UNIT AVERAGING
FOR HETEROGENEOUS PANELS

Christian Brownlees† Vladislav Morozov†‡∗

May 10, 2024

Abstract

In this work we introduce a unit averaging procedure to efficiently recover unit-specific parameters in a heterogeneous panel model. The procedure consists in estimating the parameter of a given unit using a weighted average of all the unit-specific parameter estimators in the panel. The weights of the average are determined by minimizing an MSE criterion we derive. We analyze the properties of the resulting minimum MSE unit averaging estimator in a local heterogeneity framework inspired by the literature on frequentist model averaging, and we derive the local asymptotic distribution of the estimator and the corresponding weights. The benefits of the procedure are showcased with an application to forecasting unemployment rates for a panel of German regions.

Keywords: heterogeneous panels, frequentist model averaging, prediction

JEL: C33, C52, C53

1 Introduction

Estimation of unit-specific parameters in panel data models with heterogeneous parameters is a topic of active research in econometrics (Maddala, Trost, Li, and Joutz, 1997; Pesaran, Shin, and Smith, 1999; Wang, Zhang, and Paap, 2019; Liu, Moon, and Schorfheide, 2020). Estimation of unit-specific parameters is relevant, for instance, when interest lies in constructing forecasts for the individual units in the panel (Baltagi, 2013; Zhang, Zou, and

---

†† Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Economics; e-mail: christian.brownlees@upf.edu;
‡‡ Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Economics; e-mail: vladislav.morozov@barcelonagse.eu. Corresponding author.
We thank Jan Ditzen, Kirill Evdokimov, Geert Mesters, Luca Neri, Katerina Petrova, Barbara Rossi, Wendun Wang, and the participants at 26th IPDC, 7th RCEA Time Series Workshop, 9th SIDE WEEE, the 2021 ERFIN workshop, and the 2023 EEA-ESEM conference for comments and discussion. Christian Brownlees acknowledges support from the Spanish Ministry of Science and Technology (Grant MTM2012-37195) and the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075).
Liang, 2014; Wang et al., 2019; Liu et al., 2020), which typically arises in the analysis of international panels of macroeconomic time series (Marcellino, Stock, and Watson, 2003). Other unit-specific parameters of interest include individual coefficients (Maddala et al., 1997; Maddala, Li, and Srivastava, 2001; Wang et al., 2019) and long-run effects of a change in a covariate (Pesaran and Smith, 1995; Pesaran et al., 1999).

There are three natural strategies for estimating unit-specific parameters (Baltagi, Bresson, and Pirotte, 2008). The simplest approach consists in estimating each unit-specific parameter from its individual time series. While this strategy typically leads to approximately unbiased estimation, such estimators suffer from large estimation variability when the time dimension is small. In the second approach, an assumption of parameter homogeneity is imposed and a common panel-wide estimator is used for all unit-specific parameters. This strategy leads to small variability; however, it suffers from large bias in the presence of heterogeneity. The third strategy is a compromise between the first two. It uses panel-wide information to reduce the variability of the individual estimator to obtain an estimator with favorable risk properties (Maddala et al., 2001; Wang et al., 2019; Liu et al., 2020). This is appealing when the time dimension is moderate in the sense that there is a nontrivial bias-variance trade-off between individual-specific and panel-wide estimation.

In this paper we propose a novel compromise estimator for unit-specific “focus” parameters — the unit averaging estimator. Focus parameters considered are smooth transformations of unit-specific parameters, including the examples mentioned above. The unit averaging estimator for the unit-specific focus parameter is defined as a weighted average of all the unit-specific focus parameter estimators in the panel. The weights are chosen by minimizing one of the two unit-specific mean squared error (MSE) criteria we derive. One of the criteria can leverage prior information about similarities between cross-sectional units in terms of their parameters. The other criterion is agnostic and requires no prior information. In both cases, the weights solve a straightforward quadratic optimization problem. The estimator is fairly general and is designed for possibly nonlinear and dynamic panel models estimated by M-estimation.

We analyze the theoretical properties of the our unit averaging methodology. We focus on a moderate-\(T\) setting — a setting in which the amount of information in each time series is limited and the variance of individual estimators is of the same order of magnitude as the coefficients. In this setting, we derive the leading terms of the MSE of the unit averaging estimator. We do so using a limited information local asymptotic technique under a local heterogeneity framework, in which the unit-specific coefficients are local in the time dimension to a common mean. This theoretical device emulates a moderate-\(T\) setting and the trade-off between unit-specific and panel-wide information. It is inspired by the local misspecification technique used in the frequentist model averaging literature for analyzing
finite-sample properties of estimators (Hjort and Claeskens, 2003a; Liu, 2015; Hansen, 2016).

We propose and analyze minimum MSE weights that minimize an estimator of the leading terms of the MSE. As we show, these minimum MSE weights minimize an appropriately defined notion of the population MSE contaminated by a noise component that we characterize explicitly. We obtain the limiting distribution of the minimum MSE unit averaging estimator in a local heterogeneity setting, similarly to Liu (2015). Finally, we argue that the minimum MSE weights also have desirable properties in a large-$T$ setting, in which the amount of information in each time series grows without bound.

In a simulation study, we assess the finite sample properties of the our methodology. We compare our minimum MSE unit averaging estimator against the unit-specific and mean group estimators, along with AIC and BIC weighted averaging estimators (Buckland, Burnham, and Augustin, 1997). The proposed methodology performs favorably relative to these benchmarks. Gains in the MSE are possible without prior information about unit similarity. However, leveraging prior information may lead to stronger improvements.

An application to forecasting regional unemployment in Germany showcases the methodology (Schanne, Wapler, and Weyh, 2010). Unemployment forecasting is a natural application of the unit averaging methodology since the literature documents both evidence of regional heterogeneity and the benefits of pooling data (Schanne et al., 2010; de Graaff, Arribas-Bel, and Ozgen, 2018). We find that unit averaging using minimum MSE weights improves prediction accuracy. The gains in the MSE are larger for shorter panels.

This paper is related to two strands of the literature. First, it contributes to the literature on estimation of unit-specific parameters. Important contributions in this area include Zhang et al. (2014), Wang et al. (2019), Issler and Lima (2009) and Liu et al. (2020). In contrast to these contributions, we focus on a setting where the time dimension is moderate (as opposed to either large or small). Moreover, the existing literature largely focuses on linear models under strict exogeneity (Baltagi et al., 2008; Wang et al., 2019) whereas our framework allows for nonlinear and dynamic models. Second, our paper is related to the literature on frequentist model averaging. Important contributions in this area include Hjort and Claeskens (2003a), Hansen (2007), Hansen (2008), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), Liu (2015), and Gao, Zhang, Wang, and Zou (2016), among others. Gao et al. (2016); Yin, Liu, and Lin (2021) deal with model averaging estimators specifically tailored for panel models. The main difference with respect to these contributions is that we focus on averaging different units with the same model whereas these papers average different models for a given fixed unit or the pooled data.

The rest of the paper is structured as follows. Section 2 introduces the unit averaging methodology. Section 3 studies the theoretical properties of the procedure. Section 4 contains the simulation study. Section 5 contains the empirical application. Concluding
remarks follow in section 6. All proofs are collected in the proof appendix. Further theoretical, numerical, and empirical results are collected in an online appendix.

2 Methodology

We introduce our unit averaging methodology within the framework of a fairly general class of panel data models with heterogeneous parameters. Let \( \{z_{it}\} \) with \( i = 1, \ldots, N \) and \( t = 1, \ldots, T \) denote a panel where \( z_{it} \) denotes a random vector of observations taking values in \( Z \subset \mathbb{R}^d \). For each unit in the panel, we define the unit-specific parameter \( \theta_i \in \Theta \subset \mathbb{R}^p \) as

\[
\theta_i = \arg \max_{\theta \in \Theta} \mathbb{E} \left( \frac{1}{T} \sum_{t=1}^{T} m(\theta, z_{it}) \right),
\]

where \( m : \Theta \times Z \to \mathbb{R} \) is a smooth criterion function.

Our interest lies in estimating the unit-specific “focus” parameter \( \mu(\theta_i) \) for a fixed unit \( i \) with minimal MSE, where \( \mu : \Theta \to \mathbb{R} \) is a smooth function (similarly to the setup in Hjort and Claeskens (2003a)). For example, \( \mu(\theta_i) \) may denote a component of \( \theta_i \), the conditional mean of a response variable given the covariates, or the long-run effect of a covariate. To simplify exposition and without loss of generality, we focus on the problem of estimating the focus parameter \( \mu(\theta_1) \) for unit 1. In this paper we consider the case in which the focus function \( \mu \) is scalar-valued. It is straightforward to generalize the framework to a focus function taking values in \( \mathbb{R}^q \) for some \( q > 1 \).

To estimate \( \mu(\theta_1) \) we consider the class of unit averaging estimators given by

\[
\hat{\mu}(w) = \sum_{i=1}^{N} w_i \mu(\hat{\theta}_i),
\]

where \( w = (w_i) \) is a \( N \)-vector such that \( w_i \geq 0 \) for all \( i \) and \( \sum_{i=1}^{N} w_i = 1 \), and \( \hat{\theta}_i \) is the unit-specific estimator of unit \( i = 1, \ldots, N \), given by

\[
\hat{\theta}_i = \arg \max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} m(\theta, z_{it}).
\]

The class of estimators in (1) is fairly broad and contains a number of important special cases. It includes the individual estimator of unit 1 \( \hat{\mu}_1 = \mu(\hat{\theta}_1) \) and the mean group estimator \( \hat{\mu}_{MG} = N^{-1} \sum_{i=1}^{N} \mu(\hat{\theta}_i) \). It also includes estimators based on smooth AIC/BIC weights (Buckland et al., 1997) as well as Stein-type estimators (Maddala et al., 1997).

The class of estimators in (1) may be motivated by the following representation for the
individual parameters $\theta_i$. Assume that $\theta_i$ can be written as $\theta_i = \theta_0 + \eta_i$, where $\theta_0$ is a common mean component and $\eta_i$ is a zero-mean random component. All units in the panel carry information on $\theta_0$, and so all units may be useful for estimating $\theta_1 = \theta_0 + \eta_1$. The vector of weights $w$ controls the balance between the bias and the variance of estimator (1). Assigning a large weight to unit 1 leads to low bias but may also lead to excessive variability. Alternatively, assigning larger weights to units other than unit 1 induces bias but may substantially reduce variability. This bias-variance trade-off is most relevant in a moderate-$T$ setting, defined as the range of values of $T$ for which the variability of the individual estimators $\hat{\theta}_i$ is of the same order of magnitude as $\eta_i$ (see remark 1 below for a heuristic criterion for detecting a moderate-$T$ setting).

In this work we introduce two weighting schemes — the fixed-$N$ and the large-$N$ minimum-MSE unit averaging estimators. The key practical difference between the two is that the large-$N$ estimator uses prior information about the similarity of cross-sectional units in terms of the focus parameter. In contrast, the fixed-$N$ estimator requires no prior information (see the discussion following eq. (6) explaining the names of the approaches) These estimators seek to strike a balance between the bias and variance of the unit averaging estimator. For both, the weights are chosen by minimizing an estimator of the local asymptotic variance of $\hat{\theta}_i$ when $N < \infty$ is justified in detail in the next section.

The fixed-$N$ approach provides an agnostic way to determine the weights. It imposes no structure on the weights. All of the weights are determined only by the data. Formally, let $\tilde{N} < \infty$ be the number of units. Let $w^\tilde{N} = (w_i^\tilde{N})$ be a $\tilde{N}$-vector such that $w_i^\tilde{N} \geq 0$ for all $i$ and $\sum_{i=1}^{\tilde{N}} w_i^\tilde{N} = 1$. The fixed-$N$ LA-MSE estimator associated with $w^\tilde{N}$ is given by

$$LA-MSE_N(w^\tilde{N}) = \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} w_i^{\tilde{N}} \hat{\Psi}_{[i,j]} w_j^{\tilde{N}},$$

(3)

where $\hat{\Psi}_{[i,j]} \in \mathbb{R}^{\tilde{N} \times \tilde{N}}$ with entries $\hat{\Psi}_{[i,i]} = \nabla \mu(\hat{\theta}_1)'(T(\hat{\theta}_1 - \hat{\theta}_1))(\hat{\theta}_1 - \hat{\theta}_1)' + \hat{V}_i \nabla \mu(\hat{\theta}_1)$ and $\hat{\Psi}_{[i,j]} = \nabla \mu(\hat{\theta}_1)'T(\hat{\theta}_1 - \hat{\theta}_1)(\hat{\theta}_j - \hat{\theta}_1)'\nabla \mu(\hat{\theta}_1)$ when $i \neq j$. Here $\hat{V}_i$ is an estimator of the asymptotic variance of $\hat{\theta}_i$, and $\nabla \mu(\cdot)$ is the gradient of $\mu$. The terms $\nabla \mu(\hat{\theta}_1)'T(\hat{\theta}_1 - \hat{\theta}_1)(\hat{\theta}_i - \hat{\theta}_1)'\nabla \mu(\hat{\theta}_1)$ and $\nabla \mu(\hat{\theta}_1)'\hat{V}_i \nabla \mu(\hat{\theta}_1)$ are estimators of, respectively, the squared bias and variance of $\mu(\hat{\theta}_1)$ as estimators of $\mu(\theta_1)$. The fixed-$N$ minimum MSE weights are defined as

$$w^\tilde{N} = \arg \min_{w \in \Delta^{\tilde{N}}} LA-MSE_N(w),$$

(4)

where $\Delta^{\tilde{N}} = \{w \in \mathbb{R}^{\tilde{N}} : \sum_{i=1}^{\tilde{N}} w_i = 1, w_i \geq 0, i = 1, \ldots, \tilde{N}\}$. 

5
Alternatively, the researcher may have prior information on which units are potentially more important for estimating $\mu(\theta_1)$ (in terms of having a similar $\mu(\theta_i)$ or being similar in observables, see below). Accordingly, units are partitioned into two sets – a set of $\bar{N} \geq 0$ unrestricted potentially important units, and a set of the remaining $N - \bar{N}$ restricted units. The number of restricted units $N - \bar{N}$ is assumed to be at least somewhat large for the partition of units to have a meaningful impact on the resulting estimator.

The large-$N$ estimator leverages prior information expressed through these two sets. Intuitively, the weights of the unrestricted units are freely determined by the data. For the restricted units, the optimization problem determines only the total mass assigned to the whole restricted set. This mass is then equally split over its members, though we note that other weighting schemes are allowed for the restricted units; see theorem 3 below. Formally, let $\mathbf{w}^{N,\infty} = (w_i^{N,\infty})$ be an $N$-vector and assume that the weights of the unrestricted units are placed in the first $\bar{N}$ positions. The vector of weights $\mathbf{w}^{N,\infty}$ is such that $w_i^{N,\infty} \geq 0$ for all $i$, $\sum_{i=1}^{N} w_i^{N,\infty} = 1$, and the weights of the restricted units ($i > \bar{N}$) are equal and given by $w_i^{N,\infty} = (1 - \sum_{j=1}^{\bar{N}} w_j^{N,\infty})/(N - \bar{N})$. Let $\mathbf{w}^{\bar{N},\infty} = (w_i^{\bar{N},\infty})$ be a $\bar{N}$-vector such that $w_i^{\bar{N},\infty} = w_i^{N,\infty}$ for $i = 1, \ldots, \bar{N}$. These are the weights of the unrestricted units. The large-$N$ LA-MSE estimator associated with $\mathbf{w}^{N,\infty}$ is controlled by $\mathbf{w}^{\bar{N},\infty}$ and given by

$$
\text{LA-MSE}_{\infty}(\mathbf{w}^{N,\infty}) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{\bar{N}} w_i^{\bar{N},\infty} [\bar{\Psi}_{N}]_{ij} w_j^{\bar{N},\infty} + \left( 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \left( \sqrt{T} \nabla \mu(\theta_1)' \left( \theta_1 - \frac{1}{N} \sum_{i=1}^{N} \theta_i \right) \right) \\
- 2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \nabla \mu(\theta_1) \sqrt{T} \left( \theta_i - \bar{\theta}_1 \right) \left( 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \left( \sqrt{T} \nabla \mu(\theta_1)' \left( \theta_1 - \frac{1}{N} \sum_{i=1}^{N} \theta_i \right) \right).
$$

The above approximation to the MSE assumes that the number $N - \bar{N}$ of restricted units is large. In this case the restricted units have an impact on the bias of the estimator, but only a negligible contribution to its variance (asymptotically as $N \to \infty$).

