Standard Errors for Calibrated Parameters*

Matthew D. Cocci
Princeton University

Mikkel Plagborg-Møller
Princeton University

June 12, 2019

Abstract: Calibration, a popular way to discipline the parameters of structural models using data, can be viewed as a version of moment matching (minimum distance) estimation. Existing standard error formulas for such estimators require knowledge of the correlation structure of the matched empirical moments, which is often unavailable in practice. Given knowledge of only the variances of the individual empirical moments, we develop 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 an over-identification test and a joint test of parameter restrictions. All procedures are quickly and easily computable in standard software packages.

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. Moment matching is popular in diverse fields of economics, including macroeconomics, international trade, industrial organization, public economics, and structural labor 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.

*Emails: mcocci@princeton.edu, mikkelpm@princeton.edu. We are grateful for comments from Isaiah Andrews, Bo Honoré, Michal Kolesár, Pepe Montiel Olea, Ulrich Müller, and seminar participants at Chicago and Princeton.

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 known. 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 scheme that minimizes the worst-case estimator variance amounts to a moment selection problem whose solution is easily computable. 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, 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 it has exact minimax coverage probability, 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. The solution to this problem can be obtained from the coefficients in a median regression (i.e., Least Absolute Deviation regression). The median regression is applied to an artificial data set that is easily computable. Standard quantile regression software efficiently performs median regression. The optimal 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 apply our results to 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 "unmatched" 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 set-up. 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 easily computable using off-the-shelf software.

Finally, as extensions, we discuss settings with more detailed knowledge of the variance-covariance matrix of empirical moments. This includes settings where the correlation structure of a subset of moments is known, or where certain moments are known to be independent.

Literature. This paper is related to the literature on correlation matrix completion, see Georgescu et al. (2018) and references therein. Several papers have proposed methods for computing positive definite correlation matrices that satisfy various optimality criteria, such as the maximum entropy principle. We differ from these papers by tying the optimization over correlation matrices to the objective of finding the worst-case variance of estimated structural parameters in an economic model. Moreover, our derivations of the worst-case optimal weight matrix and joint testing procedure do not seem to have parallels in the matrix completion literature.

The viewpoint that calibration is a version of minimum distance estimation has been articulated by Hansen & Heckman (1996). The conventional asymptotic analysis of minimum distance estimators is reviewed by Newey & McFadden (1994).

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. In such cases, Hahn et al. (2018a) provide advanced tools for doing inference with a mix of cross-sectional and time series data. Their methods, unlike ours, generally require access to the underlying data. Hahn et al. (2018b) 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 set-up. Section 3 derives the worst-case standard errors and the optimal moment weighting/selection. Section 4 presents two simple geometric and analytical illustrations of our basic results. Section 5 develops tests of parameter restrictions and of over-identifying restrictions. Section 6 discusses extensions. Appendix A contains technical lemmas and other details.

2 Set-up

Consider a standard moment matching (minimum distance) estimation framework. Let $\mu_0 \in \mathbb{R}^p$ be a vector of reduced-form parameters ("moments"), and $\theta_0 \in \Theta \subset \mathbb{R}^k$ 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. We have access to an estimator $\hat{\mu}$ ("empirical moments") such that

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

for a $p \times p$ symmetric positive semidefinite variance-covariance matrix V.[1] Let $\hat{W}$ be a $p \times p$ symmetric matrix satisfying $\hat{W} \xrightarrow{p} W$. Then a "moment matching" (i.e., minimum distance) estimator of θ_0 is given by

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

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

[1] Here and below, all limits are taken as the sample size $n \to \infty$. If the different elements of $\hat{\mu}$ are computed on different data sets or using different econometric methods, the convergence rates may differ across elements. Here we work in an asymptotic framework where all convergence rates are asymptotically proportional. The factors of proportionality are implicitly reflected in V. If the factor of proportionality for some element $\hat{\mu}_j$ is zero (faster rate of convergence than the slowest rate), then $V_{jj} = 0$.

