

cmenet

Simon Mak, Prof. C. F. Jeff Wu

Georgia Institute of Technology

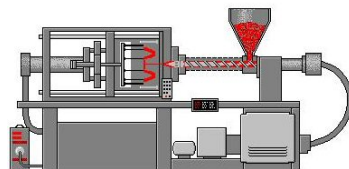
(Submitted to JASA)

Section 1

Motivation

What are CMEs?

- A **conditional main effect (CME)** captures the conditional effect of a factor at a fixed level of another factor
- CMEs are **interpretable** phenomena for many applications in genomics, social sciences and engineering:
 - CMEs can quantify the **activation behavior** of gene-gene interactions
 - CMEs also arise naturally in many **physical processes**, e.g., for injection molding, the effect of mold temperature only at a high level of holding pressure



History of CMEs

- First introduced in Wu (2015) as a way to **disentangle** effects which are **aliased** in a designed experiment
- For **designed experiments**, Su and Wu (2017) developed a variable selection method for CMEs
 - To obtain an **orthogonal** model, this method exploits the natural grouping of CMEs into twin, sibling and family effects



Our contribution

- For **observational data**, the goal is not to de-alias effects, but to separate **active** effects from correlated groups of **inert** effects
 - This motivates a **different CME grouping** than in Su and Wu (2017)
 - **Bi-level variable selection** is needed to perform **between-group** and **within-group** selection of CMEs



Section 2

Methodology

CME groupings

A CME is formally defined as follows:

Definition

Let $\tilde{\mathbf{x}}_j \in \{-1, +1\}^n$ be the covariate vector for **main effect (ME)** J , $j = 1, \dots, p$. The **CME** $J|K+$ quantifies the effect of covariate vector $\tilde{\mathbf{x}}_{j|k+} = \tilde{\mathbf{x}}_j \circ (\tilde{\mathbf{x}}_k > 0)$, where \circ is the Hadamard product.

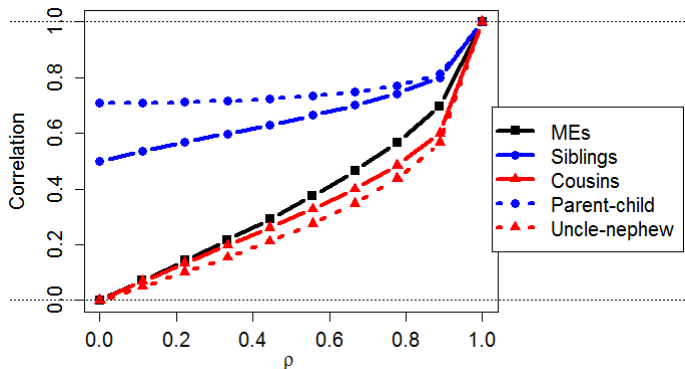
- J and K are the **parent** and **conditioned** effect of $J|K+$

Consider the following effect groups:

- **Siblings**: CMEs with same parent effect, e.g., $A|B+$ and $A|C+$
- **Cousins**: CMEs with same conditioned effect, e.g., $B|A+$ and $C|A+$
- **Parent-child**: A CME and its parent, e.g., $A|B+$ and A
- **Uncle-nephew**: A CME and its conditioned effect, e.g., $B|A+$ and A

CME groupings

- Suppose the main effects $\{\mathbf{x}_j\}_{j=1}^p$ are generated from normally distributed latent variables with **unit variance** and **covariance ρ**
- The figure below plots the **pairwise correlation** within each effect group:



cmenet: Bi-level variable selection criterion

We propose the following **selection criterion**:

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_S(\boldsymbol{\beta}) + P_C(\boldsymbol{\beta}) \right\}$$

- $\mathbf{y} \in \mathbb{R}^n$ is the **response** vector
- $\mathbf{X} \in \mathbb{R}^{n \times p'}$ is the **normalized** model matrix, with $p' = p + 4\binom{p}{2}$
- $\boldsymbol{\beta} \in \mathbb{R}^{p'}$ is the **coefficient vector**
- $P_S(\boldsymbol{\beta})$ and $P_C(\boldsymbol{\beta})$ are the **sibling** and **cousin** penalty functions:

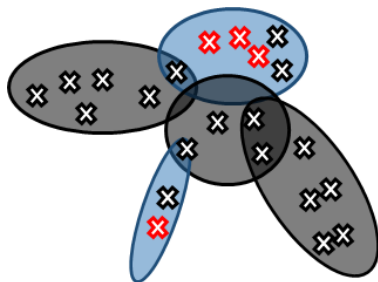
$$P_S(\boldsymbol{\beta}) = \sum_{j=1}^p f_{o,S} \left\{ \sum_{k \in \mathcal{S}(j)} f_{i,S}(\beta_k) \right\}, P_C(\boldsymbol{\beta}) = \sum_{j=1}^p f_{o,C} \left\{ \sum_{k \in \mathcal{C}(j)} f_{i,C}(\beta_k) \right\}$$

- $\mathcal{S}(j) = \{J, J|A+, J|A-, J|B+, J|B-, \dots\}$ is the **sibling group** of J
- $\mathcal{C}(j) = \{J, A|J+, A|J-, B|J+, B|J-, \dots\}$ is the **cousin group** of J

cmenet: Inner and outer penalties

The **outer** and **inner** penalties f_o and f_i parametrize the **bi-level selection** of CMEs:

- f_o controls **between-group** selection (e.g., selecting which CME groups are active)
- f_i controls **within-group** selection (e.g., selecting which CMEs are active within a group)



cmenet: CME coupling and reduction

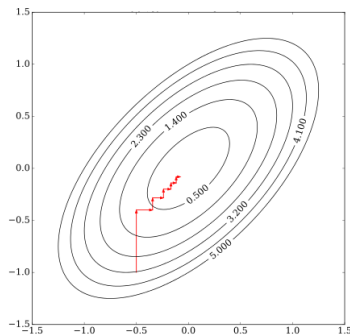
The proposed criterion offers two appealing features called **CME coupling** and **CME reduction**:

- CME coupling allows $J|K+$ to enter the model **more easily** if effects in $\mathcal{S}(j)$ or $\mathcal{C}(k)$ have been **selected**
 - **Intuition**: If $A|B+$ is active, then its siblings $A|C+$, $A|D+$, \dots are more likely to be active
- CME reduction allows the main effect J to enter the model **more easily** if effects in $\mathcal{S}(j)$ and $\mathcal{C}(j)$ have been **selected**
 - **Intuition**: When many siblings are selected, its parent effect is most likely active instead
- Parallels the principles of **effect heredity** and **effect hierarchy**, which are used to guide model selection in designed experiments

cmenet: Coordinate descent and threshold functions

The selection criterion $Q(\beta)$ can be efficiently minimized using a technique called **coordinate descent**:

- Idea is to minimize $Q(\beta)$ for β_1, β_2, \dots , cycling through this **iterative optimization** until β converges
- Computational efficiency comes from a **closed-form minimizer** for these iterative updates
- For $Q(\beta)$, a first-order **Taylor expansion** of f_o provides this closed-form minimizer in the form of a **threshold function**



Section 3

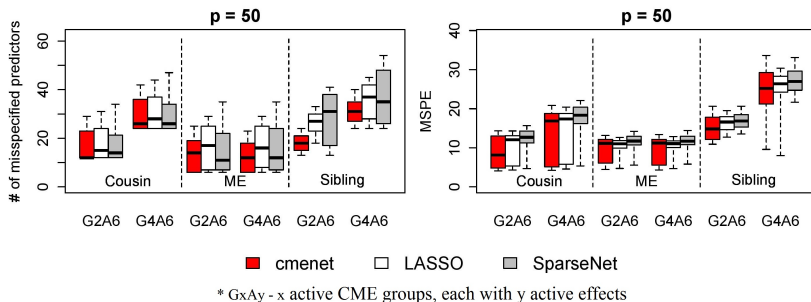
Simulations

Simulation set-up

The **simulation set-up** is as follows:

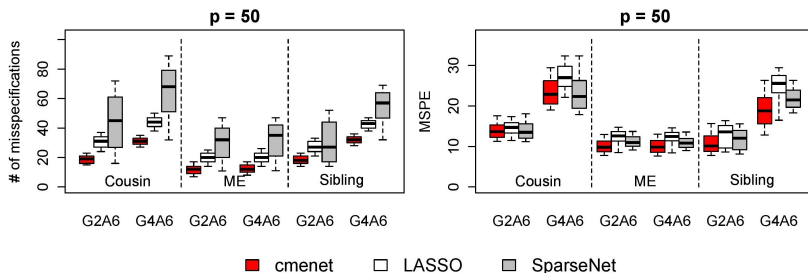
- $n = 20$ observations, $p = 25$ or $p = 50$ main effects
- Model matrix \mathbf{X} simulated from the equicorrelated normal distribution earlier, with $\rho = 0$ or $\rho = 1/\sqrt{2}$
- Effect type = siblings, cousins or main effects
- Varying # of active groups
- Varying # of active effects within a group
- cmenet is compared with **LASSO** (Tibshirani, 1996) and **SparseNet** (Mazumder et al., 2011), all performing selection on the same MEs and CMEs
- Judged on:
 - # of misspecified effects: false-positives and true-negatives
 - Mean-squared prediction error (MSPE)

No correlation ($\rho = 0$)



- cmenet provides **slightly improved** selection and prediction performance to LASSO and SparseNet
- Selection improvement most pronounced in the **sibling** case, which is not surprising given its **high correlations** at $\rho = 0$

Moderate correlation ($\rho = 0.7$)



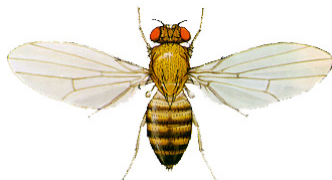
* GxAy - x active CME groups, each with y active effects

- cmenet provides **considerably improved** selection performance to LASSO and SparseNet
- This improvement can be seen in all three cases (ME, cousin, sibling), which is expected given the **grouped structure** for CME correlation is **most prominent** for moderate $\rho > 0$

Section 4

Gene association study

Background



- Single nucleotide polymorphisms (SNPs) serve as **biological markers** for many organism characteristics
- A data-driven exploration of this gene expression behavior provides valuable insight on the **underlying biological process**
 - Specifically, CME analysis reveals the **activation behavior** of gene-gene interactions, i.e., which genes are **conditionally active**, and which are important in **activating other genes**
- We illustrate cmenet on a **gene association** study for the wing shape of *Drosophila Melanogaster*, the common fruit fly
 - $n = 701$ observations collected from $p = 48$ polygene markers

Predictive error

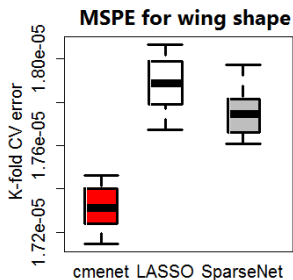


Figure: *Boxplots of the 10%, 25%, 50%, 75% and 90% quantiles for MSPE.*

- cmenet (with MEs and CMEs) is compared with LASSO and SparseNet
- The latter two perform selection on MEs and **two-way interactions**, which is typical for gene-gene interaction analysis
- cmenet provides the **best predictive model**, with its highest 10% error quantile sizably smaller than the lowest 10% for LASSO and SparseNet

Selected effects

	<i>Some selected effects</i>	
cv.cmenet	$V4 V1+, V4 V33+,$ $V10 V4+, V31 V4+$	$V43 V1-$
LASSO	$V4$	$V43, V1 * V43$
SparseNet	$V4$	$V43, V1 * V43$

- Both LASSO and SparseNet selected the fourth polygene $V4$, whereas cmenet selected the two **siblings** $V4|V1+$, $V4|V33+$ and the two **cousins** $V10|V4+$, $V31|V4+$
 - cmenet therefore gives a more **nuanced analysis** of the effect of $V4$, i.e., **conditionally active** under the 1st and 33rd gene, and **activates** the 10th and 31st gene
- Both LASSO and SparseNet selected $V43$ and the interaction $V1 * V43$, whereas cmenet selected only the CME $V43|V1-$
 - Precisely **Rule 1** of Su and Wu (2017)

Summary

- CMEs arise in many applications in [genomics](#), [social sciences](#) and [engineering](#)
- We propose a new method called `cmenet` for selecting CMEs in [observational data](#)
- `cmenet` exploits the underlying [grouped correlation structure](#) of CMEs for variable selection, specifically through the principles of [CME coupling](#) and [reduction](#)
- In both simulations and a gene expression study, `cmenet` provides [improved selection](#) and [prediction performance](#) over existing methods
- An efficient implementation of these algorithms is provided in the [R package](#) CMENET in CRAN