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

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

where

$$
\mathbf{w}^{\bar{N},\infty} = \text{arg min}_{\mathbf{w} \in \Delta^{\bar{N}}} \text{LA-MSE}_{\infty}(\mathbf{w})
$$

with $\Delta^{\bar{N}} = \{ \mathbf{w} \in \mathbb{R}^{\bar{N}} : w_i \geq 0, \sum_{i=1}^{\bar{N}} w_i = 1 \}$. Note that the optimization problem defining $\mathbf{w}^{\bar{N},\infty}$ is $\bar{N}$-dimensional and can be solved by standard quadratic programming methods.
Three comments are in order before we proceed. First, the names of the approaches come from the frameworks used to study their properties. The fixed-$N$ estimator is studied in a setting where the number of units $\bar{N}$ is held finite and fixed, regardless of whether $\bar{N}$ is small or large in practical terms. In contrast, the large-$N$ estimator is studied in a framework where the size of the restricted set grows without bound.

Second, using the large-$N$ estimator requires choosing the set of unrestricted units. In principle, this set may be chosen arbitrarily, with weights (6) adapting to the choice. However, larger reductions in bias are possible if the unrestricted set contains units with $\mu(\theta_i)$ similar to $\mu(\theta_1)$. For example, when dealing with country-level, this similarity may be established by using previous country-level studies focusing on the parameter of interest or related parameters. We explore several ways of specifying this set in sections 4-5.

Last, the fixed- and large-$N$ LA-MSE estimators have the appealing property of being applicable both when the amount of time series information in the panel is moderate or large. When the amount of time series information is moderate, the LA-MSE approximates the infeasible population problem of minimizing the MSE, along with uncertainty about individual parameters (see the discussion following theorem 2). When the amount of time series information is large, the bias term in the MSE dominates. Then the unit averaging estimator based on the minimum MSE weights converges to the individual estimator $\mu(\hat{\theta}_1)$, if the coefficients $\theta_i$ are continuously distributed (see remark 3 in the next section).

**Remark 1** (Practical criterion for a moderate-$T$ setting). In practice, the small-, moderate- and large-$T$ settings may be differentiated using the following heuristic criterion. If the realized $t$-statistic(s) of the individual-specific estimates is between 1 and 5, the setting is a moderate-$T$ one. Larger $t$-statistics signal a large-$T$ setting. If the $t$-statistics are smaller than 1 or the individual estimators cannot be computed, the setting is a small-$T$ one.

**Remark 2** (Non-MSE criteria). The quality of the estimator may also be measured using notions of risk different from the MSE. In the Online Appendix, we extend the analysis of the paper to risks of the form $R_l(\mu(\theta_1), \hat{\mu}(w_N)) = \mathbb{E}[l(\mu(\theta_1), \hat{\mu}(w_N))]$, where $l$ is some loss function. If $l$ is a strictly convex smooth function, we show that $R_l$ behaves essentially like the MSE. Weights (4) and (6) are feasible minimum risk weights for $R_l$. In contrast, if $l$ is the absolute loss, the local approximation to $R_l$ (the mean absolute deviation in this case) is different from LA-MSE. However, optimal weights may be obtained similarly.
3 Theory

3.1 Assumptions

We focus on a moderate-\(T\) setting — in which the variance of the individual estimators is of the same order of magnitude as the individual components \(\eta_i\). In this case, the amount of information in each individual time series is limited. To emulate this and the trade-off between unit-specific and panel-wide information, we make a local heterogeneity assumption.

A.1 (Local Heterogeneity). The sequence of unit-specific parameters \(\{\theta_i\}\) is such that

\[
\theta_i = \theta_0 + \frac{\eta_i}{\sqrt{T}}
\]

where \(\{\eta_i\}\) is a sequence of independent random vectors that satisfy \(E_{\eta} [\eta_i] = 0\) and \(\sup_i E_{\eta} [||\eta_i||^2] < \infty\) (here and below \(||-||\) means the 2-norm; \(E_{\eta}\) means that the expectation according to the joint distribution of \(\{\eta_i\}\)). All analysis is done conditional on \(\sigma(\eta_1, \eta_2, \ldots)\) and all statements below are conditional on \(\sigma(\eta_1, \eta_2, \ldots)\) unless specifically stated otherwise.

Scaling \(\eta_i\) by \(\sqrt{T}\) is a mathematical device that allows us to approximate a limited-information moderate-\(T\) setting using asymptotic techniques with \(T \to \infty\). Intuitively, as \(T\) becomes larger, the signal strength becomes proportionally weaker, so that the amount of information in each time series is unchanged and bounded even if \(T \to \infty\). At the same time, this assumption will permit us to apply asymptotic techniques to characterize the leading terms of the bias and the variance of the unit averaging estimator. The local heterogeneity assumption is analogous to the local misspecification device used in the frequentist model averaging literature (Hjort and Claeskens, 2003a,b; Hansen, 2016; Yin et al., 2021). It is also similar to the techniques of weak instrument asymptotics (Staiger and Stock, 1997) and local alternatives used in test evaluation (Lehmann and Romano, 2022). Like in those settings, this assumption should not be interpreted literally as meaning that the true parameters change depending on time series length (see Raftery and Zheng (2003) and Hjort and Claeskens (2003b) for some important criticism of such an interpretation of locality).

Since the focus lies on recovering the realized individual parameter \(\mu(\theta_1)\), all probability statements are implicitly conditional on \(\sigma(\eta_1, \eta_2, \ldots)\). Such conditioning is typical when individual parameters are of interest (Vaida and Blanchard, 2005; Donohue, Overholser, Xu, and Vaida, 2011; Zhang et al., 2014). Importantly, all the results are shown to hold with \(\eta\)-probability 1 (for almost any realization of \(\{\eta_i\}\)).

In this paper we assume that the cross-sectional units are independent.

A.2 (Independence). For each \(i, j_1, \ldots, j_k, k\) such that \(i \neq j_1, \ldots, j_k\) \(\{\{z_{it}\}_{t=0}^\infty, \eta_i\}\) and \(\{\{z_{j_1t}\}_{t=0}^\infty, \eta_{j_1}\}, \ldots, \{\{z_{j_kt}\}_{t=0}^\infty, \eta_{j_k}\}\) are independent.
Note that together A.1 and A.2 permit cross-sectional heterogeneity. In particular, \( \eta_i \) may be heterogeneously distributed, provided the coefficients \( \theta_i \) share a common mean \( \theta_0 \).

The unit-specific estimators \( \hat{\theta}_i \) are assumed to satisfy a number of regularity conditions.

A.3 (Individual Objective Function).

(i) The parameter space \( \Theta \) is convex.

(ii) The function \( m(\theta, z) : \Theta \times Z \rightarrow \mathbb{R} \) is twice continuously differentiable in \( \theta \) for each value of \( z \). \( m(\theta, z) \) is measurable as a function of \( z \) for every value of \( \theta \).

(iii) There exists a positive finite constant \( T_0 \) (which does not depend on \( i \)) such that for all \( i \) and \( T > T_0 \) it holds that the unit-specific estimator satisfies \( \hat{\theta}_i \in \text{int}(\Theta) \) a.s..

(iv) The gradient of the unit-specific objective function satisfies

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) \Rightarrow N(0, \Sigma_i),
\]

where \( \Sigma_i = \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} \mathbb{E} \left[ \left( \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) \right) \left( \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) \right)' \right] \).

(v) There exist a positive finite constant \( C_{\nabla m} \) (which does not depend on \( i \) or \( T \)) such that, for all \( i \) and all \( T > T_0 \) and for some \( \delta > 0 \), it holds that

\[
\mathbb{E} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) \right\|^{2(1+\delta)} \leq C_{\nabla m}.
\]

(vi) The Hessian of the unit-specific objective function satisfies

\[
\sup_{\theta \in [\theta_i, \hat{\theta}_i]} \left\| \frac{1}{T} \sum_{t=1}^{T} \nabla^2 m(\theta, z_{it}) - H_i \right\| \overset{p}{\to} 0,
\]

where \( H_i = \lim_{T \to \infty} \mathbb{E}(T^{-1} \sum_{t=1}^{T} \nabla^2 m(\theta_i, z_{it})). \)

(vii) Let \( D_{iT} = \sup_{\theta \in [\theta_i, \hat{\theta}_i]} \left\| \left( T^{-1} \sum_{t=1}^{T} \nabla^2 m(\theta, z_{it}) \right) H_i^{-1} - I \right\|_\infty \). \( D_{iT} < 1 \) a.s. for all \( i \) and all \( T > T_0 \). There exists a positive constant \( C_{\nabla^2 m} \) such that, for all \( i \) and all \( T > T_0 \) and for \( \delta \) as in (v), it holds that

\[
\mathbb{E} \left[ \left( \frac{D_{iT}}{1 - D_{iT}} \right)^{\frac{2(1+\delta)(1+\delta)}{\delta}} \right] \leq C_{\nabla^2 m}.
\]

(viii) The matrices \( \Sigma_i \) and \( H_i \) satisfy \( \Delta_{\Sigma} \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq \Lambda_{\Sigma} \) and \( \Delta_{H} \leq \lambda_{\min}(H_i) \leq \lambda_{\max}(H_i) \leq \Lambda_{H} \) where \( \Delta_{\Sigma}, \Lambda_{\Sigma}, \Delta_{H} \) and \( \Lambda_{H} \) are positive constants that do not depend on \( i \).

(ix) Let \( V_i = H_i^{-1} \Sigma_i H_i^{-1} \). Then, there is a sequence of estimators \( \{\hat{V}_i\} \) such that, for all
\( i, \hat{V}_i \) is consistent for \( V_i \), and, for all \( T > T_0 \), \( \lambda_{\min}(\hat{V}_i) > 0 \) holds almost surely.

A.3 requires the unit-specific estimators to be consistent, asymptotically normal and to satisfy a number of regularity conditions. This assumption allows for a fair amount of dependence, heterogeneity, and non-stationarity in the unit-specific time series; we refer to ch. 11 of Pötscher and Prucha (1997) for a catalog of low-level conditions. Assumption A.3\((iii)\) states that the unit-specific estimator lies in the interior of the parameter space almost surely. If the problem is linear or defined by a convex smooth objective function and continuous covariates, the parameter space can be taken to be \( \mathbb{R}^p \), and the condition holds automatically. Assumption A.3\((iv)\) is standard in the M-estimation literature, it requires the gradient of the objective function evaluated at \( \theta_i \) to satisfy a CLT. Assumption A.3\((v)\) is a moment condition on the gradient of the objective function. In an i.i.d. setting such an assumption translates into a moment condition on the individual gradients. More generally, this would be implied by appropriate moment and dependence assumption on the individual gradients. Assumption A.3\((vi)\) is also standard in the M-estimation literature; it requires the Hessian to satisfy a uniform law of large numbers. Assumption A.3\((vii)\) effectively requires that the sample Hessian is nonsingular in a small enough neighborhood of \( \theta_i \). In a scalar problem, \((vii)\) restricts the possible range of the second derivative as \( \theta \) ranges over a shrinking interval around \( \theta_i \). In addition, \((vii)\) places an assumption on the moments of deviation from the population limit Hessian. In case of linear regression, the sample and population Hessians do not depend on the slope parameters and \((vii)\) is an assumption on moments of covariates. Assumption A.3\((viii)\) implies a uniform restriction on the asymptotic variance \( V_i \) of the individual estimators. Assumption A.3\((ix)\) states that there exists a sequence of nonsingular estimators \( \{\hat{V}_i\} \) for the asymptotic variance-covariance matrix of the individual estimator. We remark that Assumptions A.3\((iii)\) and \((vii)\) state that the sequence of unit-specific estimation problems satisfies appropriate uniformity conditions. Such conditions allow us to distill the key arguments relevant to our averaging theory and, in a sense, should be interpreted as a simplifying approximation. In general, \((iii)\) and \((vii)\) would hold with probability approaching one for each unit. In this case all our results would still hold, though under appropriate rate conditions on \((N, T)\) and trimming to ensure certain well-behavedness of individual estimators. We further note that assumptions \((iii)\) and \((vii)\) might hold in practice in certain special cases regardless (such as linear or nonlinear models with a convex and smooth objective function and continuous covariates).

A.4 (Unit-specific Bias). There exists a constant \( C_{Bias} \), which does not depend on \( i \), such that
\[
\left\| \mathbb{E}[\hat{\theta}_i - \theta_i] \right\|_1 \leq C_{Bias}/T \enspace \text{for all} \enspace T > T_0.
\]

Assumption A.4 requires that the bias of individual estimators for their own parameters is bounded uniformly in \( i \). The order of the bias is consistent with the results obtained
by Rilstone, Srivastava, and Ullah (1996) and Bao and Ullah (2007). The higher order terms can be subsumed into the $T^{-1}$ term for a sufficiently large $C_{\text{Bias}}$. Assumption A.4 is satisfied for linear models under assumption A.3. For nonlinear models it is sufficient that for all $s$ and $i$ it holds that $\mathbb{E}(\|\nabla^s m(\theta, z_i t)\|^2) \leq C_s < \infty$ (Bao and Ullah, 2007).

A.5 (Focus Parameter). The focus function $\mu : \Theta \rightarrow \mathbb{R}$ is twice-differentiable. There exists a constant $C_{\nabla \mu}$ such that $\|\nabla \mu(\theta)\| < C_{\nabla \mu}$ for all $\theta \in \Theta$. There exists a constant $C_{\nabla^2 \mu}$ such that for all $\theta \in \Theta$ the largest and smallest eigenvalues of the Hessian $\nabla^2 \mu(\theta)$ are bounded in absolute value by $C_{\nabla^2 \mu}$. Let $d_0 = \nabla \mu(\theta_0)$ be the gradient of $\mu$ at $\theta_0$. Then $d_0 \neq 0$.

Assumption A.5 lays out mild smoothness assumptions on $\mu$. For simplicity we assume that $\mu$ is a scalar focus parameter. However, all our results can be extended to the case in which $\mu$ is a vector focus parameter.

3.2 Properties of the Minimum MSE Unit Averaging Estimator

We begin with a lemma that establishes the properties of the unit-specific estimators $\hat{\mu}(\hat{\theta}_i)$ as estimators for the target parameter $\mu(\theta_1)$ of unit 1 in limited-information local setting.

Lemma 1. Assume that assumptions A.1–A.5 are satisfied. Let the unit-specific estimators $\hat{\theta}_i$ for $i = 1, 2, \ldots$ be defined as in eq. (2). Then

$$\sqrt{T} \left( \hat{\theta}_i - \theta_1 \right) \Rightarrow N(\eta_i - \eta_1, V_i) =: Z_i,$$

$$\sqrt{T} \left( \mu(\hat{\theta}_i) - \mu(\theta_1) \right) \Rightarrow N(d_0'(\eta_i - \eta_1), d_0'V_i d_0) =: \Lambda_i$$

holds as $T \rightarrow \infty$ for $i = 1, 2, \ldots$. Convergence is joint (that is, with respect to the product topology), and all $Z_i$ and $\Lambda_i$ are independent across $i$.

Lemma 1 approximates the exact moderate-$T$ bias and variance of $\mu(\hat{\theta}_i)$ with their leading terms, which appear as the mean and variance of $\Lambda_i$. This approximation relies on the locality assumption A.1: as $T \rightarrow \infty$, the amount of information in each individual time series remains limited (see the discussion after A.1). Consequently, both the asymptotic mean and variance are non-negligible and of the same order.

We now establish a local asymptotic approximation to the MSE (LA-MSE) of the unit averaging estimator (1). Let $\{w_1, w_2, \ldots\}$ be a (non-random) sequence where $w_k$ is a $k$-vector of weights. Suppose that $w_N$ converges to some $w \in \mathbb{R}^\infty$ in the sense defined below. In what follows we treat $w_k = (w_{ik})$ as an element both in $\mathbb{R}^k$ and in $\mathbb{R}^\infty$ (with coordinates $i > k$ restricted to zero). Consider the unit averaging estimator $\hat{\mu}(w_N)$ (1).
Theorem 1. Let assumptions A.1–A.5 be satisfied. Let \( \{\mathbf{w}_1, \mathbf{w}_2, \ldots \} \) be such that (i) for each \( N, \mathbf{w}_N \) is measurable with respect to \( \sigma(\mathbf{\eta}_1, \ldots, \mathbf{\eta}_N) \), (ii) for each \( N, w_{iN} \geq 0 \) for all \( i \), \( \sum_{i=1}^{N} w_{iN} = 1, w_{jN} = 0 \) for \( j > N \), (iii) \( \sup_i |w_{iN} - w_i| = o(N^{-1/2}) \) where \( \mathbf{w} = (w_i) \in \mathbb{R}^\infty \) is a vector such that \( w_i \geq 0 \) and \( \sum_{i=1}^{\infty} w_i \leq 1 \). Let \( T_0 \) be as in assumption A.3.

