cmenet: a new method for bi-level variable selection of conditional main effects

Simon Mak & C. F. Jeff Wu

View supplementary material

Accepted author version posted online: 14 Mar 2018.

Submit your article to this journal

Article views: 81

View related articles

View Crossmark data

cmenet: a new method for bi-level variable selection of conditional main effects

Simon Mak[*], C. F. Jeff Wu[*][†][‡]

March 1, 2018

Abstract

This paper introduces a novel method for selecting main effects and a set of reparametrized effects called conditional main effects (CMEs), which capture the conditional effect of a factor at a fixed level of another factor. CMEs represent interpretable, domain-specific phenomena for a wide range of applications in engineering, social sciences and genomics. The key challenge is in incorporating the implicit grouped structure of CMEs within the variable selection procedure itself. We propose a new method, cmenet, which employs two principles called CME coupling and CME reduction to effectively navigate the selection algorithm. Simulation studies demonstrate the improved CME selection performance of cmenet over more generic selection methods. Applied to a gene association study on fly wing shape, cmenet not only yields more parsimonious models and improved predictive performance over standard two-factor interaction analysis methods, but also reveals important insights on gene activation behavior, which can be used to guide further experiments. Efficient implementations of our algorithms are available in the R package CMENET in CRAN.

Keywords: Conditional effects, coordinate descent, gene association, interaction analysis, variable selection

---

[*]School of Industrial and Systems Engineering, Georgia Institute of Technology
[†]Corresponding author

1 Introduction

This paper proposes a new method for selecting *main effects* (MEs) and a set of reparametrized effects called *conditional main effects* (CMEs) from observational data. A CME can be described as follows. Let A and B denote two binary factors with levels $+$ and $-$. The CME $A|B+$ is then defined as the effect A when effect B is at the $+$ level, and 0 when B is at the $-$ level. In words, such an effect quantifies the influence of A *only* when B is at the level $+$. The CME $A|B-$ can be defined analogously.

The appeal for CMEs as basis functions for variable selection comes from its interpretability in a wide range of applications, including genomics and the social sciences. For example, in gene association studies, where the goal is to identify important genetic contributions for a trait or disease, the CME $A|B+$ quantifies the significance of gene A *only when* gene B is present. Such conditional effects are biologically interpretable and meaningful, as noted in Chari and Dworkin (2013): "[the examination] of how one mutation behaves when in the presence of a second mutation forms the basis of our understanding of genetic interactions, and is part of the fundamental toolbox of genetic analysis." Viewed this way, the selection of CMEs can therefore serve as an effective tool for investigating the activation and inhibition behavior of gene-gene interactions, namely, which genes are *conditionally* active, and which are important in *activating* or *inhibiting* other genes. CMEs also arise naturally in many engineering applications. For example, in an injection molding experiment with two settings for mold temperature A and holding pressure B (pg. 352 of Montgomery, 2008), the CME $A|B+$ measures the effectiveness of mold temperature *only at* a high level of holding pressure. This conditional effect may be a result of material properties for the molding liquid, and the discovery of such effects can provide valuable insight on the injection process.

The idea of CMEs was first introduced in Wu (2015) as a way to disentangle effects which are fully-aliased (i.e., perfectly correlated) in a *designed* experiment. Ever since the

2

pioneering work of Finney (1945), it has been widely accepted in the design community that fully-aliased effects in a regular, two-level design cannot be "de-aliased" without adding more experimental runs. Such a belief was shown to be false in Wu (2015), where the author employed a reparametrization of these fully-aliased effects into CMEs, and allowed for the selection of the resulting conditional effects. A variable selection method for designed experiments is further developed in Su and Wu (2017), making use of the natural groupings of CMEs into so-called twin, sibling and family effects. In this paper, we generalize this CME selection framework to *observational data*, by exploiting the implicit structure of CMEs to form new effect groups and to motivate a novel penalized selection criterion.

For penalized variable selection methods, the usual procedure for two-level factors is to first normalize each factor to zero mean and unit variance (Tibshirani, 1997). Treating these rescaled factors as continuous variables, standard variable selection techniques using the l_1-penalty in LASSO (Tibshirani, 1997) or non-convex penalties (e.g., Frank and Friedman, 1993; Fan and Li, 2001; Zhang, 2010) can then be used to identify significant effects. For the problem at hand, however, such methods are inappropriate, because they do not account for the implicit group structure present in CMEs. Grouped selection techniques, such as the group LASSO (Yuan and Lin, 2006) or the overlapping group LASSO (Jacob et al., 2009), are also not suitable here, because such methods select *all* effects from an active group, whereas only a handful of effects may be active within a CME group.

In this light, a *bi-level* selection strategy is needed to select both *active CME groups* and *active effects within CME groups*. In recent years, there have been important developments on bi-level variable selection, including the sparse group LASSO (Wu and Lange, 2008; Simon et al., 2013) and the group exponential LASSO (Breheny and Huang, 2009; Breheny, 2015). We extend the latter framework here, because it allows us to encode within the penalization criterion two important selection principles called *CME coupling* and *CME reduction*. These two principles guide the search for good CME models, and can be seen as an extension of effect heredity and effect hierarchy (Wu and Hamada, 2009), two guiding

3

principles used for model selection in designed experiments.

The paper is organized as follows. Section 2 provides some motivation for the problem at hand, including the implicit collinearity structure of CME groups and its effect on selection inconsistency. Section 3 proposes a new penalization criterion for CME selection, and illustrates two appealing selection principles (CME coupling and CME reduction) encoded within this criterion. Section 4 introduces a coordinate descent optimization algorithm using threshold operators, and presents an efficient tuning procedure for penalty parameters. Section 5 outlines several simulations comparing the CME selection performance of `cmenet` to existing variable selection methods. Section 6 then demonstrates the usefulness of the proposed method in a gene association study, and Section 7 concludes with directions for future research.

2   Background and motivation

2.1   CME and CME groups

We first define some notation. Let $\mathbf{y} \in \mathbb{R}^n$ be a vector of n observations, and suppose p main effects are considered. For effect J, let $\tilde{\mathbf{x}}_j = (x_{1,j}, \cdots, x_{n,j}) \in \{-1, +1\}^n$ be its binary covariate vector, $j = 1, \cdots, p$. The tilde on $\tilde{\mathbf{x}}_j$ distinguishes the binary covariate from its normalized analogue $\mathbf{x}_j$, which is introduced later. A CME can then be defined as follows:

Definition 1 (Conditional main effect). *The* conditional main effect (CME) *of J given K at level $+$, denoted as $J|K+$, quantifies the effect of covariate vector $\tilde{\mathbf{x}}_{j|k+} = (\tilde{x}_{1,j|k+}, \cdots, \tilde{x}_{n,j|k+})$, where:*

$$\tilde{x}_{i,j|k+} = \begin{cases} \tilde{x}_{i,j}, & \text{if } \tilde{x}_{i,k} = +1 \\ 0, & \text{if } \tilde{x}_{i,k} = -1 \end{cases}, \quad \text{for } i = 1, \cdots, n.$$

The CME $J|K-$ can be defined in a similar manner.

Throughout this paper, the effects J and K are respectively referred to as the *parent effect*

4

A	B	$A\|B+$	$A\|B-$	$B\|A+$	$B\|A-$
+1	+1	+1	0	+1	0
+1	-1	0	+1	-1	0
-1	+1	-1	0	0	+1
-1	-1	0	-1	0	-1

Table 1: *Model matrix for the two MEs A and B, and its four CMEs $A|B+, A|B-, B|A+, B|A-$.*

and the *conditioned effect* of $J|K+$. Using this terminology, $J|K+$ quantifies the effect of parent J, given its conditioned effect K is at level $+$. For illustration, Table 1 shows the four possible CMEs constructed from two main effects A and B.

Restricted to two-level, fractional factorial designed experiments, Su and Wu (2017) identified three important CME groups for selecting an *orthogonal* model, in which active effects are orthogonal to each other. These three groups are: (a) *sibling* CMEs: CMEs with the same parent effect, (b) *twin* CMEs: CME pairs with the same parent and conditioned effect, but with the sign for the latter flipped, (c) *family* CMEs: CMEs with fully-aliased interaction effects. Leveraging this group structure, Su and Wu (2017) proposed three rules for selecting a parsimonious and orthogonal model. Rule 1 (the most important selection rule) relies on the two simple mathematical identities:

$$\tilde{\mathbf{x}}_{j|k+} = \frac{1}{2}(\tilde{\mathbf{x}}_j + \tilde{\mathbf{x}}_{j*k}) \quad \text{and} \quad \tilde{\mathbf{x}}_{j|k-} = \frac{1}{2}(\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_{j*k}). \tag{1}$$

Here, $\tilde{\mathbf{x}}_{j*k} = \tilde{\mathbf{x}}_j \circ \tilde{\mathbf{x}}_k$ is the covariate vector for the traditional two-factor interaction (2FI) $J*K$, where $\circ$ is the Hadamard (entry-wise) product. From (1), the CME $J|K+$ can then be viewed as an average of the main effect for J and the interaction effect for $J*K$; a similar interpretation holds for the CME $J|K-$. Motivated by this interpretation, Rule 1 of Su and Wu (2017) replaces a selected ME J and 2FI $J*K$ with either (a) the CME $J|K+$, if the signs for J and $J*K$ are identical and their effect magnitudes are similar, or (b) the CME $J|K-$, if the signs for J and $J*K$ are different and their effect magnitudes are

5

similar. Such a rule (along with Rules 2 and 3) allows for the disentangling of fully-aliased interaction effects in a designed experiment.

The above CME groupings, however, are not suitable for analyzing observational data, because an orthogonal model is most likely not attainable for this more general setting. Instead, by exploring the correlation structure of CMEs, the following new groupings can be derived:

1. *Sibling* CMEs: CMEs which share the same parent effect, e.g., $\{A|B+, A|B-, A|C+, A|C-, A|D+, A|D-, \cdots\}$. This is the same as in Su and Wu (2017).

2. *Parent-child* pairs: An effect pair consisting of a CME and its parent ME, e.g., $\{A|B+, A\}, \{A|C+, A\}, \cdots$.

3. *Cousin* CMEs[1]: CMEs which share the same conditioned effect, e.g., $\{B|A+, B|A-, C|A+, C|A-, D|A+, D|A-, \cdots\}$.

We first outline the justification for these groups in terms of collinearity, then discuss why such groupings are appealing from a selection consistency perspective.

