User's Guide for Metastudy Estimation Code

Brit Sharoni

August 20, 2020

This is a guide to R and Matlab code implementing the GMM approach discussed in the paper "Identification of and Correction for Publication Bias" by Andrews and Kasy.

In what follows, we discuss the arguments for the code, as well as the packages used.[1]. Finally, we provide an illustrative example of how to use the code, and briefly discuss interpretation of the output.

PublicationBiasGMM

This is the central function producing the results. It has three parts: (1) descriptive statistics; (2) estimation of the publication probability function (p in the notation of the paper); (3) corrections to study-level estimates based on the results in (2).

In order to run the function, the user should type:
PublicationbiasGMM(X,sigma,symmetric,cluster_ID,cutoffs,Studynames) if using either Matlab or R.

Packages needed for R users Here we list the packages the code is building on. Installation is required for the program to run smoothly.

1. RColorBrewer
2. latex2exp

[1] For readers that are using the R code, please note that in order to run the code at once, all the packages must be installed in advance.

3. xtable

4. here

Arguments We next discuss the arguments for the function. We specify what each argument stands for, as well as its dimensions.[2]

$\underline{X}$: The treatment effect estimates. Should be a column vector of size $n \times 1$, for n being the number of estimates.
For R users, this argument is of type "matrix."

$\underline{\textit{sigma}}$: The associated standard errors of the estimates. Should be the same dimension as X.
For R users, this argument is of type "matrix."

$\underline{\textit{symmetric}}$: A dummy variable. $\textit{symmetric} = 1$ imposes the assumption the publication probability p is symmetric around zero. By contrast, $\textit{symmetric} = 0$ allows the publication probability to be asymmetric around zero.
For R users, this argument is of type "numeric."

$\underline{\textit{cluster\_ID}}$: A vector that allows the reader to indicate whether any of the results are taken from the same study, in which case standard errors will be clustered by study, as in the minimum wage application discussed in Andrews and Kasy. Should be the same dimension as X, and entries should be assigned the same value if and only if they're in the same cluster.
For R users, this argument is of type "matrix."

> For example, if $X = [x_1\, x_2\, x_3]'$ and x_1 and x_3 are estimates from the same study, and x_2 is not, we should have $\textit{cluster\_ID} = [1\, 2\, 1]'$.

$\underline{\textit{cutoffs}}$: A strictly increasing vector whose entries are the thresholds for the steps of the publication probability p. Should be a column vector of size $k \times 1$, for k being the number

[2] For R users we will also specify the type of the arguments (I.e., matrix, character, etc.)

of thresholds.

For R users, this argument is of type “matrix.”

Studynames: A vector that lists the names of studies the estimates are taken from. This is a row vector of size n.

For R users, this argument’s entries are of type “character.”

For Matlab users, this argument is a “character array.”

As a default, users may use vector of numbers, or an empty vector.

Outputs

Here we specify what the code produces by default. All outputs are produced as files. The files will be saved in the “FiguresandTables” folder that is part of the codes package. We will turn to discuss each of the files separately:

GMMresultsScatter.pdf These figures plot the estimate X against its standard error Σ. The grey lines in the figure mark $\left|X\right|/\Sigma = 1.96$. The grey dots mark estimates that are not statistically significant at the 5% level. In contrast, the purple dots mark the statistically significant estimates.

GMMresultsScatterHist.pdf This figure is composed of two sub-plots. The right panel is identical to the one described previously. The left panel shows a binned density plot for the z-statistic X/Σ.

GMMEstimatesAllSelectionModel.tex This file stores a table with all the estimates, along with standard errors, which are in parenthesis on the second row The first two columns correspond to the estimated mean and standard deviation of true effect size across latent studies, respectively. The remaining columns correspond to the publication probability step values, where the publication probability above the highest step is normalized to one.

GMMresultsSelectionModel.tex This file stores a table with the publication probability step values and clustered standard errors only.

