seven methods

In the simulations we compare the performance of the WALS predictor to six commonly-used methods: unrestricted maximum likelihood (ML), two pretesting methods (PT), ridge regression (Ridge), least absolute shrinkage and selection operator (Lasso), and Mallows model averaging (MMA). We briefly describe each method below.

Unrestricted maximum likelihood simply estimates the unrestricted model (with *all* auxiliary regressors). There is no model selection here, and hence no noise associated with the model selection procedure. On the other hand, the noise associated with the estimation procedure will be large because of the large number of parameters.

Pretest estimation is a long-standing practice in applied econometrics, perhaps because pretest estimators are 'logical outcomes of the increased diagnostic testing of assumptions advocated in many econometric circles' (Poirier, 1995, p. 522). Pretest estimators and predictors do not follow textbook OLS or GLS properties, because the reported predictor is biased and its variance is only correct *conditional* on the selected model. One would expect the unconditional ('true') variance to be larger, because of the model selection noise. Giles and Giles (1993) provide a comprehensive review of the pretest literature. In pretest prediction one first selects the model based on diagnostic testing, and then predicts under the selected model. The choice of critical values of the pretest has received much attention (Toyoda and Wallace, 1976; Ohtani and Toyoda, 1980; Wan and Zou, 2003). Here we use the *stepwisefit* routine in Matlab (PT$_{sw}$), one of the most popular pretest methods. This routine begins with a forward selection procedure based on an initial model, then employs backward selection to remove variables. The steps are repeated until no additions or deletions of variables are indicated. We treat the model that includes only the focus regressors as the initial model and let the routine select the auxiliary regressors according to statistical significance. We choose the significance level for adding a variable to be 0.05 and for removing a variable to be 0.10. We also consider another model selection procedure which tests β^*_{2j} and selects the X^*_{2j} whose $|\hat{\beta}^*_{2j}|$ are larger than 1. This is a one-step pretesting method, and we denote it as PT$_{os}$.

Ridge regression (Hoerl and Kennard, 1970) is a common shrinkage technique, originally designed to address multicollinearity. Since the focus parameters are always in the model, we only penalize the auxiliary parameters. The ridge estimator is then obtained by minimizing

$$\phi(\beta_1, \beta_2) = (y - X_1\beta_1 - X_2\beta_2)'(y - X_1\beta_1 - X_2\beta_2) + \kappa\beta_2'\beta_2. \tag{40}$$

Letting

$$E_1 = \begin{pmatrix} I_{k_1} & 0_{k_1\times k_2} \\ 0_{k_2\times k_1} & 0_{k_2\times k_2} \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 0_{k_1\times k_1} & 0_{k_1\times k_2} \\ 0_{k_2\times k_1} & I_{k_2} \end{pmatrix}, \tag{41}$$

the solution can be written as

$$\hat{\beta}(\kappa) = (X'X + \kappa E_2)^{-1}X'y, \tag{42}$$

where κ is the tuning parameter. Alternatively we obtain the ridge estimator in a Bayesian framework as the mean in the posterior distribution of $\beta|(X'X)^{-1}X'y$ by combining the data density $(X'X)^{-1}X'y|\beta \sim \mathrm{N}(\beta, \sigma^2(X'X)^{-1})$ with the partially informative prior $\beta/\sigma \sim \mathrm{N}(0, (1/\epsilon)E_1 + (1/\kappa)E_2)$ and letting $\epsilon \to 0$. Following Golub et al. (1979), we choose the tuning parameter κ by minimizing the generalized cross validation criterion

$$\mathrm{GCV}(\kappa) = \frac{(y - \Xi(\kappa)y)'(y - \Xi(\kappa)y)}{(N - \mathrm{tr}\,\Xi(\kappa))^2}, \qquad \Xi(\kappa) = X\left(X'X + \kappa E_2\right)^{-1}X'. \tag{43}$$

As an alternative to ridge regression we also consider the predictor using the Lasso. The Lasso shrinks some coefficients and sets others equal to zero; it can be thought of as a combination of subset selection and ridge regression.

Finally, Mallows model averaging, proposed by Hansen (2007), averages over estimators using weights obtained by minimizing the Mallows criterion

$$C(\lambda) = (y - P(\lambda)y)'(y - P(\lambda)y) + 2\sigma^2\,\mathrm{tr}\,P(\lambda), \tag{44}$$

where $\lambda = (\lambda_1, \dots, \lambda_{2^{k_2}})$, $P(\lambda) = \sum_i \lambda_i X^{(i)}(X^{(i)\prime}X^{(i)})^{-1}X^{(i)\prime}$, and $X^{(i)}$ is the regressor matrix in model $\mathcal{M}_i$. Note that we do not assume an explicit ordering of the regressors, as Hansen does. An explicit ordering has the computational advantage that it reduces the number of weights from 2^{k_2} to k_2, but it is typically not practical in applications. (WALS also reduces the computational burden from 2^{k_2} to k_2, but through a semi-orthogonalization which does not require further assumptions.) When the submodels are strictly nested, Hansen (2007) proved that the MMA estimator is asymptotically optimal in a given class of model averaging estimators. Wan et al. (2010) extended the optimality to non-nested models, and showed the superiority of this method to smoothed AIC, weigthed BIC, Bates-Granger combination, among others; see Hansen (2008) for details. Further research may compare WALS with more recent model averaging techniques, such as jackknife model averaging (Hansen and Racine, 2012), optimal weighting (Liang et al., 2011), and other methods.

All predictors explicitly account for possible correlation in the random disturbances. In particular, the WALS predictor is obtained using Definition 1, and the predictors of the other four predictors are all computed from

$$\hat{y}_f = X_f\hat{\beta} + C_f\Omega^{-1}(y - X\hat{\beta}), \tag{45}$$

where $\hat{\beta}$ depends on the chosen method. For ML (unrestricted model, no model selection), the predictor is linear in y and the associated variance is easily computed. For PT$_{sw}$ and Ridge, the predictor is not linear in y, but the reported variance is calculated as if the predictor were linear in y, following common practice. The variance for WALS is estimated from (36) while the variance for MMA cannot be computed.

8.2 Data-generation process

We generate the data in three steps. First, we design the regressor matrix $X = (X_1 : X_2 : X_3)$, where X_1 and X_2 contain the focus and auxiliary variables, while X_3 contains the regressors that are omitted by the researcher (from *every* model) either because of ignorance or because of data limitations. The DGP and the largest (unrestricted) model are therefore not necessarily the same in the simulations. This is important because it brings us one step closer to econometric practice. In the benchmark DGP we consider six regressors with $k_1 = 2$, $k_2 = 3$, and $k_3 = 1$, such that

$$X_1 = (x_1, x_2), \qquad X_2 = (x_3, x_4, x_5) \qquad X_3 = (x_6), \tag{46}$$

where x_1 is the intercept. Since $k_2 = 3$ we have $2^3 = 8$ possible models. In the benchmark, x_2 is generated by independent standard-normal distributions, while X_2 and X_3 are generated by multivariate normal distributions with correlation 0.3. All regressors are treated as fixed, so that each replication uses the same realization of the regressors once they have been generated. In Section 11 we shall consider extensions where we have a large number of regressors and the regressors are autocorrelated or non-normally distributed.