2.2 Group structure for collinearity

To explore the group structure of CMEs, consider the following latent model for the main effects $\{\tilde{\mathbf{x}}_j\}_{j=1}^p \subseteq \{-1, +1\}^n$. Define the latent matrix $\mathbf{Z} = (z_{i,j})_{i=1\,j=1}^{n\,\,\,\,p} \in \mathbb{R}^{n \times p}$, where each row of $\mathbf{Z}$ is drawn independently from the equicorrelated normal distribution $\mathcal{N}\{\mathbf{0}, \rho\mathbf{J}_p + (1-\rho)\mathbf{I}_p\}$. Here, $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\mathbf{J}_p$ is the $p \times p$ matrix of ones, and $\rho \in [0, 1]$. We then assume the following form for the binary covariates $\{\tilde{\mathbf{x}}_j\}_{j=1}^p$:

$$\tilde{x}_{i,j} = \mathbf{1}\{z_{i,j} > 0\} - \mathbf{1}\{z_{i,j} \leq 0\}, \quad i = 1, \cdots, n, \; j = 1, \cdots, p. \tag{2}$$

---

[1]From a purely linguistic point-of-view, these effects are not cousins, because their parent effects are unrelated. However, the notion of cousin nicely encapsulates a weaker form of a sibling relationship, which is the intended meaning here.

6
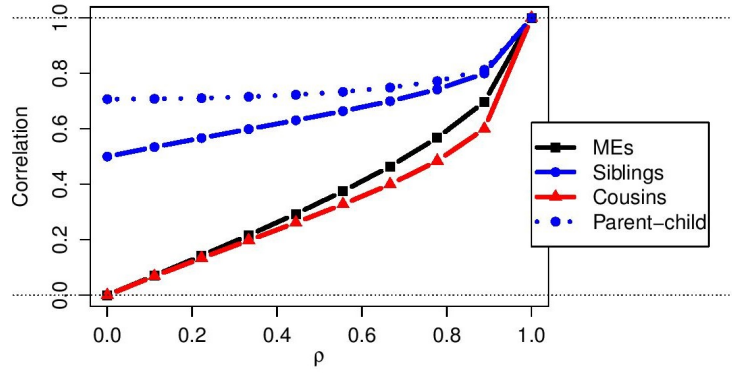
Figure 1: *Pairwise correlations within the four effect groups as a function of latent correlation ρ.*

Note that a larger value of ρ induces a higher correlation between the binary main effects.

Without loss of generality, assume here that the conditioned effects are set at the $+$ level for all CMEs. With the above model, the following theorem reveals an interesting group structure for CMEs. For brevity, proofs of all technical results are deferred to the Appendix.

Theorem 1 (Pairwise correlation within groups)**.** *Under the latent model* (2) *for main effects, the four effect groups have the following pairwise correlations:*

Group	Pairwise correlation		Group	Pairwise correlation
Main effects	$\frac{2\sin^{-1}\rho}{\pi}$		Parent-child	$\frac{1}{2\sigma_c}$
Siblings	$\frac{1}{\sigma_c^2}\left\{\frac{1}{4}+\frac{\sin^{-1}\rho}{2\pi}-\left(\frac{\sin^{-1}\rho}{\pi}\right)^2\right\}$		Cousins	$\frac{1}{\sigma_c^2}\left\{\frac{\sin^{-1}\rho}{\pi}-\left(\frac{\sin^{-1}\rho}{\pi}\right)^2\right\}$

where $\sigma_c^2 = 1/2 - (\sin^{-1}\rho/\pi)^2$.

Figure 1 plots the pairwise correlations in Theorem 1 as a function of the latent correlation parameter ρ. Two key observations can be made. First, the magnitudes of these correlations impose a natural hierarchy on the effect groups. For all values of $\rho \in (0,1)$, parent-child pairs have the largest correlations, followed by sibling pairs, then main effect and cousin pairs. Second, the correlation group structure can vary considerably for different choices

7

of ρ. In the independent setting of $\rho = 0$, sibling and parent-child pairs exhibit high correlations of 0.5 and $1/\sqrt{2}$ (≈ 0.71), respectively, whereas the remaining two groups have zero correlation. For moderately large choices of ρ, say, $\rho = 1/\sqrt{2}$ (≈ 0.71), these correlations become larger and more distinct between different groups, thereby amplifying the underlying CME group structure.

In light of this complex collinearity structure, one may suspect that standard variable selection techniques, such as the LASSO, would perform poorly for CME selection, because such methods impose the same regularization penalty over all variables, and ignore the implicit grouped correlation structure. This is indeed the case, and we demonstrate its poor selection performance in the following section and in the simulations of Section 5.

2.3   Selection inconsistency

An important property of a selection method is its *consistency* in choosing the correct model. Put mathematically, a method is (sign-)*selection consistent* if $\lim_{n\to\infty} \mathbb{P}(\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}) = 1$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the true coefficient vector, $\hat{\boldsymbol{\beta}}_n$ is the estimated vector from n observations, and $=_s$ denotes equality in sign (see Zhao and Yu, 2006 for a precise definition). The following theorem shows that LASSO is indeed inconsistent for simple CME models:

Theorem 2 (Selection inconsistency of LASSO)**.** *Under the latent model* (2)*, the LASSO is selection inconsistent in the following situations: (a) for $\rho \geq 0$, a model with $q \geq 3$ active siblings, (b) for $\rho \geq 0.27$, a model with $q = 2$ active main effects, and (c) for $\rho \geq 0.29$, a model with $q \geq 6$ active cousins.*

Theorem 2 demonstrates the poor selection of LASSO for simple CME models, even when little-to-no latent correlation is present. Part (a) says that, even in the uncorrelated setting of $\rho = 0$, LASSO yields poor selection whenever three (or more) siblings are present; part (b) says that, for mild correlations as low as 0.27, the same poor selection arises for two active MEs; part (c) says that, for correlations lower than 0.29, LASSO enjoys good selection

8

even when many cousins (up to 5) are active – this is not too surprising, because cousins experience the lowest pairwise correlations of the four groups. The proof of this theorem relies on the *irrepresentability condition* (Zhao and Yu, 2006), which shows that the LASSO is selection inconsistent when active variables are highly correlated with non-active ones.

3   `cmenet`: Penalization framework

To address these selection concerns, we propose a novel bi-level variable selection method called `cmenet`, which can identify both active CME groups and active effects within such groups. Similar to popular selection methods such as the elastic net (Zou and Hastie, 2005) and `SparseNet` (Mazumder et al., 2012), the name `cmenet` draws an analogy between the proposed method's ability to select active variables amongst non-active ones, and a fishing net's ability to catch larger fish amongst smaller ones. The penalization scheme for `cmenet` encodes two important principles, called *CME coupling* and *CME reduction*, which, as we show in this section, help guide the selection procedure for CMEs.

3.1   Selection criterion

We first introduce the selection criterion. Let $\mathbf{x}_j \in \mathbb{R}^n$ be the normalized vector for the binary main effect covariate $\tilde{\mathbf{x}}_j$, with $\mathbf{x}_j^T \mathbf{1}_n = 0$ and $n^{-1}\|\mathbf{x}_j\|_2^2 = 1$, along with a similar notation for CME covariates. Further let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_{p'}) \in \mathbb{R}^{n \times p'}$ be the full model matrix consisting of these normalized ME and CME effects, where $p' = p + 4\binom{p}{2}$ is the total number of effects considered. For simplicity, assume all considered effects are MEs and CMEs for the following exposition; Section 4.1.2 gives a simple extension for selecting these effects along with *other* covariate factors. Let $\boldsymbol{\beta} \in \mathbb{R}^{p'}$ be the coefficient vector, with β_j and $\beta_{j|k\pm}$ its corresponding coefficients for ME J and CME $J|K+$. Finally, assume that $\mathbf{y}$ is centered, i.e., $\mathbf{y}^T \mathbf{1}_n = 0$.

For effect groups, define $\mathcal{S}(j) = \{J, J|A+, J|A-, J|B+, J|B-, \cdots\}$ as the *sibling group*

for parent effect j, and $\mathcal{C}(j) = \{J, A|J+, A|J-, B|J+, B|J-, \cdots\}$ as the *cousin group* for conditioned effect j, $j = 1, \cdots, p$. We propose the following selection criterion, which can be viewed as an extension of the hierarchical framework in Breheny and Huang (2009):

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_{\mathcal{S}}(\boldsymbol{\beta}) + P_{\mathcal{C}}(\boldsymbol{\beta}) \right\},$$

$$P_{\mathcal{S}}(\boldsymbol{\beta}) \equiv \sum_{j=1}^{p} f_{o,\mathcal{S}} \left\{ \sum_{k \in \mathcal{S}(j)} f_{i,\mathcal{S}}(\beta_k) \right\}, \quad P_{\mathcal{C}}(\boldsymbol{\beta}) \equiv \sum_{j=1}^{p} f_{o,\mathcal{C}} \left\{ \sum_{k \in \mathcal{C}(j)} f_{i,\mathcal{C}}(\beta_k) \right\} \tag{3}$$

Here, $f_{o,\mathcal{S}}$ and $f_{i,\mathcal{S}}$ (similarly, $f_{o,\mathcal{C}}$ and $f_{i,\mathcal{C}}$) are *outer* and *inner* penalties which control the *between-group* and *within-group* selection for sibling (similarly, cousin) groups, respectively. While the specific penalty functions are left arbitrary in (3), we will introduce `cmenet` for the specific choice of the exponential penalty in Breheny (2015) for outer penalty, and the (scaled) minimax concave-plus penalty (MC+) in Zhang (2010) for inner penalty:

$$Outer: f_{o,\mathcal{S}}(\theta) = \eta_{\lambda_s,\tau}(\theta), \ f_{o,\mathcal{C}}(\theta) = \eta_{\lambda_c,\tau}(\theta), \ where \ \eta_{\lambda,\tau}(\theta) = \frac{\lambda^2}{\tau} \left\{ 1 - \exp\left(-\frac{\tau\theta}{\lambda}\right) \right\},$$

$$Inner: f_{i,\mathcal{S}}(\beta) = g_{\lambda_s,\gamma}(\beta), \ f_{i,\mathcal{C}}(\beta) = g_{\lambda_c,\gamma}(\beta), \ where \ g_{\lambda,\gamma}(\beta) = \int_0^{|\beta|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx. \tag{4}$$

This inner penalty is a scaled version of the MC+ penalty $\lambda g_{\lambda,\gamma}(\beta)$ in Zhang (2010) without the scaling factor λ; such a factor is accounted for in the outer exponential penalty $\eta_{\lambda,\tau}(\theta)$.

The appeal for the "exponential-MC+" framework in (4) is that it provides a concise parametrization of the grouped collinearity structure in Section 2. First, the penalty parameters $\lambda_s > 0$ and $\lambda_c > 0$ allow for differing regularization within sibling and cousin groups, respectively, with larger penalty values reducing the number of selected effects in each group. Assuming such parameters are tuned via cross-validation, a smaller tuned value of λ_s suggests many sibling effects are present in the data, while a smaller λ_c suggests the same for cousin effects. Second, the parameter $\gamma > 1$ controls the non-convexity of the inner MC+ penalty, and provides a "bridge" between the l_0-penalty (obtained when

10

$\gamma \to 1^+$) and the l_1-penalty in LASSO (obtained when $\gamma \to \infty$). In view of the selection problems for LASSO (see Theorem 2), such a parameter allows for improved selection of the highly correlated CMEs, say, within a sibling group. Lastly, the parameter τ provides two appealing principles called CME coupling and reduction, which we introduce below.

