cemmap
centre for microdata methods and practice

How to weight in moments matchings: A new approach and applications to earnings dynamics

Xu Cheng
Alejandro Sánchez-Becerra
Andrew Shephard

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP13/23

UKRI Economic and Social Research Council

How to Weight in Moments Matching: A New Approach and Applications to Earnings Dynamics*

Xu Cheng† Alejandro Sánchez-Becerra‡ Andrew Shephard§

This Version: June 25, 2023

Abstract

Following the seminal paper by Altonji and Segal (1996), empirical studies have widely embraced equal or diagonal weighting in minimum distance estimation to mitigate the finite-sample bias caused by sampling errors in the weighting matrix. This paper introduces a new weighting scheme that combines cross-fitting and regularized weighting matrix estimation. We also provide a new cross-fitting standard error, applying cross-fitting to estimate the asymptotic variance. In a many-moment asymptotic framework, we demonstrate the effectiveness of cross-fitting in eliminating a first-order asymptotic bias due to weighting matrix sampling errors. Additionally, we demonstrate that some economic models in the earnings dynamics literature meet certain sparsity conditions, ensuring that the proposed regularized weighting matrix behaves similarly to the oracle weighting matrix for these applications. Extensive simulation studies based on the earnings dynamics literature validate the superiority of our approach over commonly employed alternative weighting schemes.

Keywords: cross fitting, covariance structure model, earnings dynamics, graphical lasso, many moments, minimum distance estimation, weighting matrix

JEL Codes: C01, C13, C18

*We thank Natasha Gandhi for her excellent research assistance. We are grateful to numerous seminar and conference participants for their valuable comments and questions.

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

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

§University of Pennsylvania. E-mail: asheph@econ.upenn.edu.

1 Introduction

Minimum distance estimation is a popular approach to estimate structural econometric models through moments matching. For example, labor economists posit different models of earnings dynamics and estimate them by minimizing the weighted distance between the sample cross-period covariance and model implied population covariance.[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 result in substantial finite-sample bias in the minimum distance estimator, as documented by Altonji and Segal (1996) and others.

This paper proposes a new approach to weight the moments, aiming for better finite-sample estimation and inference of the structural parameters. This new approach employs a combination of cross-fitting estimation and regularized weighting matrix estimation. We also suggest using cross-fitting to estimate the asymptotic variance of the minimum distance 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 provide a better approximation to the finite-sample bias due to sampling errors in the $p \times p$ dimensional weighting matrix. To accommodate applications in the earnings dynamics literature, we focus on the case $n, p \to \infty$ and $p/n \to 0$ such that the number of moments is large and it is substantially smaller than the sample size. We compare the proposed method with common alternative weighting schemes and demonstrate its desirable theoretical properties and excellent finite-sample performances.

Cross-fitting uses independent data splits to compute the weighting matrix and the sample moments, then ensembles these sample-splitting estimators to achieve the efficiency of a full sample estimator. 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 minimum distance estimator. Specifically, the full sample estimator has a substantial first order asymptotic bias unless the weighting matrix converges fast enough. In contrast, this asymptotic bias is eliminated by the cross-fitting procedure without any requirement on the rate of convergence of the weighting matrix. To capture this difference in the first order asymptotic bias between the full sam-

[1] See Abowd and Card (1989), MaCurdy (1982), Meghir and Pistaferri (2004), Guvenen (2007), and Altonji, Smith, and Vidangos (2013) for some examples on 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).

ple estimator and the cross-fitting estimator, it is essential to let p grow along with n. In a standard fixed p asymptotic setup, this difference disappears. The theoretical advantages of the cross-fitting estimator suggest that it is more robust against sampling errors in the weighting matrix, and this is reflected in the performance of cross-fitting in our simulation exercises.

Cross-fitting also yields significant benefits when applied to estimate the asymptotic variance of the minimum distance estimator. Theoretically, the cross-fitting estimator and the full sample estimator have the same asymptotic variance. Therefore, one can construct standard errors using either the full sample method or the cross-fitting method. However, their finite-sample performances differ considerably. In simulations, we find that confidence intervals based on cross-fitting standard errors perform well, but those based on full sample standard errors often undercover severely. Indeed, cross-fitting minimum distance estimation together with the optimal weighting matrix (inverse sample covariance) was considered by Altonji and Segal (1996) as their independently weighted optimal minimum distance estimator (IWOMD). However, they used the full sample method to calculate the standard error, which resulted in poor coverage probabilities. Applying the cross-fitting standard error results in a remarkable improvement. Nevertheless, we suggest combining cross-fitting with a regularized weighting matrix, rather than this optimal weighting matrix.

Regularized weighting matrix estimation could be viewed as a data-dependent extension of the extreme regularization achieved by equal weighting and diagonal weighting. All methods control 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 estimate them together with the diagonal elements. The degree of regularization is data driven.

We show that in several examples motivated by the earnings dynamics literature, such as the covariance structure model, the oracle weighting matrix is sparse, i.e., it only has a small number of non-zero off-diagonal elements. This oracle weighting matrix is defined as the inverse of the true population covariance of the moments, and the sparse pattern is an implication of the economic model. While the oracle weighting matrix delivers the smallest variance for the minimum distance estimator, it is infeasible. Under the sparsity condition shown in these examples, the GLasso weighting matrix is a consistent estimator of the oracle weighting matrix. In consequence, the cross-fitting minimum distance estimator based on the GLasso weighting matrix is asymptotically unbiased and achieves the same level of efficiency as one with oracle weighting. This is our recommended estimator.

We investigate the finite-sample properties of our proposed estimator with two sets of

simulation studies. 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 coverage probability of the confidence interval. We also show that the different estimators have an important impact when we replicate an earnings inequality decomposition exercise from Baker and Solon (2003).

Relation to the Literature. Altonji and Segal (1996) investigate the small-sample bias of the optimal minimum distance estimator due to the correlation between sampling errors in the weighting matrix and sample errors in the moments. 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. Both papers contain references to earlier research that report difficulties in empirical applications of the optimal minimum distance estimator, particularly in covariance structure models. This paper provides a new solution to this long-existing problem and investigates it both in theory and in simulation.

The many-moment asymptotic framework where p increases with n has been used to study many instrumental variables (e.g., Bekker, 1994).[3] Han and Phillips (2006) and Newey and Windmeijer (2009) study an asymptotic bias due to a large number of moment conditions in a generalized method of moments (GMM) framework, when the weighting matrix is non-random. This source of bias is irrelevant for a minimum distance estimator even with many moments. Estimation based on many moments typically requires p to be much smaller than n.[4] One notable exception is Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018), who suggest a new regularized minimum distance 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, the question in Altonji and Segal (1996) and here.

This cross-fitting minimum distance estimator is related to the double/debiased machine learning methods, where cross-fitting is applied to attenuate the over-fitting bias

[3]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).

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

after estimating a high-dimensional function by machine learning methods, see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for a review on double/debiased machine learning and references therein for other applications of sample-splitting methods in machine learning and semi-parametric estimation. Here our nuisance parameter is a large scale weighting matrix. Cross-fitting relaxes regularity conditions in both cases but operates through distinct channels.[5]

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 method works under certain notions of sparsity. To link these machine learning methods to structural economic applications, it is crucial to demonstrate the economic model satisfies the particular sparsity condition required by the chosen method. We make considerable efforts in this direction.

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 supply (Abowd and Card, 1989), consumption and savings (Gourinchas and Parker, 2002), and portfolio choice behaviour (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 rest 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 is based on the design in Altonji and Segal (1996) and one is based on a fully-fledged empirical model from Baker and Solon (2003). Section 6 concludes. The appendix contains proofs and additional supplementary materials.

[5]Another important method in double/debiased machine learning is using moments that satisfy the Neyman orthogonality condition. This condition is automatically satisfied once we view the weighting matrix as a nuisance parameter in a minimum distance estimation problem.

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 the i.i.d. observed data $m_i : i = 1, \ldots, n$, and some known function $f(\theta) : \Theta \to \mathbb{R}^p$. We can estimate θ_0 by the minimum distance 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 is possibly data-dependent. As we consider models that are over-identified, the weighting matrix plays a crucial role in the asymptotic and finite-sample properties of this minimum distance estimator.[6] 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 practice, commonly used weighting matrices include (i) equal weighting, with $\widehat{W}$ the identity matrix; (ii) diagonal weighting, with $\widehat{W}$ only taking 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 the inverse covariance weighting is also called the 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 large bias in the minimum distance estimates.

We propose two channels to reduce bias in the minimum distance estimator through weighting matrix choices. The first channel is cross-fitting, a sample splitting method that ensures independence of the weighting matrix and the sample moments by construction. We also provide a cross-fitting estimator of the asymptotic variance of the minimum distance estimator. In the many-moment framework 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 estimation of the weighting matrix. Exploiting the sparse patterns 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 suggest 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.

[6] See Chamberlain (1984) for a detailed analysis of the minimum distance estimator in a standard framework.

