# Improving prediction, variable selection and treatment effect estimation by the adaptive, multi-penalty elastic net

Mark van de Wiel, Joint work with Mirrelijn van Nee

Epidemiology & Data Science, Amsterdam University Medical Centers, Netherlands

Armitage Workshop, Cambridge, 2021

# Content

- Part I: Adaptive elastic net, prognostic setting
  - Response: $\boldsymbol{y} \approx g^{-1}(X\boldsymbol{\beta})$
  - Variables (features): $X = X_{n\times p}, p > n$; High-dimensional (HD) setting
  - $g$: link function
  - $\boldsymbol{\beta}$: estimated by adaptive elastic net
  - Penalty parameters adapt to prior information
  - Main challenge: estimate penalty parameters (equiv: prior parameters)

# Content

- Part I: Adaptive elastic net, prognostic setting
  - Response: $\boldsymbol{y} \approx g^{-1}(X\boldsymbol{\beta})$
  - Variables (features): $X = X_{n \times p}, p > n$; High-dimensional (HD) setting
  - $g$: link function
  - $\boldsymbol{\beta}$: estimated by adaptive elastic net
  - Penalty parameters adapt to prior information
  - Main challenge: estimate penalty parameters (equiv: prior parameters)
- Part II: Average treatment effect (ATE) estimation using adaptive EN
  - Estimator for ATE in HD settings
  - Link to Part I

# Part I: Adaptive elastic net, clinical prognostic setting

- Research goals:
  - Predict clinical outcomes, like therapy response or tumour relapse
  - Select variables for a parsimonious predictor
- Elastic net simultaneously selects and estimates effect of variables

# Part I: Adaptive elastic net, clinical prognostic setting

- Research goals:
  - Predict clinical outcomes, like therapy response or tumour relapse
  - Select variables for a parsimonious predictor
- Elastic net simultaneously selects and estimates effect of variables
- Co-data: known grouping(s) of variables
- Groups of variables may differ in predictive strength and dimension
- Group-adaptive elastic net learns from co-data to improve prediction and variable selection

Genome (DNA) — $p_G$ = **50** specific mutations

Transcriptomics (RNA) — $p_T$ = **20,000** genes

Methylomics — $p_M$ = **800,000** locations

Proteomics — $p_P$ = **2,000** targeted proteins

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
  - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
    - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables → may not be flexible enough compared to group-adaptive methods

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
  - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables $\rightarrow$ may not be flexible enough compared to group-adaptive methods
  - `ipflasso` (Boulesteix et al. 2017): use cross-validation on predefined set of possible penalties

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
    - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables → may not be flexible enough compared to group-adaptive methods
    - `ipflasso` (Boulesteix et al. 2017): use cross-validation on predefined set of possible penalties → computationally expensive

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:

  - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables $\rightarrow$ may not be flexible enough compared to group-adaptive methods
  - `ipflasso` (Boulesteix et al. 2017): use cross-validation on predefined set of possible penalties $\rightarrow$ computationally expensive
  - `gren` (Münch et al. 2019): empirical-variational Bayes approximation

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
  - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables $\rightarrow$ may not be flexible enough compared to group-adaptive methods
  - `ipflasso` (Boulesteix et al. 2017): use cross-validation on predefined set of possible penalties $\rightarrow$ computationally expensive
  - `gren` (Münch et al. 2019): empirical-variational Bayes approximation $\rightarrow$ computationally expensive

# Previous work on group-adaptive elastic net

- Each group of variables obtains a group-specific elastic net penalty
- Groups that are relatively more important should obtain a smaller penalty
- Finding optimal penalties is hard:
  - `fwelnet` (Tay et al. 2020): boils down to (non-adaptive) group elastic net for grouped variables $\rightarrow$ may not be flexible enough compared to group-adaptive methods
  - `ipflasso` (Boulesteix et al. 2017): use cross-validation on predefined set of possible penalties $\rightarrow$ computationally expensive
  - `gren` (Münch et al. 2019): empirical-variational Bayes approximation $\rightarrow$ computationally expensive
- The presented method `squeezy`: group-adaptive, fast and for generic response

# Model

- Generalised linear model (GLM) for response $\boldsymbol{Y}$:

$$\boldsymbol{y} \sim \pi(\boldsymbol{Y}|X, \boldsymbol{\beta}),\ E(\boldsymbol{Y}) = g^{-1}(X\boldsymbol{\beta}),$$

with $g(\cdot)$ the link function. Plus: scaling parameter.

# Model

- Generalised linear model (GLM) for response $\boldsymbol{Y}$:

$$\boldsymbol{y} \sim \pi(\boldsymbol{Y}|X, \boldsymbol{\beta}),\ E(\boldsymbol{Y}) = g^{-1}(X\boldsymbol{\beta}),$$

with $g(\cdot)$ the link function. Plus: scaling parameter.