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 and confidence intervals for the estimator $\hat{\theta}$. Under standard regularity conditions (see below)

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

provided that $\hat{\theta} \xrightarrow{p} \theta_0$ (see Newey & McFadden, 1994, for lower-level identification and smoothness conditions that ensure consistency). Here $G \equiv \partial h(\theta_0)/\partial\theta' \in \mathbb{R}^{p\times k}$.

Although it would be easy to construct standard errors if the correlation structure of $\hat{\mu}$ were known, this is not the case in many applications. Often the moments $\hat{\mu}$ are obtained from a variety of 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. In these cases, it is impossible to consistently estimate all elements in the matrix V, rendering the usual minimum distance formula useless.

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 obtained from. Thus, assume that we have access to standard errors $\hat{\sigma}_1, \ldots, \hat{\sigma}_p \geq 0$ satisfying

$$\sqrt{n}\hat{\sigma}_j \xrightarrow{p} V_{jj}^{1/2}, \quad j = 1, \ldots, p. \tag{3}$$

For ease of reference, we summarize our technical assumptions and regularity conditions in the following statement.

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 (3).*

iii) *$h(\cdot)$ is continuously differentiable in a neighborhood of θ_0, and $G \equiv \partial h(\theta_0)/\partial\theta'$ has full column rank.*

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

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

Assumptions (ii)–(v) are standard regularity conditions that are satisfied in smooth, locally identified models. We will now discuss the key assumption (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. In this case, it generally does not make sense to think of the underlying uncertainty as arising from random sampling. Instead, a model-based (e.g., shock-based) uncertainty concept may be more appropriate.

There are several cases in which the joint normality assumption appears reasonable. For example:

1. *The moments are obtained from observational "macro" data sets with similar time span and geographical coverage, but the underlying data for some of the moments is not available.* For example, some moments may derive from time series regressions or cross-country panel regressions applied to proprietary data or to moments reported in previous papers.

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.*

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, see Section 6.

5. *The moments are computed from the same data set 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. Moreover, the usual micro standard errors will not take into account the *combined* uncertainty in the macro shock and idiosyncratic micro noise.

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 moments may fail.

In practice, we recommend that researchers take an explicit stand on the sampling framework, i.e., the different sources of uncertainty. This process must necessarily be application-specific.

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.

Suppose we care about the linear combination $\lambda'\theta_0$ of the structural parameters, for some $\lambda \in \mathbb{R}^k$. By the delta method, the estimator $\lambda'\hat{\theta}$ is asymptotically equivalent with a certain linear function $x'\hat{\mu}$ of the empirical moments, where $x = (x_1, \ldots, x_p)' \equiv WG(G'WG)^{-1}\lambda$. We thus seek to bound the variance of a linear combination of $\hat{\mu}$, knowing the variance of each component $\hat{\mu}_j$ but not the correlation structure. The worst-case variance is attained when all components of $\hat{\mu}$ are perfectly correlated, 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 the appendix.[2]

[2]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.

We can thus construct an estimate of the worst-case standard errors for the linear combination $\lambda'\hat{\theta}$ as

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

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

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

has asymptotic coverage probability of at least $1-\alpha$ for the true linear combination $\lambda'\theta_0$. 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). These results follow from the fact

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

as shown in Lemma 1 in the appendix. 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 Optimal moment selection

We now derive a weight matrix that minimizes the *worst-case* variance of the estimator, derived above. We show that this weight matrix only puts weight on k moments, so the procedure amounts to optimal 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.

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

$$\begin{aligned}
&\min_{W\in\mathcal{S}_p}\max_{\tilde{V}\in\mathcal{S}(\mathrm{diag}(V))}\sqrt{\lambda'(G'WG)^{-1}G'W\tilde{V}WG(G'WG)^{-1}\lambda} \qquad (4)\\
&\quad=\min_{W\in\mathcal{S}_p}\max_{\tilde{V}\in\mathcal{S}(\mathrm{diag}(V))}(x(W)'\tilde{V}x(W))^{1/2}\\
&\quad=\min_{W\in\mathcal{S}_p}\sum_{j=1}^{p}V_{jj}^{1/2}|x_j(W)|,
\end{aligned}$$

where the last equality uses Lemma 1, as in the previous subsection. Lemma 2 in the appendix shows that the solution to the final optimization problem above is given by

$$\min_{W\in\mathcal{S}_p}\sum_{j=1}^{p}V_{jj}^{1/2}|x_j(W)| = \min_{z\in\mathbb{R}^{p-k}}\sum_{j=1}^{p}|\tilde{Y}_j-\tilde{X}_j'z|, \qquad (5)$$

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 the $p\times(p-k)$ matrix of eigenvectors of $I_p-G(G'G)^{-1}G'$ corresponding to its nonzero eigenvalues, and the notation $A_{j\bullet}$ means the j-th row of matrix A.

