How to Weight in Moment Matching: An ML Approach with Applications to Earnings Dynamics *

Xu Cheng† Alejandro Sánchez-Becerra‡ Andrew Shephard§

This Version: January 8, 2026

Abstract

Following the seminal paper by Altonji and Segal (1996), empirical studies commonly adopt equal or diagonal weighting in minimum distance estimation to mitigate finite-sample bias arising from sampling error in the weighting matrix. We propose a new weighting scheme that combines cross-fitting with regularized estimation of the weighting matrix, in the spirit of de-biased machine learning. We also propose a new formula for cross-fitted standard errors. We show that several canonical models in the earnings dynamics literature satisfy exact or approximate sparsity conditions that can be exploited by graphical lasso estimation of the weighting matrix. Within a many-moment asymptotic framework, we characterize the asymptotic distribution of the structural parameters. Extensive simulation studies demonstrate that our approach outperforms commonly used alternative weighting schemes. Finally, an empirical application using data from the Panel Study of Income Dynamics illustrates the practical gains of our method.

Keywords: cross-fitting, covariance structure model, de-biased machine learning, earnings dynamics, graphical lasso, many moments, minimum distance estimation, weighting matrix

JEL Codes: C01, C13, C18

1 Introduction

Minimum distance (MD) estimation is a popular approach for estimating structural econometric models by matching moments. For example, labor economists posit dif-

*We thank Natasha Gandhi for her excellent research assistance. We are grateful to numerous seminar and conference participants at the Barcelona Summer Forum, Bristol Econometrics Group, Brown University, Johns Hopkins University, KU Leuven, LACEA/LAMES, National University of Singapore, SETA Conference, Singapore Management University, Toulouse School of Economics, University of Pennsylvania, University of Southern California, and Washington University in St. Louis.

†University of Pennsylvania. E-mail: xucheng@econ.upenn.edu.

‡Emory University. E-mail: alejandro.sanchez.becerra@emory.edu.

§KU Leuven. E-mail: andrew.shephard@kuleuven.be.

ferent models of earnings dynamics and estimate them by minimizing the weighted distance between the sample cross-period covariance matrix and the model-implied counterpart.[1] Empirical researchers most frequently use either an identity weighting matrix, which assigns equal weights to the moments, or an inverse-variance diagonal weighting matrix.[2] The inverse sample covariance matrix of the moments, despite being the optimal weighting matrix in a standard asymptotic framework, is susceptible to large estimation errors and may lead to substantial finite-sample bias in the MD estimator, as documented by Altonji and Segal (1996) and others.

This paper proposes a new approach for weighting the moments, aiming for better finite-sample estimation and inference for the structural parameters. This new approach combines cross-fitting with regularized estimation of the weighting matrix. We also suggest using cross-fitting to estimate the asymptotic variance of the MD estimator. We consider a many-moment asymptotic framework where the number of moments p and the sample size n increase simultaneously. Compared to a standard fixed p asymptotic framework, this setup allows us to derive asymptotic results that better approximate the finite-sample bias due to sampling errors in the $p \times p$ weighting matrix. To accommodate applications in the earnings dynamics literature, we focus on the case $n, p \to \infty$ and $p/n \to 0$, so the number of moments is large but still substantially smaller than the sample size.[3] We compare the proposed method with common alternative weighting schemes and demonstrate its desirable theoretical properties and excellent finite-sample performance.

We show that in several examples motivated by the earnings dynamics literature, such as the covariance structure model, the high-dimensional oracle weighting matrix is either exactly sparse (containing only a small number of non-zero off-diagonal elements) or approximately sparse (asymptotically well-approximated by an exactly sparse matrix). Here, the oracle weighting matrix is defined as the inverse of the true population covariance of the moments. Crucially, its sparsity pattern is implied by the underlying economic model, which motivates using machine learning methods to exploit this structure and improve efficiency.

Our construction of the cross-fitted MD estimator based on a regularized weighting matrix estimator is akin to double/de-biased machine learning (DML) methods, where cross-fitting is applied to attenuate overfitting bias after estimating a high-dimensional function with machine learning methods; see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018a) for a review. In our setting, the

[1] See Abowd and Card (1989), MaCurdy (1982), Meghir and Pistaferri (2004), Guvenen (2007), and Altonji, Smith, and Vidangos (2013) for examples of earnings dynamics models and their estimation.

[2] Recent papers that apply equal-weighting include Baker and Solon (2003) and Meghir and Pistaferri (2004). Applications of diagonal weighting include Hyslop (2001), Blundell, Pistaferri, and Preston (2008), and Autor, Kostøl, Mogstad, and Setzler (2019).

[3] For consistency of the proposed weighting matrix, we may need some additional condition on p and n, such as $p \log p = o(n)$, which is only slightly stronger than $p = o(n)$.

nuisance component is a large-scale weighting matrix rather than a function of high-dimensional covariates. While machine learning estimators and cross-fitting methods relax regularity conditions in both contexts, they operate through distinct channels.[4]

Cross-fitting uses independent data splits to compute the weighting matrix and the sample moments, and then ensembles the resulting sample-splitting estimators to achieve the efficiency of a full-sample estimator, see, e.g., Angrist and Krueger (1995) for instrumental variable estimation and Altonji and Segal (1996) for MD estimation. It can be combined with any method to construct the weighting matrix. In the many-moment framework, we show that this cross-fitting procedure reduces the asymptotic bias of the MD estimator. Specifically, the full-sample estimator may have a substantial first-order asymptotic bias even when the weighting matrix estimator is consistent. In contrast, this asymptotic bias is eliminated by the cross-fitting procedure without requiring a particular rate of convergence for the weighting matrix. To capture this difference in first-order asymptotic bias between the full-sample estimator and the cross-fitted estimator, it is essential to let p grow along with n. In a standard fixed p asymptotic setup, this difference disappears in a first-order asymptotic analysis. The theoretical advantages of the cross-fitted estimator suggest that it is more robust to sampling errors in the weighting matrix, which is reflected in its performance in our simulation exercises.

Cross-fitting also yields significant benefits when used to estimate the asymptotic variance of the MD estimator. Theoretically, the cross-fitted estimator and the full-sample estimator have the same asymptotic variance, given by a sandwich formula that depends on the covariance matrix of the moments, the Jacobian matrix, and the weighting matrix. One can construct standard errors using either the full-sample method or the cross-fitting method. The cross-fitting method uses one fold to compute the covariance matrix of the moments and the Jacobian matrix and the remaining folds to compute the weighting matrix, yielding one estimate of the variance, and then aggregates these variance estimates across folds. The finite-sample performance of the full-sample variance estimate and the cross-fitting variance estimate differs considerably. Indeed, cross-fitted MD estimation together with the optimal weighting matrix (inverse sample covariance) was considered by Altonji and Segal (1996) as their independently weighted optimal MD estimator (IWOMD). However, they used the full-sample method to calculate the standard error. Using cross-fitted standard errors leads to a remarkable improvement in the finite-sample coverage of confidence intervals.

Regularized weighting matrix estimation could be viewed as a data-dependent

[4]Another important component of double/de-biased machine learning is to use scores that satisfy a Neyman orthogonality condition. This condition is automatically satisfied in our paper once we view the weighting matrix as a nuisance parameter in an MD estimation problem.

extension of the extreme regularization achieved by equal weighting and diagonal weighting. All methods control the sampling noise in the weighting matrix by reducing the number of parameters to estimate. The new weighting matrix proposed here is based on the graphical lasso (GLasso) estimator of an inverse covariance matrix (Friedman, Hastie, and Tibshirani, 2008). It allows some off-diagonal elements to be non-zero and estimates them together with the diagonal elements. The degree of regularization is data-driven.

We extend the theoretical results on the GLasso estimator in Rothman, Bickel, Levina, and Zhu (2008) from models with exact sparsity to those with approximate sparsity, see Belloni, Chernozhukov, and Hansen (2013) for a discussion of approximate sparsity in high-dimensional regression models. We show that the GLasso weighting matrix consistently estimates the oracle weighting matrix across a wide range of economic models where these sparsity conditions are verifiable—either through analytical derivations, as shown in our examples, or via the numerical methods we propose for more complex cases. Ultimately, our cross-fitted GLasso-weighted estimator achieves oracle-level efficiency without inflating asymptotic bias. This is our recommended estimator.

We investigate the finite-sample properties of our proposed estimator using two sets of simulations. In the first simulation study, we revisit the original design of Altonji and Segal (1996) and compare the proposed estimator with the estimators considered there under both their original design and the many-moment design. In the second simulation study, we consider the model in Baker and Solon (2003) to study earnings dynamics in an empirically rich environment. This model captures transitory and permanent income shocks, autoregressive lag dependence, time-varying volatilities, life-cycle effects, and cohort effects. We draw the simulated data using their structural model and parameter estimates. Overall, we find that the proposed estimator has the best performance in terms of bias, root-mean-square error, and the coverage probability of confidence intervals. We also show that different estimators have an important impact on the results when we replicate an earnings inequality decomposition exercise from Baker and Solon (2003).

Our simulation studies are complemented by an empirical application using data from the Panel Study of Income Dynamics (PSID). We apply our proposed estimator to decompose the variance of U.S. log earnings into permanent and transitory components. Our empirical results demonstrate that while different weighting schemes generally align on a long-term upward trend in income inequality, our proposed cross-fitted GLasso-weighted estimator provides the most precise estimates.

Our paper contributes to several strands of literature. Altonji and Segal (1996) investigate the small-sample bias of the optimal MD estimator due to the correlation between sampling errors in the weighting matrix and sampling errors in the mo-

ments. They provide extensive simulation evidence in favor of the equally weighted estimator. Clark (1996) supports this recommendation with additional simulation evidence based on nonlinear models. To overcome this bias, Horowitz (1998) proposes bootstrap bias correction and bootstrap confidence intervals for the optimal weighting estimator.

To improve the small-sample properties of the generalized method of moments (GMM) estimator and the MD estimator, many alternative methods have been proposed. Newey and Smith (2004) study the generalized empirical likelihood (GEL) estimator and develop a higher-order asymptotic approximation for the GEL, GMM, and MD estimators. In particular, they identify a higher-order bias stemming from the weighting matrix that corresponds to the finite-sample bias studied in this paper.[5] They show that some GEL estimators, such as the EL estimator, avoid this bias by bypassing weighting matrix estimation, while for the other estimators, the bias can be corrected analytically. Our approach differs in two ways: (i) we study it as a first-order bias when the number of moments grows with the sample size, rather than a higher-order bias when the number of moments is fixed; and (ii) we remove this bias via cross-fitting and regularization rather than analytical bias correction or alternative GEL objectives. For the two-step GMM estimator, Windmeijer (2005) shows that noise in the initial consistent estimator of the structural parameter enlarges the finite-sample variance of the two-step estimator and proposes a correction for it. Because our MD estimator does not rely on an initial structural estimate to construct the weighting matrix, the channel of our bias—and its resolution—is distinct.

The many-moment asymptotic framework where p increases with n has been used to study many instrumental variables (e.g., Bekker, 1994).[6] Han and Phillips (2006) and Newey and Windmeijer (2009) study an asymptotic bias due to a large number of moment conditions in a GMM framework, when the weighting matrix is non-random. They show that this bias can be removed by GEL estimators, and that this specific bias is irrelevant for a MD estimator even with many moments. Estimation based on many moments typically requires p to be much smaller than n.[7] One notable exception is Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018), who suggest a new regularized MD estimator for cases where the number of moments and parameters could be both much larger than n. None of these papers study estimation bias due to weighting matrix sampling errors.

Sample splitting and jackknife estimation have been widely applied to instrumental variable (IV) estimation; see Angrist and Krueger (1995) and Angrist, Imbens, and

[5] This bias is denoted by B_Ω in Theorem 4.1 of Newey and Smith (2004).

[6] This framework is used extensively to study many weak instruments, see, e.g., Chao and Swanson (2005), Andrews and Stock (2007), Newey and Windmeijer (2009), Mikusheva and Sun (2021).

[7] E.g., Newey and Windmeijer (2009) consider $p = o(n^{1/2})$ for consistency and $p = o(n^{1/3})$ for asymptotic normality for the continuous updating estimator in a linear model with heteroskedasticity.

Krueger (1999). To deal with noise in weighting matrix estimation in a MD problem, Altonji and Segal (1996) introduce the sample splitting estimator, and Kezdi, Hahn, and Solon (2002) develop jackknife estimation methods. For nonlinear problems, Kezdi, Hahn, and Solon (2002)'s estimator is different from our estimator even when we set the number of cross-fitting folds to n as in a jackknife estimator.[8] Our analysis further differs from that in Kezdi, Hahn, and Solon (2002) by incorporating regularized estimation (GLasso) of the weighting matrix and studying its asymptotic properties in a high-dimensional setting. In our simulation study, we compare our estimator to several alternative methods, including the bootstrap method of Horowitz (1998), the high-order bias correction in Newey and Smith (2004), and the jackknife methods by Kezdi, Hahn, and Solon (2002).

The regularized weighting matrix we use is taken from the machine learning literature on estimation of high-dimensional inverse covariance matrices. In addition to the GLasso estimator we adopt, many other estimators are available; see, e.g., Bickel and Levina (2008), Cai, Liu, and Luo (2011), Fan, Liao, and Mincheva (2011), among others. Each of these methods typically works under certain notions of sparsity. To link these machine learning methods to structural economic applications, it is crucial to demonstrate that the economic model satisfies the particular sparsity condition required by the chosen method. We make considerable efforts in this direction.

In the existing literature, regularization of the inverse of a covariance matrix has been successfully applied to improve estimation in linear models with many instruments, sometimes in conjunction with the jackknife method to mitigate many-instrument biases; see Hansen and Kozbur (2014) and Carrasco and Doukali (2017), for example. Carrasco and Nayihouba (2022) utilize a regularized inverse of the large-scale covariance matrix to achieve efficient estimation in dynamic panel models. Hausman, Lewis, Menzel, and Newey (2011) improve upon the continuous updating estimator by shrinking the weighting matrix toward the identity matrix. However, in each of these cases, the regularization techniques and the econometric contexts differ fundamentally from our approach, which focuses on exploiting model-implied sparsity in the weighting matrix within a cross-fitted MD framework.

This paper also contributes to empirical studies of earnings dynamics. The nature of earnings risk, and its separation into persistent and transitory components, is consequential for many decisions that individuals and households make over the life-cycle. For example, earnings risk is important in determining life-cycle labor

[8]For nonlinear problems, Kezdi, Hahn, and Solon (2002) estimates the Jacobian and weighting matrices on the same subsample while evaluating moments separately. In contrast, our GW estimator decouples the weighting matrix by estimating the Jacobian and moment conditions together on one subsample and the weighting matrix on a separate subsample. In addition, Kezdi, Hahn, and Solon (2002)'s estimator aggregates the first-order condition across folds, and our estimator aggregates the estimates across folds. Kezdi, Hahn, and Solon (2002) also introduce a different estimator for the linear model only, which is the same as IWOMD in a linear setup by setting the number of folds to n.

supply (Abowd and Card, 1989), consumption and savings (Gourinchas and Parker, 2002), and portfolio choice behavior (Angerer and Lam, 2009). We provide a novel method that enables efficient estimation and robust inference for both the structural parameters and the variance decomposition. To ensure the applicability of our proposed method to these applications, we validate the required conditions and conduct simulation studies using an empirical model from this literature.

The remainder of the paper is organized as follows. Section 2 discusses the proposed estimator based on cross-fitting and regularized weighting matrix estimation. An algorithm for the proposed estimator is provided at the end of this section. Section 3 provides a theoretical justification of the proposed estimator in a many-moment asymptotic framework. Sections 4 and 5 contain two simulation studies: one based on the design in Altonji and Segal (1996) and one based on a fully-fledged empirical model from Baker and Solon (2003). Section 6 provides an empirical application of the proposed method. Section 7 concludes. The Appendix provides extensions of our theoretical results, the main proofs, and a numerical demonstration of the model-implied sparsity patterns. The Supplementary Appendix contains auxiliary proofs, specific implementation details, and additional numerical results. The notations are collected as follows. For a vector a, let $||a||$ denote its Euclidean norm and a_r denote its row r. For a matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues, $||A|| = \sqrt{\lambda_{\max}(A'A)}$ denote the spectral norm, $||A||_F$ denote the Frobenius norm, and $A_{r\ell}$ denote its element in row r column ℓ.

2 Minimum Distance Estimation and Weighting

A structural model posits that, at the true parameter vector θ_0, the moment condition $\mathbb{E}[m_i] = f(\theta_0)$ holds for i.i.d. observed data $m_i : i = 1, \ldots, n$, and a known function $f(\theta) : \Theta \to \mathbb{R}^p$. We can estimate θ_0 by the MD estimator

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \left(\overline{m} - f(\theta)\right)' \widehat{W} \left(\overline{m} - f(\theta)\right), \tag{2.1}$$

where $\overline{m} = n^{-1} \sum_{i=1}^{n} m_i$ is the sample average of the observed moments and $\widehat{W}$ is a symmetric and positive definite weighting matrix that may be data-dependent. Because we consider over-identified models, the weighting matrix plays a crucial role in the asymptotic and finite-sample properties of this MD estimator.[9] Let $\Sigma = \mathrm{Var}(m_i)$. The optimal weighting matrix is $W^O = \Sigma^{-1}$. We call this the oracle weighting matrix because it is typically unknown in practice.

Let $\widehat{\Sigma} = n^{-1} \sum_{i=1}^{n} (m_i - \overline{m})(m_i - \overline{m})'$ denote the sample covariance matrix. In

[9] See Chamberlain (1984) for a detailed analysis of the MD estimator in a standard framework.

practice, commonly used weighting matrices include (i) equal weighting, with $\widehat{W}$ the identity matrix; (ii) diagonal weighting, with $\widehat{W}$ retaining only the diagonal elements of $\widehat{\Sigma}^{-1}$; (iii) inverse covariance weighting, i.e., $\widehat{W} = \widehat{\Sigma}^{-1}$. In a standard asymptotic framework where p is fixed, $\widehat{\Sigma}^{-1}$ is a consistent estimator of W^O, and inverse covariance weighting is also referred to as optimal weighting. In finite samples, $\widehat{\Sigma}^{-1}$ is susceptible to large sampling errors, especially when the dimension p is large. Furthermore, noisy estimation of the weighting matrix can translate into a large bias in MD estimates.

We propose two channels to reduce bias in the MD estimator through weighting matrix choices. The first channel is cross-fitting, a sample-splitting method that ensures independence between the weighting matrix and the sample moments by construction. We also provide a cross-fitted estimator of the asymptotic variance of the MD estimator. In the many-moment regime, where $p \to \infty$, we show that the cross-fitting approach eliminates a first-order asymptotic bias due to weighting matrix sampling errors. The second channel is regularized weighting matrix estimation. Exploiting the sparsity implied by many economic models, we suggest data-dependent regularization that allows for a small number of non-zero off-diagonal elements. In practice, we recommend combining these two approaches. We discuss cross-fitting and regularized weighting matrix estimation in Section 2.1 and Section 2.2, respectively. We conclude this section by summarizing the recommended algorithm.

2.1 Cross-Fitting and Bias Reduction

For a given weighting matrix $\widehat{W}$, we propose estimating θ via cross-fitting and constructing standard errors using the cross-fitting formula given below. We first describe the cross-fitting procedure and provide heuristic arguments for why it reduces first-order asymptotic bias of the MD estimator in the many-moment framework. A specific regularized choice of $\widehat{W}$ is provided in the following subsection.

Cross-Fitted Estimator. Split the sample randomly into a fixed number of K disjoint subsets of size $n_k = n/K$ for $k \in \{1, \ldots, K\}$. We refer to the observations in each subset as folds. Let $\mathcal{I}_k$ denote the set of indices in each fold $k \in \{1, \ldots, K\}$ and $\mathcal{I}_{-k} = \{1, \ldots, n\} \backslash \mathcal{I}_k$.

For observations in each fold k, we let $\overline{m}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} m_i$ be the sample mean and similarly define $\widehat{\Sigma}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} (m_i - \overline{m}_k)(m_i - \overline{m}_k)'$ as the sample covariance. Using observations in the other folds, i.e., $i \in \mathcal{I}_{-k}$, we compute the weighting matrix $\widehat{W}_{-k}$. The fold-k MD estimator is

$$\widehat{\theta}^{(k)} = \arg\min_{\theta \in \Theta} \left(\overline{m}_k - f(\theta)\right)' \widehat{W}_{-k} \left(\overline{m}_k - f(\theta)\right), \tag{2.2}$$

where the sample moments $\overline{m}_k$ and the weighting matrix $\widehat{W}_{-k}$ are independent. The K-fold cross-fitted estimator is

$$\widehat{\theta}^* = \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}^{(k)}. \tag{2.3}$$

Following the convention in the double/de-biased machine learning literature (e.g., Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018b), we treat the number of folds, K, as a fixed constant in our asymptotic analysis.

While the asymptotic analysis does not formally differentiate the finite-sample trade-offs when varying K, we can intuitively see that varying K has two competing effects. First, increasing K allows for a larger sample size ($n - n/K$) for weighting matrix estimation, which improves the precision of the regularized estimator and the efficiency of the MD estimator—we refer to this as the regularization effect. Second, increasing K reduces the estimation sample size ($n_k = n/K$) for each fold, which can inflate finite-sample bias in the structural parameters due to moment nonlinearity—this is the sample-splitting effect. Even after aggregation across K folds, the finite-sample bias in $\hat{\theta}^* = K^{-1}\sum_{k=1}^{K}\hat{\theta}_k$ may still increase with K in nonlinear models.[10] In sum, the regularization effect (which improves weighting matrix precision) favors a large K, whereas the sample-splitting effect (which minimizes finite-sample bias in the structural parameters) favors a small K. Our simulations, reported in Section 5, confirm these competing intuitions. However, we find that the estimation results are relatively robust to the choice of K even in moderate samples. Consequently, we adopt $K = 2$ as our default for the numerical results throughout the paper. Developing a formal asymptotic framework to derive the optimal choice of K is beyond the scope of the present study and is left for future research.

In recent work, Velez (2024) investigates de-biased machine learning estimators in an asymptotic framework where the number of cross-fitting folds K increases with the sample size n. The paper develops higher-order asymptotic results to characterize how K affects bias, variance, and MSE, using a device similar to that in Newey and Smith (2004).[11] However, the class of estimators in that paper does not encompass our estimator.[12]

Following the typology of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen,

[10]This increase in finite-sample bias can be explained by the higher-order asymptotic bias formula in Newey and Smith (2004): $\text{Bias}[n^{1/2}(\hat{\theta} - \theta_0)] = n^{1/2}\frac{1}{K}\sum_{k=1}^{K}\text{Bias}[\hat{\theta}_k - \theta_0] = \frac{CK}{n^{1/2}}$, for some constant C, following $\text{Bias}[\hat{\theta}_k - \theta_0] = B_I = C/n_k = CK/n$ by Theorem 4.1 and Theorem 4.6 of Newey and Smith (2004). Here, the bias term B_I is due to non-linearity of the moments.

[11]The findings in Velez (2024) are consistent with the higher-order approximations discussed in Footnote 10.

[12]Specifically: (i) Velez (2024) focuses on moments that are linear in the structural parameter and the structural parameter can be identified as the ratio of two expectations, whereas in our applications moments are generally nonlinear in structural parameters with no analytical solution; (ii) his nuisance parameter is a finite-dimensional function of some covariates, while ours is a high-dimensional weighting matrix whose dimension grows with the sample size n independently of covariates.

Newey, and Robins (2018b), our proposed estimator is a form of DML1 estimator. Our cross-fitting procedure involves solving a sequence of fold-specific MD problems and then aggregating the resulting estimates. In contrast, a DML2 estimator aggregates the fold-specific criterion functions into a single objective before solving one optimization problem. That is

$$\widehat{\theta}_{\text{DML2}} = \arg\min_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} \left(\overline{m}_k - f(\theta)\right)' \widehat{W}_{-k} \left(\overline{m}_k - f(\theta)\right). \tag{2.4}$$

These estimators are all first-order equivalent by our theory with fixed K. Our simulation results in Section 5 demonstrate that DML1 typically outperforms DML2 in the context of earnings dynamics models. Characterizing their higher-order discrepancies, and how these differences vary with K, remains a promising avenue for future research.

Cross-Fitted Standard Error. One can show that the asymptotic variance of the cross-fitted estimator is the usual sandwich formula $\Omega = (F'WF)^{-1}F'W\Sigma WF(F'WF)^{-1}$, where $F = \partial f(\theta_0)/\partial\theta$ is the Jacobian matrix and W is the limit of $\widehat{W}$; see Theorem 3.1 below. We suggest estimating Ω using the following cross-fitted variance estimator

$$\widehat{\Omega}^* = \frac{1}{K}\sum_{k=1}^{K} \widehat{\Omega}^{(k)}, \text{ where } \widehat{\Omega}^{(k)} = (\widehat{F}_k'\widehat{W}_{-k}\widehat{F}_k)^{-1}\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\widehat{F}_k(\widehat{F}_k'\widehat{W}_{-k}\widehat{F}_k)^{-1}, \tag{2.5}$$

where $\widehat{F}_k = \partial f(\widehat{\theta}^{(k)})/\partial\theta$, $\widehat{\Sigma}_k$, and $\widehat{W}_{-k}$, are all computed specifically for fold k. This variance estimator $\widehat{\Omega}^*$ delivers the cross-fitted standard error.

In the literature, a standard practice is to estimate the variance matrix Ω by replacing F, Σ, and W, with $\widehat{F} = \partial f(\widehat{\theta})/\partial\theta$, $\widehat{\Sigma}$, and $\widehat{W}$, respectively. All of these are computed using the full sample. This full-sample variance estimator delivers the full-sample standard error. Although both methods yield consistent estimates of the asymptotic variance, we show that confidence intervals based on cross-fitted standard errors have significantly better finite-sample performance than those based on full-sample standard errors.

Bias Reduction. Through the lens of a many-moment framework, we now illustrate heuristically how cross-fitting reduces estimation bias due to noise in weighting matrix estimation. The goal is to show that the MD estimator can exhibit first-order bias even when the weighting matrix estimator is consistent. Consider the linear model $f(\theta) = F\theta$ with $m_i \sim \mathcal{N}(f(\theta_0), \Sigma)$. We assume $||\widehat{W} - W|| \to_p 0$ for some non-random

matrix W. The full-sample MD estimator is

$$\sqrt{n}(\widehat{\theta} - \theta_0) = (F'\widehat{W}F)^{-1} \left[\underbrace{F'W\sqrt{n}\bar{g}(\theta_0)}_{A} + \underbrace{F'(\widehat{W} - W)\sqrt{n}\bar{g}(\theta_0)}_{B} \right], \tag{2.6}$$

where $\sqrt{n}\bar{g}(\theta_0) = n^{-1/2}\sum_{i=1}^{n}(m_i - f(\theta_0)) \sim \mathcal{N}(0, \Sigma)$. The first term, denoted by A, is based on the non-random limit W and follows a zero-mean normal distribution. However, the second term, denoted by B, can have different asymptotic limits for the full-sample estimator and the cross-fitted estimator in the many-moment case where $p \to \infty$. For the cross-fitted estimator, we always have $B \to_p 0$ because $\widehat{W} - W$ and $\sqrt{n}\bar{g}(\theta_0)$ are independent by construction. We prove this result through a conditioning argument in Theorem 3.1 below.

In contrast, the bias term B may not converge to 0 in probability for the full-sample estimator even if $\widehat{W}$ is consistent. This differs from the standard result for a finite number of moments. The estimation error in $\widehat{W} - W$ could be correlated with the sampling error in $\sqrt{n}\bar{g}$. As this correlation effect accumulates across moments, it results in a non-zero limit when the dimension p is high.

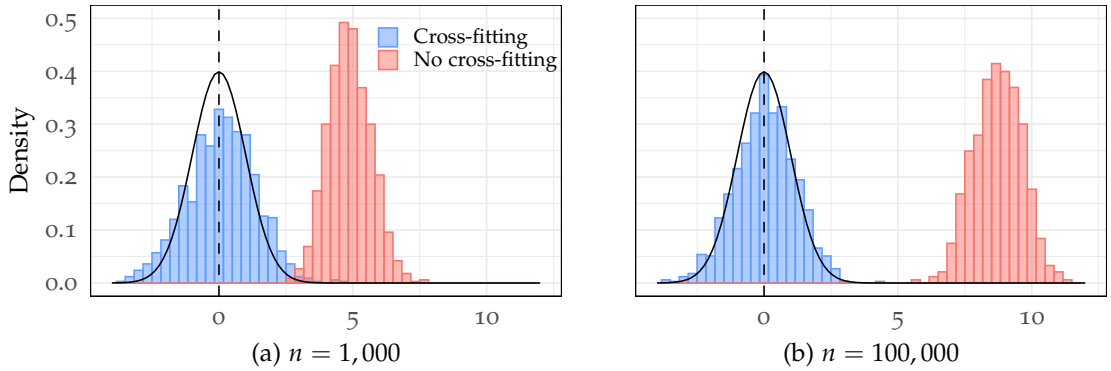
To illustrate this bias for the full-sample estimator, we consider the linear example above with a simple toy setup in which the bias can be calculated analytically. Let θ be a scalar, $\Sigma = I_p$, $F = (1, 0, \ldots, 0)' \in \mathbb{R}^p$, and let p be the nearest integer to $\sqrt{n}$. The weighting matrix estimator is $\widehat{W} = W^O + \Delta$. To facilitate the calculation, we artificially construct a symmetric positive definite matrix Δ such that (i) $||\Delta|| = o_p(1)$ so that $\widehat{W}$ is consistent, and (ii) $F'\Delta$, which selects the first row of Δ, is correlated with $\sqrt{n}\bar{g}(\theta_0)$. Under this setup, $B = (F'\Delta)\sqrt{n}\bar{g}(\theta_0)$, as defined in (2.6), can be studied analytically and shown to be non-zero as $n, p \to \infty$. One simple construction of such Δ is to set its first column and first row equal to $c_0 p^{-1}\sqrt{n}\bar{g}(\theta_0)$ and its transpose, respectively, for some constant c_0. We set $c_0 = 10$ in the simulation below so that the bias is large enough for a clear demonstration in Figure 1. All other elements of Δ are set to 0.[13] With this construction of Δ, it immediately follows that the bias term $B = c_0 p^{-1}||\sqrt{n}\bar{g}(\theta_0)||^2 \to_p c_0$.[14]

We conduct a Monte Carlo simulation based on this example. Figure 1 presents histograms of the t-statistic for two methods: (i) the full-sample estimator coupled with the full-sample standard error and (ii) the $K = 2$ cross-fitted estimator coupled

[13] In this toy example, both F and Δ are chosen for the purpose of mathematical illustration through simple calculations. The construction of Δ does not correspond to the standard optimal weighting matrix. However, the mathematical nature of the bias is captured by this toy example, and it suffices to show that consistency of $\hat{W}$ does not guarantee the absence of first-order bias B.