GMMresultsSelectionModelCS.tex This file will be produced only if the number of thresholds equals to one (i.e., the *cutoffs* vector is of length one.) It plots a table with idenification-robust confidence sets.

GMMresultsCS.pdf This file will be produced only if the number of thresholds equal to two (i.e., the *cutoffs* vector is of length two.) It plots a 95% identification-robust confidence set.

GMMresultsOriginalAndAdjusted.pdf This figure plots the original and adjusted estimates, as well as the associated confidence sets. Original estimates and confidence sets are in purple, while adjusted ones are provided in black.

GMMresultsOriginalAndAdjustedBonferroni.pdf This figure plots the original and adjusted estimates, as well as the associated confidence sets, including Bonferroni corrections. Original estimates and confidence sets are in purple, while the adjusted ones are in black. Two black dots at both ends of the adjusted confidence sets mark the Bonferroni corrections.

GMMresultsCorrection_plot.pdf This figure plots 95% confidence bounds and the median unbiased estimator as a function of the original estimate. The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

Additional options

Here we discuss variables the reader can modify directly in the code. These are variables that are passed to the function.

Psihat0: A vector whose entries equal the starting values for optimization of the publication probability steps' values. Should be a row vector of dimension $1 \times k$, for k the length of the vector *cutoffs*. The default value is $\underbrace{[1 \cdots 1]}_{k}$.

For R users, this argument is of type "matrix."

Psihat0_theta: A vector whose entries equal the starting values for optimization of the mean and standard error of the distribution of the true treatment effects. Should be a row vector of dimension 1 × 2. The default value is $[0, 1]$.

For R users, this argument is of type "matrix."

name: This argument is relevant only for Matlab users. It specifies the names of the files which are outputs of code.[3] The default name is "GMMresults."

dofigures: A dummy variable. *dofigures*= 1 will result in an output of descriptive statistics presented in figures. The default value is 1.

For R users, this argument is of type "numeric."

doestimates: A dummy variable. *doestimates*= 1 will result in an output in the form of a latex file with the estimates for the publication probability steps' values, the estimates for the mean and standard errors for the distribution of the true treatment effect, as well as robust standard errors. The default value is 1.

For R users, this argument is of type "numeric."

dorobust: A dummy variable. *dorobust*= 1 will result in an output in the form of a latex file with the robust confidence sets for the estimates for the publication probability steps' values given there are at most two cutoffs values (either in a latex file, or as a figure.) The default value is 1.

For R users, this argument is of type "numeric."

docorrections[4]: A dummy variable. *docorrections*= 1 will result in an output in the form of figures that account for the publication bias and correct for it. The default value is 1.

For R users, this argument is of type "numeric."

[3] For R users this modification is more complicated. The names can be modified directly in the code, separately in each file. Either in the "sink()" function, for latex files, or in the "pdf()" functions for figure.

[4] Note that this code depends on the outputs produced in the estimation. Therefore, it must be associated with *doestimation*= 1, with plugged in estimation values.

Example

To illustrate how to use the code, let us consider one of the datasets that are used in the paper. The dataset we will be using is of the deworming application.[5]

The dataset can be found in the materials supplied by the authors. It is in the deworming folder, under applications, and the file name is "cleaned_deworming_data.csv".

Once the file has been uploaded, the reader can notice that there are three columns. The first corresponds to the estimates, the second to the standard errors associated with these estimates, and the third to the number of the study it belongs to. Thus, following the explanations of the arguments, the first column should be the variable X, the second the variable *sigma*, and the third the variable *cluster_ID*.

