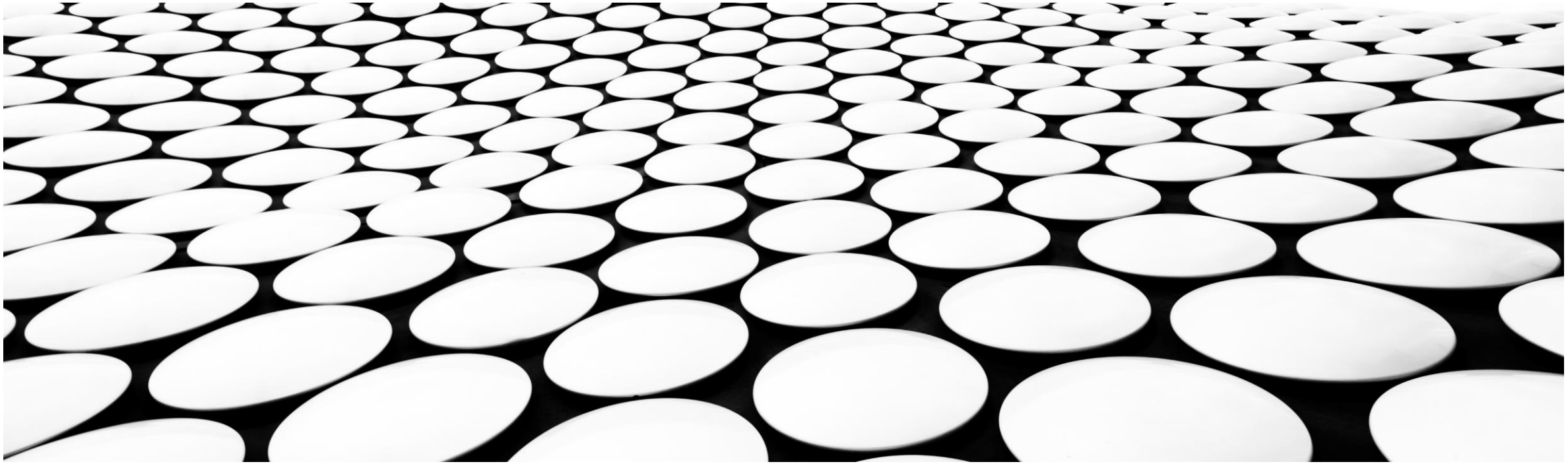
MODEL SELECTION AND SEMIPARAMETRIC ESTIMATION

DENIS NEKIPELOV, UNIVERSITY OF VIRGINIA

STEPHEN P. RYAN, WASHINGTON UNIVERSITY IN ST. LOUIS, CESIFO, AND NBER



GOING FOR A WALK IN THE (RANDOM) FOREST

RESEARCH OBJECTIVE

- **Machine Learning** (ML) methods have made several recent inroads in econometrics
- One key area of research has been around model selection
  - Many ML methods can be conceptualized as having **two steps**:
    - [**Selection**] What goes into the model?
    - [**Estimation**] What are the parameters post selection?
- Chernozhukov and co-authors have expanded both theoretical foundations and applications of ML:
  - **Double LASSO** for selecting instruments in high-dimensional IV
  - **Orthogonalization** of machine learning plug-in methods for inference
- This paper: expand the methods of Nekipelov, Novosad, and Ryan (2020) to **semiparametric domains**

BASIC IDEA

- NNR considered the problem of how to assign parameters to observations in the following moment model:

$$E\left[Y - m\left(X, \theta(Z)\right)\right] = 0$$

- Here the m is a moment function known up to a finite vector of parameters, θ, Y are outcomes, X are observable covariates, and Z are observables that govern the assignment of parameters to observations
- We proposed a *moment forest* for estimating this problem
  - Moment forest is an ensemble of moment trees
  - Moment trees are generalization of regression/classification trees with moment functions in each "leaf"
  - Generates a recursive partitioning of the *Z* space; in each partition, solve the moment above
  - We previously proved consistency and asymptotic normality of this estimator
- Here, we extend that analysis to incorporate **homogeneous** parameters

DIGRESSION ON TREES

- Decision Trees are nested binary partitions
- Grow through a greedy search at each node
- Each node has two potential children
- Keep splitting the sample until reach some limit (hyperparameters):
  - Number of observations in each split
  - Minimum level of improvement in objective function
  - Maximum depth of tree
- That controls the complexity of the tree
- Use cross-validation to determine hyperparameters
- Use honest trees: one sample to grow the tree structure, one sample to fill in values
- Universal approximators

DIGRESSION ON TREES

- Trees have some cool properties
- First is that they are universal approximators, but they work in the characteristic space (X)
- Start off with simple models and build complexity
- We do not have to specify the relationship between outcomes and explanatory variables
- But, unlike usual nonparametric estimators, you have a **parametric function** conditional on tree structure
- Under some regularity conditions, you can actually achieve **faster than parametric rates of convergence** in first step
- This means that asymptotics are governed by second step estimator, which is really nice
- **Random forest** extension is to take iid resamples and grow many trees, average together -> also, caps run time with complexity
- Random part is not only sample, but which covariates you split on -> reduces link across trees -> improve variance of estimates (in limit down to irreducible error!)

