Regularized Empirical Likelihood as a Solution to the No Moment Problem: The Linear Case with Many Instruments

Pierre Chaussé *†

November 29, 2017

Abstract

In this paper, we explore the finite sample properties of the generalized empirical likelihood for a continuum, applied to a linear model with endogenous regressors and many discrete moment conditions. In particular, we show that the estimator from this regularized version of GEL has finite moments. It therefore solves the issue regarding the no moment problem of empirical likelihood. We propose a data driven method to select the regularization parameter based on a cross validation criterion, and show that the method outperforms many existing methods when the number of instruments exceeds 20.

Classification JEL: C13, C30
Keywords: Many Instruments, Weak Instruments, Regularization, Cross-validation

1 Introduction

When we want to estimate linear models with endogenous regressors for which there exists many valid instruments, we often rely on the two-stage least squares (2SLS) for iid observations or the two-step generalized method of moments (GMM) of Hansen (1982) for more general data generating processes. However, the bad small sample properties of their estimators when the number of instruments is large or relatively weak, redirected our attention toward alternative methods such as the continuously updated GMM (CUE) of Hansen et al. (1996) or the generalized empirical likelihood

*Department of Economics, University of Waterloo, Waterloo ON, Canada, N2L 3G1. Email: pchausse@uwaterloo.ca, Phone: +1-519-888-4567 x32422

†The author wants to acknowledge financial supports from the Social Sciences and Humanities Research Council of Canada (Grant number: 114666).

(GEL) of Smith (1997). These alternative methods have the potential to outperform GMM if we compare their second order asymptotic properties derived by Newey and Smith (2004). In fact the authors show that the bias of CUE is smaller than the bias of GMM, and the bias of empirical likelihood (EL), which is a subset of the GEL family of estimators, exhibits the smallest bias. Also, the bias-corrected version of EL is the most efficient in terms of the root mean squared error (RMSE).

The issue raised by many is that the theoretical results, especially when we have many potentially weak instruments, can hardly be reproduced in small samples. It is now well established that CUE in such cases, as it is also true for the limited information maximum likelihood method (LIML), suffers from the no moment problem. In fact, simulations show that the CUE estimates are sometimes unbounded. The same problem was detected by Guggenberger (2008) for the case of EL estimators. As a result, many have suggested modifications in an attempt to solve this no moment problem. For example, the Jackknife version of 2SLS (J2SLS) of Hahn et al. (2004) is an attempt to reduce the bias of 2SLS to a level comparable to LIML, without creating unbounded moments. Also, Hausman et al. (2011) propose to stabilize the CUE (RCUE for regularized CUE) estimates by regularizing both the vector of coefficients and the inverse of the covariance matrix, a method that relies on the choice of two regularization parameters. They show, through simulations, that there exists a pair of parameters that solves the no moment problem, but they do not provide us with a method for selecting them. Finally, a similar approach is taken by Carrasco and Tchuente (2015) for the case of LIML. The method is a special case of the GMM for a continuum (CGMM) of Carrasco and Florens (2000), and relies on only one regularization parameter. They provide a data driven method for selecting the parameter based on the second order approximation of the RMSE.

In this paper, we develop a method for the case of GEL which is based on GEL for a continuum (CGEL) of Chaussé (2011). As it is the case for CGMM or RCUE, CGEL is a regularized version of GEL. CGEL was first developed to address the ill-posed issue created by moment conditions defined on a continuum, but we show that it can also be applied to the case of a finite number of moment conditions. As for RCUE, CGEL regularizes the first order condition. However, the regularization is only applied to the vector of Lagrange multipliers, λ, associated with the moment conditions. Therefore, it only relies on one regularization parameter, α. Since CUE is a member of the family of GEL estimators, it is a good alternative to RCUE. We show that our method approaches the one step GMM with the identity matrix as α increases, which implies that there is a regularization parameter that solves the no moment problem of all GEL methods. We then propose a method for selecting α that regularizes the solution only when it is necessary, and therefore preserves as much as possible the higher order properties of GEL obtained by Newey and Smith (2004). Finally, we analyze the finite sample properties of CGEL using Monte Carlo experiments based on the data generating processes used by Guggenberger (2008) and Hausman et al. (2011).

First, our simulations show that regularizing λ is sufficient to solve the no moment problem of GEL. Second, we find that CGEL estimators, compared with RCUE and the two-step GMM, are always the least biased, and have the smallest RMSE when the number of instruments exceeds 20 and the error term is homoscedastic. Although CGEL remains the least biased, its RMSE is greater then GMM in presence of heteroscedasticity, when there are few extremely weak instruments (a $R^2 = 0.002$ in the first stage regression). However, as the number of irrelevant instruments increases, holding the concentration parameter constant, the RMSE of CGEL becomes comparable to GMM.

In the next section, we explain what we mean by "the no moment problem", and show why GEL estimates are sometimes unbounded. In Section 3, we explain the difference between GEL and CGEL, and show why the regularization procedure solves the no moment problem of GEL. In Section 4, we present our procedure for selecting α, and show that it solves the issue raised by Guggenberger (2008). In Section 5, we compare CGEL with the RCUE method of Hausman et al. (2011), and we conclude in Section 6.

2 The no moment problem of GEL

We first review the results obtained by Guggenberger (2008), who finds that the distribution of the empirical likelihood (EL) estimator, like the one from the limited information maximum likelihood estimator (LIML), have extreme heavy tails, which suggests that their moments don't exist. Indeed, for some samples the estimates are unbounded. This finding contradicts the asymptotic results of Newey and Smith (2004) which suggest that EL should outperformed the two-step GMM. In order to see what happens for those samples, let us look at the shape of the objective function. The model considered is

$$y_i = \delta x_i + \varepsilon_i, \tag{1}$$

with

$$x_i = z_i'\pi + u_i, \tag{2}$$

where the $q \times 1$ vector π is equal to $\{\eta, \eta, \cdots, \eta\}$ and $Corr(\varepsilon_i, u_i) = \rho \neq 0$. The theoretical R^2 of the first stage regression is:

$$R^2 = \frac{q\eta^2}{q\eta^2 + 1}. \tag{3}$$

We can then control the strength of the instruments by selecting a value for η associated with the desired R^2. Since this is the data generating process (DGP) used by Guggenberger (2008), we will refer to it as GUG. The generalized empirical likelihood estimator (GEL) of δ is defined as follows:

$$\hat{\delta} = \arg\min_{\delta} \left[\arg\max_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \rho(\lambda' g_i(\delta)) \right], \tag{4}$$

where

$$g_i(\delta) = (y_i - \delta x_i) z_i,$$

and $\rho(v)$ is either $\log(1-v)$ for the Empirical Likelihood (EL) of Owen (1988), $\exp(-v)$ for the exponential tilting (ET) of Kitamura and Stutzer (1997), or $(-v - v^2/2)$ for the continuously updated GMM (CUE) or the Euclidean Empirical Likelihood (EEL) of Antoine et al. (2007). In Equation (4), λ is the Lagrange multiplier associated with the moment conditions (6) in the following primal problem:

$$\hat{\delta} = \arg\min_{\delta, p_i} \sum_{i=1}^{n} h_n(p_i) \tag{5}$$

subject to

$$\sum_{i=1}^{n} p_i g_i(\delta) = 0 \tag{6}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{7}$$

where $h_n(p_i)$ is a discrepancy function measuring the distance between p_i and $1/n$ (see Newey and Smith (2004) for more details).

We want to look at the shape of the following function:

$$P(\delta) = \frac{1}{n} \sum_{i=1}^{n} \rho\big[\lambda(\delta)' g_i(\delta)\big],$$

where

$$\lambda(\delta) = \arg\max_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \rho\big[\lambda' g_i(\delta)\big].$$

To see what happens when the estimate is out of bound, we generated the above model with $\delta = 0$, $\rho = 0.5$, $q = 20$, $R^2 = 0.002$ and the sample size n equals to 250 until the estimate of δ becomes out of bound. Figure 1 shows the shape of $P(\delta)$ when the method diverges (Figures 1c and 1d) and when it converges (Figures 1a and 1b). It seems that when the instruments are weak, we sometimes end up with a sample containing so little information that the function is either not convex in the neighborhood of the true value or the local minimum is not global. When it is the case, the numerical algorithm may converge to a flat region of the objective function[1]. Such flat regions exist because $\lambda(\delta)$ converges to a constant when δ gets far away from its

[1]Guggenberger (2008) solves the EL model with a grid search to obtain a global minimum. It is therefore not surprising that he obtains unbounded solutions in such samples

true value, as shown on Figure 2. If our sample behaves like the data used to produce Figure 1c, we have some hope because there is a local minimum around the true value. However, the ability for our numerical algorithm to reach that local minimum will depend heavily on the starting value. Also, in models with more regressors, it could be a difficult task.

This shape is only present in the GEL case, which includes the EL, the ET, and the CUE methods. We obtain a well defined objective function with a unique global minimum in the case of the two-step GMM. Table 1 shows the properties of $\hat{\delta}$ from the different methods based on a simulation, and Figure 3 illustrates the no moment problem of the three GEL methods. Since the coefficient δ is a scalar, it is estimated by the Brent method with upper and lower bounds equal to $\pm$1,000. Clearly, the distribution of all GEL estimators have heavy tails suggesting unbounded moments.

