Generalized Sensitivity Analysis and Application to Quasi-experiments

Masataka Harada
New York University
masatakaharada@nyu.edu
August 27, 2013

Abstract

The generalized sensitivity analysis (GSA) is a sensitivity analysis for unobserved confounders characterized by a computational approach with dual sensitivity parameters. Specifically, GSA generates an unobserved confounder from the residuals in the outcome and treatment models, and random noise to estimate the confounding effects. GSA can deal with various link functions and identify necessary confounding effects with respect to test-statistics while minimizing the changes to researchers’ original estimation models. This paper first provides a brief introduction of GSA comparing its relative strength and weakness. Then, its versatility and simplicity of use are demonstrated through its applications to the studies that use quasi-experimental methods, namely fixed effects, propensity score matching, and instrumental variable. These applications provide new insights as to whether and how our statistical inference gains robustness to unobserved confounding by using quasi-experimental methods.

*I thank Christopher Berry, Jennifer Hill, Luke Keele, Robert LaLonde, Andrew Martin, Marc Scott, Teppei Yamamoto for helpful comments on earlier drafts of this paper.
1 Introduction

Political science research nowadays commands a wide variety of quasi-experimental techniques intended to help identify causal effects. Researchers often begin with collecting extensive panel data to control for unit heterogeneities with difference in differences or fixed effects. When panel data are not available or the key explanatory variable has no within-unit variation, researchers typically match treated units with control units from a multitude of matching techniques. Perhaps, they are fortunate to find a good instrumental variable which is effectively randomly assigned and affects the outcome only through the treatment variable. Researchers who use these techniques are tempted to report the results causally.

Of course, none of these methods ensures that the findings are causal. In contrast to randomized experiments (assuming randomization is pristine), which randomize treatment assignments in terms of both observed and unobserved confounders, these quasi-experiments all rely on untestable assumptions (Heckman, Ichimura and Todd 1998; Smith and Todd 2005). In order to infer causality, researchers need to assume some version of ignorability (Rubin 1978), also known as conditional independence or the absence of omitted variables, and defend its validity to the readers. Although political scientists acknowledge that the assumption of ignorability is “a strong condition” (Ho et al., 2007), few studies in the discipline rigorously test this assumption as “it is one about which social scientists are deeply knowledgeable” (Ho et al., 2007).

However, the cost of blind acceptance of ignorability is nontrivial. LaLonde’s (1986) pioneering work shows that the treatment effects estimated with observational data are strikingly different from those of the randomized experiment, no matter what econometric techniques are used. Thus, if researchers’ initial motivation to perform quasi-experimental

---

1 See Imbens and Wooldridge (2009) for the survey on quasi-experimental method.
2 However, see Clarke (2005), Blattman (2009), and Caughey and Sekhon (2012), for example, that perform sensitivity analysis to verify the ignorability assumption. Also, see Imai and Yamamoto (2010) and Imai et al. (2011) for the application of sensitivity analysis to measurement error problem and causal mediation analysis.
3 Note, however, that using propensity score matching (Rosenbaum and Rubin, 1983), Dehejia and Wahba (2002) show that the discrepancy between experimental and non-experimental results in LaLonde (1986) can
methods is to produce estimates comparable to those produced by experiments, the demon-
stration of the robustness of their estimates against the confounding from unobserved con-
founders is an unavoidable step. The obvious problem is that we cannot observe unobserved confounders and therefore cannot measure their confounding effects.

Sensitivity analysis offers a partial solution to this problem through a thought ex-
p[444] periment. It answers how strong a confounding effect researchers need to assume for an unobserved confounder to change the treatment effect by a given amount. A very basic sen-
itivity analysis goes as follows. First we posit an unobserved covariate, \( U \), that if included with our group of observed confounders, \( X \), would allow us to satisfy ignorability. Then consider the regression we could fit of our outcome, \( Y \), on \( U, X \), and treatment variable \( Z \) if \( U \) were indeed observed. The estimates from this regression (assuming it was specified correctly) would allow us to estimate the true treatment effect (the coefficient of \( Z \) in this regression). We can then consider what characteristics of \( U \) (generally expressed with regard to it’s conditional association with \( Z \) and \( Y \)) would be necessary to change the original estimate, \( \hat{\tau} \), to the target value of the researchers’ choice (\( \tau \)). If the characteristics required to change \( \hat{\tau} \) to the target value are unlikely to occur in practice, such an unobserved confounder is unlikely to exist. Thus, the treatment effect is defended against this type of unobserved confounder.

The goal of this study is to introduce the generalized sensitivity analysis (GSA) I developed and to demonstrate that GSA is applicable to a wide range of research questions requiring minimal change to the researchers’ original estimation models. A number of sen-
sitivity analyses have been proposed since the seminal work by Cornfield et al. (1959), but what make GSA appealing are its simplicity and versatility. Broadly speaking, GSA gen-

---

4See Rosenbaum (2005) for the survey of sensitivity analysis. The term “sensitivity analysis” refers to different exercises depending on the discipline and the context. In this paper, it refers to its most common definition, the robustness check against unobserved confounder.

5Target values are typically set at the values that researchers can easily interpret such as $1,000 less earning (Imbens 2003) or half of the original treatment effect (Blattman 2009), or at the critical values of the statistical significance (e.g. 1.96) if the method allows using test statistics as a target value.
erates an unobserved confounder from the residuals in the outcome and treatment models, and random noise. GSA only asks the researchers to add this generated unobservable as a regressor to their estimation models, so this process does not require any further knowledge of advanced statistics. Thus, GSA is available to any applied researchers and applicable to any estimation models where the residuals contain useful information to recover unexplained variations of the model. With GSA, researchers can also set the target value in terms of test statistics. In this paper, GSA are mostly applied to quasi-experiments, but they are not a prerequisite for the sensitivity analysis. Indeed, GSA can be performed more easily in non-quasi experimental settings. However, I focus on quasi-experiments because this is the situation in which reducing the concern on unobserved confounders benefits the research the most, nevertheless the sensitivity analysis is rarely performed as a standard routine.

The remainder of this paper is organized as follows. The next section surveys existing sensitivity analyses and summarizes their relative strengths and weaknesses. The third section introduces GSA. The fourth section presents a basic application of GSA using Fearon and Laitin (2003). The following three sections present the applications of GSA to fixed effects models, propensity score matching and instrumental variable approach using the datasets of Fearon and Laitin (2003), LaLonde (1986) and Acemoglu, Johnson and Robinson (2001), respectively. Then, the final section concludes. Monte Carlo evidence that GSA estimates the unbiased sensitivity parameters and the detailed procedure of GSA are available in the appendices.

2 Sensitivity Analysis

In this section, I briefly survey existing sensitivity analyses. To understand the differences between existing sensitivity analyses, it would be helpful to focus on two primary characteristics. The first characteristic is how to define the sensitivity parameters that represent the size of confounding effects. Some methods assign a single parameter as the sensitivity parameter. For example, Rosenbaum (2002) defines the sensitivity parameter as the exponential
value, $\gamma$, that bounds the confounding effects on the odds ratio of the treatment assignment from $1/e^\gamma$ to $e^\gamma$. Single parameter specification makes it possible to present the results in a simple bivariate summary such as two-dimensional diagram or cross table. However, single parameter specification often requires additional constraints or assumptions. For example, Altonji et al. (2005) define the sensitivity parameter as the correlation between the error terms in the outcome and treatment assignment equations, which means that an unobserved confounder affects the two equations by the same amount.

Imbens (2003), on the other hand, considers two parameters, the partial effect of the unobserved confounder on the treatment and that on the outcome, as the sensitivity parameters. The advantage of this specification is that, if we think sensitivity analysis as a class of omitted variable bias problem, these parameters account for the two main components of the omitted variable bias. To see this point, let us state the omitted variable bias of a bivariate regression with a single unobserved confounder (Wooldridge 2009).

\[\begin{align*}
y &= \beta_0 + \tau z + \delta u + \epsilon \\
y &= \tilde{\beta}_0 + \tilde{\tau} z + \tilde{\epsilon} \\
u &= \gamma_0 + \gamma_1 z + \epsilon \\
Bias(\tilde{\tau}) &= E(\tilde{\tau}) - \tau = \delta \cdot \gamma_1
\end{align*}\]

where the bias is defined as the product of the partial effect of an unobserved confounder on the outcome ($\delta$) and the regression coefficient of the treatment on the unobserved confounder ($\gamma_1$). These two quantities are essentially equivalent to the two parameters in Imbens (2003). Dual parameter specification also allows researchers to provide the information about which of the outcome or the treatment model is more likely to be the source of confounding. The drawback is that three dimensional relationships between two sensitivity parameters and the treatment effect make the presentation of the result more complicated.

The second key characteristic of sensitivity analysis approaches is how to obtain a
given amount of a confounding effect that corresponds to the sensitivity parameters. Broadly speaking, the existing sensitivity analyses can be classified into three approaches in this regard, namely algebraic-based, likelihood-based, and computational approaches. The algebraic method first derives the formula for the effect of the unobserved confounder on the treatment effect. Then, it calculates the confounding effect by entering the values of sensitivity parameters to the formula. The simplest example is the aforementioned Equation (4) for omitted variable bias. This approach provides a concise solution, but requires researchers to make formulae with a set of appropriate assumptions between variables, which is particularly difficult when the model involves link function.\footnote{See Imai and Yamamoto (2010) as another example of algebraic-based approach.} For example, Rosenbaum (2002) developed the highly model-independent sensitivity analysis in which the effect of the binary unobserved confounder on the odds ratio can be represented as a parameter, $\gamma$, regardless of the parametric model specification. However, we must assume that the treatment group and the control group are perfectly balanced in observed covariates and that the unobserved variable is nearly perfectly correlated with the outcome variable.\footnote{See Rosenbaum (2005) and Becker and Caliendo (2007) for an introductory application of Rosenbaum’s (2002) sensitivity analysis.}

Imbens (2003) developed the likelihood-based approach, in which he specifies the complete data likelihood that would exist if $U$ following Bernoulli(.5) were observed. Because the two sensitivity parameters, the partial effect of $U$ on $Z$ and that on $Y$, are explicitly embedded in the likelihood function, it can be maximized for any combination of sensitivity parameters. Thus, this approach requires a small number of iterations in finding a set of $\alpha$ and $\delta$ sufficient to draw the contour. Also, the key assumptions such as ignorability are guaranteed to hold because of this built-in setup. However, the likelihood function often has a hard time in converging at the global maximum particularly when the estimation models involve link functions. Moreover, standard error estimates obtained from Imbens (2003) sensitivity are not appropriate as discussed in Appendix A.

In computational approach, a pseudo unobservable (PU, henceforth) is drawn from a
“semi-random” distribution, and the confounding effect is calculated through the estimation of the models including the PU. I call the distribution a “semi-random” because the PU is generated using some information in the treatment model and the outcome model. This is because, without relying on these models, it is difficult to generate PUs that affect the treatment effect by a non-negligible amount. For example, Ichino, Mealli and Nannicini (2008) extends Rosenbaum’s (2002) sensitivity analysis to the outcome model and draw PUs from the conditional distribution of PU given the expected values of the binary outcome and the binary treatment.

As Imbens (2003) points out, the identification of this approach is weak because a given draw of the PU may or may not represent the hypothesized properties of the unobserved confounder. Therefore, the identification of the estimators must be verified through iteration. For example, GSA requires about at least 100 successful draws to estimate a contour. This means that if a successful draw (as explained later) occurs every 100 draws, the model needs to be estimated for 10,000 times.\(^8\) Thus, computational approach is more time-consuming than the other two. Computational approach, however, simplifies the process of calculating the size and uncertainty of the confounding effects, which is particularly useful when the baseline model (i.e. the model to which the researchers perform the sensitivity analysis) is complex.

