Seminar, so hello everyone.
My name is Qingyuan Zhao, I'm currently a University Lecturer in Statistics in University of Cambridge.
I visited Yale Biostats, briefly last year in February.
And so it's nice to see every guest very shortly this time.
And today I'll talk about sensitivity analysis for observational studies, looking back and moving forward.
So this is based on ongoing work with several people Bo Zhang, Ting Ye and Dylan Small at University of Pennsylvania, and also Joe Hogan at Brown University.
So sensitivity analysis is really a very broad term and you can find in almost any area that uses mathematical models.
So, broadly speaking, what it tries to do is it studies how the uncertainty in the input of a mathematical model or system, numerical or otherwise can be apportioned to different sources of uncertainty in it's input. So it's an extremely broad concept.
And you can even fit statistics as part of a sensitivity analysis in some sense. But here, there can be a lot of kinds of model inputs.
So, in particular, it can be any factor that can be changed in a model prior to its execution.
One example is structural...
epistemic sources of uncertainty. And this is sort of the things we’ll talk about. So basically, what our talk about today is those things that we don’t really know. I mean, we made a lot of assumptions about when proposing such a model. So in the context of observational studies, a very common and typical question that requires sensitivity analysis is the following. How do the qualitative and or the quantitative conclusions of the observational study change if the no unmeasured confounding assumption is violated? So this is really common because essentially, in the vast majority of observational studies, it’s essential to assume this no unmeasured confounding assumption, and this is an assumption that we cannot test with empirical data, at least with just observational data. So any, if you do any observational studies, so you’re almost bound to be asked this question, that, what if this assumption doesn’t hold? And I’d like to point out that this question is fundamentally connected to missing not at random in the missing data literature. So what I will do today is I’ll focus on sensitivity analysis for observational studies, but a lot of the ideas are drawn from the missing data literature. And most of the ideas that I’ll talk about
today can be also applied there and to related problems as well. 
So, currently, a state of the art of sensitivity analysis for observational studies is the following. 
There are many, many masters gazillions of methods of exaggeration, but certainly many many methods that are specifically designed for different kinds of sensitivity analysis.

It often also depends on how you analyze your data under unmeasured confounding assumption. There are various forms of statistical guarantees that have been proposed. And oftentimes, these methods are not always straightforward to interpret, at least for inexperienced researchers, it can be quite complicated and confusing.

The goal of this talk is to give you a high level overview. This is not a talk where I’m gonna unveil a lot of new methods. This is more of an overview kind of talk that just to try to go through some of the main ideas in this area.

So in particular, what I wanted to address is the following two questions. What is the common structure behind all these sensitivity analysis methods? And what are some good principles and ideas we should follow and perhaps extend when we have similar problems?