2.1 Cross Fitting and Bias Reduction

For a given weighting matrix $\widehat{W}$, we propose to estimate θ through cross-fitting and calculate the standard error of this estimator using the cross-fitting formula provided below. We first describe the cross-fitting procedure and provide some heuristic arguments on why it reduces the first-order asymptotic bias of the minimum distance estimator in the many-moments framework. A specific regularized weighting matrix $\widehat{W}$ is provided in the following subsection.

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

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 minimum distance estimator is

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

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

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

Cross-Fitting Standard Error. One can show that the asymptotic variance of the cross-fitting estimator is the usual sandwich formula $\Omega = (F'WF)^{-1} F'W\Sigma WF(F'WF)$, 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-fitting estimator

$$\widehat{\Omega}^{*} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\Omega}^{(k)}, \text{ where } \widehat{\Omega}^{(k)} = (\widehat{F}_k' \widehat{W}_{-k} \widehat{F}_k)^{-1} \widehat{F}_k' \widehat{W}_{-k} \widehat{\Sigma}_k \widehat{W}_{-k} \widehat{F}_k (\widehat{F}_k' \widehat{W}_{-k} \widehat{F}_k)^{-1}, \tag{2.4}$$

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

In the literature, it is a standard practice 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 with the full sample. This full sample variance estimator delivers the full sample standard error. Although both methods yield consistent estimation of the asymptotic variance, we show that confidence intervals based on cross-fitting standard errors have significantly

better finite-sample performances than those based on full sample standard errors.

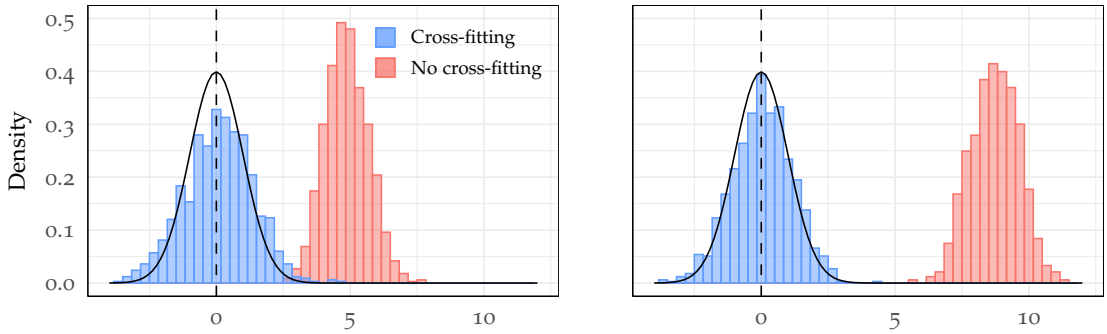
Bias Reduction. Through the lens of a many-moment framework, we now heuristically illustrate how the cross-fitting method reduces estimation bias due to weighting matrix estimation noise. Consider the linear model $f(\theta) = F\theta$ with $m_i \sim \mathcal{N}(f(\theta_0), \Sigma)$. We assume $||\widehat{W} - W|| \to_p 0$ for some non-random matrix W. The full-sample minimum distance estimator is

$$\sqrt{n}(\widehat{\theta} - \theta_0) = (F'\widehat{W}F)^{-1} \left[\underbrace{F'W\sqrt{n}\overline{g}(\theta_0)}_{A} + \underbrace{F'(\widehat{W} - W)\sqrt{n}\overline{g}(\theta_0)}_{B} \right], \tag{2.5}$$

where $\sqrt{n}\overline{g}(\theta_0) = n^{-1/2}\sum_{i=1}^{n}(m_i - f(\theta_0)) \sim \mathcal{N}(0, \Sigma)$. The first term, denoted by A, is based on the non-random limit W, and it 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-fitting estimator in the many-moment case where $p \to \infty$. For the cross-fitting estimator, we always have $B \to_p 0$ because $\widehat{W} - W$ and $\sqrt{n}\overline{g}(\theta_0)$ 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 $\sqrt{n}\overline{g}$. As this correlation effect accumulates across moments, it results in a non-zero limit when the dimension p is high.

To illustrate this bias, consider the linear example above with the parameter θ being a scalar, $\Sigma = I_p$, and $F = (1, 0, \ldots, 0)' \in \mathbb{R}^p$. The weighting matrix estimator is $\widehat{W} = W^O + \Delta$. The first column and first row of Δ are $c_0 p^{-1}\sqrt{n}\overline{g}(\theta_0)$ and its transpose, respectively, for some constant c_0, and all the other elements of Δ are zero. The weighting matrix estimator $\widehat{W}$ is consistent because $||\Delta|| = O_p(p^{-1/2})$ with $p \to \infty$, where $||\cdot||$ denotes the spectral norm of a matrix. In this case, the bias term $B = c_0 p^{-1}||\sqrt{n}\overline{g}(\theta_0)||^2 \to_p c_0$. For the bias term B to converge to 0 in probability for the full sample estimator, the weighting matrix has to converge faster than $p^{-1/2}$. In our design, it converges at exactly $p^{-1/2}$ rate for $c_0 \neq 0$. In (2.5), the leading term A follows a standard normal distribution and the normalization factor $(F'\widehat{W}F)^{-1} \to_p 1$. We conduct a Monte Carlo simulation based on this example with p being the nearest integer to $\sqrt{n}$ and $c_0 = 10$. Figure 1 presents histograms of the t-statistic for two methods: (i) full sample estimator coupled with the full sample standard error and (ii) cross-fitting estimator coupled with the cross-fitting standard error with $K = 2$. This figure confirms that the full sample estimator is biased and the bias does not disappear with a large sample. The cross-fitting method eliminates



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

Figure 1: Bias reduction under cross-fitting. 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-fitting estimator with the cross-fitting standard error. Black line is the standard normal density. See the text for a description of the many-moment simulation design.

the bias effectively.

In sum, in the many-moment framework where $p \to \infty$, the cross-fitting estimator is more robust than the full sample estimator. The cross-fitting 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, could have a substantial bias. This provides a theoretical justification for our observation that the cross-fitting 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. These studies postulate different notions of sparsity, and use various shrinkage methods to achieve dimension reduction and improved estimation accuracy. Therefore, before suggesting a specific regularized estimator for W^O, we first present some empirical examples of interest and demonstrate the sparsity patterns in them. Here we define the sparsity measure as the number of non-zero off-diagonal elements in the $p \times p$ dimensional matrix W^O relative to the number of moments p. Although the total number of off-diagonal elements increases at the rate p^2, we show that most of them are zeros in these examples and that the number of non-zero elements increases linearly with p.

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 such a setting, the moment m_i may comprise means or autocovariances over time for individual i. In our examples

here, we provide very stylised illustrations to show how the time series properties of the moments yield the desired sparse structure. We also illustrate that zero partial correlation among moments from the same time period could also 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: $x_{i,t} = \mu_t + \rho(x_{i,t-1} - \mu_t) + u_{i,t}$, where μ_t is the fixed effect, $u_{i,t}$ 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 auto-correlation is ρ^{t-s} for periods t and s. However, the oracle weighting matrix $W^O = [\text{Var}(m_i)]^{-1}$ is sparse with a band-diagonal structure

$$W^O = \sigma_u^{-2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & (1+\rho^2) & -\rho & \cdots & 0 \\ 0 & -\rho & (1+\rho^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{2.6}$$

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

$$W^O = \sigma_u^{-2} \begin{pmatrix} \frac{\sigma_0^2+2\sigma_u^2}{\sigma_0^2+\sigma_u^2} & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \tag{2.7}$$

□

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 $x_{i,t}$ is determined by a combination of a time fixed effect μ_t and a weighted average of past shocks: $x_{i,t} = \mu_t + \sum_{s=1}^{t} a_{t,s}' u_{i,s}$, where $a_{t,s} \in \mathbb{R}^M$ is a vector of loadings and $u_{i,s} \in \mathbb{R}^M$

is a vector of mean zero independent shocks.[7]

Let $X_i = (x_{i,1}, \ldots, x_{i,T})' \in \mathbb{R}^T$ and $U_i = (u'_{i,1}, \ldots, u'_{i,T})' \in \mathbb{R}^{MT}$ respectively denote vectors of individual time series and sequence of shocks. Similarly, let μ denote the vector of time fixed effects. The time series process can be represented in a matrix form as $X_i = \mu + AU_i$, where A is a $T \times MT$ coefficient matrix with block entries a'_{ts}. The autocovariance structure of X_i is determined by that of AU_i.[8] Let $\text{vec}(\cdot)$ and $\text{vech}(\cdot)$ denote the vectorization and half-vectorization of a symmetric matrix, with the selector matrix Γ converting the vectorization to its half-vectorization and the selector matrix Γ_* converting the half-vectorization to the vectorization. To match the autocovariance structure of X_i, we have

$$m_i = \text{vech}(AU_iU_i'A') = \Gamma(A \otimes A)\,\text{vec}(U_iU_i') = \Gamma(A \otimes A)\Gamma_*\,\text{vech}(U_iU_i'). \tag{2.8}$$

