Testing the Coherence of Data and External Intervals via an Imprecise Sargan-Hansen Test

Martin Jann MARTIN.JANN@UNI-HAMBURG.DE
Department of Research Methods and Statistics, Institute of Psychology, University of Hamburg, Germany

Abstract

When information about a population is sparse, it is difficult to test whether a data set originated from that population. In applied research, however, researchers often have access to external information in the form of (central) statistical moments such as mean or variance. To compensate for the uncertainty in the external point values, this paper uses external intervals instead to represent the information about moments. The Sargan-Hansen test from the generalized method of moments framework is used, which exploits point-valued external information about moments in the presence of a statistical model to test whether data and external information are in conflict. For the Sargan-Hansen test, a separability result is derived with respect to the model and the external information. This result leads to a simplification of the test in terms of its analytical form and the calculation of the test statistics. To allow the use of external intervals instead of point values, an imprecise version of the Sargan-Hansen test is created using the Gamma-maximin decision rule. Assuming that the variables are normally distributed, a small sample version of this imprecise Sargan-Hansen test is derived. The power and type I errors of the developed tests are analyzed and compared in a simulation study in different small sample scenarios.

Keywords: imprecise external information, information-data conflict, generalized method of moments, Sargan-Hansen test, credal set, robustness

1. Introduction

The use of (external) prior information on parameters has frequently been studied. Well-known techniques for incorporating external information into statistical analysis include informed prior distributions in Bayesian statistics [3] and constraints on the parameter space imposed by the external information, leading to constrained optimization (see, e.g. Knopov and Korkhin [11] for the case of multiple linear regression). However, in some research areas, there may not be enough information to determine the feasible region or a prior distribution. The following example is provided to support this assertion:

Example 1 *Suppose we have a simple linear regression model* $y = \beta_1 + x\beta_2 + \epsilon$ *under Gauss-Markov assumptions and only the expected value* $E(y) = 100$ *is known externally. Under the model assumptions,* $E(y) = 100$ *becomes a constraint on the parameter,*

$$100 = E(y) = \beta_1 + E(x)\beta_2, \tag{1}$$

which is a linear constraint on intercept β_1 *and slope* β_2*. However, if* $E(x)$ *is not known, we cannot use Equation (1) directly as a constraint in the optimization. Equation (1) is also not sufficient to identify (the moments of) a prior distribution, since there are usually several different distributions that satisfy this condition.*

The fact that the external information in Example 1 is in the form of a moment motivates another method of using external information. According to an idea proposed by Imbens and Lancaster [9], this type of external information implies moment conditions that can be combined with the moment conditions used to estimate a statistical model. In general, the resulting overidentified system of moment conditions does not have an exact solution, but the Generalized Method of Moments (GMM) [7] can be used to find estimators that are 'as close as possible' to a solution with respect to some norm. Imbens and Lancaster [9] showed for multiple linear models that the estimators found in this way generally have lower variances than the corresponding OLS estimators, provided that the external information is correct. This paper examines the opposite question: Given the combined moment conditions of the model and the external information, is the external information correct (for a given data set)? This concept is similar to the prior-data conflict in Bayesian statistics and will be referred to hereafter as *information-data conflict*. To answer this question in the GMM framework, the Sargan-Hansen test is typically used because it is a test for overidentifying restrictions [18, 7].

However, its role as a test for misspecification has been criticized in current research, especially with respect to models that use instrumental variables [14, 10]. Therefore, the results of this paper should be interpreted as a test of the coherence of external information and data rather than a test of misspecification of a model. This argument is supported by a small thought experiment. There are

two statements: *"The model assumptions are true."* and *"The expected value of the dependent variable is 100."* Both statements are logically independent, one is neither necessary nor sufficient for the other to be true. How might a model specification test benefit from this kind of external information? A mathematical formulation of this logical independence is proved in Section 2.

Most external information depends on population, time, and many other aspects, which makes the use of point values for the external information risky because the results of Imbens and Lancaster [9] depend on the correctness of these point values. To reduce the risk of potentially misspecified external information, this paper addresses the case where an interval is given that contains the true value of the external moments, but its exact position inside the interval is unknown. This epistemic uncertainty about the true value of the external moments leads directly to the use of imprecise probabilities in the form of credal sets, as we show in Section 2.

2. The Sargan-Hansen Test with External Information

2.1. The Point-Valued Case

We assume that the external information only consists of point values of the respective moments. The notation from Newey and McFadden [12] is adopted. In the following, italic lowercase letters are for (random) scalar values, bold lowercase letters are for (random) vectors, and bold uppercase letters are for (random) matrices, unless otherwise indicated. Now let $\mathbf{z}$ be a random variable over $\mathbb{R}^k$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be $n > 1$ i.i.d. random variables distributed like $\mathbf{z}$.[1] Further let q be an integer and $\boldsymbol{\Theta} \subset \mathbb{R}^q$, then let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be a possible value for a (fixed) parameter of a statistical model, where $\boldsymbol{\theta}_0$ is the true value. Given a function $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ with the property $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$, one can try to estimate the parameter by the method of moments. Practically, this is done by formulating the equivalent sample moment conditions $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}$ and solving for $\boldsymbol{\theta}$.

To explain this method, let's consider Example 1. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be an i.i.d. sample of random variables distributed like y, and let

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,q-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,q-1} \end{pmatrix}$$

be the design matrix containing the covariates assumed to be an i.i.d. sample of the random variable $\mathbf{x}$. The sample moment conditions for the OLS estimator can be derived by setting the mixed moment of the independent variables and the error term to zero, i.e. $E(\mathbf{g}(\mathbf{z}, \boldsymbol{\beta}_0)) = E(\mathbf{x}\epsilon) = \mathbf{0}$. The sample moment conditions are therefore $\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ in this case denotes the parameter [4, p. 172].