[5] We chose this data set as it is smaller than the other ones.

$$X = \begin{bmatrix} -0.76 \\ -0.45 \\ -0.38 \\ -0.05 \\ 0 \\ 0 \\ 0.01 \\ 0.027 \\ 0.03 \\ 0.04 \\ 0.05 \\ 0.13 \\ 0.13 \\ 0.15 \\ 0.16 \\ 0.17 \\ 0.19 \\ 0.29 \\ 0.35 \\ 0.7 \\ 0.9 \\ 0.98 \end{bmatrix} \qquad sigma = \begin{bmatrix} 0.44 \\ 0.17 \\ 0.23 \\ 0.08 \\ 0.27 \\ 0.14 \\ 0.09 \\ 0.175 \\ 0.15 \\ 0.06 \\ 0.05 \\ 0.11 \\ 0.15 \\ 0.09 \\ 0.08 \\ 0.07 \\ 0.45 \\ 0.09 \\ 0.13 \\ 0.45 \\ 0.18 \\ 0.15 \end{bmatrix} \qquad cluster\_ID = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 7 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \\ 18 \\ 19 \\ 20 \end{bmatrix}$$

All vectors are of size 22 × 1.

The next variable is *symmetric*. The authors model the publication function as symmetric around zero, and therefore we set *symmetric*$= 1$. Moreover, the authors model the publication probability as depending only on the significance of the results (and not, e.g., their positivity), and thus we set *cutoffs*$= 1.96$. Lastly, we must specify an array of characters that corresponds to the names of the studies the estimates were taken from. Using the material supplied by the authors we have

$$Studynames = \begin{bmatrix} 'Krugeretal.(1996)' \\ 'Watkins, CruzandPollitt(1996)' \\ 'Donnenetal.(1998)' \\ 'Awasthi, PandeandFletcher(2000)' \\ 'DossaandAtegbo(2001)' \\ 'DossaandAtegbo(2001)' \\ 'Aldermanetal.(2006)' \\ 'Awasthietal.(2008)' \\ 'AwasthiandPande(2001)' \\ 'Halletal.(2006)' \\ 'Suretal.(2005)' \\ 'Willett, KilamaandKihamia(1979)' \\ 'Josephetal.(2015)' \\ 'MiguelandKremer(2004)' \\ 'Ndibazzaetal.(2012)' \\ 'GuptaandUrrutia(1982)' \\ 'GuptaandUrrutia(1982)' \\ 'Ostwaldetal.(1984)' \\ 'Gateff, LemarinierandLabusquiere(1972)' \\ 'Wiriaetal.(2013)' \\ 'Liuetal.(2016)' \\ 'Stephensonetal.(1993)' \end{bmatrix}$$

We now turn to show how to run the code in both Matlab and R.

```
%After defining the variables in the workspace, type in this command into the command line
PublicationbiasGMM(X,sigma,symmetric,cluster_ID,cutoffs,Studynames)
```

Listing 1: Matlab Example

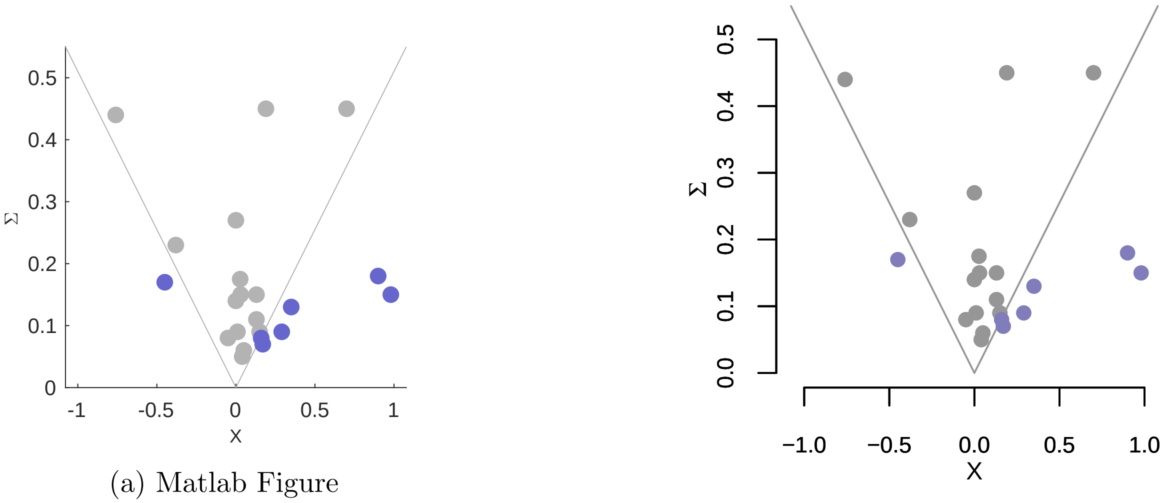
```
#Make sure you download all the packages needed
install.packages(RColorBrewer)
library(RColorBrewer)

install.packages(latex2exp)
library(latex2exp)

install.packages(xtalbe)
library(xtalbe)

install.packages("here")
library(here)

#After installing all the packages needed, type in this commend into the commend line
PublicationbiasGMM(X,sigma,cluster_ID,symmetric,cutoffs,Studynames)
```