3.2   CME coupling and reduction

Consider first a CME $J|K+$ which has yet to be selected, and assume without loss of generality that $\mathbf{x}_{j|k+}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n > 0$. Taking the derivative of $Q(\boldsymbol{\beta})$ with respect to $\beta_{j|k+}$, and setting $\beta_{j|k+} = 0$ (as $J|K+$ is not in the model), we get:

$$\frac{\partial}{\partial \beta_{j|k+}} Q(\boldsymbol{\beta})\bigg|_{\beta_{j|k+}=0} = -\frac{1}{n} \mathbf{x}_{j|k+}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \Delta_{\mathcal{S}(j)} + \Delta_{\mathcal{C}(k)},$$
$$\text{where}\ \ \Delta_{\mathcal{S}(j)} = \lambda_s \exp\left\{-\frac{\tau\|\boldsymbol{\beta}_{\mathcal{S}(j)}\|_{\lambda_s,\gamma}}{\lambda_s}\right\}\ \text{and}\ \Delta_{\mathcal{C}(k)} = \lambda_c \exp\left\{-\frac{\tau\|\boldsymbol{\beta}_{\mathcal{C}(k)}\|_{\lambda_c,\gamma}}{\lambda_c}\right\}. \tag{5}$$

Here, $\boldsymbol{\beta}_g \in \mathbb{R}^{|g|}$ denotes the coefficient vector for an effect subset $g \subseteq \{1, \cdots, p'\}$, and $\|\boldsymbol{\beta}_g\|_{\lambda,\gamma} \equiv \sum_{l \in g} g_{\lambda,\gamma}(\beta_l)$ denotes its "norm" under the inner MC+ penalty. (For completeness, a full derivation of the subgradient for $Q(\boldsymbol{\beta})$ – which is quite technical and requires several applications of the chain rule – is found in equation (4) of the Appendix.)

Equation (5) reveals an appealing selection property of `cmenet` called CME coupling, which we describe below. Note that, when more effects have been selected in the sibling group $\mathcal{S}(j)$ (or cousin group $\mathcal{C}(k)$), the effect norms $\|\boldsymbol{\beta}_{\mathcal{S}(j)}\|$ (or $\|\boldsymbol{\beta}_{\mathcal{C}(k)}\|$) become larger. This then results in a smaller linearized slope $\Delta_{\mathcal{S}(j)}$ (or $\Delta_{\mathcal{C}(k)}$), which generates a decrease in the derivative $\frac{\partial}{\partial \beta_{j|k+}} Q(\boldsymbol{\beta})$ in (5). Since the goal is to minimize the selection criterion $Q(\boldsymbol{\beta})$, a smaller derivative allows for greater decrease in $Q(\boldsymbol{\beta})$ when $\beta_{j|k+}$ enters the model. In other words, the CME $J|K+$ has a greater chance of entering the model when other effects in its sibling group $\mathcal{S}(j)$ or its cousin group $\mathcal{C}(k)$ have already been selected; the selection of sibling or cousin effects can *couple* in the selection of the CME $J|K+$. We call

11

this property *CME coupling*, following the idea of effect coupling in Breheny (2015).

Consider next a ME J which has yet to be selected, and assume again that $\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n > 0$. Taking the derivative of $Q(\boldsymbol{\beta})$ with respect to β_j, and setting $\beta_j = 0$ (as J is not in the model), we get:

$$\frac{\partial}{\partial \beta_j} Q(\boldsymbol{\beta})\Big|_{\beta_j = 0} = -\frac{1}{n}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \Delta_{\mathcal{S}(j)} + \Delta_{\mathcal{C}(j)}. \tag{6}$$

The interpretation of equation (6) is similar to that for (5). When more effects have already been selected in the sibling group $\mathcal{S}(j)$ (or the cousin group $\mathcal{C}(j)$), the linearized slopes $\Delta_{\mathcal{S}(j)}$ (or $\Delta_{\mathcal{C}(j)}$) become smaller, which then decreases the derivative $\frac{\partial}{\partial \beta_j} Q(\boldsymbol{\beta})$ in (6). Hence, the ME J enters the model more easily when effects in its sibling group $\mathcal{S}(j)$ or its cousin group $\mathcal{C}(j)$ have already been selected; the selection of many sibling or cousin effects can then *reduce* to its underlying main effect. We refer to this phenomenon as *CME reduction*.

The notions of CME coupling and reduction are quite intuitive to expect in many CME applications. Consider the gene expression example in the Introduction, where the selection of the CME $A|B+$ indicates the effectiveness of gene A only when gene B is present. When several sibling CMEs of A, say, $A|B+$ and $A|C+$, are already selected in the model, one naturally expects gene A to be conditionally active under more genes as well. In other words, conditional effects with parent A are more likely to be active compared to conditional effects with no selected siblings – this is precisely the principle of CME coupling. However, when many sibling effects of gene A have already been selected, one may suspect that the underlying parent effect for gene A is active instead of these selected siblings – this is precisely the principle of CME reduction. A similar intuition holds for cousin effects.

An interesting parallel can also be made connecting CME coupling and reduction with the two guiding principles for model selection in designed experiments (Wu and Hamada, 2009). The first principle, called (weak) *effect heredity*, states that higher-order interactions can be selected only when either of its parent main effects are in the model. This idea is

quite similar to CME coupling, which allows for easier selection of a CME when effects with either the same parent or conditioned ME have been selected. Furthermore, note that a CME can be interpreted as a component of an interaction effect, because the difference of the two CMEs $A|B+$ and $A|B-$ is precisely the two-factor interaction $A * B$ (Su and Wu, 2017). Coupling can therefore be seen as an *extension* of effect heredity, after breaking an interaction effect (which is often difficult to interpret) into more interpretable conditional effects. The second principle, called *effect hierarchy*, states that lower-order interactions are more likely active than higher-order ones. This is akin to CME reduction, which encourages the reduction of selected sibling (or cousin) CMEs to its parent (conditioned) effect when too many siblings (cousins) are in the model.

4   `cmenet`: Optimization framework

With the proposed penalty $Q(\boldsymbol{\beta})$ in hand, we now present an optimization framework for `cmenet` in three parts. We first introduce the optimization algorithm for minimizing $Q(\boldsymbol{\beta})$, then describe several computational techniques for tuning penalty parameters, and finally conclude with several novel CME screening rules for speeding up the tuning procedure.

4.1   Optimization algorithm

4.1.1   Coordinate descent and threshold operators

We first develop the algorithmic framework for minimizing the selection criterion $Q(\boldsymbol{\beta})$. A key tool in this optimization algorithm is *coordinate descent*, which can be explained as follows. Viewing $Q(\boldsymbol{\beta})$ as a function of only the first coefficient β_1 (call this $Q_1(\beta_1)$), we first update β_1 as the minimizer of $Q_1(\cdot)$, keeping the remaining $p'-1$ coefficients fixed. The same procedure is then applied cyclically over $\beta_2, \cdots, \beta_{p'}$, and repeated until the full coefficient vector $\boldsymbol{\beta}$ converges. In recent years, coordinate descent has become widely used

13

in the variable selection literature (see, e.g., Fu, 1998; Friedman et al., 2007; Mazumder et al., 2011), due to its simplicity and efficiency for high-dimensional problems. The key to efficiency lies in the existence of a *closed-form* minimizer for the coordinate-wise objective $Q_j(\cdot)$, also known as a *threshold function* from signal processing (Donoho, 1995). We derive below such a threshold function for $Q(\boldsymbol{\beta})$.

Before delving into details, we first investigate the convexity properties of $Q(\boldsymbol{\beta})$:

Proposition 1 (Strict convexity)**.** $Q(\boldsymbol{\beta})$ *is strictly convex whenever* $\tau + 1/\gamma < \lambda_{min}(\mathbf{X}^T\mathbf{X})/(2n)$, *where* $\lambda_{min}(\cdot)$ *returns the minimum eigenvalue. Also, assuming each column* $\mathbf{x}_j$ *of* $\mathbf{X}$ *is normalized (i.e.,* $\mathbf{x}_j^T\mathbf{1} = 0$ *and* $n^{-1}\|\mathbf{x}_j\|_2^2 = 1$ *for any* $j = 1, \cdots, p'$*), it follows that* $Q_j(\beta_j)$ *is strictly convex for any* $j = 1, \cdots, p'$*, whenever* $\tau + 1/\gamma < 1/2$.

In words, this shows that a sufficiently small choice of $\tau + 1/\gamma$ is needed to ensure some form of convexity for the objective $Q(\boldsymbol{\beta})$. The first part of this proposition shows a unique global minimum exists for $Q(\boldsymbol{\beta})$ when $\tau + 1/\gamma < \lambda_{min}(\mathbf{X}^T\mathbf{X})/(2n)$. Such a result is quite restrictive, because it applies only to the low-dimensional setting of $n \leq p'$, where $\lambda_{min}(\mathbf{X}^T\mathbf{X})$ is strictly positive. The second part guarantees the coordinate-wise objective $Q_j(\beta_j)$ is strictly convex whenever $\tau + 1/\gamma < 1/2$, a result which holds in the high-dimensional setting of $n > p'$. This coordinate-wise convexity is important for deriving the threshold function below.

