Unit Averaging for Heterogeneous Panels

Christian Brownlees
UPF and Barcelona GSE

Vladislav Morozov
UPF and Barcelona GSE

Contents

Motivating Example: Prediction in Linear Model

Linear heterogeneous panel:

$$y_{it} = \boldsymbol{\theta}_i' \boldsymbol{x}_{it} + u_{it}$$

Units independent, T moderate, u_{it} sequentially exogenous

Goal is prediction: estimating $\mathbb{E}(y_{1T+1}|\boldsymbol{X}_i, \boldsymbol{\theta}_i) = \boldsymbol{\theta}_1' \boldsymbol{x}_{1T+1}$ with minimal MSE (e.g. Baltagi (2013); Zhang et al. (2014); Wang et al. (2019); Liu et al. (2020))

Interest in unit-specific parameter, not population average. Other examples of unit-specific parameters: slopes (Maddala et al., 1997; Wang et al., 2019), long-run effects (Pesaran and Smith, 1995; Pesaran et al., 1999), etc.

Motivating Example: Prediction in Linear Model

How to estimate $\theta_1' \boldsymbol{x}_{1T+1}$?

- With T is small, individual estimator $\hat{\theta}_1$ is too imprecise. The only option is some kind pooling at the price of bias
- If T is large, $\hat{\theta}_1$ is good, and pooling bias is too strong $\Rightarrow$ use individual estimator
- If T is moderate, there is a nontrivial bias-variance trade-off.
  There is potential for interesting compromise estimators that use some panel-wide information.

In general, existing studies focus on long/short setups (Maddala et al., 1997; Wang et al., 2019; Liu et al., 2020).

$\Rightarrow$ We investigate exploiting the trade-off in a moderate T setting.

Motivating Example: Why Use Panel-Wide Information

Individual slope can be written

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_i, \quad \mathbb{E}(\boldsymbol{\eta}_i) = 0$$

Every unit carries information about common mean $\boldsymbol{\theta}_0$. This information is valuable for estimating $\boldsymbol{\theta}_1' \boldsymbol{x}_{1T+1} = \boldsymbol{\theta}_0' \boldsymbol{x}_{1T+1} + \boldsymbol{\eta}_1' \boldsymbol{x}_{1T+1}$. Bias-variance trade-off: using information on other units reduces uncertainty about $\boldsymbol{\theta}_0$, but creates bias due to $\boldsymbol{\eta}_i$.

Idea: consider linear combinations of all units:

$$\tilde{\boldsymbol{\theta}}_1(\boldsymbol{w}) = \sum_{i=1}^{N} w_i \hat{\boldsymbol{\theta}}_i, \quad w_i \geq 0, \sum_{i=1}^{N} w_i = 1$$

Simple average of slopes – all bias, individual estimator $\hat{\boldsymbol{\theta}}_1$ – all variance. Averaging estimator – compromise

Problem and Estimator Considered

In the paper we consider a heterogeneous M-estimation problem. Define the individual estimator for unit i as

$$\hat{\boldsymbol{\theta}}_i = \underset{\theta_i \in \Theta_i \subset \mathbb{R}^p}{\arg\min}\, T^{-1} \sum_{t=1}^{T} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})$$

Interest in estimating some smooth parameter $\mu_1 \equiv \mu(\boldsymbol{\theta}_1)$ for T small with minimal MSE (examples: policy multiplier, choice probability, etc.)

We decompose each coefficient into a mean and individual component $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_i$, $\mathbb{E}(\boldsymbol{\eta}_i) = 0$. Each unit has information about $\boldsymbol{\theta}_0$.

Unit Averaging Estimator Definition

We define unit averaging estimator with weights $\boldsymbol{w}_N$ as

$$\hat{\mu}_1(\boldsymbol{w}_N) = \sum_{i=1}^{N} w_{i,N}\mu(\hat{\boldsymbol{\theta}}_i), \quad w_{i,N} \geq 0, \sum_{i=1}^{N} w_{i,N} = 1. \tag{1}$$

Our paper: study risk properties and distribution $\hat{\mu}_1(\boldsymbol{w}_N)$ for moderate T, characterize an "optimal" choice for $\boldsymbol{w}_N$ and its properties.

Possible Averaging Estimators: Weight Example

Unit averaging estimators of form (1) are a rich class controlled by the weight vectors $\boldsymbol{w}_N$. Examples:

1. Individual estimator: $w_1 = 1$
2. Mean group/simple average: $w_i = N^{-1}$ – equal weights
3. Stein-like
$$\hat{\mu}_{Stein} = w_1 \mu(\hat{\boldsymbol{\theta}}_1) + (1 - w_1)\frac{1}{N}\sum_{i=1}^{N} \mu(\hat{\boldsymbol{\theta}}_i), \quad w_1 \in [0, 1]$$
4. Information criteria based: AIC/BIC (Buckland et al., 1997), MMA (Hansen, 2007; Wan et al., 2010), etc.
5. Optimal: $\boldsymbol{w}^o = \arg\min \text{MSE}$

Main Technique: Local Heterogeneity

Question: how to pick weights? To minimize the MSE, we need the MSE

No useful exact finite-sample results at such level of generality. Instead use an approximation by assuming local heterogeneity:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta} + \frac{\boldsymbol{\eta}_i}{\sqrt{T}}$$