Listing 2: R Example

Outputs

We next discuss the files that are produced and how one should interpret the results.

GMMresultsScatter.pdf and **GMMresultsSlidesScatter.pdf**



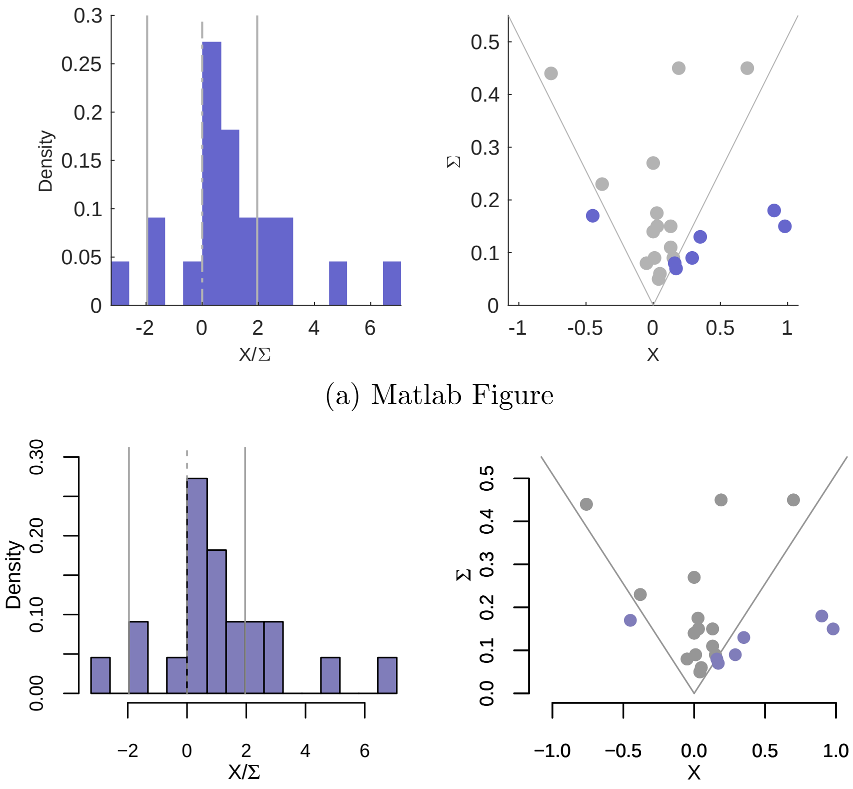
(a) Matlab Figure

(b) R Figure

Figure 1: GMMresultsScatter.pdf

This figure plots the estimates X against its standard error Σ. The grey diagonal lines in the figure mark the critical value for statistical significance in the 5% level. The purple dots mark the estimates that are statistically significant, while the grey ones mark the non-significant estimates. As you can see, the estimates values are bounded in $[-1, 1]$, and the standard errors on $[0, 0.5]$.

