A Framework for Learning $g \times e$ from Data and Application to Household Stress

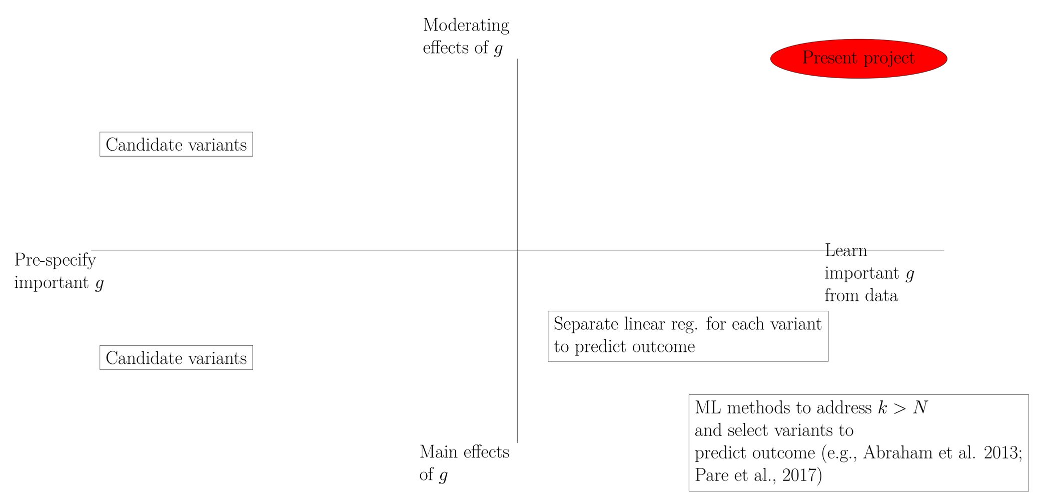
Rebecca Johnson, Lisa Schneper, Sara McLanahan, Dalton Conley, and Daniel Notterman

Princeton University

Definitions and specific case

1. **Main effect of an environment (e) on an outcome:** does household economic stress increase child behavior problems (internalizing anxious/withdrawal behaviors)?
2. **Main effect of a genetic variant (g) on an outcome:** does between-child variation in single nucleotide polymorphism (SNP) minor allele counts (e.g., $AA > AT > TT$) increase/decrease child behavior problems?
3. **Genetic moderation ($g \times e$):** does genetic variation (# 2) moderate the effect of household stress (# 1) on child behaviors? Two types of moderation:
   1. *Exacerbating:* stress has a larger effect on children with a higher minor allele count for the variant
   2. *Buffering*: stress has a smaller effect on children with a higher minor allele count

Motivation: learn $g \times e$ from data



Complements three existing approaches to $g \times e$:

1. Interact candidate variants with e: researchers use theory to select a small set (e.g., 1-2) variants that are thought to moderate an environmental stressor (e.g., Beaver 2013; Kim & Weeland, 2015)
2. Interact polygenic scores (PGSs) optimized to predict mean levels of a trait with e
3. Quantify variability in the outcome possibly caused by unobserved $g \times e$ (e.g., Yang et al. 2012; Dumitrascu et al. 2015; Conley et al. 2018)

Data, Feature Selection, and Estimation

1. Process features

Data: Fragile Families and Child Wellbeing Study (FFCW)
$N = 2813$; all analyses stratified by race and conducted in 80% training set
Features: ~ 592,200 variants measured using Illumina PsychChip

Randomly select 1000 SNPs | Linkage-disequilibrium based pruning (for pairs of correlated SNPs ($r^2 > 0.25$) remove one from pair; random selection of 1000 *see: Next Steps* | Feature filtering

2. Interact treatment (worsening consumer sentiment index pre-interview) with minor allele count for SNP $1, 2, \dots k$ (de-meaned by city and survey wave fixed effects):
3. Create feature matrix composed of treatment (e main effects), SNPS (g main effects), and 3. treatment × SNP interactions from step #2
4. To address $k > N$, use LASSO (Tibshirani, 1996); implemented using `R glmnet` to perform variable selection on #3 when predicting behavior problems; chose λ using 5-fold cross-validation. More formally, where β represents a coefficient on a term, k indexes a coefficient, n represents the number of participants, i indexes a participant, p indicates a vector of term predictors, lasso aims to solve the following constrained minimization problem: $min(\frac{1}{n}\Sigma_{i=1}^{N}(\text{behaviors}_i - \beta_0 - p_i^T\beta)^2)$. Subject to: $\Sigma_{k=1}^{p}|\beta_k| < \alpha$. We used 5-fold cross-validation–fitting the model on a portion of the data and using the coefficients to predict in a test portion–to select a regularization parameter.