REVIEW OF MOMENT FOREST CONSTRUCTION

- Moment Forest is composed of iid resampled Moment Trees
- A Moment Tree is a rectangular partitioning of the Z space
- Estimate a separate moment in each partition
- Partitions are found by greedy search at each node
- Tree growth is stopped when convergence criteria are met (number of observations in each child node, minimum objective function improvement)

PARTIALLY HOMOGENEOUS MODELS

- We want to allow for the following:

$$E\left[Y - m\left(X; \bar{\theta}, \hat{\theta}(Z)\right)\right] = 0$$

- Where the parameter vector may include components which are homogeneous across the entire domain of Z
  - At one extreme is the standard GMM model, other is NNR
- Why would we do this?
  - **Efficiency** (variance) versus **flexibility** (bias)
  - Imposing (correctly) homogeneity helps improve the precision of our estimates
- The issue is that we don't know which parameters are heterogeneous…

ESTIMATION METHOD

- We extend NNR to allow for partially homogeneous parameters
- Sketch of idea:
  - Set hyperparameters of moment forest using cross validation
  - Grow an unrestricted moment tree
  - Test for homogeneity parameter-by-parameter across all terminal leaves, correcting for multiple hypothesis testing using Holm-Bonferroni
  - Imposing homogeneity for parameters that fail to reject null, estimate a nested fixed point:
    - In an outer loop, search for homogeneous parameters
    - In an inner loop, condition on those and grow an (optimal) moment forest

TESTING

- The Holm-Bonferroni procedure is a method for controlling the family-wise error rate in a multiple hypothesis testing setting
  - Control Type I errors (rejecting the null when it is true)
  - Cannot control Type II errors in frequentist settings (that I'm aware of; there are "agnostic test statistics" that make an attempt…?)
- Grow an unrestricted tree with *v* final splits
- Run constrained estimation imposing that $\theta_{left}^{k} = \theta_{right}^{k}$ for all splits simultaneously (akin to SUR)
- Compute all the *p*-values under hypothesis that parameters are equal using GMM distance metric test (or Wald3)
- Sort from lowest to highest value, reject null and continue to next highest *p*-value if:

$$P_k < \frac{\alpha}{m - k + 1}$$

- Otherwise, stop testing

ECONOMETRIC THEORY: HEURISTIC OVERVIEW

- First, we view our model as an “overparameterized” semiparametric model
- The conditional moment function depends on an unknown function of the covariates, $\eta(\cdot)$
- However, there exists a representation $\eta(x) = f\big(\psi(x)\big)$ where $\psi(x)$ is a known function of the covariates and f is an unknown low-dimension function
- Implies that model without constraints is correctly specified, as is any correct reduction in complexity
- Example: if $\psi(x) = \{\psi_1(x_1), \psi_2(x_2), \dots, \psi_K(x_K)\}$, then the model $\eta(x) = g(\psi_1(x_1), x_2, \dots, x_K)$ is correctly specified, where g is unknown

ECONOMETRIC THEORY: HEURISTIC OVERVIEW

- Second, we use the moment forest as a classifier to determine target function $\eta(\cdot)$ with reduced complexity
- Moment forest is a uniformly convergent universal approximator
- Can verify that moment functions fit on the low-dimensional function almost everywhere
- With guarantee on the rate of convergence of the moment forest, correct specification of model established with probability that is ***exponential*** in convergence rate
- This is a consequence of standard exponential inequalities for sample means
  - Subtle point: we may select **inefficient** models, but we will (essentially never) select ***incorrect*** models

ECONOMETRIC THEORY: HEURISTIC OVERVIEW

- Third, given correct specification, model becomes a standard semiparametric model
- Under independent splitting of the sample between classification and estimation, nonparametric component can be recovered at the specified rate
- Simultaneously, convergence of parametric component can be established conditionally on the nonparametric component (as is standard in semiparametric theory models)
- In parametric models, this means we can obtain $\sqrt{n}$ convergence!

ECONOMETRIC THEORY: HEURISTIC OVERVIEW

- Fourth, estimator for parametric component will generally depend on error in estimation of nonparametric part
- To restore parametric rates of convergence on parametric component, we orthogonalize the model
- In linear settings, this is the standard “partialling out” of the nonparametric components
- In nonlinear models, a linear offset can be created using a pilot estimate for the target parameter
- Moment models that are linear in parameters can be orthogonalized by residualizing variables corresponding to homogeneous components and subtracting their mean conditional on the nonparametric components

ECONOMETRIC THEORY: DETAILS

The data is characterized by the joint distribution of random variables (X, Z) with $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ that may share some of the components. The econometric model is defined by the moment function $\rho(\cdot;\cdot) : \mathcal{X} \times \mathbb{R}^p \mapsto \mathbb{R}^p$. The model is parameterized by the general functional parameter $\eta : \mathbb{R}^e \mapsto \mathbb{R}^p$, such that each component $\eta_i(\cdot)$ is a function mapping from $\mathbb{R}^e$ or its subset to $\mathbb{R}$ belonging to a separable space endowed with $\mathbf{L}_2(\mathbb{R}^e)$ norm. Without loss of generality we set the dimension of the moment vector $\rho(\cdot;\cdot)$ to be equal to the dimension of the parameter vector η.

The conditional moment linking the functional parameter to the distribution of the data is

$$\mathbb{E}[\rho(X\,;\,\eta(\cdot)) \,|\, Z = z] = 0. \tag{2.13}$$

ECONOMETRIC THEORY: DETAILS

ASSUMPTION 1. *1. There exists a partition* $X = (X^1, X^2)$*, a parametric function* $g_0(\cdot;\theta)$*, a parameter* $\theta_0 \in \Theta \subset \mathbb{R}^k$*, and an unknown function* $h_0(\cdot) \in \mathcal{H}_0$ *where* $\mathcal{H}_0$ *is a separable space with* $\mathbf{L}_2(\mathbb{R}^e)$ *norm such that*

$$\mathbb{E}[\rho\left(X^2\,;\, g_0(X^1;\theta_0),\, h_0(\cdot)\right) \mid Z = z] = 0.$$

2. For each $j = 0, \ldots, p$ *there exists a family of models indexed by* $\eta^j(\cdot) = \left(g^j\left(\cdot;\theta^j\right), h^j(\cdot)\right)$ *defined by parametric functions* $g^j(\cdot;\cdot)$ *indexed by* θ^j *taking values in a convex compact set* $\Theta \subset \mathbb{R}^j$ *and unknown functions* $h^j(\cdot)$ *belonging to a separable space* $\mathcal{H}^j$ *with* $\mathbf{L}_2(\mathbb{R}^e)$ *norm such that*