GMMresultsScatterHist.pdf



(a) Matlab Figure

(b) R Figure

Figure 2: GMMresultsScatterHist.pdf

The right panel of the figure is identical to the plot described previously. The purple columns in the histogram plotted in the left panel show a binned density plot for the z-statistic X/Σ. That is, they show the probability the z-statistic falls in the specified intervals. For example, the probability that the z-statistics falls in the interval $[0, 1.33)$ is a bit higher than $\frac{1}{4}$. The solid grey lines mark the critical values, 1.96 and -1.96, above which the estimate is said to be statistically significant. The dashed grey line marks zero.

GMMEstimatesAllSelectionModel.tex

Θ	Σ	β_p
0.052	0.237	0.251
(0.073)	(0.087)	(0.236)

(a) Matlab Table

Θ	Σ	β_p
0.052	0.237	0.251
(0.073)	(0.087)	(0.236)

(b) R Table

Table 1: GMMEstimatesAllSelectionModel.tex

These tables presents the estimation results. The first two columns correspond to the estimation of the mean and standard error of the true effect across latent studies. Following Definition 1 in the paper, we have

$$X^* \mid \Theta^*, \Sigma^* \sim \mathcal{N}\left(\Theta^*, \Sigma^{*2}\right)$$

So that Θ and Σ are the estimates for Θ^* and Σ^*. The last column is the publication probability step value. Recalling Definition 1 in the paper, we have that

$$D \mid X^*, \Theta^*, \Sigma^* \sim \mathrm{Ber}\left(p\left(Z^*\right)\right)$$

for p the publication probability and $Z^* = X^*/\Sigma^*$. Note that Z^* is the z-statistic that corresponds to the draw (X^*, Σ^*), and we assume that p is a step function with thresholds defined by the vector of cutoffs. Thus, $\beta_p = 0.251$ means that a paper with a non-significant result (i.e., $|X^*/\Sigma^*| < 1.96$) will be published with probability proportional to 0.251. As a normalization, a paper with statistically significant result will be published with probability proportional to 1. More precisely,

$$p\left(Z\right)\propto\begin{cases}\beta_p & \text{if } \left|Z\right|<1.96\\ 1 & \text{if } \left|Z\right|\geq1.96\end{cases}$$

The numbers on the second line in parentheses correspond to the standard errors of the estimations.

GMMresultsSelectionModel.tex

β_p
0.251
(0.236)

(a) Matlab Table

β_p
0.25
(0.24)

(b) R Table

Table 2: GMMresultsSelectionModel.tex

This table is identical to the previous table, after deleting the estimates for Θ and Σ.