Next, we simulate the parameters β_j $(j = 1, \ldots, 6)$ corresponding to regressors $x_1, \ldots, x_6$. For the auxiliary and omitted regressors $x_3, \ldots, x_6$ we set these parameters indirectly by controlling the 'theoretical' t-ratios, as follows. If we estimate the focus variables and just one auxiliary variable x_j, we obtain an estimated coefficient $\hat{\beta}_j$ with variance $\text{var}(\hat{\beta}_j) = (x_j' M_1^* x_j)^{-1}$. This implies a t-ratio $\hat{t}_j = \hat{\beta}_j \sqrt{x_j' M_1^* x_j}$. The 'theoretical' t-ratio is now defined as

$$t_j = \beta_j \sqrt{x_j' M_1^* x_j} \qquad (j = 3, \ldots, 6). \tag{47}$$

The values of the t_j are important (especially whether $|t_j| > 1$ or $|t_j| < 1$), because they determine whether adding an auxiliary regressor to the model will increase or decrease the root mean squared prediction error (the square root of the mean squared prediction error); see Theorem 1. We consider five combinations, as follows:

	Auxiliary			Omitted
T	t_3	t_4	t_5	t_6
T_1	1.2	0.9	1.1	0.0
T_2	1.2	1.7	0.7	0.9
T_3	1.2	0.9	1.0	2.5
T_4	2.0	2.5	2.7	0.0
T_5	0.4	0.2	0.5	0.0

Given x_j and t_j, we then obtain the parameters β_j $(j = 3, \dots, 6)$. Three of the five cases (T_1, T_4, T_5) have no omitted variables. In T_1 the t-ratios of the auxiliary variables are close to 1, in T_4 the t-ratios are large, and in T_5 they are small. The other two cases (T_2, T_3) have an omitted variable. The value of t_6 is either close to one (T_2) or large (T_3).

Regarding the focus parameters we let $\beta_1 = \beta_2 = \nu\sqrt{\sum_{j=3}^{6} \beta_j^2}$ for three values of ν: 1, 2, and 3. Since the prediction performance is hardly affected by this choice, we shall report for $\nu = 2$ only.

Finally, we generate the error terms, based on (3), from a normal distribution with mean zero and variance Ω_{all}. We consider six specifications of Ω_{all}: two for homoskedasticity, two for heteroskedasticity, and two for autocorrelation. More precisely,

- homoskedasticity: $\Omega_{all} = \sigma^2 I_{n+n_f}$ with $\sigma^2 \in \{0.25, 1.00\}$;
- heteroskedasticity: $\Omega_{all} = \operatorname{diag}[\exp(\tau x_2)]$ with $\tau \in \{0.2, 0.7\}$;
- autocorrelation: AR(1) with $\sigma^2 = 1.0$ and $\rho \in \{0.3, 0.7\}$.

8.3 Comparison of prediction methods

We evaluate the seven methods by comparing the predictors and the estimated variances of the predictors. To compare the predictors produced by the seven methods, we consider the deviation between the predictor $\hat{y}_f$ and the true value y_f. A direct comparison is, however, misleading because there is a component common to all procedures. Hence we compute a modified version of the root mean squared prediction error,

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_f^{(r)} - y_f^{(r)} + (u_f - C_f\Omega^{-1}u)\right)'\left(\hat{y}_f^{(r)} - y_f^{(r)} + (u_f - C_f\Omega^{-1}u)\right)} \quad (48)$$

where $\hat{y}_f^{(r)}$ and $y_f^{(r)}$ are the predictor and the true value in the r-th replication. We follow Hansen (2008) and subtract $u_f - C_f\Omega^{-1}u$ from the prediction error, because it is common across prediction methods and independent of u, hence independent of $\hat{\beta} - \beta$.

To compare the prediction variances is more subtle. We could just compare the magnitudes of

$$\frac{1}{R}\sum_{r=1}^{R}\operatorname{var}(\hat{y}_f^{(r)}), \quad (49)$$

which would tell us whether one method reports more precise predictions than another. This is of interest, but more important than whether the reported prediction variance is small is whether the prediction variance is *correct*. It is easy to find predictors with small variances, but this does not make them good predictors.

Thus we wish to determine how close the estimated variance is to the 'true' variance, and this is measured by the RMSE of the prediction variance,

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\operatorname{var}(\hat{y}_f^{(r)}) - V_T\right)^2}, \tag{50}$$

where V_T denotes the 'true' variance, that is, the actual variance of the predictor. Since different methods give different predictors, the 'true' variance of the predictor varies across methods. We estimate V_T by obtaining $R_v = 100$ predictors from the replications, and then computing the sample variance of these predictors,

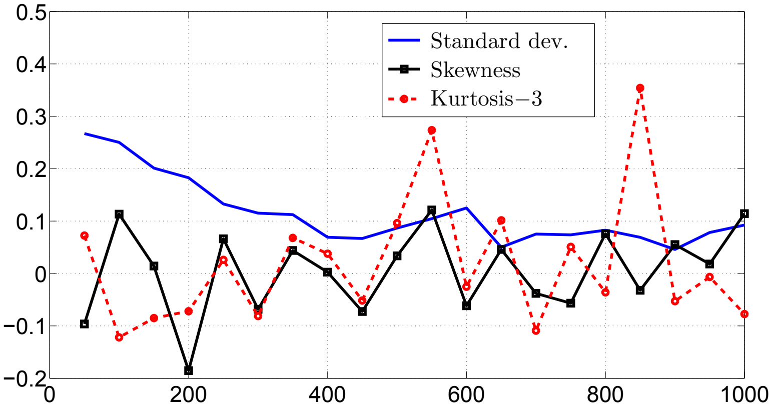
$$V_T := \frac{1}{R_v - 1}\sum_{r=1}^{R_v}\left(\hat{y}_f^{(r)} - \frac{1}{R_v}\sum_{r=1}^{R_v}\hat{y}_f^{(r)}\right)^2. \tag{51}$$

We consider training samples of size $N = 100$ and $N = 300$, and a prediction sample of size $N_f = 10$. The simulation results are obtained by computing averages across $R = 3000$ draws.

9 Simulation results: The benchmark

Before we compare the finite-sample properties of the seven methods, we first examine briefly the asymptotic behavior of the WALS predictor. Figure 1 presents three moments of the WALS predictor (computed based on the empirical distribution) as the number of observations increases to 1000. The figure shows

Figure 1: Empirical moments of the WALS predictor as a function of sample size



that, as the number of observations increases, the standard deviation decreases, suggesting consistency of the WALS predictor. However, the skewness and kurtosis do not seem to converge to zero, which suggests that the WALS predictor is asymptotically not normally distributed, probably due to the random weights employed in WALS. In our set-up we cannot verify Leeb and Pötscher's (2006) critique that the distribution estimate of a post-model-selection estimator is not uniformly consistent.

We now turn to the finite sample behavior, which is our motivating interest. We compare the predictors by considering two sample sizes ($N = 100$, $N = 300$), five sets of parameter values ($T_1, \ldots T_5$), six specifications of Ω_{all}, and seven methods. Each method is presented relative to ML, that is, we present the RMSE of each method divided by the RMSE of ML. An entry smaller than one thus indicates a superior performance relative to the ML method.

TABLE 1

The RMSEs of the predictors are presented in Table 1. We omit the results of PT$_{os}$ and Lasso in Table 1 (available upon request) since the performance of PT$_{os}$ is highly similar to PT$_{sw}$, and the Lasso predictor is not as good as ridge in most cases except in T_5 with medium heteroskedastic errors ($\tau = 0.7$) and medium autocorrelated errors ($\rho = 0.7$). We see that WALS comes out best in 42 out of 60 cases (70%), followed by Ridge (15%), and ML (10%). There are three cases (5%) where Lasso outperforms WALS and Ridge: Heteroskedasticity T_5 ($N = 300, \tau = 0,7$), autocorrelation T_5 ($N = 100, \rho = 0,7$), and autocorrelation T_5 ($N = 300, \rho = 0,7$). The pretest and MMA predictors never dominate. The dominance of WALS occurs for each of the specifications of Ω_{all}, though slightly less in the autocorrelation case than in the homo- and heteroskedastic cases. One reason why WALS is superior over MMA is that WALS makes use of the information in the error structure while MMA does not.