Table 1: *Properties of $\hat{\delta}$ for different methods using the Guggenberger model with $R^2 = 0.002$, $n = 200$, $\rho = 0.5$ and $k = 20$, based on 1,000 replications*

	Mean-Bias	Median-Bias	RMSE	S-dev	Interquartile range
Two-step	0.4898	0.4933	0.5344	0.2139	0.2832
EL	-2.8376	0.3801	47.8746	47.8144	2.1298
ET	2.3704	0.4305	42.1462	42.1006	2.0314
CUE	1.4783	0.4375	35.6503	35.6374	2.2079

There exists no formal proof of the no moment problem for GEL. Formal proofs would rely on assumptions about the joint distribution of our observations. For example, Mariano and Sawa (1972) show that limited information maximum likelihood estimators (LIML) have no moment, by deriving the extact distribution of the estimator under the normality assumption. Although the proof is informative, researchers were already aware of the instability of LIML. What matters in our case is that GEL can produce unreliable estimates in small samples when the instruments are relatively weak. Whether the instability of GEL is due to estimators having no finite moments or that they are simply very volatile is irrelevant. What we want to propose is a way of stabilizing the GEL estimator, while keeping as much as possible its desirable properties.

We want to argue that the volatility of the estimator of δ can be reduced by controlling the instability of $\hat{\lambda}$. In the GEL method, the moment conditions must be satisfied exactly for all n. This is done by adjusting the probabilities that are assigned to each observation. For a given δ, each of them is defined as

$$p_i = \frac{\rho'(\lambda(\delta)' g_i(\delta))}{\sum_{j=1}^{n} \rho'(\lambda(\delta)' g_j(\delta))},$$

and condition (6) must be satisfied. Therefore, λ must react to any change in δ. If we can somehow restrict λ from being too volatile, we would simultaneously stabilize $\hat{\delta}$.

One way to smooth λ is to use the generalized empirical likelihood for a continuum (CGEL) of Chaussé (2011). We don't have a continuum of moments in our model but the method applies also to discrete moments by defining inner products in Euclidean space instead of defining them in a functional space. In CGEL, the above definition of $\lambda(\delta)$ is replaced by the following:

$$\lambda_\alpha(\delta) = \min_\lambda \left\| \frac{1}{n} \sum_{i=1}^{n} \rho'\big[\lambda' g_i(\delta)\big] g_i \right\|^2 + \alpha \|\lambda\|^2, \tag{8}$$

with $\|x\|^2 = x'x$. The first term is the GEL first order condition (FOC) for $\lambda(\delta)$. Therefore, we try to make the FOC as close as possible to zero and impose a penalty on the volatility of the λ. The regularized objective function for δ becomes:

$$P(\delta; \alpha) = \frac{1}{n} \sum_{i=1}^{n} \rho\big[\lambda_\alpha(\delta)' g_i(\delta)\big]. \tag{9}$$

As it is clear from the definition of $\lambda_\alpha(\delta)$, the moment condition (6) is no longer satisfied. In fact, the method moves away from GEL as α increases. As we will see in Section 3, increasing α makes the GEL estimator shrinks toward the first-step GMM with weighting matrix equals to the identity matrix, which explains why the no moment problem disappears.

Figure (4) shows the effect of α on $P(\delta; \alpha)$, using the same problematic samples used to produce Figures 1c and 1d. We can see that as α increases, the shape becomes more and more compatible with an unbounded solution. The method is therefore potentially a solution to the no moment problem of GEL. We also see that the required value of α is different across different samples. We will discuss its impact on the properties of the estimator in the next section.

Other studies have addressed the issue of the no moment problem of CUE and LIML. For example, Hahn et al. (2004) propose a Jackknife version of the 2SLS (J2SLS) as an alternative to LIML. It is defined as:

$$\hat{\delta}_{J2SLS} = n\hat{\delta}_{2SLS} - \frac{n-1}{n} \sum_{i=1}^{n} \hat{\delta}^{(i)}_{2SLS},$$

where $\hat{\delta}^{(i)}_{2SLS}$ is the 2SLS estimator of δ obtained by removing the i^{th} observation. The bias is comparable to LIML and it does not suffer from the no moment problem since it is a weighted sum of 2SLS estimators, which are known to have finite moments (see Sawa, 1969). It is also a good alternative to the Bias-corrected 2SLS of Donald and Newey (2001) (B2SLS), which also suffers from the no moment problem. Hausman et al. (2011) derive a version of CUE (RCUE) in which the first order condition is regularized. In the method they propose, there are two regularization parameters:

one to regularize the inverse of the covariance matrix of the moment conditions, and one to regularize the vector of coefficients. They show through simulations that the method seems to solve the no moment problem for a given choice of the regularization parameters. But they do not provide us with procedures to select them. Also, they only use an approximation of the first order condition of the CUE which is known to be invalid in small samples with weak instruments (see Kleibergen (2005)). Finally, Carrasco and Tchuente (2015) derive a regularized version of LIML with a data driven method for selecting the regularization parameter.

We will denote the regularized version of GEL as CGEL to be consistent with the notation of Chaussé (2011). The case in which $\rho(v)$ is quadratic will be referred as CEEL[2], and CEL and CET will denote the regularized empirical likelihood and exponential tilting respectively. Table 2 shows the result[3] of a simulation using the same datasets used to produce Table 1. For now, we just set the regularization parameter α arbitrarily to 1.7, to illustrate the potential of CGEL even for fixed α. We compare the different CGEL to GMM, J2SLS and B2SLS. We can see that CGEL reduces considerably the dispersion of $\hat{\delta}$. All the CGEL are comparable to GMM in terms of RMSE, with CEL being the best. Also, they dominate the B2SLS method, and perform marginally better than J2SLS. Therefore, the modification we propose seems to solve the issue raised by Guggenberger (2008). The fixed α, however, is likely to affect the properties of $\hat{\delta}$. The main reason for selecting a large α was to avoid having too many unnstable estimates. In practice, we want to regularize only when it is necessary, because we want to preserve as much as possible the propoerties of GEL. We will present a data driven method for selecting α in Section 4.

Table 2: *Properties of $\hat{\delta}$ for different methods using the Guggenberger model with $R^2 = 0.002$, $n = 200$, $\rho = 0.5$ and $k = 20$, based on 1,000 replications. For the regularized methods, α is fixed and equal to 1.7.*

	Mean-Bias	Median-Bias	RMSE	S-dev	Interquartile range
Two-step	0.4898	0.4933	0.5344	0.2139	0.2832
CEEL	0.4589	0.4724	0.5845	0.3622	0.2941
CET	0.4729	0.4748	0.5498	0.2806	0.2698
CEL	0.4819	0.4786	0.5312	0.2238	0.2664
B2SLS	-0.2192	0.4907	25.7137	25.7257	1.1548
J2SLS	0.4632	0.4606	0.6301	0.4274	0.4941

[2]Chaussé (2011) shows that CGEL is not CCUE when $\rho(v)$ is quadratic. It is a result that holds only when we have a finite number of moment conditions and when we do not regularize the first order condition (see Newey and Smith (2004)).

[3]All estimations are performed in R using the gmm package of Chaussé (2010). By setting the argument *alpha* of the gel() function to a non-null value, the function estimates the model using the CGEL method.

3 CGEL vs GEL

Before proposing a method for selecting the regularization parameter, α, we analyze in this section its impact on the $\hat{\delta}$. In large samples, Chaussé (2011) shows that all CGEL methods are equivalent to the CGMM of Carrasco and Florens (2000), if α converges to zero at a certain speed. We are not interested in asymptotic results here because the no moment problem occurs in small and fixed samples. We rather want to see what happens when α increases for a given sample size. We will consider CUE versus CEEL because in the case of CUE, $\lambda(\delta)$ has a close form representation. Explicitly, the CUE estimator of the model represented by Equations (1) and (2), in the context of the GEL, is defined as:

$$\hat{\delta} = \arg\min_{\delta} \left[\arg\max_{\lambda} \frac{-1}{n} \sum_{i=1}^{n} \left(\varepsilon_i z_i'\lambda + \frac{1}{2}(\varepsilon_i z_i'\lambda)^2 \right) \right],$$

The FOC for $\lambda(\delta)$ is:

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i z_i + (\varepsilon_i z_i'\lambda)\varepsilon_i z_i) \\ &= \frac{Z'\varepsilon}{n} + \frac{[Z' D_{\varepsilon^2} Z]}{n}\lambda, \end{aligned}$$

where Z is the $n \times q$ matrix with the i^{th} row equals to z_i, ε is the $n \times 1$ vector of ε_i's, and D_{ε^2} is an $n \times n$ diagonal matrix with $[D_{\varepsilon^2}]_{ii} = \varepsilon_i^2$. It follows that:

$$\lambda(\delta) = -[Z' D_{\varepsilon^2} Z]^{-1} Z'\varepsilon.$$

If we substitute in the objective function for δ we obtain:

$$\begin{aligned} P(\delta) =& \frac{1}{n} \sum_{i=1}^{n} \left(\varepsilon' Z [Z' D_{\varepsilon^2} Z]^{-1} [\varepsilon_i z_i] - \frac{1}{2} \varepsilon_i^2 \varepsilon' Z [Z' D_{\varepsilon^2} Z]^{-1} z_i z_i' [Z' D_{\varepsilon^2} Z]^{-1} Z'\varepsilon \right) \\ =& \frac{\varepsilon' Z [Z' D_{\varepsilon^2} Z]^{-1} Z'\varepsilon}{n} - \frac{1}{2} \frac{\varepsilon' Z [Z' D_{\varepsilon^2} Z]^{-1} Z' D_{\varepsilon^2} Z [Z' D_{\varepsilon^2} Z]^{-1} Z'\varepsilon}{n} \\ =& \frac{1}{2} \frac{\varepsilon' Z [Z' D_{\varepsilon^2} Z]^{-1} Z'\varepsilon}{n} \\ \equiv& \frac{1}{2} \frac{\varepsilon' Z \Omega^{-1} Z'\varepsilon}{n}, \end{aligned}$$

where $\Omega = [Z' D_{\varepsilon^2} Z]$. We can see that $P(\delta)$ is indeed the objective function of CUE with the assumption that the error terms ε_i's are uncorrelated.