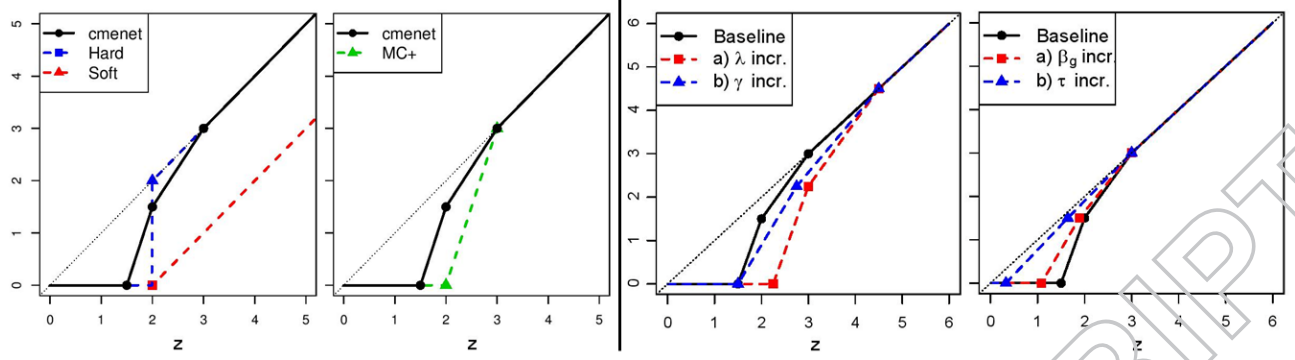For a main effect J, consider now its coordinate-wise minimization:

$$\min_{\beta_j} Q_j(\beta_j) = \min_{\beta_j} \left[\frac{1}{2n}\|\mathbf{r}_{-j} - \mathbf{x}_j\beta_j\|_2^2 + \eta_{\lambda_s,\tau}\left\{\|\boldsymbol{\beta}_{\mathcal{S}(j)}\|_{\lambda_s,\gamma}\right\} + \eta_{\lambda_c,\tau}\left\{\|\boldsymbol{\beta}_{\mathcal{C}(j)}\|_{\lambda_c,\gamma}\right\} \right], \quad (7)$$

where $\mathbf{r}_{-j} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_j\beta_j$ is the residual vector fitted without $\mathbf{x}_j$. Similarly, for a CME $J|K+$, its coordinate-wise minimization becomes:

$$\min_{\beta_{j|k+}} Q_{j|k+}(\beta_{j|k+}) = \min_{\beta_{j|k+}} \left[\frac{1}{2n}\|\mathbf{r}_{-(j|k+)} - \mathbf{x}_{j|k+}\beta_{j|k+}\|_2^2 + \eta_{\lambda_s,\tau}\left\{\|\boldsymbol{\beta}_{\mathcal{S}(j)}\|_{\lambda_s,\gamma}\right\} + \eta_{\lambda_c,\tau}\left\{\|\boldsymbol{\beta}_{\mathcal{C}(k)}\|_{\lambda_c,\gamma}\right\} \right].$$
$$(8)$$

An optimization technique called *majorization-minimization* (MM, see Chapter 12 of Lange,

14

Figure 2: *(1st and 2nd plots) A comparison of the baseline threshold function S_{λ_1,λ_2} (baseline setting: $(\lambda_1, \lambda_2, \gamma, \tau) = (1, 0.5, 3, 0.05)$ with no selected group effects) with soft-, hard- and MC+ thresholding. (3rd plot) A comparison of the baseline threshold function with two new settings $(1.5, 0.75, 3.0.05)$ and $(1, 0.5, 4.5, 0.05)$, all with no selected group effects. (Last) A comparison of the baseline threshold with two new settings $(1, 0.5, 3, 0.05)$ and $(1, 0.5, 3, 0.25)$, the latter with grouped norms $\|\beta_g\|_{\lambda_1,\gamma} = \|\beta_g\|_{\lambda_2,\gamma} = 5$.*

2010) can now be used to derive a threshold function. The main idea of MM is as follows. Instead of minimizing the original objective function, one first obtains a majorizing surrogate function which lies above the desired objective. This surrogate is then minimized in place of the original objective. Under certain conditions, the solution iterates generated by repeating this procedure converge to a minimizer for the original problem (Lange, 2010). For Q_j and $Q_{j|k+}$, a simple first-order expansion yields a nice majorizing surrogate function which can be minimized in closed form, as the following theorem demonstrates:

Theorem 3 (Threshold function). *Suppose $\tau + 1/\gamma < 1/2$. For fixed $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{p'}$, define $\bar{Q}_j(\cdot|\tilde{\boldsymbol{\beta}})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\boldsymbol{\beta}})$ as:*

$$\bar{Q}_j(\beta_j|\tilde{\boldsymbol{\beta}}) = \frac{1}{2n}\|\mathbf{r}_{-j} - \mathbf{x}_j\beta_j\|_2^2 + \eta_{\lambda_s,\tau}\left\{\|\tilde{\boldsymbol{\beta}}_{\mathcal{S}(j)}\|_{\lambda_s,\gamma}\right\} + \eta_{\lambda_c,\tau}\left\{\|\tilde{\boldsymbol{\beta}}_{\mathcal{C}(j)}\|_{\lambda_c,\gamma}\right\}$$
$$+ \tilde{\Delta}_{\mathcal{S}(j)}\left\{g_{\lambda_s,\gamma}(\beta_j) - g_{\lambda_s,\gamma}(\tilde{\beta}_j)\right\} + \tilde{\Delta}_{\mathcal{C}(j)}\left\{g_{\lambda_c,\gamma}(\beta_j) - g_{\lambda_c,\gamma}(\tilde{\beta}_j)\right\}, \ and$$

$$\bar{Q}_{j|k+}(\beta_{j|k+}|\tilde{\boldsymbol{\beta}}) = \frac{1}{2n}\|\mathbf{r}_{-(j|k+)} - \mathbf{x}_{j|k+}\beta_{j|k+}\|_2^2 + \eta_{\lambda_s,\tau}\left\{\|\tilde{\boldsymbol{\beta}}_{\mathcal{S}(j)}\|_{\lambda_s,\gamma}\right\} + \eta_{\lambda_c,\tau}\left\{\|\tilde{\boldsymbol{\beta}}_{\mathcal{C}(k)}\|_{\lambda_c,\gamma}\right\}$$
$$+ \tilde{\Delta}_{\mathcal{S}(j)}\left\{g_{\lambda_s,\gamma}(\beta_{j|k+}) - g_{\lambda_s,\gamma}(\tilde{\beta}_{j|k+})\right\} + \tilde{\Delta}_{\mathcal{C}(k)}\left\{g_{\lambda_c,\gamma}(\beta_{j|k+}) - g_{\lambda_c,\gamma}(\tilde{\beta}_{j|k+})\right\},$$

15

where $\tilde{\cdot}$ indicates the quantity is computed with $\tilde{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$. Then:

a) $\bar{Q}_j(\cdot|\tilde{\boldsymbol{\beta}})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\boldsymbol{\beta}})$ are majorization functions for $Q_j(\cdot)$ and $Q_{j|k+}(\cdot)$, respectively,

b) The unique minimizers of $\bar{Q}_j(\cdot|\tilde{\boldsymbol{\beta}})$ and $\bar{Q}_{j|k+}(\cdot|\tilde{\boldsymbol{\beta}})$ are given by $S_{\lambda_s,\lambda_c}(\mathbf{x}_j^T\mathbf{r}_{-j}/n; \tilde{\Delta}_{\mathcal{S}(j)}, \tilde{\Delta}_{\mathcal{C}(j)})$
   and $S_{\lambda_s,\lambda_c}(\mathbf{x}_{j|k+}^T\mathbf{r}_{-j|k+}/n; \tilde{\Delta}_{\mathcal{S}(j)}, \tilde{\Delta}_{\mathcal{C}(k)})$, respectively. Here, $S_{\lambda_1,\lambda_2}(\cdot; \Delta_1, \Delta_2)$ is the thresh-
   old function:

$$
S_{\lambda_1,\lambda_2}(z; \Delta_1, \Delta_2) = \begin{cases}
z & \text{if } z \in [\lambda_{(1)}\gamma, \infty), \\[2mm]
\mathrm{sgn}(z)\left(|z| - \Delta_{(1)}\right) / \left(1 - \dfrac{\Delta_{(1)}}{\lambda_{(1)}\gamma}\right) \\[4mm]
\qquad \text{if } z \in \left[\lambda_{(2)}\gamma + \Delta_{(1)}\left(1 - \dfrac{\lambda_{(2)}}{\lambda_{(1)}}\right), \lambda_{(1)}\gamma\right), \\[4mm]
\mathrm{sgn}(z)\left(|z| - \Delta_{(1)} - \Delta_{(2)}\right) / \left(1 - \dfrac{\Delta_{(1)}}{\lambda_{(1)}\gamma} - \dfrac{\Delta_{(2)}}{\lambda_{(2)}\gamma}\right) \\[4mm]
\qquad \text{if } z \in \left[\Delta_{(1)} + \Delta_{(2)}, \lambda_{(2)}\gamma + \Delta_{(1)}\left(1 - \dfrac{\lambda_{(2)}}{\lambda_{(1)}}\right)\right), \\[4mm]
0, & \text{otherwise.}
\end{cases}
\tag{9}
$$

where $\lambda_{(1)} = \max(\lambda_1, \lambda_2)$ and $\lambda_{(2)} = \min(\lambda_1, \lambda_2)$, with $\Delta_{(1)}$ and $\Delta_{(2)}$ its corresponding slopes.

To better understand the shrinkage behavior of this new threshold function, the left two plots in Figure 2 show a baseline setting of the `cmenet` threshold $S_{\lambda_1,\lambda_2}(z; \Delta_1, \Delta_2)$, compared with the soft-threshold function (corresponding to the shrinkage operator in LASSO), the hard-threshold function (corresponding to best-subset selection; see Friedman et al., 2001), and the MC+ threshold function (Mazumder et al., 2011). The baseline setting for the proposed threshold $S_{\lambda_1,\lambda_2}(z; \Delta_1, \Delta_2)$ is set as $(\lambda_1, \lambda_2, \gamma, \tau) = (1, 0.5, 3, 0.05)$, with $\|\boldsymbol{\beta}_g\|_{\lambda_1,\gamma} = \|\boldsymbol{\beta}_g\|_{\lambda_2,\gamma} = 0$ (i.e., no selected grouped effects). We see that the proposed threshold function is continuous and piecewise linear in four segments. Beginning from the left, the first segment is a horizontal line at zero, and represents the inner-product values for which a coefficient is shrunk to zero after thresholding. The last segment, which

matches the identity line, represents the values for which the full coefficient signal is retained without any shrinkage. The middle two segments provide a two-step transition between these two extremes, with slopes controlled by the sibling and cousin penalties. Similar to the MC+ threshold, the `cmenet` threshold bridges the gap between the two extremes of full shrinkage and no shrinkage; however, the former threshold accomplishes this transition in one step, while the latter achieves this in two steps. This two-step transition for `cmenet` is a consequence of the two-tiered coupling effect from sibling and cousin groups.

Consider next the right two plots of Figure 2, which investigate the sensitivity of the proposed threshold $S_{\lambda_1,\lambda_2}(z; \Delta_1, \Delta_2)$ to changes in penalty parameters. From the first plot, an increase in λ_1, λ_2 or γ appears to yield greater shrinkage of the coefficient signal. This is expected, because a larger choices of λ_1 and λ_2 induce greater regularization, and a larger γ generates a "more convex" penalty (see Mazumder et al., 2011). From the second plot, an increase in the coupling parameter τ in the presence of selected group effects appears to greatly reduce signal shrinkage. This observation nicely demonstrates the earlier CME coupling principle in Section 3.2, where the selection of sibling or cousin effects increases the chances of a CME entering the model.