Then (i) \( \sum_{i=1}^{\infty} w_i \mathbf{d}_i^0 \mathbf{\eta}_i \) and \( \sum_{i=1}^{\infty} w_i^2 \mathbf{d}_i^0 \mathbf{V}_i \mathbf{d}_0 \) exist; (ii) for any \( N \) and \( T > T_0 \) the MSE of the averaging estimator is finite; and (iii) as \( N, T \to \infty \) jointly it holds that

\[
T \times MSE(\hat{\mathbf{\mu}}(\mathbf{w}_N)) \to \left( \sum_{i=1}^{\infty} w_i \mathbf{d}_i^0 \mathbf{\eta}_i - \mathbf{d}_0^0 \mathbf{\eta}_1 \right)^2 + \sum_{i=1}^{\infty} w_i^2 \mathbf{d}_i^0 \mathbf{V}_i \mathbf{d}_0. =: LA-MSE(\mathbf{w}). \tag{7}
\]

Theorem 1 provides a local approximation to the MSE (LA-MSE) of the averaging estimator. The LA-MSE consists of the leading terms of the moderate-\( T \) bias and variance of the estimator. This result parallels local approximations for the finite-sample risk in the model averaging literature (e.g. Hjort and Claeskens (2003a); Hansen (2016)).

The LA-MSE highlights the bias-variance trade-off associated with the choice of the weights. The two extremes of the trade-off correspond to the individual estimator \( \hat{\mathbf{\mu}}(\mathbf{\hat{\theta}}_1) \) of the first unit and the mean group estimator \( \hat{\mathbf{\mu}}_{MG} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{\mu}}(\mathbf{\hat{\theta}}_i) \). \( \mathbf{\mu}(\mathbf{\hat{\theta}}_1) \) is obtained by setting \( w_{1N} = 1 \) for all \( N \). It is asymptotically unbiased, and its LA-MSE is equal to \( \mathbf{d}_0^0 \mathbf{V}_i \mathbf{d}_0 \), the asymptotic variance of the individual estimator. The mean group estimator is obtained by setting \( w_{iN} = (N)^{-1} \mathbb{1}_{i \leq N} \) for \( i = 1, \ldots, N \) for all \( N \). The variance term for \( \hat{\mathbf{\mu}}_{MG} \) is zero, and the LA-MSE is equal to \( (\mathbf{d}_0^0 \mathbf{\eta}_1)^2 \).

The weight convergence condition (iii) characterizes the spaces of weights over which the MSE is validly approximated by the LA-MSE. (iii) requires the sequence \( \{\mathbf{w}_1, \mathbf{w}_2, \ldots \} \) of weight vectors to converge uniformly to some limit \( \mathbf{w} \) as the cross-section grows. Note that the sum of the limit \( \mathbf{w} \) can be less than one, as is the case for the mean group estimator.

We now specialize the LA-MSE expression to the fixed-\( N \) and large-\( N \) averaging approaches of section 2. In the fixed-\( N \) case, suppose that only the first \( \bar{N} \) units are being averaged, where \( \bar{N} \) is fixed and finite. Only these units affect the bias and the variance of the estimator, and both sums in eq. (7) are finite sums. The LA-MSE is a quadratic function of the weights. Formally, for all \( N \geq \bar{N} \), let \( \mathbf{w}_N = (w_{iN}) \) satisfy two conditions. First, set \( w_{iN} = 0 \) for all \( i > \bar{N} \). Second, let \( \mathbf{w}^\bar{N} \) be a \( \bar{N} \)-vector that satisfies \( \sum_{i=1}^{\bar{N}} w_{iN}^\bar{N} = 1, w_{iN}^\bar{N} \geq 0 \). Then let \( w_{iN} = w_i^\bar{N} \). The condition that \( N \to \infty \) becomes superfluous and condition (iii) holds automatically. The LA-MSE is controlled by the \( \bar{N} \)-vector \( \mathbf{w}^\bar{N} \) and can be written as

\[
LA-MSE_{\bar{N}}(\mathbf{w}^\bar{N}) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{\bar{N}} w_{iN}^\bar{N} [\mathbf{\Psi}^\bar{N}]_{ij} w_{jN}^\bar{N} \equiv \mathbf{w}^\bar{N} \mathbf{\Psi}^\bar{N} \mathbf{w}^\bar{N},
\]

where \( \mathbf{\Psi}^\bar{N} \) is an \( \bar{N} \times \bar{N} \) matrix with elements \([\mathbf{\Psi}^\bar{N}]_{ii} = \mathbf{d}_0^0 ((\mathbf{\eta}_i - \mathbf{\eta}_1) (\mathbf{\eta}_i - \mathbf{\eta}_1)' + \mathbf{V}_i) \mathbf{d}_0 \) and
\([\Psi_N]_{ij} = d_0'(\eta_i - \eta_1)(\eta_j - \eta_1)'d_0\) when \(i \neq j\).

In the large-\(N\) case, let the \(N\) unrestricted units be placed in the first \(\bar{N}\) positions, with the \(N - \bar{N}\) remaining units forming the restricted set. By eq. (6), the individual weights of the restricted units converge to 0 uniformly and satisfy \((iii)\). The restricted units contribute only to the bias component of the LA-MSE. The LA-MSE itself is fully determined by the individual weights of the unrestricted units and the total mass assigned to the restricted set. Formally, let \(w_{\bar{N},\infty}\) be a \(\bar{N}\)-vector that satisfies \(\sum_{i=1}^{\bar{N}} w_{i,\infty}^{\bar{N},\infty} = 1, w_{i,\infty}^{\bar{N},\infty} \geq 0\); the vector \(w_{\bar{N},\infty}\) holds the weights of the unrestricted units. Set \(w_{i,N} = w_{i,\infty}^{\bar{N},\infty}\) for \(i \leq \bar{N}\) and \(w_{i,N} = (1 - \sum_{j=1}^{\bar{N}} w_{j,\infty}^{\bar{N},\infty})/(N - \bar{N}), i \in \{\bar{N} + 1, \ldots, N\}\). Let \(w = (w_i)\) where \(w_i = w_{i,\infty}^{\bar{N},\infty}, i \leq \bar{N}\) and \(w_i = 0, i > \bar{N}\). Then \(\sup_i |w_{i,N} - w_i| = O(N^{-1})\). Note that the mass of the restricted units \((1 - \sum_{i=1}^{\bar{N}} w_{i,\infty}^{\bar{N},\infty})\) may lie anywhere between 0 and 1 (the latter being the case for the mean group estimator). The LA-MSE is controlled by \(w_{\bar{N},\infty}\) as

\[
LA-MSE_{\infty}(w_{\bar{N},\infty}) = \sum_{i=1}^\bar{N} \sum_{j=1}^\bar{N} w_{i,\infty}^{\bar{N},\infty}[\Psi_N]_{ij} w_{j,\infty}^{\bar{N},\infty}
+ \left(1 - \sum_{i=1}^\bar{N} w_{i,\infty}^{\bar{N},\infty}\right) d_0'(\eta_1 - 2 \sum_{i=1}^\bar{N} w_{i,\infty}^{\bar{N},\infty} d_0'(\eta_i - \eta_1)) \left(1 - \sum_{i=1}^\bar{N} w_{i,\infty}^{\bar{N},\infty}\right) d_0'(\eta_1).
\]

The same expression for the LA-MSE can be obtained with other weighting schemes for the restricted set. The weights in \(w_N\) beyond \(\bar{N}\) can display strong variations in orders of magnitude, with some weights decaying like \(N^{-1/2-\varepsilon}\), and some at a faster rate.

The above arguments also show that it is internally consistent to use the fixed-\(N\) and large-\(N\) approaches to minimize the MSE. These approaches minimize (an estimator) of the LA-MSE. The weights returned lie within the class of weights for which the LA-MSE provides a valid approximation to the MSE.

The quantities \(\hat{LA-MSE}_N\) and \(\hat{LA-MSE}_{\infty}\) used to define the minimum MSE weights introduced in section 2 are estimators of the population expressions for the LA-MSE given above. In the rest of the section we focus on the properties of these estimators as well as the optimal weights (4) and (6) associated with them.

We begin by noting that in our framework the population LA-MSE cannot be consistently estimated. Under local heterogeneity the idiosyncratic components \(\eta_i\) cannot be consistently estimated, as the amount of information in each time series is finite and bounded under A.1 (Hjort and Claeskens, 2003a). Instead, we form \(\hat{LA-MSE}_N\) and \(\hat{LA-MSE}_{\infty}\) by plugging in asymptotically unbiased estimators for \(\eta_i - \eta_1\) and \(\eta_i\) (Hjort and Claeskens, 2003a). Such estimators are provided by \(\sqrt{T}(\hat{\theta}_i - \theta_1)\) and \(\sqrt{T}(\hat{\theta}_1 - N^{-1} \sum_{i=1}^N \hat{\theta}_i)\), respectively:
Lemma 2. Let assumptions A.1-A.5 hold. Then as \( N, T \to \infty \) jointly, it holds that

\[
\sqrt{T} \left( \hat{\theta}_i - \theta_1 \right) \Rightarrow N(\eta_i - \eta_1, V_i + V_1) = Z_i - Z_1, \\
\sqrt{T} \left( \theta_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i \right) \Rightarrow N(\eta_1, V_1) = Z_1 + \eta_1.
\]

Convergence is joint for all \( i \).

The following two theorems establish the properties of our LA-MSE estimators and the associated minimum MSE weights (4) and (6). The theorem also characterizes the asymptotic distribution of the minimum MSE unit averaging estimators. First, we state a result for the fixed-\( N \) estimator. Recall that \( \Delta^N = \{ \mathbf{w} \in \mathbb{R}^N : \sum_{i=1}^{N} w_i = 1, w_i \geq 0, i = 1, \ldots, N \} \).

Theorem 2 (Fixed-N Minimum MSE Unit Averaging). Let assumptions A.1-A.5 hold and \( \bar{N} < \infty \) be a fixed positive integer.

(i) For any \( \mathbf{w}^N \in \Delta^N \) it holds that \( \text{LA-MSE}_N(\mathbf{w}^N) \Rightarrow \text{LA-MSE}_{\bar{N}}(\mathbf{w}^N) := \mathbf{w}^N \mathbf{\Psi}_N \mathbf{w}^N \) as \( T \to \infty \), where \( \mathbf{\Psi}_N \) is an \( \bar{N} \times \bar{N} \) matrix with 

\[
[\mathbf{\Psi}_N]_{ij} = d_0^i((Z_i - Z_1)(Z_i - Z_1)' + V_1) d_0 \quad \text{when } i = j \text{ and } d_0^i((Z_i - Z_1)(Z_j - Z_1)') d_0 \quad \text{when } i \neq j; \text{ and } Z_i \text{ is as in lemma 1}.
\]

(ii) As \( T \to \infty \), the minimum MSE weights satisfy

\[
\mathbf{w}^N = \arg \min_{\mathbf{w}^N \in \Delta^N} \text{LA-MSE}_N(\mathbf{w}^N) \Rightarrow \mathbf{\bar{w}}^N = \arg \min_{\mathbf{w}^N \in \Delta^N} \text{LA-MSE}_{\bar{N}}(\mathbf{w}^N).
\]

(iii) As \( T \to \infty \), for \( \Lambda_i \) of lemma 1, the minimum MSE unit averaging estimator satisfies

\[
\sqrt{T} \left( \hat{\mu}(\mathbf{w}^N) - \mu(\theta_1) \right) \Rightarrow \sum_{i=1}^{\bar{N}} \bar{w}_i^N \Lambda_i.
\]

The quantity \( \text{LA-MSE}_{\bar{N}} \) plays the same role to \( \text{LA-MSE}_N \) as \( Z_i \) does to \( \sqrt{T}(\theta_i - \theta_1) \) in lemma 1. \( \text{LA-MSE}_{\bar{N}} \) uses a local approximation to express \( \text{LA-MSE}_N \) in terms of the leading components of the MSE and the approximate distribution of the individual estimators. We can see that \( \text{LA-MSE}_{\bar{N}} \) is composed of the population LA-MSE, a bias term, and a noise component. In fact, the entries of the matrix \( \mathbf{\Psi}_N \) may be expressed as

\[
[\mathbf{\Psi}_N]_{ii} = [\Psi_N]_{ii} + d_0^i(V_i + V_1) d_0 + d_0^i e_i d_0,
\]

\[
[\mathbf{\Psi}_N]_{ij} = [\Psi_N]_{ij} + d_0^j V_1 d_0 + d_0^j e_{ij} d_0, \quad i \neq j,
\]

where \( e_{ij} = (Z_i - Z_1)(Z_j - Z_1)' - \mathbb{E}((Z_i - Z_1)(Z_j - Z_1)') \). The noise terms \( e_{ij} \) may be interpreted as the result of the fact that in a moderate-\( T \) setting there is limited information about the idiosyncratic components \( \eta_i \). These terms are mean zero and independent
conditional on unit 1. The bias terms guarantee that $\overline{\Psi}_N$ is positive definite and arise as a consequence of using the biased positive definite estimator $\hat{\Psi}_N$ (see remark 4 below). The bias can be split into two components. The $d_0' V_1 d_0$ is common for all elements of $\overline{\Psi}_N$ and does not affect the solution of the MSE minimization problem. The second component $d_i' V_i d_0$ only affects the diagonal of $\overline{\Psi}_N$ and measures the individual variances. This component does not modify the ordering of the estimators in terms of their variances.

Result (iii) shows that the minimum MSE unit averaging estimator has a nonstandard asymptotic distribution in the local heterogeneity framework. The limit distribution is a randomly weighted sum of independent normal random variables. This result is somewhat similar to the distributional results for model averaging estimators (Liu, 2015). In the Online Appendix, we show how to construct confidence intervals based on theorem 2.

The following theorem establishes an analogous result for the large-$N$ estimator.

**Theorem 3 (Large-$N$ Minimum MSE Unit Averaging).** Let assumptions A.1-A.5 hold and $\bar{N} < \infty$ be a fixed non-negative integer.

(i) For any $w^{N,\infty} \in \bar{\Delta}^N$ it holds that $\overline{LA-MSE}_\infty(w^{N,\infty}) \Rightarrow \overline{LA-MSE}_\infty(w^{N,\infty})$ as $N, T \rightarrow \infty$ jointly where $\bar{\Delta}^N = \{ w \in \mathbb{R}^\bar{N} : w_i \geq 0, \sum_{i=1}^N w_i \leq 1 \}$ and

$$\overline{LA-MSE}_\infty(w^{N,\infty}) = \begin{pmatrix} w^{N,\infty} \Psi_N \bar{w}^{N,\infty} + \left( 1 - \sum_{i=1}^N \bar{w}_i^{N,\infty} \right) d_0'(\eta_1 + Z_1) \\
2 \sum_{i=1}^N \bar{w}_i^{N,\infty} d_0'(Z_i - Z_1) \end{pmatrix} \left( 1 - \sum_{i=1}^N \bar{w}_i^{N,\infty} \right) d_0'(\eta_1 + Z_1).$$

(ii) As $N, T \rightarrow \infty$, the minimum MSE weights satisfy

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

(iii) Let $v_{N-\bar{N}} = (v_{\bar{N}N}, \ldots, v_{NN})$ be a $(N - \bar{N})$-vector such that $\sup_i v_{iN-\bar{N}} = o(N^{-1/2})$, $v_{iN-\bar{N}} \geq 0$, for each $N$ it holds that $\sum_{i=1}^{N-\bar{N}} v_{iN-\bar{N}} = 1$. Then as $N, T \rightarrow \infty$ jointly

$$\sqrt{T} \left( \sum_{i=1}^{\bar{N}} \hat{w}_i^{N,\infty} \mu(\hat{\theta}_1) + \left( 1 - \sum_{i=1}^{\bar{N}} \hat{w}_i^{N,\infty} \right) \sum_{j=\bar{N}N}^{N} v_{jN-\bar{N}} \mu(\hat{\theta}_j) - \mu(\theta_1) \right)$$

$$\Rightarrow \sum_{i=1}^{N} \overline{w}_i^{N,\infty} \Lambda_i - \left( 1 - \sum_{i=1}^{N} \overline{w}_i^{N,\infty} \right) d_0' \eta_1.$$ (8)

Note that the estimator in equation (8) is a valid averaging estimator, with weights summing to unity. The exact way $v_N$ is picked does not matter, as long as the decay condition holds. All admissible choices lead to the same limit. In particular, we may pick equal weights.
\[ v_{iN} = 1/(N - \bar{N}), \] as we do in eq. (6). Also note that the convergence result (ii) applies to the vector \( \hat{w}^{N,\infty} \) of the weights of the unrestricted units, a vector of fixed length \( \bar{N} \).