The final optimization problem (5) 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 only 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 (5) such that at least $p-k$ out of the p median regression residuals

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

equal zero. Hence, any optimal weight matrix W^* that achieves the minimum in (5) will yield a linear combination vector $x(W^*)=(V_{11}^{-1/2}e_1^*,\ldots,V_{pp}^{-1/2}e_p^*)'$ that attaches weight to at most k out of the p empirical moments $\hat{\mu}$. In other words, the solution to the optimal moment *weighting* problem is achieved by an optimal moment *selection*.

ALGORITHM. In practice, 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 $W_{jj} = \hat{\sigma}_j^{-2}$.

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

iii) Solve the median regression (5), substituting $\hat{G}$ for G and $\hat{\sigma}_j$ for $V_{jj}^{1/2}$. Compute the residuals $\hat{e}_j^*$, $j = 1, \ldots, p$, from this median regression. (In the non-generic case where multiple solutions to the median regression exist, select one that yields at least $p - k$ residuals that are zero.)

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

v) To compute an efficient estimator of $\lambda'\theta_0$, either:

   a) Compute the just-identified efficient minimum distance estimator $\hat{\theta}_{\text{eff}}$ 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 $\lambda'\theta_0$ by $\lambda'\hat{\theta}_{\text{eff}}$. Or:

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

vi) The worst-case standard error of $\lambda'\hat{\theta}_{\text{eff}}$ and of $\hat{\theta}^{\lambda}_{\text{eff-1S}}$ is given by the value of the median regression (5) (i.e., the minimized objective function).

Options (a) and (b) in the above procedure 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. One can optionally re-run the median regression with an updated $\hat{G}$ based on $\hat{\theta}_{\text{eff}}$, but this does not increase asymptotic efficiency.

REMARKS.

1. Since all operations involved in computing the optimal linear combination $\hat{x}^*$ are continuous, $\hat{x}^*$ converges in probability to the population optimal linear combination $x(W^*)$. The only exception may be where the population median regression (5) does not have a unique minimum, which is a non-generic case. Even in this case, however, the optimal worst-case standard errors will be consistent (when multiplied by $\sqrt{n}$) under Assumption 1(i)–(iii), by a standard application of the maximum theorem.

2. The full-information (infeasible) optimal weight matrix that exploits knowledge of all of V is known to equal $W = V^{-1}$. This weight matrix in general attaches weight to all moments, unlike the limited-information optimal solution derived above. The worst-case asymptotic standard deviation (5) given limited information is of course larger than the standard deviation $(\lambda'(G'V^{-1}G)^{-1}\lambda)^{1/2}$ of the full-information efficient estimator of $\lambda'\theta_0$. The limited-information efficient estimators $\hat{\theta}_{\text{eff}}$ and $\hat{\theta}^{\lambda}_{\text{eff-1S}}$ depend on the linear combination vector λ, unlike the full-information efficient estimator.

3. When running the median regression (5), the quantile regression software must be instructed to omit the intercept.

4. 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 (i.e., estimators that remain asymptotically unbiased under locally drifting parameter sequences in θ_0), we need $HG = I_k$. Among all estimators $\hat{\vartheta}$ satisfying these requirements, the smallest possible worst-case asymptotic standard deviation of $\lambda'\hat{\vartheta}$ 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 the appendix shows that the solution to this problem is precisely the value of the median regression (5). In other words, the minimum distance estimator $\hat{\theta}$ with (limited-information) optimal weight matrix delivers the smallest possible worst-case standard errors in a large class of estimators.