(a) For each fixed partition $X = (X^{1*}, x^*, X^{2*})$ *with a single element* x^* *model indexed by* $j+1$ *implies model indexed by* j*. In other words, if there exist* $(\theta^{(j+1)*}, h^{(j+1)*}(\cdot))$ *such that*

$$\mathbb{E}[\rho\left(X^{2*}\,;\, g(X^{1*}, x^*; \theta^{(j+1)*}), h^{(j+1)*}(\cdot)\right) \mid Z = z] = 0.$$

then there necessarily exists $(\theta^{j*}, h^{j*}(\cdot))$ *such that*

$$\mathbb{E}[\rho\left(X^{2*}, x^*\,;\, g^j(X^{1*};\theta^{j*}), h^{j*}(\cdot)\right) \mid Z = z] = 0.$$

(b) Model $g_0(\cdot;\theta_0)\ h_0(\cdot)$ *belongs to the family of models above indexed by* k*.*

ECONOMETRIC THEORY: DETAILS

ASSUMPTION 2. *1. Almost everywhere on* $z \in \mathcal{Z}$ $\mathbb{E}[\rho(X\,;\,\eta(\cdot)) \mid Z = z]$ *is Frechet-differentiable in each* $\eta^i(\cdot)$ *for* $i = 1, \ldots, p$ *and the Jacobian matrix* $J(z)$ *with elements*

$$J_{ij}(z) = \frac{\partial}{\partial \eta_j} \mathbb{E}[\rho^i(X\,;\,\eta(\cdot)) \mid Z = z]$$

is non-singular with bounded eigenvalues and its determinant does not change sign. Moreover,the corresponding Hessian exists and its universally bounded.

2. *For each* $h(\cdot) \in \mathcal{H}$, $\rho(X\,;\,\theta,\, h(\cdot))$ *is* L_h*-Lipschitz continuous in* θ *with* $L_h \leq \overline{L} < \infty$.

3. *For each* $\theta \in \Theta$ *with probability 1* $\rho(X\,;\,\theta,\, h(Z))$ *is* ω*-Hölder continuous in* $h(\cdot)$, *i.e. there exist constants* H *and* ω *such that for each* $h_1(\cdot), h_2(\cdot) \in \mathcal{H}$

$$\|\rho(X\,;\,\theta,\, h_1(\cdot)) - \rho(X\,;\,\theta,\, h_2(\cdot))\| \leq H\|h_1 - h_2\|^{\omega}.$$

REGULARITY CONDITIONS ON DGP

We impose the following assumption on the random forest estimator.

ASSUMPTION 3. *1. The data is split into proportional independent subsamples for each θ one subsample is used to produce a random forest estimator while the other one is used to produce θ for each random forest estimator.*

2. There exists an algorithm that induces the sequence of partitions Π indexed by the sample size n that yields the random forest estimator attaining the minimax convergence rate $n^{\gamma)}$ for each k component $h_k(\cdot)$ of $h(\cdot)$ where γ is a function of dimension of w_3 and the properties of function $h_k(\cdot)$.

LEMMA 3. *Let the total sample size be* $2n$ *with sample of size* n *used to produce the random forest estimator (2.14) and remaining sample is used to estimate parameter* θ*. Then for each dimension* k *of* θ, $\|\widehat{\theta}_k - \theta_{0k}\| = O_p(n^{-\omega\gamma})$.

Proof: For a given instance of the random forest estimator $\tilde{h}$,

$$\sup_{\theta\in\Theta} \|\frac{1}{n}\sum_{i=1}^{n} \rho(x_i^2\,;\, g(x_i^1;\theta),\tilde{h}) - \mathbb{E}[\rho(X^2\,;\, g(X_1;\theta,\tilde{h})]\| = O_p((n)^{-1/2})$$

due to Lipschitz-continuity of $\rho(\cdot)$ in θ, validity of the symmetrization lemma and exponential bound conditional on $\tilde{h}$ and universal bound on the Lipschitz constanstant which guarantees, due to independence of $\tilde{h}$, that the corresponding bound also holds in expectation over $\tilde{h}$. Next, using the first-order Taylor expansion of $\mathbb{E}[\rho(X^2\,;\, g(X_1;\theta,\tilde{h})]$ we can express

$$\mathbb{E}[\rho(X^2\,;\, g(X_1;\theta,\tilde{h})] - \mathbb{E}[\rho(X^2\,;\, g(X_1;\theta_0,\tilde{h})] = \mathbb{E}[J(Z)](\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$$

due to the boundedness of the Hessian of the conditional expectation of the moment function. Due to Hölder-continuity of the moment function in $h(\cdot)$, we can also express that

$$\mathbb{E}[\rho(X^2\,;\, g(X_1;\theta_0,\tilde{h})] = O(n^{-\gamma\omega})$$

given that $\|\tilde{h} - h_0\| = O_p(n^{-\gamma})$. Collecting terms, we conclude that

$$\widehat{\theta} - \theta_0 = O_p(n^{-1/2}) + O(n^{-\gamma\omega})$$

LEMMA 4. *Suppose that model indexed by $j+1$ as discussed above is correctly specified then for $\widehat{\theta}_1^j$ and $\widehat{\theta}_1^j$ in (2.15) and (2.16). Then for verification on an independent sample of size n :*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho\left(\left(x_i^{2j}\setminus x_i^{2j}(k)\right)\;;\;g^{j+1}(\left(x_i^{1j}\cup x_i^{2j}(k)\right);\widehat{\theta}_1^j),\widetilde{h}(\cdot)\right)\right.\right.$$
$$\left.\left.-\frac{1}{n}\sum_{i=1}^{n}\rho\left(\left(x_i^{2j}\setminus x_i^{2j}(k)\right)\;;\;g^{j+1}(\left(x_i^{1j}\cup x_i^{2j}(k)\right);\widehat{\theta}_2^j),\widetilde{h}(\cdot)\right)\right|>T\right)\to 0.$$

Proof: Denote $\rho_{il}=\rho\left(\left(x_i^{2j}\setminus x_i^{2j}(k)\right)\;;\;g^{j+1}(\left(x_i^{1j}\cup x_i^{2j}(k)\right);\widehat{\theta}_l^j),\widetilde{h}(\cdot)\right),\ l=1,2$. Then due to independence of $\widetilde{h}(\cdot)$ we first compute probability

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\rho_{i1}-\rho_{i2})\right|>T\,\middle|\,\widetilde{h}(\cdot)\right).$$

Given that $|\rho_{il}|<\overline{L}\mathrm{diam}(\Theta)$ we can apply standard exponential bound inequality to assess this probability. The magnitude of the exponent is $\exp(-\frac{2nT^2}{\overline{L}^2\mathrm{diam}(\Theta)^2})$ and it does not depend on $\widetilde{h}(\cdot)$. Therefore, we can take expectation over it which yields the statement of the lemma. ■

THE COUP DE GRACE

THEOREM 1. *The Holm-Bonferroni procedure selects the correct model with probability approaching 1 as* $n \to \infty$.