However, sometimes the number of the moment conditions is larger than the dimension of the parameter. As a classic example from econometrics, we present estimation using instrumental variables, following the presentation of Cameron and Trivedi [4, p. 170]. As before, we assume a linear model. If some of the independent variables in $\mathbf{x}$ are correlated with the error term, then the Gauss-Markov assumptions are incorrect, and therefore OLS will not provide a consistent estimate of the regression parameter. A common idea to solve this problem is to find other variables that are correlated with $\mathbf{x}$ but uncorrelated with the error term. These variables are called instruments, and we represent their sample realizations by the $(n \times s)$ matrix $\mathbf{D}$. Similar to the OLS case, we can set the mixed moment of the instruments and the error term to zero. The corresponding sample moment conditions are $\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. If the number of potential instruments is greater than the dimension of the parameter, the sample moment conditions are generally not solvable for $\boldsymbol{\beta}$, the system of equations is overidentified. Not using all the instruments would result in a loss of efficiency. Instead of solving the equations, the idea of the GMM is to find a value for $\boldsymbol{\beta}$ that makes $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ as small as possible in terms of quadratic loss, i.e, by minimizing

$$(\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))^T \mathbf{W} (\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})),$$

where $\mathbf{W}$ is a chosen positive-definite weighting matrix. Note that this is a generalization of the case of solvable sample moment conditions, since a positive quadratic form in $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is zero if and only if $\frac{1}{n}\mathbf{D}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$. In general, different $\mathbf{W}$ lead to different estimators. In the GMM approach, there is a way to choose the optimal weighting matrix with respect to the efficiency of the estimator. This optimality is achieved by $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ with $\boldsymbol{\Omega} = E(\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})^T)$ [7]. This optimal $\mathbf{W}$ is almost always unknown and must be estimated by a random matrix $\hat{\mathbf{W}}$. Taken together, this leads to the following definition:

Definition 1 *[12, p. 2116] Let $p \geq q$ be an integer and $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{\theta})$ be a vector valued function with values in $\mathbb{R}^p$, that meets the moment conditions $E[\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{\theta}_0)] = \boldsymbol{0}$. Further let $\hat{\boldsymbol{W}} \in \mathbb{R}^{p,p}$ be a positive semi-definite (and hence symmetric) random matrix such that $(\boldsymbol{r}^T \hat{\boldsymbol{W}} \boldsymbol{r})^{1/2}$ is almost surely a norm for all $\boldsymbol{r} \in \mathbb{R}^p$. Then a **GMM-estimator** $\hat{\boldsymbol{\theta}}_{ex}$ is defined as a $\boldsymbol{\theta}$, that maximizes the following objective function:*

$$\hat{Q}_n(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(z_i, \boldsymbol{\theta}))^T \hat{\boldsymbol{W}} (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}(z_i, \boldsymbol{\theta})) \,. \qquad (2)$$

[1] Some entries of $\mathbf{z}$ could possibly be fixed, as long as at least one entry is random.

Under mild regularity conditions, the GMM-estimator is point-identified, consistent, and asymptotically normally distributed [12, Theorem 3.4]. To emphasize the generality of the GMM, we give some examples. Special cases of GMM-estimators range from OLS estimators to maximum likelihood estimators (MLE) [4, p. 172] to estimators derived by generalized estimating equations [4, p. 790]. To see that GMM is an extension of MLE, note that maximizing the log-likelihood function implies setting the score function to zero. This corresponds to the first-order conditions for MLE and has exactly the form of sampling moment conditions. In addition, the regularity conditions of the MLE require that the expected value of the score function be zero at the true parameter value, which is exactly the requirement $E[\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$ in Definition 1. This property of the score function is central to establishing the consistency and asymptotic normality of the MLE. For the mathematical details of incorporating the MLE into the GMM, see Cameron and Trivedi [4, p. 140]. Finally, there is also an important connection to robust statistics, since M-estimators with differentiable ρ (those of the $\psi-$type) are also derived by sample moment conditions and thus represent a special case of GMM estimators [4, p. 118].

Following Imbens and Lancaster [9], we include in $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ not only the moment conditions for the model, but also those for the external information, resulting in an overidentified system of moment conditions. Let $\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})$ denote the $p_1 \geq q$ moment conditions for the model and $\mathbf{h}(\mathbf{z})$ denote the p_2 moment conditions for the external information, which are assumed to be expressible as functions of the data alone, then $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta}) = (\mathbf{m}(\mathbf{z}, \boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T)^T$. For example, the condition for the external moment in Example 1 is $h(\mathbf{z}) = y - 100$. If one of the moment conditions for the external information depends only on the parameter, then the results derived here will not hold in general.

For the overidentified case $p > q$, under the null hypothesis that all moment conditions are correct, it holds that $-n\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) \xrightarrow{d} \chi^2_{p-q}$ if the regularity conditions hold and if $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W} = \boldsymbol{\Omega}^{-1}$. The χ^2-test that results from this distribution property is called the Sargan-Hansen test [18, 7]. For simplicity, in the remainder of this paper we assume that $\hat{\mathbf{W}}$ is invertible almost surely and therefore positive-definite almost surely by Definition 1. All the following results are derived for this almost sure case of invertible $\hat{\mathbf{W}}$ and thus hold almost surely. If $\hat{\mathbf{W}}$ is singular for certain data, one should first check whether the moment conditions are linearly dependent, and accordingly delete some conditions, so that the remaining ones are not linearly dependent. Otherwise, one could add random noise to $\hat{\mathbf{W}}$ to try to make it invertible, or use its Moore-Penrose inverse [20].

Let $\hat{\boldsymbol{\Omega}}$ be the inverse of $\hat{\mathbf{W}}$. For the sake of brevity we define $\overline{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^n \mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})$ and $\overline{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)$. By $\mathbf{A}/\mathbf{B}$ we denote the Schur complement of the block $\mathbf{B}$ of the matrix $\mathbf{A}$ and obtain

Lemma 2 *(Separability) From the premises of Definition 1 and $\mathbf{g}(z, \boldsymbol{\theta}) = (\boldsymbol{m}(z, \boldsymbol{\theta})^T, \boldsymbol{h}(z)^T)^T$ it follows that $\hat{\boldsymbol{\Omega}}$ has the block form*

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} \hat{\boldsymbol{\Omega}}_m & \hat{\boldsymbol{\Omega}}_r^T \\ \hat{\boldsymbol{\Omega}}_r & \hat{\boldsymbol{\Omega}}_h \end{pmatrix},$$

where $\hat{\boldsymbol{\Omega}}_m \in \mathbb{R}^{p_1,p_1}$ and $\hat{\boldsymbol{\Omega}}_h \in \mathbb{R}^{p_2,p_2}$. Further,

$$-\hat{Q}_n(\boldsymbol{\theta}) = (\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}(\overline{\boldsymbol{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}}) + \overline{\boldsymbol{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\boldsymbol{h}}.$$