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

$$W^O = [\Gamma(A \otimes A)\Gamma_*\,\text{Var}(\text{vech}(U_iU_i'))\Gamma_*'(A' \otimes A')\Gamma]^{-1}, \tag{2.9}$$

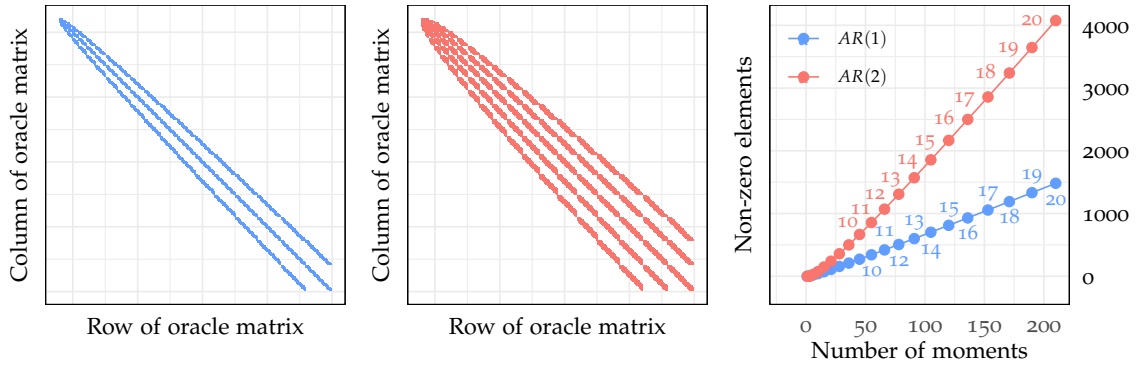
where A is lower triangular because the time series only depends on past shocks. Given that $u_{i,t}$ are independent across t, the matrix $\text{Var}(\text{vech}(U_iU_i'))$ has a sparse structure, and it is diagonal in the $M = 1$ case.

Although it is difficult to study the sparsity pattern of W^O in (2.9) analytically for an arbitrary matrix A, it is easy to compute this formula for specific dynamic processes to confirm that only a small fraction of its off-diagonal elements are non-zero. 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 sparse patterns of the oracle weighting matrices. Panel (a) and (b) plot the non-zero elements of the oracle weighting matrix for $AR(1)$ and $AR(2)$ processes, respectively, for the case $T = 20$. 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 $p-1$

[7] For example, suppose $M = 1$ for a single shock and $x_{i,t} - \mu_t$ follows a stationary $AR(1)$ process with dependence parameter ρ. Write $x_{i,1} = u_{i,1}$, where $u_{i,1}$ is drawn from the stationary distribution of this $AR(1)$ process. For $t \geq 2$, we have $x_{i,t} = \mu_t + a_{t,t-1}u_{i,t-1} + a_{t,t}u_{i,t}$, where $a_{t,t-1} = \rho$ and $a_{t,t} = 1$.

[8] 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 B.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. Panel (a) and Panel (b) respectively show the non-zero elements of the oracle weighting matrix when $T = 20$ with an $AR(1)$ and $AR(2)$ process. 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 is varied.

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 rotating different elements in m_i to take the role of y_i. As such, W^O is sparse if there are many pairwise zero partial correlation among these moments.

Graphical models view each element of m_i as a vertex and encode the partial correlation between them as edges. This graphical relationship among m_i is entirely captured by the oracle weighting matrix. Sparse graphical models could be generated by the time structure in Example 1 and Example 2 as well as conditional independence among other types of spatial or network relationships. □

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) and 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 specificially, we adopt the correlation-based version suggested by Rothman, Bickel, Levina, and Zhu (2008), 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 worked particularly well for us as a weighting matrix. Moreover, it automat-

ically 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.

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

$$\widehat{Q}_G = \underset{Q \in \mathcal{W}}{\arg\max}\ \log(\det(Q)) - \text{tr}(Q\widehat{R}) - \lambda \sum_{j \neq j'} |Q_{jj'}|, \tag{2.10}$$

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

$$\widehat{W}_G = \widehat{D}^{-1}\widehat{Q}_G\widehat{D}^{-1}. \tag{2.11}$$

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 (2.10) is equal to the log-likelihood of Q for a normal distribution plus a penalty for the correlation coefficient in the off-diagonal elements. 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 B.2. To obtain the GLasso estimation in (2.10), 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.

Rothman, Bickel, Levina, and Zhu (2008) show that the GLasso weighting matrix above is a consistent estimator of the oracle weighting matrix under the key condition that the number of non-zero off-diagonal elements, denoted by s, is much smaller than the sample size n. This condition holds in the examples studied above, where s increases proportionally with p, both are at a slower rate than n. Lemma 2.1 below restates this result. Let $\varepsilon_i = m_i - \mathbb{E}[m_i]$. Let c_0, c_1, c_2, c, C, δ denote some constants.

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, and $A_{r\ell}$ denote its element in row r column ℓ.

Lemma 2.1. $||\widehat{W}_G - W^O|| = O_p(\sqrt{(s+1)n^{-1}\log p})$ *given the following conditions: (i) the tuning parameter satisfies* $\lambda = c_0(n^{-1}\log p)^{1/2}$*; (ii)* $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$*; (iii)* $P[n^{-1}|\sum_{i=1}^n(\varepsilon_{i,r}\varepsilon_{i,\ell} - \Sigma_{r\ell})| \geq v] \leq c_1\exp(-c_2nv^2)$ *for* $|v| < \delta$.

Rothman, Bickel, Levina, and Zhu (2008) assume m_i are normally distributed, which satisfies the exponential-type tail condition (iii) used in the proof. Ravikumar, Wainwright, Raskutti, and Yu (2011) establish consistency of $\widehat{W}_G$ under more general tail conditions, including polynomial-type tail conditions.

When the GLasso estimator is used to compute the fold k weighting matrix for cross-fitting estimation, we denote it by $\widehat{W}_{G,-k}$, as it is computed with data from $\mathcal{I}_{-k}$. The resulting cross-fitting 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 2.1. As a consequence, the cross-fitting estimator $\widehat{\theta}^*_G$ has the same asymptotic distribution as the oracle estimator based on W^O. We establish this result in Corollary 3.1 below. Algorithm 1 summarizes the steps to compute this cross-fitting estimator $\widehat{\theta}^*_G$ and the cross-fitting estimator of its asymptotic variance denoted by $\widehat{\Omega}^*_G$. For numerical results in Section 4 and Section 5, we use $K = 2$.[9]

Algorithm 1: Cross-Fitting 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 (2.4);
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 (2.10) and (2.11). ; `/* 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 (2.4) with $\widehat{\theta}^{(k)} = \widehat{\theta}^{(k)}_G$ and $\widehat{W}_{-k} = \widehat{W}_{G,-k}$
end
$\widehat{\theta}^*_G \leftarrow K^{-1}\sum_{k=1}^K \widehat{\theta}^{(k)}_G$, $\quad \widehat{\Omega}^*_G \leftarrow K^{-1}\sum_{k=1}^K \widehat{\Omega}^{(k)}_G$.

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

[9] Sampling errors in the weighting matrix and in the moments are about the same asymptotic order when s is proportional to p, see Lemma 2.1 and Lemma A.1(a).

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. 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-fitting minimum distance estimator $\widehat{\theta}^*$, defined in (2.3), with a convergent weighting matrix. A special case is the recommend estimator $\widehat{\theta}^*_G$ based on the GLasso weighting matrix. We also show consistency of the cross-fitting variance estimator defined in (2.4). 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-fitting estimator as long as it holds for the full sample estimator. The asymptotic theory below does not distinguish between two cross-fitting 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_r and θ_ℓ is denoted by $f_{\theta\theta,r\ell}(\theta) \in \mathbb{R}^{p \times 1}$. 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 $f(\theta_0) = \mathbb{E}[m_i]$.

Assumption R. (i) $||f_\theta(\theta)|| \leq C$ for any $\theta \in \Theta$. (ii) $||f_{\theta\theta,rl}(\theta)|| \leq C$ for any $||\theta - \theta_0|| \leq \delta$ for some $\delta > 0$. (iii) $\lambda_{\min}(F'F) \geq c$. (iv) $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$.

Assumptions ID and R ensure strong identification of θ_0. 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-fitting 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 GLasso weighted estimator with $W = W^O$ by Lemma 2.1, 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 size n. For the cross-fitting 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 3.1. *Suppose Conditions (i)–(iii) of Lemma 2.1 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-fitting variance estimator $\widehat{\Omega}^*$ in (2.4) 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 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 automatically applies to $\widehat{\Omega}^*_G$ in Algorithm 1 for the cross-fitting GLasso-weighted estimator under conditions in Lemma 2.1, because $\widehat{\Omega}^*_G$ is a special case of $\widehat{\Omega}^*$. Note that it is different 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.

Finally, we consider a simple extension of the regularity condition presented in Assumption R 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 below. 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 R to Assumption R^+ as follows. The constant a_n in Assumption R^+ is $\sqrt{p}$ in this example.