For the CGEL case, $\lambda_\alpha(\delta)$ is defined as:

$$\begin{aligned}\lambda_\alpha(\delta) =& \arg\min_\lambda \left\| \frac{Z'\varepsilon}{n} + \frac{[Z'D_{\varepsilon^2}Z]}{n}\lambda \right\|^2 + \alpha\|\lambda\|^2 \\ =& \arg\min_\lambda \left[\frac{2\varepsilon' Z[Z'D_{\varepsilon^2}Z]\lambda}{n^2} + \frac{\lambda'[Z'D_{\varepsilon^2}Z][Z'D_{\varepsilon^2}Z]\lambda}{n^2} + \alpha\lambda'\lambda \right].\end{aligned}$$

The FOC and solution are:

$$\begin{aligned}0 =& \frac{2[Z'D_{\varepsilon^2}Z]Z'\varepsilon}{n^2} + \frac{2[Z'D_{\varepsilon^2}Z][Z'D_{\varepsilon^2}Z]\lambda}{n^2} + 2\alpha\lambda \\ =& \frac{2[Z'D_{\varepsilon^2}Z]Z'\varepsilon}{n^2} + \left(\frac{2[Z'D_{\varepsilon^2}Z][Z'D_{\varepsilon^2}Z]}{n^2} + 2\alpha I \right)\lambda \\ =& \frac{2[Z'D_{\varepsilon^2}Z]Z'\varepsilon}{n^2} + \left(\frac{2[Z'D_{\varepsilon^2}Z]^2}{n^2} + 2\alpha I \right)\lambda \\ \lambda_\alpha(\delta) =& -\left(\left[\frac{Z'D_{\varepsilon^2}Z}{n}\right]^2 + \alpha I \right)^{-1} \left[\frac{Z'D_{\varepsilon^2}Z}{n}\right] \left[\frac{Z'\varepsilon}{n}\right] \\ \equiv& -\left(\Omega^2 + \tilde{\alpha} I\right)^{-1} \Omega Z'\varepsilon,\end{aligned}$$

where $\tilde{\alpha} = \alpha n^2$. The objective function for δ is therefore defined has:

$$\begin{aligned}P(\delta;\alpha) =& \frac{\varepsilon' Z\Omega \left(\Omega^2 + \tilde{\alpha} I\right)^{-1} Z'\varepsilon}{n} - \frac{1}{2}\frac{\varepsilon' Z\Omega \left(\Omega^2 + \tilde{\alpha} I\right)^{-1} \Omega \left(\Omega^2 + \tilde{\alpha} I\right)^{-1} \Omega Z'\varepsilon}{n} \\ =& \frac{\varepsilon' Z\Omega_\alpha^{-1} Z'\varepsilon}{n} - \frac{1}{2}\frac{\varepsilon' Z\Omega_\alpha^{-1}\Omega\Omega_\alpha^{-1} Z'\varepsilon}{n},\end{aligned}$$

where Ω_α^{-1} is the regularized inverse of Ω. It is almost the CUE with a regularized inverse of Ω. The extra term results front the fact that $\Omega_\alpha^{-1}\Omega \neq I$. The equality is obtained only if $\alpha = 0$. Therefore $P(\delta)$ is just a special case of $P(\delta;\alpha)$ with $\alpha = 0$.

We can rewrite $P(\delta;\alpha)$ in terms of the singular value decomposition of Ω. Let μ_i and ϕ_i be respectively the i^{th} eigenvalue and eigenvector of Ω. Then,

$$\begin{aligned}P(\delta;\alpha) =& \frac{1}{n}\sum_{i=1}^{q} \frac{\mu_i}{\mu_i^2 + \tilde{\alpha}} < \phi_i, Z'\varepsilon >^2 - \frac{1}{2n}\sum_{i=1}^{q} \frac{\mu_i^3}{(\mu_i^2 + \tilde{\alpha})^2} < \phi_i, Z'\varepsilon >^2 \\ =& \frac{1}{n}\sum_{i=1}^{q} \frac{0.5\mu_i^3 + \tilde{\alpha}\mu_i}{(\mu_i^2 + \tilde{\alpha})^2} < \phi_i, Z'\varepsilon >^2 \\ \equiv& \frac{1}{n}\sum_{i=1}^{q} W(\mu_i, \alpha) < \phi_i, Z'\varepsilon >^2\end{aligned}$$

We can think of the above expression as being a generalization of the GMM objective function. For $W(\mu_i, \alpha) = 1$, it is the objective function of the first-step GMM with

the weighting matrix equals to the identity matrix, $W(\mu_i, 0)$ generates the objective function of CUE, and $W(\tilde{\mu}_i, 0)$ the objective function of the two-step GMM, where $\tilde{\mu}_i$ is the i^{th} eigenvalue of the first-step estimate of Ω. $W(\mu_i, \alpha)$ controls the efficiency of $\hat{\delta}$ by putting small weights on more volatile linear combinations of the moment conditions. Figure 5 shows what happens to $W(\mu_i, \alpha)$ when δ moves away from the first-step GMM solution. It only shows the ones associated with the highest and lowest eigenvalues because all others lie between them. In a bad sample, the peaks of $W(\mu_i, \alpha)$ are so high that it offsets the well-behaved U-shape of $\|Z'\varepsilon\|^2$, which results in an objective function without global minimum. Figures 5c and 5d presents the same curves when $\alpha = 2$. We see that a higher α implies less sensitive $W(\mu_i, \alpha)$ to variations of the parameter δ, and a global minimum for $P(\delta, \alpha)$ comparable to the one-step GMM. Therefore, the regularisation solves the no moment problems for CUE. However, if α is too high, we no longer penalize highly volatile linear combinations of the moment conditions, which will likely result in less efficient estimators. It must therefore be carefully selected.

It is more difficult to generalize the analysis to all CGEL methods, because in general $\lambda(\delta)$ does not have a closed form expression and must be obtained numerically. However, we can easily show that if we use the Newton method, the first iteration for $\lambda(\delta)$ and $\lambda_\alpha(\delta)$, using the starting value $\lambda_0 = 0$, corresponds to CUE and CEEL respectively. We should therefore expect similar results regarding the effect of α on the objective function.

4 Selecting α

Selecting the regularization parameter is comparable to selecting the number of instruments or the number of regressors in linear regression models. In most cases, the selection is based on a large sample approximation of the RMSE. For example, Donald and Newey (2001) derive a second order approximation of the RMSE for 2SLS, and Donald et al. (2008) provide the second order approximation for GEL. In both cases, the optimal number of instruments, assuming that they are correctly ordered, is the one that minimizes the RMSE. Other methods control the number of instruments through a regularization parameter similar to CGEL, and therefore don't require any instruments ordering. Once again, the parameter selection is based on the asymptotic RMSE. For example, Carrasco (2012) proposes a data driven procedure for selecting the parameter in the case of instrumental variable with either many or a continuum of instruments, while Carrasco and Tchuente (2015) proceed similarly for LIML.

The above methods may eventually be valid as the sample size increases, but they cannot necessarily solve the no moment problem that we observe in small and fixed samples. Indeed, simulations such as the ones performed by Guggenberger (2008) contradict the second order properties of GMM and GEL derived by Newey and Smith (2004). It is therefore reasonable to assume that something is lost when we let the

sample size grows to infinity.

The first thing we want to do is to analyze the relationship between the RMSE and the value of the regularization parameter. This experiment will tell us whether there exists an optimal one or not. Figure 6 presents the case of CEL using the same model as in the previous section, and 500 iterations. We clearly see that the model cannot be estimated without regularization. The RMSE for EL ($\alpha = 0$) is much higher than the RMSE of CGEL with $\alpha = 0.2$, say. Unfortunately, we do not see a shape that would suggest us to look for one optimal α. We even have a few spikes for $\alpha > 0.6$.

