Standard Errors for Calibrated Parameters*

Matthew D. Cocci
Princeton University

Mikkel Plagborg-Møller
Princeton University

This version: September 16, 2021
First version: June 12, 2019

Abstract: Calibration, the practice of choosing the parameters of a structural model to match certain empirical moments, can be viewed as minimum distance estimation. Existing standard error formulas for such estimators require a consistent estimate of the correlation structure of the empirical moments, which is often unavailable in practice. Instead, the variances of the individual empirical moments are usually readily estimable. Using only these variances, we derive conservative standard errors and confidence intervals for the structural parameters that are valid even under the worst-case correlation structure. In the over-identified case, we show that the moment weighting scheme that minimizes the worst-case estimator variance amounts to a moment selection problem with a simple solution. Finally, we develop tests of over-identifying or parameter restrictions. We apply our methods empirically to a model of menu cost pricing for multi-product firms and to a heterogeneous agent New Keynesian model.

Keywords: calibration, data combination, minimum distance, moment selection, semidefinite programming. *JEL codes:* C12, C52.

*Emails: mcocci@princeton.edu and mikkelpm@princeton.edu. Address for both authors: Julis Romo Rabinowitz Building, Princeton, NJ 08544, USA. We are grateful for comments from Fernando Alvarez, Isaiah Andrews, Kirill Evdokimov, Bo Honoré, Michal Kolesár, Elena Manresa, Pepe Montiel Olea, Ulrich Müller, and seminar participants at Bocconi, Chicago, Cleveland Fed, NYU, Princeton, and the SED 2021 Annual Meeting. We thank Minsu Chang and Silvia Miranda-Agrippino for supplying the moments used in our second empirical application. Eric Qian provided excellent research assistance. Plagborg-Møller acknowledges that this material is based upon work supported by the NSF under Grant #1851665. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

1 Introduction

Researchers often discipline the parameters of structural economic models by calibrating certain model-implied moments to the corresponding moments in the data (Kydland & Prescott, 1996; Nakamura & Steinsson, 2018). This calibration strategy can be viewed as a version of moment matching (or more generally, minimum distance) estimation, as argued by Hansen & Heckman (1996). Moment matching is popular in diverse fields of applied structural economics.

Standard moment matching inference requires knowledge of the variance-covariance matrix of the empirical moments, but in practice this matrix is often only partially known. When the empirical moments are obtained from different data sets, different econometric methods, or different previous papers, it is usually hard or impossible to estimate the off-diagonal elements of the variance-covariance matrix. Nevertheless, the diagonal of the matrix – the variances of the individual empirical moments – is typically estimable. In this paper, we show that the diagonal suffices to obtain practically useful worst-case standard errors for the moment matching estimator. Moreover, in the over-identified case, we show that the moment weighting that minimizes the worst-case estimator variance amounts to a moment selection problem with a simple solution. Hence, our methods allow researchers to choose their moments and data sources freely without giving up on doing valid statistical inference.

We show that worst-case standard errors for the structural parameters (or smooth transformations thereof), using only the empirical moment variances, are easy to compute. They are given by a weighted sum of the standard errors of individual empirical moments, where the weights depend on the moment weight matrix and the derivatives of the moments with respect to the structural parameters. The derivatives can be obtained analytically, by automatic differentiation, or by first differences. Using these worst-case standard errors, one can construct a confidence interval that is valid even under the worst-case correlation structure. The confidence interval is generally conservative for specific correlation structures, but its minimax coverage is exact, i.e., under the worst-case correlation structure, which amounts to perfect positive/negative correlation. The confidence interval is likely to be informative in many empirical applications, as it is at most $\sqrt{p}$ times wider than it would be if the moments were known to be independent, where p is the number of moments used for estimation.

Given knowledge of only the individual empirical moment variances, we show that the moment weighting scheme that minimizes the worst-case estimator variance amounts to a moment *selection* problem. That is, the efficient minimum distance weight matrix attaches zero weight to some of the moments. The efficient selection of moments can be conveniently computed by running a median regression (i.e., Least Absolute Deviation regression) on a particular artificial data set. The efficient estimator given knowledge of only the moment variances is generally different from the familiar full-information efficient estimator that requires knowledge of the entire moment correlation structure.

To understand the intuition behind our results, consider the analogy of portfolio selection in finance. This analogy is mathematically relevant, as it is well known that any minimum distance estimator is asymptotically equivalent to a linear combination of the empirical moments – a "portfolio" of moments – with a linear restriction on the weights to ensure unbiasedness. When constructing a minimum-variance financial portfolio that achieves a given expected return, it is usually optimal to diversify across all available assets, *except* if the assets are perfectly (positively or negatively) correlated. In the latter extreme case, it is optimal to entirely disregard assets with sufficiently high variance relative to their expected return. But it is precisely the extreme case of perfect correlation that delivers the worst-case variance of a given portfolio. Thus, the portfolio with the smallest worst-case variance across correlation structures is a portfolio that *selects* a subset of the available assets. We further illustrate the analogy between portfolio selection and moment selection in Section 4.

We derive joint tests of parameter restrictions as well as tests of over-identifying restrictions. A common form of over-identification test used in the empirical literature is to check whether the estimated structural parameters yield a good fit of the model to "non-targeted" moments, i.e., moments that were not exploited for parameter estimation. We show how to implement a formal statistical test based on this idea in our setup. For joint testing of parameter restrictions, we propose a Wald-type test. The proof of the validity of this test relies on tail probability bounds for quadratic forms in Gaussian vectors from Székely & Bakirov (2003), but the test statistic and critical value are simple and easy to compute.

Finally, we extend our procedures to settings with more detailed knowledge of the covariance matrix of empirical moments. This includes settings where the entire correlation structure is known for some subsets of the moments, or where certain moments are known to be independent of each other.

We illustrate the usefulness of our procedures through two empirical applications. In the first one, we estimate and test the Alvarez & Lippi (2014) model of menu cost price setting in multi-product firms, by matching moments of price changes. In the second application, we estimate and test a heterogeneous agent New Keynesian model developed by McKay, Nakamura & Steinsson (2016) and Auclert, Bardóczy, Rognlie & Straub (2021), by matching impulse responses for macro time series and cross-sectional micro moments. Our worst-case standard errors allow for informative inference on several parameters of interest in both applications. A Monte Carlo simulation study calibrated to the first application indicates that our methods perform well in finite samples.

LITERATURE. Unlike the literature on correlation matrix completion, we solve the explicit problem of finding the worst-case correlation structure when estimating parameters in a structural model. Existing papers (see Georgescu, Higham & Peters, 2018, and references therein) instead compute positive definite correlation matrices that satisfy various reduced-form optimality criteria, such as the maximum entropy principle. Our derivations of the worst-case efficient weight matrix and joint testing procedure do not seem to have parallels in the matrix completion literature.

While we focus on cases where it is difficult to estimate the correlation structure of different moments, in some applications it may be possible to model and exploit the precise relationship between the moments. The literature on estimating heterogeneous agent models in macroeconomics has recently developed procedures for combining macro and micro data, as discussed further in Section 7.2. Hahn, Kuersteiner & Mazzocco (2020b) provide advanced tools for doing inference with a mix of cross-sectional and time series data. These methods, unlike ours, generally require access to the underlying data, rather than just the moments and their standard errors. Imbens & Lancaster (1994) consider a microeconometric setting where certain macro moments are known without error, which is a special case of our framework. Hahn, Kuersteiner & Mazzocco (2020a) give examples of structural models where both time series and cross-sectional data are required for identification of structural parameters. Their insights may help inform the choice of moments for the methods that we develop below.

OUTLINE. Section 2 defines the moment matching setup. Section 3 derives the worst-case standard errors and the efficient moment weighting/selection. Section 4 presents simple geometric and analytical illustrations of our results. Section 5 develops tests of parameter restrictions and of over-identifying restrictions. Section 6 extends our methods to settings where some off-diagonal elements of the moment covariance matrix are known. Section 7 contains two empirical illustrations. Section 8 concludes. Appendix A contains proofs and other technical details. Code for implementing our procedures is available online.[1]

[1]Matlab: https://github.com/mikkelpm/stderr_calibration_matlab. Python: https://github.com/mikkelpm/stderr_calibration_python

2 Setup

Consider a standard moment matching (minimum distance) estimation framework. Let $\mu_0 \in \mathbb{R}^p$ be a vector of reduced-form parameters, which we will refer to as "moments", though the method applies more generally. Let $\theta_0 \in \Theta \subset \mathbb{R}^k$ be a vector of structural model parameters. According to an economic model, the two parameter vectors are linked by the relationship $\mu_0 = h(\theta_0)$, where $h \colon \Theta \to \mathbb{R}^p$ is a known function implied by the model. The map $h(\cdot)$ may be computed either analytically or numerically. We have access to an estimator $\hat{\mu}$ ("empirical moments") that satisfies

$$\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, V) \tag{1}$$

for a $p \times p$ symmetric positive semidefinite variance-covariance matrix V.[2] Let $\hat{W}$ be a $p \times p$ symmetric matrix satisfying $\hat{W} \xrightarrow{p} W$ (we discuss the choice of $\hat{W}$ in Section 3.2). Then a "moment matching" or "minimum distance" estimator of θ_0 is given by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, (\hat{\mu} - h(\theta))' \hat{W} (\hat{\mu} - h(\theta)). \tag{2}$$

This estimation strategy is sometimes referred to as "calibration".

If we were able to estimate the covariance matrix of the empirical moments $\hat{\mu}$ consistently, it would be straight-forward to construct standard errors for any smooth function of the estimator $\hat{\theta}$. Suppose we are interested in the scalar transformed parameter $r(\theta_0)$, where $r \colon \Theta \to \mathbb{R}$. For example, $r(\cdot)$ may equal a particular counterfactual quantity in the structural model, or we could simply set $r(\theta) = \theta_i$ for some index i. Under the standard regularity conditions listed below in Assumption 1,

$$\begin{aligned} \sqrt{n}(r(\hat{\theta}) - r(\theta_0)) &= \lambda'(G'WG)^{-1}G'W\sqrt{n}(\hat{\mu} - \mu_0) + o_p(1) \\ &\xrightarrow{d} N\left(0, \lambda'(G'WG)^{-1}G'WVWG(G'WG)^{-1}\lambda\right), \end{aligned} \tag{3}$$

where $G \equiv \partial h(\theta_0)/\partial \theta' \in \mathbb{R}^{p \times k}$ and $\lambda \equiv \partial r(\theta_0)/\partial \theta \in \mathbb{R}^k$. See Newey & McFadden (1994)

[2] Here and below, all limits are taken as the sample size $n \to \infty$. We implicitly think of the sample sizes for the different moments as being proportional, with the factors of proportionality reflected in V. If some element $\hat{\mu}_j$ converges at a faster rate than $\sqrt{n}$, then $V_{jj} = 0$. Sample sizes and convergence rates only enter into our practical procedures through their implicit effect on the calculation of the moment standard errors $\hat{\sigma}_j$ (discussed below), which is handled automatically by econometric software.

for details. If the entire asymptotic covariance matrix V of $\hat{\mu}$ were consistently estimable, the above display would allow computation of standard errors, confidence intervals, and hypothesis tests.

Unfortunately, the full correlation structure of $\hat{\mu}$ is difficult or impossible to estimate in certain applications. This may be the case, for example, when the moments $\hat{\mu}$ are obtained from a variety of different data sources or econometric methods, or from previous studies for which the underlying data is not readily available. Moreover, if the moments involve a mix of time series and cross-sectional data sources, it can be difficult conceptually or practically to estimate correlations across data sources, whether through the bootstrap or the Generalized Method of Moments (GMM). While the structural model could in some cases be exploited to estimate the moment covariance matrix, this may require stronger assumptions than what is needed for point estimation of the structural parameters.[3] We illustrate all these points in the empirical application in Section 7.2. Even in cases where a convenient estimator of the full asymptotic covariance matrix V is available, the off-diagonal elements may not be accurately estimated in finite samples (Altonji & Segal, 1996).

Yet, it is often the case that the standard errors of each of the components of $\hat{\mu}$ are available. These marginal standard errors may be directly computable from data, or they may be reported in the various papers that the individual elements of $\hat{\mu}$ are gleaned from. Thus, assume that we have access to standard errors $\hat{\sigma}_1, \dots, \hat{\sigma}_p \geq 0$ satisfying

$$\sqrt{n}\hat{\sigma}_j \xrightarrow{p} V_{jj}^{1/2}, \quad j = 1, \dots, p. \tag{4}$$

We show in the next section that these marginal standard errors suffice for doing informative inference on $r(\theta_0)$.