Predictors

treat	snp1	snp2	…	snpk	*treat : snp1*	*treat : snp2*	…	*treat : snpk*
1	1	0	…	2	1	0	…	2
0	1	1	…	0	0	0	…	0
⋮								
1	2	0	…	1	1	2	…	0

5. Genetic moderator k satisfies criteria:

1. **Present in model where λ retains $\beta_{stress} \neq 0$**
2. **SNP has non-zero main effect on behavior problems**: $\beta_k \neq 0$
3. **Non-zero interaction effect between SNP and stressor**: $\beta_{kt} \neq 0$
4. **SNPs that increase behavior problems have exacerbating effects; SNPS that decrease behavior problems have buffering effects**: $sign(\beta_k) = sign(\beta_{kt})$
5. Replicates in held-out 20% test set

Does Increased Household Stress Cause More Child Behavior Problems? (e)

Approach: to control for confounders associated with household stress and child outcomes (e.g., parent genotype), exploit random variation in when families are surveyed that exposes the families to fluctuations in the national consumer sentiment index (NCSI) in the 3-months prior to the interview that others have shown increases harsh parenting (Lee et al., 2013) and partner violence (Schneider et al. 2015)

Figure: *Left panel*: shows within-city variation in interview timing relative to the NBER Great Recession window; *Right panel*: shows fluctations in consumer sentiment during that period, with the jittered colors at the bottom referring to respondents from different sample cities.