Finally, the limitation of sensitivity analysis must be also addressed. The limitation is that sensitivity analysis “does not provide an objective criterion (Imai et al., 2011)” for how strong effect the researchers need to assume about an unobserved confounder because an unobserved confounder in sensitivity analysis is nothing but a figment of a researcher’s imagination. Thus, defining the strength of confounding effect in less subjective way is one of the most important tasks in sensitivity analysis. Altonji et al. (2005), for example, postulate that the size of the bias due to unobserved confounders is the same as that due to observed

---

\(^8\) These numbers are provided as a guidance for the readers based on the author’s experience. The actual number of iterations varies by the dataset and the model specifications. In general, the more successful draws result in the more accurate analysis.
covariates and show that this requires a set of assumptions weaker than those required by OLS. However, in many cases, researchers have little idea about the strength of unobserved confounders before (and even after) the sensitivity analysis. Thus, whenever researchers can avoid the problem of unobserved confounders by adopting alternative research design, sensitivity analysis should be considered as an auxiliary means.

3 GSA

In this section, I explain the outline of GSA. GSA adopts a computational approach with dual sensitivity parameters and presents the results in a similar way to Imbens (2003). In addition to the advantages of computational approach and dual parameter specification, GSA has three distinct strengths. First, GSA precisely estimates the sensitivity parameters necessary to produce a given amount of the confounding effect. Second, GSA accurately recovers the standard errors of the treatment coefficient obtained from the model that includes true $U$, thereby enables researchers to set the target value using test statistics. Because social scientists are mostly interested in whether their hypothesis is supported rather than how large the effect is, this broadens the scope of research to which sensitivity analysis can be applied. I prove these properties in Appendix A through Monte Carlo simulations. Finally, GSA minimizes the change researchers need to make to their original estimation models in performing the sensitivity analysis because researchers only need to add a PU as a regressor in the treatment and the outcome models.

---

9Although I have independently developed GSA, a prototype of the computational approach was first reported in the supplemental file of Bingenheimer et al. (2005). I call this a “prototype” because their setup is similar to GSA but has some crudities. To list them, they generate only 12 artificial confounders of which parameters are constrained non-exhaustive ways, these artificial confounders are not orthogonal to covariates, and most importantly the identification is not warranted.

10On the other hand, Ichino, Mealli and Nannicini (2008) develop a computational approach with dual sensitivity parameters based on Rosenbaum (2002), which is highly model independent. However, their method requires (mostly propensity score) matching technique and dichotomization of the outcome and the treatment.
GSA uses a subset of the assumptions of Imbens (2003), which are,

\[ Y(1), Y(0) \perp Z \mid X \]  
(5)

\[ Y(1), Y(0) \perp Z \mid X, U \]  
(6)

\[ X \perp U \]  
(7)

where \( Y(\{0,1\}) \) is the potential outcome when \( Z = \{0,1\} \), \( Z \) is the treatment, \( X \) is a set of covariates, and \( U \) is an unobserved confounder. Equation (6) means that the ignorability assumption holds only when an unobserved \( U \) is included in the outcome model. The stable unit treatment value assumption (SUTVA) and other regression assumptions that do not conflict with this setup must be also satisfied. Researchers can choose any combination of the outcome and the treatment models as long as the right hand side variables are linearly additive.\(^{11}\) Thus, the true models is represented as,

\[ \mathbb{E}(Y \mid Z, X, U) = g_{y}^{-1}(\tau Z + X\beta + \delta U) \]  
(8)

\[ \mathbb{E}(Z \mid X, U) = g_{z}^{-1}(X\gamma + \alpha U) \]  
(9)

where \( g_{\{y,z\}}^{-1} \) is a link function, \( \tau \) is a target value, and \( \alpha, \beta, \gamma, \) and \( \delta \) are coefficients. However, what researchers estimate under the assumption in Equation (5) are the coefficients from the models below,

\[ \mathbb{E}(\hat{Y} \mid Z, X) = g_{y}^{-1}(\hat{\tau} Z + X\hat{\beta}) \]  
(10)

\[ \mathbb{E}(\hat{Z} \mid X) = g_{z}^{-1}(X\hat{\gamma}) \]  
(11)

where \(^{-}\) indicates that the coefficients are estimated with the bias due to the omission of \( U \).

When the outcome model is linear, the treatment model is logit, and \( U \sim Bin(1, 0.5) \), GSA

\(^{11}\)If this condition does not hold, the difference between the left hand side variable and its predicted value is not necessarily associated with the unexplained variations. Thus, the construction of PU is difficult.
is comparable to Imbens’ (2003) sensitivity analysis, but these two methods disagree about their estimates of the standard errors.\textsuperscript{12}

Now, I briefly explain the procedure of GSA along with Figure 1. This process is also discussed further in Appendix B. The goal of GSA is to find sufficient and various sets of the sensitivity parameters \((\alpha, \delta)\), which change the treatment effect (or its test statistics) from \(\tilde{\tau}\) to the target value through the generation of multiple PUs. The first step is to define the target value in terms of coefficient of the treatment variable or its test statistics. If the researchers observe a positive and statistically significant treatment effect in Equation (10) and seek to identify the combinations of sensitivity parameters corresponding to a \(U\) that

\textsuperscript{12}See Appendix A for the comparison of these two approaches.
would yield a statistically marginally insignificant treatment effect estimate, the target value would be $t = 1.96$ (or a value slightly smaller than 1.96).

Next, GSA generates a candidate variable of a PU from a weighted average of the deviance residuals from the treatment and outcome models plus random noise. This variable is further residualized by covariates to obtain the candidate variable. In GSA, these weights of the deviance residuals, $c_1$ and $c_2$, are the two key location parameters. Specifically, the larger the $c_1$, the larger the $\alpha$ because the residual in the treatment model is associated with the unexplained variations in the treatment model. Likewise, the larger the $c_2$, the larger the $\delta$. Thus, GSA indirectly adjusts the sizes of two sensitivity parameters, $\alpha$ and $\delta$, through the manipulation of $c_1$ and $c_2$.

Then, Equations (8) and (9) are estimated replacing $U$ with a candidate variable, $\hat{U}$, and $\hat{U}$ is adopted as a PU, if the estimand is close enough to the target value. For example, when the target value is set at 1.96 in test statistics and the error tolerance level is 1%, a candidate variable is accepted as a PU if the test statistics of the model with the candidate variable is $1.94 < t < 1.98$. PUs are generated in this way for a sufficient number of times (usually more than 100 times), and the partial effects of PUs are recorded for each successful draw.

Now, a contour curve like the one in Figure 2 can be produced by plotting the sets of the partial effects. Superimposing a fitted fractional polynomial or another function to these scatter plots is also informative if the researchers need to know a set of point estimates of the partial effects in a form of a contour line. Following Imbens (2003), the final step is to add the plots of the partial effects of each covariate to the figure, which provides the contour curve the relative scale of the confounding effects.

---

13The error tolerance level smaller than 1% usually does not change the result. See Appendix B for the discussion on the appropriate error tolerance level.
4 A basic application of GSA with Fearon and Laitin (2003)

In this section, a basic application of GSA is presented using the dataset from Fearon and Laitin (2003), one of the most cited works in political science.

**Background.** Fearon and Laitin (2003) (henceforth, FL) challenge the conventional wisdom in comparative politics that civil wars are rooted in ethnic fractionalization (Horowitz 1985). Specifically, no matter how ethnic fractionalization is severe, civil wars will not occur if rebel leaders know their plots will fail. Thus, focusing on feasibility of civil wars rather than motive and the fact that most post-1945 civil wars have been initiated as small-scaled insurgencies, FL argue that mountainous terrain is the true determinant of civil war onset because it makes governments’ counterinsurgency operations difficult.

**Preliminary analysis.** FL’s main claim is that the positive association between ethnic fractionalization and civil war onset is spurious correlation owing to the omission of economic variable, and mountainous terrain is a persistent predictor that explains civil war onset. Their claim is tested with pooled cross-country data from 1945 to 1999 that contain a sample of 6,610 country years, which is confirmed in Columns 1 and 2 of Table 1. Without controlling for lagged per capita GDP (in Model 1), the coefficient of ethnic fractionalization is statistically significant and positive, and this is the basis of the conventional wisdom that ethnic fractionalization matters for civil war onset. However, once lagged per capita GDP is added as a regressor of the model in Column 2, which reflects FL’s argument, the statistical significance disappears, while logged estimated % mountainous terrain retains statistically significant and positive association with civil war onset.

**Sensitivity analysis.** Given the above discussion, the following models are assumed for the sensitivity analysis that examines the robustness of the coefficient of ethnic fractionalization.

---

14 The original article presents the contour plot of civil war onset against ethnic fractionalization and mountainous terrain, but I use a parametric approximation to compare different models.
Table 1: Estimated baseline coefficients of Fearon and Laitin (2003). Logistic regression. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Covariates not listed above the table are ln(population), Noncontinuous state, Oil exporter, New state, Instability, Democracy, Religious fractionalization, Anocracy. Column 3 is estimated with the artificial cross-sectional dataset.

<table>
<thead>
<tr>
<th>Dependent Variable: Civil war onset</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic Fractionalization</td>
<td>0.880**</td>
<td>0.166</td>
<td>-0.404</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td>(0.374)</td>
<td>(0.373)</td>
<td>(0.660)</td>
<td>(0.474)</td>
</tr>
<tr>
<td>In(Estimated % mountainous terrain)</td>
<td>0.263***</td>
<td>0.219***</td>
<td>0.336**</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.085)</td>
<td>(0.131)</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Inversed lagged per capita GDP</td>
<td>0.344***</td>
<td>0.402***</td>
<td>0.404***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.097)</td>
<td>(0.093)</td>
<td></td>
</tr>
<tr>
<td>Regional dummies</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Year dummies</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Control variables</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>6,526</td>
<td>6,327</td>
<td>239</td>
<td>6,217</td>
</tr>
</tbody>
</table>

where onset is the dummy variable of civil war onset, ethfrac is the continuous treatment variable of ethnic fractionalization, lmtnest is logged estimated % mountainous terrain, $x$ is the set of the covariates listed in the caption of Table 1, $\Lambda$ is logit function, $\tau, \theta, \beta, \delta, \lambda, \gamma$ and $\alpha$ are coefficients, and $u$ is an unobserved confounder. With these models and the target value set at 1.96 in test statistics, the sensitivity analysis provides the necessary sizes of the sensitivity parameters ($\alpha$ and $\delta$) to change the positive and significant coefficient of ethfrac to the one that is statistically marginally insignificant. On the other hand, when $\alpha = \delta = 0$, these models produce identical estimates as those of Column 1 in Table 1. Similarly, the

\[ E(\text{onset}) = \Lambda(\tau \text{ethfrac} + \theta \text{lmtnest} + \beta x + \delta u) \]  
\[ E(\text{ethfrac}) = \lambda \text{lmtnest} + \gamma x + \alpha u \]

\[ 15 \text{Lagged per capita GDP} \text{gdpenl is assumed to be unobservable in Model 1.} \]
sensitivity analysis that examines FL’s argument can be set up as follows:

\[
\begin{align*}
\mathbb{E}(\text{onset}) &= \Lambda(\tau \text{lmtnest} + \theta \text{ethfrac} + \vartheta \text{gdpenl} + x\beta + \delta u) \\
\mathbb{E}(\text{lmtnest}) &= \lambda \text{ethfrac} + \omega \text{gdpenl} + x\gamma + \alpha u
\end{align*}
\] (14) (15)

where \(\text{gdpenl}\) is inversed lagged per capita GDP,\(^{16}\) and \(\tau, \theta, \vartheta, \beta, \delta, \lambda, \omega, \gamma\) and \(\alpha\) are coefficients. Because FL emphasizes the persistence of mountainous terrain as a predictor of civil war onset, logged estimated % mountainous terrain (\text{lmtnest}) is now the treatment variable, and ethnic fractionalization (\text{ethfrac}) is one of the observed covariates.