Proof We take advantage of the fact that $\hat{\mathbf{W}}$ is symmetric, positive-definite, and can be written in block form

$$\hat{\mathbf{W}} = \begin{pmatrix} \hat{\mathbf{W}}_m & \hat{\mathbf{W}}_r^T \\ \hat{\mathbf{W}}_r & \hat{\mathbf{W}}_h \end{pmatrix},$$

where $\hat{\mathbf{W}}_m \in \mathbb{R}^{p_1,p_1}$ and $\hat{\mathbf{W}}_h \in \mathbb{R}^{p_2,p_2}$. The first statement follows from the fact that $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^{-1}$ and the block form of $\hat{\mathbf{W}}$. For the second statement, note that $\hat{\mathbf{W}}$ is positive-definite and so is $\hat{\boldsymbol{\Omega}}$, so the Schur complement $\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Omega}}_m - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \hat{\boldsymbol{\Omega}}_r$ is invertible. Now $\hat{\mathbf{W}}$ can be expressed by Schur complements:

$$\begin{aligned}
\hat{\mathbf{W}}_m &= (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}, \\
\hat{\mathbf{W}}_r &= -\hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}, \\
\hat{\mathbf{W}}_h &= \hat{\boldsymbol{\Omega}}_h^{-1} + \hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}.
\end{aligned}$$

It follows that

$$\begin{aligned}
-\hat{Q}_n(\boldsymbol{\theta}) &= (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}))^T \hat{\mathbf{W}} (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})) \\
&= \overline{\mathbf{m}}^T \hat{\mathbf{W}}_m \overline{\mathbf{m}} + 2\overline{\mathbf{m}}^T \hat{\mathbf{W}}_r^T \overline{\mathbf{h}} + \overline{\mathbf{h}}^T \hat{\mathbf{W}}_h \overline{\mathbf{h}} \\
&= \overline{\mathbf{m}}^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\overline{\mathbf{m}} - 2\overline{\mathbf{m}}^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}} \\
&\quad + \overline{\mathbf{h}}^T (\hat{\boldsymbol{\Omega}}_h^{-1} + \hat{\boldsymbol{\Omega}}_h^{-1}\hat{\boldsymbol{\Omega}}_r(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}\hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1})\overline{\mathbf{h}} \\
&= (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T\hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}) \\
&\quad + \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}.
\end{aligned}$$

■

Lemma 2 can be interpreted as a separability result, since $\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}$ is not a function of $\boldsymbol{\theta}$ if a suitable $\hat{\boldsymbol{\Omega}}_h$ is used, e.g, $\hat{\boldsymbol{\Sigma}}_h = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T$ or the sample covariance matrix $\hat{\mathbf{S}}_h = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})^T$. In these cases, $\hat{\boldsymbol{\Omega}}_h$ can be calculated from the data and external information

alone. Note that the matrix $\hat{\mathbf{S}}_h$ can be computed even without knowing the true external value. Both matrices are asymptotically identical if the null hypothesis of correctly specified external values holds, but different if it does not. The following important result holds for these examples.

Theorem 3 *Let the premises and notation of Lemma 2 be given. If $\hat{\boldsymbol{\Omega}}_h$ is not a function of $\boldsymbol{\theta}$ and if there is a $\boldsymbol{\theta}_h \in \boldsymbol{\Theta}$, for which $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\boldsymbol{h}} = \mathbf{0}$ holds, it follows that*

$$-\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) = \overline{\boldsymbol{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\boldsymbol{h}}.$$

Proof By Definition 1 we get

$$-\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex}) = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{Q}_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} -\hat{Q}_n(\boldsymbol{\theta}).$$

For $\boldsymbol{\theta}_h$ given in the premises, it follows from Lemma 2, that $-\hat{Q}_n(\boldsymbol{\theta}_h) = \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$. Since $\hat{\boldsymbol{\Omega}}$ is positive-definite, $(\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1}$ is also positive-definite. Therefore, $(\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}})^T (\hat{\boldsymbol{\Omega}}/\hat{\boldsymbol{\Omega}}_h)^{-1} (\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}})$ is a positive quadratic form and reaches its global minimum at 0, which is achieved by the given $\boldsymbol{\theta}_h$. Since $\overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ is not a function of the parameter $\boldsymbol{\theta}$, the proof is complete. ■

Theorem 3 shows the reduction of the Sargan-Hansen test based on external information to a test of the fit of the external information and the data alone, without the model. Moreover, under the conditions of Theorem 3 the test statistic $-n\hat{Q}_n(\hat{\boldsymbol{\theta}}_{ex})$ has the form of a Wald statistic, and the Sargan-Hansen test is then equivalent to a Wald test of linear restrictions [4, p. 136]. The condition $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$ is equivalent to the main separability result of Ahu and Schmidt [1], if the external information is interpreted as a parameter with only one possible value. Their result gives an indication of the meaning of $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$, since they proved that it always holds when the first-order conditions for the GMM are satisfied. As an important special case, this result applies to OLS estimation in multiple linear models when the design matrix $\mathbf{X}$ has full rank, because the result then has the form $\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} = \mathbf{0}$, which can be directly resolved to $\boldsymbol{\beta}$. This is the mathematical form of logical independence mentioned in Section 1.

Finally, $\overline{\mathbf{h}}$, if the external information is correct, will in general almost surely be arbitrarily close to $\mathbf{0}$ for $n \to \infty$ as $E(\overline{\mathbf{h}}) = \mathbf{0}$, in which case the disturbance term $\hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ vanishes. Overall, the case $\overline{\mathbf{m}} - \hat{\boldsymbol{\Omega}}_r^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}} \neq \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ seems to be rather pathological for models that are just-identified by their moment conditions, which is why it will not be treated in the rest of the paper and only the case of just-identified models, $q = p_1$, will be treated.

2.2. The Interval-Valued Case

The assumption of point-value external information is now weakened by the assumption that a (possibly multidimensional) closed interval $\mathbf{I}_{ex}$ is known, for which we want to test the null hypothesis that it contains the true value of the external moments. The nature of this external interval is that it is based on external data that is affected by random noise. Thus, it reflects the current state of knowledge about the (moments of the) variables. Now the regularity conditions of the GMM apply to this true value, but it is not known which value in $\mathbf{I}_{ex}$ it is. Therefore, $\mathbf{I}_{ex}$ can be interpreted as coarse data, and cautious data completion can be applied to the test statistic $n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ to derive the set of possible test statistics without further assumptions [2, p. 182]. If $\mathbf{I}_{ex}$ is bounded and the test statistic is a continuous function of the external information, the result is a bounded and closed interval $[\underline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}, \overline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}]$, since in this case $\mathbf{I}_{ex}$ is compact and connected. The interval $[\underline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}, \overline{n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}}]$ is denoted by $[\underline{\chi^2}, \overline{\chi^2}]$. However, if $\mathbf{I}_{ex}$ is unbounded, the cautious data completion may result in a right-unbounded interval $[\underline{\chi^2}, \infty)$, e.g, if $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$ is used. The test statistic interval cannot be left-unbounded because the test statistic $n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}$ is a positive-definite quadratic form and therefore cannot be less than zero. In the following, we will focus on the case where the set of possible test statistics is an interval $[\underline{\chi^2}, \overline{\chi^2}]$.