Assumption R^+**.** For some sequence of constants a_n that satisfies $a_n \to \infty$ and $a_n = O(p^{1/2})$, 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 R holds with $f_\theta(\theta)$, F, and $f_{\theta\theta,rl}(\theta)$ replaced by $f^+_\theta(\theta)$, and F^+, $f_{\theta\theta,rl}(\theta)^+$, respectively.

Theorem 3.3. *Suppose Assumption R is replaced with Assumption R^+ in Theorem 3.1, Corollary 3.1, and Theorem 3.2. Then, Theorem 3.1, Corollary 3.1, 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 3.3 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.

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, denoted as $i = 1, \dots, n$, 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 $t \in \{1, \dots, T\}$, 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 Simulation Results

We replicate the analysis of Altonji and Segal (1996) by considering nine different distributions for $x_{i,t}$,[10] 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 the cross-fitting GLasso-weighted (GW) minimum distance estimator proposed in this paper. Its confidence interval is based on the cross-fitting standard error proposed in this paper.

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)

[10] 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 (obtained as an equally-weighted mixture of two unit variance normally distributed random variables, with means -2 and 2).

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 all the other estimators considered, which assign different weights to the sample variances from the different time periods. We are therefore interested in how the alternative estimators perform relative to the EW estimator.

Table 1: Alton 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	0.001	-0.123	-0.124	0.009	0.086	0.142	0.145	0.122	0.873	0.309	0.292	0.843
t(10)	100	-0.001	-0.063	-0.064	0.011	0.056	0.086	0.087	0.074	0.885	0.579	0.556	0.870
t(15)	100	-0.000	-0.052	-0.053	0.010	0.052	0.076	0.078	0.066	0.883	0.646	0.615	0.861
Normal	100	-0.000	-0.036	-0.037	0.009	0.045	0.060	0.062	0.055	0.895	0.733	0.707	0.878
Uniform	100	-0.000	-0.006	-0.006	0.010	0.030	0.031	0.032	0.032	0.895	0.876	0.848	0.875
Log normal	100	0.015	-0.473	-0.480	0.103	0.357	0.488	0.494	0.886	0.786	0.009	0.007	0.739
Exp	100	0.000	-0.160	-0.162	0.046	0.095	0.186	0.189	0.175	0.882	0.276	0.265	0.864
Half-normal	100	-0.001	-0.063	-0.064	0.022	0.055	0.086	0.088	0.078	0.884	0.575	0.541	0.897
Bimodal	100	-0.001	-0.011	-0.011	0.010	0.028	0.030	0.031	0.031	0.905	0.845	0.832	0.890
t(5)	1000	0.001	-0.025	-0.025	0.002	0.028	0.035	0.035	0.033	0.909	0.658	0.657	0.896
t(10)	1000	0.001	-0.007	-0.007	0.002	0.018	0.019	0.019	0.019	0.886	0.854	0.854	0.881
t(15)	1000	0.000	-0.005	-0.005	0.002	0.016	0.017	0.017	0.017	0.889	0.862	0.859	0.894
Normal	1000	0.000	-0.003	-0.003	0.001	0.013	0.014	0.014	0.014	0.906	0.896	0.899	0.899
Uniform	1000	-0.001	-0.001	-0.001	0.000	0.009	0.009	0.009	0.009	0.914	0.904	0.903	0.920
Log normal	1000	0.005	-0.157	-0.157	0.015	0.122	0.171	0.171	0.178	0.848	0.147	0.143	0.807
Exp	1000	-0.000	-0.024	-0.024	0.002	0.028	0.038	0.038	0.033	0.892	0.720	0.718	0.881
Half-normal	1000	0.001	-0.006	-0.006	0.003	0.017	0.018	0.018	0.018	0.884	0.849	0.848	0.883
Bimodal	1000	-0.000	-0.001	-0.001	0.001	0.009	0.009	0.009	0.009	0.900	0.892	0.891	0.888

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-fitting GLasso-weighted, GW). EW is the oracle benchmark.

Firstly, both the DW and OW estimators have comparatively similar performance to each other, but with both exhibiting non-negligible negative bias. The bias is largest in the case of the student-$t(5)$ distribution (which is thick-tailed and symmetric) and for log normal and exponential distributions (longer-tailed and skewed). In addition to the bias, the root-mean squared errors are larger relative to EW and the coverage probabilities of the 90% confidence intervals are typically far below 0.9, such that inference about the parameter estimates is much less accurate 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 Appendix C 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 the strong performance of GW relative to DW and OW is primarily achieved through the cross-fitting estimation procedure and cross-fitting variance estimation formula in this simulation design.

4.2 Cross-Fitting Standard Error

As part or our procedure, in (2.4) we propose a cross-fitting standard error. It has been applied in calculating the 90% coverage probabilities for 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-fitting standard error in Table 2.

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 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 the cross-fitting standard error is important for reliable inference.

Table 2: Altonji and Segal (1996) Design: Importance of Cross-Fitting Standard Error

		OW			GW		
Distribution	n	No CF	CF-Full	CF-CF	No CF	CF-Full	CF-CF
t(5)	100	0.292	0.568	0.839	0.306	0.597	0.843
t(10)	100	0.556	0.675	0.881	0.579	0.722	0.870
t(15)	100	0.615	0.709	0.869	0.644	0.748	0.861
Normal	100	0.707	0.760	0.881	0.729	0.793	0.878
Uniform	100	0.848	0.820	0.887	0.876	0.858	0.875
Log normal	100	0.007	0.193	0.736	0.009	0.234	0.739
Exp	100	0.265	0.470	0.868	0.275	0.487	0.864
Half-normal	100	0.541	0.648	0.880	0.574	0.689	0.897
Bimodal	100	0.832	0.805	0.883	0.845	0.860	0.890
t(5)	1000	0.657	0.787	0.898	0.658	0.786	0.896
t(10)	1000	0.854	0.852	0.875	0.854	0.851	0.881
t(15)	1000	0.859	0.878	0.893	0.861	0.883	0.894
Normal	1000	0.899	0.897	0.906	0.895	0.892	0.899
Uniform	1000	0.903	0.915	0.921	0.904	0.919	0.920
Log normal	1000	0.143	0.468	0.811	0.146	0.478	0.807
Exp	1000	0.718	0.813	0.882	0.719	0.828	0.881
Half-normal	1000	0.848	0.863	0.884	0.847	0.868	0.883
Bimodal	1000	0.891	0.882	0.888	0.892	0.883	0.888

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-fitting estimator is coupled with the full sample standard error, while CF-CF indicates the use of (2.3) for the estimator and (2.4) for the cross-fitting standard error. $T = 10$ for all cases.

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 data set of yearly tax records.[11] The richness of their data allowed a flexible earnings process to be specified, whose estimated parameter values rejected a number of common restrictions that have been commonly imposed on the covariance structure (such as the absence

[11]See Ostrovsky (2010) for an extension of the Baker and Solon (2003) analysis using data from 1985 to 2005.

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.[12] 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 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 $(\sigma_\alpha^2, \sigma_{\alpha\beta}, \sigma_\beta^2)$. 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_r^2 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

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

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 $v_{ibt} = v_{ibt}^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_b^2)$ in the first year that cohort b is observed in the sample. Note that the variance σ_b^2 is cohort-specific. This captures the fact that they start

[12] All our analysis 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.

at different ages when the sample begins.[13]

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

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

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 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 have 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. 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 C 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

[13] 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 $(t - b - 26)\sigma_r^2$, the distribution of a random walk that has been accumulating since age 26.

comprises 60 parameters, and we provide parameter-level performance statistics for a cohort sample size of 400 in Appendix C. While there are some exceptions, starting with a cohort sample size of 400 the results across the alternative weighting schemes can essentially be summarized as follows: the absolute bias of the parameters is typically highest under DW, followed by EW and OW, and is much lower under GW; the coverage probabilities of the 90% confidence intervals are well-below 0.9 under OW, improve a lot under DW and EW, and are always close to 0.9 under GW; the root-mean squared error is typically highest under DW, followed by EW, OW, and is lowest under GW. Thus, GW typically dominates across the range of estimator performance statistics that we consider here.

The same qualitative patterns exist with larger cohort sample sizes. We visually summarize our results in the violin plot in Figure 3. This shows the distribution of the three performance statistics over the set of model parameters for each of the considered cohort sample sizes. For example, the figure in the first row and first column shows that, for $n_b = 400$, the bias can be as large as 0.3 for the EW estimator and 0.5 for the DW estimator for some of the 60 parameters. In contrast, the GW estimator exhibits a bias of less than 0.025 for all 60 parameters. While the difference across estimators with alternative weighting regimes becomes smaller as the sample size increases, it is still the case that our proposed GW estimator based on cross-fitting and GLasso weighting always performs most favorably. In Appendix C we show the importance of both cross-fitting and GLasso weighting in obtaining these results.