The two panels in Figure 2 present the results of these sensitivity analysis using GSA. Interestingly, these graphs show that the effect estimate of logged estimated % mountainous terrain is at best as robust as that of ethnic fractionalization to unobserved confounding. To explain the figure first, the sets of \(\alpha\) and \(\delta\) in form of partial correlations are displayed as blue hollow circles. They are transformed in this way to make the confounding effects of PUs easily comparable with the counterparts of the observed covariates. Thus, all of these scatter plots represent partial correlations of pseudo unobservables that change the test statistics of the treatment effects to 1.96 plus or minus 0.01. These contour plots are hyperbolic curves because the size of confounding effect is largely determined by the product of these two partial correlations as shown in Equation (4). The contour in Figure 2 is wide because \(\alpha\) and \(\delta\) necessary to change the target value by a given amount change depending on the coefficients of other covariates with non-identity link function. Because \(U \perp X\) does not necessarily imply \(U \perp X \mid Y\) or \(U \perp X \mid Z\), the coefficients of the observed covariates change slightly for every generation of \(PU\).\(^{17}\) The partial effects of covariates that are included in the model are indicated in plus signs (+), and the partial effects of covariates that are

\(^{16}\)This variable is inversed to make the coefficient positive.

\(^{17}\)The discrepancies between the estimated values and the target value are another source for the width of the contour. However, in this application, the error tolerance level in accepting PUs is sufficiently small (0.5%) to eliminate this possibility. See Appendix B for the relationship between the error tolerance level and the width of the contour.
omitted from the model are indicated in hollow diamond (○).

To look at the left panel first, the variables of new state (nwstate) and prior war (warl) are located above the contour. This indicates that, if there exists an unobserved covariate of which the confounding effects are as strong as those of the variable of new state or prior war, the effect of ethnic fractionalization becomes statistically insignificant. In general, when the observed covariates have stronger partial effects than those represented by the contour, the robustness of the treatment effect estimate is questioned. This is because one reasonable assumption about an unobserved confounder is that its confounding effect is as strong as that of the most influential observed covariate (Altonji et al. 2005). Indeed, as Fearon and Laitin (2003) point out, inverted lagged per capita GDP (gdpenl) turns out to be the key confounding variable in the left panel, which is plotted far above the contour. This suggests if inverted lagged per capita GDP were to be added in the estimation in Column 1, the statistical significance of ethnic fractionalization would have disappeared as indicated in Column 2.
However, the sensitivity analysis in the right panel of Figure 2 shows that Model 2 is at best as robust as Model 1 against unobserved confounders. The right panel shows that an unobserved confounder that is as powerful as logged population (\textit{lpop11}) would also change the coefficient of logged estimated % mountainous terrain statistically insignificant. To demonstrate that such an unobserved confounder is likely to exist, I created 16 regional dummy variables and plotted their multiple correlation (i.e. square root of their partial R-square) in the right panel, which is displayed as hollow diamond.\textsuperscript{18} Although they do not control for unit heterogeneities, the regional fixed effects control for important regional heterogeneities such as geographical features (not limited to mountainous terrain), history and culture that affect regional security.\textsuperscript{19} The multiple correlation is plotted far above the contour indicating that regional dummies would be a very strong confounders against their claim if it were to be added to the model. Indeed, once these regional dummies are added to the model (Model 3 of Table 1), the coefficient of logged estimated % mountainous terrain becomes half and is no longer statistically significant.

Some readers might wonder what the advantages of looking at Figure 2 over comparing coefficients in Table 1 are. The following three advantages of using GSA are worth being emphasized. The first advantage is, of course, that sensitivity analysis provides a glimpse into the potential impact of an unobserved covariate. Second, dual parameter specification enables researchers to identify which of the outcome model or the treatment model is more likely to be confounded. This is an important information because it affects researchers research design when their original estimation strategy turns out to be not robust according to the sensitivity analysis. For example, one may wish to employ quasi-experimental techniques, but they typically reduce the confounding in the treatment model. Finally, it provides a relative scale of unobserved confounders compared to observed confounders and

\textsuperscript{18}The breakdown of 16 regions are: North/Central America, South America, North/West Europe, Southern Europe, East Europe, Former Soviet, West Africa, Central Africa, East Africa, Southern Africa, North Africa, Middle East, East Asia, South Asia, Southeast Asia, and Oceania.

\textsuperscript{19}For example, African countries may have more civil war onsets because of the borders imposed by the former colonial powers. FL also used regional fixed effects in the robustness check but, contrary to my finding, reported that their findings are not affected by them.
vice versa. Without sensitivity analysis, it is hard to tell an unobserved covariate in the right panel, for instance, needs to have the confounding effect as strong as that of logged population.

5 Application to Fixed Effects with Fearon and Laitin (2003)

Background. The previous section hints the role of so-called fixed effects, which refer to a set of binary regressors that are created according to the multilevel or cluster structure of a given dataset. Although recent studies have found that causal inference using fixed effects approach with panel data is more difficult than it has been thought or requires fairly strict assumptions such as sequential ignorability (Sobel 2012), the fixed effects are still one of the most popular empirical approaches because they remove the observed and unobserved unit heterogeneities that are otherwise difficult to control for.

In applying GSA to fixed effects model, two questions are of particular interests. First, how the results of the sensitivity analysis differ between the case in which the treatment is robust to the inclusion of the fixed effects and the case in which it is not. This is a practically quite important question because researchers who have some cross-sectional data often debate whether to obtain the additional panel data to remove unit heterogeneities. The other question is whether and how the inclusion of the fixed effects makes the model robust.

Sensitivity analysis 1: when should researchers worry about unit heterogeneities? One way to answer the first question is to perform the sensitivity analysis to each of two types of treatment variables with a cross-sectional dataset, in which one of the treatments loses its statistical significance after including the fixed effects in panel data setting while the other does not. As seen in Columns 2 and 4 of Table 1, logged estimated % mountainous terrain loses its statistical significance once the regional fixed effects are added to the model, while inversed lagged per capita GDP does not. Thus, the artificial cross-sectional dataset
Figure 3: Sensitivity analysis to logged % mountainous terrain (left) and inversed lagged per capita GDP (right) with the artificial cross-sectional data. The baseline regression is presented in Column 3 of Table 1.

was constructed based on FL’s dataset by merging the observations in the year 1997 with the observations of all years that record the occurrence of civil war onset. The result of the baseline regression estimated with this dataset is shown in Column 3 of Table 1. The comparison of Column 3 with Column 4 shows that the statistical significance of logged estimated % mountainous terrain disappears once panel data is collected and the fixed effects are added. Equations (14) and (15) are assumed for the sensitivity analysis, and the sensitivity parameters are searched with respect to each of two treatment variables setting the target value at 1.96 in test statistics.

In Figure 3, the result using logged estimated % mountainous terrain as the treatment is displayed in the left panel and that using inversed lagged per capita GDP is displayed in the right panel. In the left panel, lagged per capita GDP (\textit{gdpenl}) is above the contour (displayed as blue hollow circles), which means an unobserved confounder as strong as lagged per capita GDP can make the coefficient of logged estimated % mountainous terrain statistically marginally insignificant. On the other hand, the right panel shows that the unobserved
confounder must be about twice as strong as Polity IV score of democracy ($\text{polity21}$) to change the coefficient of inversed lagged per capita GDP to statistically insignificant. This comparison indicates that the strength of the partial effects of the observe covariates relative to the contour can be a good barometer to determine whether researchers need to worry about the confounding due to unit heterogeneities. In other words, the absolute size of the partial effects of observed covariates are not necessarily informative. In fact, the partial correlations of the covariates in the right panel are slightly larger than the counterparts in the left panel.

**Sensitivity analysis 2: whether and how do the fixed effects make the model robust?** The second question is explored by comparing the result of the sensitivity analysis of the model with the fixed effects and the counterpart without the fixed effects using a treatment of which the statistical power does not change very much between the two models in order to set the same target value. According to Table 2, the coefficient of inversed lagged per capita GDP does not change very much by whether the fixed effects are included (Column 4) or not (Column 2). Thus, using FL’s original dataset, the sensitivity analysis of inversed lagged per capita GDP is performed with and without the regional and year fixed effects.$^{20}$

The sensitivity analysis for the model without the fixed effects is based on Equations (14) and (15), and that for the fixed effects models is set up as follows:

\[
E(\text{onset}) = \Lambda (\tau \text{lmtnest} + \theta \text{ethfrac} + \vartheta \text{gdpenl} + x \beta + R \beta_r + T \beta_t + \delta u) \tag{16}
\]
\[
E(\text{gdpenl}) = \lambda \text{ethfrac} + \omega \text{gdpenl} + x \gamma + R \gamma_r + T \gamma_t + \alpha u \tag{17}
\]

where $R$, $\beta_r$ and $\gamma_r$ are the regional fixed effects and their coefficients and $T$, $\beta_t$ and $\gamma_t$ are the year fixed effects and their coefficients.

$^{20}$Although inversed lagged per capita GDP is not time invariant, country fixed effects are not used because using country fixed effects drops more than half of the observations due to the rarity of civil war onset.
Figure 4: Sensitivity analysis with and without the fixed effects. The baseline regression coefficients are presented in Columns 2 and 4 in Table 1.

Figure 4 presents the comparison of the two sensitivity analysis. The contour drawn based on the models without the fixed effects are displayed as red hollow triangles, and the counterpart with the fixed effects are displayed as blue hollow circles. The partial correlations of the covariates of the models with and without the fixed effects are displayed as blue plus sign (+) and red cross sign (×) respectively. The counterfactual partial correlations of the fixed effects for the models that do not include the fixed effects are displayed as red hollow diamond (○).

The contour curve of the fixed effects models are slightly below its counterpart without the fixed effects. This reflects the fact that the test statistics of the treatment in Table 2 is smaller when it is estimated with the fixed effects model than when it is estimated without the fixed effects. However, the fixed effects models gain robustness through the improved balances of the covariates at the same time. That is, the most scatter plots of the covariates moved toward the left when the fixed effects are included in the models. Putting the plots of the fixed effects aside, the reduction of the partial correlations is particularly large for the powerful covariates in the figure such as Oil, ethfrac, and polity2l. Thus, although
the fixed effects are often powerful confounders themselves, they make the model robust to unobserved confounding by making the treatment less dependent on the covariates.\textsuperscript{21}

6 Application to Propensity Score Matching with the Data from the National Supported Work Experiment

Here I apply GSA to propensity score matching and demonstrate that GSA is useful to visualize whether and how much propensity score matching helps increase the robustness of the model.\textsuperscript{22}

Background. The National Supported Work Demonstration (NSW) was a social experiment, in which participants were randomly assigned to participate in a job training program. The effect of the program on yearly earnings in 1978 was estimated. LaLonde (1986) used the data from this evaluation to construct an observational study by merging the non-experimental control group from the Population Survey of Income Dynamics (PSID) with the experimental treatment group from NSW. He demonstrated that the sophisticated econometric techniques that were commonly used in the 1980’s could not recover the estimates obtained using the randomized experiment. Since then, the dataset has been used as a touchstone for new statistical methods that intend to remove selection bias in observational studies. Among them, Dehejia and Wahba (2002) revisited LaLonde’s (1986) study using propensity score matching and recovered the estimates that were close to those obtained using the randomized experiment. Their findings, however, hold only for a limited number of propensity score matching techniques with specific combinations of higher-order and interaction terms of the covariates. Thus, their study is still not conclusive proof that ignor-

\textsuperscript{21}Another important point is that the partial correlations of the fixed effects can be considered as another reference points in evaluating the strength of the confounding covariates because it is rare that a single covariate has the confounding effects stronger than those of the fixed effects, which are the total effects of multiple dummy variables.