Allows using asymptotic analysis techniques to approximate a finite-sample setting. Intuitively: overall signal strength is fixed as $T \to \infty$. Bias remains bounded and nontrivial. This creates a bias-variance trade-off asymptotically

Similar to frequentist model averaging approach (used by Hjort and Claeskens (2003); Claeskens and Hjort (2008)) or Hansen (2016, 2017) for shrinkage estimators

Summary of Results

- We obtain an local approximation to the MSE of the unit averaging estimator.
- There are two important averaging regimes: when N is large and when N is small.
- In both regimes deterministic weights $\Rightarrow$ estimator (1) is normal under local asymptotics.
- We provide a plug-in estimator for MSE and corresponding optimal weights. Optimal weights satisfy a regularity property – they solve the infeasible problem + noise we characterize. The estimator is not normally distributed in the limit.
- In addition, we provide a procedure to construct valid CIs with optimal weights.
- Simulations: estimator (1) performs favorably relative to equal and information criteria-based weights.
- Empirical application: nowcasting GDP in Eurozone+UK. We find notable improvements in nowcasting performance with our optimal weights (average improvement $\approx 7\%$).

Key Assumptions: Local Heterogeneity, Fixed Effects, Independence

A1 Heterogeneity is local: $\boldsymbol{\eta}_i$ is iid and

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \frac{\boldsymbol{\eta}_i}{\sqrt{T}}.$$

$\mathbb{E}\left[\boldsymbol{\eta}_i\right] = 0$ and $\boldsymbol{\eta}$ has finite 12th moments.

Fixed effects: all analysis done conditional on $\sigma(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$. Importantly, we show that all our results hold for almost any sequence $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ (i.e. with $\boldsymbol{\eta}$-probability 1).

A2 Data $\{\boldsymbol{z}_{it}, \boldsymbol{\eta}_i\}$ is independent over i

Fixed effects assumption is natural – we are interested in a realized parameter value for $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_1/\sqrt{T}$. We show that our results hold for (almost) all values of $\boldsymbol{\eta}_1$.

Key Assumptions: Smoothness Assumptions

Other assumptions are fairly standard smoothness assumption:

A3 M-estimation niceness

A4 Uniform bound on *own*-parameter bias (deterministic variation)

A5 μ is twice-differentiable with $\|\nabla^2 \mu\|$ uniformly bounded

We allow cross-sectional heterogeneity. However, we place some limits on how different the laws of different units are.

Asymptotic Distribution of Individual Estimators

Basic building block of averaging – things to be averaged.

Lemma

As $T \to \infty$, the individual estimators satisfy

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) \Rightarrow N(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1, \boldsymbol{V}_i) = \boldsymbol{Z}_i$$

$$\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow N(\boldsymbol{d}_1'\left(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right), \boldsymbol{d}_1'\boldsymbol{V}_i\boldsymbol{d}_0) = \Lambda_i$$

$\boldsymbol{d}_0$ – gradient of μ at $\boldsymbol{\theta}_0$. Convergence is joint, all $\Lambda_i, \boldsymbol{Z}_i$ are independent.

Important: $T \to \infty$ is taken in the local approximation sense of assumption A1. Amount of information is each time series is limited and not growing.

MSE of Individual Estimators

From above lemma, we get that

Lemma

Local asymptotic approximation to the MSE of using $\mu(\hat{\boldsymbol{\theta}}_i)$ as an estimator for $\mu(\boldsymbol{\theta}_1)$ is

$$LA - MSE\left(\hat{\theta}_i\right) = \boldsymbol{d}_0' \left((\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)' + \boldsymbol{V}_i\right) \boldsymbol{d}_0$$

Controlled by distance $(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)$ from unit 1 and individual variances $\boldsymbol{V}_i$. Differences between units – ground for trade-off

MSE of Unit Averaging Estimator: Statement

Theorem

Let $\{\boldsymbol{w}_N\}$ be such that (i) for each N $\boldsymbol{w}_N$ is measurable w.r.t. $\sigma(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N)$, (ii) for each N $w_{i\,N} \geq 0$ for all i, $\sum_{i=1}^{N} w_{i\,N} = 1$, $w_{j\,N} = 0$ for $j > N$, (iii) $\sup_i |w_{i\,N} - w_i| = o(N^{-1/2})$ where $\boldsymbol{w} = (w_i) \in \mathbb{R}^\infty$ is a vector such that $w_i \geq 0$ and $\sum_{i=1}^{\infty} w_i \leq 1$.

Then (i) $\sum_{i=1}^{\infty} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i$ and $\sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0$ exist; (ii) for any N and $T > T_0$ the MSE of the averaging estimator is finite; and (iii)

$$T \times MSE\left(\hat{\mu}(\boldsymbol{w}_N)\right) \to \left(\sum_{i=1}^{\infty} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i - \boldsymbol{d}_0' \boldsymbol{\eta}_1\right)^2 + \sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0$$

holds as $N, T \to \infty$ jointly.