**Remark 3** (Large-\( T \) properties). Minimizing \( \text{LA-MSE}_N \) is natural even in a non-local (fixed parameters) setting where we drop assumption A.1 and allow the amount of information in each time series to grow as \( T \to \infty \). Asymptotically, this approach will place zero weights on units with \( \theta_i \neq \theta_1 \), while the weights on units with \( \theta_i = \theta_1 \) will follow theorem 2. Specifically, for all \( i \) such that \( \theta_i \neq \theta_1 \), the bias estimators \( \sqrt{T}(\theta_i - \hat{\theta}_1) \) will diverge. In contrast, for the units with \( \theta_i = \theta_1 \), the bias estimators \( \sqrt{T}(\theta_i - \hat{\theta}_1) \) will instead behave as in lemma 2 (with \( \eta_i - \eta_1 = 0 \)). Accordingly, asymptotically no weight will be assigned to units with \( \theta_i \neq \theta_1 \). Similarly, \( \sqrt{T}(\theta_i - N^{-1} \sum_{i=1}^{N} \theta_i) \) will diverge, leading the approach to place no weight on the restricted set, if it is present. Such a result has a parallel in fixed parameter asymptotics for model averaging (Zhang and Liu, 2019, 2024). The units with \( \theta_i \neq \theta_1 \) play the role of under-fitted models (asymptotically zero weights), while the units \( \theta_i = \theta_1 \) correspond to the just-fitted and over-fitted models (random weights characterized by a normal vector). Moreover, the difference between the averaging estimator with minimum MSE weights and the individual estimator will converge to zero in probability if there are no other units \( i \) with \( \theta_i = \theta_1 \) (as would happen if the distribution of \( \eta \) is continuous).

**Remark 4** (Bias in \( \hat{\Psi}_N \) and an alternative estimator for \( \Psi_N \)). The matrix \( \hat{\Psi}_N \) of equations (3) and (5) is a biased estimator of \( \Psi_N \). Such a bias ensures that \( \text{LA-MSE} \) is nonnegative for all admissible weight vectors. An asymptotically unbiased estimator \( \tilde{\Psi}_N \) instead would have elements \( [\tilde{\Psi}_N]_{ij} = d_i'(T(\hat{\theta}_i - \hat{\theta}_1)(\hat{\theta}_j - \hat{\theta}_1)' - (\hat{V}_i \mathbb{1}\{i = j\} + \hat{V}_i))d_j \). However, \( \hat{\Psi}_N \) can fail to be positive definite, as it involves a difference of positive definite matrices, leading to the undesirable possibility of negative estimates of the LA-MSE.

### 4 Simulation Study

In this section, we study the performance of our minimum MSE unit averaging estimator for a variety of sample sizes via a simulation exercise. We consider a model similar to the one we use in our empirical application – a linear dynamic heterogeneous panel model defined as

\[
y_{it} = \lambda_i y_{i t-1} + \beta_i x_{it} + u_{it}, \quad u_{it} \overset{i.i.d.}{\sim} N(0, \sigma_i^2), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{9}
\]

The error \( u_{it} \) is cross-sectionally heteroskedastic, with variance \( \sigma_i^2 \) drawn independently from an exponential(1) distribution. \( u_{it} \) is independent from the coefficients and the covariates. The exogenous variable \( x_{it} \) is independently drawn from a \( N(0, 1) \) distribution. The initial conditions \( y_{i0} \) are drawn from a \( N(0, (\beta_i^2 + \sigma_i^2)/(1 - \lambda_i^2)) \) distribution to ensure that \( \{y_{it}\}_t \) is
covariance stationary. The two components of the parameter \( \theta_i = (\beta_i, \lambda_i)' \) are independently drawn from a \( N(0, 1) \) and a Beta(5,5) distribution on \([0.2, 0.8]\), respectively. Note that, in order to measure the impact of increasing information and to compare results across \( T \), we model the distribution of \( \theta_i \) as independent from \( T \). Under this (fixed parameter) approach, the amount of information in each time series increases as \( T \) grows.

We study both moderate-\( T \) and large-\( T \) settings for a variety of cross-sectional sample sizes \( N \). Specifically, we consider \( N = 50, 150, 450 \), and \( T = 50, 60, 600 \). \( T = 30 \) and \( T = 60 \) are moderate values of \( T \), according to the heuristic criterion of remark 1: the average \( t \)-statistic of the parameter estimates is 2 for \( T = 30 \) and 3.5 for \( T = 60 \). In contrast, \( T = 600 \) is a large value of \( T \), with an average \( t \)-statistic value of 10. We also note that \( N = 150, T = 60 \) is one of the estimation sample sizes in our empirical application.

The measures of interest are the MSE, bias, and variance of the unit averaging estimators (see below) for the focus parameter \( \mu(\theta_1) = \lambda_1 \). Specifically, we evaluate the MSE of the form \( \mathbb{E}[(\hat{\mu}(w) - \mu(\theta_1))^2|\lambda_1 = c] \), where \( c \) ranges through a grid of values in \([0.2, 0.8]\), and the expectation is over the distribution of data, \( \beta_1 \), and the parameters of units 2-N. The bias and variance of interest are defined similarly. We draw 10000 datasets for each value of \( c \) and \((N, T)\). For each sample, we estimate eq. (9) by OLS, compute the estimators, and record the estimates and estimation errors. The MSE is approximated with the average square Monte Carlo estimation error; we compute biases and variances similarly.

We estimate the focus parameter using the fixed-\( N \) and large-\( N \) minimum MSE estimators. We consider three specifications for the large-\( N \) estimator.

1. For the most similar specification, an oracle selects the 10% units whose parameter vector \( \theta_i \) is most similar to \( \theta_1 \) in terms of the 2-norm. These units are set as the unrestricted units. This approach measures the impact of prior information on unit similarity.

2. For the Stein-like specification, only the target unit is unrestricted.

3. For the top units approach, we first run the fixed-\( N \) estimator. The units are then sorted by the estimated weights. The top 10% units are set to be the unrestricted units. Note that the latter specification is data-driven and thus not directly covered by theorem 3. The corresponding tuning parameter (number of top units) matches the empirical application; in the Online Appendix we explore the impact of this choice.

The performance of the minimum MSE estimator is benchmarked against the individual estimator of unit 1, the mean group estimator, as well as the unit averaging estimator based on AIC/BIC weights (Buckland et al., 1997; Vaida and Blanchard, 2005). AIC and BIC generate the same likelihood-based weights, as each unit has the same number of coefficients.

Our key result is that the minimum MSE estimators generally have lower MSE for both moderate and large-\( T \). As fig. 1 shows, all of the minimum MSE estimators (bar the Stein-like one) perform favorably throughout most of the parameter space for all \((N, T)\).
Figure 1: MSE of unit averaging estimators relative to the individual estimator

Gains in the MSE are possible without prior information, as shown by the agnostic fixed-$N$ estimator, and the data-driven top unit large-$N$ specification. However, leveraging prior
information may lead to stronger improvements for some parameter values (the “most similar” line).

Fig. 1 shows a trade-off between stronger improvements for more typical values of $\lambda_1$ vs. for less typical ones (closer to $E[\lambda_1] = 0.5$ vs. closer to the boundary of the support of $\lambda_1$). This trade-off is controlled by the flexibility of the estimator, determined by the number of free weights it has. Importantly, this trade-off is not identical to the bias variance trade-off (fig. 2). More flexible estimators (such as the fixed-$N$ estimator) have uniformly lower bias for all values of $\lambda_1$. However, more flexible estimators also have lower variance for more extreme values of $\lambda_1$, while less flexible estimators have lower variance for $\lambda_1$ close to $E[\lambda_1]$.

Increasing $N$ has a twofold effect. First, it strictly improves the performance of the similarity-based large-$N$ estimator. For larger $N$, more units will lie within any given neighborhood of $\lambda_1$ on average, reducing bias. Second, more flexible estimators offer a stronger gain for less typical $\lambda_1$, as larger cross-sections will have more units with similar $\lambda_i$. At the same time, the region around $E[\lambda_i]$ in which improvements are modest grows.

![Figure 3: Average weight of target unit (unit 1). Select values of $(N, T)$](image)

Gains in MSE are strongest for smaller values of $T$. The impact is not symmetric around $E[\lambda_1] = 0.5$, with stronger improvements in the left tail than in the right one. This asymmetry is due to the increase in the convergence rate of the individual estimator as $\lambda_1$ moves into a near-unit root region. At the extreme, if $\lambda_1 \approx 0.8$, most of the other units will have smaller values of $\lambda_i$. Their own individual estimators will converge at a rate closer to $T^{-1/2}$. Accordingly, for larger values of $T$, the variance of the individual estimator of unit 1 may be significantly smaller than the variance of most of the individual estimators. This effect has little impact for $T = 30, 60$, but is more notable for $T = 600$.

As $T$ increases, more weight is placed on the individual estimator of unit 1, in line
with the discussion after theorem 2 (fig. 3). This effect is more pronounced in smaller cross-sections, for less flexible estimators, more extreme values of $\lambda_1$, and values of $\lambda_1$ where the individual estimator is more efficient ($\lambda_1 \approx 0.8$).

Additional simulation results are reported in the Online Appendix. We consider an additional data-driven large-$N$ specification, further focus parameters; perform simulations for the intermediate case $T = 180$; analyze the choice of tuning parameters for large-$N$ estimators; and examine the estimated weights. The evidence emerging from these simulations is in line with the results presented above.

5 Empirical Application

We illustrate our averaging methodology with an application to forecasting monthly unemployment rates for a panel of German regions. This setting provides a natural application for two reasons. First, the unemployment dynamics of German regions are heterogeneous due to differences in sectoral composition, regional laws, and historical trends such as the East-West divide (de Graaff et al., 2018). At the same time, using data on other regions at least partially improves prediction. (Schanne et al., 2010). Second, the performance of our methodology can be explicitly measured against realized unemployment rates. Our application contributes to the growing literature on forecasting regional unemployment (Schanne et al., 2010; Patuelli, Schanne, Griffith, and Nijkamp, 2012; Wozniak, 2020; Aaronson, Brave, Butters, Fogarty, Sacks, and Seo, 2022).

The regions of interest are the 150 German labor market districts (Arbeitsagenturbezirke, AABs) of the German Federal Employment Agency. Each AAB is medium-size region, between a NUTS-2 and a NUTS-3 region in size. Together, the 150 AABs cover all of Germany. The AABs are grouped into 10 regional directorates (RDs). These RDs correspond either to German federal states or unions of two states (NUTS-2).

We make use of monthly AAB-, RD-, and Germany-wide seasonally adjusted unemployment data from May 2007 to February 2024 (a total of 202 time series observations). The resulting panel is balanced with $N = 150$. All data is freely available from the Federal Employment Agency.

We model the AAB-level unemployment rate as a function of the past values of AAB-, RD-, and national-level unemployment rates. Specifically, let $y_{it}^{AAB}$ be the unemployment rate in the $i$th AAB at month $t$. Let $y_{it}^{RD}$ be the unemployment rate of the RD to which the $i$th AAB belongs. Finally, let $y_{t}^{DE}$ be the unemployment rate in Germany. Then $y_{it}^{AAB}$ is modeled as:

$$y_{jt}^{AAB} = \theta_0 + \theta_1 y_{i,t-1}^{AAB} + \theta_2 y_{i,t-1}^{RD} + \theta_3 y_{t-1}^{DE} + u_{i,t}, \quad \mathbb{E}[u_{i,t}|y_{i-1,t}^{AAB}, y_{i-1,t}^{RD}, y_{i-1,t}^{DE}] = 0. \quad (10)$$
In model (10), we allow both idiosyncratic and regional dynamics to drive the AAB-level unemployment rate, following Schanne et al. (2010). These dynamics may be heterogeneous between AABs, and all coefficients are AAB-specific.

For each AAB, we forecast $y_{i}^{AAB}$ with its conditional mean $E[y_{i}^{AAB}|y_{i-1}^{AAB}, y_{i-1}^{RD}, y_{i-1}^{DE}]$ implied by eq. (10). Formally, the target parameter for the $i$th AAB in month $t$ is $\mu(\theta_{i}) = \theta_{i0} + \theta_{i1}y_{i-1}^{AAB} + \theta_{i2}y_{i-1}^{RD} + \theta_{i3}y_{i-1}^{DE}$. Observe that the period ($t-1$) unemployment rates are treated as part of the parameter $\mu$.

The key measure of interest in our study is the out-of-sample forecasting MSE of our unit averaging approaches (see below). To estimate this MSE, we adopt a rolling-window approach. The data is split into all possible contiguous subsamples of window sizes $T = 40, 60, 80$ months (between 3 and 7 years of data). On each window we estimate the individual parameters of eq. (10) with OLS. We compute the one-step-ahead out-of-sample unit averaging forecasts and record the forecast error. These errors are used to estimate the MSE for each AAB and averaging approach. Note that estimating the MSE from rolling windows implicitly assumes that individual parameters are stable over time, see remark 5 below for evidence in favor of this. We also note that the values of $T$ considered satisfy the heuristic criterion for moderate-$T$ of remark 1. The average $t$-statistic across coefficients, AABs, and $T$s is approximately 2.

We estimate the conditional mean using our fixed-$N$ and large-$N$ minimum MSE estimators. For the large-$N$ approach, we consider two specifications. For the Stein-like specification, only the target AAB is unrestricted. For the top units specification, we first run the fixed-$N$ estimator. The 15 AABs (10% of total) with the largest weights are set as unrestricted units, and the rest are restricted, and the large-$N$ estimator is then ran (see also the discussion in section 4). The choice of the number of top units is explored in the Online Appendix. The pre-averaging fixed-$N$ procedure is done for every AAB in every window subsample. The performance of our minimum MSE unit averaging estimator is benchmarked against the individual, mean group, and AIC-weighted averaging estimators.

| $T$ | Fixed-N | Large-N (Stein) | Large-N (top units) | Mean Group | AIC |
|-----|---------|-----------------|---------------------|------------|-----|
| 40  | 0.62    | 1.08            | 0.66                | 1.05       | 2.56|
| 60  | 0.76    | 1.21            | 0.76                | 1.20       | 2.35|
| 80  | 0.91    | 1.34            | 0.87                | 1.33       | 2.02|

Table 1: Average (across AAB) MSE of unit averaging estimators relative to the individual estimator.

Figures 4-6 visualize our results for the MSE. Fig. 4 provides a box plot for the MSE for all averaging approaches relative to the MSE of the individual estimator, along with a box plot of the (absolute) MSE of the individual estimator. Table 1 complements fig. 4 with
Figure 4: Left panel: distribution of relative MSEs across AABs. Split by different averaging strategies and estimation window size. Right panel: (absolute) MSE of the individual estimator. Both: whiskers – 10th and 90th percentiles; box boundaries – 25th and 75th percentiles; box crossbar – median.

Figure 5: Geographic distribution of MSE to $T = 40$. Thin lines denote borders of AABs. Left and right panels: MSE of minimum MSE fixed-N and individual estimators respectively. Middle panel: ratio of MSE of fixed-N estimator to individual estimator.

Figure 6: Best averaging approach for every AAB for $T = 40, 60, 80$. Thin lines denote borders of AABs. The average relative MSEs. The underlying geographic distribution of the MSE is plotted on fig. 5 for $T = 40$. Finally, on fig. 6 we compare the individual and the minimum MSE estimators, and depict the best performing approach for each AAB and each value of $T$. 

22
Maps for all of the averaging approaches and $T$ are provided in the Online Appendix.

Our key finding is that averaging with minimum MSE weights generally improves forecasting performance. For most AABs, at least one minimum MSE approach outperforms the individual estimator for all $T$, as can be seen on fig. 6. The gain in MSE can be substantial, as fig. 4 and table 1 show. These gains are stronger for regions where the individual estimator does relatively poorly (fig. 5); these regions are predominantly concentrated in the former East Germany. The improvement is also stronger for smaller values of $T$, although it is also non-negligible even for $T = 80$.

The fixed-$N$ and the top units large-$N$ minimal MSE estimators emerge as the leading averaging approaches, in line with the simulation evidence of section 4. Both offer roughly similar gains in MSE (fig. 4). For $T = 40$, greater flexibility makes the fixed-$N$ approach the overall best, as fig. 6 shows. For $T = 80$, the leading option is the top units large-$N$ estimator, which has only 15 unrestricted units.

The other averaging methods considered perform somewhat worse. Mean group and AIC weights do not improve forecasting performance on average, although they offer an improvement for a non-trivial share of AABs. The Stein-like large-$N$ performs similarly to the mean group estimator, but with smaller variation in the MSEs across AABs.