5.4 Variance Decomposition Analysis

We are interested in the extent to which the different estimates obtained under the alternative weighting schemes is 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 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 4. 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

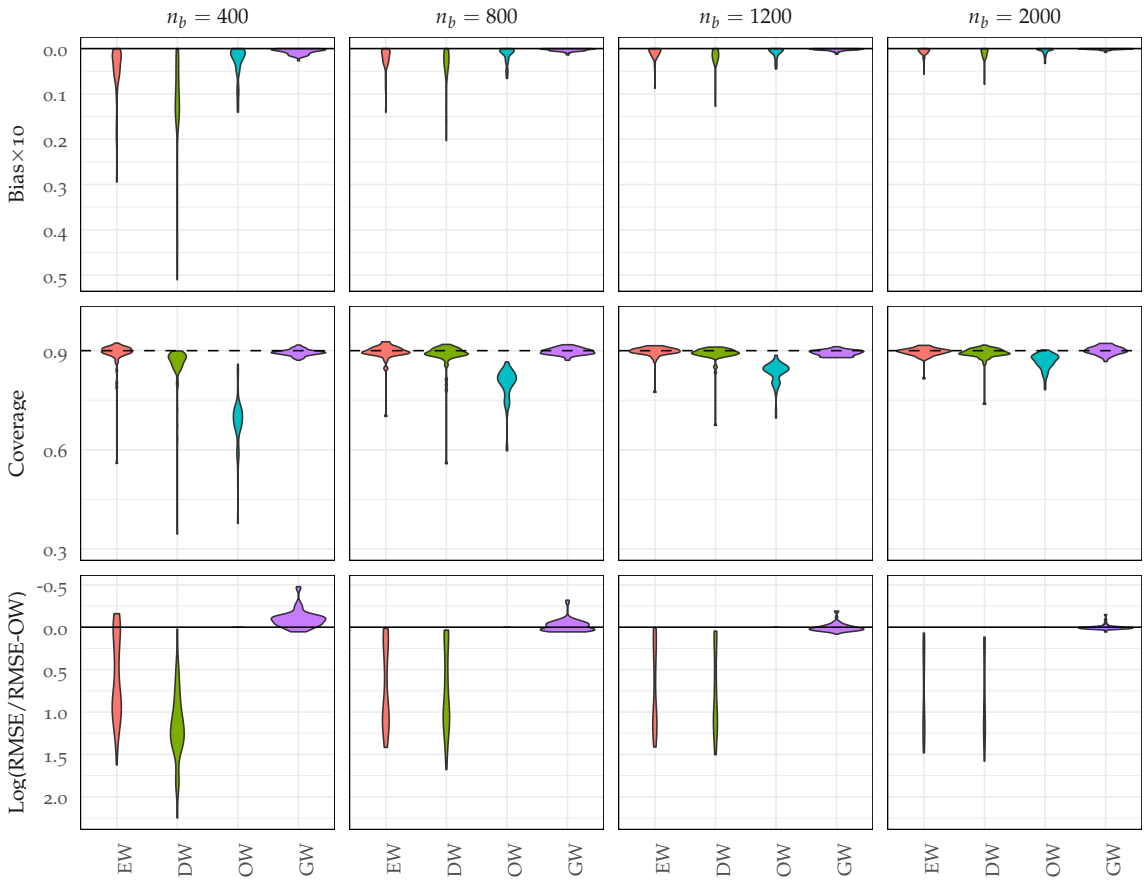


Figure 3: Violin plots for Baker and Solon (2003) parameters, showing absolute bias, 90% confidence interval coverage probabilities, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting. Figure derived from 1,000 replications. Weighting denoted EW (equally-weighted), DW (diagonally-weighted), OW (optimally-weighted), and GW (cross-fitting GLasso-weighted).

5% and 95% quantiles of the estimates obtained with different simulated samples. The broken black lines indicating the true data-generating decomposition.[14] 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

[14] 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.

well: the predicted variance amounts (overall and by persistent/transitory status) almost perfectly coincides with that implied by the true data-generating process. Furthermore, the confidence bands are considerably narrower than were obtained under both EW and DW.[15]

6 Conclusion

In the conclusion of Altonji and Segal (1996), they highlight several desirable features of a future weighting matrix: (i) it is a robust weighting matrix estimator that is superior to the conventional optimal weighting matrix or the independently weighting optimal weighting matrix; (ii) it may incorporate prior information about which sets of moments are likely to be highly correlated to reduce the dimension of the weighting matrix estimation; (iii) it may transition between the equal weighting matrix and the optimal weighting matrix; (iv) it applies to nonlinear models. The proposed weighting method is a modern answer to all of these requests.[16] The regularized weighting matrix adapts to the data to determine which elements of the weighting matrix are estimated. The cross-fitting method also substantially reduces estimation bias. The asymptotic framework allows the number of moments to increase along with the sample size, ensuring the small-sample issue considered by Altonji and Segal (1996) stays relevant in a big data environment. Using simulation designs based on earnings dynamics models, we show that the proposed weighting scheme that combines cross-fitting and regularized weighting matrix estimation compares extremely favorably to popular alternative weighting schemes widely used in the empirical literature.

[15]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.

[16]On feature (iii), our estimator transitions between diagonal weighting, rather than equal weighting, to the optimal weighting.

Figure 4: 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-fitting GLasso-weighted, GW). Shaded regions indicated 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.

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 (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.

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.

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

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

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.

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.

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

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.

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.

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 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.

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

Appendices

A Proofs

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

Lemma A.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

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

By the mean-value expansion,

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

for some $\widetilde{\theta}^{(k)}$ between $\widehat{\theta}^{(k)}$ and θ_0 and thus $\widetilde{\theta}^{(k)} \to_p \theta_0$ by Lemma A.2. The second-order derivative in (A.2) converges in probability to $(F'WF)^{-1}$ by Lemma A.3.

The first-order derivative satisfies

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

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).

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 $\mathcal{I}_{-k}$,

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

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 $\delta > 0$,

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

Let $\mathcal{E} = \{||\widehat{W}_{-k} - W|| \leq \varepsilon\}$ for any give $\varepsilon > 0$. Then by (A.4), (A.5), and the law of iterated expectations, we have

$$
\Pr\left(|B_k| > \delta|\mathcal{E}\right) \leq \frac{C\varepsilon^2}{\delta^2}. \tag{A.6}
$$

This shows $B_k = o_p(1)$ because $Pr(\mathcal{E}) \to 1$. This completes the proof for (A.1).

The cross-fitting estimator satisfies

$$
\begin{aligned}
\sqrt{n}(\widehat{\theta}^* - \theta_0) &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_k}(\widehat{\theta}^{(k)} - \theta_0) \\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(F'WF\right)^{-1}\sqrt{n_k}F'W\overline{g}_k(\theta_0) + o_p(1) \\
&= \left(F'WF\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}F'W\left(m_i - \mathbb{E}[m_i]\right) + o_p(1),
\end{aligned} \tag{A.7}
$$

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 (A.1), and the last equality holds because sample splitting implies

$\sum_{k=1}^{K} n_k \overline{g}_k(\theta_0) = \sum_{i=1}^{n} (m_i - \mathbb{E}[m_i])$. Note that $\xi_i = F'W(m_i - \mathbb{E}[m_i])$ 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. random variables and Slutsky's theorem.

□

Proof of Corollary 3.1 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 (i) $||\widehat{W}_G - W^O|| \to_p 0$ by Lemma 2.1 and (ii) 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

$$\begin{aligned}
&\left\|\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\widehat{F}_k - F'W\Sigma WF\right\| \leq H_1 + H_2, \text{ where} \\
H_1 &= \left\|F'W(\widehat{\Sigma}_k - \Sigma)WF\right\|, \\
H_2 &= \left\|F'\left(\widehat{W}_{-k} - W\right)\widehat{\Sigma}_k WF\right\| + \left\|F'\widehat{W}_{-k}\widehat{\Sigma}_k\left(\widehat{W}_{-k} - W\right)F\right\| + \\
&\quad \left\|\left(\widehat{F}_k - F\right)'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}F\right\| + \left\|\widehat{F}_k'\widehat{W}_{-k}\widehat{\Sigma}_k\widehat{W}_{-k}\left(\widehat{F}_k - F\right)\right\|. \qquad \text{(A.8)}
\end{aligned}$$

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 $\varepsilon_i = m_i - \mathbb{E}(m_i)$ and $\vartheta_i = \varepsilon_i\varepsilon_i' - \mathbb{E}[\varepsilon_i\varepsilon_i']$, we have $\widehat{\Sigma}_k - \Sigma = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\vartheta_i - \overline{\varepsilon}^{(k)}\overline{\varepsilon}^{(k)'}$. Write

$$\begin{aligned}
F'W\left(\widehat{\Sigma}_k - \Sigma\right)WF &= R_1 - R_2, \text{ where} \\
R_1 = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k} F'W\vartheta_i WF, &\quad R_2 = F'W\overline{\varepsilon}^{(k)}\overline{\varepsilon}^{(k)'}WF. \qquad \text{(A.9)}
\end{aligned}$$