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 optimal 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 $k = 1$, $p = 2$,

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

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:

$$\hat{\theta}(x_1, x_2) \equiv x_1\hat{\mu}_1 + x_2\hat{\mu}_2, \qquad x_1 + x_2 = 1. \tag{6}$$

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 (6). Note that this constraint ensures that $\hat{\theta}(x_1, x_2)$ is unbiased.

In this example it is obvious that the worst-case optimal 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 (6), which we represent 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,

$$\text{Var}[\hat{\theta}(x_1, x_2)] = \sigma_1^2 x_1^2 + \sigma_2^2 x_2^2 + 2\rho\sigma_1\sigma_2 x_1 x_2. \tag{7}$$

which depends upon the unknown correlation parameter ρ. Dashed ellipses in the figure represent level sets of the estimator variance (7).

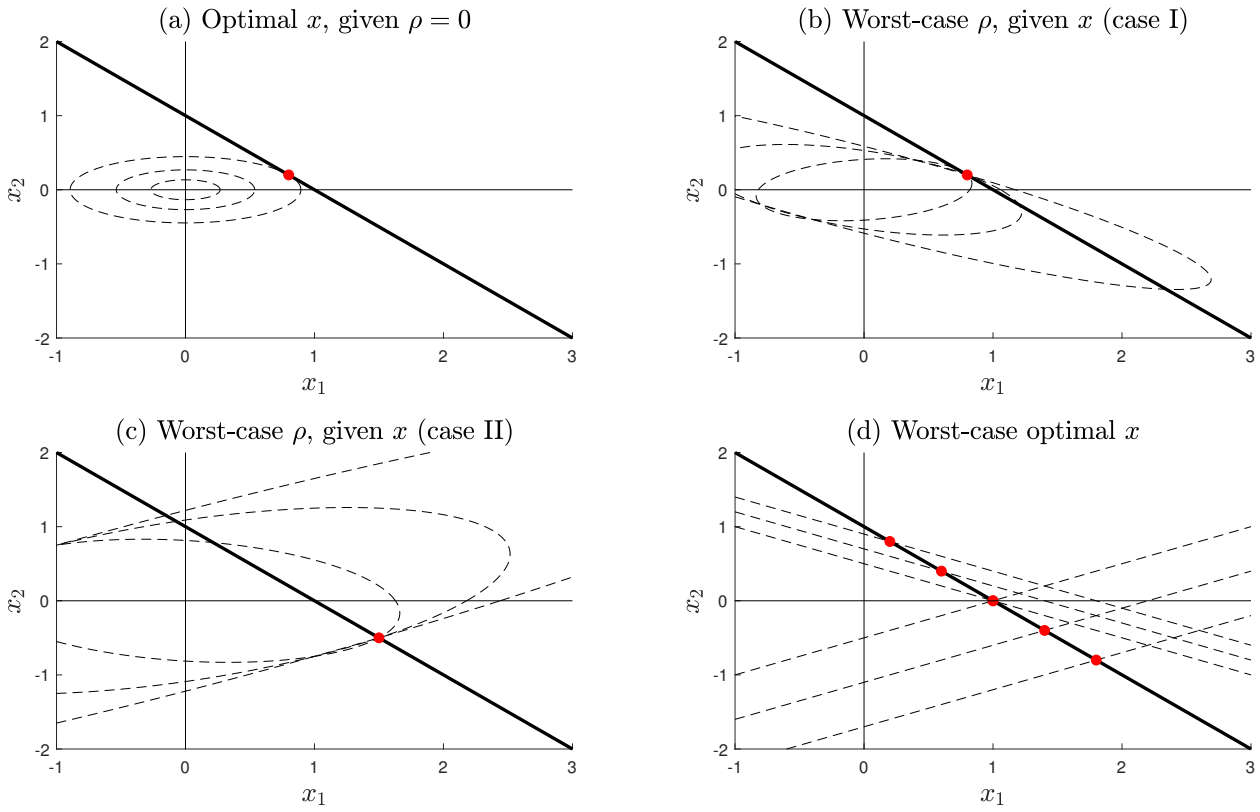

Figure 1: Geometric illustration of worst-case optimal estimator. (a): Optimal estimator given known correlation structure. (b)+(c): Worst-case correlation structure for two different estimators. (d) Worst-case optimal estimator. See main text for explanation.