For ease of reference, we summarize our technical assumptions here:

Assumption 1.

i) The empirical moment vector $\hat{\mu}$ is asymptotically normal, as in (1), and $V \neq 0_{p \times p}$.

ii) The standard error estimators $\hat{\sigma}_j$ are consistent, as in (4).

iii) $h(\cdot)$ and $r(\cdot)$ are both continuously differentiable in a neighborhood of θ_0 (which lies in the interior of Θ), $G \equiv \partial h(\theta_0)/\partial \theta'$ has full column rank, and $\lambda \equiv \partial r(\theta_0)/\partial \theta \neq 0_{k \times 1}$.

[3] For example, if we exploit the model's predictions about second moments for estimation, model-based estimation of V would require believing the model's predictions about fourth moments.

iv) $\hat{\theta} \xrightarrow{p} \theta_0$.

v) $\hat{W} \xrightarrow{p} W$ *for a symmetric positive semidefinite matrix* W.

vi) $G'WG$ *is nonsingular.*

Conditions (ii)–(vi) are standard regularity conditions that are satisfied in smooth, locally identified models (Newey & McFadden, 1994). Note that we allow for the possibility that some moments are known with certainty ($V_{jj} = 0$) as in Imbens & Lancaster (1994). We will now discuss the key condition (i).

Discussion of joint normality assumption. When the elements of the moment vector $\hat{\mu}$ are obtained from different data sets, the joint normality assumption (1) requires justification. This ensures not only that a normal distribution is the appropriate reference distribution for obtaining critical values, but also that the vector $\hat{\mu}$ can reasonably be viewed as arising from some joint, repeatable experiment for which the given standard errors $\hat{\sigma}_j$ capture all sources of uncertainty. The joint normality assumption is most easily understood and justified under a model-based (e.g., shock-based) perspective on uncertainty. In this view, there exists a coherent data generating process with both aggregate and idiosyncratic shocks that affect all of the observed data. The empirical application of Section 7.2 is a prototypical example of this framework, but such applications are not the only use case.

There are several cases in which the joint normality assumption appears reasonable, but estimation of the full moment covariance matrix V could be challenging. For example:

1. *The moments are obtained from the same or similar data sets, but the underlying data for some of the moments is not available.* For example, some moments may be reported in previous papers that use proprietary data. See Section 7.1 for an empirical example.

2. *Some of the moments are computed from aggregate time series and others from panel data spanning similar time periods.* If the clustering procedure of the panel data regressions allows for aggregate shocks, and these aggregate shocks also affect the time series data, then the panel regressions will have correct standard errors but the coefficients may be correlated with the time series moments.

3. *The moments stem from time series data observed at various frequencies, or from regional data with various levels of geographic aggregation.* While careful econometric analysis may allow the estimation of the full covariance matrix of the moments, this could be cumbersome in practice.

4. *We use a combination of aggregate time series moments and micro moments from surveys, and the latter measure time-invariant parameters that are not affected by macro shocks in the sample.* In this case, it is often reasonable to assume that the uncertainty in the micro moments (arising purely from idiosyncratic noise) is independent of the uncertainty in the macro moments. Such extra information can be incorporated in our procedures, as shown in Section 6.

5. *The moments are all computed from the same data set, but using a variety of complicated procedures.* In this case, it may be difficult to estimate the correlation structure analytically using, say, GMM, and the bootstrap may be computationally impractical.

However, in certain cases the joint normality assumption may fail. For example:

1. *We use a combination of aggregate time series moments and micro moments from surveys, but the latter are affected by aggregate macro shocks that shift the whole micro outcome distribution.* In this case, standard cross-sectional moments may not even be *consistent* for the true underlying population moments, since the aggregate shocks do not get averaged out (Hahn et al., 2020a, Section 3). Moreover, the usual micro standard errors will not take into account the *combined* uncertainty in the macro shock and idiosyncratic micro noise. Nevertheless, as long as one appropriately accounts for all types of uncertainty when computing moments and their standard errors, our methods below can be applied. We illustrate this empirically in Section 7.2.

2. *The data used to compute some of the moments is very heavy tailed, or the estimation procedures are not asymptotically regular.* In this case, even *marginal* normality of the individual empirical moments may fail. We leave extensions to non-normal limit distributions as an interesting topic for future research.

3 Standard errors and moment selection

We first derive the worst-case standard errors for a given choice of moment weight matrix. Then we show that the weighting scheme that minimizes the worst-case standard errors amounts to a moment selection problem with a simple solution.

3.1 Worst-case standard errors and confidence intervals

We first compute the worst-case bound on the standard error of the moment matching estimator, given knowledge of only the variances of the empirical moments. Although the argument relies on a straight-forward application of the Cauchy-Schwarz inequality, it appears that the literature has not realized the practical utility of this result.

Recall that we seek to do inference on the scalar parameter $r(\theta_0)$. By the standard delta method expansion (3) under Assumption 1, the estimator $r(\hat{\theta})$ is asymptotically equivalent to a certain linear function $x'\hat{\mu}$ of the empirical moments, where $x = (x_1, \dots, x_p)' \equiv WG(G'WG)^{-1}\lambda$. We thus seek to bound the variance of an (asymptotically) known linear combination of $\hat{\mu}$, knowing the variance of each component $\hat{\mu}_j$ but not the correlation structure. This worst-case variance is attained when all components of $\hat{\mu}$ are perfectly positively or negatively correlated (depending on the signs of the elements of x), yielding the worst-case variance $(\sum_{j=1}^{p} |x_j| \operatorname{Var}(\hat{\mu}_j)^{1/2})^2$. This elementary result is proved in Lemma 1 in Appendix A.1.[4]

We can thus construct an estimate of the worst-case standard error of $r(\hat{\theta})$ as

$$\widehat{\mathrm{se}}(\hat{x}) \equiv \sum_{j=1}^{p} \hat{\sigma}_j |\hat{x}_j|,$$

where $\hat{x} = (\hat{x}_1, \dots, \hat{x}_p)' \equiv \hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}\hat{\lambda}$, $\hat{G} \equiv \frac{\partial h(\hat{\theta})}{\partial \theta'}$, and $\hat{\lambda} \equiv \frac{\partial r(\hat{\theta})}{\partial \theta}$. In practice, the partial derivatives may be computed analytically, by automatic differentiation, or by finite differences. Let $\Phi(\cdot)$ denote the standard normal distribution function. Then the confidence interval

$$\left[r(\hat{\theta}) - \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}(\hat{x}), r(\hat{\theta}) + \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}(\hat{x})\right]$$

covers $r(\theta_0)$ with probability at least $1-\alpha$ asymptotically. The asymptotic coverage probability is exactly $1-\alpha$ if V happens to have the worst-case structure, i.e., when all elements of $\hat{\mu}$ are perfectly correlated asymptotically (so V has rank 1). Formally, these results follow from the fact that, under Assumption 1,

$$\sqrt{n}\,\widehat{\mathrm{se}}(\hat{x}) \xrightarrow{p} \sum_{j=1}^{p} V_{jj}^{1/2}|x_j| = \max_{\tilde{V} \in \mathcal{S}(\operatorname{diag}(V))} \sqrt{\lambda'(G'WG)^{-1}G'W\tilde{V}WG(G'WG)^{-1}\lambda},$$

[4] The basic insight is that $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y) \leq \operatorname{Var}(X) + \operatorname{Var}(Y) + 2(\operatorname{Var}(X)\operatorname{Var}(Y))^{1/2} = (\operatorname{Var}(X)^{1/2} + \operatorname{Var}(Y)^{1/2})^2$ by Cauchy-Schwarz.

where the equality uses Lemma 1 in Appendix A.1. Here $\mathcal{S}(\text{diag}(V))$ denotes the set of matrices $\tilde{V}$ that are $p \times p$ symmetric positive semidefinite and with diagonal elements $\tilde{V}_{jj} = V_{jj}$ for all j.

Remark.

1. By Jensen's inequality, $\sum_{j=1}^{p} \hat{\sigma}_j |\hat{x}_j| \leq p^{1/2} (\sum_{j=1}^{p} \hat{\sigma}_j^2 \hat{x}_j^2)^{1/2}$. Hence, the worst-case standard errors are at most $\sqrt{p}$ times larger than the standard errors that assume all elements of $\hat{\mu}$ to be mutually uncorrelated.

3.2 Efficient moment selection

We now derive a weight matrix W that minimizes the *worst-case* variance of the estimator, derived above. We show that this weight matrix puts weight on at most k moments, so the procedure amounts to efficient moment *selection*. Since the weight matrix W only matters in the over-identified case, we assume $p > k$ in this section. Let $\mathcal{S}_p$ denote the set of $p \times p$ symmetric positive semidefinite matrices W such that $G'WG$ is nonsingular.[5]

We seek a weight matrix W that minimizes the worst-case asymptotic standard deviation of $r(\hat{\theta})$. Let $x(W)$ denote the vector x defined in Section 3.1, viewed as a function of W. Then we solve the problem

$$\begin{aligned} &\min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \sqrt{\lambda'(G'WG)^{-1} G'W\tilde{V}WG(G'WG)^{-1}\lambda} \qquad (5) \\ &= \min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} (x(W)'\tilde{V}x(W))^{1/2} \\ &= \min_{W \in \mathcal{S}_p} \sum_{j=1}^{p} V_{jj}^{1/2} |x_j(W)|, \end{aligned}$$

where the last equality uses the result in Section 3.1 (cf. Lemma 1 in Appendix A.1). Lemma 2 in Appendix A.1 shows that the solution to the final optimization problem above is given by

$$\min_{W \in \mathcal{S}_p} \sum_{j=1}^{p} V_{jj}^{1/2} |x_j(W)| = \min_{z \in \mathbb{R}^{p-k}} \sum_{j=1}^{p} |\tilde{Y}_j - \tilde{X}_j' z|, \qquad (6)$$

[5]The latter constraint ensures that the true parameter vector θ_0 is a locally unique minimum of the population minimum distance objective function.

where we define

$$\tilde{Y}_j \equiv V_{jj}^{1/2} G_{j\bullet}(G'G)^{-1}\lambda \in \mathbb{R}, \quad \tilde{X}_j \equiv -V_{jj}^{1/2} G_{j\bullet}^{\perp\prime} \in \mathbb{R}^{p-k},$$

$G^\perp$ is any $p \times (p-k)$ matrix with full column rank satisfying $G'G^\perp = 0_{k\times(p-k)}$, and the notation $A_{j\bullet}$ means the j-th row of matrix A. The intuition for the equality (6) is that the set of all minimum distance estimators for various weight matrices is asymptotically equivalent to the set of all estimators that are linear combinations of the p moments $\hat{\mu}$, subject to k asymptotic unbiasedness constraints. We can therefore optimize over a $(p-k)$-dimensional linear space.

The final optimization problem (6) is a median regression (Least Absolute Deviation regression) of the artificial "regressand" $\{\tilde{Y}_j\}$ on the $p-k$ artificial "regressors" $\{\tilde{X}_j\}$. This regression can be executed efficiently using standard quantile regression software.

The solution to the median regression amounts to optimally selecting at most k of the p moments for estimation. Theorem 3.1 of Koenker & Bassett (1978) implies that there exists a solution z^* to the median regression (6) such that at least $p-k$ out of the p median regression residuals

$$e_j^* \equiv \tilde{Y}_j - \tilde{X}_j' z^*, \quad j = 1, \dots, p,$$

equal zero. Hence, an efficient weight matrix W^* that achieves the minimum in (6) will yield a linear combination vector $x(W^*) = (V_{11}^{-1/2} e_1^*, \dots, V_{pp}^{-1/2} e_p^*)'$ that attaches nonzero weight to at most k out of the p empirical moments $\hat{\mu}$. In other words, the solution to the efficient moment *weighting* problem is endogenously achieved by an efficient moment *selection*. We may pick an arbitrary weight matrix that attaches nonzero weight to only the efficiently selected moments (the magnitudes of the weights do not matter asymptotically, as the selected set of moments is just-identified).