[14] To see that this convergence holds, note that $\sqrt{n}\bar{g}(\theta_0) \sim N(0, I_p)$, and therefore $||\sqrt{n}\bar{g}(\theta_0)||^2$ is the sum of p i.i.d. random variables with χ^2_1 distribution. The convergence then holds by the law of large numbers. In addition, we have $||\Delta|| = O_p(p^{-1/2}) = o_p(1)$ with this construction because $||\Delta||_F = O(\|p^{-1}\sqrt{n}\bar{g}(\theta_0)\|) = O_p(p^{-1/2})$ and $p \to \infty$.

with the cross-fitted standard error. This figure confirms that the full-sample estimator is biased and that the bias does not disappear as the sample size grows. The cross-fitting method effectively eliminates the bias.



(a) $n = 1,000$ (b) $n = 100,000$

Figure 1: Bias reduction under cross-fitting. This figure presents histograms of the t-statistic under two sample sizes when using (i) the full-sample estimator with the full-sample standard error and (ii) the $K = 2$ cross-fitted estimator with the cross-fitted standard error. The black solid line is the standard normal density. See the text for a description of the many-moment simulation design.

In sum, in the many-moment framework where $p \to \infty$, the cross-fitted estimator is more robust than the full-sample estimator. The cross-fitted estimator has zero asymptotic bias as long as the weighting matrix is consistent, without any requirement on the rate of convergence. The standard full-sample estimator, in contrast, can exhibit substantial bias. This provides a theoretical justification for our observation that the cross-fitted estimator has better finite-sample performance than its full-sample counterpart that uses the same weighting matrix estimation method.

2.2 Regularized Weighting Matrix Estimation

Regularized estimation of large inverse covariance matrices is well studied in the statistics and machine learning literature; see Fan, Liao, and Liu (2016) for a review. This literature considers different notions of sparsity and uses shrinkage methods to reduce dimensionality and improve estimation accuracy. Accordingly, before proposing a specific regularized estimator for W^O, we present examples motivated by our empirical application and describe their sparsity patterns.

2.2.1 Examples of Sparse Weighting Matrices

Let $S = \{(i,j) : W^O_{i,j} \neq 0,\ i \neq j\}$ with $s = |S|$ denoting the number of non-zero off-diagonal elements in the $p \times p$ matrix W^O. Although the total number of off-diagonal elements increases at rate p^2, we show that most are zero in these examples.

Motivated by the literature on earnings dynamics, we consider panel data $x_{i,t}$ for $i = 1,\ldots,n$ and $t = 1,\ldots,T$, where n is much larger than T. In this setting,

the moment m_i may comprise means or autocovariances over time for individual i. Here we provide stylised illustrations showing how the time-series properties of the moments yield the desired sparse structure. We also illustrate how zero partial correlations among moments from the same time period can lead to sparse structures in the weighting matrix. A fully-fledged empirical model is presented in Section 5.

Example 1. Matching Mean Structure Across Time. Consider the following $AR(1)$ process with additive time fixed effects: $x_{i,t} = \rho x_{i,t-1} + u_{i,t}$, where $u_{i,t}$ is i.i.d. across i and t with $\mathbb{E}[u_{i,t}] = 0$ and $\text{Var}[u_{i,t}] = \sigma_u^2$. Let $X_i = (x_{i,1}, \ldots, x_{i,T})'$ denote a vector of time-series processes for individual i. We will consider matching the mean of X_i to the prediction of the model. That is, $m_i = X_i$.

First, we consider a stationary time series with $|\rho| < 1$. In this case, the covariance matrix of m_i is dense because the autocorrelation is $\rho^{t-t'}$ for periods t and t'. However, the oracle weighting matrix $W^O = [\text{Var}(m_i)]^{-1}$ is sparse with a band-diagonal structure

$$W^O = \sigma_u^{-2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & (1+\rho^2) & -\rho & \cdots & 0 \\ 0 & -\rho & (1+\rho^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{2.7}$$

Second, we consider the unit-root case where $\rho = 1$. We assume that the process has an initial condition with finite variance $\sigma_0^2 = \text{Var}(x_{i,0})$. The oracle weighting matrix for the random-walk process also has a band-diagonal structure

$$W^O = \sigma_u^{-2} \begin{pmatrix} \frac{\sigma_0^2+2\sigma_u^2}{\sigma_0^2+\sigma_u^2} & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \tag{2.8}$$

Example 2. Matching Covariance Structure Across Time. Following Abowd and Card (1989), a large literature on earnings dynamics fits the sample autocovariance of wage data by imposing structural assumptions on the time series process. We assume that the data $x_{i,t}$ are determined by a weighted average of past shocks: $x_{i,t} = \sum_{t'=1}^{t} a_{t,t'}' u_{i,t'}$, where $u_{i,t'} \in \mathbb{R}^M$ is a vector of M mean-zero independent shocks and $a_{t,t'} \in \mathbb{R}^M$ is a vector of loadings. For example, consider $M = 1$ for a single shock. For an $AR(1)$ process, we can write $a_{t,t'} = \rho^{t-t'}$, assuming $x_{i,1} = u_{i,1}$. For an $MA(1)$

process, we can write $a_{t,t} = 1$, $a_{t,t-1} = \rho$, and $a_{t,t'} = 0$ for $t' < t-1$.

Let $X_i = (x_{i,1}, \ldots, x_{i,T})' \in \mathbb{R}^T$ and $U_i = (u'_{i,1}, \ldots, u'_{i,T})' \in \mathbb{R}^{MT}$ denote, respectively, the vector of individual time series and the vector of shocks. The time series process can be represented in a matrix form as $X_i = AU_i$, where A is a $T \times MT$ coefficient matrix with block entries $a'_{t,t'}$. The autocovariance structure of X_i is determined by that of AU_i.[15] Let $\text{vec}(\cdot)$ and $\text{vech}(\cdot)$ denote the vectorization and half-vectorization of a symmetric matrix. Let Γ denote the selector matrix that converts vectorization to half-vectorization, and let Γ_* denote the selector matrix that converts half-vectorization to vectorization. To match the autocovariance structure of X_i, we have

$$m_i = \text{vech}(AU_iU_i'A') = \Gamma(A \otimes A)\,\text{vec}(U_iU_i') = \Gamma(A \otimes A)\Gamma_*\,\text{vech}(U_iU_i'). \tag{2.9}$$

Therefore, the oracle weighting matrix $W^O = \text{Var}(m_i)^{-1}$ takes the form

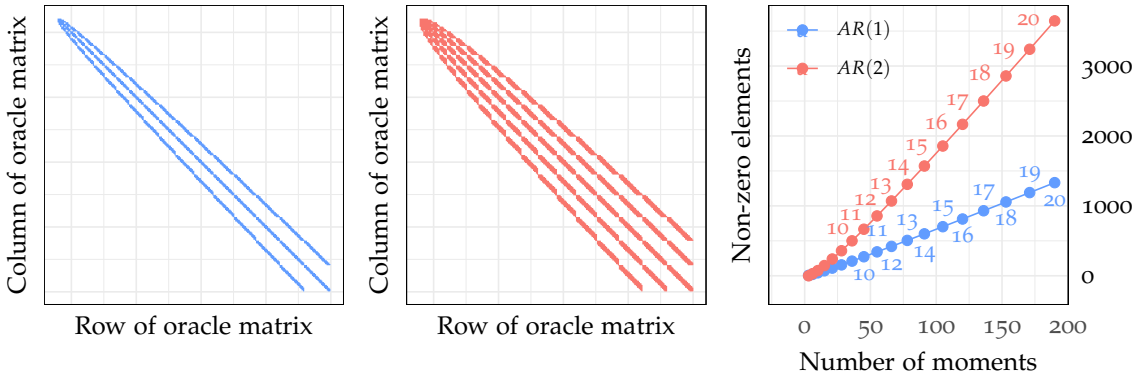
$$W^O = [\Gamma(A \otimes A)\Gamma_*\,\text{Var}(\text{vech}(U_iU_i'))\Gamma_*'(A' \otimes A')\Gamma]^{-1}, \tag{2.10}$$

where A is lower triangular because the time series only depends on past shocks. Because $u_{i,t}$ are independent across t, the matrix $\text{Var}(\text{vech}(U_iU_i'))$ has a sparse structure, and it is diagonal in the $M = 1$ case.

Although it is difficult to study the sparsity pattern of W^O in (2.10) analytically for an arbitrary matrix A, it is straightforward to compute this expression for specific dynamic processes and confirm its sparse structure. Here we consider a single shock and both an $AR(1)$ process and an $AR(2)$ process for the shock component, with $\rho = 0.9$ and $T = 1, \ldots, 20$. The number of moments is $p = T(T+1)/2$. Figure 2 illustrates the sparsity patterns of the oracle weighting matrices. Panels (a) and (b) plot the non-zero elements of the oracle weighting matrix for the $AR(1)$ and $AR(2)$ processes, respectively, when $T = 20$. Panel (c) shows that the number of non-zero elements in the oracle weighting matrix increases linearly with the number of moments, confirming the desired sparse pattern.

Example 3. Conditional Correlation Among Cross-Sectional Moments. In addition to matching moments over time, researchers often include multiple moments from a single time period. Write $m_i = (y_i, x_i')'$, where y_i is a scalar and $x_i = (x_{i,1}, \ldots, x_{i,p-1})'$ is a $p-1$ vector. Let $\beta \in \mathbb{R}^{p-1}$ denote the population coefficients of a regression of y_i on x_i. An element of β, denoted by β_r, is zero if and only if y_i and $x_{i,r}$ have zero partial correlation, i.e., they are uncorrelated conditional on all of the other variables in x_i. This zero element in β translates to a zero element in W^O following the block matrix inverse formula. We have p such regressions by cycling the role of y_i across

[15] In practice, the sample covariance matrix is computed with the mean μ replaced by the cross-sectional mean of X_i. This difference is negligible asymptotically, see Appendix E.1.



(a) $AR(1)$ sparsity pattern (b) $AR(2)$ sparsity pattern (c) Non-zero element count

Figure 2: Illustration of the sparsity pattern in the oracle weighting matrix. Panels (a) and (b) show the non-zero elements of the oracle weighting matrix when $T = 20$ for an $AR(1)$ and an $AR(2)$ process, respectively. Panel (c) shows the number of moments p and the number of non-zero off-diagonal elements in the oracle weighting matrix as the panel length T varies.

the elements of m_i. As such, W^O is sparse if there are many pairwise zero partial correlations among these moments.

Graphical models view each element of m_i as a vertex and encode partial correlations as edges. This graphical relationship among the elements of m_i is entirely captured by the oracle weighting matrix. Sparse graphical models can arise from the time structure in Examples 1 and 2, as well as from conditional independence in other types of spatial or network relationships.

2.2.2 Approximate Sparsity

The examples so far demonstrate exact sparsity in the sense that, after accounting for s non-zero elements, the remaining elements of W^O are exactly zero. In practice, however, many examples exhibit approximate sparsity patterns: the weighting matrix is dense, but it is well approximated by a sparse matrix, and the approximation error is sufficiently small.

Define

$$r_n = \sqrt{\frac{(p+s)\log(p)}{n}} \to 0. \tag{2.11}$$

When W^O is exactly sparse, Rothman, Bickel, Levina, and Zhu (2008) show that the graphical lasso estimator of W^O converges at the rate r_n in the Frobenius norm under suitable tail conditions. We introduce this estimator and a variation in the next subsection. To show that W^O is approximately sparse, we write $W^O = W^* + R$, where W^* is a symmetric, positive-definite, sparse matrix with s non-zero elements and R is the approximation error that satisfies

$$\|R\|_F = O(r_n), \tag{2.12}$$

i.e., the approximation error is no larger than the estimation error of an exactly sparse matrix. Below we verify this approximate sparsity condition with a few examples. The first example is based on a closed-form representation and analytical calculations. The subsequent examples are more complex and are based on a proposed numerical algorithm.

Example 4. Analytical Approximation with an Exponential Decay Rate. Consider an $MA(1)$ process $x_{i,t} = \theta u_{i,t-1} + u_{i,t}$, where $u_{i,t} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2)$. Suppose we match the mean of this MA process, i.e., $m_i = X_i$, where X_i is the $T \times 1$ vector that collects $x_{i,t}$. The covariance matrix of m_i, denoted by Σ, is a band matrix where $\Sigma_{t,t} = 1 + \theta^2$, $\Sigma_{t,t'} = \theta$ for $|t - t'| = 1$, and $\Sigma_{t,t'} = 0$ for $|t - t'| > 1$. In general, the inverse of Σ does not admit a simple expression for finite T. To illustrate the ideas, we focus on a convenient parameterization such that $\Sigma_{1,1}$ and $\Sigma_{T,T}$ are replaced by 1 in the large $T \times T$ matrix.[16] Now W^O, the precision matrix, takes the simple closed form

$$W^O_{t,t'} = \frac{(-\theta)^{|t-t'|}}{1-\theta^2}, \qquad t, t' = 1, \ldots, T. \tag{2.13}$$

The precision matrix of this $MA(1)$ process is dense, but its off-diagonal entries decay exponentially with the distance $|t - t'|$.

Given this band structure of W^O, we decompose it into two components: a sparse approximation W^* that captures the dominant local dependence and a remainder R that collects the exponentially small terms. Specifically, the sparse approximation W^* contains the k diagonal bands, i.e., $W^*_{t,t'} = W^O_{t,t'}$ for $|t - t'| \leq k$ and $W^*_{t,t'} = 0$ otherwise. For illustration, when $k = 1$, we have

$$W^O = \underbrace{\frac{1}{1-\theta^2}\begin{pmatrix} 1 & -\theta & 0 & 0 & 0 \\ -\theta & 1 & -\theta & 0 & 0 \\ 0 & -\theta & 1 & -\theta & 0 \\ 0 & 0 & -\theta & 1 & -\theta \\ 0 & 0 & 0 & -\theta & 1 \end{pmatrix}}_{\text{Sparse Approximation } (W^*)} + \underbrace{\frac{1}{1-\theta^2}\begin{pmatrix} 0 & 0 & 0 & -\theta^3 & \theta^4 \\ 0 & 0 & 0 & \theta^2 & -\theta^3 \\ \theta^2 & 0 & 0 & 0 & \theta^2 \\ -\theta^3 & \theta^2 & 0 & 0 & 0 \\ \theta^4 & -\theta^3 & \theta^2 & 0 & 0 \end{pmatrix}}_{\text{Remainder } (R)}.$$

We consider sequences in which the number of bands k increases with p and n at suitable rates to verify the approximate sparsity condition. As k increases, the number of non-zero off-diagonal elements in W^* increases and the norm of the approximation error $\|R\|_F$ decreases. Lemma 2.1 below suggests that, with an exponential decay in the off-diagonal elements, it suffices for k to grow slowly to approximate W^O well by

[16] This is compatible with $x_{i,1}$ and $x_{i,T}$ being $MA(1)$ processes with some different parameter values.

a sparse matrix with k diagonal bands. Under the approximate sparsity condition, the remainder is negligible and $W^* = W^O - R$ is positive definite for n large enough.

Lemma 2.1. *Approximate the matrix W^O in (2.13) with its k diagonal bands W^*. Suppose $p(\log n)(\log p) = o(n)$. Then, the approximation error $R = W^O - W^*$ satisfies the approximate sparsity condition in (2.12) if the choice of k satisfies $\log n = o(k)$. Furthermore, $r_n^2 = n^{-1}(p+s)\log p = o(1)$ if $pk \log p = o(n)$, where s is the number of non-zero off-diagonal elements in W^*.*

Numerical Approximation for General Cases. We propose a numerical algorithm for verifying approximate sparsity in more complicated cases where W^O does not have a simple structure (e.g., band diagonal). Given a user-specified data generating process, we compute W^O analytically for different values of p, e.g., using the formula in (2.10) for covariance-matching as in Figure 2. We then suggest a numerical algorithm to compute W^* based on a threshold rule and provide conditions to verify the approximate sparsity requirement in (2.12). To this end, we first rank all the off-diagonal elements of W^O by their absolute values. Let $\mathcal{I}_s$ collect the indices of the s largest off-diagonal elements of W^O. Then, $W^*_{i,j} = W^O_{i,j}$ if $(i,j) \in \mathcal{I}_s$ or $i = j$, and $W^*_{i,j} = 0$ otherwise. That is, the approximation error $R = W^O - W^*$ is composed of all $p^2 - p - s$ small off-diagonal elements of W^O measured in absolute value.

Inspired by the MA(1) process, we suggest $s = \sqrt{pn \log n}$. By construction, we have $r_n^2 = n^{-1}(p+s)\log p = o(1)$ under the mild condition $p(\log n)(\log p)^2 = o(n)$.[17] It is only slightly stronger than $p \log p = o(n)$. Note that neither the construction of W^* nor the choice of s is unique because the approximate sparsity condition in (2.12) is asymptotic. The proposal here is one way to obtain such an approximation.

To demonstrate that the approximation error $\|R\|_F$ is sufficiently small compared to r_n, we report

$$\|\widetilde{R}\|_F^2 \equiv \frac{\|R\|_F^2}{\overline{\text{Diag}}(W^*)^2 \times r_n^2}, \tag{2.14}$$

where $\overline{\text{Diag}}(W^*)$ is the average of the diagonal elements of W^* used for normalization. To verify the approximate sparsity of W^O through the condition in (2.12), we show $\|\widetilde{R}\|_F^2$ is bounded for large n and p.

Next, we use this numerical method to verify the approximate sparsity of W^O in a few examples. We first investigate two leading cases: matching first moments of an $AR(1)$ process and an $MA(1)$ process, respectively, as discussed in Examples 1 and 4 above. These are interesting cases because we have shown analytically that W^O

[17] Under this condition and the suggested choice of s for the numerical algorithm, Lemma 2.1 holds for the MA process with the number of diagonal bands k replaced by s/p, since $pk \log p/n = \sqrt{p(\log n)\log(p)^2/n}$.

exhibits exact sparsity for the $AR(1)$ process in Example 1 and approximate sparsity for the $MA(1)$ process in Example 4. The algorithm, however, does not make use of such knowledge when computing the approximation matrix W^*. These two cases are reported in Figures 3(a) and 3(b).[18] In addition, we investigate matching the second moments of the $AR(1)$ and $MA(1)$ processes as in Example 2; see Figures 3(c) and 3(d). We have previously demonstrated the exact sparsity of W^O when matching the second moments for the $AR(1)$ process.

Figure 3 plots the normalized approximation sparsity measure $\|\widetilde{R}\|_F^2$ for these four cases. In each case, we investigate different relative sizes of p and n, e.g., $n = p^\kappa \log p$ for different values of κ. Figures 3(a) and 3(c) confirm that numerical approximation for an exactly sparse matrix is very easy and that the approximation is accurate even for a relatively small sample size n given p. Figures 3(b) and 3(d) show that the numerical methods also work well for the approximately sparse cases. The approximation error becomes smaller when both n and p increase, particularly when n is relatively large given p.

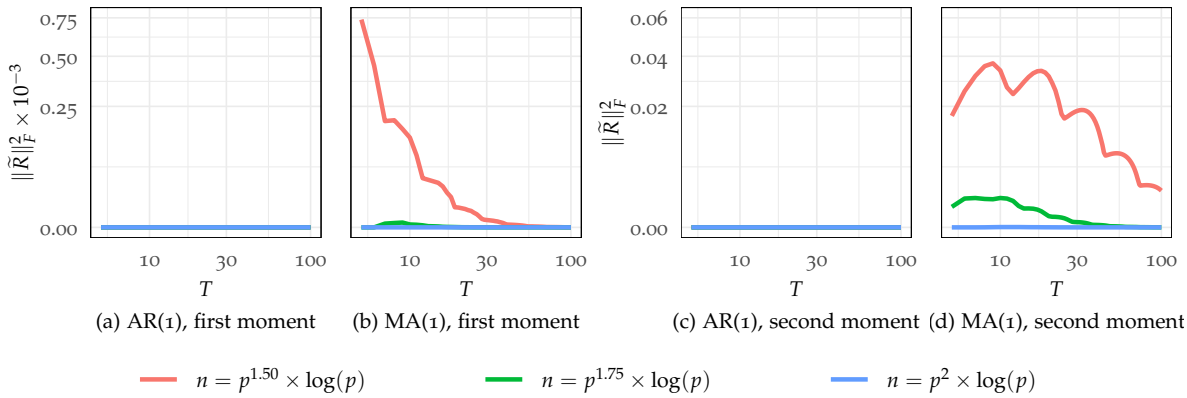


Figure 3: Approximate sparsity verification for $AR(1)$ and $MA(1)$ processes. The horizontal axis measures the number of time periods T, and the vertical axis measures the normalized approximate sparsity measure $\|\widetilde{R}\|_F^2$. For the first moment, $p = T$ and for the second moment $p = T(T-1)/2$.

Finally, we apply this numerical method to examples with both a transitory shock and a permanent shock, as well as individual heterogeneity, drawn from the earnings dynamics literature. In each case, we compute the weighting matrix W^O analytically, e.g., using the formula in (2.10) for the second moments matching, and use the proposed numerical methods to assess its approximate sparsity. We consider matching the first or second moments of $y_{i,t} = x_{i,t} + z_{i,t} + w_i$, where $x_{i,t}$ is the transitory shock that is specified as either an $AR(1)$ or $MA(1)$ process, where $z_{i,t} = z_{i,t-1} + v_{i,t}$ is the permanent shock, and w_i models individual heterogeneity.[19]

[18] The $AR(1)$ process is $x_{i,t} = \rho x_{i,t-1} + u_{i,t}$ with $\rho = 0.5$ and $x_{i,0}$ drawn from the stationary distribution. The $MA(1)$ process is $x_{i,t} = \theta u_{i,t-1} + u_{i,t}$ with $\theta = 0.3$. For both the AR and MA cases, we set $\{u_{i,t}\}_{t=0}^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

[19] The $AR(1)$ process is $x_{i,t} = \rho x_{i,t-1} + \sigma u_{i,t}$ with $\rho = 0.5$, $\sigma = \sqrt{0.5(1-\rho^2)}$, x_{i0} drawn from the

Figure 4 plots the normalized approximation sparsity measure $\|\widetilde{R}\|_F^2$ for the four cases described above with different combinations of n and p. These figures clearly demonstrate the approximate sparsity of W^O in all cases. For the first moment matching, the patterns in Figures 4(a) and 4(b) are similar to those from the correspond panels in Figure 3. For the second moments matching, adding a permanent shock and individual heterogeneity to an $AR(1)$ process increases the approximation error based on a comparison between Figures 3(c) and 4(c). Nevertheless, the approximation error measured by $\|\widetilde{R}\|_F^2$ stays below 1% in all cases, and it exhibits a downward-sloping pattern when n and p are large. Comparing Figure 3(d) to Figure 4(d), we see that adding a permanent shock and individual heterogeneity to an $MA(1)$ process does not change the overall pattern of the plots, and $\|\widetilde{R}\|_F^2$ stays below 0.5% in all cases. Figures 3 and 4 together show that the proposed numerical approximation method can effectively assess the sparse patterns of the optimal weighting matrices W^O generated from empirical examples.

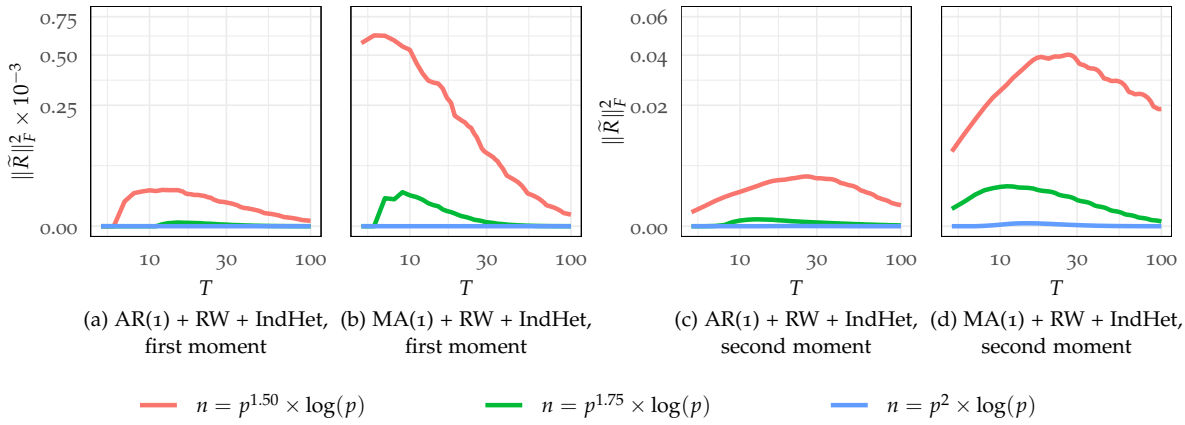


Figure 4: Approximate sparsity verification for $AR(1)$ + random walk + individual heterogeneity and $MA(1)$ + random walk + individual heterogeneity processes. The horizontal axis measures the number of time periods T, and the vertical axis measures the normalized approximate sparsity measure, $\|\widetilde{R}\|_F^2$. For the first moment, $p = T$ and for the second moment $p = T(T-1)/2$.

2.2.3 Weighting Matrix Estimation – Graphical Lasso (GLasso)

For applications where the oracle weighting matrix satisfies the sparsity condition demonstrated above, we provide a regularized weighting matrix estimator based on the penalized quasi-likelihood approach, (see, e.g., Yuan and Lin, 2007; Banerjee, El Ghaoui, and d'Aspremont, 2008). We follow the literature and call it the GLasso estimator based on the efficient computation algorithm proposed by Friedman, Hastie, and Tibshirani (2008) and its graphical interpretation. More specifically, we adopt the correlation-based version suggested by Rothman, Bickel, Levina, and Zhu (2008),

stationary distribution. The $MA(1)$ process is $x_{i,t} = \theta u_{i,t-1} + u_{i,t}$ with $\theta = 0.3$. For both the AR and MA cases, we set $\{u_{i,t}\}_{t=0}^T, \{v_{i,t}\}_{t=1}^T, w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and $z_{i,0} = 0$.

which only estimates and shrinks the off-diagonal correlation coefficients toward zero. Although many alternative regularized inverse covariance matrix estimators are available under the sparsity condition, this estimator has performed particularly well in our framework as a weighting matrix. Moreover, it automatically transitions between the inverse sample covariance (which is the so-called optimal weighting matrix in a fixed p framework) and the diagonal weighting matrix (which is frequently adopted by empirical researchers). The transition between these extremes is entirely data-driven, determined by the selection of the tuning parameter.

The correlation-based GLasso weighting matrix estimator is defined as follows. Let $\widehat{R} = \widehat{D}^{-1}\widehat{\Sigma}\widehat{D}^{-1}$ denote the sample correlation matrix, where $\widehat{D}$ is the diagonal matrix of sample standard deviations and $\widehat{\Sigma}$ is the sample covariance matrix. We first compute a GLasso estimator of the inverse correlation matrix

$$\widehat{Q}_G = \underset{Q \in \mathcal{W}}{\arg\max}\ \log(\det(Q)) - \operatorname{tr}(Q\widehat{R}) - \lambda \sum_{j \neq j'} |Q_{jj'}|, \tag{2.15}$$

where $\mathcal{W}$ is the space of $p \times p$ positive-definite matrices and λ is a tuning parameter and $Q_{jj'}$ denotes the element of Q in row j column j'. The correlation-based GLasso weighting matrix is

$$\widehat{W}_G = \widehat{D}\widehat{Q}_G\widehat{D}. \tag{2.16}$$

Here we use the subscript G to clarify that $\widehat{W}_G$ is estimated by the GLasso method, a specific choice for the general weighting matrix $\widehat{W}$.

The criterion function in (2.15) is equal to the log-likelihood of Q for a normal distribution plus a penalty on the off-diagonal correlation coefficients. As the tuning parameter λ moves from 0 to ∞, the solution transitions from the maximum likelihood estimator $\widehat{R}^{-1}$ to a diagonal matrix. In practice, we choose λ through cross-validation with the negative log-likelihood function as the loss function. The cross-validation procedure and computation details are described in Appendix E.2. To obtain the GLasso estimation in (2.15), we implement the R package `glassoFast` based on the algorithm in Sustik and Calderhead (2012). We configure this algorithm to only penalize the off-diagonal elements.

Let $\varepsilon_i = m_i - \mathbb{E}[m_i]$. Let c_0, c_1, c_2, c, C, δ denote some constants.