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

$$\begin{aligned}
\sum_{r=1}^{d_\theta}\sum_{\ell=1}^{d_\theta}\mathbb{E}[(R_{1,r\ell})^2] &= \frac{1}{n_k}\sum_{r=1}^{d_\theta}\sum_{\ell=1}^{d_\theta}\mathbb{E}\left[\left([F'W\vartheta_i WF]_{r\ell}\right)^2\right] = \frac{1}{n_k}\mathbb{E}[\text{tr}(F'W\vartheta_i WFF'W\vartheta_i WF)] \\
&\leq \frac{1}{n_k}||WFF'W||\,\mathbb{E}[\text{tr}(F'W\vartheta_i\vartheta_i WF)] \\
&\leq \frac{1}{n_k}d_\theta\,\|F\|^4\,\|W\|^4\,\left\|\mathbb{E}[\vartheta_i^2]\right\| \leq C\frac{p}{n}, \qquad \text{(A.10)}
\end{aligned}$$

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 $\text{tr}(\cdot)$, $\text{tr}(A) \leq \text{rank}(A)\lambda_{\max}(A)$, 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 $||R_2|| \leq ||F||^2||W||^2||\bar{\varepsilon}^{(k)}(\theta_0)||^2 = o_p(1)$ by Assumption R(i), W(ii), and Lemma A.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 A.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 A.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 Theorem 3.3. We make the following adjustments to the previous results and proofs, in addition to replacing Assumption R with Assumption R^+.

(i). Lemma A.1. Part (b), C and $O_p(1)$ are replaced by a_n and $O_p(a_n)$, respectively. Part (d) hold with $\widehat{F}_k$ and F replaced by $\widehat{F}_k^+$ and F^+, respectively.

(ii). Lemma A.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 $a_n^{-1}\overline{g}_k(\theta)$. The sample criterion function requires a different normalization. In this proof, we need $a_n^{-1}(\overline{g}_k(\theta) - g(\theta)) = o_p(1)$. By Lemma A.2(a), this condition holds for $p^{1/2}n^{-1/2} = o(a_n)$.

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

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

To prove (A.11), (A.2) is replaced by

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

where the modified second-order derivative continues to satisfy Lemma A.3 and the modified first-order derivative continues to satisfy (A.3), with F replaced by F^+. Following these adjustments, (A.7) becomes

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

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

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

where the equality follows from the definition of Ω^+ and the convergence follows from (A.13). The claim on Corollary 3.1 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

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

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^+. □

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

$$\mathbb{E}\left[\|\overline{g}_k(\theta) - g(\theta)\|^2\right] = \mathbb{E}\left[\left\|\bar{\varepsilon}^{(k)}\right\|^2\right] = \frac{1}{n_k}\sum_{r=1}^{p}\mathbb{E}\left[\varepsilon_{i,r}^2\right] \leq C\frac{p}{n}, \tag{A.16}$$

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

$$\sup_{\theta\in\Theta}\|g(\theta)\| \leq \sup_{\theta\in\Theta}\left\|f_\theta(\widetilde{\theta})\right\| \sup_{\theta\in\Theta}\|\theta - \theta_0\| \leq C \tag{A.17}$$

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

$$\left\|\widehat{F}_k - F\right\| = \left\|f_\theta(\widehat{\theta}_k) - f_\theta(\theta_0)\right\| \leq C\left\|\widehat{\theta}_k - \theta_0\right\| = o_p(1) \tag{A.18}$$

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

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

$$\begin{aligned} 2\sup_{\theta\in\Theta}|Q_{nk}(\theta) - Q(\theta)| &= \left|\overline{g}_k'\widehat{W}_{-k}\overline{g}_k(\theta) - g(\theta)'Wg(\theta)\right| \\ &\leq \left|(\overline{g}_k(\theta) + g(\theta))'\widehat{W}_{-k}\left(\overline{g}_k(\theta) - g(\theta)\right)\right| + \left|\overline{g}_k(\theta)'\left(\widehat{W}_{-k} - W\right)g(\theta)\right|, \end{aligned} \tag{A.19}$$

which converges to 0 in probability by Lemma A.1(a) – (c). Because Θ is compact and $f(\theta)$ is continuous, Assumption ID implies $\inf_{||\theta-\theta_0||\geq\varepsilon} Q(\theta) > 0$ 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 A.3 Row r and column ℓ of the left hand side is

$$\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_{nk}(\widetilde{\theta}^{(k)})\right]_{r\ell} = \left(\frac{\partial}{\partial\theta_r}f(\widetilde{\theta}^{(k)})\right)'\widehat{W}_{-k}\left(\frac{\partial}{\partial\theta_\ell}f(\widetilde{\theta}^{(k)})\right) - \frac{\partial^2}{\partial\theta_r\partial\theta_\ell}f(\widetilde{\theta}^{(k)})'\widehat{W}_{-k}\overline{g}_k(\widetilde{\theta}^{(k)}). \tag{A.20}$$

The second term on the right hand side of (A.20) is negligible because

$$\left\|\frac{\partial^2}{\partial\theta_k\partial\theta_\ell}f(\widetilde{\theta})'\widehat{W}_{-k}\overline{g}(\widetilde{\theta}^{(k)})\right\| \leq C\left\|\widehat{W}_{-k}\right\|\left\|\overline{g}(\widetilde{\theta}^{(k)})\right\| = o_p(1) \tag{A.21}$$

by Assumption R(ii), and Lemma A.1(a) – (c). The first term on the right hand side of (A.20) satisfies

$$\left\|\frac{\partial}{\partial\theta_r}f(\widetilde{\theta}^{(k)}) - \frac{\partial}{\partial\theta_r}f(\theta_0)\right\| \leq C\left\|\widetilde{\theta}^{(k)} - \theta_0\right\| = o_p(1) \tag{A.22}$$

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

□

B Additional Implementation Details

B.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 $\widetilde{m}_i = \text{vech}((X_i - \overline{X})(X_i - \overline{X}')$ and with $\text{vech}(\cdot)$ denoting 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 $m_i = \text{vech}(X_i - \mathbb{E}[X_i])(X_i - \mathbb{E}[X_i])'$. Comparing $\widetilde{m}$ and $\overline{m}$, the differences between $\overline{X}$ and $\mathbb{E}[X_i]$ in their centering term and the difference between $n-1$ and n in their normalization are negligible asymptotically to study 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 the asymptotic distribution of the minimum distance estimator with $\overline{m}$.

B.2 Tuning Parameter for GLasso

In practice, we select the tuning parameter λ for the GLasso estimator by cross-validation. For the cross-fitting estimator, $\widehat{W}_{G,-k}$ is computed with data $i \in \mathcal{I}_{-k}$. In this case, we further divide data in $\mathcal{I}_{-k}$ to L folds to choose λ for the computation of $\widehat{W}_{G,-k}$. We use $L = 10$ for the tuning parameter choice in all cases.

The cross validation procedure is conducted 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 as in (2.10), so that $-\mathcal{L}(\widehat{\Sigma}, W)$ is the loss function. For a given λ, obtain the GLasso estimator $\widehat{W}_{-\ell}(\lambda)$ following (2.10) and (2.11), for $\widehat{\Sigma} = \widehat{\Sigma}_{-\ell}$ 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:

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

In practice, we 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.

C Additional Simulation Results

C.1 High-Dimensional Extension of Altonji and Segal (1996)

In Table C.1 we present the full results from the Altonji and Segal (1996) simulations in a high-dimensional setting where the panel time dimension increases simultaneously with the cross-sectional dimension, $T = 0.2n$. As in the main text, for each of the nine distributions considered, we present the average bias, the root-mean square error, and the coverage probabilities of the 90% confidence intervals, as the sample size is varied ($n = 100$ and $n = 1000$). Separate results are presented under the alternative weighting regimes that we considered (equally-weighted, diagonally-weighted, optimally-weighted, and cross-fitting GLasso-weighted).