The perspective of this talk will be global and frequentist. By that, I mean,
There's an area in sensitivity analysis called local sensitivity analysis, where you're only allowed to move your parameter near its maximum likelihood estimate, usually. But global sensitivity analysis refer to the method that you can model your sensitivity parameter freely in a space. So that's what we'll focus on today. And also, I'll take a frequentist perspective. So I won't talk about Bayesian sensitivity analysis, which is also a big area. And I'll use this portal typical setup in observational studies, where you have iid copies of these observed data $O$, which has three parts, $x$ is the covariance, $A$ the binary treatment, $Y$ is the outcome and these observed observed data that come from underlying full data, $F$, which includes $X$ and $A$ and the potential outcomes, $Y(0)$ and $Y(1)$. Okay, so this is, if you haven't, if most of you probably have seen this many, many times already, but if you haven't seen that this is the most typical setup in observational studies. And it kind of gets a little bit boring when you see it so many times. But what we're trying to do is use this as the simplest example, to demonstrate the structure and ideas. And hopefully, if you understand these good ideas,
0:06:46.06 –> 0:06:49.78 you can apply them to your problems
0:06:49.78 –> 0:06:52.793 that are maybe slightly more complicated than this.
0:06:54.93 –> 0:06:56.758 So here’s the outline
0:06:56.758 –> 0:06:58.5 and I’ll give a motivating example
0:06:58.5 –> 0:07:01.26 then I’ll talk about three components
0:07:01.26 –> 0:07:02.85 in the sensitivity analysis.
0:07:02.85 –> 0:07:04.33 There the sensitivity model,
0:07:04.33 –> 0:07:07.633 the statistical inference and the interpretation.
0:07:09.53 –> 0:07:13.33 So the motivating example will sort of demonstrate
0:07:13.33 –> 0:07:16.24 where these three components come from.
0:07:16.24 –> 0:07:20.75 So this example is in the social sciences actually
0:07:20.75 –> 0:07:22.943 it’s about child soldiering,
0:07:29.54 –> 0:07:34.018 On the review of economics and statistics,
0:07:34.018 –> 0:07:39.018 so what they studied is this period of time in Uganda,
0:07:41.32 –> 0:07:43.572 from 1995 to 2004,
0:07:43.572 –> 0:07:45.656 where there was a civil war
0:07:45.656 –> 0:07:49.092 and about 60,000 to 80,000 youth
0:07:49.092 –> 0:07:52.223 were abducted by a rebel force.
0:07:53.12 –> 0:07:54.41 So the question is,
0:07:54.41 –> 0:07:57.98 what is the impact of child soldiering
0:07:57.98 –> 0:08:00.453 sort of this abduction by the rebel force,
0:08:01.38 –> 0:08:04.37 as on various outcomes,
0:08:04.37 –> 0:08:07.82 such as years of education,
0:08:07.82 –> 0:08:11.663 and in this paper to actually study the number of outcomes.
0:08:12.74 –> 0:08:16.59 The authors controlled for a variety of baseline covariates,
0:08:16.59 –> 0:08:19.64 like the children’s age, their household size,
0:08:19.64 –> 0:08:22.013 their parental education, et cetera.
0:08:23.21 –> 0:08:25.71 They were quite concerned about
0:08:25.71 –> 0:08:28.48 this possible unmeasured confounder.
0:08:28.48 -> 0:08:32.89 That is the child’s ability to hide from the rebel.
0:08:32.89 -> 0:08:37.89 So it’s possible that maybe if this child is smart,
0:08:38.62 -> 0:08:41.23 and if he knows that he or she knows
0:08:41.23 -> 0:08:44.01 how to hide from the rebel,
0:08:44.01 -> 0:08:48.61 then he’s less likely to be abducted
0:08:48.61 -> 0:08:50.543 to be in this data set.
0:08:51.62 -> 0:08:54.68 And he’ll probably also be more likely
0:08:54.68 -> 0:08:58.21 to receive longer education just because maybe
0:09:00.331 -> 0:09:04.023 the skin is a bit more small, let’s say.
0:09:05.71 -> 0:09:07.19 So in their analysis,
0:09:07.19 -> 0:09:10.88 they follow the model proposed by Imbens,
0:09:10.88 -> 0:09:12.43 which is the following.
0:09:12.43 -> 0:09:17.43 So basically, they assume this no unmeasured confounding
0:09:18.12 -> 0:09:21.273 after you conditional on this unmeasured confounder U.
0:09:22.48 -> 0:09:24.291 Okay, so X are all covariates
0:09:24.291 -> 0:09:25.233 that U controlled for,
0:09:26.197 -> 0:09:30.41 and U is they assumed is a binary, unmeasured confounder.
0:09:31.84 -> 0:09:34.513 That’s just a coin flip.
0:09:35.8 -> 0:09:39 And then they assume the logistic model
0:09:39 -> 0:09:44 for the probability of being abducted
0:09:44.12 -> 0:09:49.12 and the normal linear model for the potential outcomes.
0:09:49.41 -> 0:09:54.41 So notice that here the linear these terms
0:09:55.22 -> 0:09:57.83 depends on not only the observed covariance,
0:09:57.83 -> 0:10:00.92 but also the unmeasured covariates U.
0:10:00.92 -> 0:10:02.44 And of course,
0:10:02.44 -> 0:10:03.91 we don’t measure this U.
0:10:03.91 -> 0:10:08.003 So we cannot directly fit these models.
0:10:08.92 -> 0:10:12.04 But what they did is they because they made
0:10:12.04 -> 0:10:16.08 some distribution assumptions on U,
0:10:16.08 –> 0:10:19.1 you can treat U as unmeasured variable.
0:10:19.1 –> 0:10:21.0 And then, for example,
0:10:21.0 –> 0:10:23.873 fit maximum likelihood estimate.
0:10:25.47 –> 0:10:29.397 So they’re treated this two parameters lambda and delta,
0:10:29.397 –> 0:10:30.993 as sensitivity parameters.
0:10:31.98 –> 0:10:34.97 So these are the parameters that you vary
0:10:34.97 –> 0:10:37.26 in a sensitivity analysis.
0:10:37.26 –> 0:10:39.22 So when they’re both equal to zero,
0:10:39.22 –> 0:10:42.837 that means that there is no unmeasured confounding.
0:10:42.837 –> 0:10:45.81 So you can actually just ignore this confounder U.
0:10:45.81 –> 0:10:48.38 So it corresponds to your primary analysis,
0:10:48.38 –> 0:10:49.81 but in a sensitivity analysis,
0:10:49.81 –> 0:10:52.58 you change the values of lambda and U
0:10:52.58 –> 0:10:55.33 and you see how that changes your result
0:10:55.33 –> 0:10:57.03 above this parameter beta,
0:10:57.03 –> 0:10:59.783 which is interpreted as a causal effect.
0:11:01.54 –> 0:11:06 Okay, so the results can be summarized in this one slide.
0:11:06 –> 0:11:07.94 I mean they’ve done a lot more definitely.
0:11:07.94 –> 0:11:11.65 But for the purpose of this talk, basically,
0:11:11.65 –> 0:11:14.76 what they found is that the primary analysis
0:11:14.76 –> 0:11:17.443 found that the average treatment effect is -0.76.
0:11:18.6 –> 0:11:21.27 So remember the outcome was years of education.
0:11:21.27 –> 0:11:23.46 So being abducted,
0:11:23.46 –> 0:11:28.297 has a significant negative effect on education.
0:11:30.15 –> 0:11:32.16 And then it did a sensitivity analysis,
0:11:32.16 –> 0:11:35.74 which can be summarized in this calibration plot.
0:11:35.74 –> 0:11:39.71 What is shown here is that these two axis
0:11:39.71 –> 0:11:43.01 are basically the two sensitivity parameters,
0:11:43.01 –> 0:11:44.89 lambda and delta.
0:11:44.89 –> 0:11:48.35 So what the paper did is they transform it
0:11:48.35 –> 0:11:50.423 to the increase in R-squared.
0:11:51.31 –> 0:11:54.853 But that’s that can be mapped to lambda and delta,
0:11:55.72 –> 0:11:57.28 and then they compared
0:11:59.04 –> 0:12:01.78 this curve, so this dashed curve
0:12:03.384 –> 0:12:06.97 is where the values of lambda and delta such that
0:12:06.97 –> 0:12:09.773 the treatment in fact is reduced by half.
0:12:10.7 –> 0:12:12.64 And then they compare this curve
0:12:12.64 –> 0:12:15.08 with all the measured confounders, 
0:12:15.08 –> 0:12:16.76 like year and a location,
0:12:21.348 –> 0:12:25.02 And then you compare it with the corresponding coefficients
0:12:25.02 –> 0:12:30.02 of those variables in the model
0:12:30.8 –> 0:12:35.8 and then they just plot these in the same figure.
0:12:36.58 –> 0:12:39.26 What is supposed to show is that look,
0:12:39.26 –> 0:12:42.36 this is the point where the treatment effect
0:12:42.36 –> 0:12:44.12 is reduced by half,
0:12:44.1 –> 0:12:47.12 and this is about the same strength
0:12:47.12 –> 0:12:50.093 as location or birth alone.
0:12:50.093 –> 0:12:53.73 So, if you think your unmeasured confounder is in some sense
0:12:53.73 –> 0:12:57.95 as strong as the location or the year of birth,
0:12:57.95 –> 0:13:00.59 then it is possible that the treatment infact,
0:13:00.59 –> 0:13:03.993 is half of what it is estimated to be.
0:13:05.02 –> 0:13:07.76 Okay, so it’s a pretty neat way
0:13:07.76 –> 0:13:10.503 to present a sensitivity analysis.
0:13:12.22 –> 0:13:13.55 So in this example, you see,
0:13:13.55 –> 0:13:16.73 there’s three components of sensitivity analysis.
0:13:16.73 –> 0:13:19.137 First is model augmentation.
0:13:19.137 –> 0:13:23.51 And you need to expand the model used by primary analysis
Second, you need to do statistical inference. So you vary the sensitivity parameter, estimate the effect, and then control some statistical errors. 

