trying to draw on the model selection approach/analysis

Drawing on the logic of subset selection/lasso to do the model selection – can we choose the target parameter that is gonna be best estimated with the given data ? – maybe the logic will be that with the increasing sample size the estimate for that parameter will converge fastest ?? – or maybe the complexity of parameters increases with sample size, and pick the ones that are the best given the sample size ?? (that is, the cardinality of the set of parameters increases with sample size and we pick the ones that are optimal for the current sample size) ??

Can we develop a method to 'learn' the optimal function of the parameter vector that we can estimate well enough given the sample?

What does it mean for the ATE to approximate the actual joint distribution of potential outcomes? Like what is the metric? How 'many' dbns are there that could map to this ATE? Take a Dirichlet process and ask what's the probability of the dbn (in the support) to have this mean ? Which CEF (CEF with respect to what) has the highest entropy? Entropy / ability-to-learn trade-off ?

The challenge with fitting derivatives is that we don't observe them (in contrast to outcome levels).

Important to understand:

Even if you have a decent predictor levels $\hat{f}(x)$, it might not work as well for predicting $f(x') - f(x)$. Think of the average treatment effect as a result of moving from x to x'.

From the *all causes* framework, the object of interest:

$$\left\{ \frac{\partial}{\partial x} \phi(X, u) \right\}_{u \in supp\, U} \tag{1}$$

The average that's considered to be a more feasible target:

$$\frac{\partial}{\partial x} \mathbb{E}\left[\phi(X, U) | X = x \right] \tag{2}$$

Weight derivatives at different values of the domain with the density of those values:

$$\mathbb{E}\left[\frac{\partial}{\partial x} \mathbb{E}\left[\phi(X, U) | X = x \right] \right] \tag{3}$$

A note about the difference between using $\phi(X, U)$ and $\phi(x, U)$ (notational issue? maybe a lil more than that)

$\phi(X, U)$ and $\phi(x, U)$ are two different random variables, right?

What is the bias like for β_{BLP} in this setting? Can we obtain the parameter that β_{BLP} estimates as a result of some penalisation? So it would be a penalisation that leads to the bin being the complete domain, right?

Is β_{BLP} the efficient estimator of the Yitzhaki-weighted average derivative?

Is there a difference between

How much can we gain in terms of variance by imposing the bias of the sort that β_{BLP} does?

When we do lasso: we chose a particular 'space' with respect to which to penalise the model, right? What can we choose as a 'space' in this case of estimating derivatives?

This penalisation must make it optimal to have the longest bin width.

It seems weird to impose a minimum distance between the set of weights and the pdf of X, right? Maybe penalise directly penalise the distance between the set of Yitzhaki weights and the true pdf of X?

Set of weights $\omega : suppX \to \mathbf{R}_+$.

$$d(\omega, \omega_z) \leq \lambda \tag{4}$$
$$s.t. \min d(\omega, f_X) \tag{5}$$

Sudden realisation about lasso: If p is infinite, does it mean we can recover y perfectly?

Consider a scenario: We obtained an estimate of β_{BLP}, which we know what estimates in terms of weights for the average derivative. Then we can obtain another estimate of the average derivative with another set of weights. Can we then use this joint info to do better inference on the whole set of values of derivatives?

What can we learn about such a complex multidimensional object? Is it worth it to try to match all of it well? Maybe penalise some parts of it?

Bonhomme's suggestion to do Bayesian – that would be very similar to Chernozhukov stuff, right? With Dirichlet processes and such?

What I'm pursuing: what if we treat linear estimation when targeting derivatives not as a 'biased' approximation of real average but some weird function of those derivatives - so we kinda 'escape' the bias part ?.. does this angle make it any better/more useful ?