**Remark 5 (Individual parameter stability).** Estimating AAB-level MSE implicitly requires that the unemployment rate dynamics of eq. (10) are stable over time. As our sample covers 2007-2024, the key possible threat to this stability is the Covid-19 pandemic. However, we find no evidence of a corresponding change in dynamics. First, the literature finds that employment dynamics are stable across the pre-, intra-, and post-pandemic periods due to the strong German Kurzarbeit scheme, both on the regional (Aiyar and Dao, 2021) and the national level (Adams-Prassl, Boneva, Golin, and Rauh, 2020; Casey and Mayhew, 2023). Second, we find no statistical evidence of coefficient breaks with a joint Chow coefficient breakpoint test with a Bonferroni-corrected 5% level critical value.

**Remark 6 (Additional empirical results).** The Online Appendix contains further results, including detailed maps of the MSE and results for several specifications of the top units approach. We also examine the averaging weights of the minimum MSE estimators.

We also provide an application to nowcasting quarterly GDP for a panel of European countries. As above, the minimum MSE estimator improves nowcasting performance relative to competing estimators. The gains are larger for shorter panels.
6 Conclusions

In this work we introduce a unit averaging estimator to recover unit-specific parameters in a general class of panel data models with heterogeneous parameters. The procedure consists in estimating the parameter of a given unit using a weighted average of all the unit-specific parameter estimators in the panel. The weights of the average are determined by minimizing an MSE criterion. The paper studies the properties of the procedures using a local heterogeneity framework that builds upon the literature on frequentist model averaging (Hjort and Claeskens, 2003a; Hansen, 2008). An application to forecasting regional unemployment for a panel of German regions shows that the procedure performs favorably for prediction relative to a number of alternative procedures.

References

D. Aaronson, S. A. Brave, R. A. Butters, M. Fogarty, D. W. Sacks, and B. Seo. Forecasting Unemployment Insurance Claims in Realtime With Google Trends. *International Journal of Forecasting*, 38(2):567–581, 2022. ISSN 01692070. doi: 10.1016/j.ijforecast.2021.04.001.

A. Adams-Prassl, T. Boneva, M. Golin, and C. Rauh. Inequality in the Impact of the Coronavirus Shock: Evidence From Real Time Surveys. *Journal of Public Economics*, 189:104245, 2020. ISSN 00472727. doi: 10.1016/j.jpubeco.2020.104245.

S. Aiyar and M. C. Dao. The Effectiveness of Job-Retention Schemes: COVID-19 Evidence From the German States. *IMF Working Papers*, 2021(242):1, 2021. ISSN 1018-5941. doi: 10.5089/9781513596174.001.

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer Berlin Heidelberg, 3 edition, 2006. ISBN 978-3-540-29586-0. doi: 10.1007/3-540-29587-9.

B. H. Baltagi. *Panel data forecasting*, volume 2. Elsevier B.V., 2013. ISBN 9780444627315. doi: 10.1016/B978-0-444-62731-5.00018-X.

B. H. Baltagi, G. Bresson, and A. Pirotte. To Pool or Not to Pool. In *The Econometrics of Panel Data*, chapter 16, pages 517–546. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-75892-1_16.

Y. Bao and A. Ullah. The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics*, 140(2):650–669, 2007. ISSN 03044076. doi: 10.1016/j.jeconom.2006.07.007.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model Selection: An Integral Part of Inference. *Biometrics*, 53(2):603–618, 1997. doi: 10.2307/2533961.
B. H. Casey and K. Mayhew. Kurzarbeit/Short-Time Working: Experiences and Lessons from the COVID-Induced Downturn. *National Institute Economic Review*, 263:47–60, 2023. doi: 10.1017/nie.2021.46.

T. de Graaff, D. Arribas-Bel, and C. Ozgen. Demographic Aging and Employment Dynamics in German Regions: Modeling Regional Heterogeneity. In *Modelling Aging and Migration Effects on Spatial Labor Markets*, chapter 11, pages 211–231. Springer Cham, 2018. ISBN 978-3-319-68563-2. doi: 10.1007/978-3-319-68563-2_11.

M. C. Donohue, R. Overholser, R. Xu, and F. Vaida. Conditional Akaike Information Under Generalized Linear and Proportional Hazards Mixed Models. *Biometrika*, 98(3):685–700, 2011. ISSN 00063444. doi: 10.1093/biomet/asr023.

Y. Gao, X. Zhang, S. Wang, and G. Zou. Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1):139–151, 2016. ISSN 18726895. doi: 10.1016/j.jeconom.2015.07.006.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007. ISSN 00129682. doi: 10.1111/j.1468-0262.2007.00785.x.

B. E. Hansen. Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350, 2008. ISSN 0304-4076. doi: 10.1016/j.jeconom.2008.08.022.

B. E. Hansen. Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132, 2016. ISSN 18726895. doi: 10.1016/j.jeconom.2015.09.003.

B. E. Hansen and J. S. Racine. Jackknife Model Averaging. *Journal of Econometrics*, 167(1):38–46, 2012. doi: 10.1016/j.jeconom.2011.06.019.

N. L. Hjort and G. Claeskens. Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003a. ISSN 01621459. doi: 10.1198/016214503000000828.

N. L. Hjort and G. Claeskens. Rejoinder to the Focused Information Criterion and Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464):938–945, 2003b.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. ISBN 9780521548236. doi: 10.1017/CBO9781139020411.

J. V. Issler and L. R. Lima. A panel data approach to economic forecasting: The bias-corrected average forecast. *Journal of Econometrics*, 152(2):153–164, 2009. ISSN 03044076. doi: 10.1016/j.jeconom.2009.01.002.

O. Kallenberg. *Foundations of Modern Probability*. Springer Cham, 3 edition, 2021. ISBN 978-3-030-61870-4. doi: 10.1007/978-3-030-61871-1.
E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Cham, 4 edition, 2022. doi: 10.1007/978-3-030-70578-7.

C.-A. Liu. Distribution Theory of the Least Squares Averaging Estimator. *Journal of Econometrics*, 186(1):142–159, 2015. ISSN 0148-396X. doi: 10.1227/01.NEU.0000349921.14519.2A.

L. Liu, H. R. Moon, and F. Schorfheide. Forecasting with Dynamic Pane Data Models. *Econometrica*, 88(1):171–201, 2020. doi: 10.2139/ssrn.2908529.

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.

G. S. Maddala, H. Li, and V. K. Srivastava. A Comparative Study of Different Shrinkage Estimators for Panel Data Models. *Annals of Economics and Finance*, 2(1):1–30, 2001. ISSN 15297373.

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. ISSN 00142921. doi: 10.1016/S0014-2921(02)00206-4.

R. Patuelli, N. Schanne, D. A. Griffith, and P. Nijkamp. Persistence of Regional Unemployment: Application of a Spatial Filtering Approach to Local Labor Markets in Germany. *Journal of Regional Science*, 52(2):300–323, 2012. ISSN 00224146. doi: 10.1111/j.1467-9787.2012.00759.x.

M. H. Pesaran and R. P. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 6061:473–477, 1995. ISSN 0045-9801.

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. ISSN 1537274X. doi: 10.1080/01621459.1999.10474156.

B. M. Pötscher and I. R. Prucha. *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer, 1997. ISBN 3662034867.

A. E. Raftery and Y. Zheng. Discussion: Performance of Bayesian Model Averaging. *Journal of the American Statistical Association*, 98(464):931–938, 2003. ISSN 01621459. doi: 10.1198/016214503000000891.

P. Rilstone, V. K. Srivastava, and A. Ullah. The Second-Order Bias and Mean Squared Error of Nonlinear Estimators. *Journal of Econometrics*, 75(2):369–395, 1996. ISSN 03044076. doi: 10.1016/0304-4076(96)89457-7.

V. K. Rohatgi. Convergence of Weighted Sums of Independent Random Variables. *Mathematical Proceedings of the Cambridge Philosophical Society*, 69(2):305–307, 1971. doi: 10.1017/S0305004100046685.
N. Schanne, R. Wapler, and A. Weyh. Regional Unemployment Forecasts with Spatial Interdependencies. *International Journal of Forecasting*, 26(4):908–926, 2010. ISSN 01692070. doi: 10.1016/j.ijforecast.2009.07.002.

D. Staiger and J. H. Stock. Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557, 1997. ISSN 00129682. doi: 10.2307/2171753.

F. Vaida and S. Blanchard. Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, 92(2):351–370, 2005. doi: 10.1093/biomet/92.2.351.

A. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. ISBN 978-1-4757-2547-6. doi: 10.1007/978-1-4757-2545-2.

A. T. K. Wan, X. Zhang, and G. Zou. Least Squares Model Averaging by Mallows Criterion. *Journal of Econometrics*, 156(2):277–283, 2010. ISSN 0304-4076. doi: 10.1016/j.jeconom.2009.10.030.

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. ISSN 10991255. doi: 10.1002/jae.2696.

M. Wozniak. Forecasting the Unemployment Rate Over Districts With the Use of Distinct Methods. *Studies in Nonlinear Dynamics and Econometrics*, 24(2):657–666, 2020. ISSN 15583708. doi: 10.1515/snde-2016-0115.

S.-Y. Yin, C.-A. Liu, and C.-C. Lin. Focused Information Criterion and Model Averaging for Large Panels with a Multifactor Error Structure. *Journal of Business and Economic Statistics*, 39(1):54–68, 2021. doi: 10.1080/07350015.2019.1623044.

X. Zhang and C.-A. Liu. Inference After Model Averaging in Linear Regression Models. *Econometric Theory*, 35(4):816–841, 2019. ISSN 14694360. doi: 10.1017/S0266466618000269.

X. Zhang and C.-A. Liu. A Unified Approach to Focused Information Criterion and Plug-In Averaging Method. *Statistica Sinica*, 34:771–792, 2024. ISSN 10170405. doi: 10.5705/ss.202021.0266.

X. Zhang, G. Zou, and H. Liang. Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101(1):205–218, 2014. ISSN 00063444. doi: 10.1093/biomet/ast052.
Proofs of Results in the Main Text

Under assumption A.1 we work conditional on \{\eta_1, \eta_2, \ldots\}. We use \(E[\cdot]\) to denote the expectation operator conditional on \{\eta_1, \eta_2, \ldots\}, whereas \(E_{\eta}[\cdot]\) is the expectation taken with respect the distribution of \(\eta\). All results are shown to hold with probability one with respect to the distribution of \(\eta\) (denoted \(\eta\)-a.s.).

A.1 Proof of Lemma 1

Recall that the data vector \(z_{it}\) takes values in \(\mathcal{Z} \subset \mathbb{R}^d\) and define the data matrix \(z_i = (z_{i1}', \ldots, z_{iT}')'\) that takes values in \(\mathcal{Z}^T = \prod_{t=1}^T \mathcal{Z}\). Recall that that the parameter vector \(\theta = (\theta_1, \ldots, \theta_p)\) takes values in \(\Theta \subset \mathbb{R}^p\). We denote by \(\nabla m(\theta, z_{it})\) the gradient vector of \(m\) with respect to \(\theta\), by \(\nabla^2 m(\theta, z_{it})\) the Hessian matrix of \(m\) with respect to \(\theta\), by \(\nabla_{\theta_k} m(\theta, z_{it})\) the partial derivative of \(m\) with respect to \(\theta_k\), and by \(\nabla^2_{\theta_k \theta_k} m(\theta, z_{it})\) the gradient vector of \(\nabla_{\theta_k} m(\theta, z_{it})\) with respect to \(\theta\).

We establish a mean value theorem that does not require compactness of \(\Theta\).

Lemma A.1.1. Suppose assumption A.3 is satisfied. Then for each unit \(i\), any \(T\) and any \(k = 1, \ldots, p\) there exists a measurable function \(\hat{\theta}_{ik}\) from \(\mathcal{Z}^T\) to \(\Theta\) such that the individual estimator \(\hat{\theta}_i\) of eq. (2) satisfies

\[
\frac{1}{T} \sum_{t=1}^T \nabla_{\theta_k} m(\hat{\theta}_i, z_{it}) = \frac{1}{T} \sum_{t=1}^T \nabla_{\theta_k} m(\theta_i, z_{it}) + \left[ \frac{1}{T} \sum_{t=1}^T \nabla^2_{\theta_k \theta_k} m(\hat{\theta}_{ik}, z_{it}) \right]' \left( \hat{\theta}_i - \theta_i \right),
\]

where \(\hat{\theta}_{ik}\) lies on the segment joining \(\hat{\theta}_i\) and \(\theta_i\).

Further, suppose A.5 is satisfied. Then for each \(i\) and any \(T\) there exist measurable functions \(\theta_i, \hat{\theta}_i\) and \(\bar{\theta}_i\) from \(\mathcal{Z}^T\) to \(\Theta\) such that the individual estimator \(\hat{\theta}_i\) of eq. (2) satisfies

\[
\mu(\hat{\theta}_i) = \mu(\theta_i) + \nabla \mu(\bar{\theta}_i)'(\hat{\theta}_i - \theta_i) + \frac{1}{2}(\hat{\theta}_i - \theta_i)' \nabla^2 \mu(\hat{\theta}_i)(\hat{\theta}_i - \theta_i),
\]

(A.1.1)

\[
\mu(\hat{\theta}_i) = \mu(\theta_i) + d_1'(\hat{\theta}_i - \theta_i) + \frac{1}{2}(\hat{\theta}_i - \theta_i)' \nabla^2 \mu(\hat{\theta}_i)(\hat{\theta}_i - \theta_i),
\]

(A.1.2)

\[
\mu(\hat{\theta}_i) = \mu(\theta_i) + d_i'(\hat{\theta}_i - \theta_i) + \frac{1}{2}(\hat{\theta}_i - \theta_i)' \nabla^2 \mu(\hat{\theta}_i)(\hat{\theta}_i - \theta_i),
\]

(A.1.3)

where \(d_1 = \nabla \mu(\theta_1)\); \(\hat{\theta}_i\) and \(\bar{\theta}_i\) lie on the segment joining \(\hat{\theta}_i\) and \(\theta_i\); and \(\hat{\theta}_i\) lies on the segment joining \(\theta_i\) and \(\theta_i\).
Proof. Fix \( k \in \{1, \ldots, p\} \) and define the function \( f_i : \mathcal{Z}^T \times [0,1] \to \mathbb{R} \) as
\[
f_i(z_i, y) = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\theta}_i, z_{it}) - \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\theta_i, z_{it})
- \left[ \frac{1}{T} \sum_{t=1}^{T} \nabla^2_{\theta_k} m(y\hat{\theta}_i + (1-y)\theta_i, z_i) \right]' (\hat{\theta}_i - \theta_i) .
\]

A.3 implies that \( f_i \) is well-defined, as for each \( y \in [0,1] \) we have that \( y\hat{\theta}_i + (1-y)\theta_i \in \Theta \). \( f_i \) is a measurable function of \( z_i \) for every fixed value \( y \in [0,1] \), as \( \hat{\theta}_i \) and \( m \) are measurable functions of \( z_i \) and \( m \) is continuously differentiable in \( \theta \). \( f_i \) is a continuous function of \( y \) for every value of \( z_i \).

Define the correspondence \( \varphi_i : \mathcal{Z}^T \to [0,1] \) as \( \varphi_i(z_i) = \{y \in [0,1] : f_i(z_i, y) = 0\} \). The function \( f_i \) satisfies the assumptions of corollary 18.8 in Aliprantis and Border (2006), and so \( \varphi_i \) is a measurable correspondence. \( \varphi_i(z_i) \) is nonempty for every \( z_i \), as by the mean value theorem, for every fixed value of \( z_i \) there exists some \( \tilde{y} \in [0,1] \) such that
\[
\frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\theta}_i, z_{it}) = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\theta_i, z_{it})
+ \left[ \frac{1}{T} \sum_{t=1}^{T} \nabla^2_{\theta_k} m(\tilde{y}\hat{\theta}_i + (1-\tilde{y})\theta_i, z_i) \right]' (\hat{\theta}_i - \theta_i) .
\]

In addition, \( \varphi_i(z_i) \) is closed for every \( z_i \) as \( m \) is twice continuously differentiable in \( \theta \) by assumption A.3. Then by the Kuratowski-Ryll-Nardzewski measurable selection theorem (theorem 18.13 in Aliprantis and Border (2006)), \( \varphi_i(z_i) \) admits a measurable selector \( \tilde{y}_{ik} = \tilde{y}_{ik}(z_i) \). Finally, define \( \tilde{\theta}_{ik} = \tilde{y}_{ik}\hat{\theta}_i + (1-\tilde{y}_{ik})\theta_i \) and note that \( \tilde{\theta}_{ik} \) satisfies the requirements of the lemma. This establishes the first claim of the lemma.

The proof of the second claim of the lemma is analogous. \( \square \)

The following lemma is needed to prove lemmas 1 and A.2.1.