One strategy for computing $[\underline{\chi^2}, \overline{\chi^2}]$ for a given data set is to use quadratic programming, as we will show now. To reflect the dependence of $\overline{\mathbf{h}}$ on the external value $\mathbf{e} \in \mathbf{I}_{ex}$, it is now written as a function $\overline{\mathbf{h}}(\mathbf{e})$. If $\hat{\boldsymbol{\Omega}}_h^{-1}$ is not a function of $\mathbf{e}$, e.g, $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, the objective function $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e}) = \overline{\mathbf{h}}(\mathbf{e})^T (\hat{\boldsymbol{\Omega}}_h/n)^{-1} \overline{\mathbf{h}}(\mathbf{e})$ is already in quadratic form based on the variable $\overline{\mathbf{h}}(\mathbf{e})$. The feasible region becomes $\overline{\mathbf{h}}(\mathbf{I}_{ex})$, the image of $\mathbf{I}_{ex}$ under $\overline{\mathbf{h}}(\mathbf{e})$. If $\overline{\mathbf{h}}(\mathbf{e})$ can be written as $\overline{\mathbf{h}}(\mathbf{e}) = \hat{\mathbf{h}} - \mathbf{e}$, where $\hat{\mathbf{h}}$ represents the sample moment, then $\overline{\mathbf{h}}(\mathbf{I}_{ex}) = \hat{\mathbf{h}} - \mathbf{I}_{ex}$ holds (Again, $\hat{\mathbf{h}} - \mathbf{I}_{ex}$ denotes the image of $\hat{\mathbf{h}} - e$ on $\mathbf{I}_{ex}$.). In this case, the feasible region is an interval. Taken together, the optimization problem is now a quadratic programming problem.

If $\hat{\boldsymbol{\Omega}}_h$ depends on $\mathbf{e}$, for example $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h$, the optimization problem is more complex. Again, the dependence on $\mathbf{e}$ is denoted by the notation $\hat{\boldsymbol{\Omega}}_h(\mathbf{e})$. In this case, the problem is not necessarily convex, as Figure 1 shows. Another problem is that the matrix $\hat{\boldsymbol{\Omega}}_h(\mathbf{e})$ must be nonsingular for each $\mathbf{e}$ for the problem to be well-defined. In the case $\hat{\boldsymbol{\Omega}}_h(\mathbf{e}) = \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})$ both problems can be solved by

Theorem 4 *The matrix $\hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})$ is positive-definite for each $\boldsymbol{e} \in \boldsymbol{I}_{ex}$ if $\hat{\mathbf{S}}_h$ is positive-definite. Assuming that $\hat{\mathbf{S}}_h$ is*

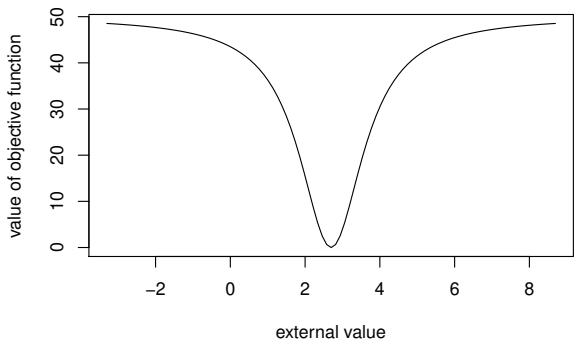

Figure 1: Graph of the objective function $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1}\overline{\mathbf{h}}(\mathbf{e})$ as a function of the external value $\mathbf{e}$ in the context of Example 1, i.e, $h(z) = y - e$, based on a sample of 50 i.i.d random variables $y_1, \ldots, y_{50}$ distributed like $N(3,1)$

positive-definite, the objective function (2) *becomes*

$$n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{\Sigma}}_h(\boldsymbol{e})^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e}) = n \cdot \frac{\overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})}{\frac{n-1}{n} + \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})}$$

and reaches its minimum over $\boldsymbol{I}_{ex}$ *at the same point as the objective function* $n \cdot \overline{\boldsymbol{h}}(\boldsymbol{e})^T \hat{\boldsymbol{S}}_h^{-1}\overline{\boldsymbol{h}}(\boldsymbol{e})$.

Proof For brevity, we will denote $\overline{\mathbf{h}}(\mathbf{e})$ by $\overline{\mathbf{h}}$ during the proof of the first statement. The first statement is clear by definition, since

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_h(\mathbf{e}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}} + \overline{\mathbf{h}})^T \qquad (3) \\ &= \frac{n-1}{n}\hat{\mathbf{S}}_h + \overline{\mathbf{h}}\overline{\mathbf{h}}^T\end{aligned}$$

is a sum of the positive-definite matrix $\frac{n-1}{n}\hat{\mathbf{S}}_h$ and the positive semi-definite matrix $\overline{\mathbf{h}}\overline{\mathbf{h}}^T$, and hence positive-definite. Using (3) and applying the formula (13.72) in Puntanen et al. [16, p. 301] to $\frac{n-1}{n}\hat{\mathbf{S}}_h + \overline{\mathbf{h}}\overline{\mathbf{h}}^T$ now yields

$$\begin{aligned}n \cdot \overline{\mathbf{h}}^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1}\overline{\mathbf{h}} = n \cdot (&\overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}} \\ &- \frac{(\overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}})^2}{1 + \overline{\mathbf{h}}^T (\frac{n-1}{n}\hat{\mathbf{S}}_h)^{-1}\overline{\mathbf{h}}}).\end{aligned}$$

The second statement follows after a little algebra. The last statement follows from the fact, that the function $f(x) = \frac{x}{\frac{n-1}{n} + \frac{x}{n}}$ is strictly increasing for every $n > 1$ in $x \geq 0$. Thus, the extrema of quadratic form $x = n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e})$ over $\mathbf{I}_{ex}$ and the extrema of $f(x)$ over $\mathbf{I}_{ex}$ are attained by the same values in $\mathbf{I}_{ex}$. ■

Theorem 4 effectively reduces the case $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})$ to the case $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, which is solvable by quadratic programming.