The above result suggests that a fixed α for a given DGP cannot be the solution. We need a data driven method capable of detecting bad samples. Carrasco and Kotchoni (2017) use a bootstrapping method to estimate the RMSE, and they apply it to CGMM. The method is as follows. Let $\hat{\delta}_0(\alpha)$ be the estimated value of δ using the original sample for a given α. Also, let $\hat{\delta}_j(\alpha)$ be the estimated δ from the j^{th} resample for a given α. Then, the estimated MSE is defined as:

$$MSE(\alpha) = \frac{1}{B}\sum_{j=1}^{B}[\hat{\delta}_j(\alpha) - \hat{\delta}_0(\alpha)]^2,$$

where B is the number of resamples. Figure 7 plots $MSE(\alpha)$ for $B = 200$, using the samples from Figures 1c and 1a. The method do find an optimal α (=1.9) for the bad sample, but fails to select one for the good sample because the function in that case is mostly decreasing. We cannot therefore rely on that bootstrapping method because it would over-regularized good samples. We want a selection method that sets α to zero when regularization is not necessary, so that the properties of GEL are preserved. We have to notice also that bootstrapping methods are not valid when moments don't exist, which is a problem when α is too small.

Since we only want to stabilize the solution, we can check the stability by using a leave-one-out cross validation method (CV). Let $\hat{\delta}_i(\alpha)$ be the estimated δ obtained by removing the i^{th} observation. The cross validation criterion is define as:

$$CV(\alpha) = \frac{1}{N}\sum_{k=1}^{N}[y_{i(k)} - \hat{\delta}_{i(k)}(\alpha)x_{i(k)}]^2,$$

where $i(k)$ is a subsequence that is chosen randomly, and N is the number of observations used to compute the criterion. This is a general way of defining $CV(\alpha)$. Since it may be computationally intensive to compute $CV(\alpha)$ for several α, it may be reasonable to use a subset of the sample. Of course, we don't want to arbitrarily choose the subsample, so a random choice is preferable. $CV(\alpha)$ measures the ability of the model to predict out of sample observations, but it also measures the stability of the solution. Figure 8 shows $CV(\alpha)$ for the same good and bad samples used above for the bootstrap method. For both cases, we chose $N = 50$, which represents 25% of

the sample. We tried larger values for N, and the results were similar. For the bad sample, $CV(\alpha)$ stabilizes around $\alpha = 0.6$. From $\alpha \geq 0.6$, the benefit of increasing α in terms of stability becomes relatively small: from 0.5 to 0.6, $CV(\alpha)$ decreases by 6.77 and from 0.6 to 0.7 it decreases by 0.0016. The variations are similar if we keep increasing α. For the good sample, $CV(\alpha)$ is upward sloping, but quite stable since $CV(1) - CV(0) \approx 0.0043$. Therefore, based of the stability of the solution, that sample does not require regularization.

The above results suggest the following procedure: We first determine a sequence of α, $\{\alpha_0, \alpha_1, \alpha_2, ...\}$, with $\alpha_0 = 0$. Then, we sequentially compute $CV(\alpha_l)$ for $l = 0, 1, 2, ...$, and stop whenever $|CV(\alpha_l) - CV(\alpha_{l-1})| < \xi$, for some tolerance level ξ. We finally set the regularization parameter α to α_{l-1}. Alternatively, we may choose a stopping rule that is independent of the scale of $CV(\alpha)$. For example, we could stop whenever, $|CV(\alpha_l) - CV(\alpha_{l-1})| < \xi CV(\alpha_{l-1})$. Of course, we don't want ξ to be too small. The objective is not to have a constant criterion. We just want to stop whenever $CV(\alpha)$ becomes relatively stable. We applied the procedure to the same model used to produce Table 2 above ($n = 200$, $R^2 = 0.002$, and 20 instruments) and also to a model with R^2 equals to 0.1 to see if the need for regularization decreases with the strenght of the instruments. For each model, we applied the procedure with $N = 50$, a sequence $\{0, 0.1, 0.2, ...\}$ for α, a stopping rule $|CV(\alpha_l) - CV(\alpha_{l-1})| < 0.1$, and 1,000 iterations. Table 3 compares the properties of three different CGEL's, and Figure 9 plots the distribution of the selected α's for CEL.

Table 3: *Properties of CGEL with automatic selection of α (The last column is the proportion of $\alpha = 0$)*

	Means Bias	Median Bias	RMSE	S-dev	Interquartile Range	Not regul.
GMM ($R^2 = 0.002$)	0.4922	0.4910	0.5384	0.2185	0.2829	
CEL ($R^2 = 0.002$)	0.4756	0.4717	0.7132	0.5317	0.5273	0.4384
CET ($R^2 = 0.002$)	0.4599	0.4701	0.7667	0.6137	0.5568	0.4224
CEEL ($R^2 = 0.002$)	0.5253	0.4707	1.4588	1.3616	0.6071	0.5265
GMM ($R^2 = 0.1$)	0.1310	0.1354	0.1822	0.1267	0.1726	
CEL ($R^2 = 0.1$)	0.0113	0.0223	0.1867	0.1865	0.2438	0.8130
CET ($R^2 = 0.1$)	0.0108	0.0198	0.1806	0.1804	0.2462	0.8150
CEEL ($R^2 = 0.1$)	-0.0081	0.0118	0.2260	0.2260	0.2716	0.9100

The last column presents the proportion of the samples that do not need to be regularized. In general, the proportions are similar between the different methods. Also, the number of samples that need regularization is reduced substantially as R^2 increases; only 10% are regularized when $R^2 = 0.1$, while about half need regularization with $R^2 = 0.002$. Furthermore, the value of α for the regularized samples are on average smaller when R^2 is higher. We should therefore expect the properties of CGEL to be similar to GEL as R^2 increases.

In terms of the properties of the estimators, CGEL with automatic selection of

α does not produce extreme values anymore. Also, the bias is smaller than GMM estimators as it is predicted by Newey and Smith (2004). For a very low R^2, however, CGEL falls behind GMM in terms of RMSE, but as R^2 increases to a more reasonable value, GMM and CGEL are comparable in terms of RMSE and all CGEL estimators are less biased.

5 CGEL versus RCUE

In this section, we want to compare CGEL with the regularized CUE (RCUE) method proposed by Hausman et al. (2011). They show that the no moment problem of CUE can be solved by regularizing both the weighting matrix and the vector of coefficients. They do not propose procedures for selecting the two regularization parameters, but they show by mean of Monte Carlo experiments that there exist values that solve the no moment problem. In order to compare the properties of CGEL with those of RCUE, we consider simulations based on the same data generating processes used by the authors (HLMN) [4], which includes the possibility of having heteroskedastic errors. The authors want to show that RCUE is still valid in presence of heteroskedasticity, as opposed to LIML, when the number of instruments becomes large. The model is:

$$y_i = x_i\beta + \varepsilon_i, \tag{10}$$

$$x_i = z_{1i}\pi + \nu_i, \tag{11}$$

and

$$\varepsilon_i = \rho\nu_i + \sqrt{1-\rho^2}\left(\phi\theta_{1i} + \theta_{2i}\sqrt{1-\phi^2}\right), \tag{12}$$

where $\nu_i \sim N(0,1)$, $\theta_{1i} \sim N(0, z_{1i}^2)$, $z_{1t} \sim N(0,1)$, and $\theta_{2i} \sim N(0,1)$. The set of K instruments is: $Z_i = \{1, z_{1i}, z_{1i}^2, z_{1i}^3, z_{1i}^4, D_{1i}z_{1i}, ..., D_{(K-5)i}z_{1i}\}'$, where D_{li} has a Bernoulli distribution with probability of success equals to 0.5. The endogeneity is controlled by ρ, the heteroskedasticity by ϕ, and the strength of the instruments by π. If we define the concentration parameter as μ^2, then π is related to it by the following expression: $\pi = \mu/\sqrt{n}$. A part from the presence of heteroskedastic errors, the model is different from the one analyzed in the previous section in the sense that in the latter, all instruments are equally weak, but also equally important. In the HLMN model, z_{1i} is the optimal instrument, and the other instruments are just functions of z_{1i}. Therefore, it includes cases in which some instruments are irrelevant. We know from the second order asymptotic properties of GMM and GEL derived by Newey and Smith (2004) that adding irrelevant instruments is likely to increase the bias of GMM estimators more than the bias of GEL estimators.

[4]The model was introduced by Hausman et al. (2007). Therefore, they provide a more detailed explanation of its specification.

First, we consider a fixed α so that we can compare the results with the RCUE of Hausman et al. (2011). Table 4 shows the properties of GMM, CUE, CEL, CEEL, B2SLS, and J2SLS, for $\alpha = 0.9$, $n = 400$, $\phi = 0$, $\mu^2 = 8$, and $\rho = 0.3$, using 1,000 replications. It is therefore a case of homoscedasticity with both weak identification, because of the small concentration parameter μ^2, and weak endogeneity.