4.1.2 Algorithm statement

Putting all the pieces together, Algorithm 1 summarizes the detailed steps for `cmenet`, which minimizes the selection criterion $Q(\boldsymbol{\beta})$ given fixed parameters λ_s, λ_c, γ and τ. Starting with an initial solution of $\boldsymbol{\beta} = \mathbf{0}_{p'}$, the threshold function in (9) is applied cyclically over each element in $\boldsymbol{\beta}$. This iterative procedure is then repeated until $\boldsymbol{\beta}$ converges. Using the majorization function in Theorem 3, one can prove the convergence of `cmenet` to a stationary solution.

Corollary 1 (Convergence of `cmenet`). *When $\tau + 1/\gamma < 1/2$, `cmenet` converges to a stationary solution $\hat{\boldsymbol{\beta}}$ satisfying $\nabla Q(\hat{\boldsymbol{\beta}}) = 0$.*

17

Algorithm 1 `cmenet`: An algorithm for bi-level CME selection

1: **function** CMENET$(\mathbf{X}, \mathbf{y}, \lambda_s, \lambda_c, \gamma, \tau, \boldsymbol{\beta} = \mathbf{0}_{p'})$ ▷ Assume columns of $\mathbf{X}$ are normalized

2:      • Initialize $\mathbf{r} \leftarrow \mathbf{y} - \bar{\mathbf{y}}$, $\Delta_{\mathcal{S}(j)} = \lambda_s$, $\Delta_{\mathcal{C}(j)} = \lambda_c$ for $j = 1, \cdots, p$

3:      **repeat**

4:          **for** $j = 1, \cdots, p$ **do**   ▷ For all main effects...

5:             • $\beta_0 \leftarrow \beta_j$, $\beta_j \leftarrow S_{\lambda_s, \lambda_c}\{\mathbf{x}_j^T \mathbf{r}/n + \beta_0; \Delta_{\mathcal{S}(j)}, \Delta_{\mathcal{C}(j)}\}$ ▷ Shrinkage

6:             • $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{x}_j(\beta_0 - \beta_j)$ ▷ Update residual

7:             • $\Delta_{\mathcal{S}(j)} \leftarrow \Delta_{\mathcal{S}(j)} \exp\{-\tau/\lambda_s [g_{\lambda_s, \gamma}(\beta_j) - g_{\lambda_s, \gamma}(\beta_0)]\}$ ▷ Update slopes

8:             • $\Delta_{\mathcal{C}(j)} \leftarrow \Delta_{\mathcal{C}(j)} \exp\{-\tau/\lambda_c [g_{\lambda_c, \gamma}(\beta_j) - g_{\lambda_c, \gamma}(\beta_0)]\}$

9:          **for** $j = 1, \cdots, p$ and $k = 1, \cdots, p$ **do** ▷ For all CMEs (both $J|K+$ and $J|K-$) ...

10:            • $\beta_0 \leftarrow \beta_{j|k+}$, $\beta_{j|k+} \leftarrow S_{\lambda_s, \lambda_c}\{\mathbf{x}_{j|k+}^T \mathbf{r}/n + \beta_0; \Delta_{\mathcal{S}(j)}, \Delta_{\mathcal{C}(k)}\}$ ▷ Shrinkage

11:            • $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{x}_{j|k+}(\beta_0 - \beta_{j|k+})$ ▷ Update residual

12:            • $\Delta_{\mathcal{S}(j)} \leftarrow \Delta_{\mathcal{S}(j)} \exp\{-\tau/\lambda_s \left[g_{\lambda_s, \gamma}(\beta_{j|k+}) - g_{\lambda_s, \gamma}(\beta_0)\right]\}$ ▷ Update slopes

13:            • $\Delta_{\mathcal{C}(k)} \leftarrow \Delta_{\mathcal{C}(k)} \exp\{-\tau/\lambda_c \left[g_{\lambda_c, \gamma}(\beta_{j|k+}) - g_{\lambda_c, \gamma}(\beta_0)\right]\}$

14:      **until** $\boldsymbol{\beta}$ converges

15: **return** the converged coefficient vector $\boldsymbol{\beta}$

As for its running time, one can show that one coordinate descent cycle in `cmenet` over all p' ME and CME coefficients requires $\mathcal{O}(np')$ work, because each coordinate descent step requires $\mathcal{O}(n)$ work. The linear running time in both sample size n and total effects p' is crucial for the computational efficiency of `cmenet`, particularly when a large number of main effects $p \gg 1$ is considered.

We mention here several extensions for `cmenet`. First, while Algorithm 1 considers only the selection and estimation of CMEs, the proposed algorithm can easily be extended for the selection of both CMEs and *other* covariate factors (whether continuous or discrete). For example, if the l_1-penalty were imposed on these latter factors, one can simply modify the coordinate descent loop in Algorithm 1 by incorporating soft-threshold updates (Donoho, 1995) to the coefficients of such factors. The algorithmic convergence for this extension is analogous to Corollary 1, and is not included for brevity. Second, we note that `cmenet`, as stated in Algorithm 1, is suitable for selecting binary CMEs – CMEs which quantify the effect of a *binary* factor at fixed levels of another factor, but not continuous CMEs – CMEs

18

which quantify the effect of a *continuous* factor at fixed levels of another factor. One way to extend `cmenet` for the latter problem is to first (a) discretize the underlying continuous factor into two levels, then (b) perform `cmenet` on the resulting binary CMEs, and finally (c) quantify the continuous component of these continuous CMEs using the *residuals* from `cmenet` as a new response vector. However, this extension requires further developments, and given the length of the current paper, we defer such an extension to future work.

4.2 Parameter tuning, warm starts and active set optimization

While Algorithm 1 provides an efficient method for minimizing the selection criterion $Q(\boldsymbol{\beta})$ given *fixed* penalty parameters λ_s, λ_c, γ and τ, such parameters are typically not known in practice, and therefore require tuning. We present below a method for performing this tuning procedure, as well as two computational tools – warm starts and active set optimization – which greatly speed up this tuning in practice.

For parameter tuning, we adopt the relatively standard procedure (see, e.g., Friedman et al., 2001; Mazumder et al., 2011) of finding the optimal penalty setting whose corresponding model (fitted using `cmenet`) minimizes some estimate of prediction error. In our implementation, called `cv.cmenet`[2], this prediction error is estimated using a technique called K-fold cross validation (or K-fold CV; see Friedman et al., 2001), which randomly splits the observed data into K parts, and uses one part of the data to validate the model fitted with the remaining $K-1$ parts. After obtaining this optimal penalty setting, the corresponding fitted model is then used for variable selection and prediction. For brevity, the specific details for `cv.cmenet` are summarized in Section 2 of the Appendix.

One practical challenge for this tuning procedure is that there are four parameters $(\lambda_s, \lambda_c, \gamma, \tau)$ to tune for in `cv.cmenet`. Some guiding rules are therefore needed to efficiently explore this 4-d parameter space. The proposition below provides one such rule for (λ_s, λ_c):

---

[2]In later sections, the tuning procedure `cv.cmenet` is often referred to as simply `cmenet` for brevity.

Proposition 2 (Search rule for (λ_s, λ_c)). *Suppose* $\lambda_s + \lambda_c \geq \max\limits_{j=1,\cdots,p'} |\mathbf{x}_j^T \mathbf{y}|/n$. *When* $Q(\boldsymbol{\beta})$ *is strictly convex, the unique minimizer of* $Q(\boldsymbol{\beta})$ *is the zero solution* $\boldsymbol{\beta} = \mathbf{0}_{p'}$.

It should be noted that, in the high-dimensional setting of $n > p'$, $Q(\boldsymbol{\beta})$ cannot be strictly convex (see discussion for Proposition 1), so $\boldsymbol{\beta} = \mathbf{0}_{p'}$ is only a stationary solution. Nonetheless, the restriction of $\lambda_s + \lambda_c < \max\limits_{j=1,\cdots,p'} |\mathbf{x}_j^T \mathbf{y}|/n$ allows for considerable reduction in the search for interesting choices of λ_s and λ_c. From Proposition 1, another rule is $\tau + 1/\gamma < 1/2$, which ensures the strict convexity of the coodinate-wise problem and therefore the numerical stability of the optimization procedure. For brevity, the incorporation of these rules in `cv.cmenet` is outlined in Section 2 of the Appendix.

Two computational tools can be used to greatly speed up the tuning procedure `cv.cmenet` in high-dimensions. The first tool, called *warm starts*, makes use the converged solution from a previous parameter setting to initialize the optimization problem for the current setting. The use of warm starts in variable selection was popularized in Friedman et al. (2007) for efficiently fitting multiple models along the full LASSO path, and we found such a tool to be equally effective for efficiently fitting multiple models over a grid of penalty parameters for `cmenet`. The second tool, called *active set optimization* (see, e.g., Meier et al., 2008; Friedman et al., 2010), performs coordinate descent updates over a small subset of *active* variables, instead of over the full set of p' variables. This technique is most effective when there are only a small number of active effects present, because one can avoid performing redundant coordinate descent updates on coefficients of inactive effects. Section 2 of the Appendix provides specific details on how these two tools can be incorporated into `cv.cmenet`.

4.3 CME screening rules

When the number of main effects p grows large, performing even one full coordinate descent over all $p' = p + 4\binom{p}{2}$ total effects can be computationally cumbersome. One effective

way of reducing computation time in such a situation is the use of screening rules, or *strong rules*, which screen out a large number of inactive variables from consideration using previously-solved coefficient solutions. The term "strong rules" is first coined in Tibshirani et al. (2012), where the authors used previously-solved solutions along the LASSO path to screen out inactive effects for subsequent optimizations. We derive below similar strong rules for screening out inactive effects for `cmenet`, and reveal an interesting connection between these screening rules and CME coupling.