- Intuitively, the idea here is that the p-values for dimensions with homogeneity approach 1 as sample grows, while p-values for heterogeneous dimensions approach 0.
- For any level of confidence, and a sufficiently large sample size, Bonferroni-Holm procedure selects model with probability approaching one

ACCELERATING RATE OF CONVERGENCE

The acceleration of the convergence rate with orthogonal moments We note that Lemma 3 implies a slow nonparametric convergence rate for the finite-dimensional parameter. The reason for this is that the moment function depends on the function estimated with the random forest in the first order which means that the non-parametric rate of that estimator dominates the sampling noise in the empirical moment itself. The acceleration of the convergence rate to the standard parametric rate is possible if the convergence rate for the random forest estimator is sufficiently high and the moment function is in the orthogonal form.

A moment function $\rho(x^2; g(x^1; \theta), h(\cdot)) : \mathcal{X}^2 \times \mathcal{X}^1 \times \Theta \times \mathcal{H} \to \mathbb{R}^p$ is orthogonal with respect to h if its pathwise derivative w.r.t h at h_0 is zero:

$$\nabla_r \mathbb{E}\left[\rho_j(x^2; g(x^1; \theta_0), r(h - h_0) + h_0)\big|_{r=0}\right] = 0, \quad \forall j \in \{1, 2, ..., p\}.$$

ORTHOGONALIZATION OF THE MOMENT FUNCTION

THEOREM 2. *Suppose that moment function* $\rho(x^2; g(x^1;\theta), h(\cdot))$ *is orthogonal with respect to* $h(\cdot)$ *and the random forest estimator* $\widetilde{h}(\cdot)$ *satisfies* $\omega\gamma > \frac{1}{4}$. *Then* $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$.

Proof: Due to orthogonality and the finiteness of the Hessian of the moment function $\mathbb{E}[\rho(X^2\,;\, g(X_1;\theta_0,\tilde{h})] = O(n^{-2\gamma\omega})$. By the assertion of the theorem, $\gamma\omega > \frac{1}{4}$, meaning that $\mathbb{E}[\rho(X^2\,;\, g(X_1;\theta_0,\tilde{h})] = o_p(n^{-1/2})$. Applying the expansion in the proof of Lemma 3, leads to the result of the theorem. ■

PARTIAL LINEAR MODEL

- The classic partial linear model (Robinson, 1988) is:

$$Y = X'\beta + g(T) + \epsilon$$

- When our moment function is linear, we have:

$$Y = X_1'\beta + X_2'\beta(Z) + \epsilon$$

- To see the connection, when $X_2 = 1$ we obtain:

$$Y = X_1'\beta + \beta(Z) + \epsilon$$

- Our estimator can produce this model as an endogenously-selected outcome

PARTIAL LINEAR MODEL

- We would like to expand the approach to account for arbitrary partially linear models
- $Y = W_1'\beta + g(W_2) + \epsilon$
- Goal is to classify W into two components: linear component and the nonparametric component
- This is different that what we were doing previously, since we are no longer imposing any linearity in the g function
- How to do this?
  - Outer loop still looking at estimating linear component
  - Inner loop is no longer a moment forest but rather a random forest with regression trees (each leaf is a constant)
- Econometrically we need to ensure that g converges at fast enough rate so that usual semiparametric theory follows through

CLUSTERED STANDARD ERRORS / HETEROSCEDASTICITY

- Our approach applies to the clustered standard errors literature as well
- Basically the idea is that we want to allow for correlations in the error terms in a model
- For example, in the linear model:

$$Y = X'\beta + \epsilon$$

- We may think that the error terms are correlated
- The general way of approaching this is to put a data-generating process on the error:

$$\epsilon \sim N(0, \sigma^2(X))$$

- Note that this is a partitioning problem!
  - Outer loop, search for first-order parameters
  - Inner loop, run the random forest
- What are we matching? Out-of-sample correlation of error terms!

RANDOM COEFFICIENTS

- The random coefficient model (ala BLP (1995)) is:

$$u_{ij} = x'\beta + x'\nu_i + \xi_j + \epsilon_{ij}$$

- This generates correlation in utilities across products that have similar characteristics
- Utility equation -> probabilities over choice set -> aggregate market shares
- We want to determine which components are fixed and which are random
- This is the higher-order version of the moment functions that we discussed previously