Lemma 2.2. *Suppose W^O is approximately sparse such that $W^O = W^* + R$, where W^* is a symmetric, positive-definite, sparse matrix with s non-zero off-diagonal elements and the approximation error satisfies $\|R\|_F = O(r_n) = o(1)$. Then $\|\widehat{W}_G - W^O\|_F = O_p(r_n)$ for i.i.d. data, provided the following conditions hold: (i) the tuning parameter satisfies $\lambda = c_0(n^{-1}\log p)^{1/2}$; (ii) the eigenvalues of the covariance matrix are bounded such that $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$; (iii) the tail condition $P\big[n^{-1}\big|\sum_{i=1}^n(\varepsilon_{i,r}\varepsilon_{i,\ell} - \Sigma_{r\ell})\big| \geq \nu\big] \leq c_1\exp(-c_2 n\nu^2)$ holds for $|\nu| < \delta$.*

Remarks. (i) This lemma generalizes Theorem 1 of Rothman, Bickel, Levina, and Zhu (2008) from exact sparsity to approximate sparsity.[20] (ii) The convergence rate r_n is tied to the exponential-type tail condition, which holds for the normal distribution considered by Rothman, Bickel, Levina, and Zhu (2008). The estimator $\widehat{W}_G$ is consistent when $r_n = o(1)$, which holds under the condition $p \log p = o(n)$ if $s = O(p)$. The requirement is only slightly stronger than $p = o(n)$ under the exponential-tail condition. (iii) Ravikumar, Wainwright, Raskutti, and Yu (2011) establish consistency of $\widehat{W}_G$ under more general tail conditions, including polynomial-type tail conditions. Hayakawa (2024) suggests a GLS procedure for the structural parameter that relies only on the second moments instead of the fourth moments. Furthermore, the trimming method in Horowitz (1998) could be combined with the regularized estimator to mitigate the influence of outliers in heavy-tailed distributions.[21]

When the GLasso estimator is used to compute the fold k weighting matrix for cross-fitting, we denote it by $\widehat{W}_{G,-k}$, as it is computed with data from $\mathcal{I}_{-k}$. The resulting cross-fitted estimator is denoted by $\widehat{\theta}^*_G$, following the definition in (2.2) and (2.3) with $\widehat{W}_{-k}$ replaced by $\widehat{W}_{G,-k}$. With a finite number of cross-fitting folds K, $\widehat{W}_{G,-k}$ is a consistent estimator of W^O by Lemma 2.2. As a consequence, the cross-fitted estimator $\widehat{\theta}^*_G$ has the same asymptotic distribution as the oracle estimator based on W^O. We establish this result in Corollary 1 below. Algorithm 1 summarizes the steps to compute this cross-fitted estimator $\widehat{\theta}^*_G$ and the cross-fitted estimator of its asymptotic variance, denoted $\widehat{\Omega}^*_G$. For numerical results in Section 4 and Section 5, we use $K = 2$.

In addition to the GLasso estimator, many other types of regularized inverse covariance matrix estimators are available. These include Bickel and Levina (2008), Cai, Liu, and Luo (2011), and Fan, Liao, and Mincheva (2011), to name just a few. These alternative estimators also serve as proper weighting matrices for the minimum distance problem if the required sparsity condition for each method is satisfied by the empirical model. Therefore, the ideal choice could be model specific for an economic application. For example, Bickel and Levina (2008) consider a banding structure where the off-diagonal coefficients decay to 0 at certain rate as the moments become more distant from each other. Fan, Liao, and Mincheva (2011) consider a factor model structure in the data, which applies to many economic applications.

[20] Theorem 2 of Rothman, Bickel, Levina, and Zhu (2008) shows that under exact sparsity $\widehat{W}_G$ converges at the rate $\sqrt{n^{-1}(s+1)\log p}$ in the spectral norm. This rate is faster than r_n if s is much smaller than p, which does not hold in the examples we investigate. Therefore, we define approximate sparsity with the rate r_n, and under this approximate sparsity, $\widehat{W}_G$ converges at the rate r_n under both the Frobenius norm and the spectral norm.

[21] We implemented the proposed GLasso estimator with a pre-trimming step following the idea in Horowitz (1998). We chose the trimming parameter and sparsity penalty via a joint likelihood cross-validation. Results in Table A-6 in the Supplemental Appendix show improvements in bias, RMSE, and coverage for simulations analogous to those in Altonji and Segal (1994) and Horowitz (1998), where the data is drawn from the log-normal distribution, particularly for small sample sizes.

Algorithm 1: Cross-Fitted Estimator and Variance with GLasso Weighting

Data: $m_i \in \mathbb{R}^p$, i.i.d. for $i = 1, \ldots, n$;
Model: $f(\theta_0) = \mathbb{E}[m_i]$;
Result: estimator $\widehat{\theta}^*_G$ defined in (2.3) and its variance $\widehat{\Omega}^*_G$ defined in (2.5);
for $k = 1, \cdots, K$ **do**
  $\widehat{W}_{G,-k} \leftarrow$ compute with data $i \in \mathcal{I}_{-k}$, follow the GLasso estimator defined in (2.15) and (2.16). ; `/* use cross-validation to choose λ */`
  $\widehat{\theta}^{(k)}_G \leftarrow$ follow (2.2) with $\overline{m}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} m_i$ and $\widehat{W}_{-k} = \widehat{W}_{G,-k}$,
  $\widehat{\Omega}^{(k)}_G \leftarrow$ follow (2.5) with $\widehat{\theta}^{(k)} = \widehat{\theta}^{(k)}_G$ and $\widehat{W}_{-k} = \widehat{W}_{G,-k}$
end
$\widehat{\theta}^*_G \leftarrow K^{-1} \sum_{k=1}^K \widehat{\theta}^{(k)}_G$, $\quad \widehat{\Omega}^*_G \leftarrow K^{-1} \sum_{k=1}^K \widehat{\Omega}^{(k)}_G$.

Furthermore, the sparsity condition could be defined as near zero rather than exact zero, see Cai, Liu, and Luo (2011). The asymptotic theory on the minimum distance estimator in Section 3 does not distinguish between two regularized weighting matrix estimators with the same asymptotic limit, similar to that in a standard setup. Overall, data-dependent regularization of the weighting matrix is a versatile and effective approach to reduce the weighting matrix estimation errors and, in turn, improve the performance of minimum distance estimators.

3 Asymptotic Analysis with Many Moments

In this section, we derive the asymptotic distribution of the minimum distance estimator under $n \to \infty$, $p \to \infty$, and $p/n \to 0$. The dimension of the structural parameter θ, denoted by d_θ, is fixed and finite. We present the asymptotic distribution of a general cross-fitted minimum distance estimator $\widehat{\theta}^*$, defined in (2.3), with a convergent weighting matrix. A special case is the recommended estimator $\widehat{\theta}^*_G$ based on the GLasso weighting matrix. We also show consistency of the cross-fitted variance estimator defined in (2.5). We first present the asymptotic results in the canonical case, followed by an extension to cover a broader class of empirical applications. Let C and c denote some generic finite positive constants that bound some quantities from above and below. They do not have to take the same values when they appear in different places.

We first provide a generic high-level assumption on the weighting matrix $\widehat{W}$. Assumption W holds for the cross-fitted estimator as long as it holds for the full-sample estimator. The asymptotic theory below does not distinguish between two cross-fitted estimators with different weighting matrices that have the same asymptotic limit W.

Assumption W. (i) For some non-random matrix W, $||\widehat{W} - W|| \to_p 0$. (ii) $c \leq$

$\lambda_{\min}(W) \leq \lambda_{\max}(W) \leq C$.

To study the asymptotic distribution of the minimum distance estimator, we impose the following regularity conditions. We assume $f(\theta)$ is twice continuously differentiable. The first-order derivative is denoted by $f_\theta(\theta) \in \mathbb{R}^{p \times d_\theta}$ and we define $F = f_\theta(\theta_0)$. The second-order derivative with respect to θ_ℓ and θ_r is denoted by $f_{\theta\theta,r\ell}(\theta) \in \mathbb{R}^{p\times 1}$. Define $F_{\ell,\theta}(\theta) = (f_{\theta\theta,1\ell}(\theta), \ldots, f_{\theta\theta,d_\theta\ell}(\theta)) \in \mathbb{R}^{p\times d_\theta}$ and $F_{\ell,\theta} = F_{\ell,\theta}(\theta_0)$. The difference between the population moments and the model moments is denoted by $g(\theta) = \mathbb{E}[m_i] - f(\theta)$. We assume the parameter space Θ is compact and θ_0 is in the interior of the parameter space.

Assumption ID. There exists a unique true value $\theta_0 \in \Theta$ such that (i) $f(\theta_0) = \mathbb{E}[m_i]$. (ii) $\liminf_{n\to\infty} \inf_{\|\theta-\theta_0\|\geq\varepsilon} g(\theta)'Wg(\theta) > 0$.

Assumption R. The data are i.i.d., and $f(\theta)$ and Σ satisfy (i) $||f_\theta(\theta)|| \leq C$ for any $\theta \in \Theta$. (ii) $||F_{\ell,\theta}(\theta)|| \leq C$ for any $||\theta - \theta_0|| \leq \delta$ for some $\delta > 0$. (iii) $\lambda_{\min}(F'F) \geq c$. (iv) $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$. (v) $\mathbb{E}[m_{i,r}^{2+\varepsilon}] \leq C$ for $r = 1, \ldots, p$ for some $\varepsilon > 0$.

Assumptions ID and R ensure strong identification of θ_0.[22] Assumption R also assumes that the total identification information has an upper bound as the number of moments increases. Thus, some moments do not provide much information about θ_0. The procedure does not require us to know which moments are more informative.

Note that $\Omega = (F'WF)^{-1}F'W\Sigma WF(F'WF)^{-1}$ is the asymptotic covariance of the full-sample estimator in a standard fixed p asymptotic framework. Now we show that the cross-fitted estimator has the same asymptotic normal distribution in the many-moment framework $p \to \infty$ and $p/n \to 0$ under the stated assumptions. In particular, Assumption W only requires convergence in probability of the weighting matrix, putting no conditions on its rate of convergence. As discussed in Section 2.1, the full-sample estimator without cross-fitting, in contrast, could be asymptotically biased without stronger assumptions on the weighting matrix.

Theorem 3.1. *Suppose Assumptions ID, R, and W hold. Then,*

$$(\Omega)^{-1/2}\sqrt{n}\left(\widehat{\theta}^* - \theta_0\right) \to_d \mathcal{N}(0, I_{d_\theta}).$$

Assumption W holds for the cross-fitted GLasso weighted estimator with $W = W^O$ by Lemma 2.2, under the sparsity condition $sn^{-1}\log(p) \to 0$, i.e., the number of non-zero elements s estimated by the GLasso estimator is smaller than the sample

[22] Under fixed p asymptotics, Assumption ID(i), W(ii), compactness of Θ, and continuity of f imply ID(ii), see Newey and McFadden (1994). With large p asymptotics the infimum of $g(\theta)'Wg(\theta)$ given $||\theta - \theta_0|| \geq \varepsilon$ can vary with (p, n), and Assumption ID(ii) ensures that the model remains identified.

size n. For the cross-fitted GLasso weighted estimator $\widehat{\theta}^*_G$, the asymptotic variance is $\Omega^O = (F'W^OF)^{-1}$, the same as that obtained with the oracle weighting matrix.

Corollary 1. *Suppose Conditions (i)–(iii) of Lemma 2.2 and Assumptions ID and R hold. Then,*

$$(\Omega^O)^{-1/2}\sqrt{n}\left(\widehat{\theta}^*_G - \theta_0\right) \to_d \mathcal{N}(0, I_{d_\theta}).$$

Next, we show the cross-fitted variance estimator $\widehat{\Omega}^*$ in (2.5) is a consistent estimator of the asymptotic variance Ω. For this purpose, we need some additional regularity conditions.

Assumption V. (i) $\mathbb{E}[m_{i,r}^4] \leq C$ for $r = 1, \ldots, p$. (ii) $\lambda_{\max}(\widehat{\Sigma}_k) = O_p(1)$.

Assumption V(i) requires the fourth moments of all entries of m_i to be uniformly bounded. Assumption V(ii) is weaker than consistency of $\widehat{\Sigma}_k$. It holds even when $p/n \to c \in [0, 1]$ in the case where $m_i \sim \mathcal{N}(0, I_p)$, following Johnstone (2001).

Theorem 3.2. *Suppose Assumptions ID, R, W, and V hold. Then,*

$$(a). \quad ||\widehat{\Omega}^* - \Omega|| \to_p 0 \quad \textit{and} \quad (b). \quad (\widehat{\Omega}^*)^{-1/2}\sqrt{n}\left(\widehat{\theta}^* - \theta_0\right) \to_d \mathcal{N}(0, I_{d_\theta}). \tag{3.1}$$

Theorem 3.2 applies to $\widehat{\Omega}^*_G$ in Algorithm 1 for the cross-fitted GLasso-weighted estimator under the conditions in Lemma 2.2, because $\widehat{\Omega}^*_G$ is a special case of $\widehat{\Omega}^*$. Note that it differs from the covariance estimator based on the simplified formula $\Omega^O = (F'W^OF)^{-1}$, because the GLasso weighting matrix is different from the inverse sample covariance matrix.

In Appendix A, we present two extensions of the theoretical results to cover a wider range of applications. The first allows the identification strength to increase with the number of moments. The second allows the number of parameters to increase with the number of moments.

4 Simulation 1: Altonji and Segal (1996)

Altonji and Segal (1996) evaluate the finite-sample performance of minimum distance estimation in a balanced panel setting. We replicate and extend their simulation design, and use it to evaluate the performance of alternative weighting schemes in both the low-dimensional setting (p is fixed as n increases) and the high-dimensional setting (p and n increase simultaneously). In their experimental design, the objective is to estimate the population variance of a scalar random variable x based on observations collected from a panel of individuals, indexed by $i = 1, \ldots, n$, over T time periods.

Let $x_{i,t} \sim F$ be i.i.d. across i and t, where F is a probability distribution normalized to have mean zero and variance one. For each period, the sample variance $\widehat{\sigma}_t^2$ can be computed using the standard unbiased estimator. That is, $\widehat{\sigma}_t^2 = \frac{1}{n-1}\sum_{i=1}^n (x_{i,t} - \overline{x}_t)^2$, for $t \in \{1, \ldots, T\}$, where $\overline{x}_t = n^{-1}\sum_{i=1}^n x_{i,t}$ is the within-period sample average. By construction, $\widehat{\sigma}_t^2$ are i.i.d. across time. Altonji and Segal (1996) are interested in estimating the intra-period variance, a scalar $\theta = \text{Var}(x_{i,t})$, and it is straightforward to see that $\theta = \mathbb{E}[\widehat{\sigma}_t^2]$. The authors proceed by stacking the estimates of the second moments into a T-dimensional vector, $\overline{m}$. The estimation problem proceeds as in (2.1), with $f(\theta) = \theta \mathbb{1}_T$. Note that since the observations from all time periods are generated independently from the same distribution and each period has an equal number of observations, the model exhibits homoskedasticity. This model is a special case of Example 2. It matches the variance only and omits the covariance across time. In this case, the researcher has the knowledge that the covariance across time contains no information about the parameter of interest.

4.1 Comparison with Different Weighting Methods

We replicate the analysis of Altonji and Segal (1996) by considering nine different distributions for $x_{i,t}$,[23] which we recall are all scaled to have a zero mean and unit variance. Here, two alternative sample sizes, 100 and 1000, are considered, and in each case, we perform 1,000 Monte Carlo replications.

We consider four candidates. The first three, equally-weighted (EW), diagonally-weighted (DW), and optimally-weighted (OW) minimum distance estimators, are commonly used in practice. They are all computed with the full sample. Their confidence intervals are based on full-sample standard errors. The fourth candidate is our proposed cross-fitted GLasso-weighted (GW) minimum distance estimator. Its confidence interval is based on our proposed cross-fitted standard error.

In Table 1 we summarize the performance of our estimators across distributions and under different scenarios. Note that the EW estimator is optimal as it imposes the (correct) restriction that the estimated sample variances from different time periods provide equal and independent information. That is, the identity matrix is the oracle weighting matrix. This contrasts with the other estimators considered, which assign different weights to the sample variances from the different time periods. We are therefore interested in how these alternatives perform relative to the EW estimator.

First, the DW and OW estimators perform similarly, but both exhibit non-negligible negative bias. The bias is largest for the student-$t(5)$ distribution (which is thick-

[23] We consider the same set of distributions as in the original Altonji and Segal (1996) study: student-$t(5)$, student-$t(10)$, student-$t(15)$, normal, uniform, log-normal, exponential, half-normal, and bimodal. That latter is obtained as an equally-weighted mixture of two unit variance normally distributed random variables, with means -2 and 2.

Table 1: Altonji and Segal (1996) Design: Comparison of Weighting Schemes, $T = 10$

		Bias				RMSE				Coverage Prob.			
Distribution	n	EW	DW	OW	GW	EW	DW	OW	GW	EW	DW	OW	GW
t(5)	100	0.002	-0.123	-0.124	0.004	0.087	0.141	0.142	0.124	0.880	0.327	0.309	0.823
t(10)	100	0.001	-0.063	-0.062	0.003	0.055	0.085	0.086	0.074	0.900	0.571	0.554	0.855
t(15)	100	0.001	-0.051	-0.051	0.003	0.051	0.074	0.075	0.065	0.901	0.657	0.630	0.861
Normal	100	0.000	-0.036	-0.037	-0.000	0.043	0.058	0.059	0.052	0.894	0.747	0.724	0.884
Uniform	100	-0.001	-0.007	-0.007	-0.001	0.028	0.030	0.031	0.029	0.912	0.887	0.861	0.920
Log norm.	100	-0.001	-0.475	-0.482	-0.024	0.354	0.490	0.496	0.582	0.786	0.013	0.012	0.662
Exp	100	-0.003	-0.166	-0.168	-0.005	0.087	0.190	0.192	0.142	0.890	0.248	0.234	0.828
Half-norm.	100	0.001	-0.060	-0.061	0.002	0.051	0.082	0.083	0.069	0.913	0.606	0.576	0.880
Bimodal	100	0.000	-0.011	-0.011	0.000	0.027	0.030	0.031	0.029	0.901	0.843	0.819	0.894
t(5)	1000	-0.001	-0.027	-0.027	-0.001	0.026	0.036	0.037	0.031	0.899	0.622	0.627	0.873
t(10)	1000	-0.000	-0.008	-0.008	-0.001	0.017	0.019	0.019	0.018	0.903	0.840	0.845	0.900
t(15)	1000	-0.001	-0.006	-0.006	-0.001	0.016	0.017	0.017	0.016	0.898	0.851	0.855	0.892
Normal	1000	-0.001	-0.004	-0.004	-0.001	0.014	0.015	0.015	0.014	0.900	0.875	0.873	0.902
Uniform	1000	0.000	-0.000	-0.000	0.000	0.009	0.009	0.009	0.009	0.905	0.899	0.893	0.904
Log norm.	1000	-0.003	-0.164	-0.164	0.000	0.098	0.177	0.177	0.151	0.842	0.136	0.135	0.785
Exp	1000	0.000	-0.022	-0.022	0.000	0.029	0.037	0.037	0.033	0.880	0.725	0.721	0.865
Half-norm.	1000	-0.001	-0.007	-0.007	-0.001	0.017	0.019	0.019	0.018	0.900	0.848	0.847	0.885
Bimodal	1000	-0.000	-0.001	-0.001	-0.000	0.009	0.009	0.009	0.009	0.901	0.894	0.890	0.900

Notes: Average bias, root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative weighting schemes (equally-weighted, EW, diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW). EW is the oracle benchmark.

tailed and symmetric) and for log-normal and exponential distributions (longer-tailed and skewed). Root-mean-squared errors are also larger relative to EW, and the 90% coverage probabilities are typically far below 0.9, so inference is much less reliable in these cases. In contrast, our proposed GW estimator performs much better than both DW and OW. Importantly, the bias is much smaller and the coverage probabilities of the 90% confidence intervals are close to 0.9. As the sample size increases to $n = 1,000$, inference is generally improved for both DW and OW (although the coverage probability for some distributions is still well-below 0.9), but with the bias (while much reduced) often non-negligible. Compared to the EW estimator (the oracle benchmark), the GW estimator has similar root mean square errors with thin-tailed distributions. In cases where the DW or OW estimators perform poorly, the GW estimator also tends to have noticeably larger root mean square errors than the EW estimator. Across all distributions examined, the GW estimator generally shows comparable bias and coverage probabilities to the EW estimator and outperforms the DW and OW estimators significantly.

In Table 2 we extend the canonical experimental design by allowing the time dimension (and therefore the number of moments) to increase together with the cross-sectional dimension by setting $T = 0.2n$. In this setting, we again achieve broadly comparable performance between EW and GW estimators. However, we do note that the coverage probabilities for both DW and OW estimators are often very poor such that inference based on these estimators is particularly problematic in these settings.

It is important to emphasize that in this simulation design, the strong performance of GW relative to DW and OW is primarily achieved through the cross-fitting estimation procedure and the cross-fitted variance estimator.[24]

Table 2: Altonji and Segal (1996) Design: Comparison of Weighting Schemes, $T = 0.2n$

		Bias				RMSE				Coverage Prob.			
Distribution	n	EW	DW	OW	GW	EW	DW	OW	GW	EW	DW	OW	GW
t(5)	100	0.000	-0.130	-0.132	0.003	0.058	0.138	0.141	0.085	0.883	0.095	0.087	0.852
t(10)	100	0.001	-0.067	-0.067	0.003	0.038	0.078	0.080	0.051	0.900	0.388	0.353	0.860
t(15)	100	0.001	-0.054	-0.056	0.002	0.036	0.066	0.069	0.047	0.887	0.482	0.427	0.873
Normal	100	0.000	-0.038	-0.039	0.000	0.032	0.050	0.053	0.038	0.899	0.620	0.559	0.878
Uniform	100	-0.000	-0.006	-0.006	-0.000	0.020	0.021	0.023	0.020	0.920	0.886	0.802	0.927
Log norm.	100	0.005	-0.497	-0.511	-0.002	0.233	0.505	0.520	0.454	0.823	0.002	0.001	0.697
Exp	100	0.001	-0.173	-0.178	-0.001	0.064	0.187	0.193	0.113	0.879	0.077	0.068	0.824
Half-norm.	100	0.001	-0.063	-0.065	0.003	0.038	0.075	0.078	0.051	0.894	0.406	0.357	0.860
Bimodal	100	-0.001	-0.012	-0.012	-0.001	0.019	0.022	0.024	0.020	0.899	0.812	0.757	0.906
t(5)	1000	0.000	-0.028	-0.028	0.000	0.006	0.029	0.029	0.007	0.891	0.000	0.000	0.883
t(10)	1000	0.000	-0.009	-0.009	0.000	0.004	0.010	0.010	0.004	0.910	0.257	0.222	0.900
t(15)	1000	0.000	-0.006	-0.006	-0.000	0.003	0.007	0.008	0.004	0.906	0.431	0.375	0.899
Normal	1000	-0.000	-0.004	-0.004	0.000	0.003	0.005	0.005	0.003	0.897	0.655	0.578	0.902
Uniform	1000	-0.000	-0.001	-0.001	-0.000	0.002	0.002	0.002	0.002	0.911	0.886	0.801	0.910
Log norm.	1000	-0.001	-0.176	-0.180	-0.002	0.023	0.177	0.181	0.034	0.893	0.000	0.000	0.848
Exp	1000	0.000	-0.025	-0.025	0.000	0.006	0.026	0.026	0.007	0.901	0.014	0.011	0.884
Half-norm.	1000	0.000	-0.007	-0.007	0.000	0.004	0.008	0.008	0.004	0.903	0.387	0.357	0.898
Bimodal	1000	0.000	-0.001	-0.001	0.000	0.002	0.002	0.002	0.002	0.894	0.835	0.755	0.891

Notes: Average bias, the root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative weighting schemes (equally-weighted, EW, diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW). EW is the oracle benchmark.

4.2 Cross-Fitted Standard Error

As part of our procedure, in (2.5) we propose a cross-fitted standard error. It has been applied in calculating the 90% coverage probabilities for the GW estimator in Table 1. In their analysis, Altonji and Segal (1996) also consider a cross-fitting version of OW, which they refer to as independently-weighted optimal minimum distance (IWOMD). Relative to our simulation results, they report much lower (and often unfavorable) 90% coverage probabilities. In obtaining their coverage probabilities, Altonji and Segal (1996) use the full-sample standard error based on the usual (asymptotically equivalent) full-sample formula. We illustrate the importance of applying the cross-fitted standard error in Table 3.

For the OW estimator, we report the 90% coverage probability for three cases: (i) cross-fitting is not applied; (ii) cross-fitting is applied to obtain the estimator but

[24]Both EW and DW impose the correct sparsity structure, whereas under OW and GW the sparsity structure is estimated. As a consequence, a cross-fitted DW estimator is expected to outperform our GW estimator. In practice, the differences are small. Table A-2 in the Supplementary Appendix shows how cross-fitting affects the different weighting regimes, and how this varies with the number of folds. Relatedly, Table A-1 (also in the Supplementary Appendix) shows that GLasso successfully recovers the oracle sparsity structure.

Table 3: Altonji and Segal (1996) Design: Importance of Cross-Fitted Standard Error

		OW			GW		
Distribution	n	No CF	CF-Full	CF-CF	No CF	CF-Full	CF-CF
t(5)	100	0.309	0.547	0.834	0.324	0.576	0.823
t(10)	100	0.554	0.684	0.858	0.570	0.722	0.855
t(15)	100	0.630	0.727	0.858	0.656	0.760	0.861
Normal	100	0.724	0.778	0.888	0.746	0.819	0.884
Uniform	100	0.861	0.847	0.913	0.887	0.904	0.920
Log normal	100	0.012	0.162	0.665	0.013	0.188	0.662
Exp	100	0.234	0.504	0.835	0.248	0.548	0.828
Half-normal	100	0.576	0.674	0.870	0.606	0.722	0.880
Bimodal	100	0.819	0.823	0.896	0.844	0.860	0.894
t(5)	1000	0.627	0.777	0.872	0.622	0.775	0.873
t(10)	1000	0.845	0.864	0.894	0.840	0.877	0.900
t(15)	1000	0.855	0.870	0.887	0.852	0.881	0.892
Normal	1000	0.873	0.887	0.894	0.875	0.894	0.902
Uniform	1000	0.893	0.895	0.902	0.899	0.902	0.904
Log normal	1000	0.135	0.442	0.791	0.136	0.452	0.785
Exp	1000	0.721	0.810	0.867	0.725	0.812	0.865
Half-normal	1000	0.847	0.866	0.882	0.848	0.862	0.885
Bimodal	1000	0.890	0.891	0.902	0.893	0.896	0.900

Notes: Coverage probabilities of the 90% confidence intervals, for optimally-weighted (OW) and GLasso-weighted (GW) estimators, when cross-fitting (CF) is applied or not. CF-Full indicates that the cross-fitted estimator is coupled with the full-sample standard error, while CF-CF indicates the use of (2.3) for the estimator and (2.5) for the cross-fitted standard error. $T = 10$ for all cases.

not to calculate the standard error, as in Altonji and Segal (1996) for their IWOMD estimator; (iii) cross-fitting is applied to both the estimator and the standard error. Similarly, we also report these three cases for the GW estimator.

We first consider the three cases of the OW estimator when $n = 100$. The table shows that without cross-fitting, i.e., case (i), the coverage probabilities of the OW estimator are typically very poor (exactly as shown in Table 1 above). In case (ii), cross-fitting reduces the bias in the estimator (not shown), but the usual full-sample standard error still yields coverage probabilities well-below 0.9. This replicates the results for the IWOMD estimator in Altonji and Segal (1996). In case (iii), the coverage probabilities become comparable to that of the EW estimator, approaching 0.9, for the majority of distributions. The importance of applying cross-fitting in standard error calculation is also apparent when $n = 1000$.

Finally, the same three cases for the GW estimator demonstrate identical patterns. This further confirms that it is important to apply cross-fitting when using a GLasso weighting matrix and that the cross-fitted standard error is important for reliable inference.

4.3 Alternative Bias Correction Methods

We now compare our proposed GW estimator with alternative bias-correction methods from the literature. The first (HO) is the bootstrap estimator of Horowitz (1998),

Table 4: Altonji and Segal (1996) Design: Comparison of Alternative Estimators, $T = 10$

		Bias					RMSE				
Distribution	n	HO	NS	JK1	JK2	GW	HO	NS	JK1	JK2	GW
t(5)	100	0.069	-0.081	0.011	0.040	0.004	0.123	0.111	0.124	0.154	0.124
t(10)	100	0.075	-0.030	0.012	0.024	0.003	0.103	0.070	0.074	0.081	0.074
t(15)	100	0.073	-0.023	0.011	0.021	0.003	0.098	0.062	0.064	0.069	0.065
Normal	100	0.067	-0.014	0.010	0.016	-0.000	0.085	0.050	0.051	0.053	0.052
Uniform	100	0.042	-0.003	0.008	0.009	-0.001	0.053	0.030	0.031	0.032	0.029
Log normal	100	-0.193	-0.416	-0.023	0.240	-0.024	0.314	0.441	0.471	0.904	0.582
Exp	100	0.042	-0.105	0.005	0.055	-0.005	0.126	0.149	0.151	0.198	0.142
Half-normal	100	0.063	-0.023	0.012	0.025	0.002	0.093	0.065	0.068	0.075	0.069
Bimodal	100	0.041	-0.004	0.009	0.011	0.000	0.051	0.030	0.031	0.032	0.029
t(5)	1000	0.007	-0.013	-0.001	0.003	-0.001	0.027	0.029	0.031	0.033	0.031
t(10)	1000	0.007	-0.002	0.000	0.001	-0.001	0.019	0.018	0.018	0.018	0.018
t(15)	1000	0.006	-0.001	0.000	0.001	-0.001	0.017	0.016	0.016	0.016	0.016
Normal	1000	0.006	-0.001	0.000	0.001	-0.001	0.015	0.014	0.014	0.014	0.014
Uniform	1000	0.004	0.000	0.001	0.001	0.000	0.010	0.009	0.009	0.009	0.009
Log normal	1000	-0.002	-0.112	-0.004	0.035	0.000	0.100	0.136	0.146	0.186	0.151
Exp	1000	0.006	-0.004	0.001	0.004	0.000	0.030	0.031	0.032	0.032	0.033
Half-normal	1000	0.006	-0.000	0.000	0.001	-0.001	0.018	0.017	0.017	0.018	0.018
Bimodal	1000	0.004	-0.000	0.001	0.001	-0.000	0.010	0.009	0.009	0.009	0.009

Notes: Average bias and root-mean square error (RMSE), under alternative estimators: HO (Horowitz, 1998), NS (Newey and Smith, 2004), JK_1 and JK_2 (Kezdi, Hahn, and Solon, 2002), and GW (cross-fitted GLasso-weighted).

which incorporates an improved optimal weighting matrix via outlier trimming. The second (NS) applies the analytical bias-correction formula of Newey and Smith (2004). Finally, JK_1 and JK_2 are the jackknife MD estimators in Kezdi, Hahn, and Solon (2002). JK_1 is designed for linear models, while JK_2 covers general nonlinear cases.[25] When applied to the linear model in this simulation, JK_1 corresponds to the DML1 estimator and JK_2 to the DML2 estimator using OW instead of the GLasso, with the number of folds K set to n.[26]

In the low-dimensional case with fixed T, Table 4 shows that relative to these alternatives, the GW estimator has the lowest bias in most cases and comparable bias otherwise. In terms of RMSE, the estimators (including GW) are broadly comparable, with HO as an important exception: its performance is more sensitive to tail thickness due to trimming. In particular, HO has a noticeably smaller RMSE under the log-normal distribution, while JK_2 has a substantially larger RMSE. By contrast, for the

[25] In Kezdi, Hahn, and Solon (2002), the linear-model jackknife JK_1 (as in Altonji and Segal, 1996) is described on p. 37 and averages estimates across folds. The general/nonlinear jackknife JK_2 is described on p. 40 and averages first-order conditions. The difference between JK_2 and our estimator is discussed in Footnote 8.