GMMresultsSelectionModelCS.tex

β_p Lower Bound	β_p Upper Bound
0.048	∞

(a) Matlab Table

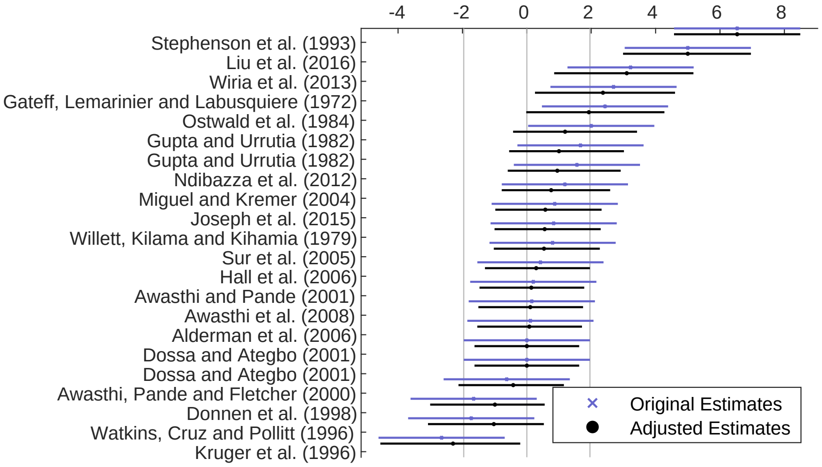
β_p Lower Bound	β_p Upper Bound
0.048	∞

(b) R Table

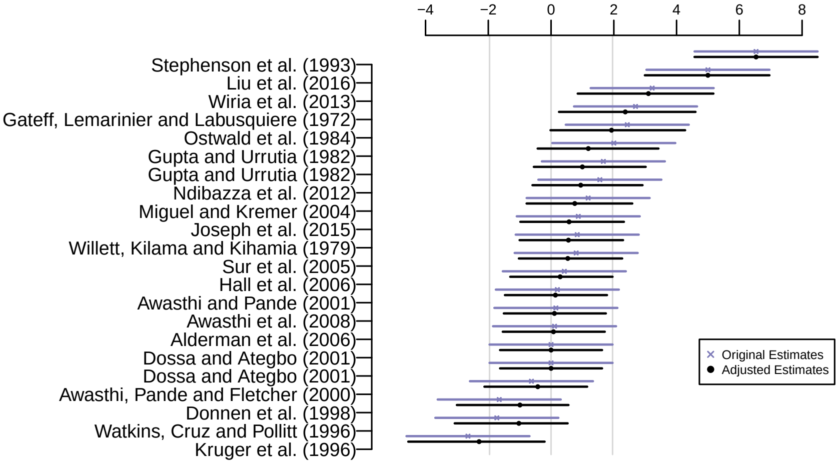
Table 3: GMMresultsSelectionModel.tex

This table plot identification-robust 95% confidence sets for β_p. As we can see, in this case, the robust confidence sets cover almost the full parameter space (which is $[0,\infty]$.)

GMMresultsOriginalAndAdjusted.pdf



(a) Matlab Figure



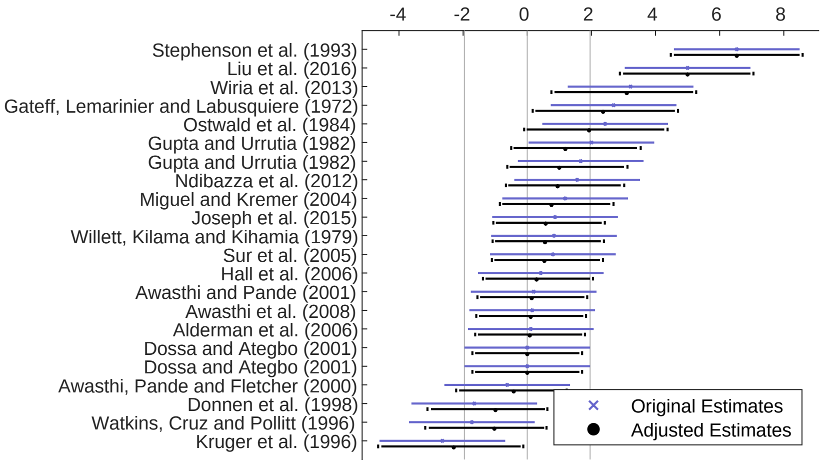
(b) R Figure

Figure 3: GMMresultsOriginalAndAdjusted.pdf

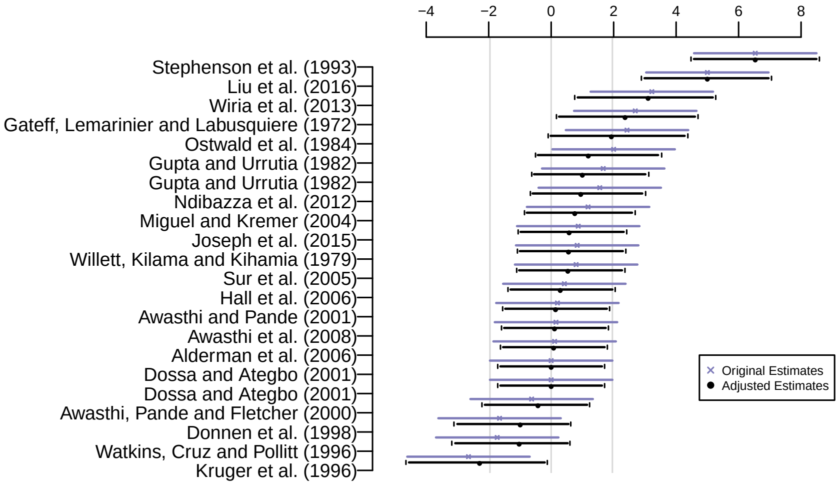
This figure plots the original estimate and 95% confidence interval, as well as the median unbiased estimate and adjusted 95% confidence interval. Adjusted intervals not accounting for estimation error in β_p are plotted with solid lines.