**Lemma A.1.2.** Suppose A.3 is satisfied. Let \( \tilde{\theta}_{ij} : \mathcal{Z}^T \to \mathbb{R}^p \) for \( j = 1, \ldots, p \) be a sequence of measurable functions that lie on the segment joining \( \theta_i \) and \( \hat{\theta}_i \), and define
\[
\hat{H}_{iT} = \begin{bmatrix}
\left[ \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_{i1}}^2 m(\tilde{\theta}_{i1}, z_{it}) \right]' \\
\vdots \\
\left[ \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_{ip}}^2 m(\tilde{\theta}_{ip}, z_{it}) \right]'
\end{bmatrix} .
\]
Then for all $T > T_0$ the matrix $\hat{H}_{iT}$ (i) is a.s. nonsingular and (ii) satisfies
\[
\mathbb{E} \left[ \left\| \hat{H}_{iT}^{-1} - \hat{H}_{iT}^{-1} \right\|_\infty^{2(2+\delta)(1+\delta)} \right] \leq p \frac{2(2+\delta)(1+\delta)}{\delta} \Delta_H \mathcal{C}_{\nabla^2m},
\]
where $H_i = \lim_{T \to \infty} \mathbb{E} \left[ T^{-1} \sum_{t=1}^T \nabla^2 m(\theta_i, z_{it}) \right]$.

Proof. The proof of assertion (i) is based on showing that $\left\| (H_i - \hat{H}_{iT}) H_i^{-1} \right\|_\infty < 1$ holds almost surely, which implies that the matrix $\hat{H}_{iT}$ is a.s. nonsingular. This result follows from the standard observation that if $\|I - A\|_\infty < 1$, then $A$ is nonsingular. Write $I = H_i H_i^{-1}$ and $A = \hat{H}_{iT} H_i^{-1}$. Then $\|I - A\|_\infty = \left\| (H_i - \hat{H}_{iT}) H_i^{-1} \right\|_\infty < 1$. The matrix $A$ is nonsingular, and $\hat{H}_{iT} = AH_i$ is a product of two nonsingular matrices.

Let $H_i^{-1} = (h_{ij})$ and observe that
\[
\hat{H}_{iT} H_i^{-1} = \begin{bmatrix}
\sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{i1}, z_{it}) h_{k1} & \cdots & \sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{i1}, z_{it}) h_{kp} \\
\sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{i2}, z_{it}) h_{k1} & \cdots & \sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{i2}, z_{it}) h_{kp} \\
\vdots & \ddots & \vdots \\
\sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{ip}, z_{it}) h_{k1} & \cdots & \sum_{k=1}^p \nabla_{\theta_i} m(\tilde{\theta}_{ip}, z_{it}) h_{kp}
\end{bmatrix}.
\]
Row $j$ of $\hat{H}_{iT} H_i^{-1} - I$ coincides with row $j$ of $\left( T^{-1} \sum_{t=1}^T \nabla^2 m(\tilde{\theta}_{ij}, z_{it}) \right) H_i^{-1} - I$. Then we have that
\[
\left\| (H_i - \hat{H}_{iT}) H_i^{-1} \right\|_\infty = \left\| \hat{H}_{iT} H_i^{-1} - I \right\|_\infty 
\leq \max_{1 \leq j \leq p} \left\| \left( T^{-1} \sum_{t=1}^T \nabla^2 m(\tilde{\theta}_{ij}, z_{it}) \right) H_i^{-1} - I \right\|_\infty 
\leq \sup_{\theta \in [\theta_i, \tilde{\theta}_i]} \left\| \left( T^{-1} \sum_{t=1}^T \nabla^2 m(\theta, z_{it}) \right) H_i^{-1} - I \right\|_\infty 
\equiv D_{iT},
\]
where the second inequality holds as all $\tilde{\theta}_{ij}$ lie on the segment joining $\theta_i$ and $\tilde{\theta}_i$ and where $D_{iT}$ is defined in A.3. A.3 implies $D_{iT} < 1$ a.s. for $T > T_0$, and thus $\left\| (H_i - \hat{H}_{iT}) H_i^{-1} \right\|_\infty < 1$ a.s. for $T > T_0$, which implies the first claim.

As $\hat{H}_{iT}$ is invertible for $T > T_0$ we have (Horn and Johnson, 2012, section 5.8)
\[
\left\| H_i^{-1} - \hat{H}_{iT}^{-1} \right\|_\infty \leq \left\| H_i^{-1} \right\|_\infty \frac{\left\| H_i^{-1} \hat{H}_{iT} - I \right\|_\infty}{1 - \left\| H_i^{-1} \hat{H}_{iT} - I \right\|_\infty} \leq \left\| H_i^{-1} \right\|_\infty \frac{D_{iT}}{1 - D_{iT}},
\]
where the last inequality follows from (A.1.4). Taking expectations, we obtain that
\[
\mathbb{E}\left[\left\| H_i^{-1} - \hat{H}_i^{-1}\right\|_{\infty}^{2(2+\delta)(1+\delta)}\right] \leq \left\| H_i^{-1}\right\|_{\infty}^{2(2+\delta)(1+\delta)} \mathbb{E}\left[\left(\frac{D_{iT}}{1 - D_{iT}}\right)^{2(2+\delta)(1+\delta)} \right]
\]
\[
\leq \frac{p^{(2+\delta)(1+\delta)}}{(1+\delta)^2} C_{\nabla^2 m}
\]

which establishes the second claim. \(\square\)

**Proof of lemma 1.** A.3 and Lemma A.1.1 imply that
\[
0 = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\theta}_i, z_{it})
\]
\[
= \frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k} m(\theta_i, z_{it}) + \left[\frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k}^{2} m(\hat{\theta}_{ik}, z_{it})\right]' (\hat{\theta}_i - \theta_i),
\]
where \(\hat{\theta}_{ik}\) lies on the segment joining \(\theta_i\) and \(\hat{\theta}_i\). Define the matrix
\[
\hat{H}_{iT} = \begin{bmatrix}
\left[\frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k}^{2} m(\hat{\theta}_{i1}, z_{it})\right]'

\vdots

\left[\frac{1}{T} \sum_{t=1}^{T} \nabla_{\theta_k}^{2} m(\hat{\theta}_{ip}, z_{it})\right]'
\end{bmatrix}.
\]

(A.1.5)

As all \(\hat{\theta}_{ik}\) lie between \(\theta_i\) and \(\hat{\theta}_i\), by lemma A.1.2 the matrix \(\hat{H}_{iT}\) is a.s. nonsingular for \(T > T_0\). Observe that \(\hat{\theta}_i - \theta_i = (\hat{\theta}_i - \theta_1) - (\theta_i - \theta_1)\). Combining the above two observations, we obtain that for \(T > T_0\) it holds that
\[
\sqrt{T} (\hat{\theta}_i - \theta_1) = -\hat{H}_{iT}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) + (\eta_i - \eta_1).
\]

By assumption A.3 and lemma A.1.2, it holds that
\[
-\hat{H}_{iT}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_i, z_{it}) \Rightarrow N(0, V_i).
\]

The convergence is joint as all units are independent by A.2.

The second assertion follows from the delta method and the observation that \(\nabla \mu(\theta_1) = \nabla \mu(\theta_0 + T^{-1/2} \eta_1) \rightarrow \nabla \mu(\theta_0) = d_0\) under the continuity assumption of A.5. \(\square\)
A.2 Proof of Theorem 1

Before presenting the proof of theorem 1 we introduce a number of intermediate results.

**Lemma A.2.1.** Suppose A.1 and A.3 are satisfied. Let \( \delta \) be as in A.3. Then there exist finite constants \( C_{\hat{\theta},1}, C_{\hat{\theta},1+\delta/2}, C_{\hat{\theta},2}, C_{\hat{\theta},2+\delta} \), which do not depend on \( i \) or \( T \), such that the following moment bounds hold for the individual estimator (2) for all \( T > T_0 \)

\[
\mathbb{E} \left[ \left\| \sqrt{T}(\hat{\theta}_i - \theta_i) \right\|^k \right] \leq C_{\hat{\theta},k}, \quad k = 1, 1 + \delta/2, 2, 2 + \delta,
\]

\[
\mathbb{E} \left[ \left\| \sqrt{T}(\hat{\theta}_i - \theta_i) \right\|^2 \right] \leq C_{\hat{\theta},2} + 2C_{\hat{\theta},1} \| \eta_i - \eta_i \| + \| \eta_i - \eta_i \|^2.
\]

**Proof.** Let the matrix \( \hat{H}_{iT} \) be defined as in eq. (A.1.5). By lemma A.1.2 the matrix \( \hat{H}_{iT} \) is non-singular for \( T > T_0 \). Then, as in the proof of lemma 1, for \( T > T_0 \) it holds that

\[
\sqrt{T}(\theta_i - \theta_i) = -\hat{H}_{iT}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it})
\]

\[
= -\hat{H}_{iT}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it}) + \left( H_{iT}^{-1} - \hat{H}_{iT}^{-1} \right) \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it}) ,
\]

where \( H_i = \lim_{T \to \infty} \mathbb{E} \left( \nabla^2 T^{-1} \sum_{t=1}^T m(\theta_i, z_{it}) \right) \). We separately bound the \((2 + \delta)\)-th moment of the norm for the two terms above. For the first term we have

\[
\mathbb{E} \left[ \left\| \hat{H}_{iT}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it}) \right\|^{2+\delta} \right]
\]

\[
\leq \mathbb{E} \left[ \left\| \hat{H}_{iT}^{-1} \right\|^{2+\delta} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it}) \right\|^{2+\delta} \right]
\]

\[
\leq \left\| \hat{H}_{iT}^{-1} \right\|^{2+\delta} \mathbb{E} \left[ \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^T \nabla m(\theta_i, z_{it}) \right\|^{2+\delta} \right]
\]

\[
\leq \lambda_{\hat{H}}^{-2-\delta} C_{\nabla m}^{2+\delta} ,
\]

where the first inequality follows from \( \| Ax \| \leq \| A \| \| x \| \), and the last line follows by assumption A.3 and by Jensen’s inequality.
For the second term we have

\[
\mathbb{E} \left[ \left\| \left( H_{i}^{-1} - \hat{H}_{i}^{-1} \right) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_{t}, z_{it}) \right\|_{2+\delta}^{2+\delta} \right]
\]

\[
\leq p^{2+\delta} \mathbb{E} \left[ \left\| \left( H_{i}^{-1} - \hat{H}_{i}^{-1} \right) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_{t}, z_{it}) \right\|_{\infty}^{2+\delta} \right]
\]

\[
\leq p^{2+\delta} \mathbb{E} \left[ \left\| H_{i}^{-1} - \hat{H}_{i}^{-1} \right\|_{\infty}^{2+\delta} \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_{t}, z_{it}) \right\|_{\infty}^{2+\delta} \right]
\]

\[
\leq p^{2+\delta} \left( \mathbb{E} \left[ \left\| H_{i}^{-1} - \hat{H}_{i}^{-1} \right\|_{\infty}^{2(2+\delta)(1+\delta)} \right] \right)^{\frac{\delta}{2(1+\delta)}} \left( \mathbb{E} \left[ \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_{t}, z_{it}) \right\|_{\infty}^{2(1+\delta)} \right] \right)^{\frac{1+\delta/2}{1+\delta}}
\]

\[
\leq p^{2+\delta} \left( \frac{(2+\delta)(1+\delta)}{\delta} \Delta_{H} \frac{2(2+\delta)(1+\delta)}{\delta} C_{\nabla^{2} m} \right)^{\frac{\delta}{2(1+\delta)}} \left( \mathbb{E} \left[ \left\| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\theta_{t}, z_{it}) \right\|_{\infty}^{2(1+\delta)} \right] \right)^{\frac{1+\delta/2}{1+\delta}}
\]

where the second inequality follows from \( \|Ax\|_{\infty} \leq \|A\|_{\infty} \|x\|_{\infty} \); the third inequality from Hölder’s inequality applied with \( p = (1 + \delta)/(1 + \delta/2) > 1 \); the fourth inequality from lemma A.1.2, and the last line follows by assumption A.3. Finally, we conclude that

\[
\mathbb{E} \left[ \left\| \sqrt{T} (\hat{\theta}_{i} - \theta_{i}) \right\|_{2+\delta}^{2+\delta} \right]
\]

\[
\leq 2^{1+\delta} \left[ \Delta_{H}^{-2-\delta} C_{\nabla^{2} m}^{2+\delta} + p^{2+\delta} \left( \frac{(2+\delta)(1+\delta)}{\delta} \Delta_{H} \frac{2(2+\delta)(1+\delta)}{\delta} C_{\nabla^{2} m} \right)^{\frac{\delta}{2(1+\delta)}} C_{\nabla^{2} \mu}^{\frac{1+\delta/2}{1+\delta}} \right]
\]

\[
\equiv C_{\hat{\theta},2+\delta}
\]

where we note that \( C_{\hat{\theta},2+\delta} \) does not depend on \( i \) or \( T \). By Jensen’s inequality we have

\[
\mathbb{E} \left[ \left\| \sqrt{T} (\hat{\theta}_{i} - \theta_{i}) \right\|^{2} \right] \leq C_{\hat{\theta},2+\delta}^{2} \equiv C_{\hat{\theta},2};
\]

\[
\mathbb{E} \left[ \left\| \sqrt{T} (\hat{\theta}_{i} - \theta_{i}) \right\|^{1+\delta/2} \right] \leq C_{\hat{\theta},2+\delta}^{2} \equiv C_{\hat{\theta},1+\delta/2};
\]

\[
\mathbb{E} \left[ \left\| \sqrt{T} (\hat{\theta}_{i} - \theta_{i}) \right\| \right] \leq C_{\hat{\theta},2+\delta}^{2} \equiv C_{\hat{\theta},1},
\]

which establishes the first part of the claim.
Next we note that
\[
E \left[ \| \sqrt{T} (\hat{\theta}_i - \theta_1) \|^2 \right] = E \left[ T (\hat{\theta}_i - \theta_1)' (\hat{\theta}_i - \theta_1) \right] \\
\leq E \left[ \| \sqrt{T} (\hat{\theta}_i - \theta_1) \|^2 \right] + 2 E \left[ T (\hat{\theta}_i - \theta_1)' (\hat{\theta}_i - \theta_1) \right] + T (\theta_i - \theta_1)' (\theta_i - \theta_1) \\
\leq E \left[ \| \sqrt{T} (\hat{\theta}_i - \theta_1) \|^2 \right] + 2 \| \eta_i - \eta_1 \| E \left[ \| \sqrt{T} (\hat{\theta}_i - \theta_1) \| \right] + \| \eta_i - \eta_1 \|^2 \\
\leq C_{\theta,2} + 2 C_{\theta,1} \| \eta_i - \eta_1 \| + \| \eta_i - \eta_1 \|^2 ,
\]
where in the first inequality we add and subtract $\theta_i$ in both parentheses, in the third inequality we apply the Cauchy-Schwarz inequality to the cross term and observe that under A.1 $\sqrt{T} (\theta_i - \theta_1) = \eta_i - \eta_1$. This establishes the second part of the claim.

**Lemma A.2.2.** Suppose A.3 and A.5 are satisfied. Let $\delta$ be as in assumption A.3. Then for all $i$ and $T > T_0$ it holds that
\[
E \left[ \| \mu (\hat{\theta}_i) \|^{2+\delta} \right] < \infty \\
E \left[ \| \sqrt{T} (\mu (\hat{\theta}_i) - \mu (\theta_i)) \|^{2+\delta} \right] \leq C_{\mu,2+\delta} C_{\theta,2+\delta}
\]

**Proof.** Equation (A.1.1) in lemma A.1.1 implies $\mu (\hat{\theta}_i) = \mu (\theta_i) + \bar{d}_i (\hat{\theta}_i - \theta_i)$, where $\bar{d}_i = \nabla \mu (\hat{\theta}_i)$ for $\hat{\theta}_i$ on the segment joining $\theta_i$ and $\hat{\theta}_i$. Raising both sides to the power of $(2 + \delta)$ and applying the $C_r$ inequality we obtain that
\[
\left| \mu (\hat{\theta}_i) \right|^{2+\delta} \leq 2^{1+\delta} \left[ \left| \mu (\theta_i) \right|^{2+\delta} + \left| \bar{d}_i (\hat{\theta}_i - \theta_i) \right|^{2+\delta} \right].
\]
By assumption A.5 and the Cauchy-Schwarz inequality it holds that $\left| \bar{d}_i (\hat{\theta}_i - \theta_i) \right|^{2+\delta} \leq \left\| \bar{d}_i \right\|^{2+\delta} \left\| \hat{\theta}_i - \theta_i \right\|^{2+\delta} \leq C_{\mu,2+\delta} \left\| \hat{\theta}_i - \theta_i \right\|^{2+\delta}$, hence by lemma A.2.1 it follows that
\[
E \left[ \| \bar{d}_i (\hat{\theta}_i - \theta_i) \|^{2+\delta} \right] \leq \frac{C_{\mu,2+\delta} C_{\theta,2+\delta}}{T^{(2+\delta)/2}},
\]
where the constants are independent on $i$. Then both claims of the lemma follow.