Table C.1: 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	0.001	-0.123	-0.124	0.009	0.086	0.142	0.145	0.122	0.873	0.309	0.292	0.843
t(10)	100	-0.001	-0.063	-0.064	0.011	0.056	0.086	0.087	0.074	0.885	0.579	0.556	0.870
t(15)	100	-0.000	-0.052	-0.053	0.010	0.052	0.076	0.078	0.066	0.883	0.646	0.615	0.861
Normal	100	-0.000	-0.036	-0.037	0.009	0.045	0.060	0.062	0.055	0.895	0.733	0.707	0.878
Uniform	100	-0.000	-0.006	-0.006	0.010	0.030	0.031	0.032	0.032	0.895	0.876	0.848	0.875
Log normal	100	0.015	-0.473	-0.480	0.103	0.357	0.488	0.494	0.886	0.786	0.009	0.007	0.739
Exp	100	0.000	-0.160	-0.162	0.046	0.095	0.186	0.189	0.175	0.882	0.276	0.265	0.864
Half-normal	100	-0.001	-0.063	-0.064	0.022	0.055	0.086	0.088	0.078	0.884	0.575	0.541	0.897
Bimodal	100	-0.001	-0.011	-0.011	0.010	0.028	0.030	0.031	0.031	0.905	0.845	0.832	0.890
t(5)	1000	0.001	-0.025	-0.025	0.002	0.028	0.035	0.035	0.033	0.909	0.658	0.657	0.896
t(10)	1000	0.001	-0.007	-0.007	0.002	0.018	0.019	0.019	0.019	0.886	0.854	0.854	0.881
t(15)	1000	0.000	-0.005	-0.005	0.002	0.016	0.017	0.017	0.017	0.889	0.862	0.859	0.894
Normal	1000	0.000	-0.003	-0.003	0.001	0.013	0.014	0.014	0.014	0.906	0.896	0.899	0.899
Uniform	1000	-0.001	-0.001	-0.001	0.000	0.009	0.009	0.009	0.009	0.914	0.904	0.903	0.920
Log normal	1000	0.005	-0.157	-0.157	0.015	0.122	0.171	0.171	0.178	0.848	0.147	0.143	0.807
Exp	1000	-0.000	-0.024	-0.024	0.002	0.028	0.038	0.038	0.033	0.892	0.720	0.718	0.881
Half-normal	1000	0.001	-0.006	-0.006	0.003	0.017	0.018	0.018	0.018	0.884	0.849	0.848	0.883
Bimodal	1000	-0.000	-0.001	-0.001	0.001	0.009	0.009	0.009	0.009	0.900	0.892	0.891	0.888

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 fitting GLasso-weighted, GW). EW is the oracle benchmark.

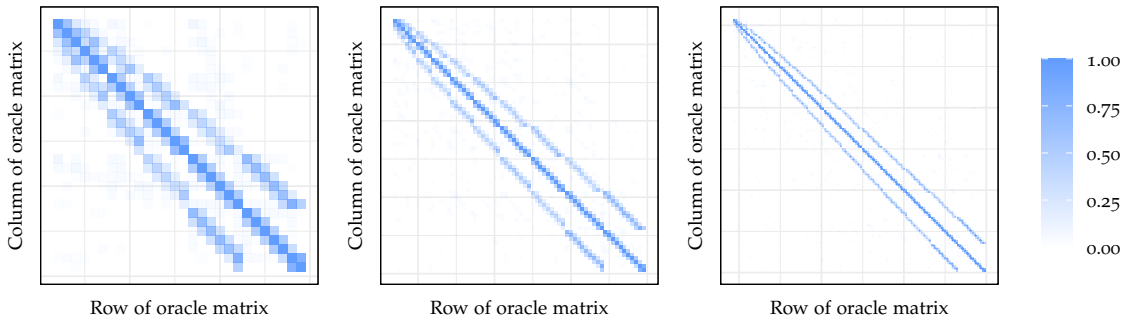
C.2 Parameter-Level Results for Baker and Solon (2003)

In Table C.2 we present the full parameter-level results from the Baker and Solon (2003) simulation design, for a cohort sample sizes 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.

C.3 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 5 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. Given 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 panel (a) and (b) in Figure 5, where the respective plots exhibit a 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 5: 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.

C.4 The Role of Cross Fitting and GLasso

We propose a cross fitting minimum distance estimator based on regularized estimation of the weighting matrix, which we construct based on the GLasso estimator. In this section we provide simulation evidence that both of these features are important in the context of our Baker and Solon (2003) simulation study.

Results are summarized in Figure 6, which presents the same style violin plots that we used in the main text. Consider first when a birth cohort sample size of $n_b = 400$ is considered. When cross-fitting is not applied, both OW and GW perform similarly with

respect to bias, however GW is slightly better in terms of coverage (but both are well-below 0.9), and much better in terms of RMSE. In both cases, applying cross fitting has an important impact on the performance of our estimators. The coverage probabilities of the 90% confidence intervals for all parameters becomes close to 0.9, and while the bias is reduced for both OW and GW, the reduction is much larger in the latter case. Finally, while RMSE worsens under OW, it is broadly unchanged under GW. Thus, the cross-fitting GW performs much better than OW with cross fitting. The same qualitative findings are true under larger birth cohort sizes, although the difference between the performance of OW and GW (both with and without cross fitting) decreases as the sample size increases.

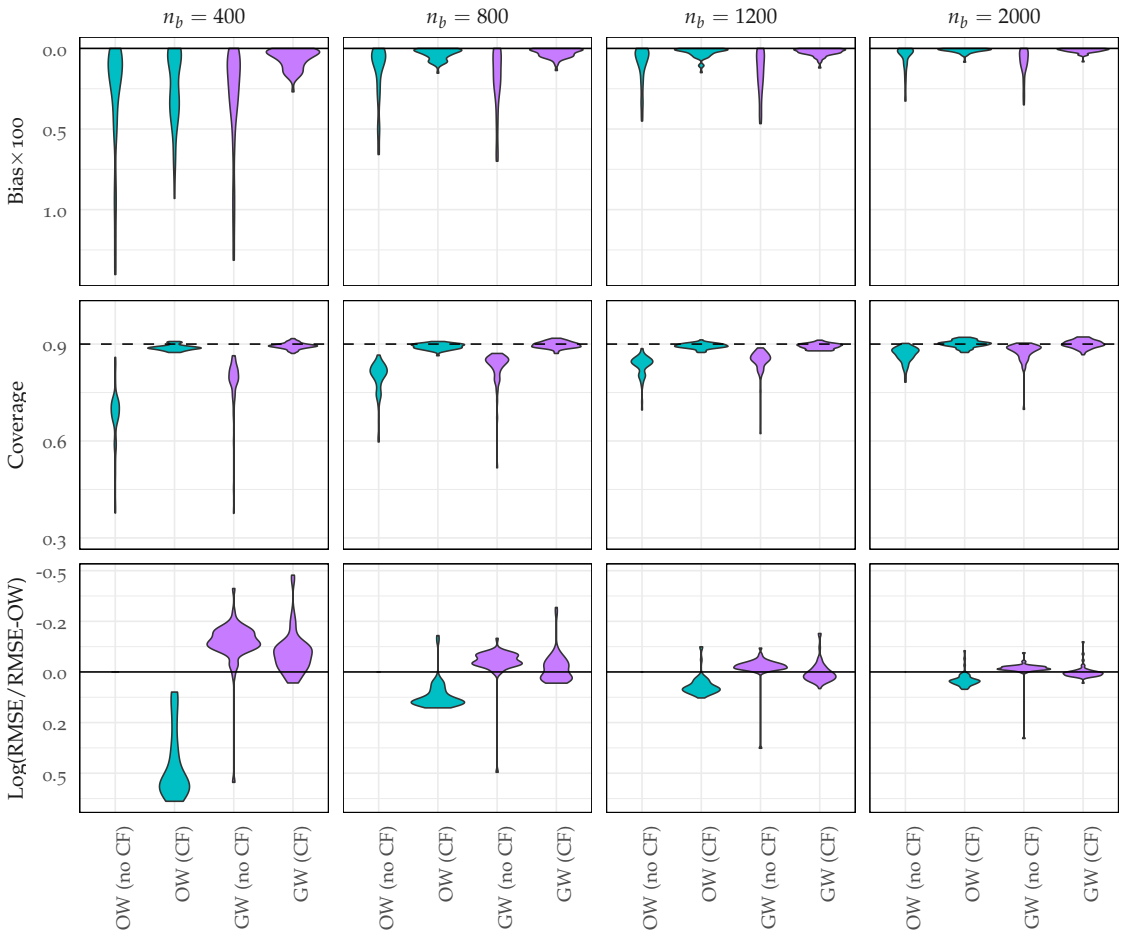


Figure 6: Violin plots for Baker and Solon (2003) parameters, showing bias, coverage probabilities of the 90% confidence intervals, and log root-mean square error (RMSE), which is relative to the RMSE under optimal weighting. Figure derived from 1,000 replications. Weighting denoted OW (optimally-weighted), and GW (GLasso-weighted), and by whether cross-fitting is applied (CF) or not (no CF).