So, where this might make a difference: when we shrink towards the linear model, we just 'handwaive' the bias, right? we just say that: yes, we allow for bias in favour of lower variance. but what if instead of looking for such 'heuristic' ways to reduce bias, we take the approach of looking for functions/functionals of parameters that are 'easier' to estimate? i think what got me hooked up with yitzhaki weights is that we basically chose another parameter of interest and obtained an estimator that has a much lower

variance. so, then, it's not unbiasedness that we sacrifice – it's kinda 'loss of info', which is requires a different metric to judge the severity of loss? i mean, admittedly bias can probably still be used; but it might be cool to reframe the econometric problem in terms of information/variance trade-off, no? It seems more intuitive in case of dimensionality reduction – entropy seems a natural measure here, no? but with yitzhaki weights, not so clear; it's not that we reduce dimensionality, right? but in the uniform case, we kinda do, maybe? btw, both average derivative and yitzhaki weights are functionals of the actual object we're interested (the whole set of derivatives).

Is entropy a norm?

make a decision on the information loss -> obtain the function form that you need to estimate (e.g., yitzhaki weights -> linear) -> then tackle bias/variance trade-off

for how 'many' dbns yitzhaki weights are an okay/horrible loss?

the average is the minimiser of the prediction error, right? what if i specify that i want to predict only on certain subsets? the bias of yitzhaki weights depends on the dbn, right? this is true for any model selection type of approach, no?

Another phrasing of what i want to do: How to rationalise the choice of Yitzhaki weights as something you want to estimate ? Basically, in the same spirit as robust contracts.

isaieh informativeness of parameter –

Just to reiterate what I should probably do in the first place: Find some evidence (simulation or theory) that choosing weird yitzhaki weights (which lead to a biased estimate of the average derivative) can be beneficial on net given very low variance of the estimator. Maybe discover that there is middle ground between targeting average derivative and yitzhaki weights. Also another aspect: analyse shrinkage vs kernel bandwidth – both can suggest linear model but in principally different ways, right?

Another big thing to keep in mind as the end goal: what was it ?.....

Think about the papers to use as a basis for simulations? Maybe Dell & Olken Java paper?

directly relevant papers

Semenova & Chernozhukov (2021)

The starting point as i see it: We have identification of what we actually want, but the chosen estimator is biased (estimator of the nuisance parameter); this can happen bc of regularisation. And I think this is where the ols thing can fit – ols is extreme regularisation, right?

Seems natural to check: Does this orthogonalisation approach to correcting bias in the estimation of η lead to lower variance compared to kernel estimation? I think this is actually very interesting: does this way of correcting bias allow not to increase the variance

to as high as kernelling does?

Also: the issue of the choice of η (e.g., how much regularisation to do) is not discussed in the paper, right? Well, there must be tuning going on, given the estimator chosen, right? (like if lasso, then the choice of λ is incorporated). But again, that's with respect to the error in predicting function values, not derivatives! Would we get anything different (in terms of how to do tuning) if we focus on targeting derivatives rather than levels?

The benefit of deriving yitzhaki weights is that we can analytically show bias as a function of the true model/dgp.

Think about this type of cases: the shape of the cef is regular enough for the re-weighting not to have dramatic consequences, but the dgp is such that kernel ols is extremely noisy.

Think about the difference between asymptotic variance and the finite sample one – even if asymptotically two estimators are equivalent, we can probably argue that one does much better/worse for finite samples, no?

A tangential idea (which goes into the territory of selection on observables): With continuous treatment, does it make sense to do regularisation on the set of controls but kernelling with respect to the treatment variable of interest?

'parameters' that should govern the choice of the estimator for the average derivative:

- cef's shape
- joint dbn of (X, U)

things to check / look into:

- how does the choice of λ in lasso relate to where we end up in the bias/variance trade-off?
- when doing asymptotics, can we decompose the deviation from the limiting normal dbn into components i specified above (cef's shape and joint dbn of (X, U))?
- how to do cross-validation when you want to target the average derivative?
- show that shrinkage can do better than local ols in finite samples
- specify the trade-off between informativeness of the target parameter and the estimator's mse
  - need to think carefully about how to define 'informativeness'. the starting point: the joint dbn of potential outcomes. one idea: there's obviously some dimensionality reduction going on

idea about how to do cross-validation when targeting the average derivative: do some sort of a hansen test? pick an unbiased estimator and see if the biased/low-variance one is too 'far' or not. The cross-validation part comes in because need to choose how much to 'shrink' towards the linear regression. Def need to figure out how the hansen test works.