We need an extension of a weighted law of large numbers due to Rohatgi (1971).

**Lemma A.2.3.** Suppose

(i) $X_1, X_2, \ldots$ is a sequence of independent random variables such that $E(X_1) = 0$ and $\sup_i E[|X_i|^{1+1/\gamma}] < \infty$ for some $\gamma \in (0, 1]$;
(ii) \( \{w_N\}_N \) with \( w_N \in \mathbb{R}^\infty \) is a sequence of weight vectors such that \( w_{iN} \geq 0 \) for \( i > 0 \), \( \sum_{i=1}^N w_{iN} \leq 1 \), and \( w_{jN} = 0 \) for \( j > N \);

(iii) \( w \in \mathbb{R}^\infty \) is a weight vector such that \( w_i \geq 0 \) for \( i > 0 \), \( \sum_{i=1}^\infty w_i \leq 1 \); and

(iv) \( \{w_N\} \) and \( w \) are such that \( \sup_i |w_{iN} - w_i| = O(N^{-\gamma}) \).

Then \( \sum_{i=1}^\infty w_{iN}X_i \) exists a.s. and \( \sum_{i=1}^N w_{iN}X_i \xrightarrow{\text{a.s.}} \sum_{i=1}^\infty w_iX_i \).

Observe that the limit sequence of weights can be defective. If \( w_{iN} = N^{-1} I_{i \leq N} \) (equal weights), the above result becomes a standard SLLN with a second moment assumption.

**Proof.** Define \( \tilde{w}_N \in \mathbb{R}^\infty \) by \( \tilde{w}_{iN} = w_{iN} - w_i \) for \( i \leq N \) and \( \tilde{w}_{iN} = 0 \) for \( i > N \). Then
\[
\sum_{i=1}^N w_{iN}X_i = \sum_{i=1}^N w_iX_i + \sum_{i=1}^N (w_{iN} - w_i)X_i = \sum_{i=1}^N w_iX_i + \sum_{i=1}^N \tilde{w}_{iN}X_i
\]
holds. For any \( n \) it holds that \( \sum_{i=1}^n \text{Var}(w_{iN}X_i) = \sum_{i=1}^n w_i^2 \mathbb{E}(X_i^2) = \mathbb{E}(X_i^2) \sum_{i=1}^n w_i^2 \leq \mathbb{E}(X_i^2) < \infty \) since \( \gamma \leq 1 \). Hence the Kolmogorov two-series theorem (Kallenberg, 2021, lemma 5.16) implies that \( \sum_{i=1}^N w_{iN}X_i \xrightarrow{\text{a.s.}} \sum_{i=1}^\infty w_iX_i \). The vector \( \tilde{w}_N \) satisfies the conditions of theorem 2 of Rohatgi (1971). Hence the same theorem implies that \( \sum_{i=1}^\infty \tilde{w}_{iN}X_i \xrightarrow{\text{a.s.}} 0 \).

The claim of the lemma then follows. \( \square \)

**Lemma A.2.4.** Suppose that the assumptions of theorem 1 are satisfied. Then (i) \( \sum_{i=1}^\infty w_i \eta_i \) exists \( \eta \)-a.s. and it holds that
\[
\sum_{i=1}^N w_{iN}(\eta_i - \eta_1) \xrightarrow{\text{a.s.}} \sum_{i=1}^\infty w_i \eta_i - \eta_1 ,
\]
and (ii) \( \sup_i \sum_{i=1}^N w_{iN} \| \eta_i - \eta_1 \|^k < \infty \) is finite \( \eta \)-a.s. for \( k = 1, 1 + \delta/2, 2, 2 + \delta \) for the choice of \( \delta \) in A.3.

**Proof.** Notice that \( \sum_{i=1}^N w_{iN}(\eta_i - \eta_1) = \sum_{i=1}^N w_{iN} \eta_i - \eta_1 \). By assumption A.1 \( \eta_i \) are independent random vectors with finite third moments and \( \sup_i |w_{iN} - w_i| = O(N^{-1/2}) \).

Lemma A.2.3 then implies that \( \sum_{i=1}^\infty w_i \eta_i \) exists \( \eta \)-a.s. and that \( \sum_{i=1}^N w_{iN} \eta_i \xrightarrow{\text{a.s.}} \sum_{i=1}^\infty w_i \eta_i \), which establishes the first claim.

Consider \( \| \eta_i - \eta_1 \|^k \) and note that the triangle and \( C_r \) inequalities imply that
\[
\| \eta_i - \eta_1 \|^k \leq (\| \eta_i \| + \| \eta_1 \|)^k \leq 2^{k-1}(\| \eta_i \|^k + \| \eta_1 \|^k) ,
\]
which, in turn, implies
\[
\sum_{i=1}^N w_{iN} \| \eta_i - \eta_1 \|^k \leq 2^{k-1} \sum_{i=1}^N w_{iN} \| \eta_i \|^k + 2^{k-1} \| \eta_1 \|^k . \tag{A.2.1}
\]
Observe that $\|\eta_i\|^k$ are independent random variables with $\sup_i \mathbb{E}_\eta [\|\eta_i\|^{3k}] < \infty$ for $k \in [1, 2 + \delta]$ by A.1. Then lemma A.2.3 applies with $\gamma = 1/2$, and $\sum_{i=1}^N w_i N \|\eta_i\|^k$ converges almost surely, which implies that $\sup_N \sum_{i=1}^N w_i N \|\eta_i\|^k < \infty \eta$-a.s.. Since $\|\eta_i\|$ is also $\eta$-a.s. finite, together with eq. (A.2.1), this implies the second claim. □

Finally, we present the proof of theorem 1.

**Proof of theorem 1.** First, from lemma A.2.2 it follows for each $N$ and $T > T_0$

$$
\mathbb{E} [\hat{\mu}(\omega_N) - \mu(\theta_1)]^2 < \infty ,
$$

establishing the second assertion of the theorem.

The MSE of the averaging estimator expressed as a sum of squared bias and variance is

$$
T \times \mathbb{E} [\hat{\mu}(\omega_N) - \mu(\theta_1)]^2 = \left( \sum_{i=1}^N w_i N \mathbb{E} \left( \sqrt{T}(\mu(\hat{\theta}_i) - \mu(\theta_1)) \right) \right)^2 + T \text{Var} \left( \sum_{i=1}^N w_i N (\mu(\hat{\theta}_i)) \right).
$$

We examine the bias and the variance separately. We first focus on the bias. By eq. (A.1.2) of lemma A.1.1, we have

$$
\mu(\hat{\theta}_i) = \mu(\theta_1) + d_1' (\hat{\theta}_i - \theta_1) + \frac{1}{2} (\hat{\theta}_i - \theta_1)' \nabla^2 \mu(\hat{\theta}_i) (\hat{\theta}_i - \theta_1),
$$

where $d_1 = \nabla \mu(\theta_1)$ and $\hat{\theta}_i$ lies on the segment joining $\hat{\theta}_i$ and $\theta_1$. The bias of $\mu(\hat{\theta}_i)$ is

$$
\sqrt{T} \mathbb{E} \left( \mu(\hat{\theta}_i) - \mu(\theta_1) \right)
= \mathbb{E} \left[ d_1' \sqrt{T}(\hat{\theta}_i - \theta_1) + \frac{1}{2} (\hat{\theta}_i - \theta_1)' \nabla^2 \mu(\hat{\theta}_i) \sqrt{T}(\hat{\theta}_i - \theta_1) \right]
= \mathbb{E} \left[ d_1' \sqrt{T}(\hat{\theta}_i - \theta_1) + \frac{1}{2} (\hat{\theta}_i - \theta_1)' \nabla^2 \mu(\hat{\theta}_i) \sqrt{T}(\hat{\theta}_i - \theta_1) \right]
+ \sqrt{T} d_0' (\hat{\theta}_i - \theta_1) + (d_1 - d_0)' \sqrt{T}(\hat{\theta}_i - \theta_1)
\quad + d_0' (\eta_i - \eta_1) + (d_1 - d_0)' (\eta_i - \eta_1) ,
$$

(A.2.3)

where in the first equality we use eq. (A.2.2); in the second equality $\theta_1$ is replaced by $\theta_i$ in the first term using $d_1' \sqrt{T}(\hat{\theta}_i - \theta_1) - d_1' (\eta_i - \eta_1) = d_1' \sqrt{T}(\hat{\theta}_i - \theta_i)$; $d_0 = \nabla \mu(\theta_0)$; and we use the locality assumption A.1 in the last equality as $\sqrt{T}(\theta_i - \theta_1) = \eta_i - \eta_1$. Define

$$
A_{i,T} \equiv \mathbb{E} \left[ d_1' \sqrt{T}(\hat{\theta}_i - \theta_1) \right] + \frac{1}{2} \mathbb{E} \left[ (\hat{\theta}_i - \theta_1)' \nabla^2 \mu(\hat{\theta}_i) \sqrt{T}(\hat{\theta}_i - \theta_1) \right] + (d_1 - d_0)' (\eta_i - \eta_1) ,
$$

36
and note that by eq. (A.2.3), the bias of the averaging estimator can be written as

\[ \sum_{i=1}^{N} w_{iN} \mathbb{E} \left( \sqrt{T}(\mu(\hat{\theta}_i) - \mu(\theta_1)) \right) = \sum_{i=1}^{N} w_{iN} d'_0(\eta_i - \eta_1) + \sum_{i=1}^{N} w_{iN} A_i T \quad \text{(A.2.4)} \]

We then proceed by showing that \( \left| \sum_{i=1}^{N} w_{iN} A_i T \right| \leq M/\sqrt{T} \to 0 \) for some constant \( M < \infty \) independent of \( N \)(recall that all statements are almost surely with respect to the distribution of \( \eta \) in line with assumption A.1, and \( M \) may depend on the sequence \( \{\eta_1, \eta_2, \ldots\} \)). Note that

1. By Hölder’s inequality, we obtain
   \[ \left| d'_i \mathbb{E} \left( \sqrt{T}(\hat{\theta}_i - \theta) \right) \right| \leq \|d_i\|_{\infty} \left\| \sqrt{T} \mathbb{E}(\hat{\theta}_i - \theta) \right\|_1 \leq C_{\nabla^2\mu} C_{\text{Bias}} T^{-1/2} \]
   where the last bound follows from assumptions A.4 and A.5;

2. By assumption A.5 the eigenvalues of \( \nabla^2\mu \) are bounded in absolute value by \( C_{\nabla^2\mu} \)
   Then
   \[ \mathbb{E}(\hat{\theta}_i - \theta)(\nabla^2\mu(\hat{\theta}_i)\sqrt{T}(\hat{\theta}_i - \theta)) \leq C_{\nabla^2\mu} T^{-1/2} \left[ C_{\hat{\theta},2} + 2C_{\hat{\theta},1} \|\eta_i - \eta_1\| + \|\eta_i - \eta_1\|^2 \right] \]
   where the bound is given by lemma A.2.1;

3. By assumption A.5, \( \|d_i - d_0\| = \|\nabla\mu(\theta_0 + T^{-1/2}\eta_i) - \nabla\mu(\theta_0)\| \leq C_{\nabla^2\mu} \|\eta_i\| T^{-1/2} \).

All the \( C \)-constants do not depend in \( i \). Combining the above results, we obtain by the triangle and Cauchy-Schwarz inequalities that

\[ |A_i T| \leq \frac{1}{\sqrt{T}} \left[ C_{\nabla\mu} C_{\text{Bias}} + C_{\nabla^2\mu} C_{\hat{\theta},2} + C_{\nabla^2\mu} \|\eta_i - \eta_1\|^2 + C_{\nabla^2\mu}(2C_{\hat{\theta},1} + \|\eta_1\|) \|\eta_i - \eta_1\| \right]. \]

Define

\[ M = C_{\nabla\mu} C_{\text{Bias}} + C_{\nabla^2\mu} C_{\hat{\theta},2} + C_{\nabla^2\mu} \sup_{N} \sum_{i=1}^{N} w_{iN} \|\eta_i - \eta_1\|^2 \]
\[ + C_{\nabla^2\mu} \left( 2C_{\hat{\theta},1} + \|\eta_1\| \right) \sup_{N} \sum_{i=1}^{N} w_{iN} \|\eta_i - \eta_1\| \]

and observe that \( M \) does not depend on \( N \) or \( T \), and by lemma A.2.4 \( M < \infty \) \( (\eta\text{-a.s.)} \).

Take the weighted average of \( A_i T \) to obtain

\[ \left| \sum_{i=1}^{N} w_{iN} A_i T \right| \leq \sum_{i=1}^{N} w_{iN} |A_i T| \leq \frac{M}{\sqrt{T}} \to 0 \text{ as } N, T \to \infty \quad \text{(A.2.5)} \]

By lemma A.2.4, \( \sum_{i=1}^{N} w_{iN} d'_0(\eta_i - \eta_1) \to \sum_{i=1}^{\infty} w_i d'_0(\eta_0 - \eta_i) \), where the infinite sum
exists. Combining this with eqs. (A.2.4) and (A.2.5), we obtain that the bias converges as \( N, T \to \infty \):

\[
\sum_{i=1}^{N} w_{i,N} \mathbb{E} \left( \sqrt{T} \left( \mu(\theta_i) - \mu(\theta_1) \right) \right) \to \sum_{i=1}^{\infty} w_i d'_0 \eta_0 - d'_0 \eta_i, \quad (\eta\text{-a.s.)} \quad (A.2.6)
\]

Now turn to the variance series and observe that

\[
T \times \text{Var} \left( \sum_{i=1}^{N} w_{i,N} (\mu(\theta_i)) \right) = T \sum_{i=1}^{N} w_{i,N}^2 \text{Var} \left( \mu(\theta_i) \right) = \sum_{i=1}^{N} w_{i,N}^2 \left[ \mathbb{E} \left( \sqrt{T} \left( \mu(\theta_i) - \mu(\theta_i) \right) \right)^2 - \left( \sqrt{T} \left( \mathbb{E} (\mu(\theta_i)) - \mu(\theta_i) \right) \right)^2 \right].
\]

We tackle the two sums separately. First we show that

\[
\sup_{N} \sum_{i=1}^{N} w_{i,N}^2 \left[ \sqrt{T} \left( \mu(\theta_i) - \mathbb{E} (\mu(\theta_i)) \right) \right]^2 = O(T^{-1})
\]

The argument is similar to that leading up to eq. (A.2.5). By eq. (A.1.3) of lemma A.1.1, we can expand \( \mu(\theta_i) \) around \( \theta_i \) to obtain that

\[
\sqrt{T} \left( \mathbb{E} (\mu(\theta_i)) - \mu(\theta_i) \right) = \mathbb{E} \left[ d'_1 \sqrt{T} (\hat{\theta}_i - \theta_i) + \frac{1}{2} (\hat{\theta}_i - \theta_i)' \nabla^2 \mu(\hat{\theta}_i) \sqrt{T}(\hat{\theta}_i - \theta_i) \right],
\]

for some \( \hat{\theta}_i \) on the segment joining \( \theta_i \) and \( \hat{\theta}_i \). Similarly to the above, we conclude by lemma A.2.1 and assumption A.4 that

\[
\mathbb{E} \left[ d'_1 \sqrt{T} (\hat{\theta}_i - \theta_i) \right] \leq \frac{C_{\nabla^2 \mu} C_{\text{Bias}}}{\sqrt{T}} \quad \text{and} \quad \mathbb{E} \left[ (\hat{\theta}_i - \theta_i)' \nabla^2 \mu(\hat{\theta}_i) \sqrt{T}(\hat{\theta}_i - \theta_i) \right] \leq \frac{C_{\nabla^2 \mu} C_{\theta_2}}{\sqrt{T}}.
\]

From this it immediately follows that

\[
\sum_{i=1}^{N} w_{i,N}^2 \left[ \sqrt{T} \left( \mathbb{E} (\mu(\theta_i)) - \mu(\theta_i) \right) \right]^2 \leq \frac{1}{T} \left[ C_{\nabla^2 \mu} C_{\text{Bias}} + C_{\nabla^2 \mu} C_{\theta_2} \right]^2, \quad (A.2.7)
\]

where the right hand side does not depend on \( i \) or \( N \).
Second, we show that

\[
\sum_{i=1}^{N} w_i^2 \mathbb{E}\left[ \sqrt{T} \left( \mu(\hat{\theta}_i) - \mu(\theta_i) \right) \right]^2 \to \sum_{i=1}^{\infty} w_i^2 d'_0 V_i d_0.
\]

Define \( X_{iT} = \mathbb{E}\left[ \sqrt{T} (\mu(\hat{\theta}_i) - \mu(\theta_i)) \right]^2 \). By lemma A.2.2 there exists a constant \( C_X < \infty \) that does not depend on \( i \) or \( T \) such that \( X_{iT} \leq C_X \) for \( T > T_0 \). Then

\[
\sum_{i=1}^{N} w_i^2 \mathbb{E}\left[ \sqrt{T} (\mu(\hat{\theta}_i) - \mu(\theta_i)) \right]^2 \\
\equiv \sum_{i=1}^{N} w_i^2 X_{iT} \\
= \sum_{i=1}^{N} w_i^2 d'_0 V_i d_0 + \sum_{i=1}^{N} (w_i^2 - w_i^2) d'_0 V_i d_0 + \sum_{i=1}^{N} (w_i^2 - w_i^2) (X_{iT} - d'_0 V_i d_0) \\
+ \sum_{i=1}^{N} w_i^2 (X_{iT} - d'_0 V_i d_0).
\]

We deal with the four sums separately:

1. By A.3, \( \sum_{i=1}^{N} w_i^2 d'_0 V_i d_0 \leq \hat{\lambda}_2 \|d_0\|^2 \). Accordingly \( \left\{ \sum_{i=1}^{N} w_i^2 d'_0 V_i d_0 \right\}_{N=1}^{\infty} \) forms a bounded non-decreasing sequence. Thus \( \sum_{i=1}^{N} w_i^2 d'_0 V_i d_0 \to \sum_{i=1}^{\infty} w_i^2 d'_0 V_i d_0 \).