\textsuperscript{22}See Ichino, Mealli, and Nannicini (2008) developed the sensitivity analysis specifically designed for matched observations based on Rosenbaum (2002)
ability was satisfied with propensity score matching, and it is still interesting to investigate the sensitivity of the results to unobserved confounders.

**Preliminary analysis.** In this exercise, I use the same datasets as those used in Imbens (2003), which are created from NSW’s male subsamples in which earnings in 1974 are non-missing, and the respondents of PSID from the same time period.\(^{23}\) These datasets are used again in Appendix A. The experimental dataset is solely created from NSW data, in which 185 people receive the treatment and 260 people are assigned to the control group. The unrestricted non-experimental dataset consists of the same treatment group in NSW and 2,490 respondents from PSID. The restricted non-experimental dataset excludes those who previously earned more than $5,000 in either 1974 or 1975 from the unrestricted dataset and contains 148 people in the treatment group and 242 people in the control group. The treatment effects are obtained from the regression of earnings in 1978 on the program participation with several pre-treatment covariates listed in the caption of Table 2. Regressions with these three datasets yield the estimated treatment effects displayed in Table 2. The last column shows the estimates from propensity score matching.

\(^{23}\)This study uses the data available at the following Professor Dehejia’s website: http://www.nber.org/~rdehejia/nswdata.html.

<table>
<thead>
<tr>
<th>Dependent Variable: Earning in 1978 ($,000)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job Training Program</strong></td>
<td>1.672 (0.634)</td>
<td>1.897 (1.120)</td>
<td>0.115 (1.007)</td>
<td>2.409 (0.946)</td>
</tr>
<tr>
<td><strong>Control Variables</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>445</td>
<td>390</td>
<td>2675</td>
<td>2675</td>
</tr>
</tbody>
</table>

**Sensitivity analysis**  First, the following models are assumed for the sensitivity analysis.

\[
E(\text{re78}) = \tau \text{treat} + x \beta + \delta u
\]  
(18)

\[
E(\text{treat}) = \Lambda (x \gamma + \alpha u)
\]  
(19)

where \text{re78} is annual earnings in 1978, \text{treat} is the treatment indicator, \(x\) is a vector of covariates listed in the caption of Table 2, \(u\) is an unobserved confounder, and \(\tau, \beta, \) and \(\gamma\) are coefficients, and \(\alpha\) and \(\delta\) are the sensitivity parameters. Following Imbens (2003), the target values are set at the treatment effect minus one.\(^{24}\) Thus, this sensitivity analysis provides the necessary confounding effect to reduce the treatment coefficient by one. The result produced with unrestricted, non-experimental dataset is compared with that with a pre-processed subset of the unrestricted, non-experimental dataset. Specifically, the observations in the treatment group is matched with those in the control group based on the propensity score using single nearest-neighbor matching without caliper.\(^{25}\)

Figure 5 presents the results, where the contour of the unrestricted dataset is displayed as red hollow triangles, and that of the matched dataset is displayed as blue hollow circles. The two sensitivity parameters, \(\alpha\) and \(\delta\), are transformed into the partial correlation with the outcome (in vertical axis) and that with the treatment (in horizontal axis). The partial correlations of the pretreatment variables in the unrestricted dataset are displayed as “\(x\)”, and the counterparts in the restricted dataset are displayed as “\(+\)”. If the matching balances the observed covariates in distribution, these covariates are expected to be uncorrelated with the treatment. Thus, these covariates will be plotted along the vertical axis without systematically affecting their correlation with the outcome. Without matching, the treatment assignment depends a lot on the covariates. In fact, the plot of the pretreatment variable, earnings in 1974 (\text{re74}), in the unrestricted dataset (shown

\(^{24}\)From Columns 3 and 4 in Table 2, the target values are \(\tau = 0.115 - 1 = -0.885\) for the unrestricted dataset and \(\tau = 2.409 - 1 = 1.409\) for the matched dataset.

\(^{25}\)The matching is performed with the Stata package, \texttt{psmatch2} (Leuven and Sianesi 2003), and the method used in this analysis is the default method of this package.
in “×”) overlaps with the contour. This means that an unobserved confounder which is as strong a confounder as earnings in 1974 would decrease the treatment effect by $1,000.

The balances, however, have noticeably improved after the matching. Most of the plots of the covariates moved to the left as indicated in “+”. This change combined with the shift of the contour to the upper right indicates that the model has become more robust to the unobserved confounding. Specifically, the effects of an unobserved confounder now needs to be approximately three times as strong as those of the most powerful observed confounder, unemployment in 1975 (unemp75). At the same time, some of the partial effects changed in non-systematic way (i.e. re75 moved downward, and unemp75 moved to the upper right). To sum up, although the propensity score matching does not improve the balances of unobserved covariates unless they are correlated with the observed covariates, it increases the balance of observed covariates significantly if not perfect.

Figure 5: Application of GSA to the propensity score matching with PSID unrestricted non-experimental dataset. See Column 4 in Table 2 for the baseline regression coefficient.
7 Application to an Instrumental Variable Approach

Background. In spite of its substantive importance, to find causal effects of institutions is one of the most difficult challenges in social science. In their influential but controversial paper, Acemoglu, Johnson and Robinson (2001) (AJR, henceforth) explore the detrimental effect of extractive institutional setting on economic development using the data about colonial countries.\(^{26}\) The obvious problem is the endogeneity of institutions as treatment variables because “rich economies choose or can afford better institutions” (ibid.), for instance. To resolve this issue, AJR use the mortality rates of European settlers more than 100 years ago as an instrumental variable (IV).

As AJR note, in order for this research strategy to work, the following three assumptions must be satisfied. First, the instrument (the mortality rates) is sufficiently correlated with the treatment (the institutional settings). Second, “the instrument is as good as randomly assigned” (Angrist and Pischke 2008) conditional on covariates. Finally, the instrument affects the outcome (logged per capita GDP in 1995) only through its correlation with the treatment.\(^{27}\) However, neither of these assumptions is easily testable with real data. For example, the possibility of weak instrument is known to be testable as both the instrument and the endogenous treatment are observed, but the test rests on the untestable assumption of instrument exogeneity.

Sensitivity analysis provides a simple alternative to evaluate the plausibility of IV estimates by characterizing an omitted exogenous unobservable as an unobserved confounder. Indeed, the effect of an endogenous treatment is unbiasedly estimated if such an unobserved confounder is explicitly controlled for (Angrist and Pischke 2008). Just like the other sensitivity analyses, the goal of this application is to evaluate the robustness of the IV estimates, and the proposed method achieves this goal by performing the sensitivity analysis to the denominator and the numerator of the Wald estimator for IV. Specifically, GSA is performed

\(^{26}\)See Albouy (2012) as an example of the criticism to AJR.

\(^{27}\)See Angrist, Imbens and Rubin (1996) for the formal definitions of these assumptions. SUTVA and monotonicity are also assumed.
to the denominator, or the first stage equation (i.e. the regression of the treatment on the instrument), and the numerator, or the reduced form equation (i.e. the regression of the outcome on the instrument). Unlike the other applications of GSA, the robustness must be demonstrated in both the first stage equation and the reduced form equation. Finally, this application is limited to a single endogenous treatment and a single instrument.

**Sensitivity analysis to the first stage equation.** The first stage equation regresses the treatment on the instrument and other observed covariates. The sensitivity analysis to the first stage equation is set up as follows:

\[
\text{avexpr} = \tau_1 \text{logem4} + x \beta_1 + \delta_1 u + \varepsilon_1
\]

\[
\text{logem4} = x \gamma_1 + \alpha_1 u + \upsilon_1
\]

(20)

(21)

where \text{avexpr} is the treatment (here, as a dependent variable), the average protection against expropriation risk from 1985 to 1995, \text{logem4} is the instrument (as a key independent variable), the logged European settler mortality, \(x\) is a set of the covariates, \(u\) is an unobserved confounder, \(\tau_1\) is the effect of the instrument on the treatment, \(\beta_1\) and \(\gamma_1\) are a vector of the coefficients, \(\alpha_1\) and \(\delta_1\) are sensitivity parameters, and \(\upsilon_1\) and \(\varepsilon_1\) are error terms.

In the Wald estimator, \(\tau_1\) is the denominator. Therefore, if the coefficient of the instrument (\(\tau_1\)) is estimated imprecisely due to an unobserved confounder, the IV estimate would be imprecise, too. The target values for GSA can be set at any value beyond which the researchers think the IV estimate is unreliable. Throughout this section, the target values are set at the half sizes of the coefficients of the key independent variables.

This means that the IV estimator is considered as unreliable if the unobserved confounder halves the partial coefficient of the instrument, or doubles the IV estimator, holding numerator constant.

---

28 This analysis uses the following four covariates: the absolute value of the latitude of capital divided by 90, Africa dummy, Asia dummy, and other continent dummy.

29 For Equation (20), \(\text{target}(\tau_1) = .292/2 = .146\) from the lower panel of Table 3.

30 The sensitivity analysis to the first stage equation also serves as the check for random assignment of the instrument. Although the instrument does not need to be uncorrelated with the observed covariates, if there...
<table>
<thead>
<tr>
<th>Simulation</th>
<th>1st stage (1)</th>
<th>Without IV (2)</th>
<th>Reduced form (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dep.var: T</td>
<td>Dep.var: Y</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>0.329</td>
<td>0.041</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.156)</td>
<td>(0.109)</td>
</tr>
<tr>
<td></td>
<td>0.231</td>
<td>0.130</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.034)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>X1</td>
<td>0.412</td>
<td>-0.382</td>
<td>-0.346</td>
</tr>
<tr>
<td></td>
<td>(0.084)</td>
<td>(0.126)</td>
<td>(0.109)</td>
</tr>
<tr>
<td></td>
<td>0.433</td>
<td>-0.498</td>
<td>-0.439</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.030)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>X2</td>
<td>0.416</td>
<td>-0.163</td>
<td>-0.181</td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.135)</td>
<td>(0.111)</td>
</tr>
<tr>
<td></td>
<td>0.373</td>
<td>-0.114</td>
<td>-0.063</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.030)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>X3</td>
<td>0.146</td>
<td>-0.045</td>
<td>-0.047</td>
</tr>
<tr>
<td></td>
<td>(0.073)</td>
<td>(0.101)</td>
<td>(0.096)</td>
</tr>
<tr>
<td></td>
<td>0.128</td>
<td>-0.175</td>
<td>-0.159</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.023)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>X4</td>
<td>0.265</td>
<td>-0.567</td>
<td>-0.572</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.104)</td>
<td>(0.092)</td>
</tr>
<tr>
<td></td>
<td>0.142</td>
<td>-0.507</td>
<td>-0.488</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.022)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.102</td>
<td>0.056</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>(0.073)</td>
<td>(0.099)</td>
<td>(0.095)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.021)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>N</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>R-sq</td>
<td>0.784</td>
<td>0.568</td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td>0.654</td>
<td>0.553</td>
<td>0.560</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AJR (2001)</th>
<th>1st stage (1)</th>
<th>Without IV (2)</th>
<th>Reduced form (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ave,protection against expropriation risk</td>
<td>Dep.var: ln(GDP per capita in 1995)</td>
<td></td>
</tr>
<tr>
<td>Inv. log(settler mortality)</td>
<td>0.292</td>
<td>0.565</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>(0.157)</td>
<td>(0.083)</td>
<td>(0.123)</td>
</tr>
<tr>
<td>Abs(latitude of capital)/90</td>
<td>0.182</td>
<td>0.111</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>(0.126)</td>
<td>(0.080)</td>
<td>(0.099)</td>
</tr>
<tr>
<td>Africa dummy</td>
<td>-0.087</td>
<td>-0.420</td>
<td>-0.345</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(0.081)</td>
<td>(0.109)</td>
</tr>
<tr>
<td>Asia dummy</td>
<td>0.113</td>
<td>-0.194</td>
<td>-0.176</td>
</tr>
<tr>
<td></td>
<td>(0.120)</td>
<td>(0.078)</td>
<td>(0.094)</td>
</tr>
<tr>
<td>Other continent dummy</td>
<td>0.154</td>
<td>0.022</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>(0.122)</td>
<td>(0.078)</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.000</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.070)</td>
<td>(0.084)</td>
</tr>
<tr>
<td>N</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>R-sq</td>
<td>0.328</td>
<td>0.714</td>
<td>0.584</td>
</tr>
</tbody>
</table>