Table 4: *Simulation based on the HLMN model:* $\alpha = 0.9$*,* $n = 400$*,* $\mu^2 = 8$*,* $K = 10$*,* $\phi = 0$*, and* $\rho = 0.3$

	Mean-Bias	Median-Bias	RMSE	S-dev	Interquartile range
2Step	0.1660	0.1652	0.3163	0.2694	0.3375
CUE	0.2070	0.0262	7.2247	7.2254	0.7983
CEEL	0.0737	0.0671	0.4480	0.4422	0.4629
CEL	0.0810	0.0843	0.4774	0.4707	0.4858
B2SLS	-0.1292	0.0684	6.8403	6.8425	0.6553
J2SLS	0.0876	0.1098	0.4682	0.4601	0.4780

Clearly, CUE and B2SLS should be avoided[5]. As for the previous model, the dispersion of the estimates is extremely high. CEEL is less biased and less volatile than both CEL and J2SLS, the latter being slighly better than CEL in terms of the RMSE. If we compare our results to the ones obtained by Hausman et al. (2011) (Table 4) for RCUE, the mean bias of CEEL and CEL is very close to RCUE_1 to RCUE_4[6]. In terms of the variance, CEEL and CEL are comparable to their best RCUE_i, but they dominate all RCUE's in terms of interquartile range. We have to notice, however, that the relative performance of any CGEL and RCUE depends on the choice of the regularization parameters. Table 4 only shows that CGEL can solve the no moment problem of CUE or EL as well as RCUE does.

In Table 5, we present the results when the number of instruments increases to 50. In that case, both CEEL and CEL outperform all other methods, including GMM, in terms of the bias and the RMSE. This is consistent with the theoretical results of Newey and Smith (2004); because the bias of GMM estimators increases with the number of instruments, the RMSE eventually falls behind GEL. To our knowledge, it is the first time this property is shown numerically. Compare to RCUE, CEEL and CEL have similar biases, but their variance and interquartile range are much smaller. In fact, CEEL is six time less volatile that the least volatile RCUE[7].

[5]We actually under-estimated the standard deviation of the CUE estimator because the Brent algorithm used to obtain the estimate has hit the boundaries ± 100 in several occasions

[6]The authors report a mean bias for CUE in the order of 10^{11}. It is much lower in our case because we have restricted the search interval.

[7]The smallest standard error among the four RCUE's is 0.52, and the smallest interquartile range is 0.774

Table 5: *Simulation based on the HLMN model:* $\alpha = 0.9$, $n = 400$, $\mu^2 = 8$, $K = 50$, $\phi = 0$, *and* $\rho = 0.3$

	Mean-Bias	Median-Bias	RMSE	S-dev	Interquartile range
2Step	0.2574	0.2597	0.2956	0.1455	0.1875
CUE	-0.2158	0.0988	13.6473	13.6524	2.1673
CEEL	0.1749	0.1685	0.2735	0.2104	0.2737
CEL	0.1505	0.1477	0.2846	0.2417	0.3092
B2SLS	0.1367	0.1534	4.6111	4.6113	0.9915
J2SLS	0.2252	0.2217	0.3382	0.2525	0.3137

Table 6: *Simulation based on the HLMN model:* $\alpha \in \{0.9, 0.4, 0.2, 0.02\}$, $n = 400$, $\mu^2 = 8$, $K = 50$, $\phi = 0.88$, *and* $\rho = 0.3$

	Mean-Bias	Median-Bias	RMSE	S-dev	Interquartile range
2Step	0.2580	0.2613	0.3159	0.1823	0.2277
CUE	-0.6577	-0.1062	19.6724	19.6713	4.2870
CEEL($\alpha = 0.9$)	0.1859	0.1837	0.3280	0.2704	0.3545
CEEL($\alpha = 0.4$)	0.1994	0.1986	0.3351	0.2695	0.3425
CEEL($\alpha = 0.2$)	0.2007	0.2085	0.3760	0.3181	0.3466
CEEL($\alpha = 0.02$)	0.1285	0.2146	1.2246	1.2185	0.5757
CEL($\alpha = 0.9$)	0.1555	0.1623	0.3541	0.3183	0.3905
CEL($\alpha = 0.4$)	0.1704	0.1743	0.3454	0.3006	0.3733
CEL($\alpha = 0.2$)	0.1815	0.1804	0.3445	0.2930	0.3573
CEL($\alpha = 0.02$)	0.1358	0.1955	0.9931	0.9844	0.4461
B2SLS	-0.1723	0.1394	10.4527	10.4565	1.3254
J2SLS	0.2127	0.2153	0.4387	0.3839	0.4701

The heteroscedasticity case ($\phi = 0.88$) is presented in Table 6 with different choices for α. Notice that both CEL and CEEL become unstable when α is too small, a result that can be avoided by using our data driven selection method. If we ignore the unstable case $\alpha = 0.02$, we can see that the bias of CEL and CEEL increases as α decreases. The effect of α on the variance is, however, less monotonic. For CEEL, the variance decreases when α goes from 0.9 to 0.4 and gets larger when α keeps decreasing. For CEL, the variance and the RMSE decrease when α goes from 0.9 to 0.2. Intuitively, we can think of a small α as being equivalent to having a large number of instruments, which results in a higher bias and smaller variance. Because the effect of the number of instruments on the bias of the EL estimator is smaller, we observe a negative relationship between α and RMSE as long as the solution remains stable. Overall for those selected α's, when heteroscedasticity is present CEL and CEEL dominate all methods in terms of the bias (we do not consider B2SLS here because of its instability), but GMM dominates in

terms of the RMSE.

In order to analyze the properties of CGEL when α is based on our data driven method, and compare them with the RCUE method of Hausman et al. (2011), we follow the authors and consider the three different concentration parameters $\mu^2 = 8, 16, 32$, the four different number of instruments $K = 5, 10, 30, 50$, $\phi = 0$ and 0.88, $\rho = 0.3$, and a sample size of 400. To select α, we set N to 50, the tolerance level to 0.05, and the step to 0.1. Finally, the number of iteration is set to 1,000. Tables 7 to 10 present different properties for all DGP's using CET, CEL, CEEL and GMM. To facilitate the comparison between our method the RCUE, we report the MSE and variance instead of the RMSE and standard deviation. We do not report the median bias because it is very similar to the mean bias for all DPG's

First, we find that CGEL with automatic selection of α solves the no moment problem. CEL is the least biased of all four methods considered, but GMM has the smallest standard error for all DGP's. In terms of the MSE, however, CEL outperforms all other methods including GMM when K is above 30 and the error is homoscedastic. For DPG's with heteroscedasticity, GMM remains the method with the smallest MSE but CEL follows closely behind. In all cases, CEL is the least biased estimator. Compare to the $RCUE_i$ and $RCUE2_i$ of Hausman et al. (2011), all CGEL have a smaller variance, MSE and interquartile range when $K > 10$, but are often more biased. Those CGEL properties, however, are likely to be affected by the stopping rule used for the selection of α. For example, reducing the tolerance level for a given step size will result in a higher α on average. Table 11 shows some summary statistics of the selected α using the stopping rule described above. We can see that the proportion of samples that need regularization increases with the number of instruments, but the maximum α is higher when K is small. Therefore, adding many irrelevant instruments seems to improve the shape of the objective function when the relevant ones are weak. It is, however, less likely to get badly-shaped objective functions when K is small as the proportion of regularized samples is much smaller with few instruments. Finally, the need for reglarization is reduced when the relevant instruments become stronger.

6 conclusion

We proposed a regularized version of GEL based on the CGEL of Chaussé (2011), for which the regularization parameter, α, is selected on the basis of the stability of a cross validation criterion. In the procedure, we gradually increase α until the criterion stabilizes. As a result, the estimator of CGEL stays as close as possible to GEL. Such proximity is important because theoretical results, such as the ones demonstrated by Newey and Smith (2004), suggest that GEL should outperform GMM.

We first investigated the properties of CGEL using the DGP used by Guggenberger (2008), and showed that the regularization approach solves the no moment problem

raised by the author. We find that the CGEL estimators are less biased than GMM in all cases, and have a smaller RMSE when the instruments are not too weak ($R^2 > 0.002$).

In another simulation experiment, we compare CGEL with the regularized CUE of Hausman et al. (2011). We find that CGEL estimators can outperform all other methods when heteroscedasticity is not present and the number of instruments exceeds 20. In presence of heteroscedasticity, CGEL is less volatile than RCUE and GMM, but more biased then RCUE. In terms of the MSE, CGEL is comparable to GMM when the number of instruments exceeds 30.

We have shown that our regularized GEL does not suffer from the no moment problem and can be a good alternative to GMM in the case of many instruments. Since we are addressing a small sample issue, however, inference should not be based on the asymptotic theory derived by Chaussé (2011) for CGEL, because it relies on the convergence of α to zero as the sample size increases. Instead, inference should be based on fixed α using some kind of sandwich matrix. Alternatively, a bootstrap method should generate reliable standard errors because we have shown that CGEL estimators have finite moments. We leave these questions to future research.