In T_1 and T_2 WALS dominates in all twelve cases. This shows that WALS performs well when the t-ratios of the auxiliary variables are close to one, even when the model possibly omits one variable with a t-ratio close to one. If the omitted variable has a stronger impact on the dependent variable, as in T_3, WALS still works best in 8/12 cases followed by ML (4/12). This suggests that omitting important regressors may affect the prediction ability of WALS under non-iid errors, and that ML using the full model without shrinking can outperform the shrinkage estimators in some cases. We shall investigate this point further in Section 11.

When the t-ratios of the auxiliary variables are much larger than one, as in T_4, then WALS is still the best, while ML also performs well in some of these cases. This makes sense, because model uncertainty plays a smaller role here. In the opposite case where the t-ratios of the auxiliary variables are much smaller than one, as in T_5, WALS is not the best. Here the Ridge or Lasso predictors dominate, and ML is always the worst. Here too there is little model uncertainty. The unrestricted model (ML) is not appropriate, while shrinkage towards the restricted estimator (with only the focus regressors) makes sense, and this is what Ridge and Lasso do.

TABLE 2

Before we study the prediction variance, we examine the performance of the WALS predictor when the structure of the error variance is misspecified. In this case the estimated variance is incorrect, which will have an effect on the properties of the predictors. The question of interest is whether this effect is large or small. We focus on two common types of misspecification: Misspecifying heteroskedasticity as homoskedasticity and ignoring autocorrelation. The results are given in Table 2. Other types of misspecification are also examined and the results are similar. We see that WALS is still the best in 25/40 cases (63%). The good performance of WALS shows that the inconsistency of the variance estimate hardly affects the *relative* prediction performance of WALS compared to the alternative predictors under consideration. We also note that the number of cases where Ridge or Lasso performs best increases under misspecification, which suggests robustness of Ridge and Lasso with respect to error variance misspecification. On the other hand, ML performs well in only one case, confirming that ML is sensitive to misspecification.

FIGURE 2

We next compare the performance of the prediction variance. We first consider the magnitude of the estimated variance itself, then we ask how close the estimated variance is to the 'true' variance. The MMA method is not included in this comparison because there is no procedure known to us to compute this variance. In the boxplots of Figure 2, the central mark is the median, the edges of each box indicate the 25-th and 75-th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

We consider six representative cases. Judging by the median of the estimated variance, ML has the largest variance, followed by WALS, while the variance of the Ridge and PT$_{sw}$ predictors are both smaller than WALS. This is in accordance with intuition, because ML includes all regressors, while pretesting and ridge are based on the selected model or the selected parameter, thus ignoring variation caused by the selection procedure. The WALS predictor has a relatively large variance (but still smaller than ML), because it does take the uncertainty in the selection procedure into account.

We note that the estimated variances for WALS and ML are more concentrated on their median values than those of Ridge and PT$_{sw}$, and that the distributions of the latter two methods are also characterized by a strong asymmetry. The difference between the four variance estimates is relatively small when there is little model uncertainty (T_4), and more pronounced when model uncertainty is large (T_1).

TABLES 3 and 4

As discussed before, a variance estimate is a good estimate, not when it is small, but when it provides the correct information about the precision of the predictor. If this precision happens to be low, then we need to provide a high value for the variance estimate. Table 3 gives the RMSE of the estimated prediction variance, as given in (50), again relative to ML. On the left side of the table (where the parameters σ^2, τ, and ρ are relatively small), the RMSE ratios (relative to ML) are, on average, 1.10 for WALS, 2.43 for Ridge, and 10.98 for PT_{sw}. On the right side (where the parameter values are larger, corresponding to more uncertainty), the RMSE ratios are 1.03 for WALS, 1.91 for Ridge, and 12.72 for PT_{sw}. The variance of PT_{os} (not reported but available upon request) has a smaller RMSE than that of PT_{sw}, but still much larger than WALS. The main conclusion from the table is therefore that ML and WALS provide the best estimates of the prediction variance, while Ridge, PT_{os}, and especially PT_{sw} generally report a variance which is misleadingly small. While WALS provides a much better estimate of the forecast than ML, the variance of the forecast is slightly more accurately estimated in ML than in WALS.

ML performs particularly well when N is large (because of the asymptotic behavior of ML estimates and predictions) and when the variance parameters are small. The relative performance of the WALS prediction variance estimates is improved by increasing the variance of the error terms. This suggests that prediction using WALS is especially attractive in the presence of model uncertainty. WALS performs even better (relative to other methods) when we allow for misspecification in the variance structure, as shown in Table 4. WALS produces the smallest RMSE in 36/40 cases (90%) while ML is the best in the remaining 4/40 cases (10%). This shows again that WALS is more robust than ML with respect to variance misspecification.

In the benchmark setup, where we have assumed deterministic regressors and coefficients, there is not much model uncertainty. If we allow more model uncertainty, for example by introducing random regressors or random coefficients or by increasing the variance of the errors, then the previous results suggest that the WALS estimates, which incorporate the model uncertainty, are more accurate than ML. The impact of model uncertainty is clearly an important issue and we analyze it in more depth in the next section.

10 Simulation results: More uncertainty

In this section we extend the benchmark setup by introducing additional randomness in the model. This is achieved by allowing for random regressors or random coefficients or by increasing the variance of errors.

10.1 Random regressors

We first consider the model with random but exogenous regressors. This is a common extension in simulation designs, and particularly useful in applications where one wishes to model dynamic economic behavior. The only difference with the benchmark is that we generate a new set of X's from $\mathrm{N}(0, \sigma_x^2)$ in every replication, so that each realization of the y-series involves a new realization of the X-series. (The introduction of σ_x^2 is unimportant, because the RMSE is invariant to its value.) The generation of X is independent of the errors.

Allowing the regressors to be random increases the RMSE of the forecast in each method (tables omitted). The relative performance of the seven predictors is similar to the benchmark case. In particular, the WALS predictor has the lowest RMSE in T_1, T_2, and T_3, about 6% lower than the RMSE of the ML predictor. In case T_5, the ridge predictor has the lowest RMSE under all error structures, around 14% lower than the ML predictor. In contrast to the benchmark results, allowing random regressors improves the relative performance of WALS over ML in T_4, because more randomness decreases the importance of the auxiliary variables.

TABLE 5

The main difference between the random regressor model and the benchmark model is in the prediction variance, and we report its RMSE in Table 5. WALS now produces the prediction variance with the smallest RMSE in all cases, including T_4 and T_5. The results are not affected by the misspecification of the error variance structure. This remarkable performance of WALS is due to the fact that randomness in the regressors raises model uncertainty, which in turn increases the variation of the predictor, that is, the true variance. The prediction variance of WALS explicitly incorporates such model uncertainty, in contrast to pretesting, ridge regression, and ML.

10.2 Random coefficients

Next we consider the situation where the coefficients of the explanatory variables are subject to random variation, that is,

$$y_t = \sum_{j=1}^{6} x_{tj}(\beta_j + v_{tj}) + u_t \qquad (t = 1, 2, \ldots, N), \tag{52}$$

where the v_{tj}'s are independent unobserved random disturbances, distributed as $\mathrm{N}(0, \sigma_v^2)$. Such models date back to Rubin (1950), Hildreth and Houck (1968), Swamy (1970), Froehlich (1973), and others, who discussed parameter estimation and provided empirical applications. Prediction in random coefficient models is studied, *inter alia*, in Bondeson (1990) and Beran (1995). We can rewrite (52) as

$$y_t = \sum_{j=1}^{6} x_{tj}\beta_j + \zeta_t, \qquad \zeta_t = \sum_{j=1}^{6} x_{tj}v_{tj} + u_t \tag{53}$$

where ζ_t is normally distributed with mean zero and variance $\sigma_\zeta^2 = \sigma_u^2 + \sigma_v^2 \sum_j x_{tj}^2$. This shows that introducing variation in the coefficients increases the variance of the errors. We assume that the researcher is ignorant of the random coefficients and misspecifies them as fixed. Hence the model is the benchmark model, but the DGP has changed. How do the predictors respond to this situation?