MSE of Unit Averaging Estimator: Discussion

At its heart, the above theorem provides a local asymptotic approximation to the MSE (LA-MSE):

1. Local heterogeneity assumption allows us to use asymptotic analysis.
2. Asymptotic approximation reduces the intractable finite sample problem to the first two moments – bias and variance. These are exactly the components of interest for analyzing the MSE

There is a bias-variance trade-off driven by weights $\boldsymbol{w}_N$ and their limit $\boldsymbol{w}$. The weights:

1. $\boldsymbol{w}$ may sum to less than 1. Example: equal weights $\boldsymbol{w}_N$: $w_{iN} = N^{-1}\mathbb{I}\{i \leq N\}$. $\boldsymbol{w}_N$ converges uniformly to $\boldsymbol{w} = 0$ (this is the mean group estimator)
2. $\boldsymbol{w}_N$ has to converge uniformly to $\boldsymbol{w}$ at a given rate: used to prevent escaping weight mass which would prevent convergence.

Two Averaging Regimes

In practice there are two main "regimes" of averaging – the fixed-N an the large-N regime. The two regimes cover all practical cases of interest.

1. In the fixed-N regime, N is treated as finite and fixed.
   The fixed-N regime is appropriate when number of cross-sectional units is not large. Alternatively, it can be used when every unit averaging can potentially have an nonnegligible weight.
2. In the large-N regime, we average over an arbitrarily large collection of units. Some units will necessarily have very small weights, since N is large and weights have to be non-negative and sum to one.
   We split units into two categories: $\bar{N} < \infty$ 'unrestricted' units which can have any weights, and $N - \bar{N}$ "restricte" units which have negligible weights.

Note: "restricted" $\neq$ "unimportant", it is just a mathematical restriction on the weights.

Fixed-N Regime: LA-MSE

The fixed-N regime is obtained by assuming that there is a fixed number $\bar{N}$ of units being averaged with weights $\boldsymbol{w}^{\bar{N}}$

In the fixed-N regime the population LA-MSE is

$$LA\text{-}MSE_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) = \boldsymbol{w}^{\bar{N}\prime} \boldsymbol{\Psi}_{\bar{N}} \boldsymbol{w}^{\bar{N}} ,$$

where $\boldsymbol{\Psi}_{\bar{N}}$ is an $\bar{N} \times \bar{N}$ matrix with elements

$$[\boldsymbol{\Psi}_{\bar{N}}]_{i\,i} = \boldsymbol{d}_0' \left((\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)' + \boldsymbol{V}_i \right) \boldsymbol{d}_0,$$
$$[\boldsymbol{\Psi}_{\bar{N}}]_{i\,j} = \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) (\boldsymbol{\eta}_j - \boldsymbol{\eta}_1)' \boldsymbol{d}_0, \quad i \neq j.$$

Large-N Regime: LA-MSE

In the large-N regime the LA-MSE is controlled by a $\bar{N}$-vector $\boldsymbol{w}^{\bar{N},\infty}$ such that $w_i^{\bar{N},\infty} \geq 0$ for all i and $\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \leq 1$. The population LA-MSE is

$$LA\text{-}MSE_\infty(\boldsymbol{w}^{\bar{N},\infty}) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{\bar{N}} w_i^{\bar{N},\infty} [\boldsymbol{\Psi}_{\bar{N}}]_{ij} w_j^{\bar{N},\infty}$$
$$+ \left(\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \boldsymbol{\eta}_1 - 2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \boldsymbol{d}_0' (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \right) \left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \boldsymbol{\eta}_1 \ .$$

The unrestricted $\bar{N}$ units contribute both variance and bias. The other restricted units only contribute bias of the form $\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right) \boldsymbol{d}_0' \boldsymbol{\eta}_1$. Observe that $\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)$ is the total weight assigned to all restricted units.

Optimal Weights: Approach

The LA-MSE expressions obtained before are unknown and depend on population quantities.

To define and characterize feasible minimum MSE weights, we proceed as follows:

1. Establish asymptotic distribution of unit averaging estimator with deterministic weights.
2. Provide estimators of LA-MSE for the two regimes. Feasible minimum MSE are minimizers of estimated LA-MSE.
3. Obtain properties of the LA-MSE estimators.
4. Transfer properties to optimal weights from the LA-MSE estimators.

Distribution For Deterministic Weight Vector: Theorem

Theorem

Let $\{\boldsymbol{w}_N\}$ be such that (i) for each N $\boldsymbol{w}_N$ is measurable w.r.t. $\sigma(\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_N)$, (ii) for each N $w_{i\,N} \geq 0$ for all i, $\sum_{i=1}^N w_{i\,N} = 1$, $w_{j\,N} = 0$ for $j > N$, (iii) for some $\bar{N} \geq 0$ it holds that $\sup_{i>\bar{N}} w_{i\,N} = o(N^{-1/2})$, and (iv) $\{w_{i\,N}\}_{i=1}^{\bar{N}} \to \{w_i\}_{i=1}^{\bar{N}}$.

Then as $N, T \to \infty$ jointly

$$\sqrt{T}\left(\hat{\mu}(\boldsymbol{w}_N) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow N\left(\sum_{i=1}^{\bar{N}} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i - \boldsymbol{d}_0' \boldsymbol{\eta}_1, \sum_{i=1}^{\bar{N}} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0\right) ,$$

Distribution For Deterministic Weight Vector: Discussion

The above theorem covers both the fixed-N and the large-N regimes:

1. For fixed-N regimes all weights $w_{i\,N} = 0$ for $i \geq \bar{N}$. The requirement that $N \to \infty$ becomes superfluous.
2. For large-N regimes units beyond $\bar{N}$ are assumed to have negligible weights. Negligibility is encoded in the decay assumption (iii).