Table 7: *Mean Bias for different methods and model specifications*

			GMM	CEL	CET	CEEL
$\phi = 0$	$\mu^2 = 8$	K= 5	0.0810	0.0155	0.0053	0.0116
		K= 10	0.1700	0.1282	0.1124	0.1235
		K= 30	0.2342	0.1668	0.1621	0.1874
		K= 50	0.2614	0.2000	0.2131	0.2294
	$\mu^2 = 16$	K= 5	0.0418	-0.0030	-0.0146	0.0008
		K= 10	0.1124	0.0623	0.0426	0.0662
		K= 30	0.1930	0.1192	0.1181	0.1380
		K= 50	0.2292	0.1492	0.1493	0.1800
	$\mu^2 = 32$	K= 5	0.0226	-0.0016	-0.0461	-0.0004
		K= 10	0.0659	0.0263	-0.0505	0.0275
		K= 30	0.1425	0.0745	0.0345	0.0946
		K= 50	0.1835	0.0994	0.0808	0.1309
$\phi = 0.88$	$\mu^2 = 8$	K= 5	0.0785	0.0074	0.0236	-0.0003
		K= 10	0.1360	0.0678	0.0745	0.0681
		K= 30	0.2407	0.1736	0.1814	0.1872
		K= 50	0.2639	0.2083	0.2108	0.2167
	$\mu^2 = 16$	K= 5	0.0399	-0.0015	-0.0074	-0.0113
		K= 10	0.0828	0.0293	0.0379	0.0266
		K= 30	0.1957	0.1204	0.1248	0.1360
		K= 50	0.2324	0.1551	0.1622	0.1751
	$\mu^2 = 32$	K= 5	0.0227	-0.0072	-0.0028	-0.0069
		K= 10	0.0437	0.0064	0.0049	0.0000
		K= 30	0.1419	0.0715	0.0732	0.0755
		K= 50	0.1873	0.1020	0.1118	0.1287

Table 8: *MSE for different methods and model specifications*

			GMM	CEL	CET	CEEL
$\phi = 0$	$\mu^2 = 8$	K= 5	0.1383	0.2946	0.2892	0.2575
		K= 10	0.1074	0.1917	0.1945	0.1740
		K= 30	0.0869	0.0970	0.1288	0.0962
		K= 50	0.0891	0.0844	0.0938	0.0949
	$\mu^2 = 16$	K= 5	0.0717	0.1171	0.1301	0.1018
		K= 10	0.0629	0.0921	0.1355	0.0870
		K= 30	0.0633	0.0677	0.0866	0.0671
		K= 50	0.0705	0.0580	0.0901	0.0693
	$\mu^2 = 32$	K= 5	0.0358	0.0481	0.1235	0.0458
		K= 10	0.0331	0.0451	0.1801	0.0411
		K= 30	0.0396	0.0400	0.1384	0.0430
		K= 50	0.0479	0.0373	0.1156	0.0475
$\phi = 0.88$	$\mu^2 = 8$	K= 5	0.2887	0.6940	0.5785	0.4959
		K= 10	0.1946	0.4518	0.3700	0.3409
		K= 30	0.1153	0.1820	0.1643	0.1557
		K= 50	0.1040	0.1403	0.1258	0.1345
	$\mu^2 = 16$	K= 5	0.1255	0.2699	0.2278	0.1976
		K= 10	0.1068	0.2237	0.1808	0.1675
		K= 30	0.0849	0.1237	0.1076	0.1098
		K= 50	0.0848	0.0973	0.1052	0.1035
	$\mu^2 = 32$	K= 5	0.0557	0.0913	0.0725	0.0670
		K= 10	0.0546	0.0852	0.0844	0.0708
		K= 30	0.0540	0.0694	0.0650	0.0627
		K= 50	0.0605	0.0619	0.0739	0.0683

Table 9: *Variance for different methods and model specifications*

			GMM	CEL	CET	CEEL
$\phi = 0$	$\mu^2 = 8$	K= 5	0.1319	0.2946	0.2894	0.2577
		K= 10	0.0786	0.1754	0.1820	0.1589
		K= 30	0.0321	0.0693	0.1027	0.0612
		K= 50	0.0208	0.0444	0.0485	0.0423
	$\mu^2 = 16$	K= 5	0.0700	0.1172	0.1300	0.1019
		K= 10	0.0503	0.0883	0.1338	0.0827
		K= 30	0.0261	0.0535	0.0727	0.0481
		K= 50	0.0180	0.0358	0.0679	0.0369
	$\mu^2 = 32$	K= 5	0.0353	0.0482	0.1215	0.0459
		K= 10	0.0288	0.0444	0.1777	0.0404
		K= 30	0.0193	0.0344	0.1373	0.0341
		K= 50	0.0143	0.0275	0.1092	0.0304
$\phi = 0.88$	$\mu^2 = 8$	K= 5	0.2829	0.6947	0.5785	0.4964
		K= 10	0.1763	0.4477	0.3648	0.3366
		K= 30	0.0574	0.1520	0.1315	0.1208
		K= 50	0.0344	0.0970	0.0814	0.0876
	$\mu^2 = 16$	K= 5	0.1240	0.2702	0.2280	0.1976
		K= 10	0.1001	0.2231	0.1796	0.1670
		K= 30	0.0466	0.1093	0.0922	0.0914
		K= 50	0.0309	0.0733	0.0790	0.0729
	$\mu^2 = 32$	K= 5	0.0553	0.0914	0.0726	0.0670
		K= 10	0.0527	0.0853	0.0844	0.0709
		K= 30	0.0339	0.0643	0.0597	0.0571
		K= 50	0.0254	0.0516	0.0615	0.0518

Table 10: *Interquartile range for different methods and model specifications*

			GMM	CEL	CET	CEEL
$\phi = 0$	$\mu^2 = 8$	K= 5	0.4106	0.5380	0.5238	0.5034
		K= 10	0.3626	0.4830	0.4745	0.4638
		K= 30	0.2401	0.3362	0.3448	0.3358
		K= 50	0.1968	0.2802	0.2656	0.2875
	$\mu^2 = 16$	K= 5	0.3174	0.3808	0.3688	0.3698
		K= 10	0.2909	0.3605	0.3635	0.3616
		K= 30	0.2136	0.2954	0.2963	0.2921
		K= 50	0.1881	0.2521	0.2400	0.2697
	$\mu^2 = 32$	K= 5	0.2430	0.2694	0.2811	0.2705
		K= 10	0.2283	0.2598	0.2731	0.2629
		K= 30	0.1919	0.2276	0.2333	0.2371
		K= 50	0.1666	0.2101	0.2086	0.2311
$\phi = 0.88$	$\mu^2 = 8$	K= 5	0.5723	0.7571	0.6817	0.6679
		K= 10	0.5172	0.7235	0.6857	0.6792
		K= 30	0.3146	0.4716	0.4627	0.4560
		K= 50	0.2537	0.3849	0.3595	0.3978
	$\mu^2 = 16$	K= 5	0.4055	0.5106	0.4725	0.4500
		K= 10	0.4082	0.5542	0.5278	0.5021
		K= 30	0.2892	0.4198	0.4033	0.4004
		K= 50	0.2422	0.3422	0.3287	0.3578
	$\mu^2 = 32$	K= 5	0.2838	0.3424	0.3182	0.3052
		K= 10	0.3093	0.3885	0.3585	0.3451
		K= 30	0.2453	0.3141	0.2929	0.3011
		K= 50	0.2204	0.2878	0.2616	0.2966

Table 11: *Summary statistics of the regularization parameter for different methods and model specifications*

			CEL			CET			CEEL		
			Prop.	Median	Max	Prop.	Median	Max	Prop.	Median	Max
$\phi = 0$	$\mu^2 = 8$	K= 5	0.3230	0.0000	14524.2000	0.3170	0.0000	16613.8500	0.2680	0.0000	18813.8500
		K= 10	0.5660	0.1000	13.5000	0.5810	0.1000	4.2000	0.5810	0.1000	6.5000
		K= 30	0.7780	0.1000	1.0000	0.8260	0.1000	2.1000	0.7850	0.1000	0.7000
		K= 50	0.7970	0.1000	0.6000	0.8720	0.1000	2.1000	0.8330	0.1000	0.4000
	$\mu^2 = 16$	K= 5	0.2010	0.0000	11113.8500	0.2150	0.0000	14413.8500	0.1690	0.0000	17713.8500
		K= 10	0.4330	0.0000	1.7000	0.4670	0.0000	2.1000	0.4270	0.0000	1.4000
		K= 30	0.6680	0.1000	0.7000	0.7700	0.1000	2.1000	0.6900	0.1000	0.7000
		K= 50	0.7650	0.1000	0.4000	0.8580	0.1000	2.1000	0.7750	0.1000	0.7000
	$\mu^2 = 32$	K= 5	0.0980	0.0000	0.4000	0.1330	0.0000	2.1000	0.0850	0.0000	0.4000
		K= 10	0.2760	0.0000	0.6000	0.3580	0.0000	2.1000	0.2770	0.0000	0.4000
		K= 30	0.5260	0.1000	0.3000	0.7180	0.1000	2.1000	0.5770	0.1000	0.4000
		K= 50	0.6500	0.1000	0.3000	0.8150	0.1000	2.1000	0.6870	0.1000	0.4000
$\phi = 0.88$	$\mu^2 = 8$	K= 5	0.4130	0.0000	6838655.2500	0.3770	0.0000	10546404103.5000	0.2930	0.0000	12537765.6000
		K= 10	0.7310	0.1000	20103.5000	0.7030	0.1000	25.6500	0.6320	0.1000	72.8500
		K= 30	0.9040	0.1000	1.7000	0.8910	0.1000	1.1000	0.8050	0.1000	1.5000
		K= 50	0.9390	0.1000	0.8000	0.9440	0.1000	1.9000	0.8550	0.1000	0.8000
	$\mu^2 = 16$	K= 5	0.3040	0.0000	13324.2000	0.2760	0.0000	3.0000	0.1850	0.0000	1.5000
		K= 10	0.6080	0.1000	2.1000	0.5720	0.1000	10.8500	0.5020	0.1000	7.7500
		K= 30	0.8700	0.1000	1.7000	0.8560	0.1000	1.3000	0.7350	0.1000	1.2000
		K= 50	0.9150	0.1000	1.3000	0.9350	0.1000	2.1000	0.8180	0.1000	0.6000
	$\mu^2 = 32$	K= 5	0.1880	0.0000	1.0000	0.1530	0.0000	1.4000	0.0850	0.0000	0.4000
		K= 10	0.4380	0.0000	0.8000	0.3990	0.0000	2.1000	0.3520	0.0000	1.0000
		K= 30	0.7550	0.1000	0.8000	0.7740	0.1000	1.3000	0.6180	0.1000	0.9000
		K= 50	0.8660	0.1000	0.4000	0.9120	0.1000	2.1000	0.7520	0.1000	0.7000