GMMresultsOriginalAndAdjustedBonferroni.pdf

This figure is almost identical to the previous one. The only difference lies in the adjusted 95% confidence intervals. In this case, the adjusted intervals not accounting for



(a) Matlab Figure

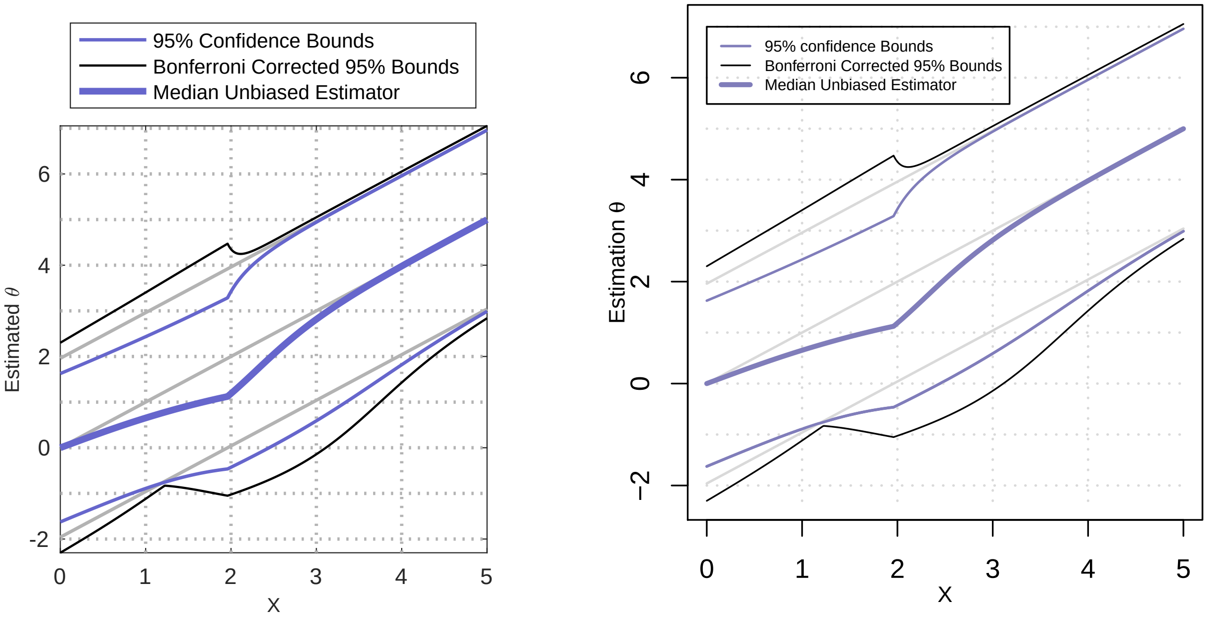


(b) R Figure

Figure 4: GMMresultsOriginalAndAdjustedBonferroni.pdf

estimation error are plotted with solid lines, while end points for intervals accounting for estimation error are marked with “|”.

GMMresultsCorrection_plot.pdf



(a) Matlab Figure

(b) R Figure

Figure 5: GMMresultsCorrection_plot.pdf

This figure plots 95% confidence bounds and the median unbiased estimator. The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.