[26] Our proposed GW estimator corresponds to a DML1 implementation with GLasso weighting with $K = 2$. Tables A-3 and A-4 in the Supplementary Appendix also report results with $K \in \{2, 5, 10, 20\}$, as well as a DML2 implementation. In small samples ($n = 100$), DML2's bias and RMSE can increase with the number of folds K for some distributions (especially log-normal), while DML1 shows no comparable trend. In larger samples the differences between DML1 and DML2 are much smaller.

normal and $t(15)$ cases, HO has a higher RMSE than the other estimators. NS also attains a small RMSE under the log-normal distribution, but it exhibits sizable bias. These differences are more pronounced at $n = 100$ than at $n = 1000$. Results from the high-dimensional case with many moments ($T = 0.2n$) are presented in Table A-5 in Section F.4 of the Supplementary Appendix, which also discusses the computational demands of the different approaches. These results are particularly encouraging for the GW estimator: it achieves the smallest bias in all scenarios. Furthermore, with a single exception (HO obtains a lower RMSE under the log-normal distribution at $n = 100$), GW achieves comparable or lower RMSE than all alternatives across all distributions and sample sizes.

5 Simulation 2: Baker and Solon (2003)

5.1 Model Description

To assess the performance of different weighting schemes in a richer empirical environment, we consider the study of Baker and Solon (2003), which examined the earnings dynamics of male workers in Canada between 1976 and 1992 using a panel dataset of yearly tax records.[27] The richness of their data allowed a flexible earnings process to be specified, whose estimated parameter values rejected a number of restrictions commonly imposed on the covariance structure (such as the absence of life-cycle variation in the variance of transitory income shocks). Here we propose a simulation study where the "true" parameters are the estimated parameters from their paper.[28] This is a good test of the performance of our estimators under a realistic model of earnings dynamics that exhibits a sparsity structure, which we now describe.

In their panel dataset, Baker and Solon (2003) identify $B = 19$ different two-year birth cohorts b, and we preserve this cohort grouping and the entrance year to the sample in our analysis (starting 1924–25 through to 1960–61). The log-earnings of individual i, in birth cohort b, at year t is specified as $Y_{ibt} = m_{bt} + y_{ibt}$, where m_{bt} is the mean log-earnings of birth cohort b in year t. They are interested in the evolution of the individual-specific deviation from this mean, y_{ibt}, which is parameterized as

$$y_{ibt} = p_t \times (\alpha_{ib} + \beta_{ib} z_{bt} + u_{ibt}) + \varepsilon_{ibt}, \tag{5.1}$$

where $z_{bt} = t - b - 26$ measures the potential labor market experience of cohort

[27] Ostrovsky (2010) extends the analysis of Baker and Solon (2003) using data from 1985 to 2005.

[28] All our analyses use the same unrounded estimates that Baker and Solon (2003) obtained in their analysis. We thank Michael Baker for sharing these with us. Rounded parameter estimates are presented in Table 4 ("Estimates of Earnings Dynamics Models") from their paper.

b at time t, p_t is a year-specific factor loading, α_{ib} is the time-invariant permanent component of earnings, and β_{ib} is the individual-specific growth rate in earnings. In the population, these heterogeneity parameters $(\alpha_{ib}, \beta_{ib})$ are normally distributed with mean zero and associated covariance parameters $(\sigma^2_\alpha, \sigma_{\alpha\beta}, \sigma^2_\beta)$. In addition, u_{ibt} is a random walk component driving permanent shocks to wages and ε_{ibt} is an $AR(1)$ process capturing transitory shocks

$$\begin{aligned} u_{ibt} &= u_{ib,t-1} + r_{ibt}, \\ \varepsilon_{ibt} &= \rho \varepsilon_{ib,t-1} + \lambda_t \nu_{ibt}, \end{aligned} \tag{5.2}$$

where λ_t is a year-specific fixed effect affecting the cross-sectional variance of transitory shocks in year t, and ρ is an auto-correlation parameter. The shocks r_{ibt} and ν_{ibt} are independent, normally distributed random variables with respective variances σ^2_r and $\text{Var}(\nu_{ibt})$. To capture potential variation in the variance of the transitory shocks over the life cycle, Baker and Solon (2003) allow $\text{Var}(\nu_{ibt})$ to depend on z_{bt} and specify a quadratic function

$$\text{Var}(\nu_{ibt}) = \gamma_0 + \gamma_1 z_{bt} + \gamma_2 z^2_{bt} + \gamma_3 z^3_{bt} + \gamma_4 z^4_{bt}. \tag{5.3}$$

The auto-regressive processes u_{ibt} and ε_{ibt} require an initial condition. For the random walk component $u_{ibt} = 0$ at age 26 (which is the age when individuals can first enter the sample), whereas $\varepsilon_{ibt} = \varepsilon^*_{ibt} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_b)$ in the first year that cohort b is observed in the sample. Note that the variance σ^2_b is cohort-specific. This captures the fact that they start at different ages when the sample begins.[29]

Recalling that there are 19 birth-cohort groups, with data from between 1976 and 1992, there are a total of 60 parameters to estimate with associated parameter vector

$$\theta = (\sigma^2_\alpha, \sigma^2_\beta, \sigma_{\alpha\beta}, \rho, \gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, p_{77}, \ldots, p_{92}, \sigma^2_{24-25}, \ldots, \sigma^2_{60-61}, \lambda_{78}, \ldots, \lambda_{92}). \tag{5.4}$$

5.2 Simulation and Estimation Design

We generate a synthetic dataset using the base model parameter estimates of Baker and Solon (2003). The observations are drawn from independent samples that are observed over different time periods, with cohort b comprising a sample of n_b individuals. We construct 19 different sample covariance matrices $\widehat{\text{Var}}(y_{ib1}, \ldots, y_{ibT_b})$, where T_b is the total number of time periods observed for each cohort. For each b, we extract the upper-triangular elements of $\widehat{\text{Var}}(y_{ib1}, \ldots, y_{ibT_b})$ and obtain the sample

[29] For individuals who do not enter the sample at age 26, u_{ibt} for their first appearance is drawn from a normal distribution with mean zero and variance $(t - b - 26)\sigma^2_r$, the distribution of a random walk that has been accumulating since age 26.

moment $\overline{m}_b$. For each cohort, the number of moments is $T_b \times (T_b + 1)/2$. Across all 19 birth cohorts, there are a total of 2,077 different moments. For each cohort b, this is exactly the same as the covariance structure model investigated in Example 2 with $x_{i,t} = y_{ibt}$ and $X_i = (y_{ib1}, \ldots, y_{ibT_b})'$. Given the model, the expectation of the moments for each cohort has a closed-form expression as a function of θ, which we denote by $f_b(\theta)$. We estimate the model using a minimum distance estimator with a cohort-specific weighting matrix $\widehat{W}_b$. To simplify comparisons, we assume that all cohorts have the same number of individuals, denoted by n_b.[30] The total number of individuals in the sample is $n = Bn_b$. Since the cohorts are independent with equal sizes, the minimum distance estimator minimizes the sum of criterion functions for each cohort as

$$\widehat{\theta} = \arg\min_{\theta} \sum_{b=1}^{B} (\overline{m}_b - f_b(\theta))' \widehat{W}_b (\overline{m}_b - f_b(\theta)). \tag{5.5}$$

As part of our experimental design, we perform 1,000 Monte Carlo replications and consider alternative birth cohort sizes (400, 800, 1200, and 2000). In Appendix F we show that the oracle weighting matrix is sparse.

5.3 Simulation Results

As in our analysis of the Altonji and Segal (1996) model in Section 4, we are interested in the performance of alternative weighting schemes. The Baker and Solon (2003) model comprises 60 parameters.[31]

Figure 5 visually summarizes the results for the 60 parameters using violin plots for each of the considered cohort sample sizes. Each violin plot acts as a smoothed, vertical histogram, mirrored to illustrate the density of the parameter-level performance measures. The widest sections represent the most frequent values, while the narrowest sections indicate the least common values across the parameter set.

The results across the alternative weighting schemes are summarized as follows. Absolute bias is highest under DW and EW (with DW exhibiting a notably longer tail than EW), followed by OW, and is substantially lower under GW. For coverage probabilities of the 90% confidence intervals, OW performs well below the nominal 0.9 level, while DW and EW show significant improvement for most parameters; GW consistently remains closest to 0.9. Finally, RMSE is typically highest under DW and EW, followed by OW, and is minimized under GW. Although the discrepancies between weighting regimes diminish as the sample size increases, our proposed GW

[30] In Figure A-2 of the Supplementary Appendix we rerun our simulations with the empirical cohort sizes from Baker and Solon (2003) and find qualitatively similar results.

[31] We provide parameter-level performance statistics for a cohort sample size of 400 in Table A-7 in the Supplementary Appendix.

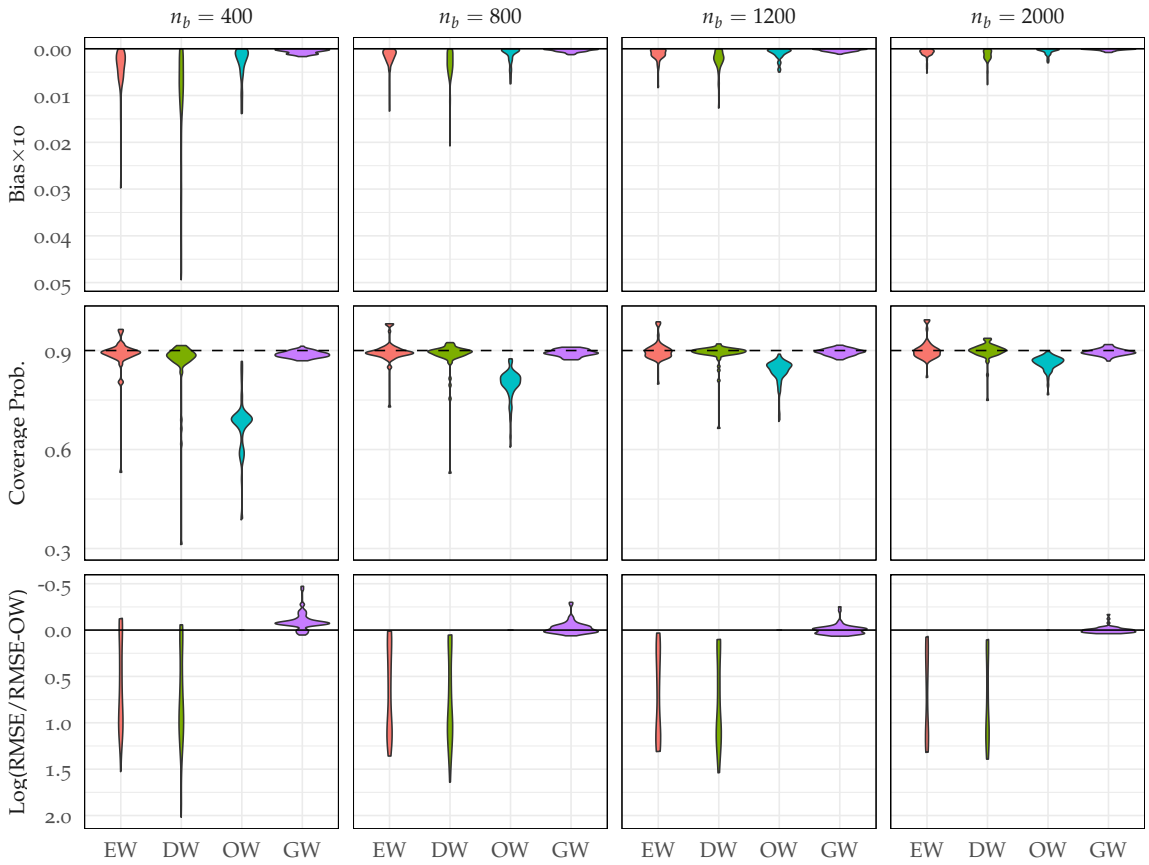


Figure 5: Violin plots for Baker and Solon (2003) parameters, showing absolute biases, 90% confidence interval coverage probabilities, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting. Figure derived from 1,000 replications with alternative cohort sample sizes. Weighting denoted EW (equally-weighted), DW (diagonally-weighted), OW (optimally-weighted), and GW (cross-fitted GLasso-weighted).

estimator dominates across the full spectrum of performance measures.[32]

Our results above were obtained by aggregating fold-specific estimates from $K = 2$ folds; this corresponds to a DML1 estimator. Figure 6 compares DML1 and DML2 estimators and examines the impact of the number of folds (for $n_b = 400$). The comparison reveals that DML1 estimators typically exhibit smaller finite-sample biases than their DML2 counterparts, while maintaining comparable coverage probabilities and RMSE. Consistent with the discussion in Section 2.1, increasing K from 2 to 5 increases bias and reduces RMSE for both the DML1 and DML2 estimators, although the quantitative magnitude of these effects is small.[33] There is little impact on cov-

[32] A caveat is that the parameters are correlated. To account for correlation across the 60 parameters, we computed an adjusted bias by normalizing the parameter vector by the inverse square root of its covariance matrix. This transformation ensures the normalized parameters are uncorrelated with unit variances. Figure A-1 in the Supplementary Appendix illustrates the results. Relative to the results here, the long tail observed under DW disappears, and we find that OW typically exhibits a higher absolute adjusted bias than DW and EW. Across all metrics, GW consistently demonstrates superior performance.

[33] This pattern reflects the sample-splitting effect: a larger K reduces the available sample size per fold (n_b/K), thereby exacerbating finite-sample bias stemming from moment nonlinearity. Conversely,

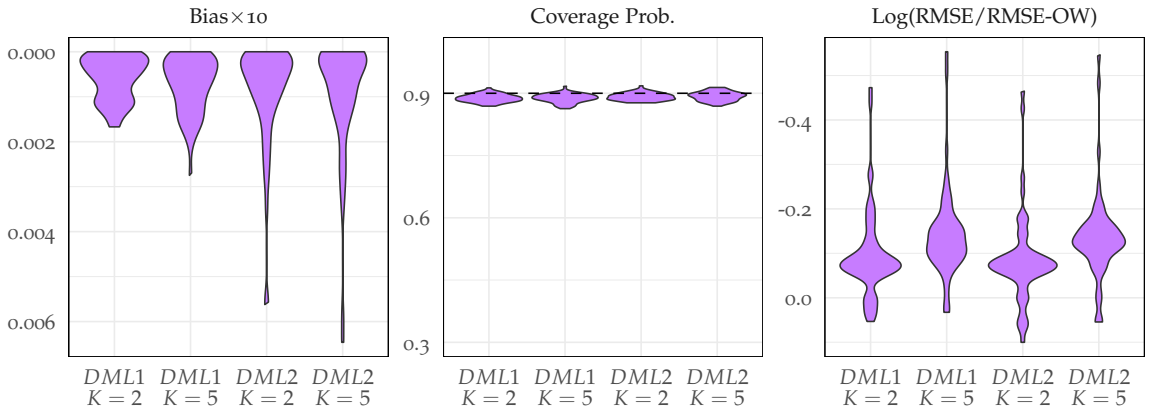


Figure 6: Violin plots for Baker and Solon (2003) parameters with cross-fitted GLasso-weighted estimation. In each panel the number of folds K is varied, and DML1 estimators are compared to DML2 estimators. Figure shows absolute biases, 90% confidence interval coverage probabilities, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting. Figure derived from 1,000 replications with a cohort sample size $n_b = 400$. The results for ($DML1$, $K = 2$) correspond to the GW results in Figure 5.

erage probabilities, which remain near nominal levels in all cases. While we adopt the DML1 estimator with $K = 2$ as our primary specification for this application, the differences here are much smaller than the impact of weighting schemes discussed above. Moreover, they do not have a quantitatively important effect on the variance decomposition that we study in the following section.

Figure 7 provides simulation evidence demonstrating that both cross-fitting and regularized estimation are critical in the context of our Baker and Solon (2003) study. In the absence of cross-fitting, OW and GW exhibit similar bias; however, GW performs slightly better in terms of coverage (though both remain well below the 0.9 target) and significantly better in terms of RMSE. Implementing cross-fitting dramatically shifts these results: coverage probabilities for all parameters approach 0.9, and while bias is mitigated for both estimators, the reduction is much larger for GW. Furthermore, while cross-fitting causes the RMSE of OW to deteriorate, the RMSE of GW remains broadly stable. Consequently, the cross-fitted GW estimator dominates the cross-fitted OW alternative. These patterns persist across larger cohort sizes, though the performance gap between them narrows as the sample size increases.

5.4 Variance Decomposition Analysis

We are interested in the extent to which the different estimates obtained under the alternative weighting schemes are consequential for economic outcomes. To this end, we replicate the decomposition exercise presented in Baker and Solon (2003), which uses the model structure to decompose the variance of log earnings into that due to

a larger K improves efficiency and reduces RMSE, as a larger proportion of the total sample is used to estimate the weighting matrix. The same qualitative findings are true with larger cohort sample sizes.

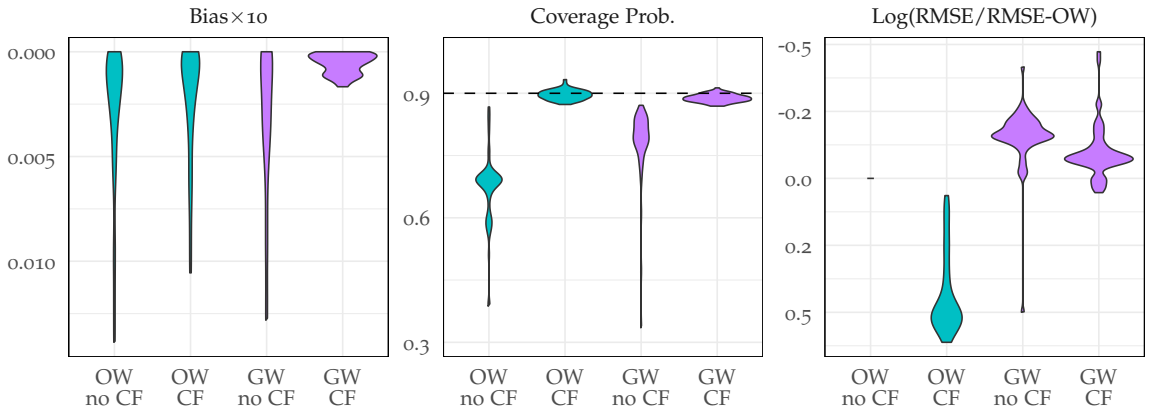


Figure 7: Violin plots for Baker and Solon (2003) parameters, showing absolute biases, coverage probabilities of the 90% confidence intervals, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting without cross-fitting. Figure derived from 1,000 replications with a cohort sample size $n_b = 400$. Weighting is denoted OW (optimally-weighted) and GW (GLasso-weighted), and results are shown by whether cross-fitting is applied (CF) or not (no CF).

the persistent and transitory components. As in Baker and Solon's (2003) analysis, we conduct this exercise with age fixed (at age 40) to abstract from any life-cycle considerations, with the variation over time induced by the changing factor loadings, as well as the initial variance for the transitory component up to age 40.

The results from this exercise when the cohort sample size is 400 are presented in Figure 8. The different panels correspond to the variance decomposition obtained when using the estimates from alternative weighting schemes. In each panel, the blue line shows the total variance of log earnings, while the red and blue lines respectively show the amount attributed to the persistent and transitory components. The shaded regions present the respective 90% pointwise confidence bands, defined as the area between the 5% and 95% quantiles of the estimates obtained with different simulated samples. The broken black lines indicate the true data-generating decomposition.[34] The figure shows that there is considerable bias under OW, with the amount of variation in log earnings systematically understated, with the true decomposition lines almost always outside of the respective confidence bands. In contrast, while DW much more closely matches the total amount of variation in log earnings, it attributes too little to the persistent component and too much to the transitory component. Note also that all the confidence bands are much wider relative to OW, especially for the early years of the analysis. Under EW the confidence bands are a similar size to those obtained under DW, while the bias (which is still present) is smaller in magnitude. Finally, we can see that GW performs exceptionally well: the predicted variance amounts (overall and by persistent/transitory status) almost perfectly coincide with that implied by the true data-generating process. Furthermore, the confidence bands

[34] By construction, the broken black lines are identical to those presented in Figure 3 from Baker and Solon (2003), which the interested reader should consult for a discussion of these inequality trends.

are considerably narrower than those obtained under both EW and DW.[35]

Figure 8: A decomposition of the variance of log earnings for males, 40 years old. The decomposition is constructed using 1,000 replications of the Baker and Solon (2003) model with a cohort sample size $n_b = 400$, under alternative weighting schemes (equally-weighted, EW, diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW). Shaded regions indicate the 90% pointwise confidence bands, defined as the area between the 5% and 95% quantiles of the estimates obtained with different simulated samples; broken black lines indicate the true data-generating decomposition.

[35]The same qualitative results are present under larger cohort sample sizes. As the sample size increases, the pointwise confidence bands are narrower in all cases, and the bias in the non-GW estimators is also reduced. While the difference across estimators is reduced, it is still the case that GW always performs the best. Full results are available upon request.

6 Estimation using the Panel Study of Income Dynamics

The empirical analysis of Baker and Solon (2003) uses longitudinal data from Revenue Canada's T-4 Supplementary tax file, which is not publicly available. To illustrate our method using accessible data, we use the Panel Study of Income Dynamics (PSID). Established in 1968, the PSID is one of the most widely used datasets for studying household income dynamics in the United States.

We construct an estimation dataset spanning the period 1970 to 2014, based on the sample in Moffitt and Zhang (2018).[36] We specify log earnings of individual i at period t as $Y_{it} = \alpha_i + z_{it}'\gamma + u_{it} + \varepsilon_{it}$. Here, α_i is the time-invariant permanent component of earnings, z_{it} contains demographic, age, and education controls, u_{it} follows a random walk, and ε_{it} is a first-order moving average process. That is

$$u_{it} = u_{i,t-1} + r_{it}$$
$$\varepsilon_{it} = \lambda v_{i,t-1} + v_{it}.$$

Our analysis focuses on residualized earnings, $y_{it} \equiv Y_{it} - z_{it}'\gamma$.[37] We construct a moment vector containing 187 elements, corresponding to the upper-triangular elements of the autocovariance matrix of residualized earnings.[38] The parameter vector θ consists of (i) the MA(1) coefficient λ, (ii) the time-varying transitory variances σ_{vt}^2, and (iii) the variance of the permanent component, $R_t \equiv \sigma_\alpha^2 + \sum_{t'=1}^{t} \sigma_{rt'}^2$, where σ_α^2 captures permanent individual heterogeneity (including the initial condition variance), and $\sigma_{rt'}^2$ are the variances of the period-specific innovations to the random walk process. By construction, the incremental random-walk variance satisfies $\sigma_{rt}^2 = R_t - R_{t-1}$.

Figure 9 reports the decomposition of the total variance of residualized earnings into its persistent and transitory components, using the minimum distance estimates under four alternative weighting schemes (EW, DW, OW, GW). The covariance matrix of sample moments is computed via bootstrap, and pointwise confidence bands are shown throughout. To facilitate comparisons, the GW decomposition is reproduced as dashed black lines in all other panels.

Our main findings are as follows. First, under EW, the overall variance of residualized earnings increased noticeably over the sample period. It increased strongly

[36] Moffitt and Zhang (2018) provide an extensive survey of the literature that has used the PSID to study income volatility. As in their analysis, we use PSID data from 1970 to 2014 to construct an unbalanced panel comprising male heads aged between 30 and 59 who are not students and have positive earnings and work hours. While the PSID was collected every year until 1997, thereafter it was only collected biannually. To accommodate this change in data structure parsimoniously, we use data every two years over the entire period.

[37] This specification is similar to that used by, e.g., Moffitt and Gottschalk (2002, 2012), who also use PSID data, but instead specify an $ARMA(1,1)$ for the transitory component. It is identical to that in, e.g., Blundell and Etheridge (2010), who uses British Household Panel Survey data.

[38] For sample size reasons, we only include autocovariances up to 10 periods (20 years) apart.

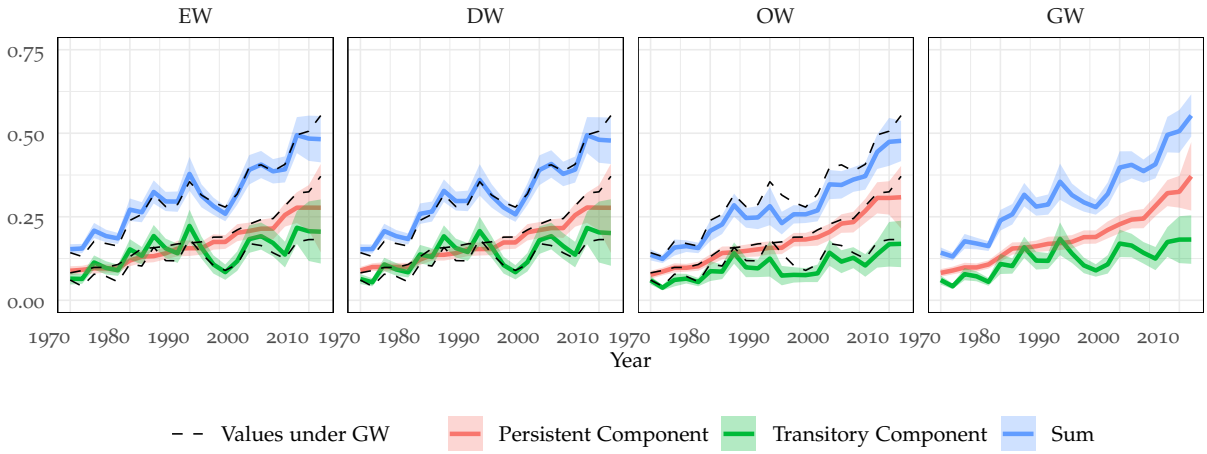


Figure 9: A decomposition of the variance of log earnings using PSID data under alternative weighting schemes (equally-weighted, EW, diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW). Shaded regions indicate the 90% pointwise confidence bands. The dashed black lines replicate the results under GW to aid comparison across specifications.

from the start of the series until the late 1980s, and again starting in the early 2000s. The overall increase in cross-sectional inequality is driven by increases in both the persistent and transitory variances. Second, DW yields near-identical estimates and confidence bands. Third, results differ somewhat under OW. The point estimates indicate a lower level of overall inequality throughout the sample period, primarily driven by lower estimated transitory variances. Finally, the GW estimates are generally very close to those under EW, except at the end of the series, where the overall inequality is higher, driven by a stronger growth in the persistent component. However, the most salient feature, which echoes the evidence from our Baker and Solon (2003) simulation study in Section 5, is that the confidence bands under GW are narrower than under EW and DW: over the entire series, the pointwise confidence bands of the persistent (transitory) component are 15.6% (7.2%) lower on average.

7 Conclusion

In their conclusion, Altonji and Segal (1996) highlight four desirable features of a future weighting matrix: (i) a robust weighting-matrix estimator that is superior to the conventional optimal weighting matrix and to the independently-weighted optimal weighting matrix; (ii) incorporation of prior information about which sets of moments are likely to be highly correlated, to reduce the effective dimension of weighting-matrix estimation; (iii) a transition between equal weighting and optimal weighting; and (iv) applicability to nonlinear models. Our proposed method provides a modern answer to each feature.[39] The regularized weighting matrix adapts to the data

[39]On feature (iii), our estimator transitions between diagonal weighting (rather than equal weighting) and optimal weighting.

by selecting which elements of the weighting matrix to estimate. Cross-fitting also substantially reduces estimation bias. Our asymptotic framework allows the number of moments to increase along with the sample size, ensuring the small-sample issue emphasized by Altonji and Segal (1996) remains relevant in a big-data environment. Using simulation designs based on earnings dynamics models, we show that an approach combining cross-fitting with regularized weighting matrix estimation performs extremely well relative to popular alternatives in the empirical literature.