Panel (a) of Figure 1 depicts the optimal estimator in the case where it is known that $\rho \equiv \text{Corr}(\hat{\mu}_1, \hat{\mu}_2) = 0$. 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 (6). Note that it is optimal to use both moments for estimation, i.e., $x_1, x_2 > 0$ (the standard diversification motive in finance, cf. the discussion 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. (7)). We see from the figure that the standard deviation of the estimator is increasing in ρ; hence, the worst case is $\rho = 1$.[3]

[3]This panel thus illustrates graphically the inner maximization in (4) for fixed W.

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.[4] 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 optimal estimator in this case is $\hat{\theta}(1, 0)$. In other words, it is optimal to throw away the second (higher-variance) moment.

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 $0_{p\times 1}$ 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).

4.2 Analytical calculations: three moments, two parameters

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

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

[4] This task corresponds to out the outer maximization in (4).

Assume $a, b, c, d \neq 0$. In the appendix we provide detailed derivations that map this toy example into the general framework and notation of Section 3.

To begin, recall from Section 3.2 that the estimator $\lambda'\hat{\theta} = \hat{\theta}_1$ with smallest variance under the worst-case correlation structure necessarily corresponds to a *just-identified* estimator $\hat{\theta}$ of the full parameter vector, using only $k = 2$ of the available $p = 3$ moments.[5] The worst-case correlation structure corresponds to *perfect* correlation among all moments, with the sign of the correlation coefficients chosen to deliver the highest variance. Therefore, we add only variance to our estimator without adding independent explanatory power whenever we use additional moments beyond the minimal number k necessary for identification and estimation of the full parameter vector θ_0.

As a result, an optimal estimator of $\theta_{0,1}$ under worst-case correlation structure uses at most $k = 2$ moments, and we now consider *which* moments should be used for estimation. 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 to also estimate $\theta_{0,2}$). In this case,

$$\hat{\theta}_1 = \frac{1}{a}\hat{\mu}_1. \tag{8}$$

Alternatively, another option is to use the second and third moments together to estimate the first parameter, which yields

$$\left.\begin{aligned} \hat{\mu}_2 &= b\hat{\theta}_1 + c\hat{\theta}_2 \\ \hat{\mu}_3 &= d\hat{\theta}_2 \end{aligned}\right\} \quad \Longrightarrow \quad \hat{\theta}_1 = \frac{1}{b}\hat{\mu}_2 - \frac{c}{bd}\hat{\mu}_3. \tag{9}$$

The choice between estimation of $\theta_{0,1}$ via expression (8) or (9) amounts to a comparison of worst-case estimator variance. Specifically, (8) is preferable under the worst-case correlation structure when

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

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.

[5] See also Section 1 for diversification intuition and Section 4.1 for a geometric illustration of this point.

We can obtain a useful interpretation of the optimal estimator by rewriting the above inequality as

$$\frac{\sigma_1}{|a|} \leq \frac{\sigma_2}{|b|} + \frac{\sigma_2/|b|}{\sigma_2/|c|} \times \frac{\sigma_3}{|d|}. \tag{10}$$

This inequality illustrates that the general approach of Section 3.2 compares "signal-to-noise" ratios when selecting the optimal identification scheme. According to (10), we use the first moment $\hat{\mu}_1$ alone for identification and estimation of $\theta_{0,1}$ when $|a|/\sigma_1$ 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 $\theta_{0,1}$.

As illustrated in this simple example, the median regression approach in Section 3.2 ultimately picks the optimal estimator by comparing the variances of all feasible just-identified estimators. This amounts to comparing signal-to-noise ratios to ensure that we use for estimation only those moments that are precisely measured or particularly sensitive to the parameters that must be identified to estimate the parameter of interest.