Suppose the parameters γ and τ are fixed, and let j index a variable of interest (ME or CME), with $\mathcal{S}$ and $\mathcal{C}$ its corresponding sibling and cousin group. Furthermore, let $\hat{\boldsymbol{\beta}}(\lambda_s, \lambda_c)$ be an optimal solution of the selection criterion $Q(\boldsymbol{\beta})$ under penalties λ_s and λ_c, and let $c_j(\lambda_s, \lambda_c) = \mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_s, \lambda_c))/n$ denote the inner-product of effect j with the current residual vector. Denoting $\lambda_s^1 > \lambda_s^2 > \cdots > \lambda_s^L$ and $\lambda_c^1 > \lambda_c^2 > \cdots > \lambda_c^M$ as the desired (decreasing) penalty sequences for λ_s and λ_c, the screening procedure can be summarized by the following three strong rules:

1. Suppose there are no active effects in $\mathcal{S}$ and $\mathcal{C}$ for penalty settings $(\lambda_s^{l-1}, \lambda_c^m)$ or $(\lambda_s^l, \lambda_c^{m-1})$. Then effect j is marked as *inactive* for penalty setting $(\lambda_s^l, \lambda_c^m)$ if:

$$|c_j(\lambda_s^{l-1}, \lambda_c^m)| < \lambda_s^l + \lambda_c^m + \frac{\gamma}{\gamma - 2}(\lambda_s^l - \lambda_s^{l-1}) \quad \text{or} \quad |c_j(\lambda_s^l, \lambda_c^{m-1})| < \lambda_s^l + \lambda_c^m + \frac{\gamma}{\gamma - 2}(\lambda_c^m - \lambda_c^{m-1}). \tag{10}$$

2. If there are no active effects in the *sibling* group $\mathcal{S}$ for penalty setting $(\lambda_s^{l-1}, \lambda_c^m)$, then effect j is marked as *inactive* for penalty setting $(\lambda_s^l, \lambda_c^m)$ if:

$$|c_j(\lambda_s^{l-1}, \lambda_c^m)| < \lambda_s^l + \Delta_{\mathcal{C}}' + \frac{\gamma}{\gamma - (\Delta_{\mathcal{C}}'/\lambda_c^m + 1)}(\lambda_s^l - \lambda_s^{l-1}), \tag{11}$$

where $\Delta_{\mathcal{C}}' = \lambda_c^m \exp\left\{-\tau\|\boldsymbol{\beta}_{\mathcal{C}}(\lambda_s^{l-1}, \lambda_c^m)\|_{\lambda_c^m, \gamma}/\lambda_c^m\right\}$.

3. If there are no active effects in the *cousin* group $\mathcal{C}$ for penalty setting $(\lambda_s^l, \lambda_c^{m-1})$, then

21

effect j is marked as *inactive* for penalty setting $(\lambda_s^l, \lambda_c^m)$ if:

$$|c_j(\lambda_s^l, \lambda_c^{m-1})| < \Delta_{\mathcal{S}}' + \lambda_c^m + \frac{\gamma}{\gamma - (\Delta_{\mathcal{S}}'/\lambda_s^l + 1)}(\lambda_c^m - \lambda_c^{m-1}), \qquad (12)$$

where $\Delta_{\mathcal{S}}' = \lambda_s^l \exp\left\{-\tau\|\boldsymbol{\beta}_{\mathcal{S}}(\lambda_s^l, \lambda_c^{m-1})\|_{\lambda_s^l, \gamma}/\lambda_s^l\right\}$.

A theoretical derivation of these rules is provided in Section 3 of the Appendix.

While these three rules may appear complicated and technical, they are in fact quite intuitive to understand. All three rules consider conditions under which it would be "safe" to screen out effect j from the optimization problem for the penalty setting $(\lambda_s^l, \lambda_c^m)$. The first rule applies when there are no active effects in $\mathcal{S}$ and $\mathcal{C}$ from previous penalty settings, and screens out effect j if the previous inner-products $c_j(\lambda_s^{l-1}, \lambda_c^m)$ or $c_j(\lambda_s^l, \lambda_c^{m-1})$ are within the upper bounds provided in (10). The intuition here is that if effect j is not correlated enough with the residual vectors at the previous penalty settings $(\lambda_s^{l-1}, \lambda_c^m)$ or $(\lambda_s^l, \lambda_c^{m-1})$, then it cannot "catch up" in time to be active for the current setting $(\lambda_s^l, \lambda_c^m)$ (see Tibshirani et al., 2012 for details). This first rule can be viewed as an extension of the MC+ strong rule in Lee and Breheny (2015) to the current model. The second rule applies when there are no active effects in the sibling group $\mathcal{S}$ (but some in cousin group $\mathcal{C}$) for the previous setting $(\lambda_s^{l-1}, \lambda_c^m)$. In such a scenario, effect j is screened out if the previous inner-product $c_j(\lambda_s^{l-1}, \lambda_c^m)$ is within the upper bound in (11). The key difference between this and the first rule is that, as more effects are selected in the cousin group $\mathcal{C}$, the linearized slope $\Delta_{\mathcal{C}}'$ decays smaller than λ_c^m, which then decreases the screening bound in (11) compared to the original bound in (10)[3]. In other words, the presence of coupled cousin effects from a previous setting can *decrease* the screening power of strong rules for the current setting. This is quite similar to the CME coupling phenomenon in Section 3.2, except instead of encouraging the *selection* of effect j, the coupled CMEs make it more diffcult to *screen out* effect j via strong rules. The third rule, which applies when there are no previously-active

---

[3]Here, we assume the last term in both (10) and (11) are nearly equal in this comparison; the discrepancy between (10) and (11) is dominated by the first two terms for most feasible parameter settings.

22

cousins in $\mathcal{C}$ (but some siblings in $\mathcal{S}$), enjoys a similar interpretation: as more siblings are coupled in from $\mathcal{S}$ at a previous setting, effect j becomes more difficult to screen out via strong rules.

Lastly, we note that while these three rules do screen out a large proportion of inert CMEs, it is possible (but highly unlikely) that an active CME is erroneously screened out. This is illustrated numerically in the following section. To prevent any false-negative screenings, we recommend that the KKT conditions (see equation (3) of the Appendix) be checked as a final step for each optimization problem.

5 Simulations

We now explore the performance of the proposed method in several simulation studies. Table 2 summarizes the test settings for these simulations, with varying sample sizes n and main effects p, varying number of active groups x and active effects within a group y (denoted as GxAy), and whether the grouped effects are siblings or cousins (main effect models are considered here as well). Active effects are assigned a value of 1 in the coefficient vector $\boldsymbol{\beta}$, and non-active effects assigned a value of 0. Each simulation case is then replicated 100 times, with the model matrix $\mathbf{X}$ simulated from the equicorrelated latent model in Section 2.2 with $\rho = 0$ and $\rho = 1/\sqrt{2}$, and the response $\mathbf{y}$ simulated independently from $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n)$. For brevity, we only report the results for $(n, p) = (50, 50), (100, 100)$ and $(150, 150)$ with G4A2 and G6A2, but similar conclusions hold for other settings.

Under such a set-up, our simulations aim to answer two questions: (a) Does the proposed method `cmenet` yield improved selection of CMEs compared to more generic selection methods? (b) For an active CME, say $J|K+$, is `cmenet` more effective at identifying this conditional, non-additive relation between J and K, compared to the more traditional 2FI analysis? To answer the first question, we compare `cmenet` with two generic variable selection techniques from the literature: the LASSO (Tibshirani, 1996) using the R package

23

Simulation parameters	Settings
Sample size	$n = 50,\ 100$ or 150
# of main effects considered (total effects considered)	$p = 50,\ 100$ or 150 $\left(p' = p + 4\binom{p}{2} = 4,950,\ 19,900 \text{ or } 44,850\right)$
# of active groups	6 or 8
# of active effects within a group	2 or 3
Effect type	Siblings, cousins, main effects[1]
Latent correlation	$\rho = 0$ or $1/\sqrt{2}$

Table 2: *Test settings for simulation study.*

[1] # of active MEs is set as # of active groups.

GLMNET (Friedman et al., 2009), and SparseNet (Mazumder et al., 2011) using the R package SPARSENET (Mazumder et al., 2012). All three methods perform selection on the *same* set of MEs and CMEs, with penalty parameters tuned using 10-fold CV. In this comparison, a better selection performance for cmenet shows that the proposed penalty $Q(\boldsymbol{\beta})$ is more appropriate for selecting CMEs compared to generic penalties. To answer the second question, we compare cmenet with a popular selection method called hierNet (Bien et al., 2013) for selecting 2FIs. A better selection performance for cmenet over hierNet thereby demonstrates the effectiveness of the proposed method in identifying the conditional, non-additive nature of CMEs.

We employ two criteria to conduct the above comparisons. The first criterion returns the number of misspecified variables: $\#\{\mathcal{A} \setminus \hat{\mathcal{A}}_n\} + \#\{\hat{\mathcal{A}}_n \setminus \mathcal{A}\}$, where $\mathcal{A}$ is the true active set of MEs and CMEs, and $\hat{\mathcal{A}}_n$ is the set of selected effects after n observations. Smaller values of this indicate better selection performance. Such a criterion is appropriate for cmenet, LASSO and SparseNet, which perform selection on the MEs and CMEs in $\mathcal{A}$, but a slight modification is needed for hierNet, which performs selection on the traditional 2FIs. To this end, let $\mathcal{A}^{(ME)}$ consist of the original MEs in active set $\mathcal{A}$ as well as the parent MEs of the CMEs in $\mathcal{A}$, and let $\mathcal{A}^{(2FI)}$ consist of the 2FIs corresponding

24

to the CMEs in $\mathcal{A}$. The misspecification criterion for `hierNet` can then be written as: $\#\{\mathcal{A}^{(ME)} \setminus \hat{\mathcal{A}}_n^{(ME)}\} + \#\{\hat{\mathcal{A}}_n^{(ME)} \setminus \mathcal{A}^{(ME)}\} + \#\{\mathcal{A}^{(2FI)} \setminus \hat{\mathcal{A}}_n^{(2FI)}\} + \#\{\hat{\mathcal{A}}_n^{(2FI)} \setminus \mathcal{A}^{(2FI)}\}$, where $\hat{\mathcal{A}}_n^{(ME)}$ and $\hat{\mathcal{A}}_n^{(2FI)}$ are the selected MEs and 2FIs from `hierNet`. Put another way, this modified criterion first translates the true CME model into its component MEs and 2FIs (see the identities in (1)), then reports the number of misspecifications for the fitted `hierNet` model based on these component effects. The second criterion is the mean-squared prediction error (MSPE): $\mathbb{E}\|\mathbf{y}_{new} - \mathbf{X}_{new}\hat{\boldsymbol{\beta}}\|_2^2$, where $(\mathbf{X}_{new}, \mathbf{y}_{new})$ is an out-of-sample dataset with $n_{new} = 20$ observations simulated from the true model $\mathcal{A}$. Smaller MSPE values suggest better predictive performance. Here, the focus is on a method which yields the best *selection* performance of CMEs (first criterion); however, such a method should have comparable *predictive* performance to other methods (second criterion).