Regarding the accuracy of the predictors, we find similar results as in the random regressor model. The WALS predictor has the lowest RMSE in cases T_1–T_4, while the ridge predictor is the best under T_5. This demonstrates good performance of the WALS predictor when the t-ratios of the auxiliary variables are close to one, even when the coefficients are misspecified.

FIGURE 3

The accuracy of the estimated prediction variance is shown in Figure 3 as a function of σ_v^2. Increasing σ_v^2 raises the model uncertainty as well as the degree of misspecification, thus lowering the accuracy of all predictions. The variance estimates obtained from pretesting have a much larger RMSE than those from other methods, and they are also more volatile. Ridge regression generally produces somewhat better variance estimates. Most accurate are ML and WALS, and their variance accuracy is close when σ_v^2 is small. When $\sigma_v^2 = 0$ (the benchmark), ML has smaller RMSE than WALS, but as σ_v^2 increases, the RMSE of WALS increases slower than the RMSE of ML, and when $\sigma_v^2 > 0.03$ the accuracy of WALS variance estimates is higher than ML. These results confirm that WALS behaves well in the presence of a large degree of model uncertainty. Viewed differently, WALS is more robust than pretesting, ridge, and ML.

10.3 Increase in the variance of errors

Finally, we consider an increase in the variance of the errors by changing a parameter in Ω_{all}. We only consider the homoskedastic and the heteroskedastic cases. Under homoskedasticity we can increase the error variance by increasing σ^2; under heteroskedasticity case by increasing τ.

FIGURE 4

Figure 4 shows how the RMSE of the prediction variance changes as the parameters σ^2 and τ increase. In both cases, WALS and ML outperform Ridge and, in particular, PT$_{sw}$. When the error variance is small, the prediction variances produced by WALS and ML show similar accuracy. But as the error variance increases, the WALS prediction variance has small RMSE than ML.

Note that increasing the error variance affects the RMSE of the prediction variance in different ways: it increases the RMSE in the homoskedastic case but reduces the RMSE in the heteroskedastic case. This is because in the design of the heteroskedastic variance, $\Omega_{all} = \exp(\tau x_2)$ is a function of x_2. Increasing τ leads to a smaller estimated coefficient $\hat{\beta}_2$ since the estimation process cannot distinguish between increasing the error variance from increasing the variation in x_2.

In summary, more model uncertainty leads to a better performance of WALS relative to the other methods.

11 Simulation results: More regressors

In Sections 9 and 10 we assumed two focus regressors, three auxiliary regressors, and one omitted regressor. In practical applications the number of regressors is likely to be larger. In this section we extend the benchmark framework by assuming $k_2 = 12$ auxiliary regressors and $k_3 = 3$ omitted regressors, while keeping the same number $k_1 = 2$ of focus regressors. The large number of auxiliary regressors will increase the model uncertainty, because we now have $2^{12} = 4096$ different models to consider compared to $2^3 = 8$ in the benchmark. When introducing new variables we have to specify the 'theoretical' t-ratios which are used to compute the values of the β-parameters. We consider four combinations, as follows:

	Auxiliary	Omitted
T	t_3–t_{14}	t_{15}–t_{17}
T_{L1}	1.2, 0.9, 1.0, 1.3, 1.2, 1.5, 1.6, 1.2, 1.1, 0.8, 1.5, 1.4	0.0, 0.0, 0.0
T_{L2}	1.2, 0.9, 1.0, 1.3, 1.2, 1.5, 1.6, 1.2, 1.1, 0.8, 1.5, 1.4	2.4, 2.8, 2.0
T_{L3}	1.2, 0.9, 1.0, 2.3, 2.2, 2.5, 2.6, 2.1, 2.0, 0.5, 2.5, 1.4	0.0, 0.0, 0.0
T_{L4}	1.2, 0.9, 1.0, 0.7, 1.2, 0.5, 0.6, 2.2, 0.3, 0.8, 0.5, 1.2	0.0, 0.0, 0.0

In T_{L1} all auxiliary variables have t-ratios close to one and there are no omitted variables. In T_{L2} we have the same t-ratios for the auxiliary variables but now there are also omitted variables. In T_{L3} many of the auxiliary variables have 'large' t-ratios, while in T_{L4} many of the t-ratios are 'small'. Only T_{L2} has omitted variables and they are all important. We combine this larger data set with the benchmark setup, random regressor DGP, and random coefficient DGP, again under each of the three error structures. We compare WALS, Ridge, and PT$_{sw}$ with ML. We do not compute MMA because the computational burden is too high when k_2 is large.

TABLE 6

Some representative simulation results are presented in Table 6. Regarding the predictor, we see that WALS and Ridge perform best, better than ML and much better than PT$_{sw}$. The number of cases where Ridge performs best is slightly larger than the number of cases where WALS is superior. Regarding the prediction variance, WALS performs best, followed by Ridge and ML, and much better than PT$_{sw}$.

We briefly consider two other extensions, both analyzed in the context of the small data set: autocorrelated regressors and non-normality. Autocorrelation is introduced through an AR(1) process, while the non-normal regressors are obtained from a Student distribution with five degrees of freedom. We experiment (separately) with these two extensions in the benchmark model and also in models with more uncertainty. The simulation results are largely similar to the case with normal and uncorrelated regressors and therefore not reported. In particular, the WALS predictor is the most accurate when t-ratios are close to one, and the WALS prediction variance is particularly reliable when there is additional uncertainty.

12 Conclusion

This paper has introduced a new method of prediction averaging using weighted-average least squares (WALS). We have argued that pretesting—the currently dominant prediction method—is dangerous, because it ignores the noise associated with model selection. Indeed, our simulation results demonstrate that pretesting performs very badly. Model averaging is an attractive method in that it allows us to combine model selection and prediction into one procedure. Within the model averaging methods we proposed the WALS predictor and also an estimate for its variance. Our predictor explicitly allows for correlation in the errors.

We have compared the WALS predictor with four competing predictors (unrestricted ML, pretesting, ridge regression, Mallows model averaging) in a wide range of simulation experiments, where we considered not only the accuracy of the predictor (measured by the root mean squared prediction error), but also the accuracy of the prediction variance. The WALS predictor generally produces the lowest mean squared error. The estimated variance of the WALS predictor, while typically larger than the variance of the pretesting and ridge predictors, has smaller RMSE, and when model uncertainty increases the dominance of WALS becomes more pronounced. These results, together with the fact that the WALS predictor is easy to compute, suggest that the WALS predictor is a serious candidate in economic prediction and forecasting.

Acknowledgements

We are grateful to the editor, associate editor, and two referees for their insightful and constructive comments. We are also indebted to Dmitry Danilov, John Einmahl, and Paolo Paruolo for useful discussions. Zhang's work was supported by the National Natural Science Foundation of China (Grant numbers 71101141 and 11271355) and took place during a visit to Texas A&M University.

References

Aiolfi, M., Timmermann, A. (2006). Persistence of forecasting performance and conditional combination strategies. *Journal of Econometrics* 135:31–53.

Bates, J. M., Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly* 20:451–468.

Beran, R. (1995). Prediction in random coefficient regression. *Journal of Statistical Planning and Inference* 43:205–213.

Bjørnland, H. C., Gerdrup, K., Jore, A. S., Smith, C., Thorsrud, L. A. (2012). Does forecast combination improve Norges Bank inflation forecasts? *Oxford Bulletin of Economics and Statistics* 74:163–179.

Bondeson, J. (1990). Prediction in random coefficient regression models. *Biometrical Journal* 32:387–405.

Buckland, S. T., Burnham, K. P., Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* 53:603–618.

Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal* 8:277–291.

Danilov, D., Magnus, J. R. (2004a). On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122:27–46.

Danilov, D., Magnus, J. R. (2004b). Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting* 23:251–274.

Elliott, G., Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics* 122:47–79.

Froehlich, B. R. (1973). Some estimators for a random coefficient regression model. *Journal of the American Statistical Association* 68:329–335.