So what they did is, they essentially varied lambda and delta, and they estimated the average treatment effect under that lambda and delta. And the third component is to interpret the results. So this paper relied on that calibration plot for that purpose. But this is often quite a tricky because the sensitivity analysis is complicated as we need to probe different directions of unmeasured confounding. So the interpretation is actually not always straightforward and sometimes can be quite complicated. There did you have there do exist two issues with this analysis. So this is just the model and rewriting it. The first issue is that actually the sensitivity parameters lambda and delta, where we vary in a sensitivity analysis are identifiable from the observed data. This is because this is a perfect parametric model. And then it’s not constructed in any way so that these lambda and delta are not identifiable. In fact, in the next slide, I’m going to show you some empirical evidence that you can actually estimate these two parameters.
So, logically it is inconsistent for us to vary the sensitivity parameter. Because if we truly believe in this model and the data actually tell us what the values of lambda and delta is. So this is the similar criticism that for Hattman selection model, for example. The second issue is a bit subtle is that in a calibration plot, what they did is they use the partial R squared as a way to measure lambda and delta in a more interpretable way. But actually the partial R squared for the observed and unobserved confounders are not directly comparable. This is because they’re they use different reference model to start with. So, actually you need to be quite careful about these interpretation this calibration quotes. So, here is what I promised that suggests you can actually identify these two sensitivity parameters lambda and delta. So here the red dots are the maximum likelihood estimators. And then these solid curves this regions, or the rejection, or I should say acceptance region for the likelihood ratio test. So this is at level 0.50, this is 0.10, this is 0.05. There is a symmetry around the origin that’s
because the U number is symmetric.
So, lambda like delta is the same
as minus lambda minus delta.
But what you see
is that you can actually estimate lambda and delta
and you can sort of estimate it
to be in a certain region.
So, something a bit interesting here
is that there’s more you can say about Delta,
which is the parameter for the outcome,
than the parameter for the treatment lambda.
But in any case,
it didn’t look like we can just vary
this parameter lambda delta freely in this space
and then expect to get different results
for each point.
What we actually can get is some estimate
of this sensitivity parameters.
So the lesson here is that
if you use a parametric sensitivity models,
then they need to be carefully constructed
to avoid these kind of issues.
So next I’ll talk about the first component
of the sensitivity analysis,
which is your sensitivity model.
So very generally,
if you think about what is the sensitivity model,
is essentially it’s a model for the full data F,
that include some things that are not observed.
So, what we are trying to do here
is to infer the full data distribution
So a sensitivity model is basically a family of distributions of the full data, is parameterized by two parameters theta and eta. So, I’m using eta to stand for the sensitivity parameters and theta is some other parameters that parameterize the distribution. So the sensitivity model needs to satisfy two properties. So first of all, if we set the sensitivity parameter eta to be equal to zero, then that should correspond to our primary analysis assuming no unmeasured confounders. So I call this augmentation. A second property is that given the value of the sensitivity prior to eta, then we can actually identify this parameters data from the observed data. So this is sort of a minimal assumption. Otherwise, this model is simply too rich, and so I call model identifiability. So the statistical problem in sensitivity analysis is that if I give you the value of eta or the range of eta, can you use observed data to make inference about some causal parameter that is a function of the theta and eta. Okay, so this is a very general abstraction. But it’s a bit too general. So let’s make it slightly more concrete.
by understanding these observational equivalence causes.

So essentially, what we’re trying to do is we observe some data, but then we know there’s an underlying full data and instead of just modeling the observed data, we’re modeling the full data set. So that makes our model quite rich, because we’re modeling something that are all observed.

For that purpose is useful to define this observationally equivalence relation between two full data distribution, which just means that their implied observed data distributions are exactly the same.

So we write this as this approximate equal to this equivalence symbol.