Algorithm. The efficient estimator and standard errors can be computed as follows:

i) Compute an initial consistent estimator $\hat{\theta}_{\text{init}}$ using, say, a diagonal weight matrix with $\hat{W}_{jj} = \hat{\sigma}_j^{-2}$.

ii) Construct the derivative matrix $\hat{G} \equiv \frac{\partial h(\hat{\theta}_{\text{init}})}{\partial \theta'}$ and vector $\hat{\lambda} \equiv \frac{\partial r(\hat{\theta}_{\text{init}})}{\partial \theta}$, either analytically or numerically.

iii) Solve the median regression (6), substituting $\hat{G}$ for G, $\hat{\lambda}$ for λ, and $\hat{\sigma}_j$ for $V_{jj}^{1/2}$.[6] Compute the residuals $\hat{e}_j^*$, $j = 1, \dots, p$, from this median regression. (In the non-generic case where multiple solutions to the median regression exist, select one that yields at most k nonzero residuals.)

iv) Construct the efficient linear combination $\hat{x}^* = (\hat{x}_1^*, \dots, \hat{x}_p^*)'$ of the p moments, given by $\hat{x}_j^* \equiv \hat{\sigma}_j^{-1} \hat{e}_j^*$ for $j = 1, \dots, p$. At least $p - k$ of the elements will be zero, corresponding to those moments that are discarded by the efficient moment selection procedure.

v) To compute an efficient estimator of $r(\theta_0)$, either:

   a) Compute the just-identified efficient minimum distance estimator $\hat{\theta}_{\text{eff}}$ of θ_0 which uses any weight matrix that attaches zero weight to those (at least) $p - k$ moments which receive zero weight in the vector $\hat{x}^*$. Then estimate $r(\theta_0)$ by $r(\hat{\theta}_{\text{eff}})$. Or:

   b) Compute the "one-step" estimator $\hat{r}_{\text{eff-1S}} \equiv r(\hat{\theta}_{\text{init}}) + \hat{x}^{*\prime}(\hat{\mu} - h(\hat{\theta}_{\text{init}}))$ of $r(\theta_0)$.[7]

vi) The worst-case standard error of the estimator from step (v) is given by the value of the median regression (6) (i.e., the minimized objective function).

Options (a) and (b) in step (v) of the algorithm are asymptotically equivalent. Option (b) is computationally more convenient as it avoids further numerical optimization, but option (a) ensures that $\hat{\theta}_{\text{eff}}$ always lies in the parameter space Θ.

Remarks.

1. One can optionally re-run the median regression with updated $\hat{G}$ and $\hat{\lambda}$ based on $\hat{\theta}_{\text{eff}}$, but this does not increase asymptotic efficiency.

2. Since all operations involved in computing the efficient linear combination $\hat{x}^*$ are continuous, $\hat{x}^*$ converges in probability to the population efficient linear combination $x(W^*)$. The only exception may be where the population median regression (6) does not have a unique minimum, which is a non-generic case. Even in this case, however, the efficient worst-case standard errors will be consistent (when multiplied by $\sqrt{n}$) under Assumption 1(i)–(iv), by a standard application of the maximum theorem.

[6] Remember to omit an intercept from the regression.

[7] See Newey & McFadden (1994, Section 3.4) for a general discussion of one-step estimators.

3. The full-information (infeasible) efficient weight matrix that exploits knowledge of all of V is known to equal $W = V^{-1}$. This weight matrix in general attaches nonzero weight to all moments, unlike the limited-information efficient solution derived above. The worst-case asymptotic standard deviation (6) given limited information is of course larger than the asymptotic standard deviation $(\lambda'(G'V^{-1}G)^{-1}\lambda)^{1/2}$ of the full-information efficient estimator of $r(\theta_0)$.

4. The efficient moment weighting/selection for the limited-information efficient estimators $r(\hat{\theta}_{\text{eff}})$ and $\hat{r}_{\text{eff-1S}}$ depends on the function $r(\cdot)$ of interest, unlike in the case of full-information efficient estimation. In practice, we can just re-run the computations for all functions $r(\cdot)$ of interest (e.g., for all components of θ).

5. It is not restrictive to consider moment matching estimators of the form (2). Consider instead any estimator $\hat{\vartheta} \equiv \hat{f}(\hat{\mu})$ of θ_0, where $\hat{f} \colon \mathbb{R}^p \to \mathbb{R}^k$ is a possibly data-dependent function with enough regularity to satisfy the asymptotically linear expansion

$$\hat{\vartheta} - \theta_0 = H(\hat{\mu} - \mu_0) + o_p(n^{-1/2}),$$

for some $k \times p$ matrix H. If we restrict attention to asymptotically regular estimators of θ_0 (i.e., estimators that remain asymptotically unbiased under locally drifting parameters), we need $HG = I_k$. Among all estimators $\hat{\vartheta}$ satisfying these requirements, the smallest possible worst-case asymptotic standard deviation of $r(\hat{\vartheta})$ is achieved by the estimator whose asymptotic linearization matrix H solves

$$\min_{H \colon HG = I_k} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} (\lambda' H' \tilde{V} H \lambda)^{1/2}.$$

Lemma 2 in Appendix A.1 shows that the solution to this problem is precisely the value of the median regression (6). In other words, the minimum distance estimator $\hat{\theta}$ with (limited-information) efficient weight matrix delivers the smallest possible worst-case standard errors in a large class of estimators.

6. Our results extend in a straight-forward manner to Generalized Minimum Distance estimation. In that setting θ_0 and μ_0 are linked through a possibly non-separable equation $g(\theta_0, \mu_0) = 0_{m\times 1}$. The setting in Section 2 is a special case with $g(\theta, \mu) = \mu - h(\theta)$, but our calculations carry over with few changes because the asymptotic expansions are essentially the same (Newey & McFadden, 1994).

4 Geometric and analytical illustrations

This section uses two simple toy examples to illustrate geometrically and analytically the calculations in the previous section. Via direct arguments in these simple examples, we will arrive at the same efficient worst-case standard errors as the median regression in Section 3.2. In this section, we remove "hats" on $\hat{\sigma}_j$ to ease notation.

4.1 Geometric intuition: two moments, one parameter

Consider first the simplest possible example of two moments $\hat{\mu}_1, \hat{\mu}_2$ that are jointly normal in finite samples and are both noisy measures of a scalar structural parameter θ_0. In our notation, this model corresponds to $k = 1$, $p = 2$, and

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \sim N\Bigg(\underbrace{\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}}_{\mu_0}, \underbrace{\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}}_{\frac{1}{n}V} \Bigg), \qquad h(\theta) = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{G} \theta, \quad \theta \in \mathbb{R}.$$

Suppose also that the standard deviations of the first and second moments are known to be $\sigma_1 = 1$ and $\sigma_2 = 2$, respectively.

In this particular linear model, it can be shown that the class of minimum distance estimators of θ_0 is precisely the class of weighted sums of the two noisy measures $\hat{\mu}_1$ and $\hat{\mu}_2$, where the weights on these moments sum to one:

$$\hat{\theta}(x_1, x_2) \equiv x_1\hat{\mu}_1 + x_2\hat{\mu}_2, \qquad x_1 + x_2 = 1. \tag{7}$$

Therefore, we seek the linear combination of the moments $\hat{\theta}(x_1, x_2)$ with the smallest possible variance, subject to the linear constraint on the weights (x_1, x_2) given in (7). Note that this constraint ensures that $\hat{\theta}(x_1, x_2)$ is unbiased.

In this example it is obvious that the worst-case efficient estimation strategy is to use only the first moment. This is because the worst-case variance of any given estimator $\hat{\theta}(x_1, x_2)$ is achieved when the two moments are perfectly correlated, and in this extreme case, there is no benefit from including the high-variance moment $\hat{\mu}_2$ in the linear combination $\hat{\theta}(x_1, x_2)$.

We will now present a geometric visualization that delivers this obvious result, illustrated in the panels of Figure 1 below. Each estimator $\hat{\theta}(x_1, x_2)$ can be represented by a point $(x_1, x_2) \in \mathbb{R}^2$. As discussed above, the set of minimum distance estimators corresponds to the subset of points (x_1, x_2) satisfying the unbiasedness constraint in (7), which we represent

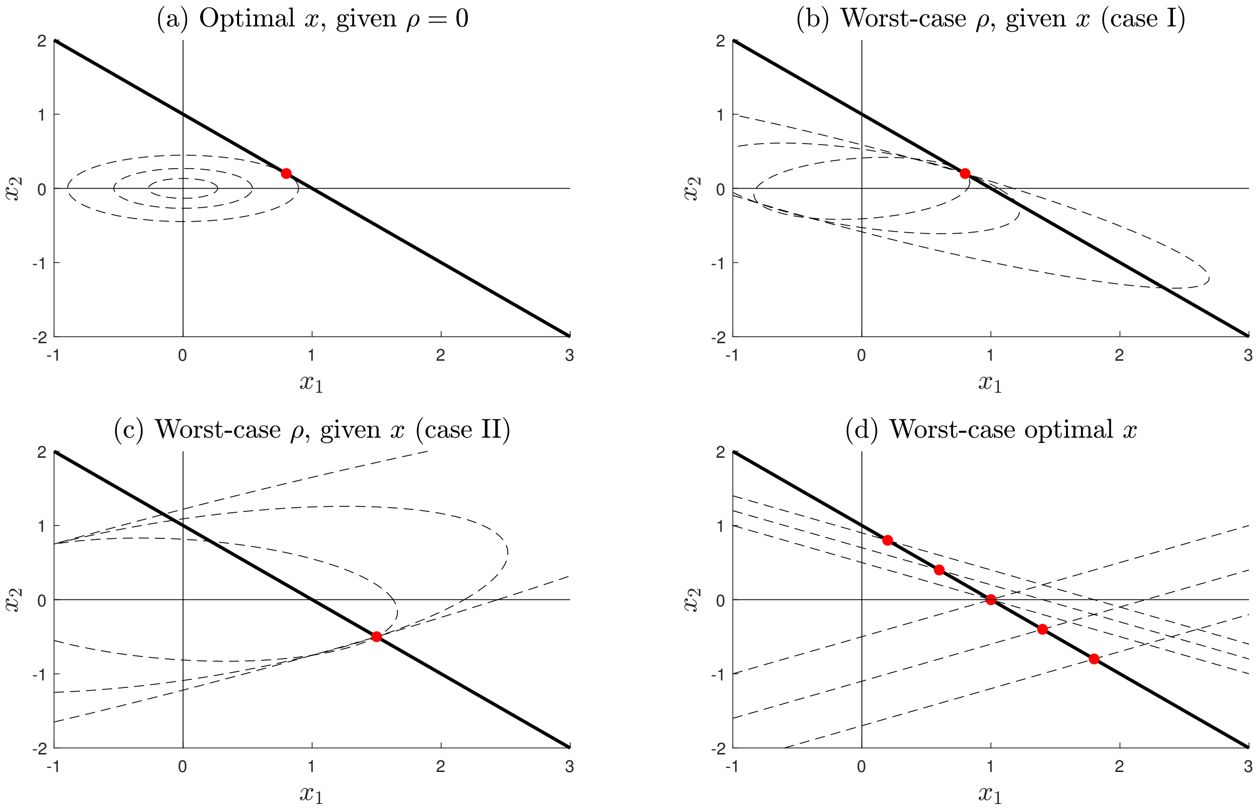

Figure 1: Geometric illustration of worst-case efficient estimator. (a): Optimal estimator given known correlation structure. (b)+(c): Worst-case correlation structure for two different estimators. (d) Worst-case efficient estimator. See main text for explanation.

by the thick straight line in all panels of the figure. For any choice of weights (x_1, x_2) (both on and off the line) we can compute the variance of the corresponding estimator,

$$\mathrm{Var}[\hat{\theta}(x_1, x_2)] = \sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 + 2\rho\sigma_1\sigma_2 x_1 x_2, \tag{8}$$

which depends upon the unknown correlation parameter ρ. Dashed ellipses in the figure represent level sets of the estimator variance (8).

Panel (a) of Figure 1 depicts the efficient estimator in the case where it is known that $\rho \equiv \mathrm{Corr}(\hat{\mu}_1, \hat{\mu}_2) = 0$. The lowest-variance estimator $\hat{\theta}(x_1, x_2)$ is found at the point (x_1, x_2) where the elliptical variance level sets are tangent to the straight line that embodies the unbiasedness constraint (7). Note that it is optimal to use both moments for estimation, i.e., $x_1, x_2 > 0$. This is the standard diversification motive in financial portfolio construction, as discussed in Section 1.

Panel (b) of Figure 1 fixes (x_1, x_2) at the optimum from panel (a) and depicts the standard deviation of the corresponding estimator $\hat{\theta}(x_1, x_2)$ for different values of ρ. The ellipses are again level sets for the variance, now corresponding to $\rho \in \{-0.2, 0.5, 0.9\}$ (the larger ρ, the larger the area of the ellipse in this plot). For a given level set ellipse, the corresponding standard deviation of the estimator is given by the x_1-coordinate at which the ellipse intersects with the positive half of the x_1-axis (since $\sigma_1 = 1$, cf. (8)). We see from the figure that the standard deviation of the estimator is increasing in ρ; hence, the worst case is $\rho = 1$.[8]

Panel (c) of Figure 1 repeats the exercise from panel (b), except that now we consider a point (x_1, x_2) with $x_2 < 0$. The three ellipses correspond to $\rho \in \{0.2, -0.5, -0.9\}$ (the more negative ρ, the larger the area of the ellipse in this plot). In this case, the figure shows that the worst-case correlation is $\rho = -1$.

