Big Data, Model Selection, Aggregation-Indexing

Esfandiar Maasoumi[1]

[1]Emory University

September 5, 2016

What Does Big Data Mean?

What Does Big Data Mean?

1. Big Data and Big Models
   - Big Data-Micro; Big Data Macro/Time series!
   - Mostly variable selection: Multiple Indicators, **Latent Objects**.

What Does Big Data Mean?

1. Big Data and Big Models
   - Big Data-Micro; Big Data Macro/Time series!
   - Mostly variable selection: Multiple Indicators, **Latent Objects**.
2. Latent Objects:
   - Well-being; Happiness, Permanent Income, Expectations.
   - What about Data Generation Process (DGP) as a "Latent Object"?

What Does Big Data Mean?

1. Big Data and Big Models
   - Big Data-Micro; Big Data Macro/Time series!
   - Mostly variable selection: Multiple Indicators, **Latent Objects**.
2. Latent Objects:
   - Well-being; Happiness, Permanent Income, Expectations.
   - What about Data Generation Process (DGP) as a "Latent Object"?
3. Common Themes
   - More Variables than Observations $p >> n$
   - RELATED: Shrinkage, Penalization, Averaging, Model Selection, model uncertainty, Misspecification

The ProtoType Case: Variable Selection

The ProtoType Case: Variable Selection

1. Outcome Y, Target T, Variable Set X of p vars
   - WRITE THE model. formulae here with definitions.
   - This is typically a linear "MODEL"!
   - Estimation and Big Data strategies: Shrinkage, LASSO, Other Penalization methods.

Inference

Inference

1. Best Practice Inference Theory
   - Exemplified by Victor Chernozhukov and coworkers.
   - Asymptotic Inference, especially for $p > n$.
   - I will advertize my old work in this area Below!!:)
   - Ridge Regression; Stein; shrinkage estimation and forecasting in systems of equations; LASSO

Inference

1. Best Practice Inference Theory
   - Exemplified by Victor Chernozhukov and coworkers.
   - Asymptotic Inference, especially for $p > n$.
   - I will advertize my old work in this area Below!!:)
   - Ridge Regression; Stein; shrinkage estimation and forecasting in systems of equations; LASSO
2. "Bayesian" Interpretation; Extraneous Statistical "Information" (important);
   - Model selection and uncertainty.
   - What do we want to learn or do? (KEY question)
   - "Causal"? Policy analysis/decision making needs this.
   - Trearment effect and Program/policy evaluation

Mechanism vs. Blackbox Prediction vs. Indexing

Mechanism vs. Blackbox Prediction vs. Indexing

1. Aggregation-Indexing Multiple Indicators
   - Aggregate-"average" all Xs. Not "causal"
   - What is "average"? An "INDEX"
   - Classic Index Number Problem
   - Aggregation-Averaging of Models is related, but not the same

Mechanism vs. Blackbox Prediction vs. Indexing

1. Aggregation-Indexing Multiple Indicators
   - Aggregate-"average" all Xs. Not "causal"
   - What is "average"? An "INDEX"
   - Classic Index Number Problem
   - Aggregation-Averaging of Models is related, but not the same
2. Models for Mechanism Learning
   - Many Moments Paradigm (GMM)
   - Empirical Likelihood-Information Theory
   - Hopeless!? All models are Misspecified
   - The Map allegory!

WHAT I HOPE TO PRESENT HERE

WHAT I HOPE TO PRESENT HERE

1. Variable selection (Brief)

WHAT I HOPE TO PRESENT HERE

1. Variable selection (Brief)
2. Examples of Penalization, Shrinkage, Ridge, LASSO, Moment selection, pre-testing, model uncertainty,...
   - Work that I have done since 1974
   - Extraneous Statistical vs. Bayesian Interpretation

WHAT I HOPE TO PRESENT HERE

1. Variable selection (Brief)
2. Examples of Penalization, Shrinkage, Ridge, LASSO, Moment selection, pre-testing, model uncertainty,...
   - Work that I have done since 1974
   - Extraneous Statistical vs. Bayesian Interpretation
3. Aggrgegation-Indexing
   - Multiple Indicators of Well-Being
   - Maasoumi (1986. Econometrica)
   - Finally: Model Averaging as Indexing, when all are misspecified
   - Gospodinov-Maasoumi (2016)

axioms discussed in the literature include additivity, differentiability, and concavity; see, e.g., Debreu (1964, 1972) and Koopmans (1972). These axioms restrict the class of admissible preference relations and thus imply important behavioral restrictions.

As is common to all fields of science, any notion of "closeness" is associated with a criterion of distance or divergence. In the case of the question above we are therefore concerned with measures of divergence between distributions. In mathematical statistics many such measures are proposed and utilized in a variety of seemingly disparate problems; see, e.g., Fisher (1925), Shannon (1948), Rao (1949), Jeffreys (1961), and Burbea and Rao (1982a, 1982b). Such measures have begun to be employed in the fields of income inequality (Cowell, 1980, 1982), multidimensional inequality (Maasoumi, 1979, 1986; Maasoumi and Nickelsburg, 1983), and international trade (Theil, 1979).

The main result of this paper, given in the next section, is that some of the most popular functional forms in economics (CES, Cobb-Douglas, Leontief, etc.) do indeed define welfare distributions which are the "closest" to the commodity distributions. We identify the criteria according to which this is so. It will be seen that such criteria are rather arbitrary and incorporate value judgments on such things as the relative importance of commodities, their substitutability, and the degree of aversion to (valuation of) distributional distortion associated with any choice of utility function or any aggregate of miscellaneous inputs (Bliss, 1975). In the next section we define a multivariate generalization of a generalized information (entropy) measure which serves as our divergence criterion. We then find the class of functional forms that includes many of the popular functions employed in economic analysis, and which minimizes the distributional divergence mentioned earlier. The last two sections offer some remarks on econometric implications and the utility of these measures in evaluating approximate regression functions.