So then we can define the equivalence class of a distribution of a full data distribution, which are all the other full data distributions in this family that are observationally equivalent to that distribution.

Then we can sort of classify these sensitivity models based on the behavior of these equivalence classes.

So, what happened in the last example is that the full data distribution full data model is not rich enough.

So these equivalence classes are just singleton’s so can actually identify the sensitivity parameter eta from the observed data.

So, this makes this model testable in some sense with the choice of sensitivity parameter testable,
and this should generally be avoided in practice.

Then there are the global sensitivity models

where you can basically freely vary

the sensitivity parameter $\eta$.

And for any $\eta$ you can always find the $\theta$

such that it is observational equivalent

to where you started from.

And then even nicer models the separable model

where basically, this $\eta$

the sensitivity parameter doesn’t change

the observation of the observed data distribution.

So for any $\theta$ and $\eta$,

$\theta$ and $\eta$ is equivalent to $\theta$ and zero.

So these are really nice models to work with.

So understand the difference between global models

and separable models.

So basically, it’s just that they have different shapes

of the equivalence classes.

So for separable models,

these equivalence classes,

needs to be perpendicular to the $\theta$ axis.

But that’s not needed for global sensitivity models.

So I’ve talked about what a sensitivity model means

and some basic properties of it,

but haven’t talked about how to build them.

So generally, in this setup,

there’s three ways to build a sensitivity model.

And then they essentially correspond

with different factorizations

of the full data distribution.

So there’s a simultaneous model
that tries to factorize distribution this way. So introduces unmeasured confounder, U, and then you need to model these three conditional probabilities. There’s also the treatment model that doesn’t rely on this unmeasured confounder U. But whether you need to specify is the distribution of the treatment given the unmeasured cofounders and x. And once you’ve specified that you can use Bayes formula to get this part. And then there’s the outcome model that factorizes this distribution in the other way. So this is basically the propensity score and the third turn is what we need to specify it’s a sensitivity parameter. So in the missing data literature, second model kind of model is usually called selection model. And the third kind of models usually called pattern mixture model, and there are other names that have been given to it. And basically different sensitivity models, they amount to different ways of specifying these either non identifiable distributions, which are these ones that are underlined. A good review is this report by a committee organized by the National Research Council. This ongoing review paper that we’re writing also gives a comprehensive review of many models that have been proposed using these factorizations.
Okay, so that’s about the sensitivity model. The next component is statistical inference. Things get a little bit tricky here, because there are two kinds of inference we can talk about in this study.

So, the first mode of inference is point identify inference. So you only care about a fixed value of the sensitivity parameter $\eta$. And the second kind of inference is partial identified inference, where you perform the statistical inference simultaneously for a range of security parameters $\eta$. And that range $H$ is given to you.

For point identified inference usually let’s say for interval estimators, you want to construct confidence intervals. And these confidence intervals depend on the observed $\theta$. And this in turn comes differences to core guarantees. So for point identified inference usually let’s say for interval estimators, you want to construct confidence intervals. And these confidence intervals depend on the observed $\theta$.

And this in turn comes differences to core guarantees. So for point identified inference usually let’s say for interval estimators, you want to construct confidence intervals. And these confidence intervals depend on the observed $\theta$.
So, it cannot depend on a specific values of the sensitivity parameter, because you only know eta is in this range H. It need to satisfy this very similar criteria. So I call this intervals that satisfy this criteria in the sensitivity interval. But in the literature people have also called this uncertainty interval and or just confidence interval. But to make it different from the first case, we’re calling a sensitivity interval here. So you can see that these two equations, two criterias look very similar, besides just that this interval needs to depend on the range instead of a particular value of the sensitivity parameter. But actually, they’re quite different. This is usually much wider. The reason is, you can actually write an equivalent form of this equation one, because this only depends on the observed data and the range H. Then for every theta in that, sorry for every eta in that range H, is missing here, eta in H and also that’s observationally equivalent to a two distribution. This interval also needs to cover the corresponding theta parameter. So in that sense, this is a much stronger guarantee that you have. So, in terms of the statistical methods,
Point identified inference is usually quite straightforward.

It’s very similar to our primary analysis.

So, primary analysis just assumes this $\eta$ equals to zero,

but this sensitivity analysis assumes $\eta$ is known.

So usually you just can just plug in

this $\eta$ in some way as an offset to your model.

And then everything works out in almost the same way

as a primary analysis.

But for partially identified analysis,

things become quite more challenging.

And there are several methods several approaches

that you can take.

So, essentially there are two big classes of methods,

one is bound estimation,

one is combining point identified inference.

The second method is basically
to try to combine the results of point identified inference.