Giles, J. A., Giles, D. E. A. (1993). Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys* 7:145–197.

Golub, G. H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21:215–223.

Granger, C. W. J. (2003). Time series concepts for conditional distributions. *Oxford Bulletin of Economics and Statistics* 65:689–701.

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

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

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

Hendry, D. F., Clements, M. P. (2004). Pooling of forecasts. *Econometrics Journal* 7:1–31.

Hildreth, C., Houck, J. P. (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association* 63:584–595.

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

Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.

Jackson, T., Karlsson, S. (2004). Finding good predictors for inflation: A Bayesian model averaging approach. *Journal of Forecasting* 23:479–496.

Johnston, J., DiNardo, J. (1997). *Econometric Methods*, Fourth Edition. New York: McGraw-Hill.

Leeb, H. and B.M. Pötscher (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19:100–142.

Leeb, H., Pötscher, R. W. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics* 34:2554–2591.

Leeb, H., Pötscher, R. W. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24:338–376.

Liang, H., Zou, G., Wan, A. T. K., Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106:1053–1066.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications* 44:293–308.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5:225–236.

Magnus, J. R., Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67:639–643.

Magnus, J. R., Powell, O., Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* 154:139–153.

Magnus, J.R., Wan, A. T. K., Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis* 55:1331–1341.

Ohtani, K., Toyoda, T. (1980). Estimation of regression coefficients after a preliminary test for homoscedasticity. *Journal of Econometrics* 12:151–159.

Poirier, D. J. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Rubin, H. (1950). Note on random coefficients. In: Koopmans, T. C. (Ed.), *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph No. 10, pp. 419–421.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.

Stock, J. H., Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23:405–430.

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

Timmermann, A. (2006). Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, Vol. 1. Amsterdam: Elsevier, pp. 135–196.

Toyoda, T., Wallace, T. D. (1976). Optimal critical values for pre-testing in regression. *Econometrica* 44:365–375.

Wan, A. T. K., Zhang, X., Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156:277–283.

Wan, A. T. K., Zou, G. (2003). Optimal critical values of pre-tests when estimating the regression error variance: Analytical findings under a general loss structure. *Journal of Econometrics* 114:165–196.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Square Methods*. London: The English Universities Press Ltd.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20:176–222.

Table 1: RMSE of predictor relative to ML, benchmark model

N	T	WALS	PT$_{sw}$	Ridge	MMA	WALS	PT$_{sw}$	Ridge	MMA
Homoskedasticity									
		$\sigma^2 = 0.25$				$\sigma^2 = 1.00$			
100	T_1	**0.8644**	1.0638	0.8456	0.9028	**0.8686**	1.1126	0.8818	0.9385
	T_2	**0.9092**	1.0690	0.9332	0.9355	**0.8472**	1.1096	0.8555	0.9270
	T_3	**0.9998**	1.1915	1.0418	1.0757	**0.9272**	1.0781	0.9541	0.9746
	T_4	**0.9280**	1.2600	0.9474	1.0889	**0.8561**	1.2851	0.9015	1.0625
	T_5	0.8203	0.8714	**0.7597**	0.8056	0.8440	0.8936	**0.7833**	0.8284
300	T_1	**0.9007**	1.1188	0.9627	0.9782	**0.8931**	1.0796	0.9016	0.9433
	T_2	**0.9226**	1.1182	0.9394	0.9963	**0.8821**	1.1189	0.9221	0.9753
	T_3	**0.9700**	1.1603	0.9862	1.0581	**0.9438**	1.1238	0.9466	1.0165
	T_4	1.0230	1.2070	1.0631	1.1564	**0.9488**	1.1910	0.9794	1.0848
	T_5	0.8574	0.8898	**0.7937**	0.8353	0.8266	0.9055	**0.7919**	0.8275
Heteroskedasticity									
		$\tau = 0.2$				$\tau = 0.7$			
100	T_1	**0.8686**	1.0948	0.9157	0.9522	**0.9411**	1.0775	0.9958	0.9833
	T_2	**0.9757**	1.1553	1.0166	1.0379	**0.8551**	1.1081	0.8675	0.9251
	T_3	**0.9674**	1.0693	0.9761	0.9970	1.0162	1.1161	1.0486	1.0602
	T_4	**0.9147**	1.2504	0.9610	1.0918	1.0669	1.1447	1.0663	1.1058
	T_5	0.8125	0.8669	**0.7416**	0.7886	0.8636	0.8971	**0.8100**	0.8428
300	T_1	**0.8968**	1.0935	0.9180	0.9583	**0.9294**	1.0963	0.9704	0.9808
	T_2	**0.9160**	1.1105	0.9281	0.9853	**0.8862**	1.1059	0.9159	0.9635
	T_3	**0.9732**	1.1454	0.9974	1.0508	1.0883	1.1669	1.1198	1.1510
	T_4	**0.9679**	1.1768	1.0043	1.1016	**0.9803**	1.1691	1.0098	1.1050
	T_5	0.8829	0.9212	**0.8426**	0.8705	0.8361	0.8906	0.7835*	0.8212
Autocorrelation									
		$\rho = 0.3$				$\rho = 0.7$			
100	T_1	**0.9454**	1.0719	0.9754	0.9806	**0.9669**	1.0303	0.9735	0.9803
	T_2	**0.9732**	1.1574	1.0655	1.0577	**0.9761**	1.0636	1.0107	1.0109
	T_3	1.0386	1.1642	1.0877	1.1041	**0.9755**	1.0323	0.9775	0.9941
	T_4	**0.9927**	1.1961	1.0337	1.1343	**0.9609**	1.1135	0.9750	1.0379
	T_5	0.8702	0.9059	**0.8320**	0.8552	0.9507	0.9635	0.9343*	0.9436
300	T_1	**0.9160**	1.0785	0.9527	0.9651	**0.9699**	1.0388	0.9897	0.9904
	T_2	**0.9617**	1.0733	1.0041	1.0017	**0.9759**	1.0551	0.9794	1.0034
	T_3	**0.9169**	1.0321	0.9179	0.9470	1.0047	1.0614	1.0153	1.0306
	T_4	**0.9560**	1.1541	0.9862	1.0734	**0.9770**	1.0867	0.9937	1.0419
	T_5	0.8870	0.9114	**0.8373**	0.8687	0.9332	0.9617	0.9175*	0.9308

Notes: The cases where the Lasso performs best are marked with *.

Table 2: RMSE of predictor relative to ML, benchmark model with misspecified variance structure

N	T	WALS	PT$_{sw}$	Ridge	MMA	WALS	PT$_{sw}$	Ridge	MMA
		True: Heteroskedasticity vs. Model: Homoskedasticity							
		$\tau = 0.2$				$\tau = 0.7$			
100	T_1	**0.8791**	1.0779	0.8818	0.9391	0.8620	0.9946	**0.8156**	0.8805
	T_2	**0.8871**	1.1248	0.8902	0.9629	**0.8816**	1.1135	0.9068	0.9707
	T_3	1.0083	1.0942	1.0367	1.0588	**0.9668**	1.1325	0.9938	1.0305
	T_4	**0.9209**	1.2617	0.9499	1.0816	**0.9712**	1.2142	1.0127	1.0958
	T_5	0.8141	0.8625	**0.7290**	0.7901	0.8814	0.8973	**0.8262**	0.8593
300	T_1	**0.9626**	1.1009	1.0397	1.0072	0.8514	0.9944	**0.8119**	0.8774
	T_2	**0.9692**	1.0633	0.9921	1.0028	**0.8996**	1.1138	0.9138	0.9718
	T_3	**0.9791**	1.1462	1.0125	1.0631	**0.9486**	1.0815	1.0018	0.9952
	T_4	**0.9792**	1.1874	1.0179	1.1055	**0.9288**	1.2552	0.9701	1.0808
	T_5	0.9059	0.9413	**0.8808**	0.9011	0.8295	0.8637	0.7621*	0.8089
		True: Autocorrelation vs. Model: Homoskedasticity							
		$\rho = 0.3$				$\rho = 0.7$			
100	T_1	**0.9282**	1.0578	0.9463	0.9565	**0.9676**	1.0004	0.9696	0.9733
	T_2	**0.9770**	1.1246	1.0475	1.0352	**0.9559**	1.0119	0.9707	0.9647
	T_3	**0.8848**	1.0384	0.8981	0.9373	0.9796*	1.0367	1.0032	0.9952
	T_4	**0.9960**	1.1484	1.0178	1.0792	**0.9608**	1.0684	0.9627	0.9961
	T_5	0.8433	0.8593	**0.7589**	0.8087	0.9298	0.9322	0.8935*	0.9186
300	T_1	0.9336	1.0364	**0.9278**	0.9549	0.9858	0.9973	**0.9847**	0.9872
	T_2	**0.9085**	1.0604	0.9101	0.9502	**0.9854**	1.0049	0.9931	0.9891
	T_3	0.8861	1.0437	**0.8840**	0.9304	**0.9801**	1.0087	0.9813	0.9857
	T_4	**0.9352**	1.1547	0.9503	1.0388	**0.9876**	1.0316	0.9940	1.0045
	T_5	0.9198	0.9336	**0.8836**	0.9052	0.9777	0.9810	0.9690*	0.9747