Finally, panel (d) of Figure 1 finds the estimator $\hat{\theta}(x_1, x_2)$ with the smallest worst-case standard deviation.[9] The figure depicts five possible choices of (x_1, x_2). For each choice, it shows the variance level set corresponding to the worst-case correlation, which is $\rho = 1$ when $x_2 > 0$ and $\rho = -1$ when $x_2 < 0$. To simplify the figure we only plot the portion of the variance "ellipse" (here a line, due to perfect correlation) that intersects the positive half of the x_1-axis. Notice that any choice (x_1, x_2) on the unbiasedness line with $x_1 \neq 1$ leads to a worst-case standard deviation (i.e., intersection with the x_1-axis) that is strictly larger than 1. However, at $(x_1, x_2) = (1, 0)$, the standard deviation at both $\rho = 1$ and $\rho = -1$ equals 1. Hence, the efficient estimator is $\hat{\theta}(1, 0)$, which discards the second (higher-variance) moment. Again, it is worthwhile recalling the portfolio analogy in Section 1, where the diversification motive disappears if the available assets are perfectly correlated.

The geometry of panel (d) of Figure 1 extends to higher-dimensional settings. The unbiasedness constraint on $\hat{\theta}$ will always amount to a linear restriction on the weights x. Meanwhile, the level sets of the worst-case standard error $\widehat{\text{se}}(x) = \sum_{j=1}^{p} \sigma_j |x_j|$ look like diamonds centered at the origin in $\mathbb{R}^p$-space (since the worst-case standard error is a weighted L_1-norm). Any point of tangency of the unbiasedness hyperplane and the diamond level sets must occur at a vertex of the diamonds. At such a vertex, some of the elements of x equal zero, corresponding to moment selection (unless the hyperplane is parallel to one of the edges of the diamond, a non-generic case).

[8] This panel thus illustrates graphically the inner maximization in (5) for fixed W.

[9] This task corresponds to the outer maximization in (5).

4.2 Analytical calculations: three moments, two parameters

We here seek to do inference on the first parameter $r(\theta_0) = \theta_{0,1}$ (without loss of generality) in the following linear model with $p = 3$, $k = 2$:

$$h(\theta) = \underbrace{\begin{pmatrix} a & 0 \\ b & c \\ 0 & d \end{pmatrix}}_{G} \theta, \quad \theta \in \mathbb{R}^2.$$

Assume $a, b, c, d \neq 0$. In Appendix A.2 we provide detailed derivations that map this toy example into the general framework and notation of Section 3. In the following we instead re-derive the optimal estimator from first principles.

As argued in Section 3.2, the task of efficient estimation under worst-case correlation structure lies in selecting *which* moments (at most $k = 2$) should be used for estimating $\theta_{0,1}$. The third moment, which does not vary with θ_1, cannot be used alone. Instead, one option is to use the first moment alone (it would be necessary to add one of the other moments if we wanted to also estimate $\theta_{0,2}$). In this case,

$$\hat{\theta}_1 = \frac{1}{a}\hat{\mu}_1. \tag{9}$$

Alternatively, one could use the second and third moments together to estimate the first parameter, which yields

$$\left.\begin{array}{l} \hat{\mu}_2 = b\hat{\theta}_1 + c\hat{\theta}_2 \\ \hat{\mu}_3 = d\hat{\theta}_2 \end{array}\right\} \quad \Longrightarrow \quad \hat{\theta}_1 = \frac{1}{b}\hat{\mu}_2 - \frac{c}{bd}\hat{\mu}_3. \tag{10}$$

The estimator (9) is preferred over (10) under the worst-case correlation structure when

$$\sqrt{\mathrm{Var}\left(\frac{1}{a}\hat{\mu}_1\right)} = \frac{\sigma_1}{|a|} \quad \leq \quad \frac{1}{|b|}\sigma_2 + \left|\frac{c}{bd}\right|\sigma_3 = \max_{\mathrm{Corr}(\hat{\mu}_1, \hat{\mu}_2)} \sqrt{\mathrm{Var}\left(\frac{1}{b}\hat{\mu}_2 - \frac{c}{bd}\hat{\mu}_3\right)}.$$

The worst-case variance expression on the right-hand side follows by recalling from Section 3.1 that we need only check whether the worst case is attained at perfect positive or perfect negative correlation.

We can obtain a useful interpretation of the general approach of Section 3.2 by rewriting the above inequality in terms of signal-to-noise ratios:

$$\frac{\sigma_1}{|a|} \leq \frac{\sigma_2}{|b|} + \frac{\sigma_2/|b|}{\sigma_2/|c|} \times \frac{\sigma_3}{|d|}. \tag{11}$$

According to (11), we use the first moment $\hat{\mu}_1$ alone for estimation of $\theta_{0,1}$ when the signal-to-noise ratio $|a|/\sigma_1$ is relatively large. This occurs when the first moment has low σ_1 and therefore is more precisely measured, or when $|a|$ is large so that the first moment is particularly sensitive to and informative about the parameter of interest $\theta_{0,1}$. In the financial portfolio analogy of Section 1, $|a|$ captures the expected return of the asset, σ_1 measures its risk, and the worst-case optimal portfolio strategy amounts to selecting the asset with the highest Sharpe ratio $|a|/\sigma$.

As illustrated in this simple example, the median regression approach in Section 3.2 ultimately picks the efficient estimator by comparing the variances of all feasible just-identified estimators. This amounts to comparing signal-to-noise ratios to ensure that we select the portfolio of moments that is most precisely measured and sensitive to the components of θ_0 that determine the scalar parameter of interest $r(\theta_0)$.

5 Testing

In this section we develop a joint test of multiple parameter restrictions as well as a test of over-identifying restrictions.

5.1 Joint testing

We propose a test of the joint null hypothesis $H_0 \colon r(\theta_0) = 0_{m\times 1}$ against the two-sided alternative. In this section, the continuously differentiable function $r \colon \Theta \to \mathbb{R}^m$ is allowed to be multi-valued. Tests of a single parameter restriction ($m = 1$) can be carried out using the confidence interval described in Section 3.1. For the case $m > 1$, we propose the following testing procedure. Let $\alpha \in (0, 1)$ denote the significance level.

i) Compute the Wald-type test statistic

$$\hat{\mathscr{T}} \equiv r(\hat{\theta})'\hat{S}r(\hat{\theta}),$$

where $\hat{S}$ is a user-specified symmetric positive definite $m \times m$ matrix, to be discussed below.

ii) Compute the critical value

$$\text{cv}_n \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \frac{1}{n} \text{trace}\left(\tilde{V} W G (G'WG)^{-1} \lambda S \lambda' (G'WG)^{-1} G'W\right) \times \left(\Phi^{-1}(1-\alpha/2)\right)^2. \tag{12}$$

In practice, we substitute the estimates $\frac{1}{n}\text{diag}(V) \approx \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2)$, $W \approx \hat{W}$, $S \approx \hat{S}$, $G \approx \hat{G} \equiv \frac{\partial h(\hat{\theta})}{\partial \theta'}$, and $\lambda \approx \hat{\lambda} \equiv \frac{\partial r(\hat{\theta})'}{\partial \theta}$.

iii) Reject $H_0\colon r(\theta_0) = 0_{m\times 1}$ if $\hat{\mathscr{T}} > \text{cv}_n$.

The maximization problem (12) is a so-called semidefinite programming problem, a special case of convex programming. Fast and numerically stable algorithms are available in many computing environments.[10]

The following proposition shows that, for conventional significance levels α, the asymptotic size of this test does not exceed α, regardless of the true correlation structure of the moments. This holds for any valid choice of weight matrix $\hat{W}$, including – but not limited to – the limited-information efficient weight matrix derived in Section 3.2.

Proposition 1. *Impose Assumption 1, except that we redefine* $\lambda \equiv \partial r(\theta_0)'/\partial\theta$ *and require this matrix to have full column rank* m*. Assume also that* $\hat{S} \xrightarrow{p} S$*,* S *is symmetric positive definite, and* $\alpha \leq 0.215$*. Then, if* $r(\theta_0) = 0_{m\times 1}$*,*

$$\limsup_{n\to\infty} P(\hat{\mathscr{T}} > \text{cv}_n) \leq \alpha.$$

Proof. Please see Appendix A.1. □

Remarks.

1. We do not have formal results on how to choose the weight matrix $\hat{S}$ in the test statistic. A pragmatic *ad hoc* choice is to set $\hat{S} = (\lambda'(G'WG)^{-1}G'W\bar{V}WG(G'WG)^{-1}\lambda)^{-1}$ (with consistent estimates plugged in), where $\bar{V} \equiv \text{diag}(V_{11}, \dots, V_{pp})$. Then the test statistic $\hat{\mathscr{T}}$ coincides with the usual Wald test statistic for the case where the moments are asymptotically independent, though the critical value differs.

2. The above test procedure is generally conservative from a minimax perspective, i.e., the size may be strictly smaller than α for all covariance matrices V of the moments.

[10] See our Matlab and Python code suites discussed in Footnote 1.

The reason is that the proof of Proposition 1 relies on a tail probability bound for quadratic forms of Gaussian random vectors (Székely & Bakirov, 2003). This bound is attained when V has rank 1, but the positive semidefinite maximum (12) need not be attained at a rank-1 matrix, to our knowledge. It is an interesting topic for future research to devise a test that has a formal minimax optimality property given the limited knowledge of V.[11]

3. The test procedure is consistent against any fixed alternative with $r(\theta_0) \neq 0_{m\times 1}$ under the conditions of Proposition 1. This follows from the standard argument that $n\hat{\mathscr{T}}$ diverges to infinity with probability 1 in this case, while $\text{cv}_n = O(n^{-1})$ since the largest eigenvalue of any matrix $\tilde{V} \in \mathcal{S}(\text{diag}(V))$ is bounded above by $\sum_{j=1}^p V_{jj}$.

5.2 Over-identification testing

The fit of the calibrated model can be evaluated using over-identification tests when we have more moments p than parameters k. In this subsection we allow for potential model misspecification by dropping the assumption in Section 2 that there exists $\theta_0 \in \mathbb{R}^k$ such that $h(\theta_0) = \mu_0$. Let an arbitrary weight matrix $\hat{W} \overset{p}{\to} W$ be given, such as the limited-information efficient weight matrix derived in Section 3.2. Define the pseudo-true parameter $\tilde{\theta}_0 \equiv \text{argmin}_{\theta\in\mathbb{R}^k}(\mu_0 - h(\theta))'W(\mu_0 - h(\theta))$, assuming the minimizer is unique. We continue to impose all the assumptions in Section 2, with $\tilde{\theta}_0$ substituting for θ_0.

Suppose we want to know whether the model provides a good fit for a particular moment. Let $j^* \in \{1,\dots,p\}$ be the index of the moment of interest. We seek a confidence interval for the model misspecification measure $\mu_{0,j^*} - h_{j^*}(\tilde{\theta}_0)$, i.e., the j^*-th element of $\mu_0 - h(\tilde{\theta}_0)$. It is standard to show that, under Assumption 1,

$$\hat{\mu} - h(\hat{\theta}) - (\mu_0 - h(\tilde{\theta}_0)) = (I_p - G(G'WG)^{-1}G'W)(\hat{\mu} - \mu_0) + o_p(n^{-1/2}). \tag{13}$$

Let $\bar{x}$ be the j^*-th column of the matrix $I_p - \hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'$. Then

$$\left[\hat{\mu}_{j^*} - h_{j^*}(\hat{\theta}) - \Phi^{-1}(1-\alpha/2)\widehat{\text{se}}(\bar{x}), \hat{\mu}_{j^*} - h_{j^*}(\hat{\theta}) + \Phi^{-1}(1-\alpha/2)\widehat{\text{se}}(\bar{x})\right]$$

[11] An alternative valid test rejects whenever $\inf_V nr(\hat{\theta})'(\lambda'(G'WG)^{-1}G'WVWG(G'WG)^{-1}\lambda)^{-1}r(\hat{\theta})$ exceeds the $1-\alpha$ quantile of a chi-squared distribution with m degrees of freedom (this approach was suggested to us by Bo Honoré). That is, we search over V for the smallest conventional Wald test statistic. Unfortunately, this optimization problem appears to be numerically challenging unless p is small.

is a confidence interval for the difference $\mu_{0,j^*} - h_{j^*}(\tilde{\theta}_0)$, with worst-case asymptotic coverage probability $1-\alpha$. Note that it can happen that $\widehat{\text{se}}(\bar{x}) = 0$, in which case it is not possible to test the over-identifying restriction corresponding to the j^*-th moment.

One common use of over-identification testing is to evaluate the estimated model's fit on "non-targeted moments". This corresponds to the special case where the weight matrix $\hat{W}$ zeroes out the corresponding rows and columns of the non-targeted moments, so that the point estimate $\hat{\theta}$ ignores these moments. Note that in this case p continues to denote the total number of moments ("targeted" plus "non-targeted"), and in particular $\hat{G}$ should contain derivatives of both kinds of moments.