The main idea is to sort of construct

interval estimators,
0:30:05.19 -> 0:30:08.09 for each individual sensitivity parameter
0:30:08.09 -> 0:30:10.68 and then take a union of them.
0:30:10.68 -> 0:30:13.63 So, these are the two broad approaches
0:30:13.63 -> 0:30:15.973 to the partially identified inference.
0:30:17.61 -> 0:30:20.15 And so, within the first approach
0:30:20.15 -> 0:30:22.01 the bound estimation approach,
0:30:22.01 -> 0:30:24.73 there are also several variety of,
0:30:24.73 -> 0:30:26.7 there are several possible methods
0:30:26.7 -> 0:30:28.163 depending on your problem.
0:30:29.48 -> 0:30:31.47 So, the first problem,
0:30:31.47 -> 0:30:34.77 the first method is called separable balance.
0:30:34.77 -> 0:30:38.93 But before that, let’s just slightly change our notation
0:30:38.93 -> 0:30:43.93 and parameterize this range H by a hyper parameter gamma.
0:30:46.526 -> 0:30:51.526 So, this is useful when we outline these methods.
0:30:51.83 -> 0:30:54.8 And then this beta L of gamma,
0:30:54.8 -> 0:30:59.683 this is the lower end of the partial identify region.
0:31:00.91 -> 0:31:04.913 So the first method is called separable bounds.
0:31:05.937 -> 0:31:10.937 What it tries to do is to write this lower end
0:31:11.418 -> 0:31:15.37 as a function of beta star and gamma,
0:31:15.37 -> 0:31:19.853 where beta star is your primary analysis estimate.
0:31:20.93 -> 0:31:23.65 So let’s say theta star zero
0:31:23.65 -> 0:31:26.55 is what you would do in a primary analysis
0:31:26.55 -> 0:31:30.413 that is observationally equivalent to the true distribution.
0:31:31.91 -> 0:31:36.91 And then, if beta star is the corresponding causal effect,
0:31:37.03 -> 0:31:38.563 from that model,
0:31:39.42 -> 0:31:42.38 and if somehow can write this lower end
0:31:42.38 -> 0:31:45.666 as a function of beta star and gamma
0:31:45.666 -> 0:31:47.16 and the function is known,
0:31:47.16 -> 0:31:49.67 then our life is quite easy,
0:31:49.67 -> 0:31:52.54 because we already know how to make inference
about beta star from the primary analysis.
And all we need to do is just plug in
that beta star in this formula,
and then we’re all done.
And we call this separable because it allows us
to separate the primary analysis
from the sensitivity analysis.
And statistical inference becomes a trivial extension
of the primary analysis.
So, some examples of this kind of method
include the classical cornfields bound
and the E-value,
if you have heard about them,
and E-value seems quite popular
these days at demonology.
The second type of bound estimation
is called tractable bounds.
we may derive this lower bound as a function
of theta star and gamma.
So we are not able to reduce it to just depend
on beta star the causal effect
under no unmeasured confounding,
but we’re able to express in terms of theta star.
And then the function gl is also some practical functions
that we can compute.
And then this also makes our lives quite a lot easier,
because we can just replace this theta star,
which can be nonparametric can be parametric,
by its empirical estimate.
And, often in these cases,
we can find some central limit theorems for the corresponding sample estimator, such that the sample estimator of the bounds converges to its truth at root and rate and it follows the normal limit. And then if we can estimate this standard error, then we can use this central limit theorem to make partial identified inference because we can estimate the bounds. There’s some examples in the literature, you’re familiar with these papers. But one thing to be careful about these kind of tractable bounds is that things that get a little bit tricky with syntactic theory. This is because in a syntactic theory, the confidence intervals or the sensitivity intervals in this case, can be point wise or uniform in terms of the sample size. So it’s possible that if the convergence, if there are statistical guarantee is point wise, then you sometimes in extreme cases, even with very large sample size, they’re still exist data distributions such that your coverage is very poor. This point is discussed very heavily in econometrics literature. And these are some references. So that’s the second type of method in the first broad approach.
The third kind of method is called stochastic programming. And this applies when the model is separable. So and we can write this parameter we’re interested in as some expectation of some function of the theta and the sensitivity parameter eta. Okay, so in this case, the bound becomes the optimal value for an optimization problem, which you want to minimize expectation of some function. And the parameter in this function is in some set as defined by U. So, this is known as stochastic programming. So, this type of problem is known as stochastic programming in the optimization literature. And what people do there is they sample from the distribution, and then they try to use it to solve the empirical version and try to use that as approximate solution to this population optimization problem, which we can’t directly U value evaluate. And the method is called sample average approximation in the optimization literature. So, what is shown there. And Alex Shapiro did a lot of great work on this, that nice problems with compact set age, and everything is euclidean. So it’s finite dimensional. Then you actually have a central limit theorem.
0:37:03.73 –> 0:37:05.693 for the sample optimal value.
0:37:07.15 –> 0:37:11.82 And this link, is a link between sensitivity analysis and this link, is a link between sensitivity analysis and stochastic programming is made in this paper and stochastic programming is made in this paper by Tudball et al.
0:37:20.33 –> 0:37:22.89 Okay, so that’s the first broad approach okay, so that’s the first broad approach with doing bounds estimation.
0:37:22.89 –> 0:37:25.003 The second broad approach is to combine the results of points identified inference.
0:37:26.29 –> 0:37:29.333 Suppose these are the confidence intervals of points identified inference.
0:37:29.33 –> 0:37:31.423 So, the first possibility is to take a union so, the first possibility is to take a union of the individual confidence intervals.
0:37:32.37 –> 0:37:36.933 Then, it is very simple to just apply a union bound then, it is very simple to just apply a union bound when the sensitivity from eta is given.
0:37:36.93 –> 0:38:01.11 and to show that if you take a union then they should satisfy the criteria and to show that if you take a union only depends on the range H, only depends on the range H,
0:37:40.02 –> 0:37:43.332 and then you just apply the union bound and then you just apply the union bound.
0:37:43.332 –> 0:37:45.282 So now, if you take a union this interval only depends on the range H, so now, if you take a union this interval only depends
0:37:46.51 –> 0:37:51.134 And this can be slightly improved then, it is very simple to just apply a union bound and this can be slightly improved to cover not just these parameters,
0:37:51.134 –> 0:37:54.06 and to show that if you take a union to cover not just these parameters, but also the entire partial identified region and to show that if you take a union,
0:37:54.06 –> 0:37:57.46 of these individual confidence intervals, of these individual confidence intervals,
0:37:57.46 –> 0:38:01.1 then they should satisfy the criteria then they should satisfy the criteria for sensitivity interval.
0:38:01.1 –> 0:38:03.355 So we discussed this in our paper.
0:38:03.35 –> 0:38:06.994 And here, so, all we need to do so we discussed this in our paper.
0:38:06.994 –> 0:38:09.06 on the range H, so we discussed this in our paper.
0:38:07.96 –> 0:38:11.511 and then you just apply the union bound and then you just apply the union bound.
0:38:11.511 –> 0:38:13.933 get this formula from the first. get this formula from the first.
0:38:13.933 –> 0:38:19.611 And this can be slightly improved and this can be slightly improved.
0:38:17.08 –> 0:38:23.27 to cover not just these parameters, to cover not just these parameters,
0:38:23.27 –> 0:38:27.21 but also the entire partial identified region but also the entire partial identified region.
0:38:27.21 –> 0:38:29.911 if the intervals if the confidence intervals if the intervals if the confidence intervals
0:38:29.91 –> 0:38:32.653 have the same tail probabilities. have the same tail probabilities.
0:38:35.05 –> 0:38:36.923 So we discussed this in our paper. So we discussed this in our paper.
0:38:38.653 –> 0:38:43.35 And here, so, all we need to do And here, so, all we need to do
0:38:43.35 –> 0:38:45.113 is to compute this union. is to compute this union.
0:38:45.97 –> 0:38:49.23 So, which essentially is an optimization problem so, which essentially is an optimization problem
0:38:49.23 –> 0:38:52.48 we’d like to minimize the lower bound.
0:38:52.48 –> 0:38:55.675 we’d like to minimize the lower bound.
0:38:55.675 –> 0:38:58.875 So, which essentially is an optimization problem
0:38:58.875 –> 0:39:01.275 we’d like to minimize the lower bound.
that the lower confidence point \( Cl \) of \( \eta \) over \( \eta \) in \( H \) and similarly for the upper bound.