Notes: The cases where the Lasso performs best are marked with *.

Table 3: RMSE of prediction variance relative to ML, benchmark model

N	T	WALS	PT$_{sw}$	Ridge	WALS	PT$_{sw}$	Ridge
				Homoskedasticity			
			$\sigma^2 = 0.25$			$\sigma^2 = 1.00$	
	T_1	**0.7705**	12.4493	2.1666	**0.7522**	11.2950	2.1449
	T_2	**0.7764**	15.1548	2.3576	**0.7713**	15.7852	2.5135
100	T_3	**0.9308**	18.6651	2.078	1.0320	18.9677	2.098
	T_4	**0.7994**	18.4211	1.5021	**0.8891**	17.9466	1.2263
	T_5	0.8860	3.9061	**0.8432**	0.9018	3.7078	**0.7918**
	T_1	1.1500	19.4101	2.9729	1.1275	17.6688	2.7914
	T_2	1.0511	22.9106	2.9839	1.0772	19.4464	2.6359
300	T_3	1.1522	25.5762	2.5351	1.2878	27.3103	2.8858
	T_4	1.0384	21.6868	1.5944	1.0836	24.784	1.5367
	T_5	1.2193	5.6784	1.2485	1.3494	5.1687	1.0811
				Heteroskedasticity			
			$\tau = 0.2$			$\tau = 0.7$	
	T_1	**0.9597**	18.4978	3.7192	1.0356	16.8279	2.9426
	T_2	**0.9841**	24.0451	3.4681	**0.9579**	19.6745	2.8953
100	T_3	**0.8782**	26.5659	3.6966	**0.9063**	25.9683	3.3147
	T_4	1.0118	28.2063	2.1269	**0.9745**	25.8139	1.9241
	T_5	**0.9999**	6.0100	1.4057	**0.9156**	4.1588	1.0565
	T_1	1.4646	21.8929	3.6805	1.1830	18.4289	3.1342
	T_2	**0.8894**	21.8063	3.4242	1.1305	21.4529	3.5049
300	T_3	1.2380	38.0474	3.7362	1.0423	25.0548	2.8665
	T_4	1.1365	25.5628	1.9827	1.0312	22.7574	1.7846
	T_5	1.3706	6.8729	1.4947	1.3548	5.1632	1.2371
				Autocorrelation			
			$\rho = 0.3$			$\rho = 0.7$	
	T_1	1.0763	2.9660	1.2644	1.0006	1.0333	1.0084
	T_2	1.0572	3.3235	1.3119	1.0020	1.0551	1.0168
100	T_3	1.0448	3.9628	1.2677	**0.9995**	1.0837	1.0208
	T_4	1.0220	3.8020	1.1374	1.0048	1.0429	1.0074
	T_5	1.0857	1.5015	1.0845	1.0025	1.0205	1.0117
	T_1	1.0090	1.6805	1.0774	**0.9967**	1.0176	1.0011
	T_2	1.0054	1.8272	1.0972	**0.9994**	1.0139	1.0003
300	T_3	**0.9969**	2.1176	1.1117	**0.9986**	1.0213	1.0006
	T_4	**0.9987**	1.9475	1.0314	1.0011	1.0144	1.0048
	T_5	1.0134	1.2023	1.0275	1.0004	1.0111	1.0071

Table 4: RMSE of prediction variance relative to ML, benchmark model with misspecified variance structure

N	T	WALS	PT_{sw}	Ridge	WALS	PT_{sw}	Ridge
		True: Heteroskedasticity vs. Model:Homoskedasticity					
			$\tau = 0.2$			$\tau = 0.7$	
100	T_1	**0.7253**	10.4231	1.8953	**0.7661**	3.4554	1.2857
	T_2	**0.7697**	12.7449	2.1457	**0.7735**	4.6074	1.5391
	T_3	**0.9335**	17.1952	1.7612	**0.8227**	5.3035	1.325
	T_4	**0.7712**	17.6055	1.2863	**0.7911**	7.2121	1.2338
	T_5	**0.8075**	3.2420	0.8858	**0.7341**	1.6100	0.8037
300	T_1	1.0160	14.3814	2.3035	**0.8610**	4.4361	1.4380
	T_2	**0.9381**	16.018	2.3322	**0.8892**	5.7947	1.5523
	T_3	1.1818	22.3939	2.2758	**0.8486**	5.9459	1.6541
	T_4	1.0859	19.5866	1.4083	**0.8569**	6.9118	1.2931
	T_5	1.2145	4.9999	1.1445	**0.8389**	2.0581	1.0155
		True: Autocorrelation vs. Model: Homoskedasticity					
			$\rho = 0.3$			$\rho = 0.7$	
100	T_1	**0.4605**	7.4888	1.7774	**0.7421**	1.7426	1.1076
	T_2	**0.5017**	9.3554	2.0489	**0.7315**	2.0768	1.2069
	T_3	**0.5758**	11.6348	1.9644	**0.7152**	2.2064	1.1793
	T_4	**0.6093**	13.4394	1.6321	**0.7654**	2.9775	1.2764
	T_5	**0.4557**	2.8689	1.0401	**0.7187**	1.3433	0.9837
300	T_1	**0.5028**	8.9684	2.1002	**0.7656**	1.8551	1.1876
	T_2	**0.5696**	11.7732	2.7547	**0.7648**	2.1344	1.2658
	T_3	**0.6350**	15.5284	2.4423	**0.7830**	2.3613	1.3001
	T_4	**0.6176**	15.1421	1.7957	**0.7954**	3.0360	1.3166
	T_5	**0.4686**	3.7499	1.3042	**0.7308**	1.5112	1.0849