Towards an Estimator of LA-MSE

Consider the fixed-N LA-MSE associated with vector $\boldsymbol{w}^{\bar{N}}$:

$$LA\text{-}MSE_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) = \boldsymbol{w}^{\bar{N}\prime}\boldsymbol{\Psi}_{\bar{N}}\boldsymbol{w}^{\bar{N}} \ ,$$

Issue: $\boldsymbol{\Psi}$ is unknown and cannot be consistently estimated. Intuition: locality in T essentially emulates T fixed, and $\boldsymbol{\eta}_i$ are parameters which can only be estimated from individual time series. However, under locality the amount information in each time series is finite and not growing.

Next best thing – use asymptotically unbiased estimators (using the above theorems and lemmas):

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right) \Rightarrow N\left(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1, \boldsymbol{V}_i + \boldsymbol{V}_1\right) = \boldsymbol{Z}_i - \boldsymbol{Z}_1,$$

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right) \Rightarrow N(\boldsymbol{\eta}_1, \boldsymbol{V}_1) = \boldsymbol{Z}_1 + \boldsymbol{\eta}_1.$$

Fixed-N: LA-MSE Estimator

The fixed-N LA-MSE estimator associated with $\boldsymbol{w}^{\bar{N}}$ is given by

$$\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{\bar{N}} w_i^{\bar{N}} [\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} w_j^{\bar{N}} ,$$

where $\hat{\boldsymbol{\Psi}}_{\bar{N}} \in \mathbb{R}^{\bar{N}\times\bar{N}}$ with entries

$$[\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{ii} = \nabla\mu(\hat{\boldsymbol{\theta}}_1)'(T(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)' + \hat{\boldsymbol{V}}_i)\nabla\mu(\hat{\boldsymbol{\theta}}_1),$$
$$[\hat{\boldsymbol{\Psi}}_N]_{ij} = \nabla\mu(\hat{\boldsymbol{\theta}}_1)' T(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_1)'\nabla\mu(\hat{\boldsymbol{\theta}}_1), i \neq j$$

and $\hat{\boldsymbol{V}}_i$ is an estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_i$.

Fixed-N: Minimum MSE Weights

The fixed-N minimum MSE weights are

$$\hat{\boldsymbol{w}}^{\bar{N}} = \operatorname*{arg\,min}_{\boldsymbol{w}\in\Delta^{\bar{N}}} \widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}) \ ,$$

where $\Delta^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : \sum_{i=1}^{\bar{N}} w_i = 1, w_i \geq 0, i = 1, \dots, \bar{N}\}$.
This is a quadratic programming problem.

Large-N: LA-MSE Estimator

Let $\boldsymbol{w}^{\bar{N},\infty} = (w_i^{\bar{N},\infty})$ be a $\bar{N}$-vector such that $w_i^{\bar{N},\infty} = w_i^{N,\infty}$ for $i = 1, \dots, \bar{N}$. The large-N LA-MSE estimator associated with $\boldsymbol{w}^{N,\infty}$ is controlled by $\boldsymbol{w}^{\bar{N},\infty}$ and given by

$$
\begin{aligned}
&\widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) \\
&= \sum_{i=1}^{\bar{N}}\sum_{j=1}^{\bar{N}} w_i^{\bar{N},\infty}[\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} w_j^{\bar{N},\infty} + \Bigg[\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\left(\sqrt{T}\nabla\mu(\hat{\boldsymbol{\theta}}_1)'\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right)\right) \\
&\quad -2\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\nabla\mu(\hat{\boldsymbol{\theta}}_1)\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right)\Bigg]\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\left(\sqrt{T}\nabla\mu(\hat{\boldsymbol{\theta}}_1)'\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right)\right)
\end{aligned}
$$

Large-N: Minimum MSE Weights

The large-N minimum MSE weights $\hat{\boldsymbol{w}}^{N,\infty} = (\hat{w}_i^{N,\infty})$ are given by

$$\hat{w}_i^{N,\infty} = \begin{cases} \hat{w}_i^{\bar{N},\infty} & i \leq \bar{N} \\ \left(1 - \sum_{j=1}^{\bar{N}} w_j^{\bar{N},\infty}\right)(N-\bar{N})^{-1} & i > \bar{N} \end{cases}$$

where

$$\hat{\boldsymbol{w}}^{\bar{N},\infty} = \operatorname*{arg\,min}_{\boldsymbol{w} \in \tilde{\Delta}^{\bar{N}}} \widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w})$$

with $\tilde{\Delta}^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : w_i \geq 0, \sum_{i=1}^{N} w_i \leq 1\}$. This is also a quadratic optimization problem. For simplicity, we assign equal weights to the restricted units (though many choices would lead to the same asymptotic result)

Fixed-N: Asymptotic Properties for Minimum MSE Weights

Theorem (Fixed-N Minimum MSE Unit Averaging)

1. For any $\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}$ it holds that $\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) := \boldsymbol{w}^{\bar{N}\prime}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N}}$ as $T \to \infty$, where $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ is an $\bar{N} \times \bar{N}$ matrix with elements $[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} = \boldsymbol{d}_1'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_i - \boldsymbol{Z}_1)' + \boldsymbol{V}_i)\boldsymbol{d}_1$ when $i = j$ and $\boldsymbol{d}_1'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)')\boldsymbol{d}_1$ when $i \neq j$; and $\boldsymbol{Z}_i$ is as in Lemma 3.1.
2. The minimum MSE weights satisfy as $T \to \infty$.