References

Abowd, J. M., and D. Card (1989): "On the covariance structure of earnings and hours changes," *Econometrica*, 57(2), 411–445.

Altonji, J. G., and L. M. Segal (1994): "Small Sample Bias in GMM Estimation of Covariance Structures," Working Paper 3255, National Bureau of Economic Research.

——— (1996): "Small-sample bias in GMM estimation of covariance structures," *Journal of Business & Economic Statistics*, 14(3), 353–366.

Altonji, J. G., J. A. Smith, and I. Vidangos (2013): "Modeling Earnings Dynamics," *Econometrica*, 81(4), 1395–1454.

Andrews, D., and J. H. Stock (2007): "Testing with many weak instruments," *Journal of Econometrics*, 138(1), 24–46.

Angerer, X., and P.-s. Lam (2009): "Income Risk and Portfolio Choice: An Empirical Study," *The Journal of Finance*, 64(2), 1037–1055.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 14(1), 57–67.

Angrist, J. D., and A. B. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13(2), 225–235.

Autor, D., A. Kostøl, M. Mogstad, and B. Setzler (2019): "Disability Benefits, Consumption Insurance, and Household Labor Supply," *The American Economic Review*, 109(7), 2613–2654.

Baker, M., and G. Solon (2003): "Earnings dynamics and inequality among Canadian men, 1976–1992: Evidence from longitudinal income tax records," *Journal of Labor Economics*, 21(2), 289–321.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008): "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *J. Mach. Learn. Res.*, 9, 485–516.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–81.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018): "High-Dimensional Econometrics and Regularized GMM," Discussion paper.

Belloni, A., V. Chernozhukov, and C. Hansen (2013): "Inference for high-dimensional sparse econometric models," in *Advances in economics and econometrics: Tenth world congress*, vol. 3, pp. 245–95. Cambridge University Press Cambridge.

Bickel, P. J., and E. Levina (2008): "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36(1), 199 – 227.

Blundell, R., and B. Etheridge (2010): "Consumption, income and earnings inequality in Britain," *Review of Economic Dynamics*, 13(1), 76–102, Special issue: Cross-Sectional Facts for Macroeconomists.

Blundell, R., L. Pistaferri, and I. Preston (2008): "Consumption Inequality and Partial Insurance," *American Economic Review*, 98(5), 1887–1921.

Brent, R. P. (1973): *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, New Jersey, 1st edn.

Cai, T., W. Liu, and X. Luo (2011): "A Constrained l1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106(494), 594–607.

Carrasco, M., and M. Doukali (2017): "Efficient Estimation Using Regularized Jackknife IV Estimator," *Annals of Economics and Statistics*, (128), 109–149.

Carrasco, M., and A. Nayihouba (2022): "Regularized Estimation Of Dynamic Panel Models," *Econometric Theory*, pp. 1—59.

Chamberlain, G. (1984): "Chapter 22 Panel data," vol. 2 of *Handbook of Econometrics*, pp. 1247–1318. Elsevier.

Chao, J. C., and N. R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673–1692.

Cheng, X., and Z. Liao (2015): "Select the valid and relevant moments: An information-based LASSO for GMM with many moments," *Journal of Econometrics*, 186(2), 443–464, High Dimensional Problems in Econometrics.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

——— (2018b): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, pp. C1–C68.

Clark, T. E. (1996): "Small-Sample Properties of Estimators of Nonlinear Models of Covariance Structure," *Journal of Business & Economic Statistics*, 14(3), 367–373.

Fan, J., Y. Liao, and H. Liu (2016): "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, 19(1), C1–C32.

Fan, J., Y. Liao, and M. Mincheva (2011): "High-dimensional covariance matrix estimation in approximate factor models," *The Annals of Statistics*, 39(6), 3320–3356.

Friedman, J., T. Hastie, and R. Tibshirani (2008): "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9(3), 432–441.

Gourinchas, P.-O., and J. A. Parker (2002): "Consumption Over the Life Cycle," *Econometrica*, 70(1), 47–89.

Guvenen, F. (2007): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?," *American Economic Review*, 97(3), 687–712.

Han, C., and P. C. B. Phillips (2006): "GMM with Many Moment Conditions," *Econometrica*, 74(1), 147–192.

Hansen, C., and D. Kozbur (2014): "Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics*, 182(2), 290–308.

Hausman, J., R. Lewis, K. Menzel, and W. Newey (2011): "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, 165(1), 45–57, Moment Restriction-Based Econometric Methods.

Hayakawa, K. (2024): "Recent development of covariance structure analysis in eco-

nomics," *Econometrics and Statistics*, 29, 31–48.

Horowitz, J. L. (1998): "Bootstrap methods for covariance structures," *Journal of Human Resources*, pp. 39–61.

Hyslop, D. R. (2001): "Rising U.S. Earnings Inequality and Family Labor Supply: The Covariance Structure of Intrafamily Earnings," *The American Economic Review*, 91(4), 755–777.

Johnstone, I. M. (2001): "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, 29(2), 295 – 327.

Kezdi, G., J. Hahn, and G. Solon (2002): "Jackknife minimum distance estimation," *Economics Letters*, 76(1), 35–45.

MaCurdy, T. E. (1982): "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of Econometrics*, 18(1), 83–114.

Meghir, C., and L. Pistaferri (2004): "Income Variance Dynamics and Heterogeneity," *Econometrica*, 72(1), 1–32.

Mikusheva, A., and L. Sun (2021): "Inference with Many Weak Instruments," *The Review of Economic Studies*, 89(5), 2663–2686.

Moffitt, R., and S. Zhang (2018): "Income Volatility and the PSID: Past Research and New Results," *AEA Papers and Proceedings*, 108, 277–80.

Moffitt, R. A., and P. Gottschalk (2002): "Trends in the Transitory Variance of Earnings in the United States," *The Economic Journal*, 112(478), C68–C73.

——— (2012): "Trends in the Transitory Variance of Male Earnings: Methods and Evidence," *The Journal of Human Resources*, 47(1), 204–236.

Newey, W. K., and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," vol. 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

Newey, W. K., and R. J. Smith (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72(1), 219–255.

Newey, W. K., and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77(3), 687–719.

Ostrovsky, Y. (2010): "Long-Run Earnings Inequality and Earnings Instability among Canadian Men Revisited, 1985-2005," *The B.E. Journal of Economic Analysis & Policy*, 10(1).

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011): "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, 5(none), 935 – 980.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008): "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494 – 515.

Sustik, M. A., and B. Calderhead (2012): "GLASSOFAST : An efficient GLASSO implementation," Discussion paper, UTCS.

Velez, A. (2024): "On the Asymptotic Properties of Debiased Machine Learning Estimators," *arXiv preprint arXiv:2411.01864*.

Windmeijer, F. (2005): "A finite sample correction for the variance of linear efficient two-step GMM estimators," *Journal of Econometrics*, 126(1), 25–51.

Yuan, M., and Y. Lin (2007): "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94(1), 19–35.

Appendices

A Theoretical Extensions

In the following we use c, C, δ and ε to denote some generic positive constants. We first consider a simple extension to cover the case where the identification information of θ_0 increases with the number of moments. This extension covers the simulation example from Altonji and Segal (1996), which is also studied in Section 4. In this example, $f(\theta) = \mathbb{1}_T\theta$ for a scalar θ, where $\mathbb{1}_T$ is a T dimensional vector of $1's$. As such, $||f_\theta(\theta)|| = \sqrt{p}$ and $||f_{\theta\theta,r\ell}(\theta)|| = 0$.

We generalize Assumption ID and R to Assumption ID$^+$ and R$^+$ as follows. Let a_n be a sequence of constants that satisfies $a_n \to \infty$ and $a_n = O(p^{1/2})$. The constant a_n is $\sqrt{p}$ in this example. The key idea is to normalize derivatives associated with $g(\theta) = \mathbb{E}[m_i] - f(\theta) = f(\theta_0) - f(\theta)$ by a_n^{-1} so that all original assumptions hold with the normalized counterparts.

Define $f_\theta^+(\theta) = a_n^{-1} f_\theta(\theta)$, $F^+ = f_\theta^+(\theta_0)$, and $f_{\theta\theta,rl}(\theta)^+ = a_n^{-1} f_{\theta\theta,rl}(\theta)$.

Assumption ID$^+$. There exists a unique true value $\theta_0 \in \Theta$ such that (i) $f(\theta_0) = \mathbb{E}[m_i]$. (ii) $\liminf_{n\to\infty} \inf_{\|\theta-\theta_0\|\geq\varepsilon} a_n^{-2} g(\theta)' W g(\theta) > 0$.

To verify that Assumption ID$^+$ holds in the case where $f(\theta) = \mathbb{1}_T\theta$, note that $a_n^{-2} g(\theta)' W g(\theta) = (\theta - \theta_0)^2 \mathbb{1}_T' W \mathbb{1}_T / p \geq \varepsilon^2 \lambda_{\min}(W)$ for any $\|\theta - \theta_0\| \geq \varepsilon$ for any n.

Assumption R$^+$. Assumption R holds with $f_\theta(\theta)$, F, and $f_{\theta\theta,rl}(\theta)$ replaced by $f_\theta^+(\theta)$, F^+, and $f_{\theta\theta,rl}(\theta)^+$, respectively.

Theorem A.1. *Suppose Assumptions ID and R are replaced with Assumptions ID$^+$ and R$^+$ in Theorem 3.1, Corollary 1, and Theorem 3.2. Then, Theorem 3.1, Corollary 1, and Theorem 3.2(b) continue to hold. The rate of convergence of $\widehat{\theta}^*$ and $\widehat{\theta}^*_G$ is $\sqrt{n}a_n$.*

Theorem A.1 shows that the normalized statistic is self-corrected when the estimator $\widehat{\theta}^*$ has a different rate of convergence that depends on a_n. The generalization in Assumption R$^+$ considers the case where $||f_{\theta,r}(\theta)||$ diverges at the same rate a_n for different parameters θ_r for $r = 1, \ldots, d_\theta$. With mixed rates, we can generalize a_n to a $d_\theta \times d_\theta$ diagonal matrix A_n such that $f_\theta^+(\theta) = f_\theta(\theta) A_n^{-1}$. Furthermore, we could allow condition (iii) in Assumption R to accommodate $\lambda_{\min}(F'F)$ converging to 0 slowly such that a consistent estimator with a slower rate of convergence is obtained. Overall, this minimum distance estimation and inference framework is flexible enough to accommodate many identification scenarios relevant in empirical work.

Next, we consider a generalization where the dimension of the structural parameter θ, denoted by d_θ, increases with the sample size. In applications to earn-

ings dynamics, there is typically a time fixed effect whose dimension is the same as $T = O(\sqrt{p})$. To accommodate this time fixed effect in covariance structure models, we let d_θ grow with n and p. As shown below, asymptotic normality requires $d_\theta = o(n^{1/3})$, which holds for the covariance structure model when $p = o(n^{2/3})$.

For notational simplicity, we omit the dependence of θ and its parameter space Θ on n. We assume that Θ is compact for any n. Define $F(\theta) = f_\theta(\theta)$ and $F = F(\theta_0)$. Let $F_\ell(\theta) \in \mathbb{R}^{p\times 1}$ denote the ℓ^{th} column of $F(\theta) \in \mathbb{R}^{p\times d_\theta}$ for $\ell = 1, \ldots, d_\theta$.

Assumption R*. The data are i.i.d., and $f(\theta)$ and Σ satisfy (i) $||f(\theta)|| \leq C$ for any $\theta \in \Theta$. (ii) For any $\|\theta - \theta_0\| \leq \delta$, the first order derivative satisfies a Taylor expansion: $F_\ell(\theta) - F_\ell(\theta_0) = F_{\ell,\theta}(\theta - \theta_0) + O(\|\theta - \theta_0\|^2)$ for $\|F_{\ell,\theta}\| \leq C$, and the $O(\cdot)$ term holds uniformly over ℓ. (iii) $\lambda_{\min}(F(\theta)F'(\theta)) \geq c$ and $\lambda_{\max}(F(\theta)F'(\theta)) \leq C$ for any $\|\theta - \theta_0\| \leq \delta$. (iv) $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$. (v) $\mathbb{E}[m_{i,r}^{2+\varepsilon}] \leq C$ for $r = 1, \ldots, p$.

Theorem A.2. *Suppose Assumptions ID, R*, and W hold, and $d_\theta = o(n^{1/3})$. For any $\gamma \in R^{d_\theta}$ and $\|\gamma\| = 1$, we have*

$$\gamma'(\Omega)^{-1/2}\sqrt{n}\left(\widehat{\theta}^* - \theta_0\right) \to_d \mathcal{N}(0,1).$$

Because the dimension of θ increases with n, we use the local perturbation method to derive asymptotic normality. Cheng and Liao (2015) use this method to study GMM estimation with an increasing number of parameters and moments, in a setup that is different from the one in this paper and without weighting matrix estimation. Theorem A.2 could also be extended to accommodate situations where the identification information of θ_0 is increasing in the number of moments, or where $||F_\ell(\theta)||$ diverges at mixed rates for different parameters θ_ℓ for $\ell = 1, \ldots, d_\theta$.

B Proofs of Asymptotic Distributions

In this section, we provide proofs for the theoretical results in Section 3. We first present some auxiliary lemmas used in the proofs of the main results. Proofs of these auxiliary lemmas are collected at the end of this section.

Define $\overline{g}_k(\theta) = \overline{m}_k - f(\theta)$ and $g(\theta) = f(\theta_0) - f(\theta)$. Write the sample and population criterion function as $Q_{nk}(\theta) = \overline{g}_k(\theta)'\widehat{W}_{-k}\overline{g}_k(\theta)/2$ and $Q(\theta) = g(\theta)'Wg(\theta)/2$.

Lemma B.1. *We have the following results.*
(a). Under Assumption R, $\sup_{\theta\in\Theta} \|\overline{g}_k(\theta) - g(\theta)\|^2 = O_p(p/n)$.
(b). Under Assumption R, $\sup_{\theta\in\Theta} \|g(\theta)\| \leq C$, $\sup_{\theta\in\Theta} \|\overline{g}_k(\theta)\| = O_p(1)$.
(c). Under Assumption W, $||\widehat{W}_{-k} - W|| = o_p(1)$ and $||\widehat{W}_{-k}|| = O_p(1)$.
(d). Under Assumption ID, R, W, $||\widehat{F}_k - F|| = o_p(1)$ and $||\widehat{F}_k|| = O_p(1)$.

Lemma B.2. *Suppose Assumption ID, R, W hold. Then, $\widehat{\theta}^{(k)}$ is consistent.*

Lemma B.3. *Suppose Assumptions R and W hold and* $\widetilde{\theta}^{(k)} \to_p \theta_0$. *We have*

$$\frac{\partial^2}{\partial\theta\partial\theta'} Q_{nk}(\widetilde{\theta}^{(k)}) = F'WF + o_p(1).$$

Proof of Theorem 3.1. First, we show that $\widehat{\theta}^{(k)}$ follows the first-order approximation

$$\sqrt{n_k}(\widehat{\theta}^{(k)} - \theta_0) = -\left(F'WF\right)^{-1} \sqrt{n_k} F'W\overline{g}_k(\theta_0) + o_p(1). \tag{B.1}$$

By the mean-value expansion,

$$\sqrt{n_k}\left(\widehat{\theta}^{(k)} - \theta_0\right) = -\left[\frac{\partial^2}{\partial\theta\partial\theta'} Q_{nk}(\widetilde{\theta}^{(k)})\right]^{-1} \sqrt{n_k}\frac{\partial}{\partial\theta} Q_{nk}(\theta_0), \tag{B.2}$$

for some $\widetilde{\theta}^{(k)}$ between $\widehat{\theta}^{(k)}$ and θ_0 and thus $\widetilde{\theta}^{(k)} \to_p \theta_0$ by Lemma B.2. The second-order derivative in (B.2) converges in probability to $F'WF$ by Lemma B.3.

The first-order derivative satisfies

$$\begin{aligned} -\sqrt{n_k}\frac{\partial}{\partial\theta} Q_{nk}(\theta_0) &= \sqrt{n_k} F'\widehat{W}_{-k}\overline{g}_k(\theta_0) = A_k + B_k, \text{ where} \\ A_k &= \sqrt{n_k} F'W\overline{g}_k(\theta_0) \to_d \mathcal{N}(0, V), \\ B_k &= \sqrt{n_k} F'(\widehat{W}_{-k} - W)\overline{g}_k(\theta_0) = o_p(1). \end{aligned} \tag{B.3}$$

The first term $A_k \to_d \mathcal{N}(0, V)$ follows from a multivariate central limit theorem for i.i.d. random variables and $||V|| \leq ||F||^2||W||^2||\Sigma|| \leq C$ by Assumption W(ii), R(i), R(iv), and ID(i).

Below we show $B_k = o_p(1)$ under the condition $||\widehat{W}_{-k} - W|| \to_p 0$ in Assumption W(i). Consider the conditional expectation given data in $\mathcal{I}_{-k}$,

$$\begin{aligned} \mathbb{E}\left[\|B_k\|^2 |\mathcal{I}_{-k}\right] &= n_k \mathbb{E}[\overline{g}_k(\theta_0)'(\widehat{W}_{-k} - W)FF'(\widehat{W}_{-k} - W)\overline{g}_k(\theta_0)|\mathcal{I}_{-k}] \\ &= \text{tr}\left[n_k \mathbb{E}\left[\overline{g}_k(\theta_0)\overline{g}_k(\theta_0)'|\mathcal{I}_{-k}\right](\widehat{W}_{-k} - W)FF'(\widehat{W}_{-k} - W)\right] \\ &\leq d_\theta \left\|\Sigma(\widehat{W}_{-k} - W)FF'(\widehat{W}_{-k} - W)\right\| \\ &\leq C\left\|\widehat{W}_{-k} - W\right\|^2, \end{aligned} \tag{B.4}$$

where we use $n_k\mathbb{E}\left[\overline{g}_k(\theta_0)\overline{g}_k(\theta_0)'|\mathcal{I}_{-k}\right] = \Sigma$ under the independence between folds and Assumption R(i), R(iv). By Markov's inequality, for any given $\delta > 0$,

$$\Pr\left(|B_k| > \delta|\mathcal{I}_{-k}\right) \leq \frac{1}{\delta^2}\mathbb{E}\left[\|B_k\|^2 |\mathcal{I}_{-k}\right]. \tag{B.5}$$

Let $\mathcal{E} = \{||\widehat{W}_{-k} - W|| \leq \varepsilon\}$ for any $\varepsilon > 0$. Then by (B.4), (B.5), and the law of iterated expectations, we have $\Pr\left(|B_k| > \delta|\mathcal{E}\right) \leq \frac{C\varepsilon^2}{\delta^2}$.This shows $B_k = o_p(1)$ because $Pr(\mathcal{E}) \to 1$. This completes the proof for (B.1).

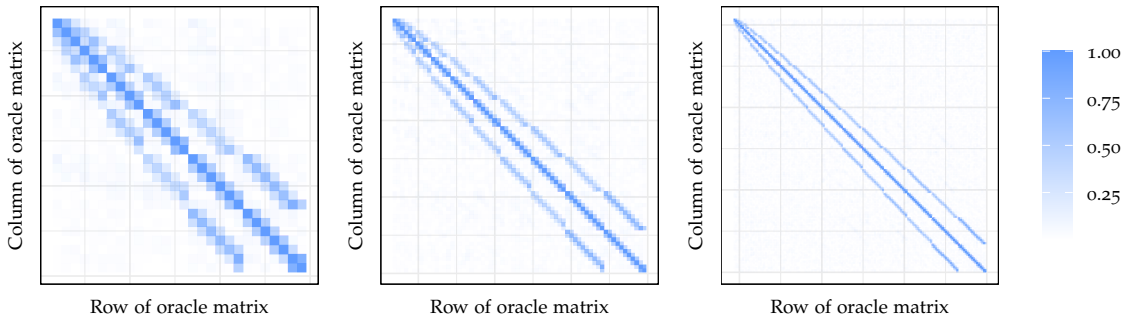
The cross-fitted estimator satisfies

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^{*}-\theta_0) &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_k}(\widehat{\theta}^{(k)}-\theta_0)\\ &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(F'WF\right)^{-1}\sqrt{n_k}F'W\overline{g}_k(\theta_0)+o_p(1)\\ &= \left(F'WF\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}F'W\left(m_i-\mathbb{E}[m_i]\right)+o_p(1),\end{aligned} \tag{B.6}$$

where the first equality follows from the definition of $\widehat{\theta}^{*}=K^{-1}\sum_{k=1}^{K}\widehat{\theta}^{(k)}$ and $n=n_kK$, the second equality uses (B.1), and the last equality holds because sample splitting implies $\sum_{k=1}^{K}n_k\overline{g}_k(\theta_0)=\sum_{i=1}^{n}\left(m_i-\mathbb{E}[m_i]\right)$. Note that $\xi_i=F'W\left(m_i-\mathbb{E}[m_i]\right)$ is a d_θ dimension random variable with mean zero and variance $V=F'W\Sigma WF$. The desired result follows from the multivariate central limit theorem for i.i.d. triangular array random variables and Slutsky's theorem. □

C Sparsity Structure in Baker and Solon (2003)

The oracle weighting matrix in the Baker and Solon (2003) model has a block structure, with the blocks corresponding to the independent birth cohorts. In Figure 10 we illustrate the sparsity structure generated by the model by plotting a normalized version of the oracle weighting matrix for three cohorts (1924–25, 1928–29, and 1934–35) evaluated at the true parameter vector θ_0. Because the model is estimated using data on a fixed number of calendar years, there are fewer moments for both the earlier and later birth cohorts in our sample. This is seen in panels (a) and (b) in Figure 10, where the corresponding heatmaps have lower resolution. In any case, the sparse structure is very evident, and this is true for all birth cohorts.



(a) Cohort 1924–25 (b) Cohort 1928–29 (c) Cohort 1934–35

Figure 10: Illustration of the sparsity pattern in the oracle weighting matrix in the Baker and Solon (2003) model for alternative birth cohorts. The heatmap indicates the absolute values of the oracle weighting matrix, which are normalized relative to the diagonal entries, evaluated at the data-generating parameter values.

Supplementary Appendix: How to Weight in Moment Matching: An ML Approach with Applications to Earnings Dynamics

Xu Cheng, Alejandro Sánchez-Becerra, Andrew Shephard

This Supplementary Appendix begins with Section D; Sections A–C are located in the Appendix of the main paper. Section D contains additional proofs of theoretical results. Section E discusses implementation details for the proposed estimators, and Section F collects all additional numerical results.

D Additional Proofs of Theoretical Results

D.1 Proofs of Asymptotic Distributions

Proof of Corollary 1. This corollary follows from Theorem 3.1 with $\widehat{W}$ and W replaced by $\widehat{W}_G$ and W^O, respectively. Assumption W follows from $||\widehat{W}_G - W^O|| \to_p 0$ by Lemma 2.2 and Assumption R(iv). □

Proof of Theorem 3.2. To show part (a), we first show $||\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\widehat{F}_k - F'W\Sigma WF|| = o_p(1)$. To this end, write

$$\begin{aligned}
&\left\|\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\widehat{F}_k - F'W\Sigma WF\right\| \leq H_1 + H_2, \text{ where} \\
H_1 &= \left\|F'W(\widehat{\Sigma}_k - \Sigma)WF\right\|, \\
H_2 &= \left\|F'\left(\widehat{W}_{-k} - W\right)\widehat{\Sigma}_k WF\right\| + \left\|F'\widehat{W}_{-k}\widehat{\Sigma}_k\left(\widehat{W}_{-k} - W\right)F\right\| + \\
&\quad \left\|\left(\widehat{F}_k - F\right)'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}F\right\| + \left\|\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\left(\widehat{F}_k - F\right)\right\|. \qquad \text{(D.1)}
\end{aligned}$$

We write H_1 and the multiple terms in H_2 separately to establish the result without requiring $||\widehat{\Sigma}_k - \Sigma_k|| \to_p 0$. Below we show both H_1 and H_2 are $o_p(1)$.

We start with the proof of $H_1 = o_p(1)$. Note that although $\widehat{\Sigma}_k$ is a $p \times p$ dimensional sample covariance matrix, $F'W\widehat{\Sigma}_k WF$ is only $d_\theta \times d_\theta$ dimensional. With $\varepsilon_i = m_i - \mathbb{E}(m_i)$, $\vartheta_i = \varepsilon_i\varepsilon_i' - \mathbb{E}[\varepsilon_i\varepsilon_i']$, and $\bar{\varepsilon}^{(k)} = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\varepsilon_i$, we have $\widehat{\Sigma}_k - \Sigma = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\vartheta_i - \bar{\varepsilon}^{(k)}\bar{\varepsilon}^{(k)'}$. Write

$$\begin{aligned}
F'W\left(\widehat{\Sigma}_k - \Sigma\right)WF &= R_1 - R_2, \text{ where} \\
R_1 = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} F'W\vartheta_i WF, &\quad R_2 = F'W\bar{\varepsilon}^{(k)}\bar{\varepsilon}^{(k)'}WF. \qquad \text{(D.2)}
\end{aligned}$$

To show $R_1 = o_p(1)$, we have

$$\sum_{r=1}^{d_\theta}\sum_{\ell=1}^{d_\theta}\mathbb{E}[(R_{1,r\ell})^2] = \frac{1}{n_k}\sum_{r=1}^{d_\theta}\sum_{\ell=1}^{d_\theta}\mathbb{E}\left[\left([F'W\vartheta_i WF]_{r\ell}\right)^2\right] = \frac{1}{n_k}\mathbb{E}[\mathrm{tr}(F'W\vartheta_i WFF'W\vartheta_i WF)]$$
$$\leq \frac{1}{n_k}||WFF'W||\,\mathbb{E}[\mathrm{tr}(F'W\vartheta_i\vartheta_i WF)]$$
$$\leq \frac{1}{n_k}d_\theta\,\|F\|^4\,\|W\|^4\left\|\mathbb{E}[\vartheta_i^2]\right\| \leq C\frac{p}{n}, \tag{D.3}$$

where the first equality holds because $R_{1,r\ell}$ is a sample average of the i.i.d. zero-mean random variable $[F'W\vartheta_i WF]_{r\ell}$, and the second equality follows from exchanging the order of $\mathbb{E}[\cdot]$ and summation, the first inequality holds because $A - ||A||\,I_p$ is negative semi-definite for a symmetric $p \times p$ dimensional matrix, the second inequality follows from exchanging the order of $\mathbb{E}[\cdot]$ and $\mathrm{tr}(\cdot)$, $\mathrm{tr}(A) \leq \mathrm{rank}(A)\lambda_{\max}(A)$, and $||AB|| \leq ||A|| \cdot ||B||$. The last inequality follows from Assumptions R(i), R(iv), W(ii), and $\left\|\mathbb{E}[\vartheta_i^2]\right\| \leq Cp$ because it is a $p \times p$ dimensional matrix with all elements uniformly bounded by Assumption V(i) and Hölder's inequality. Finally, $||R_1|| = o_p(1)$ follows from Markov's inequality and $p = o(n)$.

The remaining term R_2 satisfies $||R_2|| \leq ||F||^2||W||^2||\bar{\varepsilon}^{(k)}||^2 = o_p(1)$ by Assumption R(i), W(ii), and Lemma B.1(a). Combining it with $R_1 = o_p(1)$, we obtain $H_1 = o_p(1)$ by the triangle inequality.

To show $H_2 = o_p(1)$ is straightforward given Assumption R(i), V(ii), W(ii) and Lemma B.1(c) and (d). Using similar arguments, we have $\|\widehat{F}_k'\widehat{W}_{-k}\widehat{F}_k - F'WF\| \leq \|\widehat{F}_k\|^2\,\|\widehat{W}_{-k} - W\| + \|W\| \times \|\widehat{F}_k - F\|(\|\widehat{F}_k\| + \|F\|) = o_p(1)$ by Assumption R(i), W(ii) and Lemma B.1(c) and (d).

Because $F'WF$ is a non-singular $d_\theta \times d_\theta$ dimensional matrix by Assumption R(iii) and W(ii), we have $||\widehat{\Omega}^{(k)} - \Omega|| = o_p(1)$ by the continuous mapping theorem. This immediately gives the desired result.

Part (b) follows from Theorem 3.1, part(a), and the continuous mapping theorem given that $c \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq C$, which further follows from Assumption R(i), R(iii), R(iv) and W(ii). □

Proof of Lemma B.1. By definition, $\varepsilon_i = m_i - \mathbb{E}[m_i]$ and $\bar{\varepsilon}^{(k)} = n_k^{-1}\sum_{i\in\mathcal{I}_k}\varepsilon_i$.

$$\mathbb{E}\left[\|\bar{g}_k(\theta) - g(\theta)\|^2\right] = \mathbb{E}\left[\left\|\bar{\varepsilon}^{(k)}\right\|^2\right] = \frac{1}{n_k}\sum_{r=1}^{p}\mathbb{E}\left[\varepsilon_{i,r}^2\right] \leq C\frac{p}{n}, \tag{D.4}$$

where the inequality holds because $\mathbb{E}[\varepsilon_{i,r}^2] = \Sigma_{rr} \leq \lambda_{\max}(\Sigma) \leq C$ by Assumption R(iv). We obtain part (a) by Markov's inequality.

To prove part (b), note that

$$\sup_{\theta\in\Theta}\|g(\theta)\| \leq \sup_{\theta\in\Theta}\left\|f_\theta(\widetilde{\theta})\right\| \sup_{\theta\in\Theta}\|\theta-\theta_0\| \leq C \tag{D.5}$$

for some $\widetilde{\theta}\in\Theta$, where the last inequality follows from Assumption R(i) and the compactness of Θ. Combining it with part (a) and $p=o(n)$, we obtain $\sup_{\theta\in\Theta}\|\overline{g}_k(\theta)\| = O_p(1)$.