Table 3: Estimated baseline coefficients of Acemoglu, Johnson and Robinson (2001). OLS. Standard errors in parentheses. Quantities of interest are indicated in bold face. All variables are standardized. Upper panel: the simulation dataset. Lower panel: the dataset by AJR.
In order to know how the sensitivity analysis would look like if all IV assumptions hold, the sensitivity analysis is conducted on simulation data as well. The simulation dataset mimics the structure of the dataset of AJR. Specifically, the dataset is generated according to the following data generating process.

\[
X = (X_1, X_2, X_3, X_4) \sim MVN(0, \Sigma), \quad \Sigma = \begin{pmatrix}
1 & .6 & .3 & 0 \\
.6 & 1 & .3 & 0 \\
.3 & .3 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\] (22)

\[
Z \sim Normal(1, 0)
\] (23)

\[
U \sim Normal(1, 0), \quad V \sim Normal(1, 0), \quad E \sim Normal(1, 0)
\] (24)

\[
T = Z + 1.5X_1 + 1.5X_2 + .5X_3 + .5X_4 + U + 2E
\] (25)

\[
Y = T - 1.5X_1 - .5X_2 - .5X_3 - 1.5X_4 + U - E + V
\] (26)

where \(X_1, X_2, X_3, X_4\) are observed covariates, \(Z\) is the instrument, \(U, V,\) and \(E\) are endogenous unobserved confounders, \(T\) is the treatment, and \(Y\) is the outcome. All observed variables are standardized to zero mean and unit variance to make the baseline regression coefficients similarly sized. Although the dataset of AJR includes only 64 countries, the sensitivity analysis is performed to the simulation dataset in the small sample (\(N = 64\)) setting as well as the large sample (\(N = 1000\)) setting because of the consistency of the IV estimators. The upper panel of Table 3 reports the results of the regressions with this simulation dataset.

The results of the sensitivity analysis to the simulation dataset displayed in the left panel of Figure 6 are consistent with the properties of IV estimators.\(^{31}\) To look at the large is a non-negligible correlation between them, the instrument is unlikely to be randomly assigned conditional on the observed covariates with a few exceptions (such as a randomized block experiment in which observed covariates represent the blocks and a randomized encouragement is used as an instrument). Thus, if the instrument satisfies this assumption, the partial correlation between the instrument and observed covariates would be quite small.

\(^{31}\)The regression estimates of the first stage equation with the simulation dataset are reported in the
sample result first, the covariates indicated as L1, L2, L3 and L4, are plotted along the vertical axis, which supports the “as good as random” assignment of the instrument. Moreover, the contour displayed as blue hollow circles is located far from any observed covariates, which makes it very unlikely for any confounding covariates to halve the instrument coefficient. Thus, the instrument that satisfies all IV assumptions in the large sample setting produces a reliable estimate. However, because IV estimators are consistent estimators, the result of the sensitivity analysis looks less supportive in the small sample setting. The observed covariates indicated as S1, S2, S3 and S4 have some positive correlations with the exogenously generated instrument, which makes it somewhat more likely for unobserved confounders to halve the coefficient of the instrument.

The right panel of Figure 6 shows the result of the sensitivity analysis using the dataset of AJR (N = 64). Although the number of observation of the dataset of AJR is only 64 and the mortality rates are non-random instrument, the covariates displayed as 1, 2, 3 and 4 are all located below the contour, which shows some robustness for the denominator of the upper panel of Table 3, and the target values are set at .1645 for the small sample setting and .1155 for the large sample setting according to Columns 1 and 2.

Figure 6: Application of GSA to the first stage equation. Left panel: simulation dataset. Right panel: the dataset of AJR. See the upper panel of Table 3 for their baseline regression coefficients.
Wald estimator. However, it must be kept in mind that these covariates have moderately strong correlations with the instrument, which violates one of the IV assumptions of “as good as random” assignment of the instrument.

**Sensitivity analysis to the reduced form equation.** The reduced-form equation regresses the outcome on the instrument and other observed covariates, and the sensitivity analysis is set up as follows.

\[
\log p_{gp95} = \tau_2 \log em4 + x_2 + \delta_2 u + \varepsilon_2 
\]

\[
\log em4 = x_2 \gamma_2 + \alpha_2 u + \nu_2
\]

where \( \log p_{gp95} \) is the outcome, logged GDP per capita in 1995, \( \tau_2 \) is the coefficient of the instrument on the outcome, \( \beta_2 \) and \( \gamma_2 \) are a vector of the coefficients, \( \alpha_2 \) and \( \delta_2 \) are sensitivity parameters, and \( \varepsilon_2 \) and \( \nu_2 \) are error terms.

In the Wald estimator, \( \tau_2 \) is the numerator. Therefore, if \( \tau_2 \) is not robust against the omitted exogenous unobservable, the Wald estimate will be unreliable. The result of the sensitivity analysis is presented in the left panel of Figure 7.\(^{32}\) In the figure, the sensitivity analysis is performed in the small sample setting (displayed as green triangles) and the large sample settings (displayed as blue hollow circles). For comparison, the sensitivity analysis is also performed to the naive outcome model (i.e. the regression of the outcome on the un-instrumented treatment; displayed as red crosses).

In the large sample setting, the partial correlations of the covariates (LZ1, LZ2, LZ3 and LZ4) are plotted near the vertical axis. This makes it very unlikely that any unobserved confounder has confounding effects as strong as those displayed as blue hollow circles. In the small sample setting, the covariates (SZ1, SZ2, SZ3 and SZ4) have weak correlations with the instrument, which makes the IV estimates somewhat unreliable. On the other

---

\(^{32}\)The regression estimates of the reduced-form equation with the simulation dataset are reported in the upper panel of Table 3, and the target values are set at .0935 for the small sample setting and .061 for the large sample setting according to Columns 5 and 6.
hand, the result of the sensitivity analysis to the naive outcome model show that most of the partial correlations of the covariates (LT1, LT2, LT3 and LT4) are plotted above the contour displayed as red crosses, which indicates that the naive estimate is not robust against omitted confounders.

Furthermore, the comparison of the partial correlations of the observed covariates between the reduced form equation and the naive outcome model in the large sample setting also highlights interesting commonality between this application and the applications to other quasi-experimental techniques. Specifically, each plot of the partial correlations shifts to the left between the naive outcome model and the reduced form equation (e.g. LZ1 ← LT1). Thus, the result implies that introduction of an instrument makes the IV estimator more robust against unobserved confounders by making the treatment more independent of other covariates.

The right panel of Figure 7 presents the results of the sensitivity analysis with the dataset of AJR. According to Columns 2 and 3 of Table 3, the target values are set at .227 for the reduced form equation (displayed as blue hollow circles) and .2825 for the naive

Figure 7: Application of GSA to the reduced form equation. Left panel: simulation dataset. Right panel: the dataset of AJR. See the lower panel of Table 3 for their baseline regression coefficients.
outcome model (displayed as red crosses). The partial correlations of the covariates in the reduced-form equation (Z1, Z2, Z3 and Z4) are at least as large as those in the naive outcome model (T1, T2, T3 and T4). Furthermore, Africa dummy (Z2) is plotted near the contour displayed as blue hollow circles. This implies that an unobserved confounder as strong as those of Africa dummy can halve the coefficient of the instrument. Given that some of the covariates appear in AJR such as British colonial dummy have stronger partial effects than Africa dummy, this is fairly likely. Thus, the introduction of the settler’s mortality rate as IV does not seem to improve the robustness of the numerator of the Wald estimator against the unobserved confounding.

To sum up, when robustness is defined in terms of the possibility that the Wald estimator either doubles or is halved, the estimation strategy by AJR is not supported as a way to produce a reliable IV estimator because of the lack of robustness in the numerator of the Wald estimate.

8 Concluding remarks

In this paper, I have introduced GSA and demonstrated its versatility and simplicity of use through its application to various quasi-experimental methods. As is the case with the development of matching techniques, various sensitivity analyses have been proposed, and each of them has unique strength. I believe GSA should be added as one of such methods if not dominates other methods. Above all, GSA allows researchers to set the target value in terms of test statistics and to assume any combination of the link functions requiring minimal changes to the researchers’ original estimation models.

The applications of GSA to fixed effects model, propensity score matching, and instrumental variable approach show that successful applications of quasi-experimental techniques make the treatment assignment less dependent on the covariates as well as unobserved confounders. Unlike regression tables or conventional tests, GSA visualizes this process by showing the reduction in the partial correlation between the treatment and the covariates.
relative to the necessary confounding to change the treatment effect by a given amount.

Finally, the sensitivity analysis is still a developing field and requires further research particularly in development of reliable benchmark, which could be improved in two aspects. First, the researchers’ substantive knowledge about the treatment effect needs to be incorporated into the sensitivity analysis in evaluating the contour or in setting the target value. This would be particularly useful in some research fields where a certain size of the treatment effect is widely known such as economic return to education. Second, the objective criteria to evaluate the contour needs to be developed. Inventing a new test statistics would be challenging but promising effort.

References


University, of Michigan, National Science Foundation National Institute of Aging Institute of Child Health & Human Development. 1978. *Panel Study of Income Dynamics, public use dataset*. University of Michigan, Ann Arbor, MI.

A Checking the performance of GSA through Monte Carlo simulation

In this section, statistical properties of GSA are explored through Monte Carlo simulations. We are particularly interested in how well GSA recovers the true treatment effect when researchers know true sensitivity parameters. As a basis of comparison, the results obtained from Imbens’ (2003) sensitivity analysis (henceforth Imbens’ SA) are used. This method has been widely used as a forerunner of dual parameter sensitivity analysis.\(^{33}\)

To overview the results, GSA yields less biased treatment effects when the sensitivity parameter of the treatment model is extreme though both GSA and Imbens’ SA unbiasedly estimate the true treatment effects, otherwise. When the sensitivity parameter of the outcome model is extreme, Imbens’ SA is slightly more efficient in estimating the population-level treatment effects than GSA, while GSA yields more efficient sample-level treatment effects than Imbens’ SA. The two methods particularly disagree in the estimation of standard errors (SEs). Specifically, GSA yields the SE estimates when a true unobserved confounder \((U)\) is included in the model, while Imbens’ SA yields the SE estimates when a pseudo unobservable \((PU)\) is included in the model. I will discuss later why the former quantities are more important for the sensitivity analysis.

First, let us overview how the two methods are different using LaLonde’s (1986) experimental dataset. In Figure 8, the horizontal and vertical axes represent \(\alpha\) and \(\delta\), which are raw coefficients of PUs. The estimated contours from Imbens’ SA are displayed as red solid line, and the scatter plots from GSA are displayed as blue hollow circles. The left figure, which shows the sets of the sensitivity parameters that lower the treatment effect by $1,000, indicates that the scatter plots obtained from GSA overlap with the red contour curve obtained from Imbens’ SA. That is, both approaches identify the same (and correct) confounding effects that change the treatment effects by a given amount.