II. OPTIMAL FUNCTIONAL FORMS

Given an allocation of m goods among N individuals, X_{if}; $i = 1, \ldots, N$, and $f = 1, \ldots, m$, denote the ith individual's index of utility by S_i, $i = 1, \ldots, N$. The vector $X_i = (X_{i1}, X_{i2}, \ldots, X_{im})'$ represents the allocation of m commodities to the ith individual. $X^f = (X_{1f}, X_{2f}, \ldots, X_{Nf})'$ is the allocation of commodity f. Without loss of generality and in order to speak of our *convex* criteria as measures of divergence between distributions, we shall work with normalized variables $x_{if} \geq 0$ such that $x_{if} = X_{if}/\sum_i X_{if}$, $\sum_i x_{if} = 1$, and redefine S_i such that $\sum_i S_i = 1$.

A popular, *convex* measure of divergence between any two distributions, $\{Z_i\}$ and $\{y_i\}$, $Z_i, y_i \geq 0$, is the generalized entropy (β-order entropy) given by

$$\Delta_\beta = \frac{1}{\beta(\beta+1)} \sum_i y_i \left[\left(\frac{y_i}{Z_i} \right)^\beta - 1 \right]. \tag{1}$$

Some special forms of (1) are

$$\Delta_0 = \sum_i y_i \log \left(\frac{y_i}{Z_i} \right) \tag{2}$$

and

$$\Delta_{-1} = \sum_i Z_i \log \frac{Z_i}{y_i}. \tag{3}$$

Δ_0 and Δ_{-1} are well-known measures of "expected information" in going from distribution $\{Z_i\}$ to $\{y_i\}$. They are related to Theil's two indices of inequality, which have a central place in the analysis of single-dimensioned (e.g., income) inequality (Theil, 1967). Cowell (1980) provides a useful analysis of the properties of (1) for analyzing distributional change.

As in Maasoumi (1979, 1986), we consider a multivariate generalization of (1) to measure divergence between S_i and all the other m distributions $x^f, f = 1, \ldots, m$.[2]

$$\Delta_{\beta,m} = \frac{1}{m} \sum_f \sum_i \frac{1}{\beta(\beta+1)} S_i \left[\left(\frac{S_i}{x_{if}} \right)^\beta - 1 \right]. \tag{4}$$

Minimizing $\Delta_{\beta,m}$ with respect to S_i, we find the distribution of S_i, $i = 1, \ldots, N$, which is the "closest" to those of x^f in the sense of (4):

$$S_i = \frac{(\sum_f x_{if}^{-\beta})^{-1/\beta}}{\sum_i (\sum_f x_{if}^{-\beta})^{-1/\beta}}. \tag{5}$$

If $\sum_i S_i = 1$ is not imposed, (5) will be

$$S_i \propto \left[\frac{\beta+1}{m} \sum_f x_{if}^{-\beta} \right]^{-1/\beta} \tag{5'}$$

with $\propto$ denoting proportionality.

The solution (5') may be recognized as a constant elasticity of substitution (CES) function which contains the linear case ($\beta = -1$) and, upon the application of L'Hospital's rule, the Cobb–Douglas function ($\beta = 0$) which is a geometric mean here, and the Leontief case of fixed coefficients ($\beta \to +\infty$).

Further generalizations are straightforward. For instance, we may consider different weights for different goods distributions. Then the solution is the more general CES,

$$S_i \propto \left[\frac{(\beta+1)}{m} \sum_f \alpha_f x_{if}^{-\beta} \right]^{-1/\beta}. \tag{6}$$

Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.

Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{d_i}_{\text{treatment}} \overbrace{\alpha}^{\text{effect}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_j}_{\text{controls}} + \underbrace{\epsilon_i}_{\text{noise}}, \qquad (1)$$

$$\mathbb{E}[\epsilon_i | \underbrace{x_i, z_i}_{\text{exogenous vars}}] = 0.$$

Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{d_i}_{\text{treatment}} \overbrace{\alpha}^{\text{effect}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_j}_{\text{controls}} + \underbrace{\epsilon_i}_{\text{noise}}, \qquad (1)$$

$$\mathbb{E}[\epsilon_i | \underbrace{x_i, z_i}_{\text{exogenous vars}}] = 0.$$

- Controls can be richer as more features become available (Census characteristics, housing characteristics, geography, text data)

  $\Leftarrow$ "big" data

Introduction

- Richer data and methodological developments lead us to consider more elaborate econometric models than before.
- Focus discussion on the linear endogenous model

$$\underbrace{y_i}_{\text{outcome}} = \underbrace{d_i}_{\text{treatment}} \overbrace{\alpha}^{\text{effect}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_j}_{\text{controls}} + \underbrace{\epsilon_i}_{\text{noise}}, \qquad (1)$$

$$\mathbb{E}[\epsilon_i | \underbrace{x_i, z_i}_{\text{exogenous vars}}] = 0.$$

- Controls can be richer as more features become available (Census characteristics, housing characteristics, geography, text data)

  $\Leftarrow$ "big" data
- Controls can contain transformation of "raw" controls in an effort to make models more flexible

  $\Leftarrow$ nonparametric series modeling, "machine learning"

Introduction

- This **forces** us to explicitly consider **model selection** to select controls that are “most relevant”.
- Model selection techniques:
  - CLASSICAL: **t and F tests**
  - MODERN: **Lasso**, Regression Trees, Random Forests, Boosting

Introduction

- This **forces** us to explicitly consider **model selection** to select controls that are "most relevant".
- Model selection techniques:
  - CLASSICAL: **t and F tests**
  - MODERN: **Lasso**, Regression Trees, Random Forests, Boosting

If you are using *any* of these MS techniques directly in (1), you are doing it *wrong.*

Have to do *additional selection* to make it right.