A *joint* test of the over-identifying restrictions can be constructed by applying the idea in Section 5.1. Construct the test statistic $\hat{\mathscr{T}}_{\text{overid}} \equiv (\hat{\mu} - h(\hat{\theta}))'\hat{S}(\hat{\mu} - h(\hat{\theta}))$ for some $p \times p$ symmetric positive definite matrix $\hat{S}$ (a natural *ad hoc* choice is $\hat{S} = \hat{W}$, in which case the test statistic equals the minimized minimum distance objective function). We reject correct specification of the model at significance level $\alpha \leq 0.215$ if the test statistic exceeds the critical value

$$\begin{aligned}\text{cv}_{n,\text{overid}} \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \frac{1}{n} \operatorname{trace}\Big(\tilde{V}(I_p - WG(G'WG)^{-1}G')S(I_p - G(G'WG)^{-1}G'W)\Big) \\ \times \Big(\Phi^{-1}(1-\alpha/2)\Big)^2,\end{aligned}$$

where we plug in sample analogues for all the unknown quantities, as in Section 5.1.

6 General knowledge of the covariance matrix

We now consider the general case where any given collection of elements of the asymptotic covariance matrix V of the moments $\hat{\mu}$ is known (or consistently estimable), while the remaining elements are unrestricted. For example, if a pair of elements of $\hat{\mu}$ are known to be independent, the corresponding off-diagonal elements of V must equal zero. Hence, we may wish to impose knowledge of some off-diagonal elements in addition to the diagonal.

Letting $\tilde{\mathcal{S}}$ denote the given constraint set for V, we can compute the worst-case asymptotic standard deviation of $r(\hat{\theta})$ as

$$\sqrt{\max_{\tilde{V} \in \tilde{\mathcal{S}}} x'\tilde{V}x}, \tag{14}$$

where x was defined in Section 3.1.[12] In the case of interest to us, $\tilde{\mathcal{S}}$ is defined by equality restrictions on a subset of the elements of V, in addition to the restriction that V is symmetric positive semidefinite. In this case, the optimization (14) is a (concave) semidefinite programming problem, for which good numerical algorithms exist, as discussed in Section 5.1.

In the over-identified case $p > k$, the worst-case efficient weight matrix W can be computed through two nested convex/concave optimization problems:

$$\min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \tilde{\mathcal{S}}} x(W)'\tilde{V}x(W) = \min_{z \in \mathbb{R}^{p-k}} \max_{\tilde{V} \in \tilde{\mathcal{S}}} \{G(G'G)^{-1}\lambda + G^{\perp}z\}'\tilde{V}\{G(G'G)^{-1}\lambda + G^{\perp}z\}, \quad (15)$$

where $\mathcal{S}_p$, $x(W)$, and $G^{\perp}$ were defined in Section 3.2, and the equality follows from Lemma 2 in Appendix A.1. The inner maximization in (15) is a concave semidefinite program, as discussed in the previous paragraph. The outer minimization is an unconstrained convex program since the objective function is a pointwise maximum of convex functions in z. Once the optimal z has been computed, the corresponding optimal weight matrix is given by the matrix $W(z)$ defined in the proof of Lemma 2 in Appendix A.1.

To conduct joint hypothesis tests as in Section 5, we can simply replace the constraint set in the critical value computation (12) with $\tilde{\mathcal{S}}$.

Special case: Knowledge of the block diagonal. Suppose we know the *block* diagonal of V, while all other elements are unrestricted. That is, suppose the constraint set $\tilde{\mathcal{S}}$ is given by all symmetric positive semidefinite matrices of the form

$$V = \begin{pmatrix} V_{(1)} & ? & ? & \dots & ? \\ ? & V_{(2)} & ? & \dots & ? \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ ? & ? & \dots & ? & V_{(J)} \end{pmatrix}, \quad (16)$$

where $V_{(j)}$ are known (or consistently estimable) square symmetric matrices (possibly of different dimensions) for $j = 1, \dots, J$. Such a structure may occur if consecutive elements of $\hat{\mu}$ are obtained from the same underlying data set, facilitating computation of covariances among these elements. Partition the vector x conformably as $x = (x'_{(1)}, \dots, x'_{(J)})'$. The

[12]Similarly, we could compute the *best-case* variance by minimizing this objective function. If $p = 2$ and V_{12} is unrestricted, the best-case variance is given by $(|x_1|V_{11}^{1/2} - |x_2|V_{22}^{1/2})^2$.

worst-case asymptotic standard deviation (14), for fixed W, is then given by

$$\sqrt{\max_{\tilde{V}\in\tilde{\mathcal{S}}} x'\tilde{V}x} = \sum_{j=1}^{J}(x'_{(j)}V_{(j)}x_{(j)})^{1/2}. \tag{17}$$

This follows from the same logic as in Section 3.1 (see also Lemma 1 in Appendix A.1) once we recognize that the known block diagonal of V implies that the marginal variance of $x'_{(j)}\hat{\mu}_{(j)}$ is known for each $j = 1, \dots, J$, but the correlations among these J variables remain unrestricted.[13] Here we have also partitioned $\hat{\mu} = (\hat{\mu}'_{(1)}, \dots, \hat{\mu}'_{(J)})'$ conformably. In the over-identified case, the worst-case efficient weight matrix can be computed by substituting the formula (17) into the nested optimization (15) (with $x = G(G'G)^{-1}\lambda + G^{\perp}z$).

7 Empirical applications

We illustrate our methods through two empirical examples. First we fit a model of menu cost price setting in multi-product firms to scanner data. Then we fit a heterogeneous agent New Keynesian model to impulse responses that have been estimated from a combination of micro and macro data.

7.1 Menu cost price setting in multi-product firms

Our first application estimates the Alvarez & Lippi (2014) model of menu cost price-setting in multiproduct firms. We fit the model to moments of price changes from supermarket scanner data. This is a small-scale application with $k = 3$ parameters and $p = 4$ moments.

Though we in fact have access to the underlying data set, we emulate a hypothetical situation where the model is matched to moments that were reported in another paper. We can therefore compare full-information inference, which uses the underlying data, with limited-information inference, which uses only the moments and their marginal standard errors. We find that limited-information inference remains informative about the structural parameters. Moreover, a simulation study calibrated to this application confirms the utility of our procedures in finite samples.

[13] The maximum (17) is achieved by $V = \text{Var}(\tilde{\mu})$, where the random vector $\tilde{\mu} = (\tilde{\mu}'_{(1)}, \dots, \tilde{\mu}'_{(J)})'$ has the following representation. Let $\eta = (\eta'_{(1)}, \dots, \eta'_{(J)})'$ have the covariance matrix (16), but with zeros instead of question marks. Let $\bar{\eta}$ be a scalar random variable with variance 1 that is uncorrelated with η. Then set $\tilde{\mu}_{(j)} \equiv \frac{1}{\sqrt{x'_{(j)}V_{(j)}x_{(j)}}}V_{(j)}x_{(j)}\bar{\eta} + (I - \frac{1}{x'_{(j)}V_{(j)}x_{(j)}}V_{(j)}x_{(j)}x'_{(j)})\eta_{(j)}$, $j = 1, \dots, J$.

Model. We give a brief overview of the structural model here and refer the reader to Alvarez & Lippi (2014) for details. A firm sets prices on N products. The *desired* log prices for the products evolve in continuous time as N independent Brownian motions (without drift); however, the actual prices are fixed until the firm pays a fixed menu cost, at which point it may reset all N prices simultaneously. The firm's profit depends negatively on the squared log deviation between the current and desired price, integrated over time and averaged across the N products. The model has $k = 3$ parameters: the number N of products, the volatility of the desired prices, and the scaled menu cost (relative to the curvature of the profit function).[14] Alvarez & Lippi (2014) derive closed-form expressions for the frequency of price changes and the moments of the size of price changes. These are the moments we will match in the data.

Data. We empirically estimate the frequency and moments of price changes in scanner data from the supermarket chain Dominick's.[15] As described in detail in Appendix A.3, we clean the data following Alvarez, Le Bihan & Lippi (2016), and in particular we focus on data from a single store. Unlike those authors, we exclusively use data on beer products, which arguably increases the interpretability of the results and makes the sample size more relevant for our subsequent simulation study. The final data set contains weekly prices on 499 beer products (Universal Product Codes, henceforth UPCs), observed for an average of 76 weeks per UPC. The total sample size is $n = 37{,}916$. When computing standard errors, we treat the price changes as i.i.d. across UPCs and time.

The $p = 4$ reduced-form moments that we match to the structural model are the average number of price changes per week as well as the empirical first, second, and fourth moments of the absolute log price changes (conditional on a nonzero change).[16] We estimate the full-information covariance matrix of these moments using the usual nonparametric estimate (which depends on sample moments of price changes up to order 8). When applying our limited-information procedures, we use only the diagonal of this covariance matrix.

Results. We consider both just-identified and efficient specifications. We treat the number N of products as a parameter to be estimated, since there may not be a perfect correspon-

[14] In the notation of Alvarez & Lippi (2014), these parameters are n, σ, and $\sqrt{\psi/B}$, respectively.

[15] The data is provided by the James M. Kilts Center, University of Chicago Booth School of Business.

[16] Before computing moments, we subtract off the overall average log price change (conditional on a nonzero change), since the Alvarez & Lippi (2014) model abstracts from inflation.

Price setting application: Parameter estimates

	Just-identified specification				Efficient specification		
	# prod.	Vol.	Menu cost	Over-ID	# prod.	Vol.	Menu cost
Full-info	3.012	0.090	0.291	0.002	3.255	0.089	0.305
	(0.046)	(0.001)	(0.003)	(0.000)	(0.051)	(0.001)	(0.003)
Independ.	3.012	0.090	0.291	0.002	2.829	0.090	0.280
	(0.167)	(0.001)	(0.010)	(0.001)	(0.091)	(0.000)	(0.006)
Worst case	3.012	0.090	0.291	0.002	2.786	0.090	0.278
	(0.235)	(0.001)	(0.016)	(0.002)	(0.148)	(0.001)	(0.011)

Table 1: Estimates for the just-identified specification (uses only three moments for estimation) and the efficient specification (exploits all four moments for estimation). The rows correspond to full-information inference (exploits knowledge of $\hat{V}$), inference under independence (erroneously assumes that $\hat{V}$ is diagonal), and worst-case inference (exploits only diagonal of $\hat{V}$ without assuming off-diagonal elements are zero). Parameters: number of products ("# prod."), volatility of desired log price ("Vol."), scaled menu cost ("Menu cost"). Column "Over-ID" displays the error in fitting the non-targeted mean absolute price change moment, given the just-identified parameter estimates. Standard errors in parentheses.

dence between a UPC and the structural model's notion of a "product". The just-identified specification uses a weight matrix that attaches zero weight to the first moment of absolute price changes (i.e., the average), so that the three parameters are estimated from three moments (this estimator is available in closed form). We can then check whether the model provides a good fit for the "non-targeted" moment by carrying out the over-identification test proposed in Section 5.2. The efficient specification exploits all four empirical moments, using either the conventional full-information procedure or our limited-information procedure in Section 3.2.[17] In addition to the full-information and limited-information procedures, we report results for a procedure that (erroneously) assumes that the four empirical moments are mutually independent.

Table 1 shows that the limited-information standard errors are larger than the full-information ones, but they remain highly informative about the values of the structural parameters. In the efficient over-identified specification, the worst-case standard errors are at most 3.7 times larger than the corresponding full-information values.[18] Importantly, all

[17] The efficient estimates are computed using a one-step update of the just-identified estimates.

[18] The worst-case efficient standard errors are computed separately for each parameter, i.e., setting $r(\theta) = \theta_i$ separately for $i = 1, 2, 3$.

worst-case standard errors are arguably small relative to the economic magnitudes of the parameter estimates. Hence, taking a worst-case perspective still allows for informative inference. In this particular application, the standard errors that assume independence are mostly intermediate between the full-information and limited-information values.

Though limited-information inference is informative about the structural parameters themselves, there is a price to pay for the over-identification test in this application. In particular, Table 1 shows that the limited-information test does not reject the validity of the non-targeted moment restriction, whereas the full-information test does reject (however, the economic magnitude of the moment violation is small, as the empirical moment equals 0.145 but the error in fitting the moment is only 0.002). This illustrates the principle that full-information inference is usually preferable if it is practically feasible. Of course, if we did not have access to the underlying supermarket scanner data, there would be no alternative to the limited-information analysis.

Simulation study. In Appendix A.3 we show that our inference procedures perform well in a simulation study calibrated to the present empirical application. We simulate data from the Alvarez & Lippi (2014) model conditional on the estimated structural parameters. While our limited-information tests and confidence intervals have approximately correct size/coverage given the empirical sample size n (as do the full-information procedures), the procedures that erroneously assume independence between the reduced-form moments can over-reject/under-cover.