And usually using of syntactic theory, we can get some normal base confidence intervals for each fixed \( \eta \).

And then we just need to optimize this thing this confidence interval over \( \eta \).

But for many problems this can be computationally challenging because the standard errors are usually quite complicated and it has some very nonlinear dependence on the parameter \( \eta \).

So optimizing this can be tricky.

This is where another method of percentile bootstrap method can greatly simplify the problem.

It’s proposed by this paper that we wrote, and what it does is instead of using the syntactic confidence interval for fixed \( \eta \), we use the percentile bootstrap interval.

Where we take \( \theta \) samples, and then you estimate the causal effect \( \beta \) in each resample and then take quantiles.

Okay, so if you use this confidence interval, then there is a general, generalized minimax inequality that allows us to construct this percentile bootstrap sensitivity interval.

So what it does is this thing in the inside just the union of these percentile construct intervals for fixed \( \eta \), taken over \( \eta \) in \( H \).
And then this generalized minimax inequality allows us to interchange the infimum with quanto and the supremum of a quanto. Okay, so the infimum of a quanto is greater than equal to the quanto of infimum and that it’s always true. So it’s just a generalization of the familia minimax inequality. Now, if you look at this order interval, this is much easier to compute, because all it needs to do is you gather data resample, then you just need to repeat method 1.3. So just get the infimum of the point estimate for that resample and the supremum for that resample. Then you do this over many, many resamples and then you take the quantiles of the infimum, lower of the infimum and upper quantile of the supremum, and then you’re done. And because this union sensitivity interval is always valid, if the individual confidence intervals are valid. So you almost got a very you got a free lunch in some sense, you don’t need to show any heavy theory. All you need to show is that these percentile bootstrap intervals are valid for each fixed eta, which are much easier to establish in real problems. And this is sort of selfish,
where I’d like to compare this idea with Efron’s bootstrap, where what was found there is that you’ve got a point estimator, you resample your data, and then many times and then use bootstrap to get the confidence interval.

For partially identified inference, you need to do a bit more. So for each resample you need to get extrema optimal estimator. Then the minimax inequality allows you just sort of transfer the intuition from the bootstrap, for bootstrap from point identification to partial identification.

So the third approach in this, is a third method in this general approach is to take the supremum of key value. And this is used in Rosenbaum sensitivity analysis. If you’re familiar with that, Essentially it’s a hypothesis testing analog of the Union confidence interval method. What it does is that if you have individually valid P values for a fixed eta, then you just take the supremum of the P values over all the etas in this range. And that can be used for partially identified inference. So what Rosenbaum did, and Rosenbaum is really a pioneer in this area in the partially identify sensitivity analysis. So what he did was use randomization tests.
to construct these key values. So, this is usually done for matched observational studies and the inside of this line of work is that you can use these inequalities particularly Holley’s inequality in probabilistic combinatorics to efficiently compute these supremum of the P values. So, usually what is done there is that the Holley’s inequality gives you a way to upper bound the distribution of a that, to upper bound family of distributions in the stochastic dominance sense. So, that is used to get these supremum of the P values. And so, basically the idea is to use some theoretical tool to simplify the computation. Okay, so that’s the statistical inference. The third part, the third component is interpretation of sensitivity analysis. And this is the area that we actually really need a lot of good work at the moment. So, overall, there are two good ideas that seem to work, that seem to improve the interpretation of sensitivity analysis. The first is sensitivity value, the second is the calibration using measured confounders. So the sensitivity value is basically the value of the sensitivity parameter or the hyper parameter,
where some qualitative conclusions about your study change.