On the other hand, the right figure, which shows the sets of the sensitivity parameters

\(^{33}\text{See Clarke (2005) and Blattman (2009) for the early application of Imbens’ SA.}\)
that lower the t-value of the treatment effect from 2.606 to 1.645 (i.e. 10% significance level), shows that two methods identify different sets of sensitivity parameters. Specifically, GSA provides larger sensitivity parameters than Imbens’ SA in the upper left, and smaller ones in the lower right. In the following, I will present the results of Monte Carlo simulations and discuss the practical considerations regarding this discrepancy.

**Simulation design.** GSA and Imbens’ SA are compared through Monte Carlo simulation. First, the dataset is generated following Imbens’ (2003) original setting. That is, the identity link function is assumed for the outcome model, the logistic link function is assumed for the treatment model, and a binary unobserved confounder is assumed. The data generating
process (DGP) is:

\[ X_1, X_2, X_3 \sim \text{Normal}(0, 1) \]
\[ U \sim \text{Binomial}(1, 0.5) \]
\[ Z^* = \Lambda(0.4X_1 + 0.3X_2 + 0.2X_3 + \alpha(U - 0.5)) \]
\[ Z \sim \text{Binomial}(1, Z^*) \]
\[ \epsilon \sim \text{Normal}(0, 1) \]
\[ Y_0 = 0 + 0.5X_1 + X_2 + 1.5X_3 + \delta U + \epsilon \]
\[ Y_1 = 2 + 0.5X_1 + X_2 + 1.5X_3 + \delta U + \epsilon \]
\[ Y = Y_0 \cdot (1 - Z) + Y_1 \cdot Z \]

where \( X_1, X_2 \) and \( X_3 \) are observed covariates, \( U \) is a binary true unobserved confounder,\(^{34}\) \( Z^* \) is a propensity score, \( \Lambda \) is inverse logit function, \( Z \) is a binary treatment variable, \( \epsilon \) is a stochastic disturbance term, \( Y_0 \) and \( Y_1 \) are the potential outcome, \( Y \) is an outcome variable, and \( \alpha \) and \( \delta \) are sensitivity parameters. \( \alpha \) and \( \delta \) take the value of either 0, 2, or 4 respectively. Thus, the simulations are performed under nine different confounding effects.

The true estimation models for the outcome and the treatment models in a given dataset are stated as follows.

\[ E(y|z, x_1, x_2, x_3, u) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\tau}z + \hat{\delta}u \]  
\[ E(z|x_1, x_2, x_3, u) = \Lambda(\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 + \hat{\gamma}_3 x_3 + \hat{\alpha}u) \]

Second, GSA and Imbens’ SA are performed to the dataset. With researchers’ knowledge about true sensitivity parameters, GSA attempts to recover \( u \) and the treatment effects.

\(^{34}\)Note that we must use a binary \( U \) because Imbens’ SA cannot deal with a continuous \( U \). However, the use of a binary \( U \) puts GSA at a disadvantage because GSA’s orthogonality condition between \( X \) and \( U \) is better satisfied with a continuous \( U \). In other words, the estimated treatment coefficient and standard error obtained using a binary \( U \) are less precise than those obtained using a continuous \( U \) due to empirical correlation between a binary \( U \) and \( X \).
with the following models, which simply replace $u$ with $pu$.\(^{35}\)

\[
E(y|z, x_1, x_2, x_3, pu) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3 + \tilde{\tau} z + \tilde{\delta} pu
\]  
\[\text{(39)}\]

\[
E(z|x_1, x_2, x_3, pu) = \Lambda(\tilde{\gamma}_0 + \tilde{\gamma}_1 x_1 + \tilde{\gamma}_2 x_2 + \tilde{\gamma}_3 x_3 + \tilde{\alpha} pu)
\]  
\[\text{(40)}\]

Imbens’ SA yields the treatment effect by maximizing the likelihood function of the above models and by integrating the marginal probability of $U$ in the function. Note that we can suppose two different situations of what researchers know about the “true” sensitivity parameters. In one situation, researchers know the population-level confounding effects of $U$, or the coefficients of $U$ used in DGP. In the other situation, researchers know the confounding effects of $U$ in a given dataset, or $\hat{\alpha}$ and $\hat{\delta}$ in Equations (37) and (38). Although all simulations are performed under these two different assumptions, they do not make notable differences. Thus, I present the results obtained with population-level sensitivity parameters when the statistics represent population-level properties and those obtained with sample-level sensitivity parameters when the statistics represent sample-level properties.

Third, the simulation is iterated 1,000 times for each of 9 distinct specifications of the sensitivity parameters. The performances of GSA and Imbens’ SA are evaluated according to the following two criteria. The first criterion is how accurately sensitivity analysis recovers population parameters of the treatment. The root-mean-square error (RMSE) of the estimated coefficient with respect to $\tau$ \(\equiv \sqrt{1/1000 \sum_{i=1}^{1000} (\tilde{\tau}_i - \tau)^2}\) and the RMSE of the estimated standard errors with respect to the standard deviation of the estimated coefficients \(\equiv \sqrt{1/1000 \sum_{i=1}^{1000} (SE(\tilde{\tau}_i) - SD(\hat{\tau}))^2}\) are used as benchmark combined with the sensitivity parameters at population level. The other criterion is how accurately sensitivity analysis recovers the estimates that would be obtained if a true $U$ were observed. With this criterion, the root-mean-square deviations (RMSDs) between the estimated coefficient and $\hat{\tau}$ \(\equiv \sqrt{1/1000 \sum_{i=1}^{1000} (\tilde{\tau}_i - \hat{\tau}_i)^2}\) and the estimated standard errors against $SE(\hat{\tau})$.

\(^{35}\)A wrapper program of GSA that enables users to set $\alpha$ and $\delta$ is developed to run this simulation.
are used as benchmark combined with the sensitivity parameters at sample level.

Simulation results. The simulation results are presented from Table 4 to Table 9. For each distinct combination of the sensitivity parameters, the results of GSA are presented in the left column and those of Imbens’ SA are presented in the right column. First, let us take a look at the basic statistics. For these results, the population-level sensitivity parameters are assumed to be known. That is, $\alpha$ and $\delta$ in the tables are used as-is. Table 4 shows the mean of the estimated treatment effect over 1,000 iterations. Both methods slightly overestimate the treatment coefficients when the sensitivity parameter in the treatment model ($\alpha$) is set at an extremely large value (i.e. $\alpha = 4$), but the bias of the GSA is smaller than that of Imbens’ SA. Imbens’ SA yields the treatment effect estimates for a wide range of the sensitivity parameters even when GSA fails to generate PUs. This allows researchers to obtain extended contours, while the estimates obtained in such cases are typically erroneous as indicated by the simulation results. Both methods, otherwise, unbiasedly estimate the treatment effects as long as correct sensitivity parameters are provided.

On the other hand, the two methods return quite different estimates for the standard errors (SEs). In a simple regression of $Y$ on $Z$ and $U$, the SE of the treatment coefficient is expressed as $SE(Z) = \sigma / \left( |Z| \sqrt{1 - corr(Z,U)} \right)$, where $\sigma$ is a residual variance and $corr(Z,U)$ is the correlation between $Z$ and $U$. The larger $\alpha$ implies the larger $corr(Z,U)$ and the larger SE estimates. Table 5 shows that as $\alpha$ becomes larger the estimated SEs obtained from GSA becomes larger, while the counterparts for Imbens’ SA becomes smaller. The SE estimates of Imbens’ SA change this way because they represent the dispersion of $\tilde{\tau}$ in the sensitivity analysis not the dispersion of $\hat{\tau}$ in the true model. The difference between these two dispersions reflect the fact that, as $\alpha$ becomes larger, the more limited numbers

---

36 The average SEs are calculated as the square root of the average of the squared standard errors over 1,000 iterations.
Table 4: Mean of the estimated treatment effects. $\tau = 2$. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at population level (i.e. $\alpha$ and $\delta$ specified above are used as the sensitivity parameters as-is).

<table>
<thead>
<tr>
<th>Mean of $\bar{\tau}$</th>
<th>0</th>
<th>$\alpha$</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2.004</td>
<td>2.004</td>
<td>2.001</td>
<td>2.001</td>
</tr>
<tr>
<td>2</td>
<td>2.006</td>
<td>2.006</td>
<td>2.000</td>
<td>2.000</td>
</tr>
<tr>
<td>4</td>
<td>2.008</td>
<td>2.005</td>
<td>1.999</td>
<td>1.999</td>
</tr>
</tbody>
</table>

Table 5: Average of the standard error estimates for the treatment effect. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at population level.

<table>
<thead>
<tr>
<th>Average SE($\bar{\tau}$)</th>
<th>0</th>
<th>$\alpha$</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.131</td>
<td>0.131</td>
<td>0.145</td>
<td>0.129</td>
</tr>
<tr>
<td>2</td>
<td>0.131</td>
<td>0.146</td>
<td>0.145</td>
<td>0.141</td>
</tr>
<tr>
<td>4</td>
<td>0.131</td>
<td>0.176</td>
<td>0.145</td>
<td>0.166</td>
</tr>
</tbody>
</table>

of the candidates for PU becomes available,37 while $U$ is independent and identically distributed. As I discuss later, this makes Imbens’ SA unsuitable for using in combination with the test statistics.

Now, I discuss the results of the benchmark statistics. Tables 6 and 7 report the population-level properties, and the sensitivity parameters are set at the population level. RMSEs of $\bar{\tau}$ with respect to $\tau$ are shown in Table 6. The smaller RMSE indicates that the estimates are closer to the population-level treatment effect. The table shows that GSA yields slightly larger RMSEs when $\delta$ is extreme. This indicates that Imbens’ SA is more efficient in estimating the population-level treatment effects than GSA, but later I will show that the relationship between GSA and Imbens’ SA is reversed at the sample level. However, both GSA and Imbens’ SA yield similar RMSEs in most cases.

Table 7 reports RMSEs of $SE(\bar{\tau})$ with respect to $SD(\bar{\tau})$. The smaller RMSEs indicate that the method is good at predicting the dispersion of the treatment effect estimates in this

37In Imbens’ SA, this means that the more limited range of the treatment effect estimates is available for a given set of $\alpha$ and $\delta$ as $\alpha$ becomes larger.
Table 6: RMSE of the treatment effect estimates with respect to the true treatment effect. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at population level.

<table>
<thead>
<tr>
<th>δ</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.127</td>
<td>0.127</td>
<td>0.128</td>
<td>0.128</td>
<td>0.124</td>
<td>0.125</td>
</tr>
<tr>
<td>2</td>
<td>0.145</td>
<td>0.144</td>
<td>0.140</td>
<td>0.139</td>
<td>0.131</td>
<td>0.131</td>
</tr>
<tr>
<td>4</td>
<td>0.185</td>
<td>0.176</td>
<td>0.172</td>
<td>0.163</td>
<td>0.150</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Table 7: RMSE of the SE estimates of the treatment effect with respect to the SD of the treatment effect estimates. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at population level.

<table>
<thead>
<tr>
<th>δ</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.005</td>
<td>0.005</td>
<td>0.018</td>
<td>0.004</td>
<td>0.067</td>
<td>0.004</td>
</tr>
<tr>
<td>2</td>
<td>0.014</td>
<td>0.004</td>
<td>0.007</td>
<td>0.004</td>
<td>0.060</td>
<td>0.005</td>
</tr>
<tr>
<td>4</td>
<td>0.054</td>
<td>0.005</td>
<td>0.028</td>
<td>0.005</td>
<td>0.042</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Simulation. Unlike Table 6, Table 7 shows that RMSEs of Imbens’ SA are much smaller than the counterparts of GSA. This, however, does not mean GSA yields nonsensical estimates. I show later that GSA’s SE estimates represent the dispersion of the treatment effect estimates when true Us are observed.