RANDOM COEFFICIENTS: BASIC OVERVIEW OF BLP

- We are concerned that price and unobserved vertical quality are correlated
- However, need to get share equations in linear form so that we can apply IV
- BLP is about doing those gymnastics
- Share equation is super nonlinear:

$$s_j = \int s_j(X;\beta)dF(\beta)$$

- Key point: integrate out the nonlinear part and then use the fact that any vector of shares can be rationalized by a unique vector of numbers, δ (mean utilities)
- Regress mean utilities on X's, solve out for the unobservable, ξ, that rationalizes aggregate shares, minimize:

$$\min_{\sigma(X)} E[\xi(\sigma(X))'Z]$$

RANDOM COEFFICIENTS: BASIC OVERVIEW OF BLP

- We are concerned that price and unobserved vertical quality are correlated
- However, need to get share equations in linear form so that we can apply IV
- BLP is about doing those gymnastics
- Share equation is super nonlinear:

$$s_j = \int s_j(X;\beta)dF(\beta)$$

- Key point: integrate out the nonlinear part and then use the fact that any vector of shares can be rationalized by a unique vector of numbers, δ (mean utilities)
- Regress mean utilities on X's, solve out for the unobservable, ξ, that rationalizes aggregate shares, minimize:

$$\min_{\sigma(X)} E[\xi(\sigma(X))'Z]$$

RANDOM COEFFICIENTS: MOMENT FOREST APPROACH

- Key observation is that a random coefficient model can be consistently estimated a point using the non-RC basis functions
- In this case, the basis function is the logit function
- Note that, at a given set of characteristics:

$$s_j\left(\hat{\beta}\middle|X = x\right) = \int s_j(X, \beta) dF(\beta)$$

- The implication of this is that the moment forest will only split on variables which have a random coefficient!
- The resulting approximation will show two things:
  - Which dimensions have random coefficients
  - A low-order approximation to the integral (digression on basis functions, quadrature, FKRB, and grid methods)
- One could use the moment forest solely as classifier, run FKRB on data with restricted basis afterwards
- Seems like there is a lot of information in moment forest, however! How to use?

MONTE CARLO SETUP

- To showcase the performance of the estimator, we consider three data-generating process
- Linear model with two explanatory variables, three Z

$$Y = X'\beta(Z) + \epsilon$$

- Error is standard normal (so irreducible mean prediction error cannot be lower than 1.0)

1. Fully heterogeneous parameters: both parameters vary with Z: $\beta = \{-1,1\}$ if $Z_1 < 0$, $\{0.33, -1.05\}$ otherwise
2. One heterogeneous parameter: parameter on X2 varies with Z: $\beta = \{-1,1\}$ if $Z_1 < 0$, $\{-1, -1.05\}$ otherwise
3. Fully homogeneous parameters: parameters are fixed across Z

Table 1: Monte Carlo Results: Linear Model, No Parameters Homogeneous

		NP	OLS	Unrestricted Model		Restricted Model		
								Classification
n				Mean	S.D.	Mean	S.D.	Rate
500	$MSE(Y)$	2.16	3.14	1.07	(0.00606)	1.07	(0.00606)	
	$MSE(\beta)$		1.49	0.0433	(0.000913)	0.0433	(0.000913)	
	β_0		-0.316			0.00	(0.00)	0.00%
	β_1		-0.00809			0.00	(0.00)	0.00%
1000	$MSE(Y)$	2.00	3.13	1.04	(0.00231)	1.04	(0.00231)	
	$MSE(\beta)$		1.46	0.0230	(0.000122)	0.0230	(0.000122)	
	β_0		-0.324			0.00	(0.00)	0.00%
	β_1		-0.0129			0.00	(0.00)	0.00%
2000	$MSE(Y)$	1.87	3.12	1.02	(0.00104)	1.02	(0.00104)	
	$MSE(\beta)$		1.46	0.00906	(1.43e-05)	0.00906	(1.43e-05)	
	β_0		-0.330			0.00	(0.00)	0.00%
	β_1		-0.00962			0.00	(0.00)	0.00%
4000	$MSE(Y)$	1.61	3.12	1.01	(0.000565)	1.01	(0.000565)	
	$MSE(\beta)$		1.45	0.00463	(9.08e-06)	0.00463	(9.08e-06)	
	β_0		-0.343			0.00	(0.00)	0.00%
	β_1		0.00673			0.00	(0.00)	0.00%

These results are for the linear model without any homogeneous parameters. The numbers for the unrestricted model and restricted model are identical because the model selection procedure never classified either of the parameters are homogeneous. The irreducible squared error in Y is equal to 1.0.

- First column is nonparametric Nadaraya-Watson (NW) estimator
- Moment Forest recovers specification from smallest sample size (uniformly correct classification)
- Very close to irreducible error
- Large advantage in MSE(Y) against NW
- No advantage here for restricted model
- OLS is terrible

Table 2: Monte Carlo Results: Linear Model, One Parameter Homogeneous

		NP	OLS	Unrestricted Model		Restricted Model		
n				Mean	S.D.	Mean	S.D.	Classification Rate
500	$MSE(Y)$	2.48	4.03	1.07	(0.00556)	1.07	(0.00591)	
	$MSE(\beta)$		1.06	0.0296	(0.000222)	0.0271	(0.000274)	
	β_0		-0.987			-0.995	(0.0619)	100%
	β_1		-0.00475			0.00	(0.00)	0.00%
1000	$MSE(Y)$	2.27	4.03	1.04	(0.00257)	1.04	(0.00341)	
	$MSE(\beta)$		1.03	0.0172	(5.93e-05)	0.0137	(4.00e-05)	
	β_0		-0.993			-0.999	(0.0394)	97.5%
	β_1		-0.0134			0.00	(0.00)	0.00%
2000	$MSE(Y)$	2.08	3.99	1.02	(0.000845)	1.02	(0.000794)	
	$MSE(\beta)$		1.02	0.00785	(2.55e-05)	0.00697	(1.29e-05)	
	β_0		-0.997			-1.00	(0.0287)	100%
	β_1		-0.0106			0.00	(0.00)	0.00%
4000	$MSE(Y)$	1.78	3.99	1.01	(0.000571)	1.01	(0.000557)	
	$MSE(\beta)$		1.01	0.00320	(2.42e-06)	0.00275	(2.51e-06)	
	β_0		-1.01			-1.00	(0.0181)	100%
	β_1		0.00637			0.00	(0.00)	0.00%

These results are for the linear model with one ($\beta_1 = -1.0$) homogeneous parameter. Column NP refers to the nonparametric Nadaraya-Watson estimator. OLS is linear regression.