And in our motivating example,

this is where the estimated average treatment effect is reduced by half an Rosenbaum sensitivity analysis

if you are familiar with that.

This is where, this is the value of the gamma in his model,

where we can no longer reject the causal null hypothesis.

So, this is can be seen as kind of an extension of the idea of a P value.

So P value is used for primary analysis,

so assuming no unmeasure confounding,

and then for sensitivity analysis,

you can use the sensitivity value to sort of sorry,

that’s the P value it basically measures how likely your results,

your sort of false rejection is due to sort of random chance.

But then what a sensitivity value does

is measures how much sort of how sensitive your resources is

in some sense, so, how much deviation from the unmeasured confounding it takes

to alter your conclusion.

And for sensitivity value,

there often exists a phase transition phenomenon

for partially identified inference.

This is because if you take your hyper parameter gamma

to be very large,

then essentially your partially identify region
already covered in null.

So, no matter how large your sample size is you can never reject null.

So, this is sort of an interesting phenomenon and explained first discovered by Rosenbaum in this paper I wrote also clarified some problems some issues in both the phase transition.

So, the second idea is the calibration using measured confounders. So, you have already seen an example in a motivating study.

It’s really a very necessary and practical solution to quantify the sensitivity, because it’s not really very useful if you tell people, we are sensitive at gamma equals to two, what does that really mean?

That depends on some mathematical model. But if we can somehow compare that with what we do observe, and we have, often the practitioners have some good sense about what are the important confounders and what are not.

Then this really gives us a way to calibrate and strengthen the conclusions of a sensitivity analysis. But unfortunately, although there are some good heuristics about the calibration, they’re often suffer from some subtle issues, like the ones that I described in the beginning of the talk.

If you carefully parameterize your models
this can become easier.
And this recent paper sort of explored this in terms of linear models.
But really there’s not a unifying framework then you can cover more general cases and lots of work are needed.
And when I was writing the slides, I thought maybe what we really need is to somehow build this calibration into the sensitivity model.
Because currently our workflow is that we assume a sensitivity model, and we see where things get changed, then we try to interpret those values where things get changed.
But suppose if we somehow build that, if we left the range H eta to be defined in terms of this calibration.
Perhaps gamma directly means some kind of comparisons that measured confounders this would solve some a lot of the issues.
This is just a thought I came up when I was preparing for this talk.
Okay, so to summarize, there is number of messages, which I hope you can take home.
There are three components of a sensitivity analysis.
Model augmentations, statistical inference and the interpretation of sensitivity analysis. So sensitivity model is about parameterizing, the full data distribution.
And that’s basically about over parameterizing the observed data distribution. And you can understand these models by the observational equivalence classes. You can get different model augmentations by factorizing the distribution differently and specify different models for those that are on identifiable. And there’s a difference between point identified inference and partially identified inference, and partially identified inference is usually much harder. And there are two general approaches for partially identified inference, bound estimation and combining point identified inference. For interpretation of sensitivity analysis, there seem to be two good ideas so far, to use the sensitivity value, and to calibrate that sensitivity value using measured confounders. But overall, I’d say this is still a very open area. Even for this prototypical example that people have studied for decades, it seems there’s still a lot of questions that are unresolved. And there are methods that need to be developed for this sensitivity analysis.
0:51:44.86 –> 0:51:48.03 to be regularly used in practice.

0:51:48.03 –> 0:51:50.85 And then there are many other related problems

0:51:50.85 –> 0:51:52.81 in missing data in causal inference

0:51:53.73 –> 0:51:57.703 that need to see more developments of sensitivity analysis.

0:51:58.81 –> 0:52:00.82 So that’s the end of my talk.

0:52:00.82 –> 0:52:03.423 And there are some references that are used.

0:52:04.63 –> 0:52:08.14 I’m happy to take any questions.

0:52:08.14 –> 0:52:11.167 Still have about four minutes left.


0:52:14.307 –> 0:52:17.18 Thank you, I’m sorry I couldn’t introduce you earlier,

0:52:17.18 –> 0:52:20.627 but my connection but it did not to work.

0:52:20.627 –> 0:52:23.153 So we have time for a couple of questions.

0:52:25.56 –> 0:52:28.82 You can write the question in the chat box,

0:52:28.82 –> 0:52:30.663 or just unmute yourselves.

0:52:43.482 –> 0:52:44.315 Any questions?

0:52:53.67 –> 0:52:55.962 I guess I’ll start with a question.

0:52:55.962 –> 0:53:00.12 Yeah I guess I’ll start with a question.

0:53:00.12 –> 0:53:03.89 This was a great connection between I think,

0:53:03.89 –> 0:53:05.77 sensitivity analysis literature

0:53:05.77 –> 0:53:07.493 and the missing data literature.

0:53:08.6 –> 0:53:11.982 Which I think it’s kind of overlooked.

0:53:11.982 –> 0:53:16.982 Even when you when you run a prometric sensitivity analysis,

0:53:17.36 –> 0:53:20.27 it’s really something, like most of the times

0:53:20.27 –> 0:53:22.06 people really don’t understand

0:53:22.06 –> 0:53:24.023 how much information is given.

0:53:24.86 –> 0:53:28.56 Like, how much information the model actually gives

0:53:29.4 –> 0:53:31.66 on the sensitivity parameters.