$$\hat{\boldsymbol{w}}^{\bar{N}} = \underset{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}}{\arg\min}\, \widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{\boldsymbol{w}}^{\bar{N}} = \underset{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}}{\arg\min}\, \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}})$$

3. The minimum MSE unit averaging estimator satisfies

$$\sqrt{T}\left(\hat{\mu}(\hat{\boldsymbol{w}}^{\bar{N}}) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow \sum_{i=1}^{M} \overline{w}_i^{\bar{N}} \Lambda_i$$

Fixed-N: Asymptotic Properties for Minimum MSE Weights: Discussion I

Some comments on the above theorem:

1. $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ in be written as as

$$[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ii} = [\boldsymbol{\Psi}_{\bar{N}}]_{ii} + \boldsymbol{d}_1'(\boldsymbol{V}_1 + \boldsymbol{V}_i)\boldsymbol{d}_1 + \boldsymbol{d}_1\boldsymbol{e}_{ii}\boldsymbol{d}_1$$
$$[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} = [\boldsymbol{\Psi}_{\bar{N}}]_{ij} + \boldsymbol{d}_1'\boldsymbol{V}_1\boldsymbol{d}_1 + \boldsymbol{d}_1'\boldsymbol{e}_{ij}\boldsymbol{d}_1, \quad i \neq j,$$

where $\boldsymbol{e}_{ij} = (\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)' - \mathbb{E}((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)')$. Mean zero noise terms $\boldsymbol{e}_{ij}$ – impact of limited information in moderate T setting. $\boldsymbol{V}_1$ terms – price of having a positive definite LA-MSE estimator.

Conclusion: minimum MSE solve the ideal population problem of minimizing $\boldsymbol{w}^{\bar{N}'}\boldsymbol{\Psi}_{\bar{N}}\boldsymbol{w}^{\bar{N}}$ subject to some additional noise and bias.

Fixed-N: Asymptotic Properties for Minimum MSE Weights: Discussion II

2. $\overline{LA\text{-}MSE}_{\bar{N}}$ plays the same role to $\widehat{LA\text{-}MSE}_{\bar{N}}$ as $\boldsymbol{Z}_i$ does to $\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)$. $\overline{LA\text{-}MSE}_{\bar{N}}$ expresses $\widehat{LA\text{-}MSE}_{\bar{N}}$ in terms of the components of the MSE and the approximate distribution of the individual estimators.
3. The unit averaging estimator with minimum MSE weights has a nonstandard asymptotic distribution in the local heterogeneity framework. The limit distribution is a randomly weighted sum of independent normal random variables. In the appendix to the paper we discuss how to construct a correctly-sized confidence interval on the basis of the unit averaging estimator with minimum MSE weights.

Minimal MSE Weights In a Large T Setting

It is natural to minimize $\widehat{LA\text{-}MSE}_N$ is natural even in a non-local setting with no assumption A1 and with growing amount of information:

1. For all i with $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_1$, the bias estimators $\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)$ will diverge
2. Variance terms remain bounded.

$\Rightarrow$ Procedure will place asymptotically zero weight on all units with $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_1$.

Parallels a similar result in model averaging where (see Fang et al. (2022))

Large-N: Asymptotic Properties for Minimum MSE Weights I

A very similar result applies in the large-N regime:

1. For any $\boldsymbol{w}^{\bar{N},\infty} \in \tilde{\Delta}^{\bar{N}}$ it holds that $\widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) \Rightarrow \overline{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty})$ as $N, T \to \infty$ jointly where

$$\overline{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) = \boldsymbol{w}^{\bar{N},\infty\prime}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N},\infty} + \left[\left(1-\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\left(\eta_1+\boldsymbol{Z}_1\right) - 2\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\boldsymbol{d}_0'\left(\boldsymbol{Z}_i-\boldsymbol{Z}_1\right)\right]\left(1-\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\left(\eta_1+\boldsymbol{Z}_1\right).$$

Large-N: Asymptotic Properties for Minimum MSE Weights II

2. The minimum MSE weights satisfy

$$\hat{\boldsymbol{w}}^{\bar{N},\infty} = \arg\min_{\boldsymbol{w}\in\tilde{\Delta}^{\bar{N}}} \widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}) \Rightarrow \overline{\boldsymbol{w}}^{\bar{N},\infty} = \arg\min_{\boldsymbol{w}\in\tilde{\Delta}^{\bar{N}}} \overline{LA\text{-}MSE}_{\infty}(\boldsymbol{w})$$

as $N, T \to \infty$.

Large-N: Asymptotic Properties for Minimum MSE Weights III

3. Let $\boldsymbol{v}_{N-\bar{N}} = (v_{\bar{N}\,N}, \dots, v_{N\,N})$ be a $(N-\bar{N})$-vector such that $\sup_i v_{i\,N-\bar{N}} = o(N^{-1/2})$, $v_{i\,N-\bar{N}} \geq 0$, for each N it holds that $\sum_{i=N-\bar{N}}^{N} v_{i\,N-\bar{N}} = 1$. Then

$$\sqrt{T}\left(\sum_{i=1}^{\bar{N}} \hat{w}_i^{\bar{N},\infty}\mu\left(\hat{\boldsymbol{\theta}}_i\right) + \left(1 - \sum_{i=1}^{\bar{N}} \hat{w}_i^{\bar{N},\infty}\right)\sum_{j=N-\bar{N}}^{N} v_{j\,N-\bar{N}}\mu(\hat{\boldsymbol{\theta}}_j) - \mu(\boldsymbol{\theta}_1)\right)$$

$$\Rightarrow \sum_{i=1}^{\bar{N}} \overline{w}_i^{\bar{N},\infty}\Lambda_i - \left(1 - \sum_{i=1}^{\bar{N}} \overline{w}_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\boldsymbol{\eta}_1$$

as $N, T \to \infty$ jointly.