To extend the Sargan-Hansen test to the case of an external interval $\mathbf{I}_{ex}$, it is necessary to consider the distributional properties of the test statistic interval $[\underline{\chi^2}, \overline{\chi^2}]$. Each value $\mathbf{e} \in \mathbf{I}_{ex}$ can be specified correctly or incorrectly. If it is specified correctly, $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e}) \xrightarrow{d} \chi^2_{p_2}$, since the results of Section 2.1 apply. If it is not specified correctly, $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1}\overline{\mathbf{h}}(\mathbf{e}) \xrightarrow{d} \infty$ [4, p. 248], showing the inherent point value assumption. Only for values in a shrinking neighborhood around the true value, i.e, $\mathbf{e} = \mathbf{e}_0 + \boldsymbol{\delta}/n$, where $\mathbf{e}_0$ is the correctly specified value and $\boldsymbol{\delta}$ is a constant representing the bias, the asymptotic distribution of the test statistic is a noncentral $\chi^2_{p_2}$−distribution [4, p. 249]. The noncentral $\chi^2_{p_2}$− distribution with the noncentrality parameter λ is denoted by $\chi^2_{p_2}(\lambda)$. The interval $\mathbf{I}_{ex}$ is assumed to be constant because it is constructed outside the data, so the problem of degenerate asymptotic distributions arises. To avoid this problem, the focus is on $\underline{\chi^2}$, the minimum value of the test statistic over $\mathbf{I}_{ex}$, using the heuristic that it should not go to ∞ if $\mathbf{e}_0 \in \mathbf{I}_{ex}$. To justify this decision and to develop a test based on $\underline{\chi^2}$, two arguments are given.

First, the task is to decide whether an external interval $\mathbf{I}_{ex}$ is coherent with the data, i.e. whether it contains a value that is 'close enough' to its sample equivalent. If a test decides that this is false for $\mathbf{I}_{ex}$, it should also decide that this is false for all intervals contained in $\mathbf{I}_{ex}$ as well. For example, if a test decides that the true value is negative, one should conclude that the test would also decide that it is not in $[0,1]$. This requirement is satisfied when $\underline{\chi^2}$ is used as a single test statistic, because if $\underline{\chi^2}$ is greater than a critical value, then all values within $[\underline{\chi^2}, \overline{\chi^2}]$ are greater than it. Under the null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$, this critical value could be derived from the central χ^2−distribution to account for the fact that any value within $[\underline{\chi^2}, \overline{\chi^2}]$ could be the true one.

Second, this decision rule (reject the null hypothesis if $\underline{\chi^2}$ is greater than a critical value resulting from the central χ^2−distribution) amounts to a Γ−maximin decision rule [8, p. 193] for choosing the p-value. To recognize this, the corresponding set of gambles and the credal set must be specified. For an observed test statistic $\chi^2_{\mathbf{e}} \in [\underline{\chi^2}, \overline{\chi^2}]$, its

p-value is the probability of the event $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ under the validity of the null hypothesis, where $\mathbf{e}$ is fixed. Therefore, the indicators of the events $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ for all $\mathbf{e} \in \mathbf{I}_{ex}$ form the set of gambles. The possible asymptotic distributions for $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})$ under the null hypothesis are $\chi^2_{p_2}(\lambda)$ for $\lambda \in [0, \infty)$, so these distributions form the credal set. Now, the probabilities $P_{\chi^2_{p_2}(\lambda)}(\{\chi^2 > \chi^2_{\mathbf{e}}\})$ are increasing in λ if $\chi^2_{\mathbf{e}}$ is fixed [6], so the lower probability is reached at $\lambda = 0$. Note that this is equivalent to cumulative distribution functions that decrease pointwise in λ. But $\chi^2_{p_2}(0)$ is just the central $\chi^2_{p_2}-$distribution. Note that the above degenerate distributions at $n \to \infty$ are the limits for $\lambda \to \infty$ and thus the lower probability includes these 'distributions' as well. Finally, the lower probability $P_{\chi^2_{p_2}}(\{\chi^2 > \chi^2_{\mathbf{e}}\})$ is maximal at $\chi^2_{\mathbf{e}} = \underline{\chi^2}$, because

$$\{\chi^2 > \chi^2_{\mathbf{e}}\} \subset \{\chi^2 > \underline{\chi^2}\}$$

for all $\mathbf{e} \in \mathbf{I}_{ex}$.

Taken together, we calculate the maximum of the respective lower probabilities of the events $\{\chi^2 > \chi^2_{\mathbf{e}}\}$ for $\mathbf{e} \in \mathbf{I}_{ex}$ and compare it with the significance level α. Thus, the Sargan-Hansen test based on external intervals is

1. $P_{\chi^2_{p_2}}(\{\chi^2 > \underline{\chi^2}\}) \geq \alpha$
   $\Rightarrow$ maintain null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$
2. $P_{\chi^2_{p_2}}(\{\chi^2 > \underline{\chi^2}\}) < \alpha$
   $\Rightarrow$ reject null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$.

This test is conservative, but ensures that the asymptotic significance level is at most α, regardless of which value in $\mathbf{I}_{ex}$ is the true value under the null hypothesis.

The results obtained so far are asymptotic in nature. To derive a test for information-data conflict in small samples, distributional assumptions for $\overline{\mathbf{h}}(\mathbf{e})$ are required. So suppose that $\overline{\mathbf{h}}(\mathbf{e})$ is normally distributed for each $\mathbf{e} \in \mathbf{I}_{ex}$. If $\overline{\mathbf{h}}(\mathbf{e}) = \hat{\mathbf{h}} - \mathbf{e}$ holds, as assumed above for the application of quadratic programming, it is sufficient to assume that the sampling moment $\hat{\mathbf{h}}$ is normally distributed. Under this normality assumption, the test statistic $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})$ at a fixed $\mathbf{e} \in \mathbf{I}_{ex}$ has the scaled noncentral F-distribution $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$, where λ is again the noncentrality parameter [15, p. 889].[2] If the cumulative distribution functions of $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$ for $\lambda \in [0, \infty)$ are pointwise decreasing in λ, the same arguments used in the above construction of the Sargan-Hansen test based on external intervals can be applied. Now, the cumulative distribution function of $F_{p_2,n-p_2}(\lambda)$ is decreasing in λ [6]. This property carries over to $\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)$ since the scaling by $\frac{(n-1)p_2}{n-p_2}$ is a strictly increasing transformation and can be inverted using the definition of pushforward measures, i.e,

$$P_{\frac{(n-1)p_2}{n-p_2} F_{p_2,n-p_2}(\lambda)}(A) = P_{F_{p_2,n-p_2}(\lambda)}(\frac{n-p_2}{(n-1)p_2} \cdot A).$$

Taken together, the test for information-data conflict in small samples based on $\hat{\mathbf{S}}_h$ and the assumption of normality is

1. $P_{F_{p_2,n-p_2}}(\{\chi^2 > \frac{n-p_2}{(n-1)p_2}\underline{\chi^2}\}) \geq \alpha$
   $\Rightarrow$ maintain null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$
2. $P_{F_{p_2,n-p_2}}(\{\chi^2 > \frac{n-p_2}{(n-1)p_2}\underline{\chi^2}\}) < \alpha$
   $\Rightarrow$ reject null hypothesis $\mathbf{e}_0 \in \mathbf{I}_{ex}$.