- Now one parameter is homogeneous
- Moment Forest again recovers specification from smallest sample size (uniformly correct classification)
- Very close to irreducible error
- Large advantage in MSE(Y) against NW
- No advantage here for restricted model in MSE(Y), but ~15 percent smaller MSE(β)
- OLS is terrible

Table 3: Monte Carlo Results: Linear Model, All Parameters Homogeneous

		NP	OLS	Unrestricted Model		Restricted Model		
								Classification
n				Mean	S.D.	Mean	S.D.	Rate
500	$MSE(Y)$	1.13	1.02	1.02	(0.00361)	1.02	(0.00368)	
	$MSE(\beta)$		0.00545	0.00611	(7.41e-05)	0.00542	(5.29e-05)	
	β_0		-0.996			-0.996	(0.0645)	100%
	β_1		0.999			0.999	(0.0356)	100%
1000	$MSE(Y)$	1.09	1.00	1.00	(0.00133)	1.00	(0.00133)	
	$MSE(\beta)$		0.00215	0.00235	(8.43e-06)	0.00214	(1.21e-05)	
	β_0		-0.999			-0.999	(0.0402)	100%
	β_1		1.00			1.00	(0.0230)	100%
2000	$MSE(Y)$	1.07	1.01	1.01	(0.000686)	1.01	(0.000673)	
	$MSE(\beta)$		0.00109	0.00111	(1.60e-06)	0.00108	(1.38e-06)	
	β_0		-1.00			-1.00	(0.0289)	100%
	β_1		1.00			1.00	(0.0148)	100%
4000	$MSE(Y)$	1.05	1.00	1.01	(0.000456)	1.00	(0.000455)	
	$MSE(\beta)$		0.000421	0.000450	(3.43e-07)	0.000412	(4.74e-07)	
	β_0		-1.00			-1.00	(0.0183)	100%
	β_1		1.00			1.00	(0.00861)	100%

These results are for the linear model with two homogeneous parameters ($\beta_1 = -1.0$, $\beta_2 = 1.0$).

- Now both parameters are homogeneous
- Moment Forest again recovers specification from smallest sample size (uniformly correct classification)
- Very close to irreducible error
- NW does surprisingly pretty well here
- No advantage here for restricted model in MSE(Y), but ~10-12 percent smaller MSE(β)
- OLS is correctly specified; moment forest still does better due to resampling!

MONTE CARLO EVIDENCE: PARTIALLY LINEAR MODEL

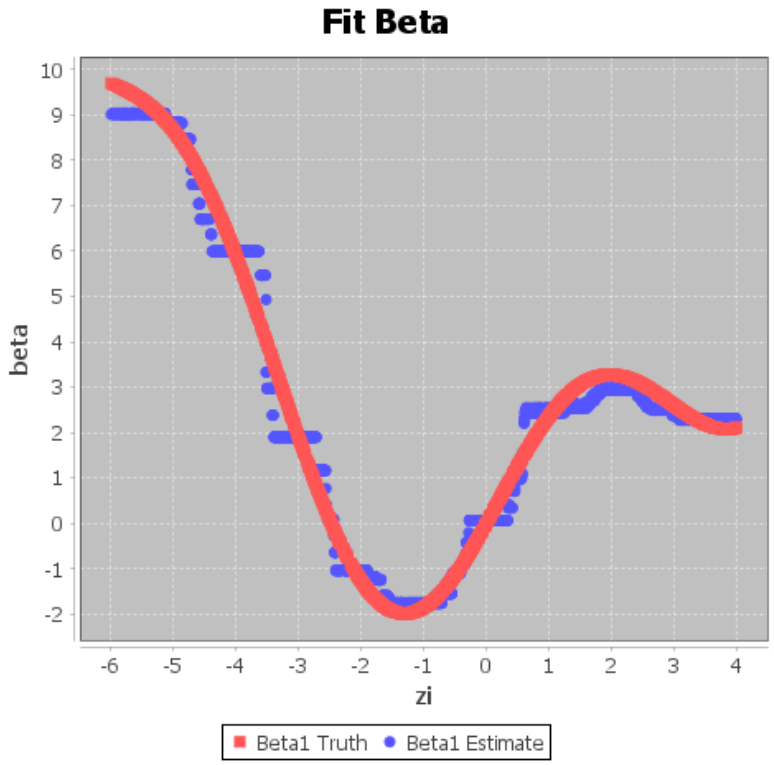
Table 4: Monte Carlo Results: Partially Linear Model

		NP	OLS	Unrestricted Model		Restricted Model		
								Classification
n				Mean	S.D.	Mean	S.D.	Rate
500	$MSE(Y)$	1.30	3.90	1.47	(0.0174)	1.47	(0.0167)	
	$MSE(\beta)$		2.86	0.445	(0.0111)	0.445	(0.0113)	
	β_0		0.260			0.00	(0.00)	0.00%
	β_1		0.990			0.996	(0.0463)	100%
1000	$MSE(Y)$	1.20	3.85	1.15	(0.00353)	1.15	(0.00390)	
	$MSE(\beta)$		2.84	0.146	(0.00133)	0.145	(0.00149)	
	β_0		0.259			0.00	(0.00)	0.00%
	β_1		0.998			0.999	(0.0232)	97.5%
2000	$MSE(Y)$	1.16	3.88	1.10	(0.00212)	1.10	(0.00198)	
	$MSE(\beta)$		2.84	0.0947	(0.000574)	0.0937	(0.000627)	
	β_0		0.248			0.00	(0.00)	0.00%
	β_1		1.00			1.00	(0.0165)	92.5%
4000	$MSE(Y)$	1.12	3.85	1.06	(0.000664)	1.06	(0.000650)	
	$MSE(\beta)$		2.83	0.0573	(0.000130)	0.0569	(0.000147)	
	β_0		0.246			0.00	(0.00)	0.00%
	β_1		1.00			1.00	(0.00881)	72.5%

These results are for the partially linear model: $Y = \beta_1(Z) + \beta_2 X_2$, with $\beta_1(Z) = 2.5 \sin Z + 0.25Z^2$ and $\beta_2 = 1.0$. Mean squared error in fitted Y and the true β are reported for both the unrestricted model (no homogeneity tested/imposed) and the restricted model (homogeneity tested/imposed).