Simulation: Setup

Consider a simple dynamic panel:

$$y_{it} = \lambda_i y_{it-1} + \beta_i x_{it} + u_{it}$$
$$\lambda_i = \mathbb{E}(\lambda_i) + \eta_i, \quad \mathbb{E}(\lambda_i) = 0$$

We compare performance of minimum MSE weights compared to AIC/BIC (Buckland et al., 1997), mean group (equal weights), MMA Hansen (2007)

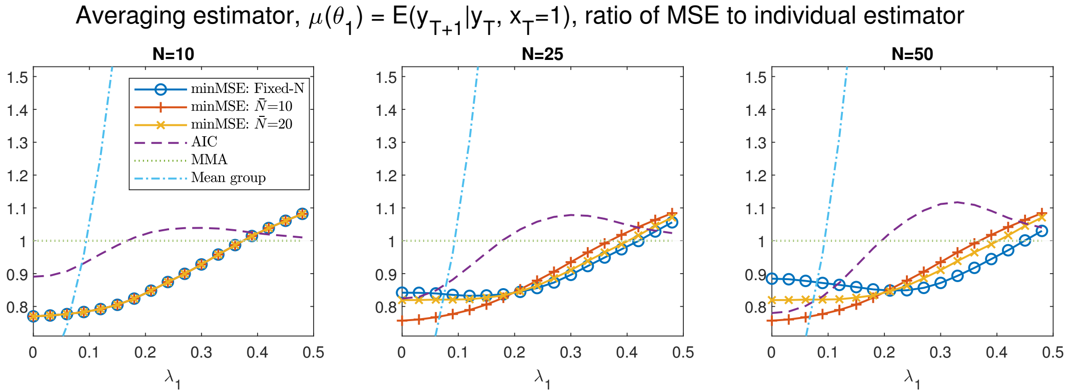
We compute ratio of MSE of different weighting schemes relative to the individual estimator of unit 1. Value above 1 means worse performance, below 1 better performance. MSE computed for a range of $\boldsymbol{\eta}_1$ such that λ_1 ranges from -0.5 to 0.5.

Simulation: Parameters of Interest

An important feature of our minimum MSE weights – they target a specific focus parameter μ. Different parameters will have difference weights (unlike other averaging schemes).

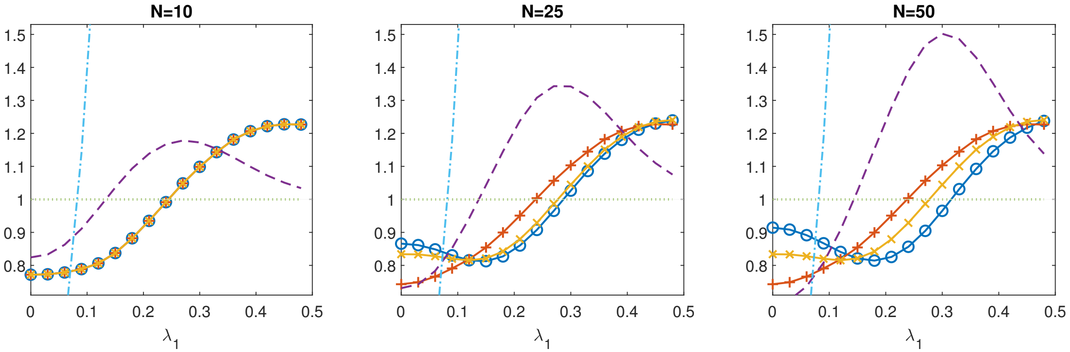
We conduct simulations for

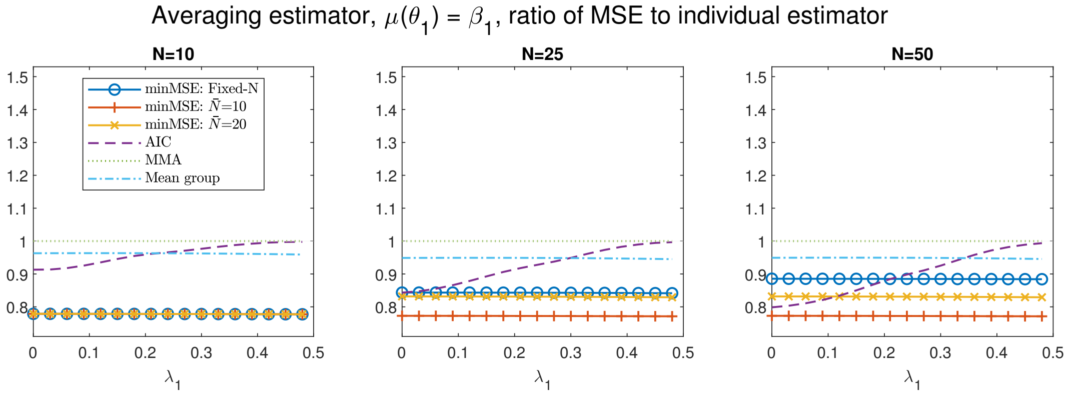
1. One-step head forecast for y
2. λ_1
3. β_1
4. Long-run effect of a change in x: $\beta_1/(1-\lambda_1)$