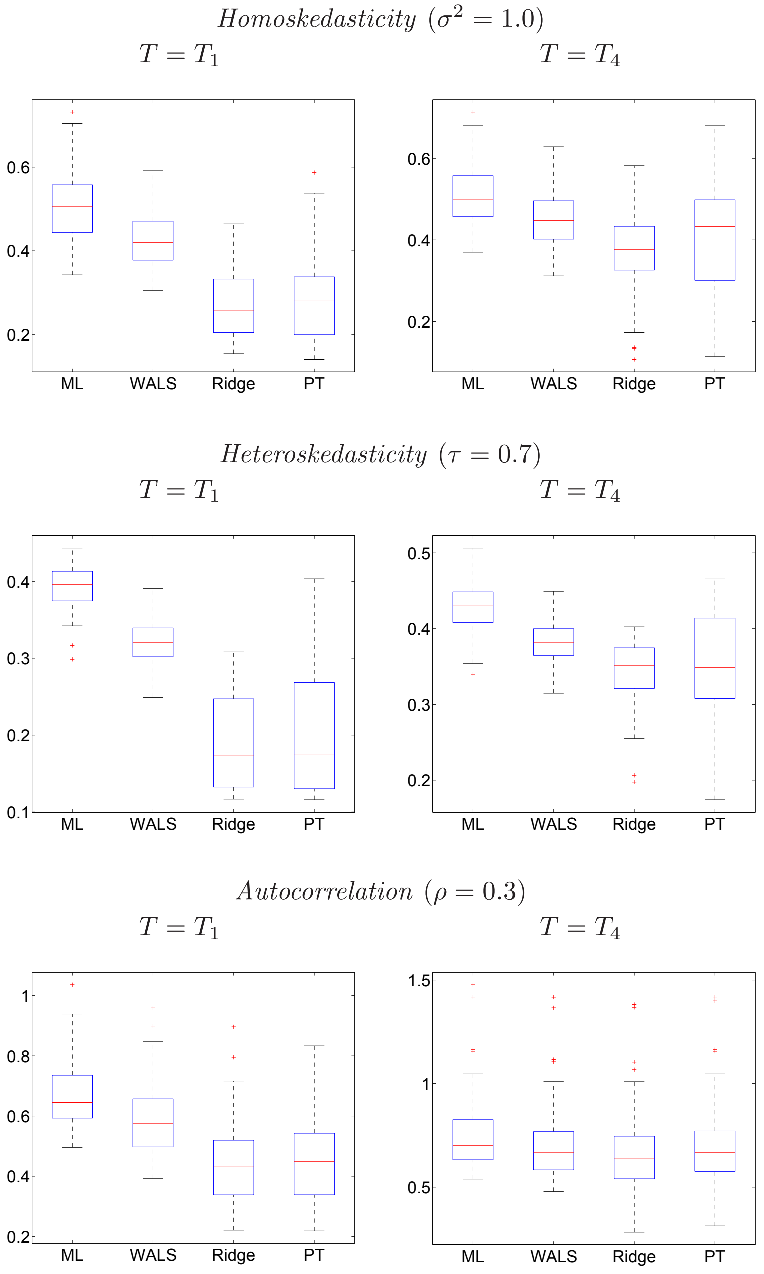
Table 5: RMSE of prediction variance relative to ML, random regressor model

N	T	WALS	PT$_{sw}$	Ridge	WALS	PT$_{sw}$	Ridge
				Homoskedasticity			
			$\sigma^2 = 0.25$			$\sigma^2 = 1.00$	
100	T_1	**0.7174**	1.0224	0.7555	**0.7169**	1.0371	0.7584
	T_2	**0.7565**	1.0243	0.7920	**0.7506**	1.0099	0.7813
	T_3	**0.8208**	1.0019	0.8408	**0.8187**	1.0047	0.8386
	T_4	**0.8520**	1.0024	0.8765	**0.8502**	0.9992	0.8760
	T_5	**0.5077**	0.9316	0.5337	**0.4996**	0.9673	0.5421
300	T_1	**0.7119**	1.0375	0.7538	**0.7161**	1.0303	0.7586
	T_2	**0.7583**	1.0123	0.7908	**0.7543**	1.0154	0.7891
	T_3	**0.8248**	1.0072	0.8435	**0.8231**	0.9993	0.8417
	T_4	**0.8555**	0.9954	0.8780	**0.8549**	0.9951	0.8771
	T_5	**0.5042**	0.9977	0.5503	**0.5018**	0.9881	0.5405
				Heteroskedasticity			
			$\tau = 0.2$			$\tau = 0.7$	
100	T_1	**0.7134**	1.0418	0.7593	**0.6708**	1.0403	0.7332
	T_2	**0.7549**	1.0181	0.7881	**0.7157**	1.0277	0.7556
	T_3	**0.8195**	1.0054	0.8386	**0.7933**	1.0140	0.8224
	T_4	**0.8483**	0.9998	0.8736	**0.8238**	0.9989	0.8466
	T_5	**0.5002**	0.9478	0.5296	**0.5114**	0.8856	0.5149
300	T_1	**0.7104**	1.0350	0.7577	**0.6824**	1.0665	0.7503
	T_2	**0.7567**	1.0109	0.7898	**0.7138**	1.0348	0.7591
	T_3	**0.8242**	1.0059	0.8456	**0.7918**	1.0115	0.8173
	T_4	**0.8515**	0.9948	0.8739	**0.8223**	1.0016	0.8417
	T_5	**0.5021**	0.9772	0.5467	**0.5219**	0.9385	0.5378
				Autocorrelation			
			$\rho = 0.3$			$\rho = 0.7$	
100	T_1	**0.7076**	1.0425	0.7543	**0.7322**	1.0715	0.8107
	T_2	**0.7472**	1.0188	0.7849	**0.7369**	1.0400	0.7898
	T_3	**0.8155**	1.0194	0.8389	**0.7844**	1.0203	0.8221
	T_4	**0.8416**	1.0003	0.8647	**0.7962**	1.0065	0.8201
	T_5	**0.5362**	0.9838	0.6091	**0.6883**	1.0634	0.8157
300	T_1	**0.7111**	1.0409	0.7612	**0.7304**	1.093	0.8164
	T_2	**0.7475**	1.0164	0.7792	**0.7465**	1.0614	0.8045
	T_3	**0.8182**	1.0101	0.8399	**0.7911**	1.0418	0.8377
	T_4	**0.8461**	0.9952	0.8667	**0.7972**	1.0176	0.8234
	T_5	**0.5359**	1.0354	0.6277	**0.7072**	1.1273	0.8631

Table 6: RMSE relative to ML, many auxiliary regressors ($N = 100$)

	Homoskedasticity			*Heteroskedasticity*			*Autocorrelation*		
	($\sigma^2 = 1.0$)			($\tau = 0.7$)			($\rho = 0.3$)		
T	WALS	PT$_{sw}$	Ridge	WALS	PT$_{sw}$	Ridge	WALS	PT$_{sw}$	Ridge
Benchmark model: fixed X, fixed β									
	Predictor								
T_{L1}	0.7591	1.2069	**0.6964**	0.8039	1.1275	**0.7702**	0.7808	1.1566	**0.7118**
T_{L2}	**0.8638**	1.1585	0.8865	0.8673	1.0734	**0.8202**	**0.9023**	1.1918	0.9102
T_{L3}	**0.8595**	1.2355	0.8926	**0.8524**	1.2380	0.8525	**0.8261**	1.1965	0.8785
T_{L4}	0.8302	1.0738	**0.7956**	0.7592	1.0172	**0.6454**	0.8086	1.0391	**0.7389**
	Prediction variance								
T_{L1}	**0.3901**	16.0771	0.7418	**0.3846**	6.3836	0.5382	**0.2605**	9.9138	0.7110
T_{L2}	1.2117	19.7130	**0.5531**	**0.6630**	9.9167	0.6657	0.7803	16.5338	**0.7236**
T_{L3}	**0.3515**	18.4750	0.7720	**0.4030**	11.791	0.7533	**0.3216**	14.9398	0.8051
T_{L4}	**0.4038**	9.1290	0.6599	**0.4031**	4.7738	0.5140	**0.2625**	6.5811	0.6292
Random regressor model: random X, fixed β									
	Predictor								
T_{L1}	0.8362	1.1644	**0.8271**	0.8293	1.1363	**0.8190**	0.8441	1.1417	**0.8380**
T_{L2}	0.9321	1.0886	**0.9321**	0.9281	1.0452	**0.9297**	0.9357	1.0851	**0.9337**
T_{L3}	**0.9029**	1.2432	0.9066	**0.8976**	1.2250	0.9104	**0.9127**	1.2372	0.9194
T_{L4}	0.8082	1.0296	**0.7860**	0.8084	1.0177	**0.7840**	0.8175	1.0266	**0.7979**
	Prediction variance								
T_{L1}	0.7624	1.0285	**0.7510**	0.7198	1.0495	**0.7098**	0.7487	1.0332	**0.7369**
T_{L2}	0.8065	1.0156	**0.8000**	0.7764	1.0297	**0.7615**	0.7933	1.0168	**0.7832**
T_{L3}	**0.8249**	1.0068	0.8389	**0.8005**	1.0178	0.8007	**0.8170**	1.0125	0.8254
T_{L4}	0.7307	1.0459	**0.7166**	0.6825	1.0724	**0.6800**	0.7181	1.0571	**0.7100**
Random coefficient model: fixed X, random β									
	Predictor								
T_{L1}	**0.9912**	0.9497	0.9954	0.9761	1.0468	**0.9717**	0.8656	1.0801	**0.8491**
T_{L2}	1.0526	1.0524	1.0487	**0.9823**	1.0302	0.9839	0.9170	1.0674	**0.9078**
T_{L3}	1.2467	1.3362	1.2592	0.9848	1.0471	**0.9784**	**0.9437**	1.1541	0.9553
T_{L4}	**0.9844**	0.9454	0.9874	0.9864	1.0338	**0.9838**	**0.8813**	1.0603	0.8878
	Prediction variance								
T_{L1}	**0.5540**	1.0402	0.6835	**0.5237**	1.0459	0.6335	**0.5418**	1.0584	0.6630
T_{L2}	**0.5484**	1.0785	0.6714	**0.5151**	1.0859	0.6276	**0.5426**	1.0837	0.6608
T_{L3}	**0.5785**	1.0762	0.7222	**0.5419**	1.0860	0.6714	**0.5700**	1.0749	0.7083
T_{L4}	**0.5478**	1.0301	0.6729	**0.5161**	1.0342	0.6200	**0.5406**	1.0454	0.6580