At first glance, one might think that the choice of $\hat{\boldsymbol{\Omega}}_h$ is always important when working with small samples. This is not necessarily the case, as we will show now. From the fact that the function $f(x)$ from the proof of Theorem 4 is strictly increasing for $x \geq 0$, it follows that for every $c \geq 0$ the inequality $n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e}) > c$ is satisfied iff

$$n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Sigma}}_h(\mathbf{e})^{-1} \overline{\mathbf{h}}(\mathbf{e}) = f(n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\mathbf{S}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})) > f(c)$$

holds. Both inequalities represent the same event in the common underlying probability space, and both are assigned the same probability. Therefore, c and $f(c)$ can be interpreted as quantiles with the same level α, and the small sample tests for information-data conflict based on $\hat{\mathbf{S}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$, respectively, are the same.

The tests developed in this section are either asymptotic or assume a normal distribution. Therefore, it is important to check their properties in small samples when there is no normal distribution. Since a conservative $\Gamma-$Maximin decision rule was used, it would be interesting to compare the expected type I error rates with the significance level α. On the one hand, the use of lower probabilities could correct for the small sample bias of the asymptotic test or for the errors caused by deviations from the normal distribution. This is due to the fact that all distributions in the credal set and their convex combinations are undercut by the lower probability. On the other hand, the type I error rate could become very low if $\mathbf{I}_{ex}$ is very broad, possibly leading to low power of the tests for a fixed n. Regarding to the use of multiple external moments, the question is how this affects the type I error rate and the power of the tests. The inclusion of additional moments increases the degrees of freedom p_2, which may increase the critical values for a given significance level α. Thus, if the interval of the added moment includes or is close to the true value, the power may decrease. On the other hand, if the interval of the added moment is far from the true value, this could increase the power. We will analyze these issues through a short simulation study.

[2] The notation of Phillips [15] is very different from ours, so we explain it here: Their T is our n, their p is 1 in our case, and their q is our p_2.

3. A Simulation Study to Investigate Small Sample Properties

First, we choose sample sizes $n = 30$ and $n = 50$ so that each scenario occurs twice and the effect of increasing sample size can be analyzed. Based on Example 1, we use a simple linear regression model under Gauss-Markov assumptions and normally distributed errors. The slope is $\beta_2 = 1$ and the intercept is $\beta_1 = 16$. The sample values for the independent variable x and the dependent variable y are drawn i.i.d. as $x \sim N(4, 4)$ and $y = \beta_1 + \beta_2 x + \epsilon$ with $\epsilon \sim N(0, 60)$, where the second terms (4 and 60) are the variances. In these settings, the actual correlation between x and y is 0.25, which is low but quite typical for applied research, e.g, in psychology. The sample values are denoted by x_i and y_i for $i = 1, \ldots, n$. As external information, the moments $E(y)$, $E(x)$, and $\mathrm{Var}(y)$ are used individually or in combination of two or more of them, resulting in 7 moment scenarios. For $\mathrm{Var}(y)$, the moment function $h(\mathbf{z}) = \frac{n}{n-1}(y - \bar{y})^2 - e$ is used, where $\bar{y}$ is the sample mean of y and $\frac{n}{n-1}$ corrects for degrees of freedom. Note that using $\mathrm{Var}(y)$ leads to $\overline{h}(e) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 - e$, which is not normally distributed. Finally, two scenarios are chosen with respect to $\mathbf{I}_{ex}$ to investigate the type I error and the power, respectively. For the first scenario, $\mathbf{I}_{ex} = [0.95 \cdot \mathbf{e}_0, 1.05 \cdot \mathbf{e}_0]$ and for the second, $\mathbf{I}_{ex} = [1.2 \cdot \mathbf{e}_0, 1.3 \cdot \mathbf{e}_0]$.

To analyze the effect of the proximity of $\mathbf{I}_{ex}$ to the true value on the power of the tests, we use distributions for x and y that differ in terms of their standardized mean difference. To justify this, note that the square root of the test statistic can be simplified when using only one of the selected moments, as follows:

$$\sqrt{n \cdot \overline{\mathbf{h}}(\mathbf{e})^T \hat{\boldsymbol{\Omega}}_h^{-1} \overline{\mathbf{h}}(\mathbf{e})} = \sqrt{n}\frac{|\hat{h} - e|}{\sqrt{\hat{\omega}_h}}, \tag{4}$$

where all expressions are not written in bold because they are now single-valued. Now, (4) resembles a t-test statistic and the typical effect size used for this test statistic is the standardized mean difference $d = \frac{|e_0 - e|}{\sqrt{\mathrm{Var}(h(\mathbf{z}))}}$ [5]. The value in $\mathbf{I}_{ex}$ that is closest to $\mathbf{e}_0$ is $1.2 \cdot \mathbf{e}_0$. For the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$, it holds for $E(x)$ that $d = \frac{|4 - 1.2 \cdot 4|}{2} = 0.4$, a small effect size, and for $E(y)$ that $d = \frac{|20 - 1.2 \cdot 20|}{8} = 0.5$, a medium effect size [5]. Thus, using $E(y)$ alone should result in higher power than using $E(x)$ alone. For $\mathrm{Var}(y)$ the calculation of d is a bit more complex. Under the above conditions, $\frac{n}{n-1}(y - \bar{y})^2$ has a scaled χ^2_1-distribution. However, d is scale invariant, so we can assume without loss of generality that $\frac{n}{n-1}(y - \bar{y})^2$ is χ^2_1- distributed. This leads to $d = 0.2\frac{1}{\sqrt{2}} = 0.1414$, which is below the threshold for small effects according to Cohen [5]. Note that the effect size for $\mathrm{Var}(y)$ does not depend on the value of any moment, which is a consequence of using a normally distributed y.

Taken together, these are 2 (sample sizes) $\times$ 7 (moment combinations) $\times$ 2 (choices of $\mathbf{I}_{ex}$) = 28 scenarios. For each scenario, the rejection rates of the null hypothesis are calculated for three tests, namely the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\mathbf{S}}_h$ (abbreviated $\mathbf{SH}(\hat{\mathbf{S}}_h)$), the Sargan-Hansen test based on external intervals using $\hat{\boldsymbol{\Omega}}_h = \hat{\boldsymbol{\Sigma}}_h$ (abbreviated $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$), and the small sample test for information-data conflict (**IDC**). The significance level is always set to $\alpha = 0.05$. For $\mathbf{SH}(\hat{\mathbf{S}}_h)$ and **IDC**, the test statistic $\underline{\chi^2}$ is computed using quadratic programming as described in Section 2.2, and for $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$ it is computed using Theorem 4.