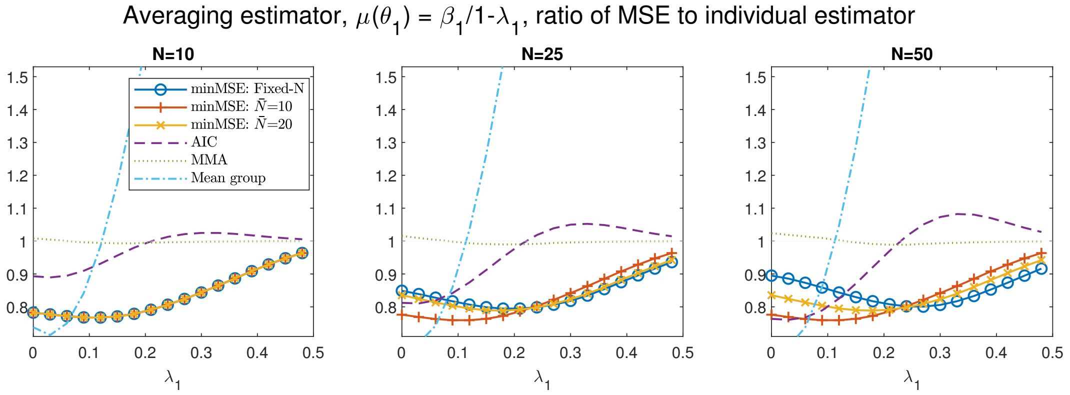


Averaging estimator, $\mu(\theta_1)$ = E(y_{T+1}|y_T, x_T=1), ratio of MSE to individual estimator
N=10
N=25
N=50
minMSE: Fixed-N
minMSE: $\bar{N}$=10
minMSE: $\bar{N}$=20
AIC
MMA
Mean group
λ_1

Averaging estimator, $\mu(\theta_1) = \lambda_1$, ratio of MSE to individual estimator




Averaging estimator, $\mu(\theta_1) = \beta_1$, ratio of MSE to individual estimator
N=10
N=25
N=50
minMSE: Fixed-N
minMSE: $\bar{N}$=10
minMSE: $\bar{N}$=20
AIC
MMA
Mean group
λ_1



Averaging estimator, $\mu(\theta_1) = \beta_1/1-\lambda_1$, ratio of MSE to individual estimator
N=10
N=25
N=50
minMSE: Fixed-N
minMSE: $\bar{N}$=10
minMSE: $\bar{N}$=20
AIC
MMA
Mean group
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0
0.1
0.2
0.3
0.4
0.5
λ_1

Empirical Application: GDP Nowcasting

Empirical application – nowcasting GDP for founding members of the Eurozone + UK.
Natural application for unit averaging:

1. Evidence of significant cross-country heterogeneity (Marcellino et al., 2003)
2. Partial pooling of data may improve performance (Garcia-Ferrer et al., 1987; Hoogstrate et al., 2000)

Nowcasting Study Setup I

We follow standard practices in nowcasting literature (e.g. Schumacher (2016)):

1. We account for delays in data publication ("ragged edge" problem (Wallis, 1986)) by having a stylized data release calendar.
2. There are different possible times for making forecasting for current quarter GDP ("vintages"). Nowcasting is possible between first day of quarter (-12 weeks to quarter end) up to one month after quarter ($+4$ weeks to quarter end).
3. Nowcasting is done using factor unrestricted MIDAS (Foroni et al., 2015). Factors estimated by EM-PCA (Stock and Watson, 1999).
   In the appendix we also consider factor bridge equations (Giannone et al., 2008; Rünstler et al., 2009) with similar results.

Nowcasting Study Setup II

4. Factors estimated using both real, financial, and survey data (up to $\approx$ 160 variables per country)
5. Estimation is done using a rolling window. Three sample sizes considered – 44 quarters, 60 quarters, 76 quarters

Empirical Application: Summary of Results

1. Using smooth data-dependent averaging weights (minimum MSE and AIC) leads to improvement in nowcasting performance.
   Using MMA and equal (mean group) weights does not lead to improvements.
2. Averaging is most beneficial for smaller samples – magnitude of improvement is shrinking as sample size increases. This is intuitive: averaging estimators converge to the individual estimator
3. Gains from averaging are heterogeneous across countries