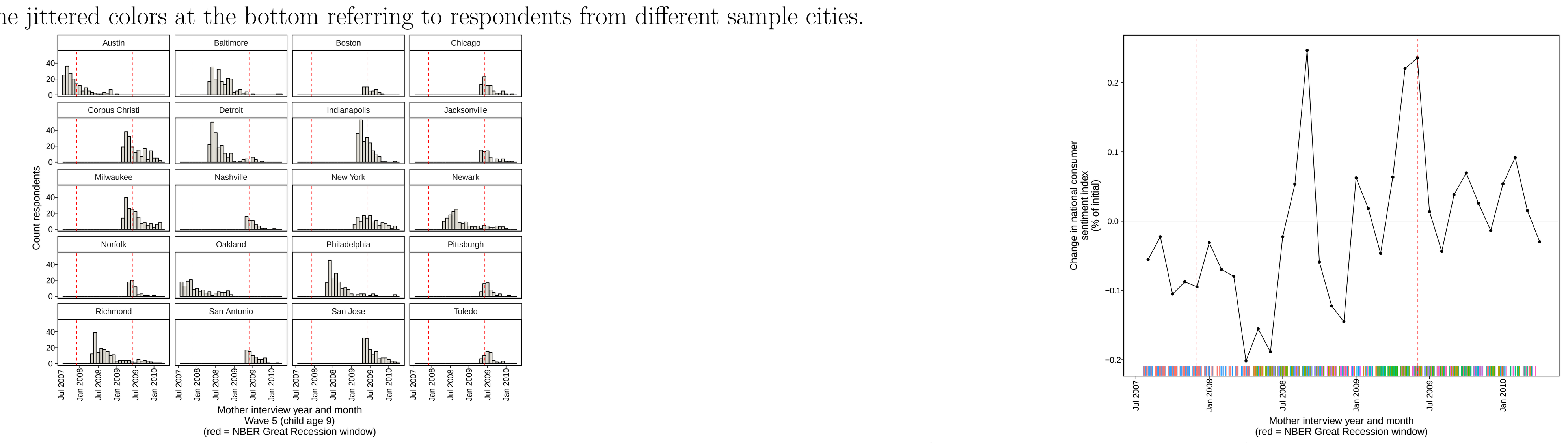
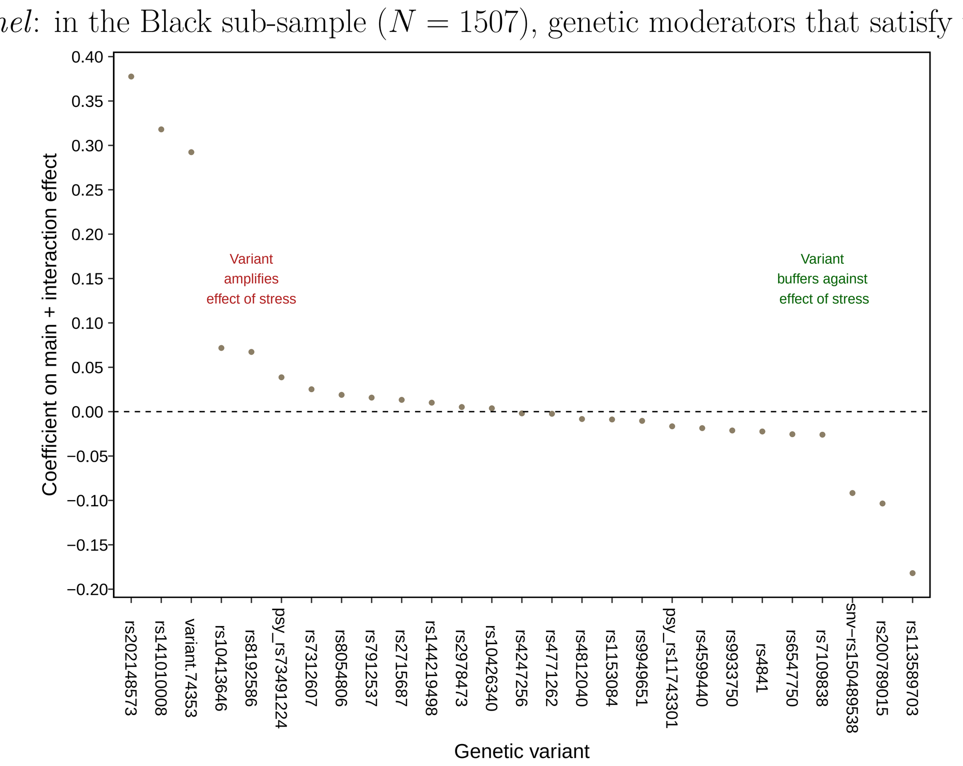


Figure: *Left panel*: in the full-sample, shows a strong main effect of deteriorating consumer sentiment (and increased household stress) on increasing behavior problems using a linear regression; *Right panel*: for the Black sub-sample (focus of present results), shows positive coefficient value at range of λ on binary measure of household stress (worsening or not) using LASSO

Which Genetic Variants Exacerbate or Buffer the Effects of this Stress ($g \times e$)?

Figure: *Left panel*: in the Black sub-sample ($N = 1507$), genetic moderators that satisfy the first four conditions; *Right panel*: biological functions of selected important moderators



- Exacerbating:
  - *rs202148573*: located on gene (BHLHE41) that 'is believed to be involved in the control of circadian rhythm and cell differentiation. Defects in this gene are associated with the short sleep phenotype' *(Entrez Gene Summary)*
- Buffering:
  - *rs113589703*: located on gene (ANP32D) that is a 'a tumor suppressor that can inhibit several types of cancers, including prostate and breast cancers' *(Entrez Gene Summary)*

Limitations and next steps

- **Main limitation:**: learns treatment-specific moderators rather than moderators that generalize across treatments/environments
- More efficient implementation of LASSO or implementation of variable screening techniques (e.g., Sure Independence Screening (SIS) (Fan and Lv, 2007) to allow for LD-based pruning of variants *without* further sub-sampling prior to model estimation
- Explore other variable-selection methods for heterogeneous treatment effects (e.g., Ratkovic and Tingley, 2017; Wager and Athey, forthcoming)
- Aggregate weights into a 'variance' polygenic score (vPGS) and compare performance as interaction term to other vPGS constructed using other methods