Figure 2: Estimated variance in the benchmark model ($N = 100$)



Homoskedasticity ($\sigma^2 = 1.0$)

$T = T_1$ $T = T_4$

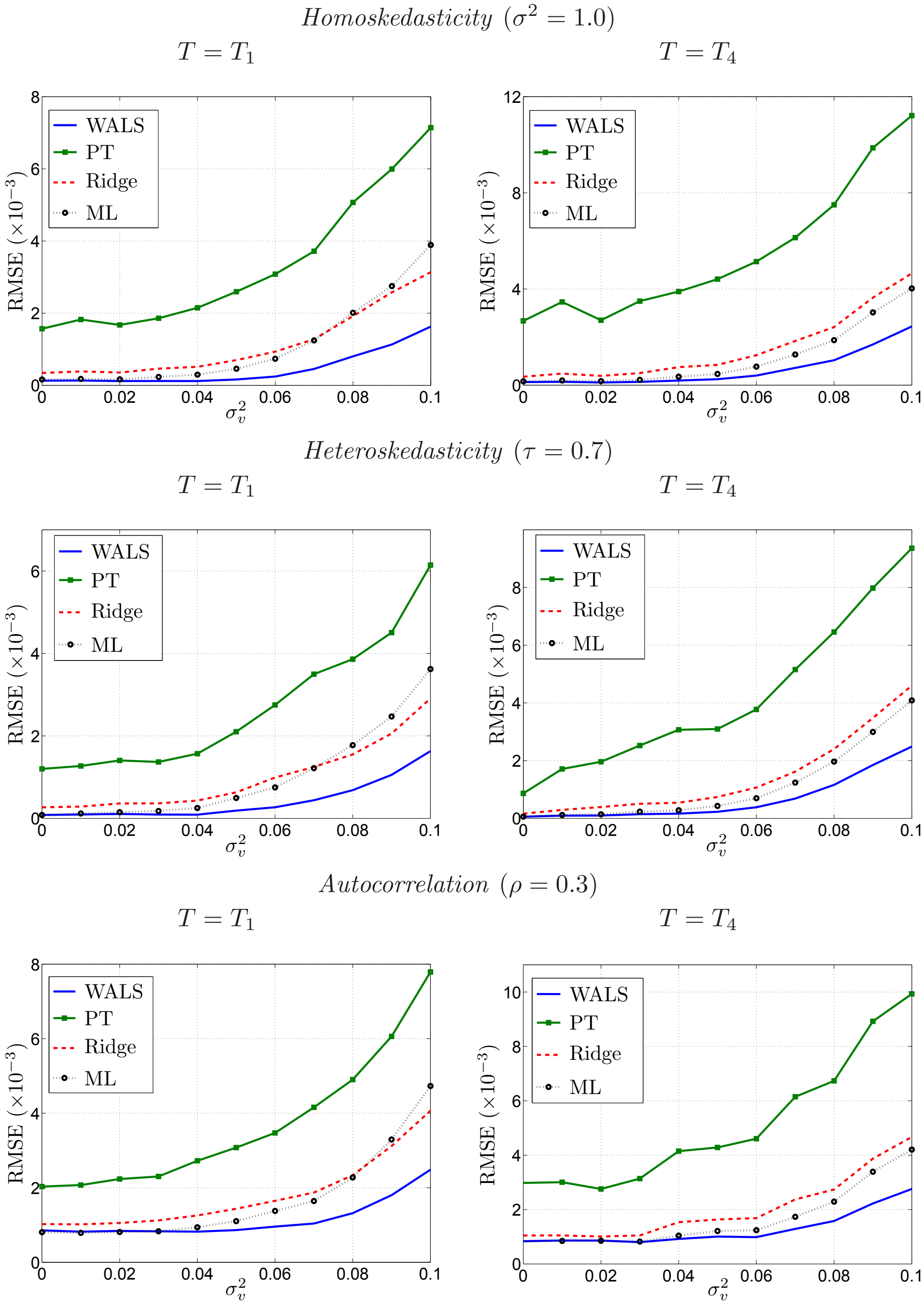
Heteroskedasticity ($\tau = 0.7$)

$T = T_1$ $T = T_4$

Autocorrelation ($\rho = 0.3$)

$T = T_1$ $T = T_4$

Figure 3: RMSE of prediction variance in random coefficient model ($N = 100$)


Homoskedasticity ($\sigma^2 = 1.0$)

$T = T_1$ $T = T_4$

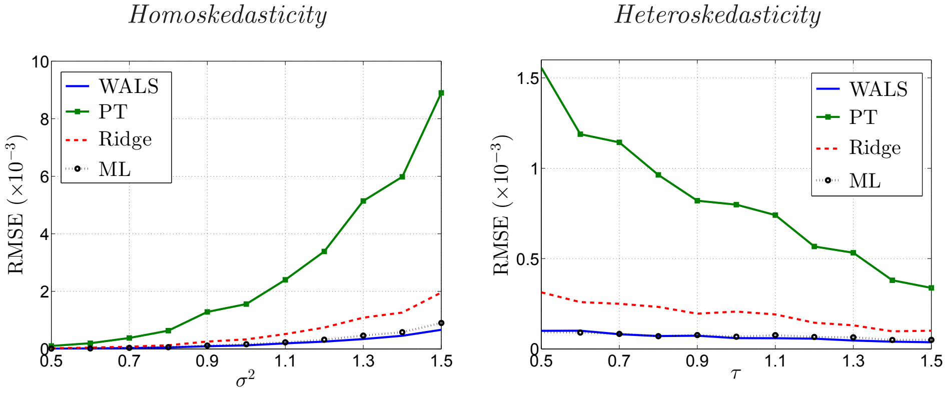
Heteroskedasticity ($\tau = 0.7$)

$T = T_1$ $T = T_4$

Autocorrelation ($\rho = 0.3$)

$T = T_1$ $T = T_4$

Figure 4: RMSE of prediction variance: homoskedastic versus heteroskedastic ($N = 100$, $T = T_1$)



Homoskedasticity

Heteroskedasticity