To prove part (c), note that $||\widehat{W}_{-k}-W|| = o_p(1)$ follows from Assumption W(i) directly because K is finite. By triangle inequality, $|\widehat{W}_{-k}|| \leq ||W|| + ||\widehat{W}_{-k}-W|| = O_p(1)$ by Assumption W(ii).

To prove part (d), we have

$$\left\|\widehat{F}_k - F\right\| = \left\|f_\theta(\widehat{\theta}_k) - f_\theta(\theta_0)\right\| \leq C\left\|\widehat{\theta}_k-\theta_0\right\| = o_p(1) \tag{D.6}$$

by Assumption R(ii) and the consistency of $\widehat{\theta}_k$ established in Lemma B.2. Then, $||\widehat{F}_k|| \leq ||F|| + o_p(1) = O_p(1)$ by Assumption R(i). □

Proof of Lemma B.2. We first show $\sup_{\theta\in\Theta}|Q_{nk}(\theta)-Q(\theta)| \to_p 0$. Note that

$$\begin{aligned} 2\sup_{\theta\in\Theta}|Q_{nk}(\theta)-Q(\theta)| &= \left|\overline{g}_k'\widehat{W}_{-k}\overline{g}_k(\theta) - g(\theta)'Wg(\theta)\right| \\ &\leq \left|(\overline{g}_k(\theta)+g(\theta))'\widehat{W}_{-k}(\overline{g}_k(\theta)-g(\theta))\right| + \left|g(\theta)'\left(\widehat{W}_{-k}-W\right)g(\theta)\right|, \end{aligned} \tag{D.7}$$

which converges to 0 in probability by Lemma B.1(a) – (c). By Assumption ID(ii), $\liminf_{n\to\infty}\inf_{||\theta-\theta_0||\geq\varepsilon} Q(\theta) > 0$ for any $\varepsilon > 0$. The desired result follows from standard arguments for the consistency of extremum estimators, see Newey and McFadden (1994). □

Proof of Lemma B.3 Row r and column ℓ of the left hand side is

$$\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_{nk}(\widetilde{\theta}^{(k)})\right]_{r\ell} = \left(\frac{\partial}{\partial\theta_r}f(\widetilde{\theta}^{(k)})\right)'\widehat{W}_{-k}\left(\frac{\partial}{\partial\theta_\ell}f(\widetilde{\theta}^{(k)})\right) - \frac{\partial^2}{\partial\theta_r\partial\theta_\ell}f(\widetilde{\theta}^{(k)})'\widehat{W}_{-k}\overline{g}_k(\widetilde{\theta}^{(k)}). \tag{D.8}$$

The second term on the right hand side of (D.8) is negligible because by stacking the d_θ rows together, we have

$$\left\|f_{\ell,\theta}(\widetilde{\theta})'\widehat{W}_{-k}\overline{g}_k(\widetilde{\theta}^{(k)})\right\| \leq C\left\|\widehat{W}_{-k}\right\|\left\|\overline{g}_k(\widetilde{\theta}^{(k)})\right\| = o_p(1) \tag{D.9}$$

by Assumption R(ii), and Lemma B.1(a), (c). The first term on the right hand side of

(D.8) satisfies

$$\left\| \frac{\partial}{\partial \theta_r} f(\widetilde{\theta}^{(k)}) - \frac{\partial}{\partial \theta_r} f(\theta_0) \right\| \leq C \left\| \widetilde{\theta}^{(k)} - \theta_0 \right\| = o_p(1) \tag{D.10}$$

by Assumption R(ii). This gives the desired results following Assumption W(ii) and Lemma B.1(c). □

D.2 Proofs of Generalization

Proof of Theorem A.1. We make the following adjustments to the previous results and proofs, in addition to replacing Assumption R with Assumption R^+.

(i). Lemma B.1. The proof of part (a) is unmodified because $\overline{g}_k(\theta) - g(\theta) = \overline{\varepsilon}^{(k)}$ regardless of the value of θ, and hence does not depend on the scaling of f. In part (b), C and $O_p(1)$ are replaced by a_n and $O_p(a_n)$, respectively. Part (c) is about convergence of the weighting matrix and does not depend on f. Part (d) holds with $\widehat{F}_k$ and F replaced by $\widehat{F}_k^+$ and F^+, respectively.

(ii). Lemma B.2. In the proof of consistency, we consider $\sup_{\theta \in \Theta} |Q_{nk}^+(\theta) - Q^+(\theta)| \to_p 0$, where $Q_{nk}^+(\theta)$ is defined similarly to $Q_{nk}(\theta)$ by replacing $\overline{g}_k(\theta)$ with $a_n^{-1}\overline{g}_k(\theta)$ and $Q^+(\theta) = a_n^{-2} g(\theta)' W g(\theta)$ is defined similarly to $Q(\theta)$ by replacing $g(\theta)$ with $a_n^{-1} g(\theta)$. The identification condition with $Q^+(\theta)$ is given in Assumption ID$^+$.

(iii). Theorem 3.1. The first-order expansion in (B.1) is replaced by

$$\sqrt{n_k} a_n (\widehat{\theta}^{(k)} - \theta_0) = -\left(F^{+\prime} W F^+\right)^{-1} \sqrt{n_k} F^{+\prime} W \overline{g}_k(\theta_0) + o_p(1). \tag{D.11}$$

To prove (D.11), (B.2) is replaced by

$$\sqrt{n_k} a_n \left(\widehat{\theta}^{(k)} - \theta_0\right) = -\left[a_n^{-2} \frac{\partial^2}{\partial \theta \partial \theta'} Q_{nk}(\widetilde{\theta}^{(k)})\right]^{-1} a_n^{-1} \frac{\partial}{\partial \theta} Q_{nk}(\theta_0), \tag{D.12}$$

where the modified second-order derivative continues to satisfy Lemma B.3 and the modified first-order derivative continues to satisfy (B.3), with F replaced by F^+. Following these adjustments, (B.6) becomes

$$\sqrt{n} a_n (\widehat{\theta}^* - \theta_0) = \left(F^{+\prime} W F^+\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F^{+\prime} W \left(m_i - \mathbb{E}[m_i]\right) + o_p(1). \tag{D.13}$$

Let $\Omega^+ = a_n^2 \Omega = (F^{+\prime} W F^+)^{-1} F^{+\prime} W \Sigma W F^+ (F^{+\prime} W F^+)^{-1}$. It is the counterpart of Ω with F replaced by F^+. We have

$$(\Omega)^{-1/2} \sqrt{n} (\widehat{\theta}^* - \theta_0) = (\Omega^+)^{-1/2} \sqrt{n} a_n (\widehat{\theta}^* - \theta_0) \to_d \mathcal{N}(0, I_{d_\theta}) \tag{D.14}$$

where the equality follows from the definition of Ω^+ and the convergence follows from (D.13). The claim on Corollary 1 follows immediately from that on Theorem 3.1.

(iv) Theorem 3.2. Let $\widehat{\Omega}^+ = a_n^2\widehat{\Omega}^*$. Then, $\widehat{\Omega}^+$ takes the same form as $\widehat{\Omega}^*$ except that $\widehat{F}_k$ is replaced by $a_n^{-1}\widehat{F}_k$. We have

$$(\widehat{\Omega}^*)^{-1/2}\sqrt{n}(\widehat{\theta}^* - \theta_0) = (\widehat{\Omega}^+)^{-1/2}\sqrt{n}a_n(\widehat{\theta}^* - \theta_0) \quad \to_d \mathcal{N}(0, I_{d_\theta}), \tag{D.15}$$

where the equality holds by the definition of $\widehat{\Omega}^+$ and the convergence follows from that of Theorem 3.2 with $\widehat{\Omega}^*$ replaced by $\widehat{\Omega}^+$ and F replaced by F^+. □

Next, we study the case where d_θ increases with n and p. Before proving Theorem A.2 on the asymptotic distribution, we first present some Lemmas.

Lemma D1. *Under Assumption R* and W,*
(a). $\|F'\widehat{W}_{-k}\overline{g}_k(\theta_0)\| = O_p(\tau_n)$, *where* $\tau_n = \sqrt{d_\theta/n}$.
(b). $\|F(\theta) - F(\theta_0)\| = O(\sqrt{d_\theta}\|\theta - \theta_0\|)$ *for* $\|\theta - \theta_0\| \leq \delta$.
(c). $\|(F(\theta) - F)'\widehat{W}_{-k}\overline{g}_k(\theta_0)\| = O_p(\sqrt{d_\theta}\tau_n\|\theta - \theta_0\|)$ *for* $\|\theta - \theta_0\| \leq \delta$.

Proof of Lemma D1. To prove part (a), we have

$$\begin{aligned}\mathbb{E}\left[\left\|F'\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\|^2 |\mathcal{I}_{-k}\right] &= \frac{1}{n_k}\mathbb{E}\left[\left\|F'\widehat{W}_{-k}\varepsilon_i\right\|^2\right] \\ &= \frac{1}{n_k}\operatorname{tr}\left[F'\widehat{W}_{-k}\Sigma\widehat{W}_{-k}F\right] \leq C\frac{d_\theta}{n}\end{aligned} \tag{D.16}$$

with probability approaching one, where the inequality follows from Assumptions W, R*(iii), R*(iv). By the law of iterated expectation and Markov's inequality, we immediately obtain the result in part (a).

To prove part (b), recall $F_\ell(\theta)$ and F_ℓ denote the ℓ^{th} column of $F(\theta)$ and $F = F(\theta_0)$, respectively. By Assumption R*(ii),

$$\|F_\ell(\theta) - F_\ell\| \leq \|\theta - \theta_0\|\,\|F_{\ell,\theta}\| + O(\|\theta - \theta_0\|^2) = O(\|\theta - \theta_0\|) \tag{D.17}$$

for $\|\theta - \theta_0\| \leq \delta$. Therefore,

$$\|F(\theta) - F(\theta_0)\|^2 \leq \sum_{\ell=1}^{d_\theta}\|F_\ell(\theta) - F_\ell(\theta_0)\|^2 = O(d_\theta\|\theta - \theta_0\|^2). \tag{D.18}$$

To part (c), the ℓ^{th} row of $(F(\theta)-F)'\widehat{W}_{-k}\overline{g}_k(\theta_0)$ satisfies

$$\begin{aligned}\left\|(F_\ell(\theta)-F_\ell)'\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\| &\leq \|\theta-\theta_0\|\left\|[F_{\ell,\theta}+O(\|\theta-\theta_0\|)]'\,\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\| \\ &= \|\theta-\theta_0\|\,O_p(\tau_n),\end{aligned} \tag{D.19}$$

where the first inequality follows from Assumption R*(ii), the second inequality follows from $F_{\ell,\theta}\widehat{W}_{-k}\overline{g}_k(\theta_0)=O_p(\tau_n)$, which follows from the same argument used to show part (a) of the Lemma, and the same arguments also hold when $F_{\ell,\theta}$ is replaced by $O(\|\theta-\theta_0\|)$ for $\|\theta-\theta_0\|\leq\delta$. The $O_p(\cdot)$ holds uniformly over ℓ. By the same argument used to show (D.18), we obtain the result in part (c). □

Lemma D2. *Suppose that $d_\theta^2=o(n)$. Under Assumptions ID, R*, and W, $\|\widehat{\theta}^{(k)}-\theta_0\|=O_p(\tau_n)$, where $\tau_n=\sqrt{d_\theta/n}$.*

Proof of Lemma D2. The consistency of the estimator follows from the same arguments as those for Lemma B.2 under Assumption ID and the uniform convergence of the sample criterion function, which holds under Lemma B.1 (a)-(c). Lemma B.1(a) and (c) follow the same arguments as in the original proof. Lemma B.1 (b) holds under Assumption R*(i) and Lemma B.1 (a). Note that

$$\overline{g}_k(\widehat{\theta}^{(k)})-\overline{g}_k(\theta_0)=g(\widehat{\theta}^{(k)})-g(\theta_0)=f(\theta_0)-f(\widehat{\theta}^{(k)}). \tag{D.20}$$

To derive the rate of convergence, we have

$$\begin{aligned}0\geq&\overline{g}_k(\widehat{\theta}^{(k)})'\widehat{W}_{-k}\overline{g}_k(\widehat{\theta}^{(k)})-\overline{g}_k(\theta_0)'\widehat{W}_{-k}\overline{g}_k(\theta_0)\\ =&\left[\overline{g}_k(\widehat{\theta}^{(k)})-\overline{g}_k(\theta_0)\right]'\widehat{W}_{-k}\left[\overline{g}_k(\widehat{\theta}^{(k)})-\overline{g}_k(\theta_0)\right]+2\left[\overline{g}_k(\widehat{\theta}^{(k)})-\overline{g}_k(\theta_0)\right]'\widehat{W}_{-k}\overline{g}_k(\theta_0)\\ =&(\widehat{\theta}^{(k)}-\theta_0)'F(\tilde{\theta})'\widehat{W}_{-k}F(\tilde{\theta})(\widehat{\theta}^{(k)}-\theta_0)-2(\widehat{\theta}^{(k)}-\theta_0)'F(\tilde{\theta})'\widehat{W}_{-k}\overline{g}_k(\theta_0)\\ \geq&\left\|\widehat{\theta}^{(k)}-\theta_0\right\|^2\lambda_{\min}(F(\tilde{\theta})'\widehat{W}_{-k}F(\tilde{\theta}))-2\left\|\widehat{\theta}^{(k)}-\theta_0\right\|\left\|F(\tilde{\theta})'\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\|\\ \geq&c\left\|\widehat{\theta}^{(k)}-\theta_0\right\|^2-2\left\|\widehat{\theta}^{(k)}-\theta_0\right\|\left(\left\|F'\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\|+\left\|(F(\tilde{\theta})-F)'\widehat{W}_{-k}\overline{g}_k(\theta_0)\right\|\right),\end{aligned} \tag{D.21}$$

with probability approaching one, where the second equality follows a mean-value expansion with $\tilde{\theta}$ between $\widehat{\theta}^{(k)}$ and θ_0 applied to (D.20), and the last inequality follows from consistency of the estimator, Assumption W, R*(iii), and triangle inequality.

Next, we consider the last two terms in the last line of (D.21). The first term is the object in Lemma D1(a), which is $O_p(\tau_n)$. Applying Lemma D1(c), we have $(F(\tilde{\theta})-F)'\widehat{W}_{-k}\overline{g}_k(\theta_0)=\|\widehat{\theta}^{(k)}-\theta_0\|O_p(\sqrt{d_\theta}\tau_n)$ because $\tilde{\theta}$ between $\widehat{\theta}^{(k)}$ and θ_0. Combining

them with (D.21), we have

$$\begin{aligned} 0 &\geq c\left\|\widehat{\theta}^{(k)}-\theta_0\right\|^2-\left\|\widehat{\theta}^{(k)}-\theta_0\right\|\left(O_p(\tau_n)+\left\|\widehat{\theta}^{(k)}-\theta_0\right\|O_p(\sqrt{d_\theta}\tau_n)\right) \\ &= c\left\|\widehat{\theta}^{(k)}-\theta_0\right\|\left[(1+o_p(1))\left\|\widehat{\theta}^{(k)}-\theta_0\right\|-O_p(\tau_n)\right], \end{aligned} \tag{D.22}$$

where the equality uses $d_\theta\tau_n^2 = d_\theta^2/n = o(1)$. This implies that $\|\widehat{\theta}^{(k)}-\theta_0\| = O_p(\tau_n)$. □

Proof of Theorem A.2. We will break down the proof into two parts. In the first part we prove that

$$\left|\gamma^{*\prime}(F'\widehat{W}_{-k}F)^{-1}F'\widehat{W}_{-k}[\sqrt{n_k}\overline{g}_k(\theta_0)-\sqrt{n_k}F(\widehat{\theta}^{(k)}-\theta_0)]\right| = o_p(1), \tag{D.23}$$

for a vector $\gamma^* \in \mathbb{R}^{d_\theta}$ such that $\|\gamma^*\| \leq C$. In the second part we build on this result to prove asymptotic normality.

Part 1: To prove (D.23) we apply a local perturbation approach. We start by introducing a set of auxiliary quantities. Let ε_n be a sequence of positive constants such that (i) $\varepsilon_n = o(n^{-1/2})$ and (ii) $\sqrt{d_\theta}\tau_n^2 = O(\varepsilon_n)$. Such a sequence can be constructed because$\sqrt{d_\theta}\tau_n^2 = o(n^{-1/2})$ under the condition $d_\theta = o(n^{1/3})$. Define a local perturbation from the estimator by ε_n:

$$\overline{\theta} = \widehat{\theta}^{(k)} + \varepsilon_n u_n^*, \tag{D.24}$$

where $u_n^* = (F'\widehat{W}_{-k}F)^{-1}\gamma^*$, for a vector $\gamma^* \in R^{d_\theta}$ and $\|\gamma^*\| \leq C$. By Assumption W and R*(iii), $\|u_n^*\| = O_p(1)$. Given $\varepsilon_n = o(n^{-1/2})$, we have $\varepsilon_n = O(n^{-1/2}) = O(\tau_n)$, where $\tau_n = \sqrt{d_\theta/n}$. Hence, $\|\varepsilon_n u_n^*\|^2 = \varepsilon_n^2\|u_n^*\|^2 = O_p(\varepsilon_n^2) = O_p(\tau_n^2)$. By Lemma D2, we have

$$\begin{aligned} \|\overline{\theta}-\widehat{\theta}^{(k)}\| &= \|\varepsilon_n u_n^*\| = O_p(\varepsilon_n), \\ \|\overline{\theta}-\theta_0\| &\leq \|\widehat{\theta}^{(k)}-\theta_0\| + \|\varepsilon_n u_n^*\| = O_p(\tau_n). \end{aligned} \tag{D.25}$$

Because the estimator minimizes the sample criterion function, we have

$$\begin{aligned} 0 \leq& \overline{g}_k(\overline{\theta})'\widehat{W}_{-k}\overline{g}_k(\overline{\theta}) - \overline{g}_k(\widehat{\theta}^{(k)})'\widehat{W}_{-k}\overline{g}_k(\widehat{\theta}^{(k)}) \\ =& \left[\overline{g}_k(\overline{\theta})-\overline{g}_n(\widehat{\theta}^{(k)})\right]'\widehat{W}_{-k}\left[\overline{g}_k(\overline{\theta})-\overline{g}_k(\widehat{\theta}^{(k)})\right] \\ &+ 2\left[\overline{g}_k(\overline{\theta})-\overline{g}_k(\widehat{\theta}^{(k)})\right]'\widehat{W}_{-k}\overline{g}_k(\widehat{\theta}^{(k)}). \end{aligned} \tag{D.26}$$

We now study the terms in (D.26).

First, note that in this MD problem, for any $\theta_1, \theta_2 \in \Theta$, we have

$$\overline{g}_k(\theta_1) - \overline{g}_k(\theta_2) = g(\theta_1) - g(\theta_2) = f(\theta_2) - f(\theta_1). \tag{D.27}$$

Therefore, we have

$$\begin{aligned}
\overline{g}_k(\overline{\theta}) - \overline{g}_n(\widehat{\theta}^{(k)}) =& g(\overline{\theta}) - g(\widehat{\theta}^{(k)}) = -F(\widetilde{\theta})\left(\overline{\theta} - \widehat{\theta}^{(k)}\right) \\
=& -F(\overline{\theta} - \widehat{\theta}^{(k)}) + J_1 \\
J_1 =& -\left[F(\widetilde{\theta}) - F\right](\overline{\theta} - \widehat{\theta}^{(k)}) = O_p(\sqrt{d_\theta}\tau_n\varepsilon_n),
\end{aligned} \tag{D.28}$$

where the first equality follows from a mean-value expansion and $\widetilde{\theta}$ is between $\overline{\theta}$ and $\widehat{\theta}^{(k)}$. The last equality follows from $\|\overline{\theta} - \widehat{\theta}^{(k)}\| = O_p(\varepsilon_n)$ and $||F(\widetilde{\theta}) - F|| = O_p(\sqrt{d_\theta}\tau_n)$, which in turn holds by Lemma D1(b) and $||\widetilde{\theta} - \theta_0|| = O_p(\tau_n)$. Consequently,

$$\left\|\overline{g}_k(\overline{\theta}) - \overline{g}_n(\widehat{\theta}^{(k)})\right\|^2 = \left\|g(\overline{\theta}) - g(\widehat{\theta}^{(k)})\right\|^2 = O_p(\varepsilon_n^2) + O_p(d_\theta\tau_n^2\varepsilon_n^2) = O_p(\varepsilon_n^2). \tag{D.29}$$

by Assumption R*(iii), $\|\overline{\theta} - \widehat{\theta}^{(k)}\| = O_p(\varepsilon_n)$, and by the theorem's assumption that $d_\theta^3/n = o(1)$, given that $d_\theta\tau_n^2 = d_\theta^2/n \leq d_\theta^3/n = o(1)$. By Assumption W and (D.29), we obtain

$$0 \leq \left[\overline{g}_k(\overline{\theta}) - \overline{g}_k(\widehat{\theta}^{(k)})\right]' \widehat{W}_{-k}\left[\overline{g}_k(\overline{\theta}) - \overline{g}_k(\widehat{\theta}^{(k)})\right] = O_p(\varepsilon_n^2) \tag{D.30}$$

Another term involved in (D.26) is $\overline{g}_k(\widehat{\theta}^{(k)})$. Using (D.27), we can write

$$\begin{aligned}
\overline{g}_k(\widehat{\theta}^{(k)}) =& \overline{g}_k(\theta_0) + g(\widehat{\theta}^{(k)}) - g(\theta_0) \\
=& \overline{g}_k(\theta_0) - F(\widehat{\theta}^{(k)} - \theta_0) + J_2 \\
J_2 =& -\left[F(\widetilde{\theta}^*) - F\right](\widehat{\theta}^{(k)} - \theta_0) = O_p(\sqrt{d_\theta}\tau_n^2),
\end{aligned} \tag{D.31}$$

where the second equality follows from a mean-value expansion with some $\widetilde{\theta}^*$ between $\widehat{\theta}^{(k)}$ and θ_0, and the last equality follows from Lemma D1(b) and $\|\widehat{\theta}^{(k)} - \theta_0\| = O_p(\tau_n)$.

Applying (D.28) and (D.31), the last term in (D.26) satisfies

$$\begin{aligned}
&\left[\overline{g}_k(\overline{\theta}) - \overline{g}_k(\widehat{\theta}^{(k)})\right]' \widehat{W}_{-k}\overline{g}_k(\widehat{\theta}^{(k)}) \\
=& \left[-F(\overline{\theta} - \widehat{\theta}^{(k)}) + J_1\right]' \widehat{W}_{-k}\left[\overline{g}_k(\theta_0) - F(\widehat{\theta}^{(k)} - \theta_0) + J_2\right] \\
=& -(\overline{\theta} - \widehat{\theta}^{(k)})'F'\widehat{W}_{-k}[\overline{g}_k(\theta_0) - F(\widehat{\theta}^{(k)} - \theta_0)] + D_1 + D_2 + D_3,
\end{aligned} \tag{D.32}$$

where

$$\begin{aligned} D_1 =& J_1' \widehat{W}_{-k} \left[\overline{g}_k(\theta_0) - F(\widehat{\theta}^{(k)} - \theta_0) \right], \\ D_2 =& -(\overline{\theta} - \widehat{\theta}^{(k)})' F' \widehat{W}_{-k} J_2, \\ D_3 =& J_1' \widehat{W}_{-k} J_2. \end{aligned} \tag{D.33}$$

Next we show

$$|D_1| = O_p(\varepsilon_n^2), \quad |D_2| = O_p(\varepsilon_n^2), \quad |D_3| = O_p(\varepsilon_n^2). \tag{D.34}$$

To study D_1, note that

$$\begin{aligned} |J_1' \widehat{W}_{-k} \overline{g}_k(\theta_0)| =& \left| (\overline{\theta} - \widehat{\theta}^{(k)})' \left[F(\widetilde{\theta}) - F(\theta_0) \right]' \widehat{W}_{-k} \overline{g}_k(\theta_0) \right| \\ =& O_p(\sqrt{d_\theta} \tau_n^2 \varepsilon_n) = O_p(\varepsilon_n^2) \end{aligned} \tag{D.35}$$

following from Lemma D1(c) and $\|\widetilde{\theta} - \theta_0\| = O_p(\tau_n)$, and the last equality holds because $\sqrt{d_\theta} \tau_n^2 = O(\varepsilon_n)$ by construction. Moreover,

$$|J_1' \widehat{W}_{-k} F(\widehat{\theta}^{(k)} - \theta_0)| \leq \|J_1\| O_p(\|\widehat{\theta}^{(k)} - \theta_0\|) = O_p(\sqrt{d_\theta} \tau_n^2 \varepsilon_n) = O_p(\varepsilon_n^2), \tag{D.36}$$

where the first equality follows from $J_1 = O_p(\sqrt{d_\theta} \tau_n \varepsilon_n)$ shown in (D.28), Assumption W, $\|\widehat{\theta}^{(k)} - \theta_0\| = O_p(\tau_n)$, and the second equality again uses $\sqrt{d_\theta} \tau_n^2 = O(\varepsilon_n)$. Together, the last two steps show $|D_1| = O_p(\varepsilon_n^2)$. By a similar argument to show (D.36), we have $D_2 = O_p(\varepsilon_n^2)$ by $J_2 = O_p(\sqrt{d_\theta} \tau_n^2)$ shown in (D.31), Assumption W, and $\|\overline{\theta} - \widehat{\theta}^{(k)}\| = O_p(\varepsilon_n)$. Finally, $|D_3| = O_p(\varepsilon_n^2)$ follows from $J_1 = O_p(\sqrt{d_\theta} \tau_n \varepsilon_n)$ shown in (D.28), $J_2 = O_p(\sqrt{d_\theta} \tau_n^2)$ shown in (D.31), Assumption W, and again $\sqrt{d_\theta} \tau_n^2 = O(\varepsilon_n)$.

Combining (D.26), (D.30), (D.32), and (D.34), we obtain

$$-O_p(\varepsilon_n^2) \leq -(\overline{\theta} - \widehat{\theta}^{(k)})' F' \widehat{W}_{-k} [\overline{g}_k(\theta_0) - F(\widehat{\theta}^{(k)} - \theta_0)]. \tag{D.37}$$

This is the main implication of (D.26). It show that some one-dimensional linear function of $F(\widehat{\theta}^{(k)} - \theta_0)$ can be approximated by the same function of the moments $\overline{g}_k(\theta_0)$, along the local perturbation given by $\overline{\theta} - \widehat{\theta}^{(k)} = \varepsilon_n u_n^*$, and the approximation error has a small lower bound.

Applying $\overline{\theta} - \widehat{\theta}^{(k)} = \varepsilon_n u_n^* = \varepsilon_n (F' \widehat{W}_{-k} F)^{-1} \gamma^*$ and $\varepsilon_n = o(n_k^{-1/2})$, which holds because $\varepsilon_n = o(n^{-1/2})$ by construction and K is finite, (D.37) leads to

$$-\gamma^{*\prime} (F' \widehat{W}_{-k} F)^{-1} F' \widehat{W}_{-k} [\sqrt{n_k} \overline{g}_k(\theta_0) - \sqrt{n_k} F(\widehat{\theta}^{(k)} - \theta_0)] \geq -o_p(1) \tag{D.38}$$

where $\gamma^* \in R^{d_\theta}$ with $\|\gamma^*\| \leq C$.

Next, define $\overline{\theta} = \widehat{\theta}^{(k)} - \varepsilon_n u_n^*$ and using the same arguments in deriving (D.38), we deduce that

$$-\gamma^{*\prime}(F'\widehat{W}_{-k}F)^{-1}F'\widehat{W}_{-k}[\sqrt{n_k}\overline{g}_k(\theta_0) - \sqrt{n_k}F(\widehat{\theta}^{(k)} - \theta_0)] \leq o_p(1), \tag{D.39}$$

which combined with (D.38), implies that

$$\left|\gamma^{*\prime}(F'\widehat{W}_{-k}F)^{-1}F'\widehat{W}_{-k}[\sqrt{n_k}\overline{g}_k(\theta_0) - \sqrt{n_k}F(\widehat{\theta}^{(k)} - \theta_0)]\right| = o_p(1). \tag{D.40}$$

Part 2: Let $\gamma^{*\prime} = \gamma'\Omega^{-\frac{1}{2}}$, where $\gamma \in R^{d_\theta}$ be an arbitrary vector with $\|\gamma\| = 1$. We have $\|\gamma^*\|^2 = \gamma'\Omega^{-1}\gamma \leq C$ by $\|\gamma\| = 1$ and Assumptions W, R*(iii), R*(iv). The approximation in (D.40) can be rewritten as

$$\begin{aligned}\sqrt{n_k}\gamma'\Omega^{-\frac{1}{2}}(\widehat{\theta}^{(k)} - \theta_0) =& \gamma'\Omega^{-\frac{1}{2}}(F'\widehat{W}_{-k}F)^{-1}F'\widehat{W}_{-k}\Sigma^{\frac{1}{2}}\left[\sqrt{n_k}\Sigma^{-\frac{1}{2}}\overline{g}_k(\theta_o)\right] + o_p(1)\\ =& \overline{\phi}'\left[\sqrt{n_k}\Sigma^{-\frac{1}{2}}\overline{g}_k(\theta_0)\right] + o_p(1),\end{aligned} \tag{D.41}$$

where $\overline{\phi}' = \gamma'\Omega^{-\frac{1}{2}}(F'\widehat{W}_{-k}F)^{-1}F'\widehat{W}_{-k}\Sigma^{\frac{1}{2}}$ by definition.

