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

**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), explicit 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 (Schaule et al., 2015).

**Figure**: Left panel: shows extimute variation in interview timing relative to the NBER Great Recession window. Right panel: shows fluctuations in consumer sentiment during that period, with the vertical axis at the bottom referring to respondents from different sample sites.

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

**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 (measuring or net) using LASSO.

**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 Li, 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., Pinto and Tingley, 2017, Wagner and Athey, forthcoming)
- **Aggregate weights into a “variance” polygenic score (vPGS) and compare performance as interaction term to other vPGS constructed using other methods

**Data, Feature Selection, and Estimation**

1. **Process features**:
   - Data: Fragile Families and Child Wellbeing Study (FFCWS)
   - N = 2013, all analyses stratified by race and conducted in 80% training set
   - Features: ~500K variants measured using Illumina PsychChip
   - Feature filtering (p-value $< 0.25$) removes one pair from pair; random selection of 1000 sec. Next Steps
   - 200 SNPs

2. **Interact treatment (worsening consumer sentiment index pre-interview) with minor allele count for SNP 1, 2, ... k (de-measured by city and survey wave fixed effects)

3. **Create feature matrix composed of treatment (e main effects), SNPs (g main effects), and 3. treatment $\times$ SNP interactions from step #2**

4. **To address $k > N$, use LASSO (Tibshirani, 1996); implemented using $\ell^1$ norm to perform variable selection on $\#3$ when predicting behavior problems; chose $\lambda$ using 5-fold cross-validation. More formally, where $\beta$ represents a coefficient on a term, $k$ indexes a coefficient, $n$ represents the number of participants, $i$ indexes a participant, $p$ indexes a vector of term predictions, $\lambda$ aims to solve the following constrained minimization problem:**

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta' x_i) + \lambda \|\beta\|^1$$

Subject to $\sum_{k=1}^{k} |\beta_k| < \alpha$.

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.

5. **Genetic moderator $k$ satisfies criteria:**
   - Present in model where $\lambda$ retains $\beta_{g,e} \neq 0$
   - SNPs non-zero main effect on behavior problems $\beta_{e} \neq 0$
   - Non-zero interaction effect between SNP and stress $\beta_{g,e} \neq 0$
   - SNPs that increase behavior problems have exacerbating effects; SNPs that decrease behavior problems have buffering effects: $\text{snp}(g) \times \text{main + interaction effect}$

**Definitions and specific case**

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

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

- 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., Bear et al., 2013, Kim & Weiland, 2015)
- Interact polygenic scores (PGSs) optimized to predict mean levels of a trait with $e$
- Quantify variability in the outcome possibly caused by unobserved $g \times e$ (e.g., Yang et al., 2012, Dimmechanus et al., 2015, Conley et al., 2018)

**Motivation: learn $g \times e$ from data**