Table C.2: 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	-0.014	-0.010	-2.2E-4	0.013	0.029	0.012	0.008	0.840	0.720	0.377	0.888
σ^2_β	9.0E-5	-7.4E-6	-1.2E-5	-2.7E-6	6.4E-7	4.0E-5	4.7E-5	2.9E-5	3.0E-5	0.907	0.869	0.714	0.893
$\sigma_{\alpha\beta}$	-0.003	4.9E-4	9.1E-4	2.6E-4	3.8E-5	8.3E-4	0.002	4.7E-4	3.9E-4	0.803	0.673	0.564	0.892
σ^2_r	0.007	-6.8E-4	-0.001	-6.4E-4	-9.7E-5	0.001	0.003	9.0E-4	6.7E-4	0.789	0.631	0.497	0.888
ρ	0.540	0.029	0.051	6.6E-4	0.002	0.039	0.072	0.008	0.008	0.559	0.345	0.726	0.884
γ_0	0.090	-8.5E-4	-0.003	-0.005	1.8E-4	0.014	0.015	0.007	0.004	0.873	0.846	0.403	0.898
γ_1	-0.005	1.6E-4	4.4E-4	4.1E-5	-5.6E-6	0.004	0.004	1.0E-3	0.001	0.897	0.870	0.738	0.901
γ_2	6.2E-5	-6.6E-6	-1.6E-5	1.8E-5	-6.0E-7	3.9E-4	4.5E-4	1.1E-4	1.1E-4	0.906	0.888	0.717	0.909
γ_3	2.2E-6	6.5E-8	1.9E-7	-8.5E-7	3.4E-8	1.7E-5	1.9E-5	4.8E-6	4.7E-6	0.906	0.883	0.722	0.905
γ_4	2.1E-9	1.0E-9	0.000	9.0E-9	0.000	2.4E-7	2.8E-7	7.0E-8	7.0E-8	0.908	0.899	0.723	0.908
p_{77}	1.035	0.001	0.002	0.003	0.001	0.012	0.015	0.015	0.013	0.894	0.879	0.673	0.893
p_{78}	1.028	0.001	0.004	0.002	9.1E-4	0.017	0.026	0.019	0.017	0.917	0.898	0.710	0.907
p_{79}	1.005	0.003	0.008	0.002	0.001	0.019	0.044	0.020	0.019	0.897	0.868	0.698	0.888
p_{80}	1.030	0.004	0.010	0.002	0.001	0.021	0.057	0.022	0.021	0.883	0.861	0.698	0.896
p_{81}	1.050	0.005	0.013	0.002	0.002	0.023	0.070	0.023	0.022	0.884	0.850	0.696	0.898
p_{82}	1.143	0.005	0.015	7.4E-4	6.9E-4	0.026	0.075	0.027	0.026	0.881	0.843	0.696	0.892
p_{83}	1.124	0.004	0.015	8.7E-4	9.0E-4	0.027	0.089	0.028	0.026	0.887	0.842	0.709	0.910
p_{84}	1.125	0.004	0.016	7.0E-4	0.001	0.027	0.102	0.028	0.026	0.889	0.847	0.705	0.895
p_{85}	1.122	0.004	0.016	3.7E-4	0.001	0.029	0.118	0.028	0.027	0.890	0.858	0.726	0.896
p_{86}	1.111	0.002	0.015	-2.6E-5	0.001	0.028	0.135	0.028	0.027	0.883	0.863	0.730	0.891
p_{87}	1.098	0.002	0.014	4.9E-6	0.001	0.028	0.154	0.028	0.027	0.913	0.884	0.725	0.896
p_{88}	1.105	0.002	0.014	2.1E-5	0.001	0.028	0.172	0.029	0.028	0.912	0.888	0.717	0.890
p_{89}	1.126	0.002	0.013	1.7E-4	0.002	0.029	0.188	0.030	0.029	0.911	0.895	0.724	0.904
p_{90}	1.127	8.7E-4	0.012	-2.3E-4	0.002	0.030	0.206	0.031	0.030	0.913	0.899	0.715	0.882
p_{91}	1.234	0.002	0.013	-6.3E-4	0.003	0.034	0.215	0.035	0.034	0.907	0.887	0.724	0.888
p_{92}	1.253	8.2E-4	0.011	-0.002	0.002	0.035	0.235	0.036	0.036	0.901	0.895	0.720	0.893
σ^2_{24-25}	0.133	0.002	0.007	-9.4E-4	3.0E-4	0.040	0.047	0.014	0.015	0.893	0.890	0.859	0.899
σ^2_{26-27}	0.084	0.003	0.011	-9.8E-4	3.5E-4	0.034	0.043	0.010	0.011	0.903	0.880	0.838	0.893
σ^2_{28-29}	0.116	0.002	0.008	-0.005	-4.1E-4	0.035	0.042	0.014	0.014	0.891	0.879	0.764	0.885
σ^2_{30-31}	0.071	0.001	0.009	-0.003	4.6E-4	0.031	0.038	0.010	0.010	0.909	0.878	0.698	0.878
σ^2_{32-33}	0.071	0.004	0.012	-0.004	3.6E-4	0.031	0.037	0.010	0.010	0.907	0.890	0.670	0.899
σ^2_{34-35}	0.127	0.001	0.006	-0.008	0.001	0.033	0.042	0.017	0.014	0.909	0.893	0.595	0.899
σ^2_{36-37}	0.085	0.002	0.009	-0.006	9.6E-4	0.031	0.039	0.013	0.011	0.898	0.874	0.584	0.878
σ^2_{38-39}	0.044	0.005	0.014	-0.003	5.7E-4	0.028	0.036	0.009	0.007	0.901	0.853	0.629	0.895
σ^2_{40-41}	0.066	0.005	0.013	-0.005	3.0E-4	0.026	0.035	0.011	0.010	0.919	0.890	0.647	0.881
σ^2_{42-43}	0.074	0.006	0.014	-0.005	6.0E-4	0.029	0.038	0.012	0.010	0.893	0.858	0.615	0.891
σ^2_{44-45}	0.054	0.006	0.015	-0.004	5.2E-4	0.026	0.037	0.010	0.009	0.912	0.875	0.589	0.887
σ^2_{46-47}	0.071	0.007	0.016	-0.005	5.2E-4	0.027	0.037	0.012	0.011	0.906	0.858	0.597	0.894
σ^2_{48-49}	0.090	0.006	0.014	-0.008	-1.3E-4	0.028	0.037	0.015	0.012	0.902	0.867	0.577	0.900
σ^2_{50-51}	0.167	0.005	0.011	-0.014	-2.2E-4	0.032	0.040	0.024	0.019	0.895	0.875	0.532	0.886
σ^2_{52-53}	0.157	0.006	0.012	-0.010	6.4E-4	0.031	0.041	0.021	0.018	0.898	0.874	0.665	0.895
σ^2_{54-55}	0.251	0.001	0.004	-0.014	8.3E-4	0.039	0.051	0.030	0.027	0.895	0.884	0.662	0.877
σ^2_{56-57}	0.295	0.004	0.007	-0.012	2.2E-4	0.044	0.061	0.033	0.032	0.904	0.887	0.744	0.871
σ^2_{58-59}	0.377	-7.5E-4	-0.004	-0.012	9.1E-5	0.050	0.071	0.037	0.036	0.896	0.887	0.792	0.895
σ^2_{60-61}	0.388	-0.004	-0.007	-0.009	-0.002	0.051	0.082	0.036	0.037	0.893	0.882	0.832	0.876
λ_{78}	1.132	-0.004	-0.009	0.001	-5.0E-5	0.044	0.056	0.026	0.023	0.923	0.888	0.680	0.909
λ_{79}	0.950	-0.009	-0.013	4.6E-4	-1.8E-4	0.046	0.065	0.023	0.020	0.892	0.838	0.685	0.889
λ_{80}	1.060	-0.003	-0.004	0.001	1.2E-4	0.060	0.073	0.025	0.022	0.888	0.881	0.665	0.894
λ_{81}	1.066	-0.008	-0.012	0.002	9.0E-4	0.060	0.073	0.024	0.022	0.910	0.864	0.695	0.914
λ_{82}	1.397	-0.006	-0.019	0.003	1.1E-4	0.080	0.086	0.032	0.028	0.899	0.851	0.660	0.893
λ_{83}	1.527	-0.017	-0.043	7.6E-4	-5.7E-4	0.087	0.105	0.035	0.031	0.893	0.813	0.670	0.893
λ_{84}	1.379	-0.019	-0.044	0.002	4.3E-4	0.086	0.108	0.030	0.027	0.867	0.799	0.687	0.917
λ_{85}	1.343	-0.019	-0.040	0.002	1.9E-4	0.084	0.109	0.031	0.027	0.863	0.795	0.672	0.897
λ_{86}	1.339	-0.011	-0.027	0.001	1.3E-4	0.084	0.109	0.030	0.027	0.895	0.840	0.676	0.890
λ_{87}	1.304	-0.016	-0.027	0.002	-4.0E-5	0.078	0.112	0.029	0.026	0.891	0.843	0.695	0.900
λ_{88}	1.285	-0.010	-0.016	0.002	1.6E-4	0.078	0.114	0.029	0.027	0.898	0.856	0.693	0.896
λ_{89}	1.260	-0.007	-0.007	0.001	-3.9E-4	0.080	0.118	0.029	0.027	0.906	0.877	0.661	0.895
λ_{90}	1.405	-0.006	-0.005	0.001	-7.7E-4	0.080	0.123	0.033	0.030	0.908	0.891	0.680	0.902
λ_{91}	1.513	-0.004	-6.7E-4	0.004	1.3E-4	0.088	0.128	0.036	0.033	0.897	0.870	0.689	0.895
λ_{92}	1.715	4.1E-4	0.003	0.002	5.2E-4	0.093	0.144	0.040	0.036	0.904	0.882	0.690	0.904

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-fitting GLasso-weighted, GW).