Simulations were performed in R, version 4.2.1 [17]. The R package *quadprog* [19] was used to perform quadratic programming. To calculate rejection rates, each simulation scenario was repeated 10000 times. The associated R script can be found in the electronic supplementary material. The results concerning type I error rates are presented in Table 1 and Table 2 and the results concerning power are presented in Table 3 and Table 4.

Table 1: Type I error rates for $n = 30$

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$	**IDC**
$E(y)$	0.0085	0.0061	0.0065
$\mathrm{Var}(y)$	0.0752	0.0657	0.0674
$E(x)$	0.0162	0.0121	0.0124
$E(y), \mathrm{Var}(y)$	0.0573	0.0429	0.0460
$\mathrm{Var}(y), E(x)$	0.0578	0.0456	0.0482
$E(y), E(x)$	0.0115	0.0057	0.0069
$E(y), \mathrm{Var}(y), E(x)$	0.0487	0.0302	0.0345

Table 2: Type I error rates for $n = 50$

Moments	$\mathbf{SH}(\hat{\mathbf{S}}_h)$	$\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$	**IDC**
$E(y)$	0.0055	0.0043	0.0044
$\mathrm{Var}(y)$	0.0554	0.0502	0.0508
$E(x)$	0.0089	0.0070	0.0075
$E(y), \mathrm{Var}(y)$	0.0330	0.0293	0.0302
$\mathrm{Var}(y), E(x)$	0.0358	0.0298	0.0305
$E(y), E(x)$	0.0041	0.0029	0.0031
$E(y), \mathrm{Var}(y), E(x)$	0.0264	0.0195	0.0213

Table 3: Power for $n = 30$

Moments	SH($\hat{\mathbf{S}}_h$)	SH($\hat{\boldsymbol{\Sigma}}_h$)	IDC
$E(y)$	0.7811	0.7519	0.7564
$\mathrm{Var}(y)$	0.2530	0.2341	0.2364
$E(x)$	0.5994	0.5604	0.5680
$E(y), \mathrm{Var}(y)$	0.7421	0.6687	0.6852
$\mathrm{Var}(y), E(x)$	0.6027	0.5209	0.5409
$E(y), E(x)$	0.8116	0.7404	0.7576
$E(y), \mathrm{Var}(y), E(x)$	0.8076	0.6885	0.7189

Table 4: Power for $n = 50$

Moments	SH($\hat{\mathbf{S}}_h$)	SH($\hat{\boldsymbol{\Sigma}}_h$)	IDC
$E(y)$	0.9416	0.9339	0.9355
$\mathrm{Var}(y)$	0.2808	0.2677	0.2699
$E(x)$	0.8048	0.7877	0.7906
$E(y), \mathrm{Var}(y)$	0.9040	0.8807	0.8860
$\mathrm{Var}(y), E(x)$	0.7929	0.7526	0.7626
$E(y), E(x)$	0.9607	0.9478	0.9508
$E(y), \mathrm{Var}(y), E(x)$	0.9499	0.9219	0.9302

4. Discussion

4.1. Summary of the Simulation Results

All type I error rates were below the α significance level, except in the cases where $\mathrm{Var}(y)$ was used. When $\mathrm{Var}(y)$ was used alone, the type I error rates of all tests were above α, indicating that the tests could not compensate for deviations from the normal distribution. A possible explanation could be that $\mathbf{I}_{ex}$ was not large enough. In practice, however, $\mathbf{I}_{ex}$ is determined externally and should not be expanded carelessly, since a broader $\mathbf{I}_{ex}$ would result in lower power. Nevertheless, a larger sample size would be a possible solution, since in all our scenarios an increase in sample size resulted in lower Type I error rates and higher power. When $\mathrm{Var}(y)$ was used in combination with other moments, the type I error rates were below α at $n = 30$ for the tests **SH($\hat{\boldsymbol{\Sigma}}_h$)** as well as **IDC**, and at $n = 50$ for all tests, showing that combinations of normally and non-normally distributed sample moments can improve the type I error rate. When in doubt, a simulation of the practical scenario should be performed to analyze whether the significance level is exceeded. For the scenarios using $E(y)$ alone, the smallest type I error rates were 0.0061 for $n = 30$ and 0.0043 for $n = 50$, showing that the tests can be much more conservative than the significance level would suggest. This is the expected consequence of using the conservative $\Gamma-$maximin rule. In all moment scenarios, there was a clear order of the tests in terms of type I error rate. The test **SH($\hat{\mathbf{S}}_h$)** always had higher type I error rates than **IDC** and **IDC** always had higher error rates than **SH($\hat{\boldsymbol{\Sigma}}_h$)**.

As for the power of the tests, their order corresponds to the order of the type I error rate. In all moment scenarios, **SH($\hat{\mathbf{S}}_h$)** had the highest power, followed by **IDC** and **SH($\hat{\boldsymbol{\Sigma}}_h$)**. As expected, $E(y)$ yielded the highest power when used alone, followed by $E(x)$ and $\mathrm{Var}(y)$, clearly reflecting the effect size d calculated in Section 3. With powers ranging from 0.7519 to 0.7811 for $n = 30$ and from 0.9339 to 0.9416 for $n = 50$, the moment $E(y)$ shows that the use of an external interval does not erase all of the power of the tests in our simulation study. Even for the small effect size exerted by the moment $E(x)$, the power ranged from 0.7877 to 0.8048 for $n = 50$. However, using combinations of moments does not always result in higher power. Combinations with $\mathrm{Var}(y)$ resulted in lower power than the same combinations without $\mathrm{Var}(y)$, with the sole exception of $\mathrm{Var}(y)$ and $E(x)$ for the test **SH($\hat{\mathbf{S}}_h$)** in the case $n = 30$. The maximum power reduction due to the inclusion of $\mathrm{Var}(y)$ was 0.0832 for $n = 30$ and 0.0532 for $n = 50$, respectively, for the moment $E(y)$ for the test **SH($\hat{\boldsymbol{\Sigma}}_h$)**. This reduction property is explained by the very small effect size when using $\mathrm{Var}(y)$, which causes the increase in the critical value due to the higher degrees of freedom p_2 to exceed the expected increase in the test statistic due to the inclusion of $\mathrm{Var}(y)$. Only for the test **SH($\hat{\boldsymbol{\Sigma}}_h$)** and $n = 30$ did the combination of $E(x)$ and $E(y)$ result in lower power than using $E(y)$ alone. In all other cases, however, the combination of $E(x)$ and $E(y)$ led to an increase in power, although not as pronounced, since for $n = 30$ the power only increased by a maximum of 0.0305.