Solution: Post-double selection

- **Post-double selection** procedure (BCH, 2010, ES World Congress, ReStud, 2013):

Step 1. Include x_i if it is a significant predictor of y_i as judged by a conservative test (t-test, Lasso etc).

Step 2. Include x_i if it is a significant predictor of d_i as judged by a conservative test (t-test, Lasso etc). [In the IV models must include x_i if it a significant predictor of z_i].

Step 3. Refit the model after selection, use standard confidence intervals.

Theorem (Belloni, Chernozhukov, Hansen: WC ES 2010, ReStud 2013)

DS works in low-dimensional setting and in high-dimensional approximately sparse settings.

Stein, Baranchik, Hoerl, Maasoumi, Andrews, Hanson, Chernozhukov, Zellner, Durbin-Theil-Goldberger,...

1.
   - Generic Reduced Forms,
   - MSRF. (Mixed Estimators)
   - 3SLS Ridge-Like (1981 JoE); $d = a/n$
   - $BY' + CZ' = AX' = U'$
   - GRF: $Y' = PZ' + V'$, $BP + C = D$, D not necessarily $= 0$

2. THE MSRF ESTIMATOR

Consider the structural model:

$$(1) \qquad BY' + CZ' = AX' = U'$$

where $A = [B : C]$ is the $n \times (n+m)$ matrix of unknown coefficients, $X = [Y : Z]$ is the matrix of T observations on the n endogenous and the m non-stochastic exogenous variables such that $\lim_{T\to\infty}(Z'Z/T) = M$ is finite and non-negative definite. U represents the T values of n serially independent disturbance terms such that $U_{.t} \sim IIN(0, \Omega_u)$ and Ω_u is non-singular. The corresponding reduced form model is:

$$(2) \qquad Y_t = PZ_t + V_t \qquad (t = 1, \ldots, T)$$

where $P = -B^{-1}C$ and $E(V_t V_t') = B^{-1}\Omega_u B'^{-1} = \Omega_v$. Denote the 3SLS estimate of A by $A^\dagger$ (with parameter constraints imposed). The following Wald type asymptotic test is employed to test the validity of the parameter constraints (and specification) on A:[4]

$$(3) \qquad \text{tr}\,(\Omega_2^{-1}A^\dagger(X'Z)(Z'Z)^{-1}(Z'X)A'^\dagger) \underset{a}{\sim} \chi^2_N.$$

Where Ω_2 is a consistent estimate of Ω_u (usually the 2SLS) and N is the total number of over-identifying degrees in (1). However, the following, asymptotically equivalent, test is developed in terms of the 3SLS reduced form estimates $P^\dagger = -B^{\dagger-1}C^\dagger$:[5]

$$(4) \qquad \phi^\dagger = \text{tr}\,[W^{-1}(\hat{P} - P^\dagger)(Z'Z)(\hat{P} - P^\dagger)'] \underset{a}{\sim} \chi^2_N$$

where $W = Y'[I - Z(Z'Z)^{-1}Z]Y/T$ is a consistent estimate of Ω_v and $\hat{P} = (Y'Z)(Z'Z)^{-1}$ is the LS estimate of P.[6]

For large sample sizes, the specifying restrictions are rejected if $\phi^\dagger$ or (3) exceed an appropriate critical value of the test. However, for small samples, asymptotic tests such as (3) and (4) lead to unduly high rates of rejection even for reasonably specified models.[7] Moreover, in this uncertain situation the unrestricted LS ($\hat{P}$) estimator may perform quite well. Given the above small sample problem we may wish to combine the unrestricted with the restricted estimator and allow the test result to determine the weights attached. Consequently the following estimator is proposed:

$$(5) \qquad P^* = \lambda P^\dagger + (1-\lambda)\hat{P}$$
$$= P^\dagger + (1-\lambda)(\hat{P} - P^\dagger).$$

[4] E.g., see E. Malinvaud [**10**, Ch. 9, pp. 358–360].

[5] Ibid., f.n. (3); also refer to Section 4 of this paper.

[6] Also note that: $A^\dagger(X'Z)(Z'Z)^{-1} = B^\dagger(Y'Z - P^\dagger Z'Z)(Z'Z)^{-1} = B^\dagger(\hat{P} - P^\dagger)$.

[7] R. L. Basmann [**3**] reports on this phenomena using a similar 2SLS test. Also the author has observed close to 35 per cent rate of rejection when small samples were applied to (4) in Monte Carlo experiments.

Note that:

$$(6) \qquad \phi^* = \operatorname{tr}\left[W^{-1}(\hat{P}-P^*)(Z'Z)(\hat{P}-P^*)'\right] = \lambda^2\phi^\dagger.$$

Then if C_p is the chosen critical value of the test, we choose λ such that

$$\lambda = \begin{cases} 1 \quad \text{if} & \phi^\dagger \leq C_p \quad \text{(hypothesis accepted)}, \\ \left(\dfrac{\phi_2}{\phi^\dagger}\right)^{\frac{1}{2}} & \text{or} \quad \left(\dfrac{\phi_2}{\phi^\dagger}\right) \quad \text{if} \quad \phi^\dagger > C_p, \end{cases}$$

where $\phi_2 \leq C_p$ and may be chosen so as to minimize a desired quadratic loss measure. The similarity of P^* to the Stein-like estimators is seen from the following:

$$(7) \qquad P^* = P^\dagger + I_{(C_p,\infty)}\left(1-\frac{\phi_2}{\phi^\dagger}\right)(\hat{P}-P^\dagger)$$

in which $I_{(\ )} = 1$ if $\phi^\dagger > C_p$ and zero otherwise.