References

B. Antoine, H. Bonnal, and E. Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and euclidean empirical likelihood. *Journal od Econometrics*, 138:461–487, 2007.

M. Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–358, 2012.

M. Carrasco and J.P. Florens. Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16:655–673, 2000.

M. Carrasco and R. Kotchoni. Efficient estimation using the characteristic function. *Econometric Theory*, 33(2):479–526, 2017.

M. Carrasco and G. Tchuente. Regularized liml with many instruments. *Journal of Econometrics*, 186(2):427–442, 2015.

P. Chaussé. Computing generalized method of moments and generalized empirical likelihood with r. *Journal of Statistical Software*, 34(11):1–35, 2010.

P. Chaussé. Generalized emprical likelihood for a continuum of moment conditions. *University of Waterloo Working Paper*, 2011.

S. Donald and W. Newey. Choosing the number of instruments. *Econometrica*, 69: 1161–1191, 2001.

S.G. Donald, G. Imbens, and W. Newey. Choosing the number of moments in conditional moment restrition models. *MIT working paper*, 2008.

P. Guggenberger. Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews*, 26:526–541, 2008.

J. Hahn, J. Hausman, and G. Kuersteiner. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *Econometrics Journal*, 7:272–306, 2004.

L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

L.P. Hansen, J. Heaton, and A. Yaron. Finit-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14:262–280, 1996.

J. Hausman, W. Newey, T. Woutersen, J. Chao, and N. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. *MIT working paper*, 2007.

J. Hausman, R. Lewis, K. Menzel, and W. Newey. Properties of cue estimator and a modification with moments. *Journal of Econometrics*, 165:45–57, 2011.

Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(5):861–874, 1997.

F. Kleibergen. Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73(4):1103–1123, 2005.

R.S. Mariano and T. Sawa. The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables. *Journal of the American Statistical Association*, 67(337):159–163, 1972.

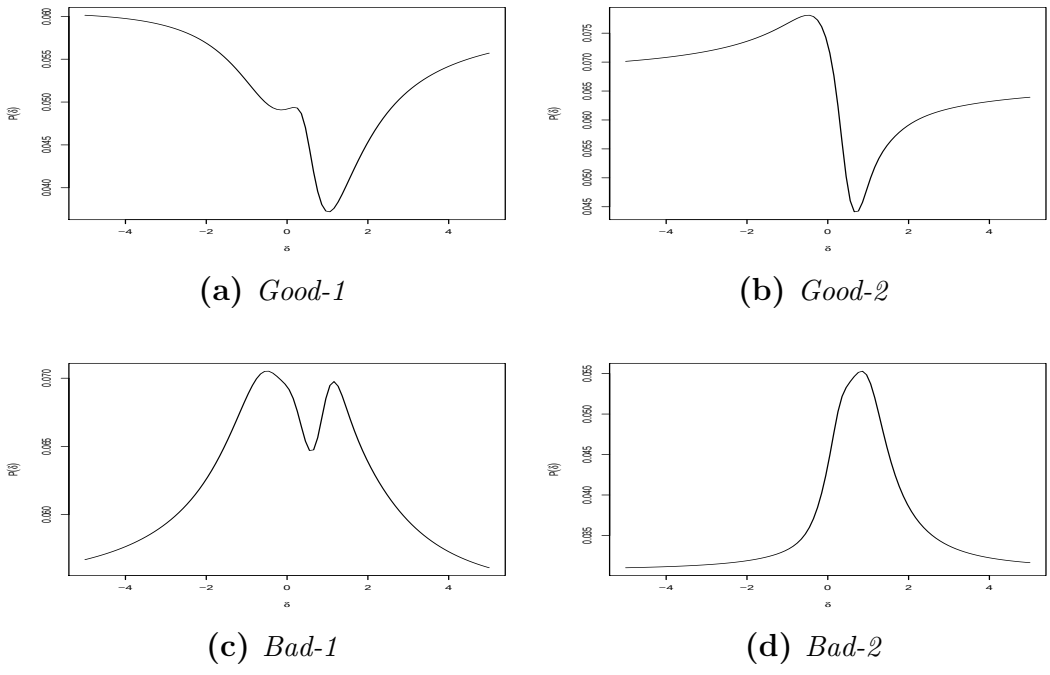
W.K. Newey and R.J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72:219–255, 2004.

A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

T. Sawa. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association*, 64(327):923–937, 1969.

R.J. Smith. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *The Economic Journal*, 107:503–519, 1997.

Figure 1: *EL objective function $P(\delta)$ for good and bad samples*



(a) *Good-1* **(b)** *Good-2*

(c) *Bad-1* **(d)** *Bad-2*

Figure 2: *The different $\lambda_i(\delta)$ for the Guggenberger model and EL objective function*

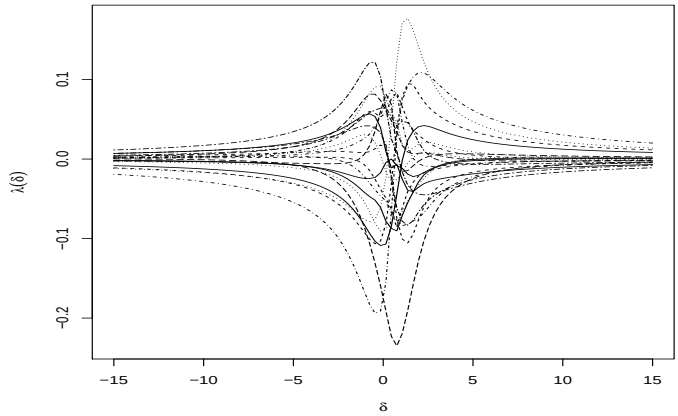
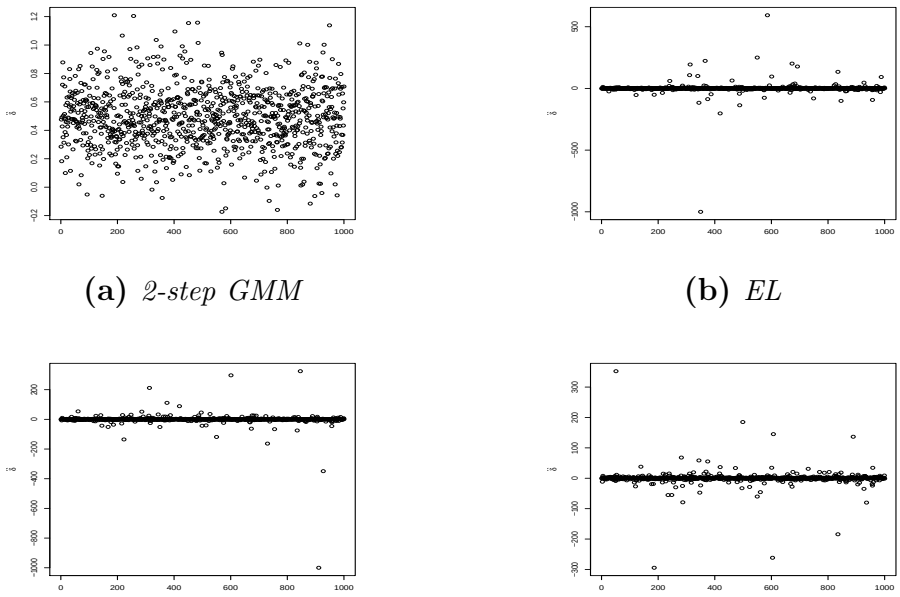