Figures 3 show the number of misspecifications and MSPE for the four methods with $\rho = 0$ and $\rho = 1/\sqrt{2}$, under the simulation settings presented earlier. Consider first the sibling and cousin models in the $\rho = 0$ setting (left part of Figure 3), where the underlying MEs are uncorrelated. For these models, `cmenet` provides noticeably improved selection performance over LASSO and `SparseNet` for nearly all simulation settings. This shows that the penalization scheme in $Q(\boldsymbol{\beta})$ is indeed more effective than generic penalties for selecting active CMEs; by accounting for the implicit group structure of CMEs, the proposed method can better guide the variable selection procedure using the novel principle of CME coupling. `cmenet` also yields a sizable selection improvement over `hierNet` for sibling and cousin models, which shows that the proposed approach can better identify the conditional, non-linear nature of CMEs compared to traditional 2FI analysis. One likely explanation is that, because a CME can be decomposed into its component ME and 2FI effects (recall the identities in (1) and Rule 1 of Su and Wu, 2017), the selection signal of an active CME is much stronger than the signals from its component ME and 2FI effects. `cmenet`, by performing selection directly on the CMEs with greater signal, can more easily identify the underlying active effects compared to `hierNet`, which performs selection on its component

25
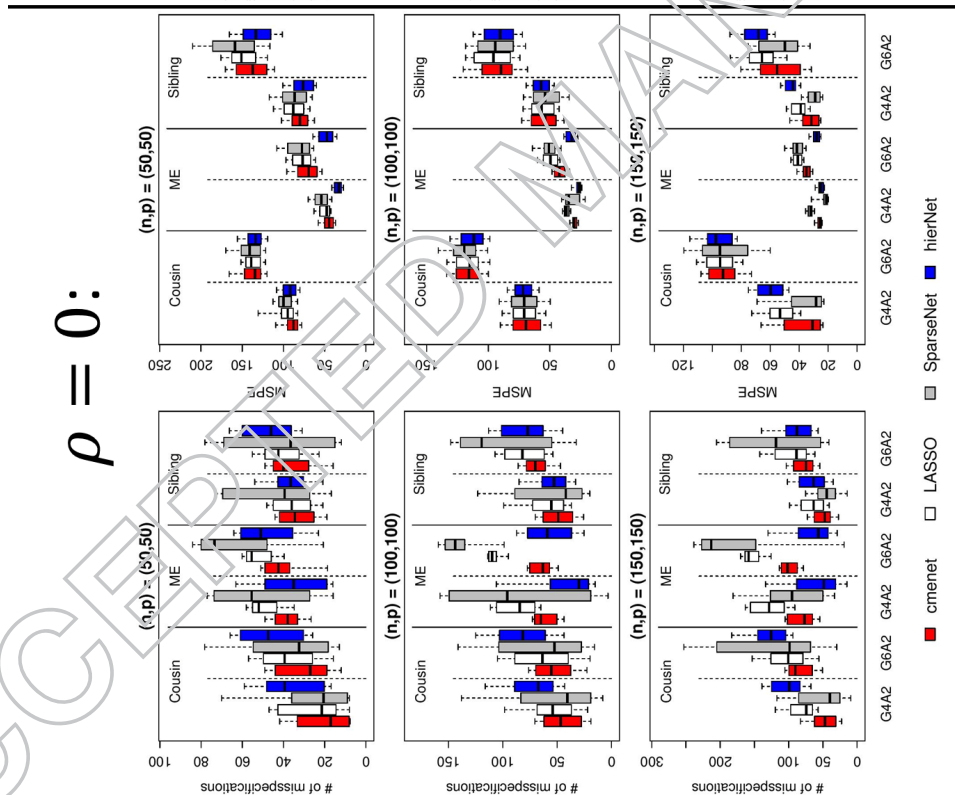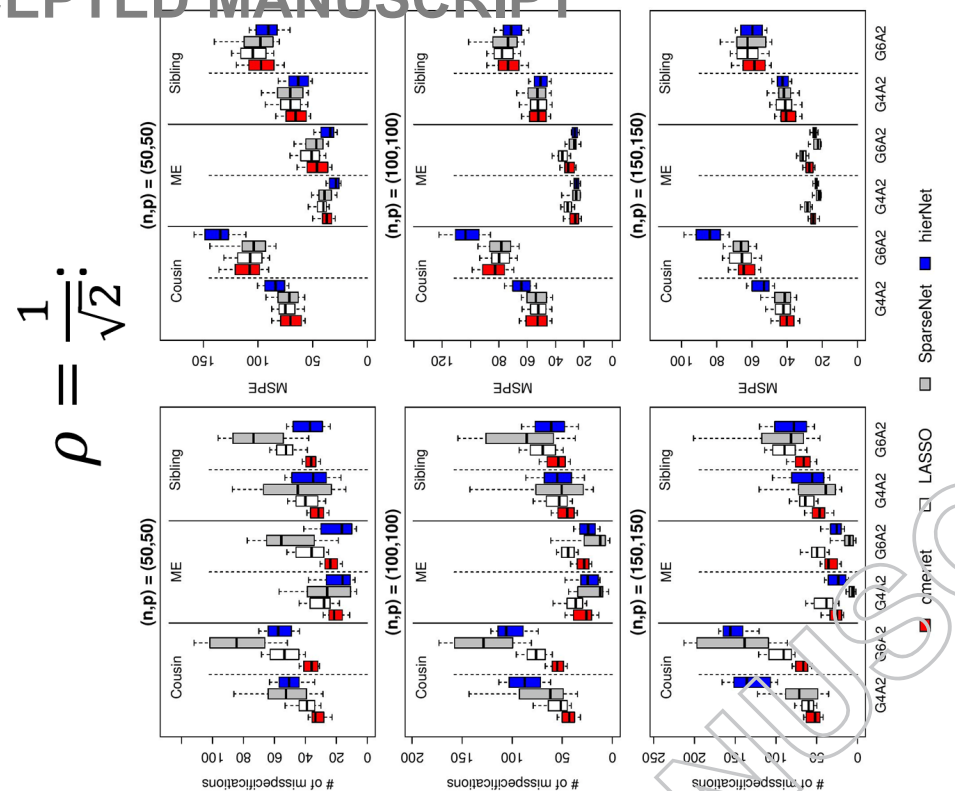
Figure 3: *Boxplots of the 10%, 25%, 50%, 75% and 90% quantiles for the # of misspecifications and MSPE, with $(n, p) = (50, 50)$ (top), $(n, p) = (100, 100)$ (middle) and $(n, p) = (150, 150)$ (bottom), using a latent correlation of $\rho = 0$ (left) and $\rho = 1/\sqrt{2}$ (right).*

26

ME and 2FI effects with diluted signals. As for MSPE, `cmenet` enjoys comparable or improved performance to the other three methods, which is as desired.

Consider next the main effect models for $\rho = 0$ (left part of Figure 3). We see that `cmenet` enjoys superior selection performance to LASSO and `SparseNet`, which demonstrates the effectiveness of the CME reduction principle in reducing selected CMEs into its underlying parent ME. Compared to `hierNet`, `cmenet` provides comparable (but slightly worse) selection for these main effect models, an observation not too surprising given that the proposed method specifically tackles the problem of CME selection. `cmenet` is therefore most effective in applications where one expects some conditional effects to be active in the model; in other words, in applications where CMEs represent interpretable, domain-specific phenomena.

Finally, consider the results for $\rho = 1/\sqrt{2}$ (right part of Figure 3), where the underlying MEs are moderately correlated. For the sibling and cousin models, `cmenet` again provides an improvement in selection performance over the other three methods, with this improvement much greater than that for the uncorrelated setting $\rho = 0$. Such an observation is expected in light of Section 2.2, because the CME group structure is most prominent for moderate choices of ρ. For the main effect models, `cmenet` and `hierNet` again provide the best selection performance, with the relative performance of `cmenet` noticeably better than that for $\rho = 0$. This again can be explained by the more pronounced CME group structure for moderate ρ, which allows for more effective CME reduction. As before, the MSPE for `cmenet` is comparable to or better than the other three methods, which is as desired.

To numerically demonstrate the effectiveness of the CME screening rules in Section 4.3, the left plot in Figure 4 shows the boxplots of the computation times for `cmenet` with $(n, p) = (200, 150)$ and $(500, 200)$, under a G2A6 sibling model with latent correlation $\rho = 0$. We see that the proposed screening rules significantly reduce computation time, with over 20% reduction in median time for $(n, p) = (200, 150)$, and 30% reduction for $(n, p) = (500, 200)$. This effectiveness appears to grow for larger sample sizes n and greater number of main effects p, which is as desired. The right plot in Figure 4 shows the proportion
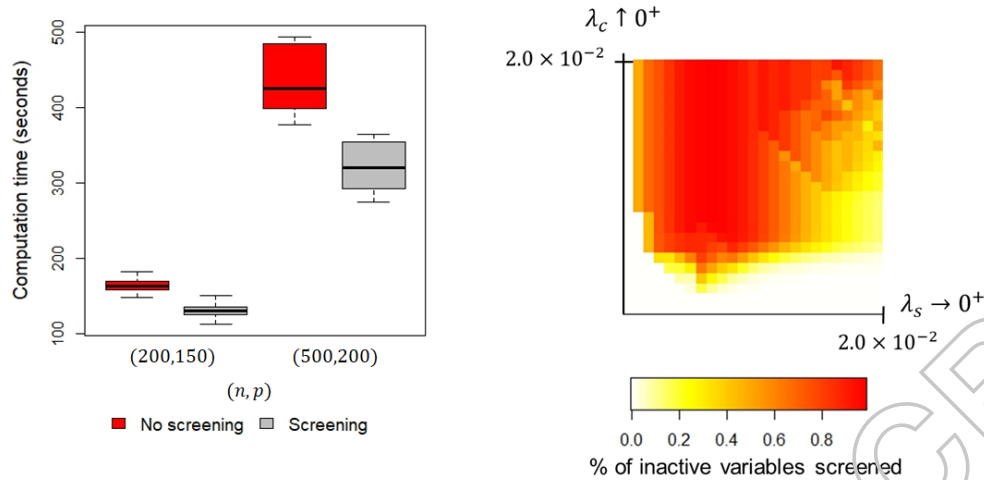
Figure 4: *(Left) Boxplots of computation times for* `cmenet` *with* $(n, p) = (200, 150)$ *and* $(500, 200)$; *(Right) Proportion of inactive variables screened for* $(n, p) = (200, 150)$.

of inactive variables removed by the screening procedure for $(n, p) = (200, 150)$ as a function of the sibling and cousin penalties λ_s and λ_c. We see that the proposed screening rules correctly remove a large proportion of inactive variables (over 80% for smaller λ_s and λ_c), which greatly speeds up the ensuing coordinate descent algorithm. In total, only 3 active variables were incorrectly screened over all values of (λ_s, λ_c) tested, and all such violations were corrected in post-convergence check of KKT conditions.

6 Polygenic association study on fly wing shape

In this section, we demonstrate the usefulness of `cmenet` for an important, real-world problem on polygenic association. Polygenes are a group of non-epistatic genes which serve as biological markers for many characteristics of interest called phenotypes (e.g., susceptibility to diabetes for youth (Rosenbloom et al., 1999) and major depressive disorders (De Moor et al., 2015)), and the association of influential polygenes to particular phenotypes is an important area of research in the biomedical community. Here, we investigate the polygenic association for the wing shape of *Drosophila Melanogaster*, the common fruit fly.