3. THE MOMENTS

The following theorem demonstrates the existence of the first $(T-n-m)$ moments of the MSRF estimator. The methodology is similar to Sargan [**14**] and could be applied to obtain comparable results for other Stein-like estimators.[8]

THEOREM 1. *The integral moments, up to order $r \leq T-n-m$, of the MSRF reduced form estimator are uniformly bounded as $T \to \infty$.*

PROOF: From the definition of P^* it follows that:

$$(8) \qquad \phi^* = \operatorname{tr}\left[TW^{-1}(\hat{P}-P^*)\left(\frac{Z'Z}{T}\right)(\hat{P}-P^*)'\right] \leq C_p.$$

The left-hand side of (8) is expanded using the identity $P^*-\hat{P} = (P^*-P)-(\hat{P}-P)$. We also note that by Cauchy's inequality:

$$\operatorname{tr}\left[TW^{-1}(P^*-P)M(\hat{P}-P)'\right] \leq \left\{\operatorname{tr}\left[TW^{-1}(P^*-P)M(P^*-P)'\right]\right\}^{\frac{1}{2}} \cdot \left\{\operatorname{tr}\left[TW^{-1}(\hat{P}-P)M(\hat{P}-P)'\right]\right\}^{\frac{1}{2}}.$$

Then:

$$(9) \qquad \left\{\operatorname{tr}\left[TW^{-1}(P^*-P)M(P^*-P)'\right]\right\}^{\frac{1}{2}} \leq \left\{\operatorname{tr}\left[TW^{-1}(\hat{P}-P)M(\hat{P}-P)'\right]\right\}^{\frac{1}{2}} + C_p^{\frac{1}{2}}.$$

Let $\gamma = \sqrt{T}f'\operatorname{vec}(P^*-P)$ be an arbitrary linear function of the elements of (P^*-P). Then we note that $[(\Omega_v^{-1}\otimes M)-ff'/\lambda_M]$ is non-negative definite where $\lambda_M = f'(\Omega_v \otimes M^{-1})f$ is the only non-zero (largest) root of $[ff'-\lambda(\Omega_v^{-1}\otimes M)]Z = 0$. Let $\phi = \operatorname{tr}\left[T\Omega_v^{-1}(P^*-P)M(P^*-P)'\right]$; it follows that:

$$(10) \qquad \gamma^2 \leq \lambda_M\phi \qquad \text{and} \qquad E(|\gamma|^r) \leq \lambda_M^{\frac{1}{2}r}E(\phi^{\frac{1}{2}r}).$$

[8] This would certainly be true for the test-based variety such as the "positive rule" and "pre-test" estimators.

Moment and IV Selection Approaches: A Comparative Simulation Study

Mehmet Caner* Esfandiar Maasoumi† Juan Andrés Riquelme‡

2016

*Ohio State University, Department of Economics,
†Emory University, Department of Economics,
‡North Carolina State University, Department of Economics.

1 Introduction

- Compare Three moment selection approaches, followed by post selection estimation strategies.
- 1. Adaptive Lasso of Zou (2006), extended by Liao (2013) to possibly invalid moments in GMM. We select valid instruments.
- 2. J test, as in Andrews and Lu (2001).
- 3. Penalized Continuous Updating (CUE), based on Hong et al. (2003) with penalized generalized empirical likelihood. [e.g., empirical likelihood, and exponential tilting].

- Final Stage Estimation:
- 1. Unpenalized GMM; Information criteria in Andrews (1999)
- 2. Unpenalized CUE
- 3. Model averaging technique of Okui (2011).
- Simulations: Which selection criterion can better select valid moments and/or eliminate invalid ones?
- Given the chosen IVs, which strategy delivers better finite sample performance?
- Bottom Line: Adaptive Lasso in model selection stage, coupled with either unpenalized GMM or moment averaging of Okui delivers generally the smallest RMSE.

2 Choices/Strategy/Literature

- Adaptive Lasso, computational advantages in large scale problems, Penalized methods in Andrews (1999) and Hong et al. (2003) not computationally advantageous, favored to determine valid IVs.
- In the final stage employ Okui (2011) model averaging for better Mean Squared Error, and smaller bias, compared to unpenalized GMM and CUE estimation.

- Shrinkage; is Eclectic, Misspecified Models, computational efficiency in high-dimensions. Hastie et al. (2009, section 3.6) conclude shrinkage better in model selection in reducing estimation error.

- Liao (2013) show GMM shrinkage procedures have the oracle property in selection; adding additional valid moments improves efficiency for strongly identified parameters.

- Cheng and Liao (2012) proposed a weighted tuning parameter to shrink invalid and redundant moments.

- Assuming Valid IVs, Belloni, Chernozhukov and Hansen (2011) utilize LASSO-type estimators in the many IV case, asymptotic oracle-efficiency. Caner (2009) and Caner and Zhang (2013) use shrinkage for model selection in a GMM.

- Weak IVs (Hausman et al. (2005); Andrews and Stock (2007)). Cheng and Liao (2012) suggest shrinkage is robust in discarding invalid IVs, but tends to retain redundant ones.

- Monte Carlo simulations allow combining several complexities:
- linear settings,
- small and large sample sizes,
- fixed and increasing number of moment conditions,
- weak and strong identification,
- local-to-zero moment conditions,
- homoskedastic and heteroskedastic errors.

3 Theoretical Framework

3.1 Moment Selection Methods

– Sequence of rv $\{Z_i\}_{i=1}^n$, unknown probability distribution.

– Selecting r valid moments from q candidates. A minimum of $s \geq p$ valid moments required to identify θ; $p \equiv \dim(\theta)$; q has two types, for $i = 1, \ldots, n$

$$E[g_S(Z_i, \theta^0)] = 0 \qquad S = \{1, \ldots, s\} \tag{1}$$

$$E[g_{S^c}(Z_i, \theta^0)] \stackrel{?}{=} 0 \qquad S^c = \{s+1, \ldots, q\} \tag{2}$$

the sign $\stackrel{?}{=}$, relationship MAY not hold for some in S^c.