0:53:31.66 –> 0:53:34.08 And as you said,

0:53:34.08 –> 0:53:35.53 like it’s kind of inconsistent
0:53:35.53 –> 0:53:37.47 to set the sensitivity parameters
0:53:37.47 –> 0:53:40.23 when sensitivity parameters are actually identified
0:53:40.23 –> 0:53:41.193 by the model.
0:53:42.94 –> 0:53:46.19 So I think like my I guess a question of like,
0:53:46.19 –> 0:53:47.603 clarifying question is,
0:53:48.67 –> 0:53:53.66 you mentioned there is this there this testable models,
0:53:53.66 –> 0:53:55.763 this testable models essentially are wherein
0:53:55.763 –> 0:53:59.69 the sensitivity model is such that
0:53:59.69 –> 0:54:03.62 the sensitivity barometer are actually point identified.
0:54:03.62 –> 0:54:04.453 Right?
0:54:04.453 –> 0:54:05.286 - Yes.
0:54:05.286 –> 0:54:07.535 So it re, so you said,
0:54:07.535 –> 0:54:10.85 you reshooting use the sensitivity analysis
0:54:10.85 –> 0:54:13.8 to actually to set the parameters
0:54:13.8 –> 0:54:16.141 if the sensitivity parameters
0:54:16.141 –> 0:54:18 are actually identified model.
0:54:18 –> 0:54:18.833 - Yeah.
0:54:18.833 –> 0:54:20.69 - Is that what you’re trying?
0:54:20.69 –> 0:54:23.48 All right, so and. - Yes, yeah.
0:54:23.48 –> 0:54:27.3 Basically what happened there is the model is too specific,
0:54:27.3 –> 0:54:29.83 and it wasn’t constructed carefully.
0:54:29.83 –> 0:54:32.57 So it’s possible to construct parametric models
0:54:32.57 –> 0:54:36.77 that are not testable that are perfectly fine.
0:54:36.77 –> 0:54:40.31 But sometimes, if you just sort of
0:54:40.31 –> 0:54:42.17 write down the most natural model,
0:54:42.17 –> 0:54:46.4 if it just extend what the parametric model
0:54:46.4 –> 0:54:50.883 you used for observed data to also model full data,
0:54:52.1 –> 0:54:53.78 then you don’t do it carefully,
0:54:53.78 –> 0:54:58.78 then the entire full data distribution becomes identifi-
0:54:59.53 –> 0:55:02.24 So it does makes sense to treat those parameters
as sensitivity parameters.

So this kind of is a reminiscent of the discussion in the 80s about the Hackmann selection model. Because in that case, there was also sir Hackmann has this great selection model for reducing or getting rid of selection bias, but it’s based on very heavy parametric assumptions. And you can adapt certainly identify the selection effect directly from the model where you actually have no data to support that identification. Which led to some criticisms in the 80s. But I think we are seeing this things repeatedly again and again in different areas. And it’s, I think it’s fine to use the power metric models that are testable, actually, if you really believe in those models, but it doesn’t seem that they should be used this sensitivity analysis, because just logically, it’s a bit strange. It’s hard to interpret those models. And but sometimes I’ve also seen people who use the sort of parameterize the model in a way that you include enough terms. So the sensitivity parameters are weakly identified in a practical example. So with a practical data set of maybe the likelihood test, Likelihood Ratio Test rejection region,
that acceptance region is very, very large. So there are suggestions like that, that kind of it’s a sort of a compromise for good practice. Right in that case you gave it either set the parameters and drag the causal effects, or kind of treat that as a partial identification problem and just write use bounds or the methods you were mentioning, I guess. Yeah. Other questions? Well I guess you can read the question? It’s a question from Kiel Sint. Sorry if I didn’t pronounce your name correctly. “In the applications of observational studies ideally, what confounders should be collected for sensitivity analysis, power sensitivity analysis for unmeasured confounding?” Thank you. So if I understand your question correctly, basically what sensitivity analysis does is you have observational study, where you for already collected confounders that you believe are important or relevant that really that are real confounders, that change the causal unchanged the treatment and the outcome. But often that’s not enough. And what sensitivity analysis does is it tries to say, “based on what the components already
you have already collected,
what if there is still something missing
that we didn’t collect?
And then if those things behave in a certain way,
does that change our results?"
So, I guess sensitivity analysis is always relative
to a primary analysis.
I think you should use the same set of confounders
that the primary analysis uses.
I don’t see a lot of reasons to vary to say
use a primary analysis with more confounders,
but a sensitivity analysis with fewer confounders.
Sensitivity analysis is really a supplement
to what you have in the primary analysis.
Just one more question if we have?
There not.
- So from Ching Hou Soo,
“How to specify the setup sensitivity parameter gamma
in the real life question?
When gamma is too large the inference results
will always be non informative?”
Yes, this is always a tricky problem any,
and essentially the sensitivity values kind of
trying to get past that.
So it tries to directly look at the value
of this sensitivity parameter that changes your conclusion.
So in some sense, you don’t need to specify
a parameter a priori.
But obviously, in the end of the day,
we need some clue about what value of sensitivity parameter is considered large.
In a practical sense, in this application. That’s something this calibration clause is trying to address.
But as I said, they’re not perfect at the moment.
So for some time, now, at the least, we’ll have to sort of live through this and
we have to understand really what the sensitivity model means,
and then use your domain knowledge to set the sensitivity parameter,
or we have to use these imperfect visualization tools to calibrate analysis.
Yeah, all right.
Thank you.
I think we need to wrap up we’ve run over time.
So thank you again Qingyuan,
for sharing your work with us.
And thank you, everyone for joining.
Thank you.
Bye bye.
See you next week.
- It’s a great pleasure.
Thank you.