- Results:
  - Get (nearly!) 100% perfect classification of both parts
    - I'm not sure what's going on with the n=4000 homogeneous classification rate
  - Models converge to irreducible error (1.0)
  - Slight but noticeable gain in precision
  - Posit: more stark when the dimensionality gets beyond 2
  - Possible efficiency gains will be higher when I implement orthogonalization

MONTE CARLO EVIDENCE: PARTIALLY LINEAR MODEL



- Make X1 a complex function now:
- 2.5*Math.sin(zi.get(0,0)) + 0.25*Math.pow(zi.get(0,0),2)

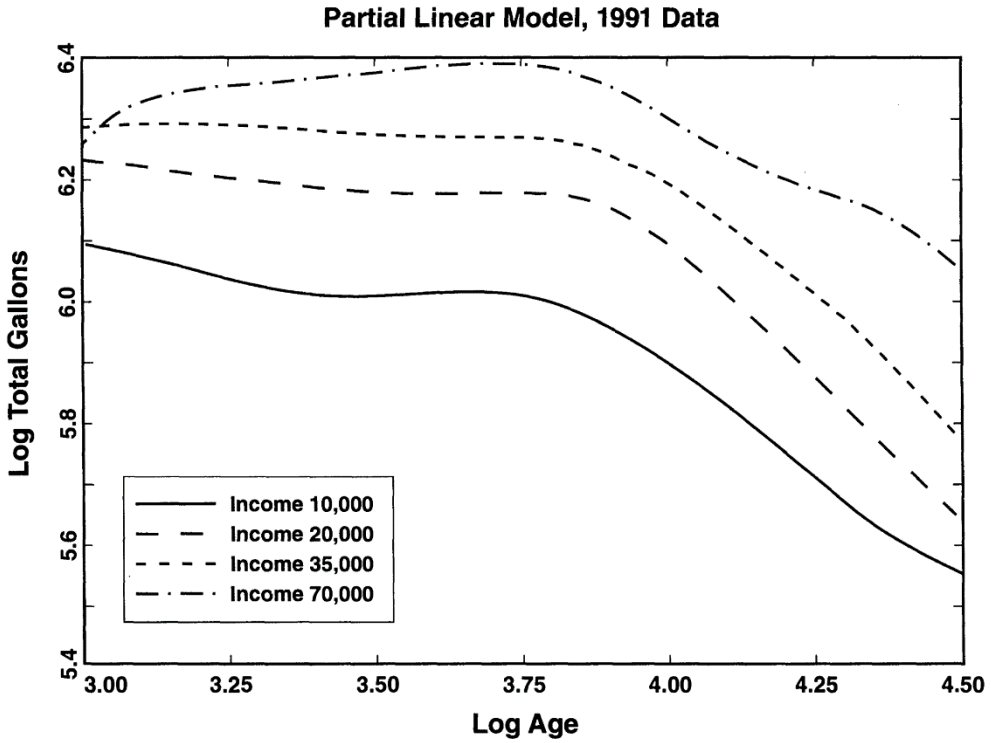
APPLICATION: GASOLINE ENGEL CURVES

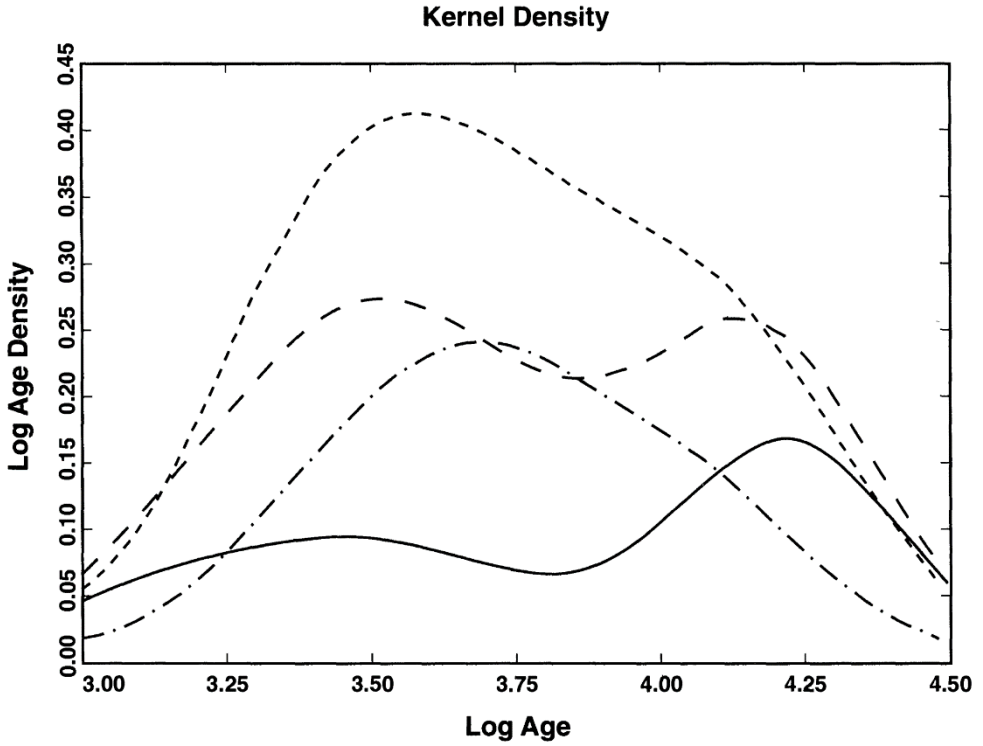
- Schmalensee and Stoker (ECMA, 1999) estimate semiparametric models of household gasoline demand
- In homage to this pioneering use of the semiparametric framework, we perform their analysis using updated data
  - Data comes from 1997 National Household Travel Survey
  - Detailed household data on age, income, vehicles, gasoline usage
- Their approach is to estimate the following consumption equation (in logs of gallons of gasoline):

$$\ln gallons = g(age, income) + x'\beta + \epsilon$$

- Where there are a vector of continuous and categorical variables in the linear component
- Our approach will mimic their empirical function. We want to answer two questions:
  - What components should be in the linear component and which should be in the nonparametric component?
  - What do the nonparametric functions looks like? Especially the Engel curve of consumption against age for various incomes