– $r = s + s_v$, where s KNOWN valid moments, s_v valid moments in the remaining $q - s$; S^c moments that may or may not be valid.

– θ^0 "the true" of dimension p;

– Empirical Moments $g_n(Z_i, \theta) : \Theta \mapsto \mathbb{R}^q = (1/n) \sum_{i=1}^n g(Z_i, \theta)$ converges in probability to $g^0(Z_i, \theta)$ as $n \to \infty$,

– a random weighting matrix W_n of dimension equal to No. of moments, Correct moment conditions is s.t. $g^0(\theta^0) = \mathbf{0}$.

– The standard GMM estimator of θ^0, $\hat{\theta}_n$ is

$$\hat{\theta}_n \equiv \underset{\theta \in \Theta}{\operatorname{argmin}}\, J(\theta, \bar{W}_n),$$

where $\bar{W}_n$ is a $p \times p$ symmetric and positive definite weight matrix and the objective function (Hansen, 1982) is

$$J(\theta, \bar{W}_n) \equiv n \cdot g_n(\theta)' \bar{W}_n g_n(\theta), \tag{3}$$

with $g_n(\theta) = n^{-1} \sum_{i=1}^{n} g(Z_i, \theta)$, and Θ is a compact subset of $\mathbb{R}^p$. Let $g(Z_i, \theta) = g_i(\theta)$.

4 The Model

–

$$y = Y\theta_0 + \varepsilon \tag{4}$$

$$Y = Z\pi_0 + u \tag{5}$$

y is $n \times 1$ vector, Y is a $n \times p$ endogenous vars, Z is an $n \times q$ instruments, ε and u are unobserved with constant second moments And Correlated with each other.

– Instruments Z_{i1} $(s \times 1)$ are valid and Z_{i2} $(q - s \times 1)$ are suspect-invalid.

$*$ The sample moments:

$$g_n(\theta, \beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta, \beta),$$

where $g_i(\theta, \beta) = (g_{i1}(\theta)', g_{i2}(\theta, \beta)')'$ with

$$g_{i1}(\theta) = Z_{i1}(y_i - Y_i\theta),$$

$$g_{i2}(\theta, \beta) = Z_{i2}(y_i - Y_i\theta) - \beta.$$

$$W_n = \frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta}, \tilde{\beta}) g_i(\tilde{\theta}, \tilde{\beta})',$$

where $\tilde{\theta}, \tilde{\beta}$ are the first step GMM estimators with I_q as the weight matrix.

* Adaptive GMM Shrinkage (Liao, 2013); advantage of selecting valid moments and estimate θ in a single step; adding a slackness parameter vector β_0 (2):

$$E\begin{bmatrix} g_{i1}(\theta_0) \\ \\ g_{i2}(\theta_0, \beta_0) \end{bmatrix} = 0.$$

* verified by testing $\beta_0 = 0$. Condition j is valid if $\beta_{0j} = 0$, for $j = 1, \cdots q - s$.

* Adaptive Lasso:

$$(\hat{\theta}_n^{alasso}, \hat{\beta}_n^{alasso}) = \underset{(\theta,\beta)\in\Theta\times B}{\operatorname{argmin}} \left[g_n(\theta,\beta)' W_n g_n(\theta,\beta) + \lambda_n \sum_{j=1}^{q-s} \hat{\omega}_j |\beta_j| \right] \tag{6}$$

weights $\hat{\omega}_j = \frac{1}{|\tilde{\beta}_j|}$, and $\tilde{\beta}_j$ is the unpenalized standard GMM estimator using all q moments.

* Adaptive lasso (ALASSO) penalizes slackness parameter by its l_1 norm; Has the oracle property (β_{0j} is shrunk to zero for valid moments); can be solved by using the LARS algorithm (Efron et al., 2004). Shrinkage by tuning parameter $\lambda_n \geq 0$; large values shrink more, and $\lambda_n = 0$ corresponds to GMM.

4.1 GMM Information Criteria

* The second MSC is a Penalization of J-
* $c \in \mathbb{R}^{q-s}$ denotes a *moment selection vector* of zeros and ones, if jth moment is valid, the jth element of c is one,
* $|c| = \sum_{j=1}^{q-s} c_j$ number of moments selected by c;
* Z_{ic} vector Z from which the jth element is deleted if corresponding jth element in c is zero.
* Corresponding weight matrix is $\bar{W}_n^c$ of dimension $s + |c| \times s + |c|$.

* The MSC estimator objective function has the general form:

$$\text{MSC}_n(c) = J_c(\theta, \bar{W}_n^c) - h(|c|)\kappa_n, \tag{7}$$

where $J_c(\theta, \bar{W}_n^c) = g_n(\theta)'\bar{W}_n^c g_n(\theta)$ uses the $s + |c|$ moments in GMM objective function. $g_n(\theta)$ is defined immediately below equation (3).

* In (7), $\bar{W}_n^c = n^{-1}\sum_{i=1}^n Z_{ic}Z_{ic}'\bar{\epsilon}_i^2$, where $\bar{\epsilon}_i = y_i - Y_i\bar{\theta}$,
* and $\bar{\theta}$ is inefficient GMM and Z_{ic}.

* EXAMPLE: consider two potentially valid IVs $Z1, Z2$; Possible combinations are $Z1$ only, $Z2$ only, $Z1, Z2$; For $Z1$ only, get inefficient GMM for weight matrix, then efficient GMM; Repeat the same for $Z2$, and then for $Z1, Z2$; choose minimizer of (7).