7.2 Heterogeneous agent New Keynesian model

Our second application estimates a heterogeneous agent New Keynesian general equilibrium macro model, following McKay et al. (2016) and Auclert et al. (2021). The matched moments are impulse response functions of macro time series and cross-sectional micro moments with respect to identified productivity and monetary policy shocks, as estimated by Chang, Chen & Schorfheide (2021) and Miranda-Agrippino & Ricco (2021). This is a medium-scale application with $k = 7$ parameters and $p = 23$ moments.

Though less efficient, impulse response matching estimation is more robust to modeling assumptions than full-information likelihood estimation.[19] This is because – in the first-order

[19]Likelihood procedures for estimation of heterogeneous agent models have been proposed by Mongey & Williams (2017), Winberry (2018), Liu & Plagborg-Møller (2020), and Auclert et al. (2021) among others.

approximation we consider – the impulse responses with respect to a monetary shock, say, do not depend on the exogenous processes for the other disturbances (e.g., shocks to the household discount rate).[20] Thus, our application only requires us to specify and estimate the exogenous processes for productivity and monetary disturbances. We remain agnostic about the number and nature of other shocks that may be driving the economy.

Likewise, our limited-information approach is simpler and less restrictive than other types of procedures that attempt to exploit more information. The only data inputs into our procedure are the impulse response point estimates and confidence intervals reported by Chang et al. (2021) and Miranda-Agrippino & Ricco (2021). We do not need access to the underlying data used in those papers, as would be required if one were to estimate the joint covariance matrix of all empirical moments via the bootstrap or GMM calculations. Unlike approaches based on bootstrapping or simulating data, we do not need to (repeatedly) re-run the impulse response estimation routines, and we do not need to fully model the relationship between the macro and micro data used by Chang et al. (2021) (e.g., by specifying all shocks).

Model. We employ the one-asset heterogeneous agent New Keynesian model described in Auclert et al. (2021, Appendix B.2); we refer to that paper for details.[21] Following McKay et al. (2016), the model features a continuum of heterogeneous households facing uninsurable idiosyncratic earnings risk. The households choose their work hours and amount of savings in a nominal Treasury bond.[22] Monopolistically competitive firms set prices subject to a quadratic adjustment cost, yielding a New Keynesian Phillips curve. Households receive lump sum distributions of government interest revenue and firm profits. The central bank sets the nominal interest rate according to a Taylor rule that depends on inflation.

The model is solved through a first-order linearization, using the numerical procedures developed by Auclert et al. (2021). We study responses to (i) a productivity shock to the exogenous AR(2) process for the log growth rate of firms' total factor productivity (TFP), and (ii) a monetary shock to an exogenous AR(2) process that enters as an additive disturbance in the Taylor rule.

[20] Impulse responses to a monetary shock are computed by holding fixed all other exogenous shocks. As a result, the linearized impulse responses depend only on parameters of the monetary disturbance process as well as model parameters that govern the endogenous transmission mechanisms.

[21] The only difference from their paper is that our monetary policy rule depends only on inflation and not output, as in the excellent GitHub repository produced by Auclert et al. (2021), which we rely on: `https://github.com/shade-econ/sequence-jacobian`

[22] It would be feasible to estimate the two-asset model in Auclert et al. (2021, Appendix B.3), which is in the spirit of Kaplan, Moll & Violante (2018), but we stick to the simpler one-asset model for clarity.

Following Auclert et al. (2021), we limit ourselves to estimating structural parameters that do not affect the steady state of the model. This allows us to avoid repeatedly recomputing the steady state, though this would be feasible to do with moderate computational effort. The steady state parameters are fixed at the values assumed by Auclert et al. (2021, Table B.2). The $k = 7$ estimated parameters are: the Taylor rule coefficient on inflation, the slope of the Phillips curve, the three parameters in the AR(2) process for TFP, and the two autoregressive coefficients for the monetary disturbance.[23]

Data. The empirical moments are obtained from two sets of Structural Vector Autoregression estimates of impulse responses to identified shocks.

Impulse responses with respect to TFP shocks are obtained from Chang et al. (2021, Fig. 7 and 9, blue lines). We use the responses of TFP itself and of GDP (output in the model), as well as the response of a cross-sectional moment estimated using data from the Current Population Survey (CPS): the fraction of people earning less than 2/3 of per capita GDP.[24] The sophisticated estimation method of Chang et al. (2021) takes into account statistical uncertainty arising from the limited sample sizes in the CPS. By relying directly on their reported results, our analysis inherits this desirable feature.

Impulse responses with respect to monetary shocks are obtained from Miranda-Agrippino & Ricco (2021, Fig. 3). We use the responses of industrial production (output in the model), the consumer price index (price level in the model), and the 1-year Treasury rate (annualized nominal interest rate in the model). Since our structural model is quarterly but the Miranda-Agrippino & Ricco (2021) data is monthly, we use the end-of-quarter impulse responses.

We focus on four impulse response horizons: the impact horizon, and the 1-, 2-, and 8-quarter horizons. When matching the model to the data, we take into account that the Chang et al. (2021) responses are with respect to a one-standard-deviation shock, while the Miranda-Agrippino & Ricco (2021) responses are normalized so that the Treasury rate increases by 100 basis points on impact.[25] Since both papers report Bayesian posterior quantiles, we appeal to the Bernstein-von Mises theorem and define the point estimates to be the reported posterior medians, while the standard errors are those implied by a normal approximation of the reported credible intervals.[26] In total, we have $p = 23$ empirical

[23] We do not need to estimate the standard deviation of the monetary shock, since this parameter does not affect the normalized impulse responses that we match (see below).

[24] The factor 2/3 approximately adjusts for the average labor share, see Chang et al. (2021, Sec. 5.1).

[25] Chang et al. (2021) actually consider a 3-standard-deviation shock, but we divide by 3.

[26] Thus, if the length of the $1 - \alpha$ credible interval for θ_j is $\hat{L}_j$, we set $\hat{\sigma}_j = \hat{L}_j/(2\Phi^{-1}(1 - \alpha/2))$.

Heterogeneous agent application: Parameter estimates

Weight matrix	TR	PC	TFP AR1	TFP AR2	TFP Std	Monetary AR1	Monetary AR2
Diagonal	1.409	0.010	0.076	-0.132	0.007	0.713	0.075
	(4.243)	(0.012)	(0.237)	(0.377)	(0.001)	(0.223)	(0.185)
Efficient	1.583	0.017	0.060	-0.078	0.007	0.723	0.014
	(3.012)	(0.010)	(0.192)	(0.282)	(0.000)	(0.170)	(0.149)

Table 2: Structural parameter estimates with diagonal weight matrix (top row) and efficient weighting (bottom row). Parameters: Taylor rule coefficient on inflation ("TR"); slope of Phillips curve ("PC"); first and second autoregressive ("AR1" and "AR2") and standard deviation ("Std") parameters of TFP and monetary disturbance processes. Worst-case standard errors in parentheses.

moments, as we discard the impact response of the bond rate, which is normalized to 1.

Results. The top row of Table 2 shows the parameter estimates obtained by using a diagonal weight matrix with $W_{jj} = 1/\hat{\sigma}_j^2$. The Taylor rule coefficient on inflation is estimated to be 1.41 with a large standard error. The Phillips curve is positively sloped but statistically insignificant at conventional significance levels. The TFP growth process is estimated to be close to white noise, while the monetary disturbance process has some persistence.

Figure 2 compares the model-implied and empirical impulse responses, at the parameter estimates discussed in the previous paragraph. We see that the model-implied impulse responses of output to a monetary shock are too small in magnitude relative to the data at the 2- and 8-quarter horizons, while the opposite is true for the responses of the price level with respect a monetary shock. To test whether these disparities are too large to be explained by statistical noise, we conduct the over-identification test proposed in Section 5.2. The vertical error bars in Figure 2 show the 90% confidence intervals for the *differences* between model-implied and empirical moments, centered at the empirical moments for visual convenience. We see that none of the model-implied impulse responses fall outside their respective intervals, meaning that we cannot reject the validity of each moment individually. The joint test of the validity of all $p = 23$ moments also does not reject at the 10% level.[27]

The efficient estimation results in the bottom row of Table 2 demonstrate the benefit of optimally weighting the moments as described in Section 3.2.[28] In particular, the t-statistic

[27] The test statistic equals $\hat{\mathscr{T}} = 21.57$, while the critical value equals $\widehat{\text{cv}}_n = 58.12$. Our choice of $\hat{S}$ follows the suggestion in the remark below Proposition 1.

[28] The efficient estimates are computed via a one-step update, cf. Section 3.2, separately for each parameter.

Heterogeneous agent application: Impulse responses

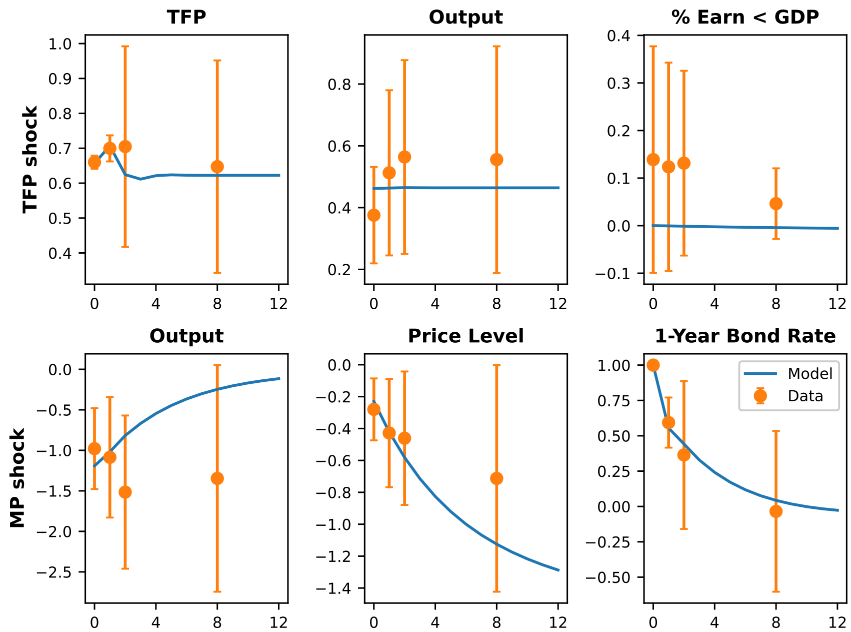

Figure 2: Model-implied impulse responses (thin curves) and corresponding empirical estimates (circles), with respect to a one-standard-deviation TFP shock (top row) or a monetary policy shock that raises the bond rate by 1 percentage point on impact (bottom row). Figure titles: response variables ("% Earn < GDP": fraction of people earning less than 2/3 of per capita GDP). Vertical axis units: percentage points. Horizontal axis units: quarters. Vertical error bars: (shifted) 90% confidence intervals for the *differences* between the empirical and model-implied moments (not confidence intervals for the empirical moments themselves).

for the slope of the New Keynesian Phillips curve increases to 1.73, from 0.77 previously. Thus, our limited-information approach yields moderately informative inference about a parameter that is often viewed as difficult to pin down in the data (Mavroeidis, Plagborg-Møller & Stock, 2014). More generally, the efficient standard errors in the bottom row of Table 2 are 19–29% smaller than the non-efficient ones in the top row.

Table 5 in Appendix A.4 shows that the optimal moment selection procedure in Section 3.2 chooses to estimate the three TFP process parameters solely off the responses of TFP with respect to a TFP shock, while the other four parameters are estimated off a combination of the three impulse response functions with respect to the monetary shock. Hence, though in principle the output responses with respect to a TFP shock are informative about five of the seven parameters, those moments are too imprecisely estimated to be useful for the purpose of limited-information efficient estimation. The same is true for the responses of the fraction of people with low earnings.

8 Conclusion

We computed simple, sharp, and informative upper bounds on the standard errors of structural parameter estimates when the correlation structure of the matched empirical moments is not fully known. In addition, we proposed an efficient moment weighting procedure in the over-identified case, as well as valid tests of parameter restrictions and over-identifying restrictions. The required inputs are minimal: Other than being able to evaluate the mapping from structural parameters to model-implied moments (at least numerically), we just need the empirical moment estimates and their individual standard errors. Our procedures are computationally tractable even in settings with many moments and/or parameters. A code suite is available online (see Footnote 1).

We believe our limited-information approach is useful for applied researchers who match their models to moments obtained from several different data sources, estimation methods, or previous papers. Our methods obviate the need to estimate the correlation structure across the various moments, which is sometimes difficult or impossible. Even when the moment correlation structure is in principle estimable, our methods may be helpful, since marginal standard errors for individual moments are typically much easier to obtain from standard econometric software than it is to figure out the joint distribution of all moments, as illustrated in our empirical applications. If nothing else, the limited-information procedures can be used to gauge whether it is worthwhile to expend the additional effort required for full-information analysis.