Tables 8 and 9 are similar to Tables 6 and 7, but the sensitivity parameters are set at the sample level. That is, estimated coefficients of \( U (\hat{\delta} \text{ in Equation (37)} \) and \( \hat{\alpha} \text{ in Equation (38)}) \) in a dataset corresponding to one of nine combinations of \( \alpha \) and \( \delta \) are used in DGP. The following two tables also report the statistics that measure its sample-level performance.

Table 8 shows RMSDs between \( \bar{\tau} \) and \( \hat{\tau} \). This quantity represents the average deviation of the treatment effect estimates from the counterparts that would be obtained if true Us were observed. Both methods yield small RMSDs when the sensitivity parameter in the outcome model (\( \delta \)) is not extreme, but Imbens’ SA yields large RMSDs when \( \delta \) is extreme. In contrast to Table 6, this indicates that GSA estimates the sample-level treatment effects more efficiently than Imbens’ SA. In general, Tables 6 and 8 indicates that two methods

42
Table 8: RMSD between the treatment effect estimates and the treatment effects in the true model. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at sample level (i.e. coefficients of a true U in a given dataset generated with the α and δ specified above are used as the sensitivity parameters).

<table>
<thead>
<tr>
<th>δ</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.001</td>
<td>0.000</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>0.002</td>
<td>0.006</td>
<td>0.009</td>
<td>0.009</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>4</td>
<td>0.004</td>
<td>0.051</td>
<td>0.016</td>
<td>0.045</td>
<td>0.031</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 9: RMSD between the SE estimates of the treatment effect and the SE of the treatment effect in the true model. Summary of 1000 simulations. Left: GSA, right: Imbens’ SA. Sensitivity parameters are set at sample level.

<table>
<thead>
<tr>
<th>δ</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
<th>GSA</th>
<th>Imbens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.016</td>
<td>0.003</td>
<td>0.065</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.015</td>
<td>0.000</td>
<td>0.005</td>
<td>0.003</td>
<td>0.058</td>
</tr>
<tr>
<td>4</td>
<td>0.001</td>
<td>0.045</td>
<td>0.001</td>
<td>0.021</td>
<td>0.003</td>
<td>0.043</td>
</tr>
</tbody>
</table>

estimate the treatment effects with similar precision in most cases. An important implication is that whether the impact of sensitivity parameters is reflected in the models through regression (as in GSA) or through integration of U’s marginal probability (as in Imbens’ SA) does not affect the size of the dispersion of their treatment effect estimates.

Table 9 reports RMSDs between $SE(\hat{\tau})$ and $SE(\tilde{\tau})$. This quantity represents the average deviation of the SE estimates of the treatment effect from the counterparts _that would be obtained if true Us were observed_. Thus, RMSDs should be small if sensitivity analysis recovers true Us precisely. The table shows that RMSDs of GSA are close to 0 for any combination of the sensitivity parameters, while Imbens’ SA yields large RMSDs. This indicates that GSA estimates the dispersion of $\hat{\tau}$ rather than that of $\tilde{\tau}$.

To summarize the results, when researchers are interested in the confounding effect of $U$ on the treatment coefficient, the two methods provide comparable and unbiased estimates in most cases. Only when the sensitivity parameter of the outcome model is extreme, Imbens’
SA is better at estimating the population-level treatment effects, while GSA is better at estimating the sample-level treatment effects. Also, when the sensitivity parameter of the treatment model is extreme, GSA yields less biased estimates. When neither of the sensitivity parameters is extreme, the two methods yield comparable estimates. Then, Imbens’ SA is computationally more efficient.

When researchers are interested in the confounding effect of $U$ on the test statistics of the treatment, I argue that GSA is more appropriate than Imbens’ SA. In short, what researchers want to know is the impact of $U$ not PU. The SE estimates of GSA approximate the SEs when the models include true $U$s, while the SE estimates of Imbens’ SA are precise but represent the dispersion of the estimates when the models include PUs. As we have seen in Table 5, $U$ and PU have different properties regarding their impacts on SE estimates. Thus, researchers should use GSA when test statistics are used as the target value. Finally, it must be noted that GSA produces more precise estimates when $U$ is assumed to be continuous than when $U$ is binary. A binary $U$ is assumed in this simulation because Imbens’ SA cannot deal with continuous $U$.

B Procedure of Generalized Sensitivity Analysis

This section describes the procedure of GSA in more detail. Figure 1 presents the overview. For the sake of clarity, I consider the following linear models, but GSA is compatible with the models with various link functions.

\begin{align*}
Z_i &= \gamma'X_i + \alpha U_i + \epsilon_i \\
Y_i &= \tau Z_i + \beta'X_i + \delta U_i + \epsilon_i
\end{align*}

(41) (42)

where $Y_i$ is an outcome variable, $Z_i$ is a treatment variable, $X_i$ is a vector of observed covariates, $U_i$ is an unobserved covariate which is orthogonal to $X_i$, $\alpha$, $\beta$, $\gamma$ and $\delta$ are coefficients, and $\epsilon_i$ and $\epsilon_i$ are error terms. GSA also assumes stable unit treatment value assumption,
which requires the independence of the treatment assignment process from individual-level treatment status. In sensitivity analysis, researchers are interested in the substantive size and/or statistical significance of the treatment effect, as well as the coefficients of the unobserved confounder, $\alpha$ and $\delta$, that change the sizes of the treatment effect to some target value.

1. **Define the quantity of interest, decide its target size, and save the partial effects of the original models.** The quantity of interest is usually defined in terms of either test statistics or coefficient of the treatment variable. If researchers are concerned about the size of confounding due to an unobserved confounder that turns a positive and statistically significant treatment effect into an insignificant one, they may want to define their quantity of interest as the t-statistics of the treatment effect and set the target value at one of the critical values (e.g. 1.645 or 1.96) or a value slightly smaller than that. On the other hand, if researchers are concerned about the confounding effect on the substantive size of the treatment effect, the quantity of interest should be defined in terms of the coefficient of the treatment variable and set the target value accordingly.

   Because the coefficients of an unobserved confounder, or the two sensitivity parameters, $\alpha$ and $\delta$, have no substantive meaning per se, they are often transformed into other forms of partial effects. For instance, Imbens (2003) suggests converting $\alpha$ and $\delta$ to the partial R-squares of an unobserved confounder in the treatment assignment equation ($R^2_{Z,\text{par}}$) and the partial R-square in the outcome equation ($R^2_{Y,\text{par}}$) respectively. To calculate the partial R-squares, the R-squares in the original models ($R^2_{Z,\alpha}$ and $R^2_{Y,\alpha}$) must be recorded beforehand. This process is not necessary if researchers convert $\alpha$ and $\delta$ to partial correlations.

2. **Generate a candidate variable for a PU.** Next, a pseudo unobservable (PU) is generated through the following process, and this is the most important part of GSA. First, an initial intermediate unobservable, $\hat{U}$, is generated from a weighted average of residuals.
from treatment and outcome models plus random noise:

\[
\hat{U} \sim \mathcal{N}(c_1 \cdot d_z + c_2 \cdot d_y, 1)
\]  

(43)

where \( \hat{U} \) is an intermediate random variable drawn from the normal distribution with the mean of \( c_1 \cdot d_z + c_2 \cdot d_y \) and a unit variance, \( c_1 \) and \( c_2 \) are the key location parameters, and \( d_y \) and \( d_z \) are the standardized deviance residuals from Equations (10) and (11).\(^{38}\) Then, a second intermediate variable, \( \bar{U} \), is obtained by residualizing the first intermediate variable by \( X \):

\[
\bar{U} = \hat{U} - \hat{\lambda}X, \text{ where } \hat{\lambda} = \arg\min_{\lambda} (\hat{U} - \lambda'X)
\]  

(44)

where \( \lambda \) is a vector of regression coefficients. Then, the second intermediate random variable, \( \bar{U} \), is adopted as a candidate variable, \( \hat{U} \), only if it is approximately orthogonal to the observed covariates (Equation (7)).\(^{39}\) When a PU is assumed to be a binary, the third intermediate variable is created from the second intermediate variable, \( \bar{U} \), by dichotomizing \( \bar{U} \) at its median. Then, whether this binary third intermediate variable satisfies the orthogonality condition is checked. Because this process requires a greater number of iterations particularly with a small number of observations, the use of a binary PU should be avoided unless researchers have a strong reason to do so.\(^{40}\)

As seen in the lower left plots displayed as red hollow circles in Figure 9, a candidate variable generated from a purely random variable can take only a limited range of \( \alpha \) and \( \delta \). Thus, GSA indirectly adjusts the sizes of two sensitivity parameters, \( \alpha \) and \( \delta \) through the manipulation of the two location parameters, \( c_1 \) and \( c_2 \). Because \( d_z \) is associated with the unexplained variations in the treatment model, the larger the \( c_1 \), the larger the \( \alpha \) as displayed as blue hollow diamonds in the lower right of Figure 9. Likewise, the larger the \( c_2 \),

\(^{38}\)See p.619 of Stata Base Reference for the list of the deviation residuals of different families.

\(^{39}\)A reasonable error tolerance level for the orthogonality condition will be discussed later.

\(^{40}\)For example, researchers may have a concrete idea about a binary unobserved confounder such as a gene. Also, a binary PU needs to be assumed to compare the results of GSA with those of Imbens’ (2003) approach.
the larger the $\delta$ (in yellow hollow triangles in the lower right of Figure 9).

3. **Adopt a candidate variable, $\hat{U}$, as a PU if the size of the estimand is close enough to the target value. Otherwise, repeat to Step 2.** Model (8) is estimated replacing $U$ with a candidate variable, $\hat{U}$, and $\hat{U}$ is adopted as a PU, $U$, if the estimand is close enough to the target value. For example, when the target value is $t = 1.96$ and the error tolerance level is 1%, a candidate variable is accepted as a PU if the test statistics of the model with the candidate variable is $1.94 < t < 1.98$.\footnote{An optimal error tolerance level is discussed later in this section.} If the difference between the estimand and the target value is more than the error tolerance level, this candidate variable is discarded, and it will be regenerated until a PU is found. The iteration algorithm is discussed in detail in Appendix C.

**Note: Two screening processes in making PUs.** Step 2 and Step 3 involves the two screening processes that make PUs closer to what they are theoretically assumed. The first mechanism is the orthogonalization. That is, a second intermediate random variable, $\tilde{U}$, is adopted as a candidate variable, $\hat{U}$, only if it satisfies the orthogonality condition in Equation (7). Of course, with real data, a random variable can never be perfectly orthogonal...
to covariates. Thus, we must set a reasonable error tolerance level that is sufficiently strict so that the violation of the assumption in Equation (7) is negligible yet not too strict to generate candidate variables becomes nearly impossible.

Specifically, GSA checks the orthogonality of an intermediate random variable in the following way. First, the treatment model is estimated including a second intermediate random variable ($\tilde{U}$) in the right hand side:

$$E(Z) = g_{\tilde{z}}^{-1}(X\tilde{\gamma} + \tilde{\alpha}\tilde{U}) \tag{45}$$

where $\tilde{\gamma}$ and $\tilde{\alpha}$ are the estimates obtained from Equation (45). If $\tilde{U}$ is orthogonal to $X$, the coefficients of $X$ obtained from Equation (45), $\tilde{\gamma}$, should not be different from those obtained from the model without $U$ (i.e. $\gamma$ in Equation (11)). Thus, the correlation between the vector of regression values of $X$ in Equation (45), $X\tilde{\gamma}$, and the counterpart in Equation (11), $X\gamma$, should be close to unity. Thus, the first check is set up as follows:\footnote{$\tilde{U}$ must be standardized}

$$\hat{U} = \tilde{U} \text{ if } \rho(X\gamma, X\tilde{\gamma}) \geq \rho_{\text{threshold}} \tag{46}$$

where $\rho$ is the correlation coefficient and $\rho_{\text{threshold}}$ is a threshold level of the correlation coefficient (e.g. 0.99).