* Andrews (1999) uses $h(|c|) = |c| - p$ and three different choices of κ_n: (AIC, BIC, Hannan-Quinn)

$$\begin{aligned}
&\text{GMMBIC:} && \text{MSC}_{\text{BIC},n}(c) = && J_c(\theta, \bar{W}_n^c) - (|c| - p) \ln n \\
&\text{GMMAIC:} && \text{MSC}_{\text{AIC},n}(c) = && J_c(\theta, \bar{W}_n^c) - 2\ (|c| - p) \\
&\text{GMMHQIC:} && \text{MSC}_{\text{HQIC},n}(c) = && J_c(\theta, \bar{W}_n^c) - 2.1\ (|c| - p) \ln \ln n
\end{aligned}$$

* We examine GMMBIC method; BIC gives consistency in both adaptive lasso and Andrews and Lu (2001).

* We only use CUE objective function due to poor performance of empirical likelihood and exponential tilting, as shown in Hong et al. (2003), weight matrix is updated continuously: $W_{n,cue} = n^{-1} \sum_{i=1}^{n} Z_{ic} Z_{ic}' \epsilon_i^2$, where $\epsilon_i = y_i - Y_i \theta$.

4.2 Parameter Estimation

* for a shrinkage parameter m define $P^m = P_{Z_I} + mP_{Z_{II}}$ and the STSLS as

$$\hat{\theta}^{stsls}_{n,s} = (Y'P^mY)^{-1}Y'P^my,$$

* Z_I is s valid moments; Z_{II} selected by an information criterion such as ALASSO or GMM. m is chosen to minimize Nagar (1959)-type approximation of MSE.

5 Monte Carlo Simulations

* The DGP in (4) and (5); Only one endogenous variable, true $\theta_0 = 0.5$; $(Z, \varepsilon, u) \sim N(\mathbf{0}, \Sigma)$ where

$$\Sigma = \begin{bmatrix} \sigma_{zz}^2 \boldsymbol{I}_q & \boldsymbol{\sigma}'_{Z\varepsilon} & \mathbf{0}'_q \\ \boldsymbol{\sigma}_{Z\varepsilon} & \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \mathbf{0}_q & \sigma_{\varepsilon u} & \sigma_u^2 \end{bmatrix}$$

is $(q+2) \times (q+2)$ symmetric, σ_{zz}^2 is variance of IVs $\boldsymbol{I}_q$ an identity matrix of order q, $\boldsymbol{\sigma}_{Z\varepsilon}$ is a $q \times 1$ vector of correlations between the instruments and the structural error, $\mathbf{0}_q$ is a $q \times 1$ vector of zeros, $\sigma_{\varepsilon u}$, σ_ε^2 and σ_u^2 are scalars.

∗ Heteroskedastic errors

$$\varepsilon_i^* = \varepsilon_i \left\| Z_i \right\|, \quad \text{with } \left\| Z_i \right\| = \sqrt{Z_{i1}^2 + \cdots + Z_{iq}^2}$$

∗ A moment is valid if $E[g(Z_i, \theta_0)] = E[Z_i'(y - Y\theta_0)] = E[Z_i'\varepsilon] = \sigma_{Z\varepsilon} = 0$;

∗ Invalid moments: Construct $\boldsymbol{\sigma}_{Z\varepsilon}$ vectors in two ways:

∗ (1) constant correlation $D \neq 0$, and (2) local to zero correlation of the form $1/n$, $1/\sqrt{n}$ and $1/\sqrt[3]{n}$ to explore different convergence rates.

∗ $\varepsilon_i^* = \varepsilon_i$ is homoscedastic.

* Setup 1: Simulate fixed number of moments: $q = 11$, $s = 3$ and $r = 7$; 11 total moments, 3 of them are valid, select from 8, 4 valid and 4 invalid; homoskedastic errors.

* Setup 2: number of valid moments increase with sample size: $q = \sqrt{n}$, $s = \sqrt{q}$ and $s_v = (q - s)/2$, choose among $q - s$ candidates, half valid. Errors are heteroskedastic.

* In Setup 1, Σ is a 13×13 matrix constructed: simulate $Z \in \mathbb{R}^{11}$ in three categories: IVs known to be strong and valid ($s = 3$); in next set of instruments first four instruments valid ($s_v = 4$), the last $q - r = 4$ invalid.

* The last elements of Σ are $\boldsymbol{\sigma}_{Z\varepsilon} = (0, 0, 0, 0, 0, 0, 0, D, D, D, D)$ in the constant correlation case, and $\boldsymbol{\sigma}_{Z\varepsilon} = (0, 0, 0, 0, 0, 0, \frac{h}{n}, \frac{h}{\sqrt{n}}, \frac{h}{\sqrt[3]{n}}$ in the local to zero scenario.

* We use three rates for the local to zero moments which are recycled as needed. We set $\sigma_\varepsilon^2 = 1$, $\sigma_u^2 = 0.5$, $D = 0.2$ and $h = 1$.

* For each correlation structure, weak and strong identification on π_0 in equation 5: Strong identification $\pi_0 = 2 \cdot \mathbf{1}_{11}$, Weak identification case $\pi_0 = (2 \cdot \mathbf{1}_3, 0.2 \cdot \mathbf{1}_8)$ with $\mathbf{1}_\ell$ being a row vector of ones of length ℓ.

* Variance of IVs $\sigma^2_{zz} = \{0.5, 1.0\} \cdot \boldsymbol{I}_q$; covariance between errors $\sigma_{\varepsilon u} = 0.5$.

* Two cases: in Case 1 $\sigma^2_{zz} = 0.5 \cdot \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.5$, in the Case 2 $\sigma^2_{zz} = 1.0 \cdot \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.5$.

* We have other cases for the covariance matrix: Case 3 $\sigma^2_{zz} = \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.5$, in Case 4 $\sigma^2_{zz} = \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.9$; Case 5 $\sigma^2_{zz} = 2 \cdot \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.5$; Case 6 $\sigma^2_{zz} = 2 \cdot \boldsymbol{I}_q$ and $\text{cov}_{ue} = 0.9$. These cases and the local-to-zero ones are available on request.