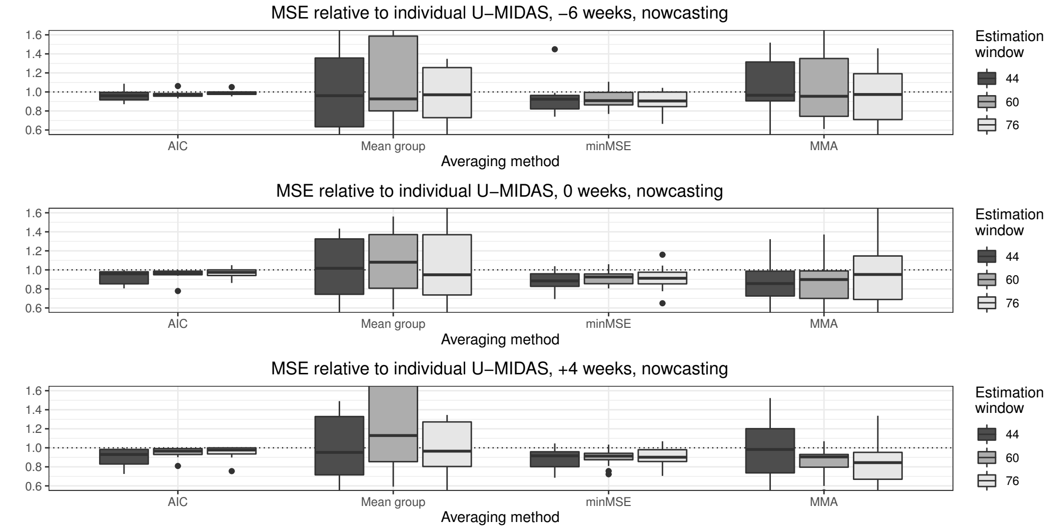


Figure: Distribution of MSEs across countries. MSE relative to individual estimator. Split by different averaging strategies and estimation window size (quarters). Nowcasting done at selected positions relative to end of quarter

Conclusions

- If T is moderate, there is bias-variance tradeoff in using other units.
- We propose and study an averaging estimator that uses all units in the panel
- We characterize a local asymptotic approximation to the MSE of averaging. There are two key regimes of averaging depending on magnitude of N and our approach.
- Minimum MSE solve the population MSE minimization problem + rank-preserving bias + mean-zero noise.
- Unit averaging with deterministic weights: asymptotic normality. With random weights: randomly weighted sum of normals.
- Estimator performs competently in simulations.
- Empirical application to GDP nowcasting: there gains from using unit averaging, especially minimum MSE weights.

References I

B. H. Baltagi. *Panel data forecasting*, volume 2. Elsevier B.V., 2013.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model Selection: An Integral Part of Inference. *Biometrics*, 53(2):603–618, 1997.

G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, 2008.

F. Fang, C. Yuan, and W. Tian. An Asymptotic Theory for Least Squares Model Averaging with Nested Models. *Econometric Theory*, pages 1–30, 2022.

C. Foroni, M. Marcellino, and C. Schumacher. Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials. *Journal of the Royal Statistical Society: Series A*, 178(1):57–82, 2015.

References II

A. Garcia-Ferrer, R. A. Highfield, F. C. Palm, and A. Zellner. Macroeconomic Forecasting Using Pooled International Data. *Journal of Business and Economic Statistics*, 5(1):53–67, 1987.

D. Giannone, L. Reichlin, and D. Small. Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 55(4):665–676, 2008.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.

B. E. Hansen. Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1): 115–132, 2016.

B. E. Hansen. Stein-like 2SLS estimator. *Econometric Reviews*, 36(6-9):840–852, 2017.

References III

N. L. Hjort and G. Claeskens. Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

A. J. Hoogstrate, F. C. Palm, and G. A. Pfann. Pooling in Dynamic Panel-Data Models: An Application to Forecasting GDP Growth Rates. *Journal of Business and Economic Statistics*, 18(3):274–283, 2000.

L. Liu, H. R. Moon, and F. Schorfheide. Forecasting with Dynamic Pane Data Models. *Econometrica*, 88(1):171–201, 2020.

G. S. Maddala, R. P. Trost, H. Li, and F. Joutz. Estimation of Short-Run and Long-Run Elasticities of Energy Demand From Panel Data Using Shrinkage Estimators. *Journal of Business and Economic Statistics*, 15(1):90–100, 1997.

References IV

M. Marcellino, J. H. Stock, and M. W. Watson. Macroeconomic Forecasting in the Euro Area: Country Specific Versus Area-Wide Information. *European Economic Review*, 47 (1):1–18, 2003.

M. H. Pesaran and R. P. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 6061:473–477, 1995.

M. H. Pesaran, Y. Shin, and R. P. Smith. Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association*, 94(446): 621–634, 1999.

G. Rünstler, K. Barhoum, S. Benk, R. Cristadoro, A. D. Reijer, A. Jakaitiene, P. Jelonek, A. Rua, K. Ruth, and C. van Nieuwenhuyze. Short-Term Forecasting of GDP Using Large Datasets: A Pseudo Real-Time Forecast Evaluation Exercise. *Journal of Forecasting*, 28(7):595–611, 2009.

References V

C. Schumacher. A Comparison of MIDAS and Bridge Equations. *International Journal of Forecasting*, 32(2):257–270, 2016.

J. H. Stock and M. W. Watson. A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. In R. F. Engle and H. White, editors, *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*, pages 1–44. Oxford University Press, 1999.

K. F. Wallis. Forecasting with an Econometric Model: The 'Ragged Edge' Problem. *Journal of Forecasting*, 5(1):1–13, 1986.

A. T. K. Wan, X. Zhang, and G. Zou. Least Squares Model Averaging by Mallows Criterion. *Journal of Econometrics*, 156(2):277–283, 2010.

References VI

W. Wang, X. Zhang, and R. Paap. To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions? *Journal of Applied Econometrics*, 34(5):724–745, 2019.

X. Zhang, G. Zou, and H. Liang. Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101(1):205–218, 2014.