Let $\phi' = \gamma'\Omega^{-\frac{1}{2}}(F'WF)^{-1}F'W\Sigma^{\frac{1}{2}}$. By construction, $\|\phi'\|^2 = \gamma'\Omega^{-\frac{1}{2}}\Omega\Omega^{-\frac{1}{2}}\gamma = 1$. Following the linear approximation in (D.41),

$$\sqrt{n_k}\gamma'\Omega^{-\frac{1}{2}}(\widehat{\theta}^{(k)} - \theta_0) = \mathcal{A}_k + \mathcal{B}_k + o_p(1), \tag{D.42}$$

where

$$\begin{aligned}\mathcal{A}_k &= \phi'\left[\sqrt{n_k}\Sigma^{-\frac{1}{2}}\overline{g}_k(\theta_0)\right] \to_d \mathcal{N}(0,1),\\ \mathcal{B}_k &= (\overline{\phi} - \phi)'\left[\sqrt{n_k}\Sigma^{-\frac{1}{2}}\overline{g}_k(\theta_0)\right] = o_p(1).\end{aligned} \tag{D.43}$$

The convergence for $\mathcal{A}_k$ follows from a triangular array central limit theorem. Note that $\mathcal{B}_k$ has a similar structure to the B_k analyzed in (B.4). Showing $\mathcal{B}_k = o_p(1)$ follows from arguments similar to those used to show $B_k = o_p(1)$ in the case when d_θ is finite, taking into account that $\mathcal{B}_k$ is a scalar and that $\|\gamma\| = 1$.

Following the definition of the estimator $\widehat{\theta}^*$ and $n = Kn_k$, we have

$$\begin{aligned}\sqrt{n}\gamma'\Omega^{-\frac{1}{2}}(\widehat{\theta}^* - \theta_0) = \sqrt{n}\frac{1}{K}\sum_{k=1}^{K}\gamma'\Omega^{-\frac{1}{2}}(\widehat{\theta}^{(k)} - \theta_0) =& \frac{1}{\sqrt{K}}\sum_{k=1}^{K}(\mathcal{A}_k + \mathcal{B}_k) + o_p(1)\\ \to_d \mathcal{N}(0,1),&\end{aligned} \tag{D.44}$$

where the second equality follows from (D.42) and the convergence follows from (D.43), $\mathcal{A}_k$ is independent across k, and K is finite. □

D.3 Proofs of Approximate Sparsity

Proof of Lemma 2.1: Because $s < 2pk$ and

$$\|R\|_F^2 = \sum_{j=1}^{p-k-1} (p-k-j)\theta^{2(k+j)} < p\frac{\theta^2}{1-\theta^2}\theta^{2k}, \tag{D.45}$$

it is sufficient to show that

$$p\frac{\theta^2}{1-\theta^2}\theta^{2k} = O\left(\frac{(p+2pk)\log p}{n}\right) = o(1). \tag{D.46}$$

The $o(1)$ term holds under the condition $\frac{pk\log p}{n} = o(1)$. Since $\theta^2 < 1$, it is sufficient to show $p\theta^{2k} \leq C\frac{pk\log p}{n}$ for some C for n and k large enough. Taking log's on both sides and rearranging, it is equivalent to showing that

$$\log(n) \leq \log(C) + \log(k) + \log(\log p) - 2k\log(\theta), \tag{D.47}$$

where $\log(\theta) < 0$. For $C = 1$, this inequality holds for large n and k as long as $\frac{\log n}{k} = o(1)$. □

Proof of Lemma 2.2: This lemma generalizes the proof of Theorem 1 of Rothman, Bickel, Levina, and Zhu (2008) from models with the exact sparsity condition to the approximate sparsity condition. We first show that $\|\widehat{W} - W^O\|_F = O_p(r_n)$ under approximate sparsity, where the estimator $\widehat{W}$ minimizes the penalized likelihood $\ell(W) = \text{tr}(W\widehat{\Sigma}) - \log|W| + \lambda|W^-|_1$, and by definition $M^- = M - \text{diag}(M)$ for any matrix M. This criterion function is based on the sample covariance matrix $\widehat{\Sigma}$. Once this result is established, we show that it also holds for the correlation-based estimator $\widehat{W}_G$ that we adopt in (2.16).

Under the approximate sparsity condition, we have $W^O = W^* + R$, where $W^O = \Sigma^{-1}$ is the oracle weighting matrix, W^* is the symmetric sparse approximation, and R is the small approximation error whose Frobenius norm satisfies (2.12). To show $\|\widehat{W} - W^O\|_F = O_p(r_n)$, it is sufficient to show $\|\widehat{W} - W^*\|_F = O_p(r_n)$ by the triangle

inequality. Let

$$\begin{aligned} Q(W) &= \operatorname{tr}(W\widehat{\Sigma}) - \log|W| + \lambda|W^-|_1 - \operatorname{tr}(W^*\widehat{\Sigma}) + \log|W^*| - \lambda|W^{*-}|_1 \\ &= \operatorname{tr}\left[(W - W^*)\widehat{\Sigma}\right] - (\log|W| - \log|W^*|) + \lambda\left(|W^-|_1 - |W^{*-}|_1\right), \end{aligned} \tag{D.48}$$

where $|a|_1$ denotes the L_1 norm of a vector a. Let $\Delta = W - W^*$. Our estimate $\widehat{W}$ minimizes $Q(W)$, or equivalently $\widehat{\Delta} = \widehat{W} - W^*$ minimizes $G(\Delta) := Q(W^* + \Delta)$.

The main idea of the proof in Rothman, Bickel, Levina, and Zhu (2008) is as follows. Consider the set $\Theta_n(C) = \{\Delta : \Delta = \Delta^T, \|\Delta\|_F = Cr_n\}$ for some $C > 0$. Note that $G(\Delta) = Q(W^* + \Delta)$ is a convex function, and $G(\widehat{\Delta}) \leq G(0) = 0$. If we can show that

$$\inf\{G(\Delta) : \Delta \in \Theta_n(C)\} > 0, \tag{D.49}$$

the minimizer $\hat{\Delta}$ must be inside the sphere defined by $\Theta_n(C)$, and hence $\|\widehat{\Delta}\|_F \leq Cr_n$. To see why, note that if instead $||\widehat{\Delta}||_F > Cr_n$, then for some $\alpha \in (0,1)$, $||\alpha\widehat{\Delta}||_F = Cr_n$. This would imply that $\alpha\widehat{\Delta} \in \Theta_n(C)$ and $G(\alpha\widehat{\Delta}) > 0$ by (D.49), but because G is convex, $G(\alpha\widehat{\Delta}) = G(\alpha\widehat{\Delta} + (1-\alpha)0) \leq \alpha G(\widehat{\Delta}) + (1-\alpha)G(0) \leq 0$, a contradiction.

Compared to the proof in Rothman, Bickel, Levina, and Zhu (2008), our analysis defines the expansion in (D.48) around the sparse matrix W^* rather than the true matrix W^O. This facilitates the exploration of the sparse structure via the penalization term $|W^{*-}|_1$. For the rest of the proof, we focus on the steps where we make adjustments to account for the approximation error R.

The expansion above is around the sparse matrix W^*. We next rewrite the log-determinant term in a way that facilitates an expansion around W^O, the true precision matrix, for which we have $[W^O]^{-1} = \Sigma$. We have

$$\log|W| - \log|W^*| = (\log|W| - \log|W^O|) - (\log|W^*| - \log|W^O|). \tag{D.50}$$

Let $\Delta_R := \Delta - R$ such that $W = W^* + \Delta = W^O + \Delta_R$. Similar to the arguments to show equation (9) of Rothman, Bickel, Levina, and Zhu (2008), we have

$$\begin{aligned} \log|W| - \log|W^O| &= \log|W^O + \Delta_R| - \log|W^O| = \operatorname{tr}(\Sigma\Delta_R) - \epsilon(\Delta_R), \\ \epsilon(\Delta_R) &= \tilde{\Delta}_R^T\left[\int_0^1 (1-\nu)(W^O + \nu\Delta_R)^{-1} \otimes (W^O + \nu\Delta_R)^{-1}\, d\nu\right]\tilde{\Delta}_R, \end{aligned} \tag{D.51}$$

where $\otimes$ is the Kronecker product, $\tilde{\Delta}_R$ is Δ_R vectorized to match the dimension of the Kronecker product, and $\epsilon(\Delta_R)$ is the remainder from a Taylor expansion.

Below we continue to establish the bound on $\epsilon(\Delta_R)$ following Rothman, Bickel, Levina, and Zhu (2008), see the arguments leading to their equation (18). By defini-

tion, $\varphi_{\min}(M) = \min_{\|x\|=1} x^T M x$. We have, for $\Delta \in \Theta_n(C)$,

$$\begin{aligned}
&\varphi_{\min}\left(\int_0^1 (1-\nu)(W^O+\nu\Delta_R)^{-1}\otimes(W^O+\nu\Delta_R)^{-1}\,\mathrm{d}\nu\right)\\
\geq&\int_0^1 (1-\nu)\varphi^2_{\min}(W^O+\nu\Delta_R)^{-1}\,\mathrm{d}\nu \geq \frac{1}{2}\min_{0\leq\nu\leq 1}\varphi^2_{\min}(W^O+\nu\Delta_R)^{-1}\\
\geq&\frac{1}{2}\min\left\{\varphi^2_{\min}(W^O+\Delta_R)^{-1} : \|\Delta_R\|_F \leq \|R\|_F + Cr_n\right\}. \qquad \text{(D.52)}
\end{aligned}$$

Let $\underline{k} > 0$ and $\bar{k} > 0$ denote the smallest and largest eigenvalue of the covariance matrix Σ. Now

$$\varphi^2_{\min}(W^O+\Delta_R)^{-1} = \varphi^{-2}_{\max}(W^O+\Delta_R) \geq (\|W^O\|+\|R\|+\|\Delta\|)^{-2} \geq \frac{1}{2}\underline{k}^2, \qquad \text{(D.53)}$$

asymptotically, since $\|R\| \leq \|R\|_F = O(r_n)$ and $\|\Delta\| \leq \|\Delta\|_F = Cr_n$. Therefore, for a large enough C, we have

$$\epsilon(\Delta_R) \geq \frac{1}{4}\underline{k}^2\|\Delta - R\|_F^2 \geq \frac{1}{4}\underline{k}^2\|\Delta\|_F^2 - \frac{1}{4}\underline{k}^2\|R\|_F^2 \geq \frac{1}{8}\underline{k}^2\|\Delta\|_F^2, \qquad \text{(D.54)}$$

for $\Delta \in \Theta_n(C)$ asymptotically, where the last inequality holds because $\|R\|_F^2 = O(r_n^2)$ and $\|\Delta\|_F^2 = C^2 r_n^2$.

Following the same arguments as in (D.51) by replacing W with W^*, we obtain similar results by setting $\Delta = 0$, such that Δ_R is replaced by $-R$,

$$\begin{aligned}
\log|W^*| - \log|W^O| = \log|W^O - R| - \log|W^O| = -\operatorname{tr}(\Sigma R) - \epsilon(-R), \qquad \text{(D.55)}\\
\epsilon(-R) = \tilde{R}^T\left[\int_0^1 (1-\nu)(W^O-\nu R)^{-1}\otimes(W^O-\nu R)^{-1}\,\mathrm{d}\nu\right]\tilde{R}.
\end{aligned}$$

By the same arguments as in (D.52), we have

$$\begin{aligned}
&\varphi_{\max}\left(\int_0^1 (1-\nu)(W^O-\nu R)\otimes(W^O-\nu R)^{-1}\,\mathrm{d}\nu\right) \qquad \text{(D.56)}\\
\leq&\int_0^1 (1-\nu)\varphi^2_{\max}(W^O-\nu R)^{-1}\,\mathrm{d}\nu \leq \frac{1}{2}\max_{0\leq\nu\leq 1}\varphi^2_{\max}(W^O-\nu R)^{-1}, \text{ where}
\end{aligned}$$

$$\varphi^2_{\max}(W^O-\nu R)^{-1} = \varphi^{-2}_{\min}(W^O-\nu R) \leq (\varphi_{\min}(W^O)-\|R\|)^{-2} \leq \left(\frac{1}{\underline{k}} - \|R\|_F\right)^{-2} \leq 2\bar{k}^2,$$

asymptotically, since $\|R\| \leq \|R\|_F = o(1)$. Therefore, for a C large enough, we have

$$\epsilon(-R) \leq \bar{k}^2\|R\|_F^2 \leq \frac{1}{16}\underline{k}^2\|\Delta\|_F^2, \qquad \text{(D.57)}$$

for $\Delta \in \Theta_n(C)$ asymptotically, where the last inequality holds because $\|R\|_F^2 = O(r_n^2)$ and $\|\Delta\|_F^2 = C^2 r_n^2$.

Putting together (D.50), (D.51), and (D.55), we have

$$\begin{aligned}\log|W| - \log|W^*| &= \operatorname{tr}(\Sigma\Delta_R) - \epsilon(\Delta_R) + \operatorname{tr}(\Sigma R) + \epsilon(-R) \\ &= \operatorname{tr}(\Sigma\Delta) + \epsilon(-R) - \epsilon(\Delta_R). \end{aligned} \tag{D.58}$$

Then, for C large enough, we have,

$$\begin{aligned} G(\Delta) =& \operatorname{tr}\left[(W - W^*)(\widehat{\Sigma} - \Sigma)\right] - (\log|W| - \log|W^*|) \\ &+ \operatorname{tr}\left[(W - W^*)\Sigma\right] + \lambda\left(|W^-|_1 - |W^{*-}|_1\right) \\ =& \epsilon(\Delta_R) - \epsilon(-R) + \left[\operatorname{tr}\left[\Delta(\widehat{\Sigma} - \Sigma)\right] + \lambda\left(|W^-|_1 - |W^{*-}|_1\right)\right] \\ \geq& \frac{1}{16}\underline{k}^2\|\Delta\|_F^2 + \left[\operatorname{tr}\left[\Delta(\widehat{\Sigma} - \Sigma)\right] + \lambda\left(|W^-|_1 - |W^{*-}|_1\right)\right], \end{aligned} \tag{D.59}$$

for $\Delta \in \Theta_n(C)$ asymptotically, where the first equality follows from (D.48), the second equality uses $\Delta = W - W^*$ and (D.58), and the inequality follows from (D.54) and (D.57).

The remainder of the proof to establish $G(\Delta) > 0$ on the boundary of $\Theta_n(C)$ for C large enough follows the same arguments as in the proof of Theorem 1 of Rothman, Bickel, Levina, and Zhu (2008); specifically, see equations (12)–(14) for the study of $\operatorname{tr}\left[\Delta(\hat{\Sigma} - \Sigma)\right]$, equation (11) for the penalty term $\lambda\left(|W^-|_1 - |W^{*-}|_1\right)$, and equations (16)–(17) for the arguments to show that $G(\Delta) > 0$ for C large enough. None of these arguments involve the approximation error R. We omit these steps here because they are identical.

Finally, showing that the same convergence holds for the correlation-based estimator $\widehat{W}_G$ follows from the same arguments as in Theorem 2 of Rothman, Bickel, Levina, and Zhu (2008). Here, we can use the Frobenius norm rather than the spectral norm because we specify the convergence rate as $r_n = \sqrt{(p+s)\log p/n}$. Because s is proportional to or slightly larger than p in our applications, we do not aim to refine the rate by replacing $p + s$ with s as they do in Theorem 2 under the spectral norm. □

E Additional Implementation Details

E.1 Verification for the Covariance Structure Model

The covariance structure model in the earnings dynamics literature, investigated in Example 2 and the Baker and Solon (2003) model, fits in the general framework of this paper. Consider $X_i \in \mathbb{R}^T$, which is i.i.d. across i, and define $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. In the covariance structure model, the observed sample moment is $\widetilde{m} = (n-1)^{-1}\sum_{i=1}^n \widetilde{m}_i$, where $\widetilde{m}_i = \text{vech}\left((X_i - \overline{X})(X_i - \overline{X})'\right)$ and $\text{vech}(\cdot)$ denotes the usual half-vectorization operator for symmetric matrices. This problem fits in our framework by considering the sample moments $\overline{m} = n^{-1}\sum_{i=1}^n m_i$, where $m_i = \text{vech}\left((X_i - \mathbb{E}[X_i])(X_i - \mathbb{E}[X_i])'\right)$. Comparing $\widetilde{m}$ and $\overline{m}$, the differences between $\overline{X}$ and $\mathbb{E}[X_i]$ in their centering terms, and the difference between $n-1$ and n in their normalizations, are asymptotically negligible for studying the asymptotic distribution of the resulting minimum distance estimator. Therefore, although the minimum distance estimator is constructed with the observed moments $\widetilde{m}$, we can derive its asymptotic distribution using $\overline{m}$.

E.2 Tuning Parameter for GLasso

In practice, we select the tuning parameter λ for the GLasso estimator by cross-validation. For the cross-fitted estimator, $\widehat{W}_{G,-k}$ is computed with data $i \in \mathcal{I}_{-k}$. In this case, we further divide the data in $\mathcal{I}_{-k}$ into L folds to choose λ for the computation of $\widehat{W}_{G,-k}$. We use $L = 5$ for the tuning parameter choice in all cases.

The cross-validation procedure is as follows. Randomly partition the sample used to estimate the weighting matrix into L folds of equal size. We compute a sample covariance matrix for the training and test folds, $\widehat{\Sigma}_{-\ell}$ and $\widehat{\Sigma}_{\ell}$, respectively. Define $\mathcal{L}(\widehat{\Sigma}, W) = \log(\det W) - \text{tr}(W\widehat{\Sigma})$ as the log-likelihood function in (2.15), so that $-\mathcal{L}(\widehat{\Sigma}, W)$ is the loss function. For a given λ, obtain the GLasso estimator $\widehat{W}_{-\ell}(\lambda)$ following (2.15) and (2.16), with $\widehat{\Sigma} = \widehat{\Sigma}_{-\ell}$ for each $\ell = 1, \cdots, L$.

We compute an optimal tuning parameter by maximizing the averaged log-likelihood (minimizing the averaged loss function) of the test samples. That is

$$\lambda^* = \underset{\lambda \in [0, \lambda_{\max}]}{\arg\max} \quad \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{L}(\widehat{\Sigma}_{\ell}, \widehat{W}_{-\ell}(\lambda)). \tag{E.1}$$

In practice, we first solve this maximization problem by defining a grid of λ values, where the highest value of λ is $\max_{j,k \in \{1,\ldots,p\}} |\widehat{\Sigma}_{jk}|$, which produces the diagonal matrix. The solution can then be refined, using Brent's method (Brent, 1973), for example, over an interval around the optimal value found on the initial grid.

F Additional Numerical Results

F.1 Recovering the Sparsity Structure Through Regularization

Table A-1 reports the proportion of zeros in the oracle weighting matrix that are successfully recovered under various distributional specifications and cross-fitting configurations. We average the fraction of zeros across folds within each simulation and then across all simulations. In this context, the label F denotes a full-sample estimate. We intentionally avoid the notation $K = 1$ for the full sample to prevent ambiguity; whereas cross-fitting utilizes $n - n/K$ observations to estimate the weighting matrix for a given fold, the full-sample approach utilizes all n observations.

Our results indicate that the cross-validated GW procedure selects a precision matrix with an average non-zero fraction of less than 2%. This sparsity becomes even more pronounced in the high-dimensional ($T = 0.2n$), large-sample ($n = 1000$) case, where the fraction drops to 0.01% or less across all distributions. These findings demonstrate that the GLasso procedure successfully recovers the sparsity pattern, which in this setup is the identity matrix. Regarding the full-sample case ($K = 1$, denoted as F in our tables), the recovery of the sparsity pattern is slightly more efficient than in the cross-fitting cases ($K > 1$). This is expected, as the precision matrix in the F specification is estimated using the full n observations rather than the $n - n/K$ observations available in each training fold.

F.2 The Impact of Sample Splitting and Number of Folds

Table A-2 demonstrates the impact of sample splitting and the number of cross-fitting folds on estimators using different weighting matrices. First, the results confirm that cross-fitting estimators consistently outperform full-sample estimators across all weighting matrices and distributions. Second, the estimators are generally robust to the choice of the cross-fitting parameter K. While the direction of change as K increases from 2 to 5 is inconclusive across different specifications, the patterns tend to be consistent across the different weighting matrices within a given specification. For instance: (i) for the normal distribution ($n = 100$), bias remains near zero for both $K = 2$ and $K = 5$, while RMSE decreases and coverage probability increases with K; (ii) for the exponential distribution ($n = 100$), bias decreases, but RMSE increases and coverage probability decreases as K rises; and (iii) for the log-normal distribution ($n = 100$), bias increases while RMSE and coverage probability both decrease as K increases. These sensitivity patterns diminish as the sample size increases to $n = 1000$. Overall, the impact of the number of folds on the final estimates is minor,

reinforcing the stability of the cross-fitting procedure.

F.3 Comparison of DML1 and DML2

In the typology of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018b), our proposed estimator is a form of a DML1 estimator. Our cross-fitting procedure relies on solving a fold-specific minimum distance problem where $Q_{nk}(\theta) = (\bar{m}_k - f(\theta))'\widehat{W}_{-k}(\bar{m}_k - f(\theta))$, $\widehat{\theta}^{(k)} = \arg\min_{\theta\in\Theta} Q_{nk}(\theta)$, and $\widehat{\theta}_{DML1} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}^{(k)}$. By contrast, a DML2 estimator aggregates the fold-specific criterion functions and solves a single optimization problem, $\widehat{\theta}_{DML2} = \arg\min_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} Q_{nk}(\theta)$.

In our main simulations, we set the number of folds to $K = 2$. In Table A-3 we report results from the Altonji and Segal (1996) simulation design, as we change the number of folds to $K \in \{2, 5, 10, 20\}$, setting $n = 100$ and $T = 10$. We do not document any substantial impact of increasing K on either bias or RMSE of our DML1 estimator, although we do see a slight increase in coverage rates, in some cases going slightly above the nominal level. In general, results for DML2 are vary comparable to DML1 when $K = 2$. However, as the number of folds increases, we find that some distributions (most notably, log-normal), experience an increase in both bias and RMSE. No such pattern is observed for DML1.

We repeat this exercise with $n = 1000$ in Table A-4 and show that the gaps between DML1 and DML2 are substantially reduced. Results with $T = 0.2n$ (not shown) are qualitatively the same.

F.4 Alternative Bias Correction Methods

High Dimensional Results. In the context of the Altonji and Segal (1996) simulation design, Section 4.3 compares *GW* to other leading methods for achieving bias correction in a low-dimensional case with fixed T. We now repeat the analysis in the high-dimensional case with $T = 0.2n$. Results are presented in Table A-5. They are very favourable to the GW estimator, which achieves the lowest bias, and the lowest RMSE (except for the log-normal distribution, where HO obtains lower RMSE).

Computational Burdens. Each of these bias-correction methods entails different computational burdens. HO re-estimates θ for different bootstrap iterations and different levels of trimming to approximate the bias. NS is an analytical correction from a single estimate of θ using the full-sample OWMD. Both JK1 and JK2 compute the weighting matrix $(n-1)$ times; JK1 computes the θ estimator $(n-1)$ times and JK2 once. GW computes both the weighting matrix and the θ estimator $K = 2$ times.

F.5 Trimming Outliers

Table A-6 presents a version of GW that trims outliers following the idea of Horowitz (1998). In Horowitz (1998), there is a criterion that ranks outlier observations and a procedure to trim the top κ proportion of observations. The trimmed GLasso runs a GLasso procedure over the trimmed data to construct $\widehat{W}^O$. Applying a GLasso penalization to the trimmed data is more numerically stable than OW, since the trimmed covariance matrix used for trimmed-OW is not guaranteed to be invertible. The trimming parameter κ and sparsity penalty λ are both chosen via likelihood cross-validation. This pre-processing step only affects the training folds; the test folds use all the data available. The results in Table A-6 show some improvements in bias, RMSE, and coverage for the log-normal distribution, particularly for a small sample size. Results in other cases are quantitatively similar.

Table A-1: Altonji and Segal (1996) Design: Average Off-Diagonal Zeros

		$T = 10$		$T = 0.2n$	
Distribution	K	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
t(5)	F	0.0059	0.0034	0.0011	0.0000
	2	0.0066	0.0045	0.0017	0.0000
	5	0.0066	0.0039	0.0013	0.0000
	10	0.0060	0.0040	0.0013	0.0000
	20	0.0060	0.0035	0.0012	0.0000
t(10)	F	0.0079	0.0100	0.0014	0.0000
	2	0.0086	0.0095	0.0019	0.0000
	5	0.0073	0.0099	0.0015	0.0000
	10	0.0080	0.0103	0.0016	0.0000
	20	0.0079	0.0103	0.0016	0.0000
t(15)	F	0.0060	0.0109	0.0013	0.0000
	2	0.0080	0.0113	0.0020	0.0000
	5	0.0072	0.0113	0.0016	0.0000
	10	0.0067	0.0108	0.0016	0.0000
	20	0.0072	0.0110	0.0016	0.0000
Normal	F	0.0083	0.0122	0.0024	0.0000
	2	0.0091	0.0142	0.0026	0.0000
	5	0.0085	0.0138	0.0024	0.0000
	10	0.0087	0.0125	0.0025	0.0000
	20	0.0082	0.0123	0.0024	0.0000
Uniform	F	0.0169	0.0168	0.0050	0.0001
	2	0.0177	0.0196	0.0048	0.0000
	5	0.0154	0.0184	0.0053	0.0000
	10	0.0153	0.0176	0.0053	0.0001
	20	0.0151	0.0175	0.0055	0.0001
Log normal	F	0.0028	0.0031	0.0007	0.0000
	2	0.0047	0.0046	0.0020	0.0000
	5	0.0032	0.0036	0.0010	0.0000
	10	0.0034	0.0038	0.0009	0.0000
	20	0.0030	0.0036	0.0008	0.0000
Exp	F	0.0033	0.0028	0.0010	0.0000
	2	0.0042	0.0020	0.0015	0.0000
	5	0.0035	0.0025	0.0010	0.0000
	10	0.0033	0.0028	0.0010	0.0000
	20	0.0034	0.0031	0.0010	0.0000
Half-normal	F	0.0047	0.0081	0.0011	0.0000
	2	0.0061	0.0066	0.0017	0.0000
	5	0.0055	0.0074	0.0012	0.0000
	10	0.0049	0.0081	0.0012	0.0000
	20	0.0051	0.0081	0.0012	0.0000
Bimodal	F	0.0124	0.0180	0.0058	0.0001
	2	0.0165	0.0186	0.0049	0.0000
	5	0.0151	0.0174	0.0054	0.0000
	10	0.0149	0.0174	0.0058	0.0001
	20	0.0143	0.0172	0.0060	0.0001

Notes: Average number of non-zero off-diagonal elements under cross-fitted GLasso-weighted weighting matrix as the number of cross-fitting folds K is varied, including F for the full-sample estimator. Results are shown when the panel time dimension is fixed, $T = 10$, and when it increases with the cross-sectional dimension, $T = 0.2n$.

Table A-2: Altonji and Segal (1996) Design: Impact of Folds, $T = 10$

			Bias			RMSE			Coverage Prob.		
Distribution	n	K	DW	OW	GW	DW	OW	GW	DW	OW	GW
t(5)	100	F	-0.123	-0.124	-0.123	0.141	0.142	0.141	0.327	0.309	0.324
	100	2	0.004	0.005	0.004	0.124	0.129	0.124	0.824	0.834	0.823
	100	5	0.002	0.003	0.002	0.118	0.121	0.119	0.838	0.835	0.837
t(10)	100	F	-0.063	-0.062	-0.063	0.085	0.086	0.085	0.571	0.554	0.570
	100	2	0.003	0.004	0.003	0.073	0.077	0.074	0.854	0.858	0.855
	100	5	0.001	0.003	0.001	0.071	0.075	0.071	0.860	0.871	0.860
t(15)	100	F	-0.051	-0.051	-0.051	0.074	0.075	0.074	0.657	0.630	0.656
	100	2	0.003	0.003	0.003	0.065	0.068	0.065	0.863	0.858	0.861
	100	5	0.002	0.002	0.002	0.062	0.064	0.062	0.874	0.882	0.875
Normal	100	F	-0.036	-0.037	-0.036	0.058	0.059	0.058	0.747	0.724	0.746
	100	2	-0.000	-0.001	-0.000	0.052	0.055	0.052	0.884	0.888	0.884
	100	5	-0.000	-0.000	-0.000	0.049	0.052	0.049	0.899	0.889	0.901
Uniform	100	F	-0.007	-0.007	-0.007	0.030	0.031	0.030	0.887	0.861	0.887
	100	2	-0.001	-0.001	-0.001	0.029	0.032	0.029	0.917	0.913	0.920
	100	5	-0.002	-0.002	-0.002	0.030	0.032	0.030	0.929	0.917	0.929
Log normal	100	F	-0.475	-0.482	-0.476	0.490	0.496	0.490	0.013	0.012	0.013
	100	2	-0.025	-0.020	-0.024	0.581	0.616	0.582	0.661	0.665	0.662
	100	5	-0.041	-0.041	-0.041	0.436	0.449	0.437	0.656	0.648	0.656
Exp	100	F	-0.166	-0.168	-0.166	0.190	0.192	0.190	0.248	0.234	0.248
	100	2	-0.005	-0.007	-0.005	0.142	0.148	0.142	0.829	0.835	0.828
	100	5	-0.002	-0.003	-0.002	0.150	0.154	0.151	0.806	0.814	0.807
Half-normal	100	F	-0.060	-0.061	-0.060	0.082	0.083	0.082	0.606	0.576	0.606
	100	2	0.002	0.003	0.002	0.069	0.074	0.069	0.880	0.870	0.880
	100	5	0.001	0.002	0.001	0.065	0.068	0.065	0.874	0.886	0.877
Bimodal	100	F	-0.011	-0.011	-0.011	0.030	0.031	0.030	0.843	0.819	0.844
	100	2	-0.000	0.001	0.000	0.029	0.031	0.029	0.896	0.896	0.894
	100	5	-0.000	-0.000	-0.000	0.030	0.031	0.030	0.910	0.903	0.910
t(5)	1000	F	-0.027	-0.027	-0.027	0.036	0.037	0.036	0.622	0.627	0.622
	1000	2	-0.001	-0.001	-0.001	0.031	0.031	0.031	0.873	0.872	0.873
	1000	5	-0.002	-0.002	-0.002	0.031	0.031	0.031	0.862	0.867	0.863
t(10)	1000	F	-0.008	-0.008	-0.008	0.019	0.019	0.019	0.840	0.845	0.840
	1000	2	-0.001	-0.001	-0.001	0.018	0.018	0.018	0.900	0.894	0.900
	1000	5	-0.001	-0.001	-0.001	0.018	0.018	0.018	0.903	0.899	0.902
t(15)	1000	F	-0.006	-0.006	-0.006	0.017	0.017	0.017	0.851	0.855	0.852
	1000	2	-0.001	-0.001	-0.001	0.016	0.016	0.016	0.894	0.887	0.892
	1000	5	-0.001	-0.001	-0.001	0.016	0.016	0.016	0.896	0.897	0.895
Normal	1000	F	-0.004	-0.004	-0.004	0.015	0.015	0.015	0.875	0.873	0.875
	1000	2	-0.001	-0.001	-0.001	0.014	0.014	0.014	0.901	0.894	0.902
	1000	5	-0.001	-0.001	-0.001	0.014	0.014	0.014	0.898	0.896	0.898
Uniform	1000	F	-0.000	-0.000	-0.000	0.009	0.009	0.009	0.899	0.893	0.899
	1000	2	0.000	0.000	0.000	0.009	0.009	0.009	0.905	0.902	0.904
	1000	5	0.000	0.000	0.000	0.009	0.009	0.009	0.902	0.897	0.901
Log normal	1000	F	-0.164	-0.164	-0.164	0.177	0.177	0.177	0.136	0.135	0.136
	1000	2	0.000	0.001	0.000	0.151	0.153	0.151	0.785	0.791	0.785
	1000	5	-0.007	-0.007	-0.007	0.141	0.142	0.141	0.764	0.765	0.763
Exp	1000	F	-0.022	-0.022	-0.022	0.037	0.037	0.037	0.725	0.721	0.725
	1000	2	0.000	0.000	0.000	0.033	0.033	0.033	0.865	0.867	0.865
	1000	5	0.000	0.000	0.000	0.032	0.032	0.032	0.872	0.874	0.872
Half-normal	1000	F	-0.007	-0.007	-0.007	0.019	0.019	0.019	0.848	0.847	0.848
	1000	2	-0.001	-0.000	-0.001	0.018	0.018	0.018	0.885	0.882	0.885
	1000	5	-0.001	-0.001	-0.001	0.018	0.018	0.018	0.877	0.889	0.876
Bimodal	1000	F	-0.001	-0.001	-0.001	0.009	0.009	0.009	0.894	0.890	0.893
	1000	2	-0.000	-0.000	-0.000	0.009	0.009	0.009	0.900	0.902	0.900
	1000	5	-0.000	-0.000	-0.000	0.009	0.009	0.009	0.906	0.909	0.905

Notes: Average bias, root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative weighting schemes (diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW) as the number of folds K is varied, including F for the full-sample estimator.