* Sample sizes $n = \{50, 100, 250\}$.1000 repetitions.

6 Results

* Focus only on most relevant and salient: Cases 1 and 2 for Setups 1 and 2, constant correlation invalid moments. Generally results hold across all the alternative setups.
* weak and strong identification cases with $\sigma_{zz}^2 = 0.5\boldsymbol{I}_q$ and $\sigma_{zz}^2 = 1\boldsymbol{I}_q$,
* The R^2 of the first stage regression is in Table **??**, Ranges from 0.533 to 0.944, depending on strength of identification and number of obs. More than NINE parameter estimates:

* For efficient GMM we have three estimators:

* GMM is the GMM estimator using the full set of moments,

* GMM_{PEN}–MA uses the penalized GMM estimator in Andrews and Lu (2001) for model selection, followed by Okui's moment averaging estimator.

* GMM_{PEN}–GMM selects moments in the same way but then parameter estimated by efficient GMM.

* CUE denotes the CUE estimator using the full set of moments,
* CUE_{PEN}–MA is obtained by selecting with penalized CUE criteria and moment averaging estimator
* and CUE_{PEN}–CUE selects moments by penalized CUE and estimates θ_0 by CUE.
* Summary is presented in Tables 1 and 2 for model selection and post selection estimation.

* In Table 1: Average ranking of each method by Probability of Selecting the Exact valid moments detailed in Tables **??** and **??** for each sample size and identification strength.

* Adaptive Lasso is the best for "perfect" moment selection.

Table 1: Summary of the Performance of the Moment Selection Techniques

	Setup 1	Setup 2
ALASSO	1.25	1.33
GMM_{PEN}	2.58	2.67
CUE_{PEN}	1.83	1.92

Average ranking based on probability of selecting the exact valid moments. The latter are in Tables **??** and **??**, by sample size and strength of identification. In case of a tie, same ranking (we can have two first or two second places). ALASSO, GMM_{PEN} and CUE_{PEN} stands for adaptive lasso, penalized GMM and penalized CUE respectively.

* Table 2 performance of the final stage estimation by RMSE.
* Rankings relative performance from Tables **??** to **??**, by sample size and identification. Estimator with Smallest value has Rank 1; Average Rankings range from 1 to 9, frequency of being in the Top Three from 0 to 12.
* From Table 2 Adaptive Lasso is best to select, followed by moment averaging (ALASSO–MA).
* Moment averaging procedure improves estimation for the three moment selection techniques.
* Worst estimators are based on CUE.
* In hetero setup (Setup 2) Adaptive Lasso-MA + moment averaging is still best in RMSE, but not as good as in the homoskedastic case (Setup 1).

Table 2: Summary of the Performance of the Post Selection Techniques

	Setup 1		Setup 2	
	Average Ranking	Times at the top three	Average Ranking	Times at the top three
ALASSO–MA	1.17	12	1.75	11
ALASSO–GMM	2.00	12	2.25	11
ALASSO–CUE	4.83	4	4.08	5
GMM	4.33	3	5.67	1
GMM_{PEN}–MA	1.67	12	2.50	9
GMM_{PEN}–GMM	3.00	8	3.00	7
CUE	6.33	0	7.08	0
CUE_{PEN}–MA	1.92	11	2.83	7
CUE_{PEN}–CUE	4.25	5	5.08	2

The performance is analyzed in terms of the RMSE. The rankings are based in the relative performance in the Tables **??** to **??**. The estimator with the smaller value takes the rank of 1. If there is a tie the estimators are given the same rank. The average ranking ranges from 1 to 9 and the times at the top three from 0 to 12. ALASSO–MA is the estimator obtained by selecting the moments using the Adaptive Lasso method in the first stage and then using Okui's moment averaging estimator in the second stage. ALASSO–GMM and ALASSO–CUE are the estimates using adaptive lasso to select the valid moments in the first stage and then use them in the efficient and unpenalized CUE and GMM respectively. For the efficient GMM we have three estimators: GMM is the GMM estimator using the full set of moments, GMM_{PEN}–MA uses the penalized GMM estimator in Andrews and Lu (2001) for model selection and then use Okui's moment averaging estimator in the first stage. GMM_{PEN}–GMM selects the moments in the same way as the previous methods but then the structural parameter is estimated using efficient GMM. In the same way, CUE denotes the CUE estimator using the full set of moments, CUE_{PEN}–MA is the estimator obtained by selecting the moments using the penalized CUE criteria and using these moments in the moment averaging estimator and CUE_{PEN}–CUE selects the moments using penalized CUE and estimates θ_0 using the CUE estimator.

References

Andrews, Donald W. K. (1999) "Consistent moment selection procedures for generalized method of moments estimation," *Econometrica*, Vol. 67, No. 3, pp. 543–564.

Andrews, Donald W. K. and Biao Lu (2001) "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models," *Journal of Econometrics*, Vol. 101, No. 1, pp. 123–164.

Andrews, Donald W.K. (1997) "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," Cowles Foundation Discussion Papers 1146R, Cowles Foundation for Research in Economics, Yale University.

Andrews, Donald W.K. and James H. Stock (2007)

"Testing with many weak instruments," *Journal of Econometrics*, Vol. 138, No. 1, pp. 24–46.

Belloni, A., V. Chernozhukov, and C.B. Hansen (2011) "Lasso Methods for Gaussian Instrumental Variables Models," working paper, Massachusetts Institute of Technology, Department of Economics.

Canay, Ivan A. (2010) "Simultaneous selection and weighting of moments in {GMM} using a trapezoidal kernel," *Journal of Econometrics*, Vol. 156, No. 2, pp. 284–303.

Caner, M. and H. Zhang (2013) "Adaptive Elastic Net GMM with Diverging Number of Moments," *Journal of Business and Economics Statistics, Forthcoming*.