28

The data employed here is collected from a study by Weber et al. (2001), where the authors considered $p = 48$ homozygous (i.e., binary[4]) polygene markers on the second chromosome of *Drosophila Melanogaster* and its effect on fly wing shape, using $n = 701$ observations collected from recombinant isogenic lines. The response of interest is a continuous index for wing shape, which incorporates both the width of the wing across the middle and the width across the base. As in simulation studies, our focus lies primarily on the *selection* of important CMEs, which here represents the effect of a gene conditional on another gene being active or absent. This is because the identification of these novel conditional effects yields valuable insight into the activation structure of gene-gene interactions, whereas the more traditional two-factor interaction analysis can be less interpretable in such a setting.

Here, we compare the analysis provided by `cmenet` with that from `hierNet`. As before, `cmenet` performs selection on MEs and CMEs ($p' = p + 4\binom{p}{2} = 4,560$ variables in total), while `hierNet` performs MEs and 2FIs ($p'' = p + \binom{p}{2} = 1,176$ variables in total). The purpose of such a comparison is to understand the practical advantages and disadvantages in employing the novel CMEs as basis functions, compared to the typical approach of using 2FIs for analyzing gene-gene interactions (Cordell, 2009). For brevity, we do not include either the LASSO or `SparseNet` selection of CMEs in this comparison, because it was already shown in Section 5 that `cmenet` enjoys better selection performance.

Consider first Table 3, which shows (a) the number of selected effects for `cmenet` and `hierNet`, and (b) some selected effects for each method, along with their corresponding p-values from a regular linear model fit. We see that the fitted model from `cmenet`, which has 21 selected effects, is much smaller than the model returned by `hierNet`, which has

---

[4]For organisms with diploid cells (including *Drosophila Melanogaster*), chromosomes are found in pairs; these chromosome pairs can be further categorized as either *heterozygous* – meaning the pair contains different alleles for each gene, or *homozygous* – meaning the pair contains identical alleles for each gene. For alleles with levels + and –, heterozygous pairs allow for four allele combinations (+,+), (+,–), (–,+) and (–,–), while homozygous pairs allow for two combinations (+,+) and (–,–). For this fly wing study, Weber et al. (2001) found very little heterozygous behavior on chromosome 2, and reported subsequent results using modified homozygous chromosomes, which are binary and fit within the framework of this paper.

Method	# of selected effects	Some selected effects (p-values)	
cmenet	21	g14|g27- (6.1×10^{-4}), g14|g38+ (2.0×10^{-2}), g17|g14- (1.6×10^{-12}), g23|g14+ (2.5×10^{-30})	g45|g10+ (7.3×10^{-7})
hierNet	129	g14 (8.3×10^{-1})	g45 (1.5×10^{-1}), g45 $*$ g10 (8.1×10^{-1})

Table 3: *Number of selected effects and some selected effects (p-values bracketed) from* cmenet *and* hierNet *in the gene association study of fly wing shape.*
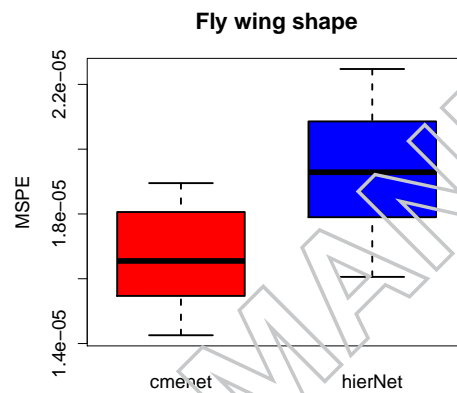
Figure 5: *Boxplots of the 10%, 25%, 50%, 75% and 90% MSPE quantiles for* cmenet *and* hierNet *in the gene association study of fly wing shape.*

129 selected effects. This model parsimony for cmenet suggests that there are indeed active CMEs for the problem at hand, i.e., there are certain polygenes which affect wing shape *only* in the presence or absence of other polygenes. Taking a closer look at some of the selected effects for cmenet and hierNet from Table 3, two interesting insights can be observed on this conditional gene association structure. From the first column of selected effects, hierNet deemed the 14-th polygene g14 to be active, while cmenet instead selected the two sibling effects g14|g27- and g14|g38+, and the two cousin effects g17|g14- and g23|g14+. In other words, under traditional analysis, gene g14 is deemed influential in all situations, whereas

30

the conclusion is more nuanced under the proposed CME analysis, with g14 influential (a) when gene g27 is absent or gene g38 is active, or (b) in inhibiting gene g17 or activating gene g23. The latter provides a more careful analysis of the signal from g14, and judging by the much smaller p-values for these conditional effects, also yields greater insight on the underlying gene activation structure. From the second column of selected effects, `hierNet` deemed both g45 and its interaction g45∗g10 to be active, while `cmenet` selected only the CME g45|g10+. This nicely illustrates why `cmenet` provides parsimonious models: by selecting the CME g45|g10+ in place of its component ME g45 and 2FI g45∗g10, we obtain a smaller model with considerably smaller p-values, which is as desired (this is akin to Rule 1 of Su and Wu, 2017 for selecting CMEs in designed experiments; see Section 2.1, especially equation (1)).

Consider next Figure 5, which shows the MSPE boxplots for `cmenet` and `hierNet` in predicting the continuous wing shape index. Here, MSPE is estimated by randomly sampling 80% of the data for model training, then using the remaining 20% to test the trained model; this procedure is then repeated 200 times to provide error variability. We see that `cmenet` enjoys considerable improvements over `hierNet` in terms of MSPE, yielding at least a 12% reduction at all five error quantiles. This again reaffirms the likely conditional nature of the underlying polygenic association structure, with certain polygenes active only in the presence or absence of other polygenes.

To summarize, this gene association study highlights two important advantages of `cmenet`. First, in applications where CMEs are interpretable phenomena, the proposed selection method can provide much more parsimonious models compared to traditional analysis using two-way interactions, and can yield greater insight on the underlying problem of interest. This is particularly true in genetic applications, where selected CMEs can be used to further investigate *why* some genes are conditionally active, and *why* some play a more supportive role in *activating* or *inhibiting* other genes. Second, when CMEs have natural domain-specific interpretations, using such effects as basis functions can also improve

31

the predictive performance of the fitted model as well.

7   Conclusion and future work

In this paper, a new method is presented for selecting binary variables and a set of reparametrized variables called conditional main effects (CMEs) from observation data. While CMEs are intuitive basis functions with appealing interpretations in many applications, existing selection methods can perform poorly due to the inherent grouped structure of these effects. We proposed a novel selection method called `cmenet`, which accounts for this underlying grouped structure using two selection principles called CME coupling and reduction; the former allows CMEs to more easily enter the model given selected siblings or cousins, and the latter encourages the selection of the underlying ME given many selected siblings or cousins. A coordinate descent algorithm is then introduced for minimizing the selection criterion, and several computational tools are proposed for efficient optimization and parameter tuning in high-dimensions. Simulation studies showed considerable improvements for `cmenet` over existing methods with respect to selection accuracy. Applied to a real-world gene association study on fly wing shape, the proposed method provides not only improved predictive performance over the standard two-way interaction analysis, but also a more parsimonious and interpretable model which reveals important insights on gene activation behavior.

Given the positive results here, there are many exciting avenues for future work. First, in the high-dimensional setting of $p \gg 1$, the tuning of the four selection parameters in $Q(\boldsymbol{\beta})$ can be computationally expensive due to the grid structure of feasible parameter combinations in `cv.cmenet`. With recent advances on the topic of optimal designs for convex spaces (e.g., Lekivetz and Jones, 2015; Mak and Joseph, 2017), it may be interesting to see whether the use of such designs as candidate settings allows for more efficient parameter tuning. Second, we are working to broaden the proposed methodology to higher-order

conditional effects, e.g., the effect of A conditional on both $B+$ and $C+$. The main challenge here is again computational efficiency, but such a direction would enable the investigation of, say, more complex activation phenomena in the earlier gene study. Lastly, we are interested in extending the current framework for selecting the continuous CMEs mentioned earlier in Section 4.1.2. This would allow the proposed methodology to be applicable for more general datasets, and we look forward to exploring this in future research.

Acknowledgements

References

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3):731–740.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369.

Chari, S. and Dworkin, I. (2013). The conditional nature of genetic interactions: the consequences of wild-type backgrounds on mutational interactions in a genome-wide modifier screen. *PLoS Genetics*, 9(8):e1003661.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404.

De Moor, M. H., Van Den Berg, S. M., Verweij, K. J., Krueger, R. F., Luciano, M., Vasquez,

A. A., Matteson, L. K., Derringer, J., Esko, T., and Amin, N. (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry*, 72(7):642–650.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Finney, D. (1945). The fractional replication of factorial arrangements. *Annals of Eugenics*, 12:291–303.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). GLMNET: Lasso and elastic-net regularized generalized linear models. *R package version 1*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer Science & Business Media.

Lee, S. and Breheny, P. (2015). Strong rules for nonconvex penalties and their implications for efficient algorithms in high-dimensional regression. *Journal of Computational and*

34

Graphical Statistics, 24(4):1074–1091.

Lekivetz, R. and Jones, B. (2015). Fast flexible space-filling designs for nonrectangular regions. *Quality and Reliability Engineering International*, 31(5):829–837.

Mak, S. and Joseph, V. R. (2017). Minimax and minimax projection designs using clustering. *Journal of Computational and Graphical Statistics*. To appear.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

Mazumder, R., Hastie, T., and Friedman, J. (2012). SPARSENET: Fit sparse linear regression models via nonconvex optimization. *R package version 1*.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71.

Montgomery, D. C. (2008). *Design and Analysis of Experiments*. John Wiley & Sons.

Rosenbloom, A. L., Joe, J. R., Young, R. S., and Winter, W. E. (1999). Emerging epidemic of type 2 diabetes in youth. *Diabetes Care*, 22(2):345–354.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Stuart, A. and Ord, J. (1994). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Arnold London.

Su, H. and Wu, C. F. J. (2017). CME analysis: a new method for unraveling aliased effects in two-level fractional factorial experiments. *Journal of Quality Technology*, 49(1):1–10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B*, 74(2):245–266.

Weber, K., Eisman, R., Higgins, S., Morey, L., Patty, A., Tausek, M., and Zeng, Z.-B. (2001). An analysis of polygenes affecting wing shape on chromosome 2 in Drosophila Melanogaster. *Genetics*, 159(3):1045–1057.

Wu, C. F. J. (2015). Post-Fisherian experimentation: from physical to virtual. *Journal of the American Statistical Association*, 110(510):612–620.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization* John Wiley & Sons.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.