ORIGINAL S&S

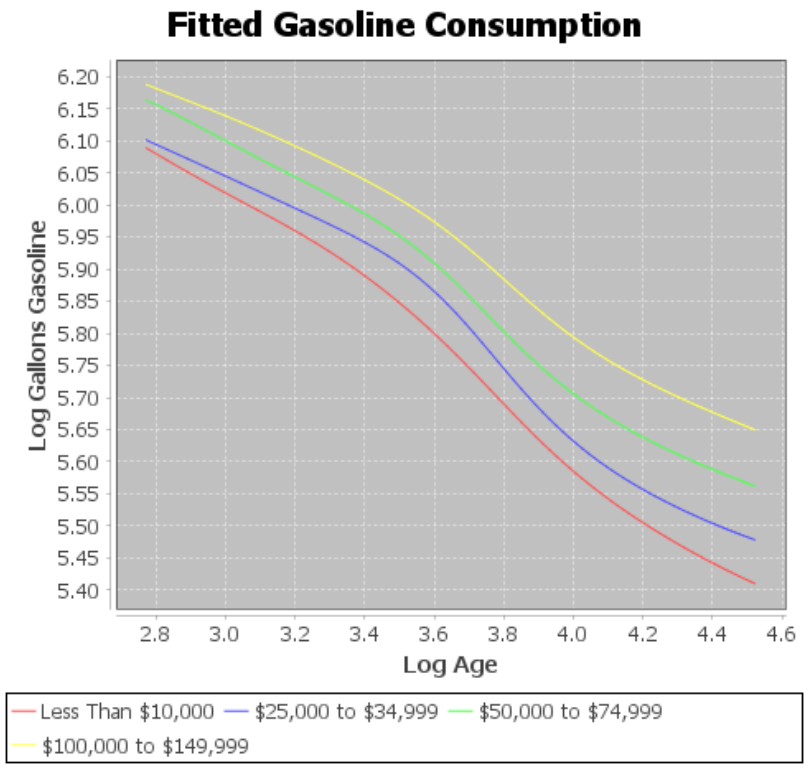

Partial Linear Model, 1991 Data
Log Total Gallons
6.4
6.2
6.0
5.8
5.6
5.4
3.00
3.25
3.50
3.75
4.00
4.25
4.50
Log Age
Income 10,000
Income 20,000
Income 35,000
Income 70,000

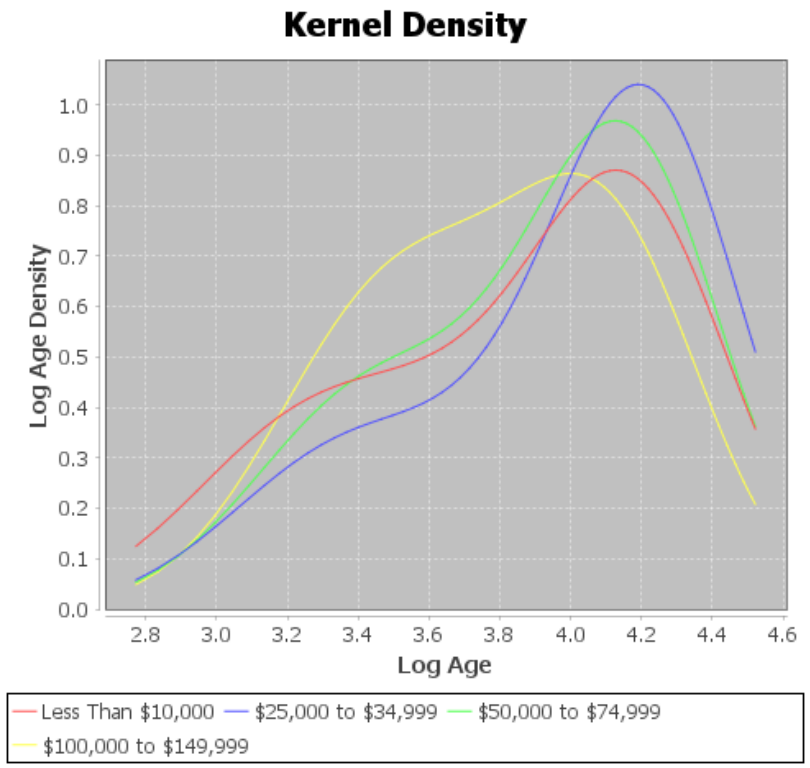

Kernel Density
Log Age Density
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
3.00
3.25
3.50
3.75
4.00
4.25
4.50
Log Age

SUMMARY OF FINDINGS

- We recreate their specification
- I allow the moment forest to split on all the X's (Z=X)
- Digression on dummy variables
- Ask the forest what variables it splits on and how often across trees
- What do I find?
- In 5% (comparable to their original sample) and 20% sample: all the parameters are homogeneous!

APPLICATION: GASOLINE ENGEL CURVES





THANK YOU!

- We propose an estimator for determining parametric and semiparametric components of nonlinear models
- We use a moment forest as a classifier and estimator (although one could use any well-behaved approach here)
- Show sufficient conditions that this approach works
- Excited about the application of the framework to some real empirical settings
  - Clustered standard errors
  - Random coefficients
  - Partially linear models

OBSERVATIONS

- Estimation with categorical variables can be treated as entering the model through a constant as a function of Z, where the categorical variables are in Z but not directly in X
- Also, can treat the categorical variables as continuous. Tree will partition if it needs to down to single groups. Much, much faster than all the combinations of bipartite partitions.
- For orthogonalization, in a sense our estimator is telling us what the structure of the model is; you can then use the two-step methods ala Robinson and S&S once that is done
- Can we formally show the equivalence of splitting on X and post-tree construction homogeneity testing?