Caner, Mehmet (2009) "Lasso-Type GMM Estimator," *Econometric Theory*, Vol. 25, pp. 270–290.

Cheng, Xu and Zhipeng Liao (2012) "Select the Valid

and Relevant Moments: A one-step procedure for GMM with many moments," PIER Working Paper Archive 12–045, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004) "Least angle regression," *Annals of Statistics*, Vol. 32, pp. 407–499.

Hansen, Bruce E. (2007) "Least Squares Model Averaging," *Econometrica*, Vol. 75, No. 4, pp. 1175–1189.

Hansen, Lars Peter (1982) "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, Vol. 50, No. 4, pp. 1029–1954.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer, corrected edition.

Hausman, Jerry, James H. Stock, and Motohiro Yogo (2005) "Asymptotic properties of the Hahn-Hausman test for weak-instruments," *Economics Letters*, Vol. 89, No. 3, pp. 333–342.

Hong, Han, Bruce Preston, and Matthew Shum (2003) "Generalized Empirical Likelihood Based Model Selection Criteria For Moment Condition Models," *Econometric Theory*, Vol. 19, No. 06, pp. 923–943.

Kuersteiner, Guido and Ryo Okui (2010) "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, Vol. 78, No. 2, pp. pp. 697–718.

Liao, Zhipeng (2013) "Adaptive GMM shrinkage estimation with consistent moment selection," *Econometric Theory*, Vol. FirstView, pp. 1–48.

Nagar, A. L. (1959) "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in

Simultaneous Equations," *Econometrica*, Vol. 27, No. 4, pp. 575–595.

Newey, W.K. and R.J Smith (2000) "Asymptotic Bias and Equivalence of GMM and GEL Estimators," Discussion paper 01/517, University of Bristol, Department of Economics.

Okui, Ryo (2011) "Instrumental variable estimation in the presence of many moment conditions," *Journal of Econometrics*, Vol. 165, No. 1, pp. 70–86.

Smith, Richard J. (1992) "Non-Nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, Vol. 60, No. 4, pp. 973–980.

Zou, Hui (2006) "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, Vol. 101, pp. 1418–1429.

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2007)

"On the "degrees of freedom" of the lasso," *The Annals of Statistics*, Vol. 35, No. 5, pp. 2173–2192.

Aggregation of Misspecified Asset Pricing Models

A Taxonomy of Misspecified Models

- Bernando and Smith (1994) characterize and taxonomize the different views regarding model comparison and selection.

1. The first perspective, that includes Bayesian model averaging and model selection, is conditioning on one of the models being 'true'.
   - in this approach, the ambiguity about the true model is resolved asymptotically and the mixture, that summarizes the beliefs about the individual models, assigns a weight of 1 to one of the models.
2. Another possibility is also to assume that a true model exists but it is too complicated or cumbersome to implement.
   - i.e., all of the candidate models are viewed as approximations of this fully-specified belief model and hence misspecified.
3. The third view dispenses completely with the notion of a true model and treats the candidate models as genuinely misspecified either because they are believed to represent different aspects of the underlying DGP or because the underlying structure is completely unknown.

Model Aggregation (Gospodinov and Maasoumi, 2016)

- Suppose there are M proposed misspecified models, $\hat{y}_i = y_i(\hat{\gamma}_i)$, $i = 1, ..., M$, for the undiscovarable true model m.
- Each model is treated as an incomplete 'indicator' of the latent DGP.
- Then, a model averaging rule would aggregate information from all of these models and construct a pseudo-true model $\tilde{y}$.
- We are interested in finding the aggregator $\tilde{y}_t$ with a distribution that is as close as possible to the multivariate distribution of $\hat{y}_i$'s.
- Maasoumi (1986) generalizes the pairwise criteria of divergence to:

$$D_\rho(\tilde{y}, \hat{y}; w) = \sum_{i=1}^{M} w_i \left\{ \sum_{t=1}^{T} \tilde{y}_t \left[\left(\frac{\tilde{y}_t}{\hat{y}_{i,t}} \right)^\rho - 1 \right] \Big/ \rho(\rho+1) \right\},$$

- The aggregator that minimizes $D_\rho(\tilde{y}, \hat{y}; w)$ subject to $\sum_{i=1}^{M} w_i = 1$ is

$$\tilde{y}_t^* \propto \left[\sum_{i=1}^{M} w_i y_{i,t}^{-\rho} \right]^{-1/\rho}.$$

  - Linear pooling of models is obtained as a special case when $\rho = -1$.

Model Aggregation

Two methods for estimating w and ρ.

1. Method 1: HJ-distance approach
   - For given $(\hat{y}_{1,t}, ..., \hat{y}_{M,t})'$, construct the pricing errors of the aggregator
   $$\tilde{e}_T(w, \rho) = \frac{1}{T} \sum_{t=1}^{T} R_t \left[\sum_{i=1}^{M} w_i \hat{y}_{i,t}^{-\rho} \right]^{-1/\rho} - 1_N.$$
   - The unknown parameters $\theta = (w', \rho)'$ are obtained by minimizing the HJ-distance of $\tilde{e}_T(\theta)$ subject to $w_i \geq 0$ and $\sum_{i=1}^{M} w_i = 1$.
2. Method 2: minimizing the distance between two distributions.
   - Let p be the density of some pivot and q denote the density of the aggregator $\tilde{y}_t(\theta) = \left[\sum_{i=1}^{M} w_i \hat{y}_{i,t}^{-\rho} \right]^{-1/\rho}$.
   - Within the Cressie-Read family, minimize the Hellinger distance
   $$\mathcal{H} = \frac{1}{2} \int \left(p^{1/2}(x) - q^{1/2}(x) \right)^2 dx,$$
   subject to the relevant restrictions, to obtain an estimate of θ.
     - the Hellinger distance is a proper measure of distance since it satisfies the triangular inequality.