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, where $r\colon \Theta \to \mathbb{R}^m$ is a continuously differentiable restriction function. Tests of a single parameter restriction ($m = 1$) can be carried out using the confidence interval described in Section 3.1 because $r(\hat{\theta})$ is asymptotically equivalent to a scalar-valued linear combination of the empirical moments. For the case $m > 1$, we propose the following testing procedure. Let α denote the significance level.

i) Compute the Wald-type test statistic

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

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

ii) Compute the critical value

$$\widehat{\text{cv}} \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \text{trace}\left(\tilde{V} W G (G'WG)^{-1} R' S R (G'WG)^{-1} G' W\right) \times \left(\Phi^{-1}(1-\alpha/2)\right)^2, \tag{11}$$

replacing all quantities with their sample analogues $\text{diag}(V) \approx n\,\text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$, $W \approx \hat{W}$, $S \approx \hat{S}$, $G \approx \hat{G} \equiv \frac{\partial h(\hat{\theta})}{\partial \theta'}$, and $R \approx \hat{R} \equiv \frac{\partial r(\hat{\theta})}{\partial \theta'}$.

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

We argue below that, as long as the significance level $\alpha \leq 0.215$ (as is usually the case), 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 optimal weight matrix derived in Section 3.2. The maximization problem (11) is a so-called semidefinite programming problem, a special case of convex programming. Fast and numerically stable algorithms are available in Matlab (e.g., the package CVX[6]) and other computing languages.

We now show that the proposed test controls size.

Proposition 1. *Assume Assumption 1, $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{S} \xrightarrow{p} S$, S is symmetric positive definite, $r(\cdot)$ is continuously differentiable, $R \equiv \frac{\partial r(\theta_0)}{\partial \theta'}$ has full column rank, and $\alpha \leq 0.215$. Then, if $r(\theta_0) = 0_{m\times 1}$,*

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

Proof. Under the null hypothesis,

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

where $V^{1/2} V^{1/2\prime} = V$, and $Z \sim N(0_{p\times 1}, I_p)$. The asymptotic null distribution of the test statistic $\hat{T}$ is therefore a Gaussian quadratic form,

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

[6] http://web.cvxr.com/cvx/doc/sdp.html

Székely & Bakirov (2003) prove that

$$P(Z'QZ \leq \text{trace}(Q) \times \tau) \leq P(Z_1^2 \leq \tau) \tag{12}$$

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,

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

□

Remarks.

1. We do not have formal results on how to choose the weight matrix $\hat{S}$ in the test statistic. A natural *ad hoc* choice is to set $\hat{S} = (R(G'WG)^{-1}G'W\bar{V}WG(G'WG)^{-1}R)^{-1}$ (or the sample analogue), where $\bar{V} \equiv \text{diag}(V_{11}, \ldots, V_{pp})$. Then the test statistic $\hat{T}$ coincides with the usual Wald test statistic for the case where the moments are asymptotically independent.

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. The reason is that the upper bound in (12) is attained when Q has rank 1, but the positive semidefinite maximum (11) need not imply a rank-1 matrix Q, 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.

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 $\hat{W} \xrightarrow{p} W$ be given, such as the limited-information optimal 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. Then all asymptotic properties of $\hat{\theta}$ mentioned in Section 2 hold, 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 $j^* \in \{1, \ldots, p\}$ 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

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

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 can be done after estimating $\hat{\theta}$ with a weight matrix $\hat{W}$ that has zeroed out the corresponding rows and columns of the non-targeted moments so that the estimator attaches zero weight to these moments. Note that p still denotes 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 $\hat{T}_{\text{overid}} \equiv n(\hat{\mu} - h(\hat{\theta}))'\hat{S}(\hat{\mu} - h(\hat{\theta}))$ 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 n times 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

$$\widehat{\text{cv}}_{\text{overid}} \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \text{trace}\left(\tilde{V}(I_p - WG(G'WG)^{-1}G')S(I_p - G(G'WG)^{-1}G'W)\right) \times \left(\Phi^{-1}(1-\alpha/2)\right)^2,$$

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

6 Extensions

We here mention extensions of our results to settings where (i) we also know some *off-diagonal* elements of the variance-covariance matrix of the empirical moments, or (ii) we use Generalized Minimum Distance for estimation.