Our work points to several future research directions. First, the joint test of multiple parameter restrictions we propose does not have a formal minimax optimality property; we hope future research will explore the optimality question. Second, it may be interesting to extend our procedures to settings where the moments have a mixed normal distribution as in Hahn et al. (2020b). Third, our worst-case standard error formula may be theoretically useful for analyzing treatment effect estimators in design-based causal inference when some aspects of the assignment mechanism are unknown (yielding an unknown correlation structure between certain terms). Finally, the worst-case standard errors might also be helpful for variational Bayesian inference when it is computationally challenging to characterize the posterior dependence across parameters.

A Appendix

A.1 Technical lemmas and proofs

Here we state and prove two technical lemmas referred to in the main text, and we provide the proof of Proposition 1.

Lemma 1. *Let $x = (x_1, \dots, x_p)' \in \mathbb{R}^p$ and $\sigma_1^2, \dots, \sigma_p^2 \geq 0$. Let $\mathcal{S}(\sigma)$ denote the set of $p \times p$ symmetric positive semidefinite matrices with diagonal elements $\sigma_1^2, \dots, \sigma_p^2$. Then*

$$\max_{V \in \mathcal{S}(\sigma)} \sqrt{x'Vx} = \sum_{j=1}^{k} \sigma_j |x_j|.$$

Proof. The right-hand side is attained by $V = ss'$, where $s = (\sigma_1 \operatorname{sign}(x_1), \dots, \sigma_p \operatorname{sign}(x_p))'$. Moreover, for any $V \in \mathcal{S}(\sigma)$,

$$x'Vx = \sum_{j=1}^{p} \sum_{\ell=1}^{p} x_j x_\ell V_{j\ell} \leq \sum_{j=1}^{p} \sum_{\ell=1}^{p} |x_j x_\ell| \, |V_{j\ell}| \leq \sum_{j=1}^{p} \sum_{\ell=1}^{p} |x_j x_\ell| \, \sigma_j \sigma_\ell = \left(\sum_{j=1}^{p} \sigma_j |x_j| \right)^2,$$

where the penultimate inequality uses that $|V_{j\ell}|^2 \leq V_{jj} V_{\ell\ell}$ for any symmetric positive semidefinite matrix V. □

Lemma 2. *Assume $p, k \in \mathbb{N}$ and $p > k$. Let $\lambda \in \mathbb{R}^k$, and let $G \in \mathbb{R}^{p \times k}$ have full column rank. Let $G^\perp$ denote any $p \times (p-k)$ matrix with full column rank such that $G'G^\perp = 0_{k \times (p-k)}$. Let $\mathcal{S}_p$ denote the set of $p \times p$ symmetric positive semidefinite matrices W such that $G'WG$ is nonsingular. Then*

$$\left\{ WG(G'WG)^{-1}\lambda \colon W \in \mathcal{S}_p \right\} = \{x \colon x \in \mathbb{R}^p, \; G'x = \lambda\} = \left\{ G(G'G)^{-1}\lambda + G^\perp z \colon z \in \mathbb{R}^{p-k} \right\}.$$

Proof. We first show that

$$\left\{ G(G'G)^{-1}\lambda + G^\perp z \colon z \in \mathbb{R}^{p-k} \right\} \subset \left\{ WG(G'WG)^{-1}\lambda \colon W \in \mathcal{S}_p \right\}. \tag{18}$$

Pick any $z \in \mathbb{R}^{p-k}$, and define

$$W(z) \equiv (G, G^\perp) \begin{pmatrix} I_k & \tilde{\lambda} z' \\ z \tilde{\lambda}' & \delta I_{p-k} \end{pmatrix} \begin{pmatrix} G' \\ G^{\perp\prime} \end{pmatrix}, \quad \tilde{\lambda} \equiv \frac{1}{\lambda'(G'G)^{-1}\lambda} \lambda,$$

where $\delta > 0$ is arbitrary but chosen large enough so that $W(z)$ is positive semidefinite. Then

$$W(z)G = (G, G^{\perp}) \begin{pmatrix} I_k & \tilde{\lambda} z' \\ z\tilde{\lambda}' & \delta I_{p-k} \end{pmatrix} \begin{pmatrix} G'G \\ 0_{(p-k)\times k} \end{pmatrix} = (G + G^{\perp} z \tilde{\lambda}')G'G.$$

Hence,

$$G'W(z)G = (G'G)^2,$$

implying

$$W(z)G(G'W(z)G)^{-1}\lambda = (G + G^{\perp} z\tilde{\lambda}')(G'G)^{-1}\lambda = G(G'G)^{-1}\lambda + G^{\perp} z,$$

and thus the statement (18) holds.

Now pick any $W \in \mathcal{S}_p$. Then $x = WG(G'WG)^{-1}\lambda$ satisfies $G'x = \lambda$. This shows that

$$\left\{WG(G'WG)^{-1}\lambda \colon W \in \mathcal{S}_p\right\} \subset \{x \colon x \in \mathbb{R}^p,\ G'x = \lambda\}. \tag{19}$$

Finally, choose any $x \in \mathbb{R}^p$ satisfying $G'x = \lambda$. Since the columns of G and $G^{\perp}$ are (jointly) linearly independent, there exist $y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{p-k}$ such that $x = Gy + G^{\perp} z$. Note that $\lambda = G'x = G'Gy$, so necessarily $y = (G'G)^{-1}\lambda$. We have thus shown that

$$\{x \colon x \in \mathbb{R}^p,\ G'x = \lambda\} \subset \left\{G(G'G)^{-1}\lambda + G^{\perp} z \colon z \in \mathbb{R}^{p-k}\right\}. \tag{20}$$

The set inclusions (18)–(20) together imply the statement of the lemma. □

Proof of Proposition 1. Under the null hypothesis,

$$\sqrt{n} r(\hat{\theta}) \xrightarrow{d} \lambda'(G'WG)^{-1}G'WV^{1/2}Z,$$

where $V^{1/2}V^{1/2\prime} = V$, and $Z = (Z_1, \ldots, Z_p)' \sim N(0_{p\times 1}, I_p)$. The asymptotic null distribution of the test statistic $\hat{\mathscr{T}}$ is therefore a Gaussian quadratic form:

$$n\hat{\mathscr{T}} \xrightarrow{d} Z'QZ, \quad Q \equiv V^{1/2\prime}WG(G'WG)^{-1}\lambda S\lambda'(G'WG)^{-1}G'WV^{1/2}.$$

Székely & Bakirov (2003) prove that

$$P(Z'QZ \le \text{trace}(Q) \times \tau) \le P(Z_1^2 \le \tau) \tag{21}$$

for *any* $p \times p$ symmetric positive semidefinite (non-null) matrix Q and any $\tau > 1.5365$. Since $(\Phi^{-1}(1-\alpha/2))^2 > 1.5365$ for $\alpha \leq 0.215$, it follows that, under the null,

$$\begin{aligned} P(\hat{\mathscr{T}} \leq \mathrm{cv}_n) &\geq P\left(n\hat{\mathscr{T}} \leq \mathrm{trace}(Q) \times (\Phi^{-1}(1-\alpha/2))^2\right) \\ &\to P\left(Z'QZ \leq \mathrm{trace}(Q) \times (\Phi^{-1}(1-\alpha/2))^2\right) \\ &\leq P\left(Z_1^2 \leq (\Phi^{-1}(1-\alpha/2))^2\right) \\ &= 1-\alpha. \end{aligned}$$

□

A.2 Details of the analytical illustration

We here provide detailed derivations that map the illustrative toy example in Section 4.2 into the general framework and notation in Section 3.

Recall that the general median regression in Section 3.2 can be written

$$\min_{z \in \mathbb{R}^{p-k}} \Psi(z;\lambda), \quad \Psi(z;\lambda) \equiv \iota' \,\mathrm{diag}(V)^{1/2} \left| G(G'G)^{-1}\lambda + G^{\perp} z \right|,$$

where $G^{\perp}$ is the $p \times (p-k)$ matrix of eigenvectors of $I_p - G(G'G)^{-1}G'$ corresponding to its nonzero eigenvalues, ι is the p-dimensional vector with all elements equal to 1, $\mathrm{diag}(V)^{1/2}$ is the $p \times p$ diagonal matrix with diagonal elements $(\sigma_1, \ldots, \sigma_p)$, and the absolute value in the final expression is taken elementwise. In the toy example we have $p - k = 1$, so the minimization is over a scalar z.

The form of G in the toy example implies

$$\begin{aligned} G(G'G)^{-1} &= \frac{1}{a^2c^2 + a^2d^2 + b^2d^2} \begin{pmatrix} ac^2 + ad^2 & -abc \\ bd^2 & a^2c \\ -bcd & a^2d + b^2d \end{pmatrix}, \\ I_p - G(G'G)^{-1}G' &= \frac{1}{a^2c^2 + a^2d^2 + b^2d^2} \begin{pmatrix} b^2d^2 & -abd^2 & abcd \\ -abd^2 & a^2d^2 & -a^2cd \\ abcd & -a^2cd & a^2c^2 \end{pmatrix}. \end{aligned}$$

An eigenvector of $I_p - G(G'G)^{-1}G'$ with eigenvalue 1 is given by

$$G^{\perp} = \frac{1}{a^2c^2 + a^2d^2 + b^2d^2} \begin{pmatrix} bd \\ -ad \\ ac \end{pmatrix}.$$

Hence, we can write the median regression objective function as

$$\Psi(z;\lambda) = \frac{1}{a^2c^2+a^2d^2+b^2d^2}\iota' \operatorname{diag}(V)^{1/2}\left|\begin{pmatrix} ac^2+ad^2 & -abc \\ bd^2 & a^2c \\ -bcd & a^2d+b^2d \end{pmatrix}\lambda + \begin{pmatrix} bd \\ -ad \\ ac \end{pmatrix} z\right|.$$

For $\lambda = (1,0)'$,

$$\begin{aligned}\Psi(z;(1,0)') &= \frac{1}{a^2c^2+a^2d^2+b^2d^2}\Big(\sigma_1\Big|ac^2+ad^2+bdz\Big| + \sigma_2\Big|bd^2-adz\Big| + \sigma_3\Big|-bcd+acz\Big|\Big) \\ &= \frac{1}{a^2c^2+a^2d^2+b^2d^2}\left(\sigma_1|bd|\left|\frac{a(c^2+d^2)}{bd}+z\right| + \Big(\sigma_2|ad|+\sigma_3|ac|\Big)\left|\frac{bd}{a}-z\right|\right).\end{aligned}$$

This latter rewritten objective function makes clear how we can usefully reduce the dimension of the problem. In particular, we will characterize the solution to the problem

$$\min_z \tilde{\Psi}(z) \equiv \min_z \varsigma_1\left|\frac{\beta}{\alpha}+z\right| + \varsigma_2|\alpha - z|, \quad \text{where} \quad \varsigma_1, \varsigma_2, \beta \in [0,\infty),\ \alpha \in (-\infty,\infty).$$

Any parameters $(\sigma_1,\sigma_2,\sigma_3,a,b,c,d)$ in the original problem map into certain parameters $(\varsigma_1,\varsigma_2,\beta,\alpha)$ in this reduced problem.