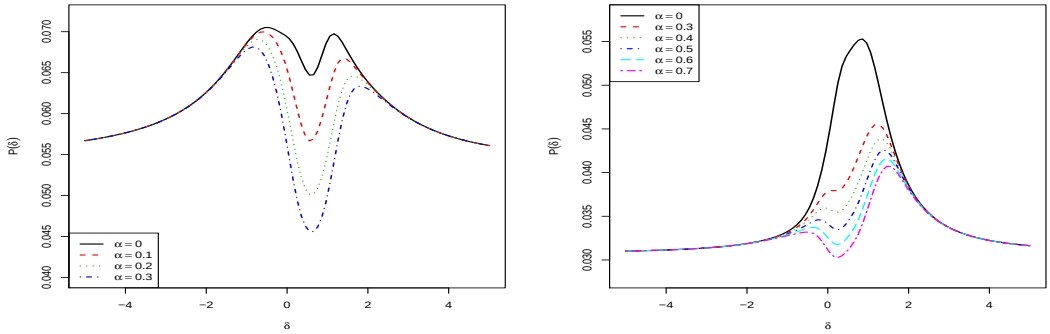
Figure 3: *Point estimates from the simulation of Table 1*



(a) *2-step GMM* **(b)** *EL*

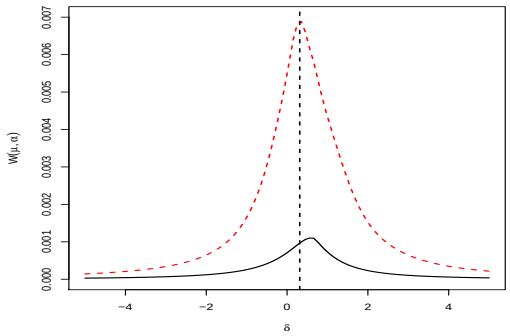
(c) *ET* **(d)** *CUE*

Figure 4: *The effect of increasing* α *on* $P(\delta;\alpha)$

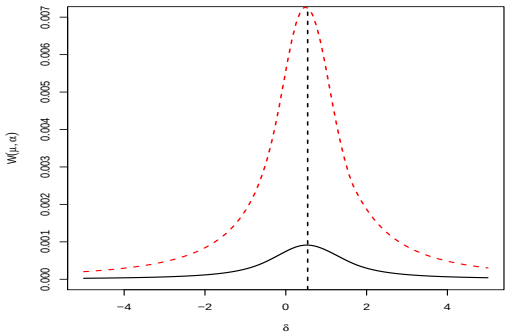


(a) *Bad-1* **(b)** *Bad-2*

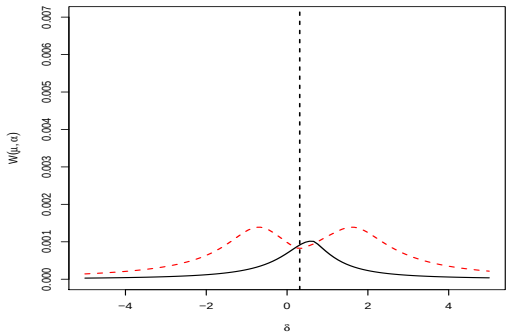
Figure 5: *Weights $W(\mu_i(\delta), \alpha)$ versus δ*
Bad-1 is from Figure 1c and Bad-2 from Figure 1d, the vertical line is the first-step estimate



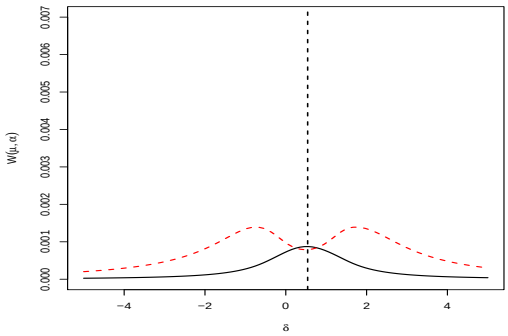
(a) *Bad-1,* $\alpha = 0$



(b) *Bad-2,* $\alpha = 0$



(c) *Bad-1,* $\alpha = 2$



(d) *Bad-2,* $\alpha = 2$

Figure 6: *RMSE as a function of a fixed* α. $R^2 = 0.002$, $n = 200$, $\rho = 0.5$, $k = 20$, *and 500 iterations*

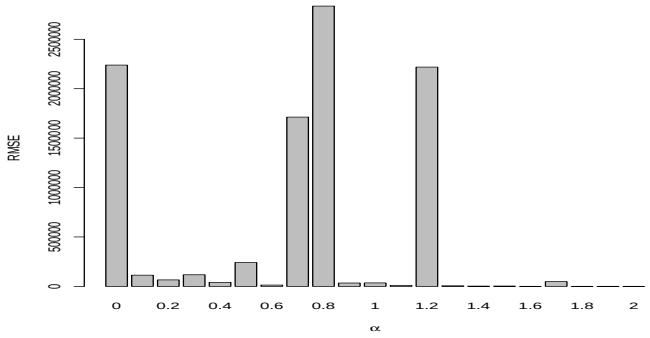
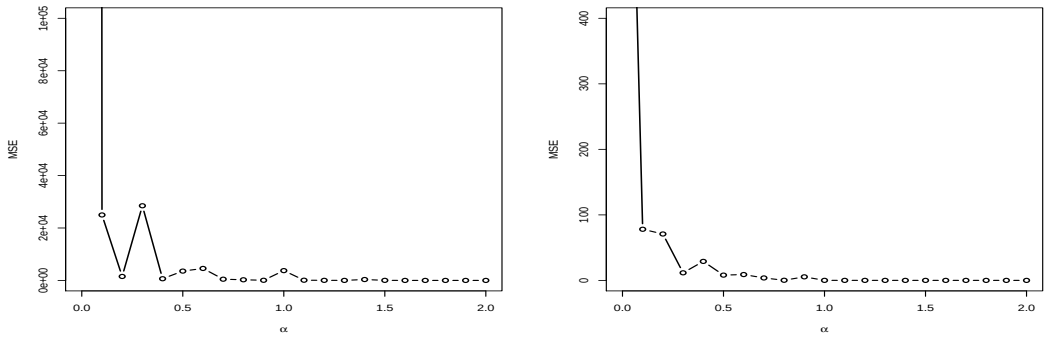
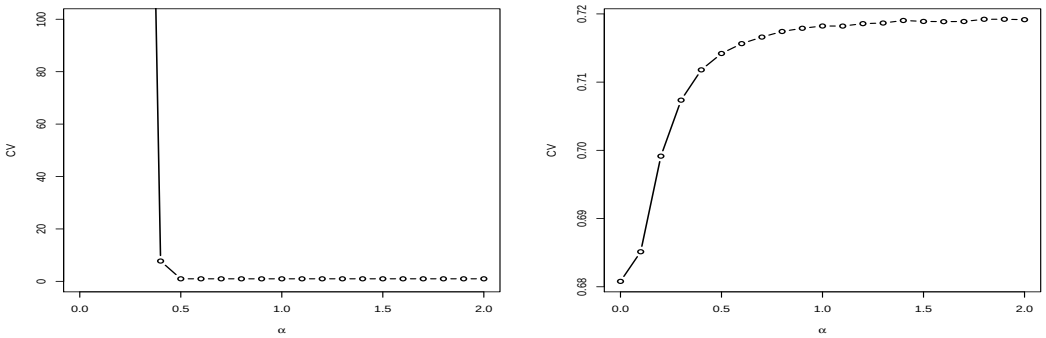


Figure 7: *Estimated MSE by Bootstrap*
Bad is from Figure 1c and Good from Figure 1a



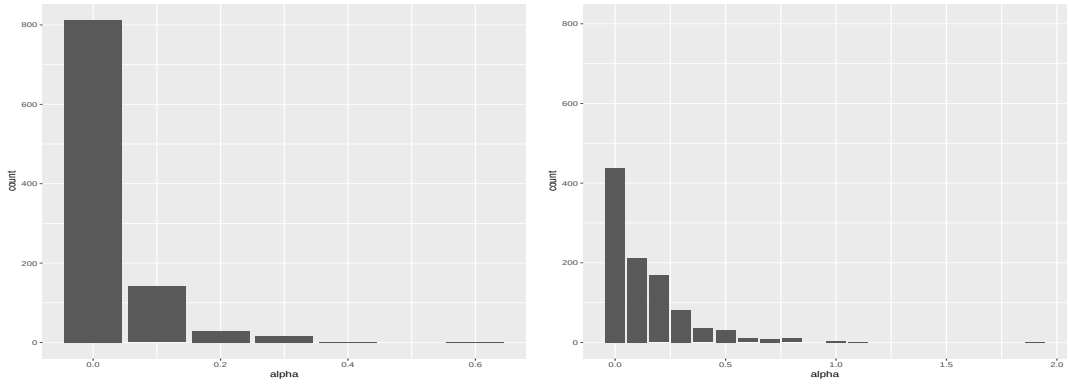
(a) *Bad* **(b)** *Good*

Figure 8: *Cross-validation with $N = 50$*
Bad is from Figure 1c and Good from Figure 1a



(a) *Bad ($CV(0) \approx 300,000$)* **(b)** *Good*

Figure 9: *Distribution of the selected α for CEL in a simulation with 1,000 iterations*



(a) $R^2 = 0.1$ **(b)** $R^2 = 0.002$