2. Consider \( \sum_{i=1}^{N} (w_i^2 - w_i^2) d'_0 V_i d_0 \)

\[
\left| \sum_{i=1}^{N} (w_i^2 - w_i^2) d'_0 V_i d_0 \right| = \left| \sum_{i=1}^{N} (w_i - w_i)(w_i + w_i) d'_0 V_i d_0 \right| \\
\leq \sup_j |w_j - w_j| \sum_{i=1}^{N} (w_i N + w_i) d'_0 V_i d_0 \\
\leq 2\hat{\lambda}_2 \|d_0\|^2 \sup_j |w_j - w_j| \to 0 ,
\]

where we have used A.3.
3. Similarly we obtain that

\[
\left| \sum_{i=1}^{N} (w_{iN}^2 - w_i^2)(X_{iT} - d_0^i V_i d_0) \right| = \sum_{i=1}^{N} |w_{iN} - w_i|(w_{iN} + w_i)|X_{iT} - d_0^i V_i d_0| \\
\leq \sup_j |w_{jN} - w_j| \sum_{i=1}^{N} (w_{iN} + w_i)|X_{iT} - d_0^i V_i d_0| \\
\leq 2 \left[ \lambda \Sigma \Delta^2 ||d_0||^2 + C_X \right] \sup_j |w_{jN} - w_j| \to 0. 
\]

4. Last, we apply the dominated convergence theorem to show that \( \sum_{i=1}^{N} w_i^2 (X_{iT} - d_0^i V_i d_0) \to 0 \).

Define \( f_{N,T} : \mathbb{N} \to \mathbb{R} \) as \( f_{N,T}(i) = w_{iN}^2 (X_{iT} - d_0^i V_i d_0) \) if \( i \leq N \) and \( f_{N,T}(i) = 0 \) if \( i > N \). For each \( i \), \( \left\{ \sqrt{T}(\mu(\hat{\theta}_i) - \theta_i), T = T_0 + 1, \ldots \right\} \) form a family with uniformly bounded \((2 + \delta)\)th moments (by lemma A.2.2). By lemma 1 \( \sqrt{T}(\mu(\hat{\theta}_i) - \theta_i) \Rightarrow N(0, d_0^i V_i d_0) \), hence by Vitali’s convergence theorem the second moments converge as \( X_{iT} \to d_0^i V_i d_0 \). This convergence is equivalent to the observation that for each \( i \) \( f_{N,T}(i) \) converges to zero as \( N, T \to \infty \).

Next, \( f_{N,T} \) is dominated: for any \( i \) it holds that \( |f_{N,T}(i)| \leq w_{iN}^2 |X_{iT} - d_0^i V_i d_0| \leq w_i (C_X + \lambda \Sigma \Delta^2 ||d_0||^2) \). The bound is summable: \( \sum_{i=1}^{\infty} w_i (C_X + \lambda \Sigma \Delta^2 ||d_0||^2) \leq (C_X + \lambda \Sigma \Delta^2 ||d_0||^2) \), which is independent of \( N \) and \( T \).

The dominated convergence theorem applies and so

\[
\sum_{i=1}^{N} w_i^2 (X_{iT} - d_0^i V_i d_0) = \sum_{i=1}^{\infty} f_{N,T}(i) \to \sum_{i=1}^{\infty} 0 = 0 \text{ as } N, T \to \infty.
\]

Combining the above arguments, we obtain that as \( N, T \to \infty \)

\[
\sum_{i=1}^{N} w_i^2 \mathbb{E} \left[ \sqrt{T} \left( \mu(\hat{\theta}_i) - \mu(\theta_i) \right) \right]^2 \to \sum_{i=1}^{\infty} w_i^2 d_0^i V_i d_0 . \tag{A.2.8}
\]

Combining together equations (A.2.6), (A.2.7), and (A.2.8) shows that as \( N, T \to \infty \)

\[
T \times \mathbb{E} [\hat{\mu}(w_N) - \mu(\theta_1)]^2 \to \left( \sum_{i=1}^{\infty} w_1 d_0^i \eta_i - d_0^i \eta_1 \right)^2 + \sum_{i=1}^{\infty} w_i^2 d_0^i V_i d_0 .
\]

\(\square\)
A.3 Proof of Lemma 2

Proof of Lemma 2. First assertion: in notation of the proof of lemma 1, for $T > T_0$

$$\sqrt{T} \left( \hat{\theta}_i - \hat{\theta}_1 \right) = \eta_i - \eta_1 + \sqrt{T} \left( \hat{H}^{-1}_{i} \sum_{t=1}^{T} \nabla m(\theta_t, z_{it}) - \hat{H}^{-1}_{1} \sum_{t=1}^{T} \nabla m(\theta_1, z_{1t}) \right).$$

By lemma 1, the term in parentheses tends to $Z_i - Z_1 \sim N(\eta_i - \eta_1, V_i + V_1)$, as $Z_i$ and $Z_i$ are independent. Convergence is joint by lemma 1 since

$$\sqrt{T} \left( \hat{\theta}_i - \hat{\theta}_1 \right) = \sqrt{T} \left( \hat{\theta}_i - \theta_i \right) - \sqrt{T} \left( \theta_1 - \theta_1 \right).$$

Now turn to the second assertion. First, it holds that

$$\sqrt{T} \left( \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i - \theta_1 \right) \overset{p}{\rightarrow} -\eta_1$$

as $N, T \rightarrow \infty$ by theorem OA.1.1 in the Online Appendix, with the $\mu$ the identity map (which satisfies condition A.5). Then

$$\sqrt{T} \left( \hat{\theta}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i \right) = \sqrt{T} \left( \hat{\theta}_1 - \theta_1 \right) + \sqrt{T} \left( \theta_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i \right) \Rightarrow Z_1 + \eta_1 \sim N(\eta_1, V_1),$$

by lemma 1 and Slutsky’s theorem.

A.4 Proof of Theorems 2 and 3

Proof of theorem 2. Lemma 2 implies that

$$\sqrt{T}(\hat{\theta}_i - \hat{\theta}_1) \Rightarrow Z_i - Z_1$$

jointly for all $i = 1, \ldots, N$. Hence jointly for all $i$ and $j$ it holds that

$$\left[ \Psi_N \right]_{ii} \Rightarrow d'_0((Z_i - Z_1)(Z_i - Z_1)' + V_i)d_0 = \left[ \Psi_N \right]_{ii},$$

$$\left[ \Psi_N \right]_{ij} \Rightarrow d'_0((Z_i - Z_1)(Z_j - Z_1)'d_0 = \left[ \Psi_N \right]_{ij}, \quad i \neq j.$$

Note that $\hat{\Psi}_N$ is finite-dimensional, and all its elements jointly converge as $T \rightarrow \infty$. Then the continuous mapping theorem readily implies that for any $w^N \in \Delta^N$

$$LA-MSE_N(w^N) \Rightarrow LA-MSE_N(w^N) := w^N' \hat{\Psi}_N w^N,$$
which establishes the first claim. The second claim is an implication of the argmax theorem (theorem 3.2.2 in Van der Vaart and Wellner (1996)). The conditions of that theorem are satisfied since we have that

1. By the first assertion of the theorem, $\lim_{N \to \infty} L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}}) \Rightarrow \lim_{N \to \infty} L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}})$ for every $\mathbf{w}^{\bar{N}}$ in the compact set $\Delta^{\bar{N}}$.

2. The limit problem $\arg\min_{\mathbf{w} \in \Delta^{\bar{N}}} \mathbf{w}^{\bar{N}}\overline{\Psi} \mathbf{w}^{\bar{N}}$ is a problem of minimizing a strictly convex continuous function on a compact convex set $\Delta^{\bar{N}}$, hence it has a unique solution. Strict convexity of the objective function follows since $\overline{\Psi} \mathbf{w}$ is positive definite. To see that $\overline{\Psi} \mathbf{w}$ is positive definite, it is sufficient to observe that for any $\mathbf{w} \neq 0 \Rightarrow \mathbf{w} \overline{\Psi} \mathbf{w} \geq \min_{i: w_i \neq 0} w_i^2 d_i V_i d_0 > 0$. The inequality follows as $\mathbf{w} \overline{\Psi} \mathbf{w}$ is formally the MSE associated with the problem with individual variances given by $V_i$ and biases of the form $(Z_i - Z_1)$. Hence $\mathbf{w} \overline{\Psi} \mathbf{w} = \text{Bias}^2(\mathbf{w}) + \text{Variance}(\mathbf{w}) \geq \text{Variance}(\mathbf{w}) \geq 0$ since $V_i$ is positive definite by assumption A.3 and $d_0 \neq 0$.

3. The weights $\mathbf{w}^{\bar{N}}$ minimize $\lim_{N \to \infty} L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}})$ over the compact set $\Delta^{\bar{N}}$ for all $T$. Then the argmax theorem applies and $\mathbf{w}^{\bar{N}} \Rightarrow \overline{\Psi} \mathbf{w}^{\bar{N}} = \arg\min_{\mathbf{w} \in \Delta^{\bar{N}}} \mathbf{w}^{\bar{N}}\overline{\Psi} \mathbf{w}^{\bar{N}}$ as $T \to \infty$.

The third claim follows from joint convergence of the weights, the estimators being averaged, and the continuous mapping theorem.

**Proof of theorem 3.** First assertion: let $\mathbf{w}^{\bar{N}, \infty} \in \Delta^{\bar{N}}$. Then by lemma 2 and Slutsky’s theorem we conclude that as $N, T \to \infty$

$$L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}, \infty}) = \mathbf{w}^{\bar{N}, \infty}'\overline{\Psi} \mathbf{w}^{\bar{N}, \infty} + \left[ 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} \right] \left( \sqrt{T} d_1' \left( \hat{\theta}_1 - \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \hat{\theta}_i \right) \right)$$

$$+ 2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} d_1' \left( \hat{\theta}_1 - \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \hat{\theta}_i \right) \left( 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} \right) \left( \sqrt{T} d_1' \left( \hat{\theta}_1 - \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \hat{\theta}_i \right) \right)$$

$$\Rightarrow L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}, \infty})$$

$$:= \mathbf{w}^{\bar{N}, \infty}'\overline{\Psi} \mathbf{w}^{\bar{N}, \infty} + \left[ 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} \right] d_0 (\eta_1 + Z_1)$$

$$+ 2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} d_0' (Z_i - Z_1) \left[ 1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N}, \infty} \right] d_0 (\eta_1 + Z_1)$$

Second assertion: follows by the same logic as in the fixed-$N$ regime (theorem 2). The objective function $L\overline{\text{MSE}}_N(\mathbf{w}^{\bar{N}, \infty})$ can be represented as a quadratic function $\mathbf{x}' Q \mathbf{x}$,
where \( x \in \Delta^{N+1} \) stands in for \( (w_{i,\infty}^N, 1 - \sum_{i=1}^{N,\infty} w_i) \), and

\[
\hat{Q} = \left( \hat{\Psi}_N \begin{bmatrix} b \\ b' \end{bmatrix} T \begin{bmatrix} \hat{d}_1' (\hat{\theta}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i) \\ \vdots \\ \hat{d}_1' (\hat{\theta}_N - \hat{\theta}_1) \end{bmatrix} \right) \Rightarrow Q = \begin{bmatrix} \bar{\Psi}^{N,\infty} \\ \bar{b} \end{bmatrix} \begin{bmatrix} \bar{d}_0 (\eta_1 + Z_1) \end{bmatrix}^2 \]

\[
\hat{b} = \begin{bmatrix} -d_1' T (\hat{\theta}_1 - \hat{\theta}_1) (\hat{\theta}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i) ' d_1 \\ \vdots \\ -d_1' T (\hat{\theta}_N - \hat{\theta}_1) (\hat{\theta}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i) ' d_1 \end{bmatrix} \Rightarrow \bar{b} = \begin{bmatrix} d_0 (Z_1 - Z_1) (\eta_1 + Z_1)' d_0 \\ \vdots \\ d_0' (Z_N - Z_1) (\eta_1 + Z_1)' d_0 \end{bmatrix}.
\]

We now verify the condition of the argmax theorem for the problem of minimizing \( x' \hat{Q} x \) over \( \Delta^{N+1} \):

1. By the first assertion of the theorem, for any \( x \) in the compact set \( \Delta^{N+1} \) it holds that \( x' \hat{Q} x \Rightarrow x' \bar{Q} x \) as \( N, T \rightarrow \infty \) jointly.

2. The limit problem \( \arg \min_{x \in \Delta^{N+1}} x' \bar{Q} x \) is a problem of minimizing a strictly convex continuous function on a compact convex set \( \Delta^{N+1} \), hence it has a unique solution. Similarly to the above, strict convexity follows from positive definiteness of \( \bar{Q} \). To establish positive definiteness, first let \( x \neq 0 \) such that at least one of first \( \bar{N} \) coordinates are nonzero. For such an \( x \) it holds that \( x' \bar{Q} x \geq \min_{i=1,...,\bar{N}, x_i \neq 0} x_i^2 d_i' V_i d_i > 0 \) where the inequality follows as in the proof of theorem 2. Alternatively, if the first \( \bar{N} \) coordinates of \( x \) are zero, then \( x' \bar{Q} w = x_{N+1}^2 (d_0' (\eta_1 + Z_1))^2 > 0 \) ((\( Z_1 \))-a.s.).

3. The vector \( \hat{x}_{N,\infty}^{\bar{N},\infty} = (w_{i,\infty}^\bar{N}, 1 - \sum_{i=1}^{\bar{N}} w_i^{N,\infty}) \) minimizes \( x' \hat{Q} x \) over the compact set \( \Delta^{N+1} \) for all \( N > \bar{N}, T \).

Then the argmax theorem shows that \( \hat{x}_{N,\infty}^{\bar{N},\infty} \Rightarrow \bar{x}_{N,\infty}^{\bar{N},\infty} := \arg \min_{x \in \Delta^{N+1}} x' \bar{Q} x \). Finally, it is sufficient to observe that \( \hat{w}_{N,\infty}^{\bar{N},\infty} \) comprises the first \( \bar{N} \)-coordinates of \( \hat{x}_{N,\infty}^{\bar{N},\infty} \), and \( \hat{w}_{N,\infty}^{\bar{N},\infty} \) comprises the first \( \bar{N} \) coordinates of \( \bar{x}_{N,\infty}^{\bar{N},\infty} \).

The last assertion follows from the joint convergence of \( (\hat{w}_{N,\infty}^{\bar{N},\infty}), \sqrt{T} (\mu(\bar{\theta}_2) - \mu(\bar{\theta}_1)), \ldots, \sqrt{T} (\mu(\bar{\theta}_N) - \mu(\bar{\theta}_1)) \) as \( N, T \rightarrow \infty \), and from the fact that \( \sqrt{T} (\sum_{j=\bar{N}+1}^{N} v_j N - N \mu(\bar{\theta}_i) - \mu(\bar{\theta}_1)) \xrightarrow{p} -d_0' \eta_1 \) by theorem OA.1.1 in the Online Appendix.