Table A-3: Altonji and Segal (1996) Design: Aggregation Comparison, $T = 10$, $n = 100$

		Bias				RMSE				Coverage Prob.			
Distribution	DML	GW_2	GW_5	GW_{10}	GW_{20}	GW_2	GW_5	GW_{10}	GW_{20}	GW_2	GW_5	GW_{10}	GW_{20}
t(5)	1	0.004	0.002	0.001	0.001	0.124	0.119	0.121	0.122	0.823	0.837	0.832	0.851
	2	0.014	0.022	0.024	0.026	0.128	0.133	0.141	0.143	0.850	0.851	0.855	0.871
t(10)	1	0.003	0.001	0.001	0.002	0.074	0.071	0.072	0.073	0.855	0.860	0.872	0.909
	2	0.010	0.011	0.012	0.012	0.076	0.075	0.076	0.077	0.862	0.878	0.885	0.909
t(15)	1	0.003	0.002	0.002	0.001	0.065	0.062	0.063	0.065	0.861	0.875	0.890	0.914
	2	0.009	0.009	0.010	0.009	0.067	0.064	0.065	0.068	0.868	0.881	0.898	0.921
Normal	1	-0.000	-0.000	-0.000	-0.000	0.052	0.049	0.050	0.053	0.884	0.901	0.908	0.942
	2	0.004	0.005	0.005	0.005	0.052	0.050	0.050	0.053	0.883	0.902	0.915	0.942
Uniform	1	-0.001	-0.002	-0.001	-0.001	0.029	0.030	0.033	0.037	0.920	0.929	0.930	0.958
	2	-0.001	-0.001	-0.001	-0.001	0.029	0.030	0.033	0.037	0.916	0.927	0.931	0.958
Log norm.	1	-0.024	-0.041	-0.040	-0.038	0.582	0.437	0.452	0.454	0.662	0.656	0.640	0.641
	2	0.001	0.078	0.126	0.160	0.590	0.600	0.681	0.726	0.686	0.721	0.718	0.720
Exp	1	-0.005	-0.002	-0.003	-0.005	0.142	0.151	0.151	0.151	0.828	0.807	0.788	0.799
	2	0.010	0.033	0.038	0.039	0.150	0.178	0.182	0.187	0.842	0.825	0.814	0.845
Half-norm.	1	0.002	0.001	0.002	0.002	0.069	0.065	0.065	0.066	0.880	0.877	0.899	0.921
	2	0.009	0.011	0.012	0.013	0.071	0.068	0.069	0.070	0.887	0.891	0.909	0.939
Bimodal	1	0.000	-0.000	-0.000	-0.000	0.029	0.030	0.032	0.036	0.894	0.910	0.928	0.958
	2	0.001	0.001	0.001	0.001	0.029	0.030	0.032	0.036	0.897	0.914	0.931	0.958

Notes: Average bias, root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative de-biased machine learning (DML) aggregation approaches with cross-fitted GLasso-weighting as the number of cross-fitting folds is varied.

Table A-4: Altonji and Segal (1996) Design: Aggregation Comparison, $T = 10$, $n = 1000$

		Bias				RMSE				Coverage Prob.			
Distribution	DML	GW_2	GW_5	GW_{10}	GW_{20}	GW_2	GW_5	GW_{10}	GW_{20}	GW_2	GW_5	GW_{10}	GW_{20}
t(5)	1	-0.001	-0.002	-0.002	-0.002	0.031	0.031	0.031	0.031	0.873	0.863	0.872	0.862
	2	0.001	0.001	0.002	0.002	0.032	0.032	0.032	0.032	0.876	0.870	0.879	0.873
t(10)	1	-0.001	-0.001	-0.001	-0.001	0.018	0.018	0.018	0.018	0.900	0.902	0.901	0.894
	2	0.000	0.000	0.000	0.000	0.018	0.018	0.018	0.018	0.905	0.913	0.902	0.902
t(15)	1	-0.001	-0.001	-0.001	-0.001	0.016	0.016	0.016	0.016	0.892	0.895	0.901	0.897
	2	-0.000	-0.000	-0.000	-0.000	0.016	0.016	0.016	0.016	0.898	0.899	0.899	0.899
Normal	1	-0.001	-0.001	-0.001	-0.001	0.014	0.014	0.014	0.014	0.902	0.898	0.901	0.904
	2	-0.000	-0.000	-0.000	-0.000	0.014	0.014	0.014	0.014	0.904	0.902	0.902	0.911
Uniform	1	0.000	0.000	0.000	-0.000	0.009	0.009	0.009	0.009	0.904	0.901	0.899	0.913
	2	0.000	0.000	0.000	0.000	0.009	0.009	0.009	0.009	0.903	0.902	0.899	0.913
Log norm.	1	0.000	-0.007	-0.007	-0.005	0.151	0.141	0.144	0.147	0.785	0.763	0.765	0.756
	2	0.015	0.021	0.025	0.031	0.170	0.161	0.172	0.180	0.812	0.806	0.800	0.805
Exp	1	0.000	0.000	0.000	0.000	0.033	0.032	0.032	0.032	0.865	0.872	0.876	0.871
	2	0.002	0.003	0.003	0.003	0.034	0.033	0.032	0.032	0.870	0.883	0.878	0.881
Half-norm.	1	-0.001	-0.001	-0.001	-0.001	0.018	0.018	0.017	0.017	0.885	0.876	0.885	0.891
	2	0.000	0.000	-0.000	0.000	0.018	0.018	0.017	0.017	0.885	0.883	0.890	0.894
Bimodal	1	-0.000	-0.000	-0.000	-0.000	0.009	0.009	0.009	0.009	0.900	0.905	0.912	0.913
	2	-0.000	-0.000	-0.000	-0.000	0.009	0.009	0.009	0.009	0.902	0.904	0.911	0.913

Notes: Average bias, root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative de-biased machine learning (DML) aggregation approaches with cross-fitted GLasso-weighting as the number of cross-fitting folds is varied.

Table A-5: Altonji and Segal (1996) Design: Comparison of Alternative Estimators, $T = 0.2n$

		Bias					RMSE				
Distribution	n	HO	NS	JK1	JK2	GW	HO	NS	JK1	JK2	GW
t(5)	100	0.058	-0.096	0.009	0.024	0.003	0.117	0.110	0.089	0.101	0.085
t(10)	100	0.063	-0.041	0.011	0.017	0.003	0.102	0.061	0.053	0.057	0.051
t(15)	100	0.055	-0.032	0.010	0.015	0.002	0.090	0.053	0.047	0.050	0.047
Normal	100	0.046	-0.020	0.011	0.014	0.000	0.079	0.042	0.040	0.042	0.038
Uniform	100	0.013	-0.004	0.010	0.010	-0.000	0.034	0.022	0.024	0.024	0.020
Log normal	100	-0.150	-0.455	0.018	0.177	-0.002	0.296	0.468	0.508	0.776	0.454
Exp	100	0.071	-0.124	0.007	0.035	-0.001	0.155	0.148	0.121	0.141	0.113
Half-normal	100	0.053	-0.033	0.014	0.022	0.003	0.093	0.056	0.054	0.058	0.051
Bimodal	100	0.019	-0.007	0.009	0.010	-0.001	0.038	0.022	0.023	0.023	0.020
t(5)	1000	0.014	-0.018	0.001	0.001	0.000	0.021	0.019	0.008	0.008	0.007
t(10)	1000	0.008	-0.004	0.001	0.001	0.000	0.012	0.006	0.005	0.005	0.004
t(15)	1000	0.007	-0.003	0.001	0.001	-0.000	0.010	0.005	0.004	0.004	0.004
Normal	1000	0.006	-0.002	0.001	0.001	0.000	0.008	0.004	0.004	0.004	0.003
Uniform	1000	0.002	-0.001	0.001	0.001	-0.000	0.003	0.002	0.002	0.002	0.002
Log normal	1000	0.041	-0.138	-0.001	0.002	-0.002	0.075	0.140	0.038	0.039	0.034
Exp	1000	0.022	-0.011	0.001	0.001	0.000	0.028	0.013	0.008	0.008	0.007
Half-normal	1000	0.008	-0.002	0.001	0.001	0.000	0.012	0.005	0.004	0.004	0.004
Bimodal	1000	0.003	-0.001	0.001	0.001	0.000	0.005	0.002	0.002	0.002	0.002

Notes: Average bias and root-mean square error (RMSE), under alternative estimators: HO (Horowitz, 1998), NS (Newey and Smith, 2004), JK_1 and JK_2 (Kezdi, Hahn, and Solon, 2002), and GW (cross-fitted GLasso-weighted).

Table A-6: Altonji and Segal (1996) Design: Comparison with Trimmed GLasso

		$T = 10$						$T = 0.2n$					
		Bias		RMSE		Cov. Prob.		Bias		RMSE		Cov. Prob.	
Distribution	n	GW	GW-Tr	GW	GW-Tr	GW	GW-Tr	GW	GW-Tr	GW	GW-Tr	GW	GW-Tr
t(5)	100	0.004	0.004	0.124	0.105	0.823	0.847	0.003	0.002	0.085	0.072	0.852	0.868
t(10)	100	0.003	0.003	0.074	0.063	0.855	0.891	0.003	0.003	0.051	0.045	0.860	0.895
t(15)	100	0.003	0.001	0.065	0.058	0.861	0.887	0.002	0.001	0.047	0.043	0.873	0.885
Normal	100	-0.000	0.000	0.052	0.048	0.884	0.901	0.000	-0.000	0.038	0.036	0.878	0.890
Uniform	100	-0.001	-0.001	0.029	0.029	0.920	0.909	-0.000	-0.000	0.020	0.021	0.927	0.931
Log norm.	100	-0.024	0.000	0.582	0.561	0.662	0.723	-0.002	0.006	0.454	0.440	0.697	0.728
Exp	100	-0.005	-0.006	0.142	0.113	0.828	0.872	-0.001	-0.001	0.113	0.096	0.824	0.850
Half-norm.	100	0.002	0.001	0.069	0.060	0.880	0.899	0.003	0.001	0.051	0.047	0.860	0.891
Bimodal	100	0.000	0.001	0.029	0.029	0.894	0.902	-0.001	-0.001	0.020	0.020	0.906	0.925
t(5)	1000	-0.001	-0.001	0.031	0.027	0.873	0.899	0.000	0.000	0.007	0.006	0.883	0.896
t(10)	1000	-0.001	-0.000	0.018	0.017	0.900	0.914	0.000	0.000	0.004	0.004	0.900	0.909
t(15)	1000	-0.001	-0.001	0.016	0.016	0.892	0.897	-0.000	-0.000	0.004	0.004	0.899	0.910
Normal	1000	-0.001	-0.001	0.014	0.014	0.902	0.902	0.000	0.000	0.003	0.003	0.902	0.902
Uniform	1000	0.000	0.000	0.009	0.009	0.904	0.902	-0.000	-0.000	0.002	0.002	0.910	0.914
Log norm.	1000	0.000	-0.003	0.151	0.100	0.785	0.847	-0.002	0.000	0.034	0.030	0.848	0.886
Exp	1000	0.000	0.000	0.033	0.029	0.865	0.885	0.000	0.000	0.007	0.007	0.884	0.906
Half-norm.	1000	-0.001	-0.000	0.018	0.017	0.885	0.906	0.000	0.000	0.004	0.004	0.898	0.903
Bimodal	1000	-0.000	-0.000	0.009	0.009	0.900	0.906	0.000	0.000	0.002	0.002	0.891	0.895

Notes: Average bias, root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under cross-fitted GLasso-weighted (GW) and trimmed cross-fitted GLasso weighted (GW-Tr) when the panel time dimension is fixed, $T = 10$, and when it increases with the cross-sectional dimension, $T = 0.2n$.

F.6 Parameter-Level Results for Baker and Solon (2003)

In Table A-7 we present the full parameter-level results from the Baker and Solon (2003) simulation design, for a cohort sample size of 400. For each of the 60 model parameters, we present the same statistics as reported in our results from the Altonji and Segal (1996) simulation study. Full results for other cohort sizes (800, 1200, and 2000) are available upon request.

F.7 Adjusted Bias Results for Baker and Solon (2003)

Figure A-1 shows the adjusted bias for our Baker and Solon (2003) simulation study, based on normalized parameter estimates, which are uncorrelated with unit variances by construction. The figure shows, across all cohort sizes, that the adjusted bias is largest for OW, followed by DW, EW and GW. The GW estimator always achieves the smallest adjusted bias.

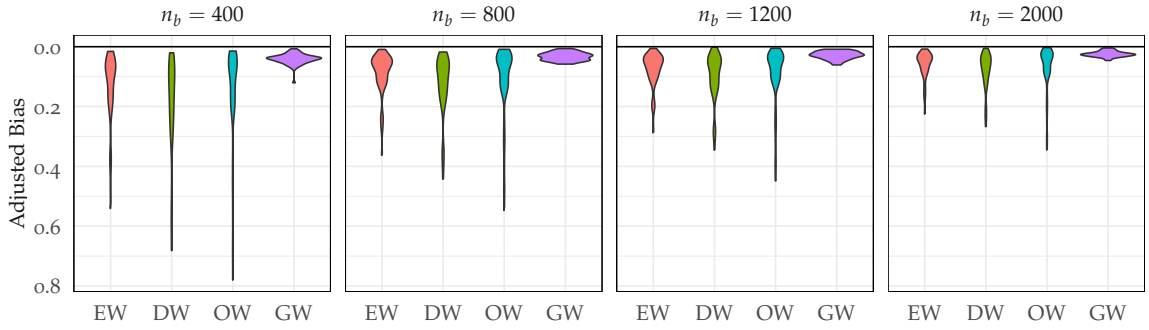


Figure A-1: Violin plots for Baker and Solon (2003) model, showing adjusted parameter biases. To maintain a clearer scale across weighting schemes, the figure presents the square root of the adjusted bias. Figure derived from 1,000 replications. Weighting denoted EW (equally-weighted), DW (diagonally-weighted), OW (optimally-weighted), and GW (cross-fitted GLasso-weighted).

F.8 Empirical Cohort Sizes for Baker and Solon (2003)

In our Baker and Solon (2003) simulation study, we focussed on fixed cohort sizes, and described how the cohort size affects results. In the actual study, the sample size varied by cohort. To illustrate the performance of our estimator in a setting closest to the actual data, we replicate the simulation exercise using the actual reported cohort sizes. Figure A-2 presents violin plots for absolute bias, coverage probability, and log RMSE for the different weighting schemes. The results align closely with those observed in our fixed cohort size simulations (it is most similar to the $n_b = 1200$ results, with GW consistently outperforming other weighting schemes across all metrics. Relative to the variance decomposition results presented in Section 5.4, we

observe the same qualitative patterns (not shown), although the absolute differences across weighting schemes is reduced.

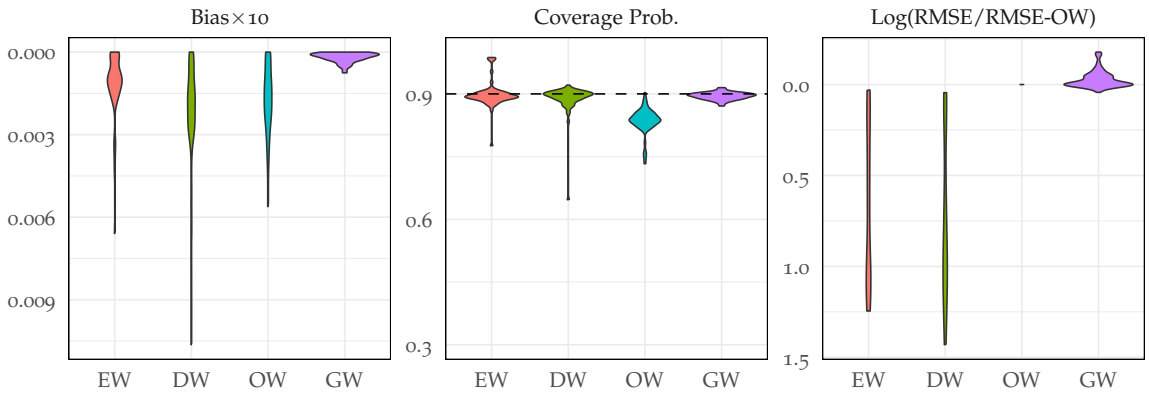


Figure A-2: Violin plots for Baker and Solon (2003) parameters, showing absolute biases, 90% confidence interval coverage probabilities, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting. Figure derived from 1,000 replications with empirical cohort sample sizes. Weighting denoted EW (equally-weighted), DW (diagonally-weighted), OW (optimally-weighted), and GW (cross-fitted GLasso-weighted).

Table A-7: Baker and Solon (2003) Simulation Results: $n_b = 400$

		Bias				RMSE				Coverage Prob.			
Param.	Value	EW	DW	OW	GW	EW	DW	OW	GW	EW	DW	OW	GW
σ^2_α	0.134	-0.007	-0.013	-0.010	-5.0E-4	0.014	0.020	0.013	0.008	0.806	0.690	0.387	0.881
σ^2_β	9.0E-5	-7.1E-6	-1.1E-5	-3.1E-6	1.7E-7	4.0E-5	4.3E-5	2.9E-5	3.0E-5	0.894	0.894	0.700	0.881
$\sigma_{\alpha\beta}$	-0.003	4.8E-4	8.2E-4	2.6E-4	3.7E-5	8.3E-4	0.001	4.8E-4	4.1E-4	0.804	0.663	0.574	0.878
σ^2_r	0.007	-6.8E-4	-0.001	-6.0E-4	-6.8E-5	0.001	0.002	9.0E-4	6.8E-4	0.803	0.617	0.511	0.876
ρ	0.540	0.030	0.049	0.001	0.002	0.039	0.064	0.008	0.008	0.531	0.312	0.654	0.878
γ_0	0.090	-0.002	-0.005	-0.005	8.2E-5	0.012	0.015	0.007	0.004	0.921	0.867	0.392	0.884
γ_1	-0.005	2.8E-4	6.6E-4	6.5E-5	1.5E-5	0.002	0.004	0.001	0.001	0.950	0.885	0.685	0.888
γ_2	6.2E-5	-1.4E-5	-3.6E-5	1.5E-5	-2.3E-6	2.2E-4	4.1E-4	1.2E-4	1.2E-4	0.959	0.900	0.689	0.883
γ_3	2.2E-6	3.2E-7	1.0E-6	-7.3E-7	1.1E-7	9.6E-6	1.8E-5	5.0E-6	5.0E-6	0.962	0.907	0.691	0.893
γ_4	2.1E-9	-3.8E-9	-1.3E-8	7.0E-9	-1.6E-9	1.5E-7	2.6E-7	7.3E-8	7.3E-8	0.964	0.903	0.683	0.902
p_{77}	1.035	0.002	0.002	8.1E-4	3.2E-4	0.012	0.013	0.014	0.013	0.895	0.892	0.700	0.904
p_{78}	1.028	0.003	0.003	8.1E-4	7.2E-5	0.018	0.019	0.019	0.018	0.897	0.887	0.676	0.888
p_{79}	1.005	0.004	0.005	3.1E-4	4.9E-5	0.019	0.021	0.020	0.019	0.894	0.870	0.702	0.893
p_{80}	1.030	0.004	0.006	5.5E-4	5.2E-4	0.021	0.025	0.022	0.020	0.898	0.871	0.698	0.899
p_{81}	1.050	0.006	0.008	4.2E-4	0.001	0.023	0.028	0.024	0.022	0.889	0.858	0.698	0.890
p_{82}	1.143	0.006	0.010	1.7E-4	0.001	0.026	0.031	0.027	0.026	0.893	0.865	0.699	0.889
p_{83}	1.124	0.004	0.008	-2.7E-4	0.001	0.027	0.032	0.030	0.027	0.887	0.864	0.674	0.874
p_{84}	1.125	0.004	0.008	-5.7E-4	0.001	0.027	0.031	0.029	0.028	0.902	0.876	0.687	0.887
p_{85}	1.122	0.003	0.007	-0.001	0.001	0.028	0.032	0.030	0.028	0.905	0.879	0.687	0.885
p_{86}	1.111	0.003	0.005	-6.5E-4	0.001	0.028	0.032	0.030	0.028	0.908	0.883	0.694	0.895
p_{87}	1.098	0.003	0.004	-6.3E-4	0.001	0.030	0.032	0.031	0.028	0.890	0.866	0.686	0.892
p_{88}	1.105	0.002	0.002	-0.001	0.001	0.030	0.033	0.032	0.029	0.884	0.867	0.681	0.875
p_{89}	1.126	0.002	8.5E-4	-0.002	9.0E-4	0.031	0.033	0.033	0.030	0.898	0.881	0.671	0.879
p_{90}	1.127	0.001	-6.6E-4	-0.002	0.001	0.032	0.034	0.034	0.032	0.890	0.892	0.664	0.870
p_{91}	1.234	0.002	-7.4E-4	-0.003	0.001	0.036	0.038	0.038	0.036	0.895	0.877	0.682	0.886
p_{92}	1.253	0.001	-0.002	-0.003	9.6E-4	0.037	0.039	0.040	0.037	0.900	0.897	0.684	0.880
σ^2_{24-25}	0.133	0.004	0.008	-7.1E-4	6.0E-4	0.040	0.046	0.014	0.015	0.894	0.884	0.867	0.900
σ^2_{26-27}	0.084	0.004	0.010	-7.4E-4	6.0E-4	0.033	0.038	0.010	0.011	0.918	0.901	0.832	0.905
σ^2_{28-29}	0.116	0.002	0.006	-0.003	9.3E-4	0.032	0.039	0.013	0.014	0.923	0.909	0.768	0.871
σ^2_{30-31}	0.071	0.002	0.008	-0.003	-5.2E-5	0.033	0.036	0.010	0.009	0.880	0.882	0.727	0.892
σ^2_{32-33}	0.071	0.004	0.010	-0.004	1.3E-4	0.032	0.035	0.011	0.010	0.883	0.887	0.651	0.891
σ^2_{34-35}	0.127	0.001	0.005	-0.009	4.9E-4	0.033	0.039	0.018	0.015	0.906	0.909	0.570	0.887
σ^2_{36-37}	0.085	0.003	0.009	-0.006	2.0E-4	0.030	0.034	0.013	0.011	0.908	0.892	0.585	0.884
σ^2_{38-39}	0.044	0.005	0.013	-0.003	2.5E-4	0.026	0.030	0.009	0.007	0.906	0.875	0.599	0.898
σ^2_{40-41}	0.066	0.006	0.013	-0.005	3.9E-5	0.030	0.033	0.011	0.009	0.883	0.864	0.600	0.895
σ^2_{42-43}	0.074	0.006	0.012	-0.005	4.7E-4	0.029	0.032	0.012	0.010	0.881	0.880	0.582	0.913
σ^2_{44-45}	0.054	0.007	0.014	-0.004	5.4E-4	0.027	0.031	0.010	0.009	0.883	0.862	0.625	0.893
σ^2_{46-47}	0.071	0.005	0.012	-0.005	8.3E-4	0.028	0.031	0.012	0.011	0.895	0.877	0.594	0.892
σ^2_{48-49}	0.090	0.006	0.012	-0.008	-2.8E-4	0.026	0.031	0.016	0.012	0.918	0.891	0.560	0.902
σ^2_{50-51}	0.167	0.004	0.007	-0.012	7.6E-4	0.031	0.035	0.024	0.019	0.907	0.899	0.586	0.880
σ^2_{52-53}	0.157	0.008	0.012	-0.010	0.002	0.032	0.036	0.021	0.019	0.879	0.885	0.605	0.869
σ^2_{54-55}	0.251	0.002	0.002	-0.014	-2.7E-4	0.039	0.045	0.028	0.025	0.889	0.890	0.702	0.901
σ^2_{56-57}	0.295	0.002	0.002	-0.013	-6.4E-4	0.046	0.056	0.032	0.030	0.897	0.887	0.753	0.891
σ^2_{58-59}	0.377	4.0E-4	-0.005	-0.012	0.001	0.049	0.062	0.036	0.037	0.892	0.890	0.808	0.885
σ^2_{60-61}	0.388	4.9E-4	-0.007	-0.006	0.002	0.050	0.062	0.036	0.037	0.904	0.894	0.849	0.896
λ_{78}	1.132	0.002	-0.006	0.002	0.001	0.057	0.052	0.025	0.024	0.859	0.874	0.695	0.875
λ_{79}	0.950	-0.003	-0.010	0.002	8.5E-5	0.051	0.049	0.022	0.020	0.874	0.873	0.700	0.881
λ_{80}	1.060	0.006	0.001	6.2E-4	-1.8E-4	0.064	0.063	0.024	0.023	0.901	0.911	0.683	0.883
λ_{81}	1.066	0.001	-0.006	0.002	7.4E-4	0.067	0.061	0.024	0.022	0.875	0.884	0.679	0.882
λ_{82}	1.397	0.007	-0.007	0.002	2.2E-4	0.089	0.080	0.031	0.029	0.895	0.899	0.691	0.880
λ_{83}	1.527	-0.004	-0.031	3.4E-4	-0.001	0.096	0.090	0.033	0.031	0.876	0.836	0.692	0.895
λ_{84}	1.379	-0.009	-0.036	4.6E-5	-3.8E-4	0.092	0.090	0.029	0.028	0.858	0.832	0.706	0.900
λ_{85}	1.343	-0.007	-0.030	0.002	4.3E-4	0.087	0.082	0.029	0.027	0.873	0.853	0.703	0.887
λ_{86}	1.339	0.002	-0.017	0.002	5.1E-4	0.088	0.079	0.028	0.026	0.892	0.885	0.702	0.904
λ_{87}	1.304	-0.006	-0.021	0.002	3.2E-4	0.083	0.075	0.029	0.027	0.884	0.880	0.692	0.890
λ_{88}	1.285	8.8E-4	-0.009	0.002	-3.9E-4	0.083	0.073	0.028	0.026	0.891	0.896	0.698	0.888
λ_{89}	1.260	0.005	-2.8E-4	0.003	-2.2E-4	0.087	0.076	0.029	0.028	0.901	0.914	0.693	0.881
λ_{90}	1.405	0.009	0.004	0.002	2.6E-4	0.089	0.078	0.032	0.029	0.896	0.914	0.693	0.887
λ_{91}	1.513	0.011	0.009	0.003	-4.8E-4	0.096	0.087	0.035	0.033	0.895	0.905	0.712	0.883
λ_{92}	1.715	0.014	0.011	0.002	-9.9E-5	0.103	0.094	0.040	0.036	0.886	0.915	0.693	0.894

Notes: Results derived from conducting 1,000 replications of the Baker and Solon (2003) model with a cohort sample size $n_b = 400$. For each parameter, it reports the values of the average bias, the root-mean square error (RMSE), and coverage probabilities of the 90% confidence intervals, under alternative weighting regimes (equally-weighted, EW, diagonally-weighted, DW, optimally-weighted, OW, and cross-fitted GLasso-weighted, GW).