Despite the conservative $\Gamma-$maximin decision rule used to construct them, the tests had good power for small sample sizes at small and medium effect sizes in our simulation scenarios. However, when a moment is not normally distributed, one should be very careful with its use, as it may lead to too high a type I error rate when used alone and to a lower power when used in combination with other moments. The simulations suggest that in scenarios such as those used here, one should select the single normally distributed moment with the largest effect size rather than using multiple moments in combination. In particular, deviations from the normal distribution, which are likely to occur frequently in practice, need to be considered in further research.

4.2. Outlook

Most importantly, the robustness of the tests to deviations from the normal distribution should be further investigated. If only the variance of y is used as the external moment, one should correct for type I error rates by deriving the

specific distribution of the test statistic in this case, given normally distributed variables.

Since the $\Gamma-$maximin decision rule is conservative, one could analyze the effect of using other p-value decision rules on the tests developed here. There are some issues regarding this endeavor. First, the upper probabilities of the events in Section 2.2 would effectively be 1. This is because the credal set includes distributions with arbitrarily large noncentrality parameters. These distributions shift the probability mass to infinity, while the interval of test statistics for a fixed n is bounded almost surely. One way to deal with this problem would be to set an upper bound on the noncentrality parameter for a fixed n. Second, note that using a $\Gamma-$maximax decision rule would result in higher p-values and thus an even more conservative test. A more liberal procedure would be to minimize the lower probabilities. Since the external interval necessarily contains values that are not the true moment value, the p-values for these would asymptotically be 0, resulting in a test that always rejects the null hypothesis even if the interval contains the true moment value.

Another way to construct a more liberal test would be to use different significance levels, possibly increasing with n, since the actual type I error rates appear to be low even at $n = 50$. However, one should keep in mind that the type I error rate depends on where the true value lies within the external interval. Therefore, it would be interesting to analyze the type I errors for several locations of the true value to calculate the worst case type I error.

Although the Sargan-Hansen test has been reduced to a Wald test as shown here, there are still ways to use information about model parameters in the tests constructed in this paper, for example implementing the OLS estimator (a function based only on the data) and an external interval that represents the external information about the regression parameter. It would be interesting to study the properties of such 'indirect' model moment conditions. In addition, other tests or frameworks for using moment-type external information could be used and compared to the tests developed here, such as the Empirical Likelihood framework [13].

Finally, the results derived here may also be useful when working with interval-valued information about moments in other research areas, since the $\Gamma-$maximin decision rule for the p-values is based on the stochastic order of the underlying family of distributions of a test statistic. This is true for many econometric and psychometric procedures, such as the Wald test for general linear and nonlinear hypotheses, the likelihood ratio test, and the Langrange multiplier test, since their test statistics are asymptotically chi-squared distributed under the null hypothesis (see Cameron and Trivedi [4] for more details). The algebraic results could help to derive analytical formulas for the externally informed estimators of Imbens and Lancaster [9] and combine them with the use of external intervals. Since these estimators are more efficient than OLS estimators and since external intervals are a more realistic and robust representation of external information, there could be an interesting interaction between the two.

Acknowledgments

The author would like to thank the three ISIPTA reviewers for their thoughtful questions and comments. They helped to improve the didactic quality of this paper and to sharpen its statistical and theoretical arguments.

References

[1] Seung C. Ahu and Peter Schmidt. A separability result for gmm estimation, with applications to gls prediction and conditional moment tests. *Econometric Reviews*, 14(1):19–34, 1995. URL https://doi.org/10.1080/07474939508800301.

[2] Thomas Augustin, Gero Walter, and Frank P. A. Coolen. Statistical inference. In *Introduction to Imprecise Probabilities*, chapter 7, pages 135–189. John Wiley & Sons, Ltd, 2014. URL https://doi.org/10.1002/9781118763117.ch7.

[3] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 1994. URL https://doi.org/10.1002/9780470316870.

[4] Adrian Cameron and Pravin Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005. URL https://doi.org/10.1017/CBO9780511811241.

[5] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. URL https://doi.org/10.1037/0033-2909.112.1.155.

[6] Bhaskar K. Ghosh. Some monotonicity theorems for chi square, F and t distributions with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(3):480–492, 1973. URL https://doi.org/10.1111/j.2517-6161.1973.tb00976.x.

[7] Lars P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. URL https://doi.org/10.2307/1912775.

[8] Nathan Huntley, Robert Hable, and Matthias C. M. Troffaes. Decision making. In *Introduction to Imprecise Probabilities*, chapter 8, pages 190–206. John Wiley & Sons, Ltd, 2014. URL https://doi.org/10.1002/9781118763117.ch8.

[9] Guido W. Imbens and Tony Lancaster. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994. URL https://doi.org/10.2307/2297913.

[10] Jan F. Kiviet and Sebastian Kripfganz. Instrument approval by the Sargan test and its consequences for coefficient estimation. *Economics Letters*, 205, 2021. URL https://doi.org/10.1016/j.econlet.2021.109935.

[11] Pavel S. Knopov and Arnold S. Korkhin. *Regression Analysis Under A Priori Parameter Restrictions*, volume 54 of *Springer Optimization and Its Applications*. Springer Science & Business Media, New York, 2011. URL https://doi.org/10.1007/978-1-4614-0574-0.

[12] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994. URL https://doi.org/10.1016/S1573-4412(05)80005-4.

[13] Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. URL https://doi.org/https://doi.org/10.2307/2336172.

[14] Paulo M. D. C. Parente and João M. C. Santos Silva. A cautionary note on tests of overidentifying restrictions. *Economics Letters*, 115(2):314–317, 2012. URL https://doi.org/10.1016/j.econlet.2011.12.047.

[15] Peter C. B. Phillips. The exact distribution of the Wald statistic. *Econometrica*, 54(4):881–895, 1986. URL https://doi.org/10.2307/1912841.

[16] Simo Puntanen, George P. H. Styan, and Jarkko Isotalo. *Matrix Tricks for Linear Statistical Models*. Springer Berlin Heidelberg, 2011. URL https://doi.org/10.1007%2F978-3-642-10473-2.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

[18] John D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26 (3):393–415, 1958. URL https://doi.org/10.2307/1907619.

[19] Berwin A. Turlach, Andreas Weingessel, and Cleve Moler. *quadprog: Functions to Solve Quadratic Programming Problems*, 2019. URL https://CRAN.R-project.org/package=quadprog. R package version 1.5-8.

[20] Zhiguo Xiao. Efficient gmm estimation with singular system of moment conditions. *Statistical Theory and Related Fields*, 4(2):172–178, 2020. URL https://doi.org/10.1080/24754269.2019.1653159.