Since the function $\tilde{\Psi}(z)$ is piecewise linear, the minimum will be at a point where the slope changes. Therefore, we need only check the points $z=-\beta/\alpha$ and $z=\alpha$:

$$\tilde{\Psi}(-\beta/\alpha) = \varsigma_2\left|\alpha+\frac{\beta}{\alpha}\right|, \quad \tilde{\Psi}(\alpha) = \varsigma_1\left|\frac{\beta}{\alpha}+\alpha\right|.$$

Thus, the solution of the reduced problem is given by

$$\tilde{\Psi}(z^*) = \begin{cases} \varsigma_1\left|\alpha+\frac{\beta}{\alpha}\right| & \text{if } \varsigma_1 \le \varsigma_2, \\ \varsigma_2\left|\alpha+\frac{\beta}{\alpha}\right| & \text{if } \varsigma_1 > \varsigma_2, \end{cases} \qquad z^* = \begin{cases} \alpha & \text{if } \varsigma_1 \le \varsigma_2, \\ -\frac{\beta}{\alpha} & \text{if } \varsigma_1 > \varsigma_2. \end{cases}$$

We can map back into a solution of the original problem:

$$\begin{aligned}\Psi(z^*;(1,0)') &= \frac{1}{a^2c^2+a^2d^2+b^2d^2}\times\begin{cases} \sigma_1|bd|\left|\frac{bd}{a}+\frac{a(c^2+d^2)}{bd}\right| & \text{if } \sigma_1|bd| \le \sigma_2|ad|+\sigma_3|ac|, \\ (\sigma_2|ad|+\sigma_3|ac|)\left|\frac{bd}{a}+\frac{a(c^2+d^2)}{bd}\right| & \text{if } \sigma_1|bd| > \sigma_2|ad|+\sigma_3|ac|, \end{cases} \\ z^* &= \begin{cases} \frac{bd}{a} & \text{if } \sigma_1|bd| \le \sigma_2|ad|+\sigma_3|ac|, \\ -\frac{a(c^2+d^2)}{bd} & \text{if } \sigma_1|bd| > \sigma_2|ad|+\sigma_3|ac|. \end{cases}\end{aligned}$$

This implies that the efficient linear combination is

$$x^* = G(G'G)^{-1}\lambda + G^\perp z^* = \begin{cases} \begin{pmatrix} \frac{1}{a} \\ 0 \\ 0 \end{pmatrix} & \text{if } \sigma_1|bd| \le \sigma_2|ad| + \sigma_3|ac|, \\ \begin{pmatrix} 0 \\ \frac{1}{b} \\ -\frac{c}{bd} \end{pmatrix} & \text{if } \sigma_1|bd| > \sigma_2|ad| + \sigma_3|ac|, \end{cases}$$

and the worst-case efficient estimator of $r(\theta_0) = \lambda'\theta_0$ is

$$r(\hat{\theta}) = (x^*)'\hat{\mu} = \begin{cases} \frac{1}{a}\hat{\mu}_1 & \text{if } \sigma_1|bd| \le \sigma_2|ad| + \sigma_3|ac|, \\ \frac{1}{b}\hat{\mu}_2 - \frac{c}{bd}\hat{\mu}_3 & \text{if } \sigma_1|bd| > \sigma_2|ad| + \sigma_3|ac|. \end{cases}$$

This confirms the heuristic derivations in Section 4.2.

A.3 Details of the price setting application and simulation study

Here we provide details of the data used for the empirical application in Section 7.1, and we conduct a simulation study calibrated to this application.

Data. We use the "movement" data set for beer products (file name `wber.csv`) on the Chicago Booth website.[29] We follow Alvarez et al. (2016) when cleaning the data. First, we keep only data for store #122. Second, we drop any observations with prices below 20 cents or above 25 dollars (the data was collected between the years 1989 and 1994). Third, we set any absolute price changes below one cent equal to zero. Fourth, we drop the largest 1% of absolute log price changes.

Table 3 shows the $p = 4$ estimated reduced-form moments, their standard errors, and their estimated correlation matrix. The sample kurtosis (fourth moment divided by squared second moment) of log price changes equals 1.80. The zero correlation between the sample frequency of nonzero price changes (a binary outcome) and the sample moments of the price change magnitudes is mechanical. Our limited-information analysis here does not exploit this fact because we want to emulate what an applied researcher might do without thinking

[29] https://www.chicagobooth.edu/research/kilts/datasets/dominicks

Price setting application: Reduced-form moment estimates

Moment	Estimate	Std. error	Pairwise correlation $E[(\Delta p)^2]$	$E[(\Delta p)^4]$	$E[\|\Delta p\|]$
Frequency	0.293	0.002	0.000	0.000	0.000
$E[(\Delta p)^2]$	0.027	0.000		0.939	0.966
$E[(\Delta p)^4]$	0.001	0.000			0.831
$E[\|\Delta p\|]$	0.145	0.001			

Table 3: Reduced-form moment estimates and their standard errors: average weekly rate of nonzero price changes (Frequency), and moments $E[|\Delta p|^j]$, $j = 1, 2, 4$ of absolute log price changes conditional on a nonzero change. "Pairwise correlation" columns: estimated pairwise correlations across the sample moments.

hard about the problem. However, the independence could be taken into account using the extensions described in Section 6.

Simulation study. We apply the inference methods to data simulated from the Alvarez & Lippi (2014) model. The simulations treat the just-identified empirical parameter estimates (columns 1–3 in Table 1) as the truth, and we use a sample size of $n = 37{,}916$ as in the real data. The binary price change indicators are drawn i.i.d. from a binomial distribution with the model-implied success probability (Alvarez & Lippi, 2014, Proposition 4). The magnitudes of the price changes are drawn from the model-implied density function (Alvarez & Lippi, 2014, Proposition 6).[30] We use 10,000 Monte Carlo repetitions. The estimation and inference procedures are the same as the ones applied to the actual data (in particular, efficient estimates are computed using the one-step approach).

Table 4 shows that the just-identified and efficient limited-information confidence intervals have coverage probabilities very nearly equal to or exceeding the nominal level of 95% for all three parameters. Though coverage is conservative, the table shows that the average length of the confidence intervals is not more than six times that of the corresponding full-information confidence intervals. This is consistent with the empirical standard errors reported in Section 7.1.

Table 4 also illustrates that the worst-case perspective is key to avoiding over-rejection in the face of limited information: Both the "efficient" t-test for the price volatility parameter

[30] We simulate from this density by numerically computing the associated quantile function on a fine grid, and then passing random uniform draws through a cubic interpolation of this function.

Monte Carlo simulation study

	Just-identified specification			Efficient specification		
	# prod.	Vol.	Menu cost	# prod.	Vol.	Menu cost
	Confidence interval coverage rate					
Full-info	94.5%	95.0%	94.7%	95.0%	95.2%	95.3%
Independence	100.0%	95.0%	100.0%	100.0%	89.1%	100.0%
Worst case	100.0%	99.4%	100.0%	100.0%	99.4%	100.0%
	Confidence interval average length					
Full-info	0.179	0.002	0.010	0.162	0.002	0.009
Independence	0.627	0.002	0.039	0.390	0.002	0.025
Worst case	0.878	0.003	0.059	0.571	0.003	0.041
	RMSE relative to true parameter values					
Full-info	1.53%	0.59%	0.86%	1.37%	0.58%	0.76%
Independence	1.53%	0.59%	0.86%	1.72%	0.59%	0.99%
Worst case	1.53%	0.59%	0.86%	1.79%	0.59%	1.03%
	Rejection rate of over-identification test					
Full-info		5.01%				
Independence		0.00%				
Worst case		0.00%				
	Rejection rate of joint test of true parameter values					
Full-info		4.79%				
Independence		7.54%				
Worst case		2.47%				

Table 4: Simulation results based on the empirically calibrated Alvarez & Lippi (2014) model. The just-identified specification uses only three moments for estimation, while the efficient specification exploits all four moments. The rows correspond to full-information inference (exploits knowledge of $\hat{V}$), inference under independence (erroneously assumes that $\hat{V}$ is diagonal), and worst-case inference (exploits only diagonal of $\hat{V}$ without assuming off-diagonal elements are zero). Estimated parameters: number of products (# prod.), volatility of desired log price (Vol.), scaled menu cost (Menu cost). The over-identification test tests the validity of the fourth non-targeted moment. The joint test of the true parameters is a Wald test (Full-info or Independence) or the test proposed in Section 5.1 (Worst case). The nominal significance level is 5%.

and the joint Wald test of the true parameter values over-reject if we erroneously assume that all the empirical moments are independent of each other. In contrast, the limited-information and full-information t-tests and joint tests are correctly sized.[31] The limited-information tests are conservative (as predicted by theory), though the joint test of parameter restrictions is only mildly conservative in this particular model.

Finally, Table 4 shows that the limited-information efficient point estimates have slightly higher root mean squared error (RMSE) than the efficient full-information estimates. It may seem surprising at first blush that the limited-information efficient estimates can have (marginally) higher RMSE than the just-identified estimates. This is because the limited-information efficient estimates are designed to have low variance under the *worst-case* correlation structure (i.e., perfect correlation of the moments), not under the true correlation structure that is unknown to the econometrician.

A.4 Details of the heterogeneous agent application

We here provide further details on the application in Section 7.2. Table 5 shows which impulse response moments are used to efficiently estimate the seven structural parameters, according to the moment selection procedure described in Section 3.2. The $p = 23$ moments are shown along the rows, while the $k = 7$ parameters are shown along the columns. A cell with an "x" indicates a non-zero efficient loading ($\hat{x}_j^*$ in the notation of Section 3.2), while empty cells indicate zero loadings.[32]

[31] We only report the joint test for the just-identified specification. This is because the joint test proposed in Section 5.2 requires a single choice of weight matrix, whereas the worst-case efficient point estimates of the three parameters correspond to three different choices of moments (selected as in Section 3.2).

[32] We define a loading to be zero if $|\hat{x}_j^*| < 10^{-4}$.

Heterogeneous agent application: Efficient moment selection

Impulse response					TFP			Monetary	
Var.	Shock	Horiz.	TR	PC	AR1	AR2	Std	AR1	AR2
TFP	TFP	0			x	x	x		
		1			x	x			
		2							
		8				x			
Output	TFP	0							
		1							
		2							
		8							
Frac	TFP	0							
		1							
		2							
		8							
Output	MP	0	x	x				x	x
		1							
		2							
		8							
Price	MP	0	x	x				x	x
		1							
		2							
		8							
Bond	MP	1	x	x				x	x
		2							
		8	x	x				x	x

Table 5: Cells with an "x" indicate that the efficient estimate of the given parameter (along columns) attaches a non-zero weight to the given empirical moment (along rows). First three columns show the impulse response variable ("Var."), shock, and quarterly horizon ("Horiz."). Variable "Frac": fraction of people earning less than 2/3 of GDP. Shock "MP": monetary shock. See parameter abbreviations in Table 2.

References

Altonji, J. G. & Segal, L. M. (1996). Small-Sample Bias in GMM Estimation of Covariance Structures. *Journal of Business & Economic Statistics*, *14*(3), 353–366.

Alvarez, F., Le Bihan, H., & Lippi, F. (2016). The Real Effects of Monetary Shocks in Sticky Price Models: A Sufficient Statistic Approach. *American Economic Review*, *106*(10), 2817–2851.

Alvarez, F. & Lippi, F. (2014). Price Setting With Menu Cost for Multiproduct Firms. *Econometrica*, *82*(1), 89–135.

Auclert, A., Bardóczy, B., Rognlie, M., & Straub, L. (2021). Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models. *Econometrica*, *89*(5), 2375–2408.

Chang, M., Chen, X., & Schorfheide, F. (2021). Heterogeneity and Aggregate Fluctuations. Manuscript, University of Pennsylvania.

Georgescu, D. I., Higham, N. J., & Peters, G. W. (2018). Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance. *Royal Society Open Science*, *5*(3), 1–11.

Hahn, J., Kuersteiner, G., & Mazzocco, M. (2020a). Estimation with Aggregate Shocks. *Review of Economic Studies*, *87*(3), 1365–1398.

Hahn, J., Kuersteiner, G., & Mazzocco, M. (2020b). Joint Time Series and Cross-Section Limit Theory under Mixingale Assumptions. *Econometric Theory*. forthcoming.

Hansen, L. & Heckman, J. (1996). The Empirical Foundations of Calibration. *Journal of Economic Perspectives*, *10*(1), 87–104.

Imbens, G. W. & Lancaster, T. (1994). Combining Micro and Macro Data in Microeconometric Models. *Review of Economic Studies*, *61*(4), 655–680.

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary Policy According to HANK. *American Economic Review*, *108*(3), 697–743.

Koenker, R. & Bassett, G. (1978). Regression Quantiles. *Econometrica*, *46*(1), 33–50.

Kydland, F. E. & Prescott, E. C. (1996). The Computational Experiment: An Econometric Tool. *Journal of Economic Perspectives*, *10*(1), 69–85.

Liu, L. & Plagborg-Møller, M. (2020). Full-Information Estimation of Heterogeneous Agent Models Using Macro and Micro Data. Manuscript, Princeton University.

Mavroeidis, S., Plagborg-Møller, M., & Stock, J. H. (2014). Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve. *Journal of Economic Literature*, *52*(1), 124–188.

McKay, A., Nakamura, E., & Steinsson, J. (2016). The Power of Forward Guidance Revisited. *American Economic Review*, *106*(10), 3133–58.

Miranda-Agrippino, S. & Ricco, G. (2021). The Transmission of Monetary Policy Shocks. *American Economic Journal: Macroeconomics*, *13*(3), 74–107.

Mongey, S. & Williams, J. (2017). Firm dispersion and business cycles: Estimating aggregate shocks using panel data. Manuscript, New York University.

Nakamura, E. & Steinsson, J. (2018). Identification in Macroeconomics. *Journal of Economic Perspectives*, *32*(3), 59–86.

Newey, W. K. & McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of Econometrics, Volume IV* chapter 36, (pp. 2111–2245). Elsevier.

Székely, G. J. & Bakirov, N. K. (2003). Extremal probabilities for Gaussian quadratic forms. *Probability Theory and Related Fields*, *126*(2), 184–202.

Winberry, T. (2018). A method for solving and estimating heterogeneous agent macro models. *Quantitative Economics*, *9*(3), 1123–1151.