6.1 Knowledge of the block diagonal of the variance matrix

Suppose we know the *block* diagonal of V, i.e., V is known to be of the form

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

where $V_{(j)}$ are known (or consistently estimable) square symmetric matrices (possibly of different dimensions) for $j = 1, \dots, J$. Such a structure may obtain 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 worst-case asymptotic standard deviation of $\lambda'\hat{\theta}$ among variance-covariance matrices V with the above structure is then given by

$$\sum_{j=1}^{J} (x'_{(j)} V_{(j)} x_{(j)})^{1/2}.$$

This follows from the same logic as in Lemma 1 in the appendix 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. Here we have also partitioned $\hat{\mu} = (\hat{\mu}'_{(1)}, \dots, \hat{\mu}'_{(J)})'$ conformably.

6.2 Additional information about the correlation structure

Consider now the general case where any given collection of elements of V is known. For example, certain off-diagonal elements are zero if the corresponding elements of $\hat{\mu}$ are known to be independent. We can generally compute the worst-case and best-case asymptotic

variance of $\lambda'\hat{\theta}$ by maximizing or minimizing $x'Vx$ subject to the given constraints on V and symmetric positive semidefiniteness.[7] This is another example of a semidefinite programming problem, for which good numerical algorithms exist, as discussed in Section 5.1.

6.3 Generalized Minimum Distance estimation

The methods in this paper extend easily to Generalized Minimum Distance estimation. In that setting θ_0 and μ_0 are known to be linked through the equation $g(\theta_0, \mu_0) = 0_{m\times 1}$, and we have available an asymptotically normal estimator $\hat{\mu}$ of μ_0. The setting in Section 2 is a special case with $g(\theta, \mu) = \mu - h(\theta)$, but our calculations carry over because the asymptotic expansions are essentially the same (Newey & McFadden, 1994).

[7] If $p = 2$ and V_{12} is unrestricted, the best-case asymptotic variance is given by $|x_1^2 V_{11} - x_2^2 V_{22}|$, which is achieved when $\hat{\mu}_1$ and $\hat{\mu}_2$ are perfectly negatively correlated. Closed-form expressions for $p > 2$ seem harder to come by.

A Appendix

A.1 Technical lemmas

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

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

Additionally, for any $z \in \mathbb{R}^{p-k}$,

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

where

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

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

Proof. We first show the second statement of the lemma. Note that

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

Hence,

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

Thus,

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

as claimed. Note that the just-proved fact also implies

$$\left\{G(G'G)^{-1}\lambda + G^\perp z\colon z \in \mathbb{R}^{p-k}\right\} \subset \left\{WG(G'WG)^{-1}\lambda\colon W \in \mathcal{S}_p\right\}. \tag{14}$$

We now prove the first statement of the lemma. Pick any $W \in \mathcal{S}_p$. Then $x = WG(G'WG)^{-1}\lambda$ satisfies $G'x = \lambda$. This shows that

$$\left\{WG(G'WG)^{-1}\lambda\colon W \in \mathcal{S}_p\right\} \subset \left\{x\colon x \in \mathbb{R}^p,\ G'x = \lambda\right\}. \tag{15}$$

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

$$\left\{x\colon x \in \mathbb{R}^p,\ G'x = \lambda\right\} \subset \left\{G(G'G)^{-1}\lambda + G^\perp z\colon z \in \mathbb{R}^{p-k}\right\}. \tag{16}$$

The set inclusions (14)–(16) together imply the first statement of the lemma. □

A.2 Details of 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

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

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, $\text{diag}(V)^{1/2}$ 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

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

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

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

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

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

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

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

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

$$z^* = \begin{cases} \frac{bd}{a} & \text{if } \sigma_1|bd| \leq \sigma_2|ad| + \sigma_3|ac|, \\ -\frac{a(c^2+d^2)}{bd} & \text{if } \sigma_1|bd| > \sigma_2|ad| + \sigma_3|ac|. \end{cases}$$

This implies that the optimal linear combination is

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

and the worst-case optimal estimator of $\lambda'\theta_0$ is

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

This confirms the heuristic derivations in Section 4.2.

References

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. (2018a). Central Limit Theory for Combined Cross-Section and Time Series. arXiv:1610.01697.

Hahn, J., Kuersteiner, G., & Mazzocco, M. (2018b). Estimation with Aggregate Shocks. *Review of Economic Studies*. Forthcoming.

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

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.

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.