The second mechanism is to accept a candidate variables as a PU only if its confounding effect changes the treatment effect close to the target value. Formally,

$$PU = \hat{U} \text{ if } ||qoi - qoi_{\text{target}}|| < \varepsilon_{\text{threshold}} \tag{47}$$

where $qoi$ is a quantity of interest such as treatment effect coefficient or t-statistics and $\varepsilon_{\text{threshold}}$ is an error tolerance level (e.g. 0.01).

A simple simulation was performed to examine the extent to which these screening
mechanisms affect the shape of the contour plot. The simulation dataset is created in such a way that the treatment assignment is highly unbalanced to contrast the results obtained with the different thresholds. Specifically, the dataset is created as follows. First, the number of observation is set to 300. Second, two observed covariates, $X_1$ and $X_2$, and one unobserved covariate, $U$, are generated with the correlation of 0.6. Then, the binary treatment, $Z$, and the continuous outcome, $Y$, are generated according to the following formulae.

$$Z^* \sim \mathcal{N}(-X_1 - X_2 - U, .5) \quad (48)$$

$$Z = \begin{cases} 1, & \text{if } Z^* \geq \text{median}(Z^*) \\ 0, & \text{if otherwise} \end{cases} \quad (49)$$

$$Y \sim \mathcal{N}(Z - X_1 - X_2 - U, 2) \quad (50)$$

Thus, the estimates obtained from the model including $u$ are:

$$y = .98(.39)z - 1.00(.17)x_1 - 1.12(.17)x_2 - 1.14(.17)u - .04(.23) \quad (51)$$

while the researchers are assumed to obtain the following wrong estimates:

$$y = 1.98(.39)z - 1.22(.18)x_1 - 1.34(.18)x_2 - .54(.23) \quad (52)$$

Then, the target value is set at $\tau = 1$, and three different values are assigned to each of $\rho_{\text{threshold}}$ and $\varepsilon_{\text{threshold}}$ while fixing the value of the other threshold. Then, GSA is performed six times.

Figure 10 shows how much the contour plot of GSA is affected by the different thresholds of these screening processes. The vertical and horizontal axes are the partial correlation between PU and the outcome and that between PU and the treatment respectively. In the left figure, the threshold for the orthogonality condition (i.e. $\rho_{\text{threshold}}$ in Equation (46)) is
Figure 10: The effects of the size of the deviation of the updated treatment effect from the target value (precision in vertical axis) and the strictness of orthogonality condition between PU and observed covariates (PU precision in horizontal axis) on the shape of the contour plot manipulated, while the error tolerance level in accepting candidate variables (i.e. $\epsilon_{\text{threshold}}$ in Equation (47)) is manipulated in the right figure. The contour plots obtained from the most strict settings (i.e. $\rho_{\text{threshold}} = .999$ in the left figure and $\epsilon_{\text{threshold}} = .001$) are displayed as blue hollow circles, those obtained from the modest settings (i.e. $\rho_{\text{threshold}} = .99$ in the left figure and $\epsilon_{\text{threshold}} = .01$) are displayed as green triangles, and those obtained from the least strict settings (i.e. $\rho_{\text{threshold}} = .9$ in the left figure and $\epsilon_{\text{threshold}} = .1$) are displayed as red crosses.

In the left figure, the partial correlations are underestimated when the threshold for the orthogonality condition is loosely set (displayed as red crosses). This implies that the contour plots are more likely to be affected by the algorithm because, as discussed in Appendix C, the algorithm searches the set of $c_1$ and $c_2$ from either horizontal or vertical axes. On the other hand, there is no noticeable difference between the plots obtained from the strict setting (displayed as blue hollow circles) and those obtained from the moderate setting (displayed as green triangles). Given computational inefficiency associated with the strict setting, the moderate setting best balances accuracy and efficiency.\footnote{Although there is no clear definition for the moderate threshold, $\rho_{\text{threshold}} = .99$ has produced sufficiently accurate results in most cases. When a binary $U$ is assumed and the number of observation is small, this...}
In the right figure, the comparison between the moderate setting (displayed as green triangles) and the loose setting (displayed as red crosses) indicates that the more strict threshold is associated with somewhat narrower contour. However, the scatter plot obtained from the most strict setting (displayed as blue hollow circles) shows greater dispersion than the other two plots, which again suggests a moderate setting. Although an optimal tolerance level is difficult to find, the error tolerance level smaller than 1% usually does not reduce the size of the dispersion. To wrap up, researchers might be tempted to narrow the tolerance level, hoping it will result in a more accurate contour curve, but this only helps remove the measurement errors. The dispersions in the contour plot that cannot be eliminated by these screening processes reflect the variance of the sensitivity parameters.

4. **Accept the candidate variable as a pseudo unobservable and calculate the partial effects**. Once we obtain the PU of which confounding changes the treatment by a similar amount that the true $U$ would change, the size of the confounding is recorded in the form of partial effects such as partial R-squares or partial correlations as well as the sensitivity parameters $\alpha$ and $\delta$.

5. **Repeat Step 2 – 4 more than 100 times**. PUs are generated in this way for a sufficient number of times (usually more than 100 times), and the partial effects of PUs are recorded for each successful draw of a PU. In general, the greater number of iterations are necessary when the contour plot shows wide dispersion. To draw a contour that covers a large part of the plane defined by the partial effects, we need to find various combinations of $\alpha$ and $\delta$ that change the treatment effect to the same target value. GSA finds such combinations by changing the values of $c_1$ and $c_2$ in systematic way, of which algorithm is discussed in detail in Appendix C.

criterion sometimes requires a large number of iterations in generating a candidate variable. Even in this case, setting $\rho_{\text{threshold}} = .98$ significantly improves the computational efficiency.
6. Draw the scatter plot using the partial effects of PUs collected in Step 5. At this point, researchers should have more than one hundred combinations of the partial effects. The contour curve, like those in Figure 2, is produced by plotting the partial effects. Superimposing fitted line is also informative to check the degree of fitness.

7. Add the plots of the partial effects of the covariates. Following Imbens (2003), the final step is to add the plots of the partial correlations or the partial R-squares of each of the covariates in the figure. Because researchers usually do not have any way to interpret the contour by itself, the plots of the covariates provide the contour with a relative scale to evaluate its size.

C Search algorithm for \( c_1 \) and \( c_2 \)

In this appendix, how the algorithm of GSA finds various sets of the two location parameters, \( c_1 \) and \( c_2 \), will be explained. The following algorithm is the current best, which reasonably balances computational efficiency and compatibility to various data, but may change in future. Each step is presented as a subfigure in Figure 11 in alphabetical order.

The step 1 is to assume hyperbolic relationship between \( c_1 \) and \( c_2 \) (Figure 11a), namely \( c_1 \cdot c_2 \equiv C^* \), where \( C^* \) is an unknown constant. This relationship is assumed as an analogy to the formula for omitted variable bias in Equation (4). However, unlike the relationship between the two sensitivity parameters, \( \alpha \) and \( \delta \), this hyperbolic relation is just a convenient approximation of the true distribution of \( c_1 \) and \( c_2 \), which depends on the model, the data and the PU. Thus, we will not bother to figure out the exact relationship.

The step 2 is to find the several sets of \( c_1 \) and \( c_2 \) that change the original treatment effect to the target value fixing \( c_1 = c_2 \) (Figure 11b). The algorithm finds these sets by gradually increasing the values of \( c_1 \) and \( c_2 \) diagonally from the origin to the intersection between the maximum of \( c_1 \) and that of \( c_2 \) (or anywhere on the diagonal line beyond which, the researchers believe, such sets cannot be found).
(a) Assume hyperbolic curve
(b) Solve diagonally
(c) Choose median
(d) Divide c2 equally
(e) Find corresponding c1
(f) Increase c2 gradually
(g) Move to next c1
(h) Repeat for all c1s
(i) Divide c1 equally
(j) Find corresponding c2
(k) Find c1 for each c2
(l) Finish

Figure 11: Search algorithm of two location parameters, $c_1$ and $c_2$, in GSA
In the step 3, this process is repeated for an odd number of times, and the set of $c_1$ and $c_2$ with the median values is selected (Figure 11c). This median set is more likely to be located closer to the hyperbolic curve, if the step 2 is repeated for the greater number of times. However, the efficiency gain from the iterations more than 10 times is small. Then, we assume the hyperbolic curve as:

$$c_1 \cdot c_2 = C^* \approx \text{median}(\hat{c}_1) \cdot \text{median}(\hat{c}_2) \quad (53)$$

In the step 4, the tentative maximum of $c_2$ ($t_{\text{max}}(c_2)$) is determined by dividing $\text{median}(\hat{c}_1) \cdot \text{median}(\hat{c}_2)$ by the value slightly smaller than $\text{median}(\hat{c}_1)$, say $0.8 \cdot \text{median}(c_1)$. Similarly, the tentative minimum of $c_2$ ($t_{\text{min}}(c_2)$) is obtained by dividing $\text{median}(\hat{c}_1) \cdot \text{median}(\hat{c}_2)$ by the maximum (not tentative) of $c_2$. Formally:

$$t_{\text{max}}(c_2) = \frac{\text{median}(\hat{c}_1) \cdot \text{median}(\hat{c}_2)}{0.8 \cdot \text{median}(\hat{c}_1)} = 1.25 \cdot \text{median}(\hat{c}_2) \quad (54)$$
$$t_{\text{min}}(c_2) = \frac{\text{median}(\hat{c}_1) \cdot \text{median}(\hat{c}_2)}{\max(c_1)} \quad (55)$$

Then, the vertical segment between $(0, t_{\text{min}}(\hat{c}_1), 0)$ and $(0, 1.25 \cdot t_{\text{max}}(\hat{c}_2))$, in the vertical axis is divided into half of the number of observations of equally sized intervals (Figure 11d). In this example, the number of observation is set at 10 for illustration purpose, but this should be set at more than 100.

In the step 5, the values of $c_1$ are calculated where the hyperbolic curve estimated in the step 3 intersects with each of $c_2$ obtained in the step 4 (Figure 11e). The algorithm determines the values of $c_1$ this way so that PUs are drawn more frequently where the curvature of the contour is high.

In the step 6, the value of $c_2$ corresponding to the leftmost $c_1$ set at the step 5 is searched. In doing so, the algorithm first divides the vertical segment between $(t_{\text{min}}(\hat{c}_1), 0)$ and $(t_{\text{min}}(\hat{c}_1), t_{\text{max}}(\hat{c}_2))$ into a certain number of equidistant intervals.\(^{44}\) In the figure, the

\(^{44}\)In the program, this number is specified as \texttt{resolution}.
number is set at 9, but this number is usually set at more than 100 because GSA cannot
directly find the set of $\alpha$ and $\delta$ that changes the original treatment effect to the target value,
so PUs must be drawn multiple times. Then, the value of $c_2$ is gradually increased up to the
tentative maximum of $c_2$, $t_{max}(\hat{c}_2)$, or until the value of $c_2$ is found (Figure 11f).

In the step 7, the algorithm finds the value of $c_2$ in the same way as the step 6 by
fixing $c_1$ at the next value (Figure 11g). In the step 8, the algorithm repeats this process
until $c_2$ is searched for all the fixed values of $c_1$ (Figure 11h).

From the step 9 to the step 11, the procedure from the step 4 to the step 8 is repeated
by switching the roles of $c_1$ and $c_2$ (Figures 11i, 11j, and 11k). We obtain the entire sets of
$c_1$ and $c_2$ by merging the sets of $c_1$ and $c_2$ obtained from the step 4 to the step 8 and those
obtained from the step 9 to the step 11 (Figure 11l).