- Group-adaptive elastic net prior for the regression coefficients $\boldsymbol{\beta}$:

$$\pi(\beta_k|\alpha, \lambda_{g_k}) \propto \exp\left(-\lambda_{g_k}\left(\alpha|\beta_k| + (1-\alpha)\frac{1}{2}\beta_k^2\right)\right)$$

with hyperparameters:

  - $\alpha$: fixed mixing parameter between ridge ($\alpha = 0$; normal) and lasso ($\alpha = 1$)
  - $\lambda_{g_k}$: group-specific penalty for variable $k$ belonging to group $g_k \in \{1, .., G\}$

# Estimation of model parameters

- Estimate penalty parameters by empirical Bayes

# Estimation of model parameters

- Estimate penalty parameters by empirical Bayes
  - Marginal likelihood, equivalent to an exhaustive cross-validation (Fong and Holmes 2019)

# Estimation of model parameters

- Estimate penalty parameters by empirical Bayes
  - Marginal likelihood, equivalent to an exhaustive cross-validation (Fong and Holmes 2019)
  - Maximum marginal likelihood estimates for the group-specific penalties:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ \pi(\boldsymbol{y}|X, \alpha, \boldsymbol{\lambda}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ \int_{\boldsymbol{\beta}} \pi(\boldsymbol{y}|X, \boldsymbol{\beta}) \pi(\boldsymbol{\beta}|\alpha, \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\beta}$$

  - Relatively hard: use approximation

# Estimation of model parameters

- Estimate penalty parameters by empirical Bayes
  - Marginal likelihood, equivalent to an exhaustive cross-validation (Fong and Holmes 2019)
  - Maximum marginal likelihood estimates for the group-specific penalties:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ \pi(\boldsymbol{y}|X, \alpha, \boldsymbol{\lambda}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ \int_{\boldsymbol{\beta}} \pi(\boldsymbol{y}|X, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\alpha, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\beta}$$

  - Relatively hard: use approximation

- Once $\boldsymbol{\lambda}$ known: estimate $\boldsymbol{\beta}$ using standard implementation (e.g. `glmnet` for fast point estimate)

# Optimising the marginal likelihood

Optimise in three steps:

1. Consider group-adaptive *ridge* (normal prior): obtain fast ridge penalty estimates
2. For group-adaptive *elastic net*: show that the prior distribution for the linear predictors is approximately (multivariate) normal
3. Transform ridge penalties to elastic net penalties

## Step 1: fast ridge estimates

- Low-dimensional representation of marginal likelihood in terms of linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta} \in \mathbb{R}^n$ (Veerman, Leday, and Wiel 2019):

$$\pi(\boldsymbol{y}|X, \alpha, \boldsymbol{\lambda}) = \int_{\boldsymbol{\beta}} \pi(\boldsymbol{y}|X, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\alpha, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\beta} = \int_{\boldsymbol{\eta}} \pi(\boldsymbol{y}|\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\alpha, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\eta}$$

## Step 1: fast ridge estimates

- Low-dimensional representation of marginal likelihood in terms of linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta} \in \mathbb{R}^n$ (Veerman, Leday, and Wiel 2019):

$$\pi(\boldsymbol{y}|X, \alpha, \boldsymbol{\lambda}) = \int_{\boldsymbol{\beta}} \pi(\boldsymbol{y}|X, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\alpha, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\beta} = \int_{\boldsymbol{\eta}} \pi(\boldsymbol{y}|\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\alpha, \boldsymbol{\lambda})\mathrm{d}\boldsymbol{\eta}$$

- For ridge models ($\alpha = 0$) with group-specific ridge penalties $\boldsymbol{\lambda}_R$, the prior distribution of the linear predictors is known:

$$\boldsymbol{\eta} := X\boldsymbol{\beta} \sim N\left(\mathbf{0}, \sum_{g=1}^{G} \lambda_{R,g}^{-1} X_g X_g^T\right)$$

with $X_g$ the observed data for variable group $g$, $X_g X_g^T \in \mathbb{R}^{n \times n}$

## Step 1: fast ridge estimates

- Low-dimensional representation of marginal likelihood in terms of linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta} \in \mathbb{R}^n$ (Veerman, Leday, and Wiel 2019):

$$\pi(\boldsymbol{y}|X,\alpha,\boldsymbol{\lambda}) = \int_{\boldsymbol{\beta}} \pi(\boldsymbol{y}|X,\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\alpha,\boldsymbol{\lambda})\mathrm{d}\boldsymbol{\beta} = \int_{\boldsymbol{\eta}} \pi(\boldsymbol{y}|\boldsymbol{\eta})\pi(\boldsymbol{\eta}|\alpha,\boldsymbol{\lambda})\mathrm{d}\boldsymbol{\eta}$$

- For ridge models ($\alpha = 0$) with group-specific ridge penalties $\boldsymbol{\lambda}_R$, the prior distribution of the linear predictors is known:

$$\boldsymbol{\eta} := X\boldsymbol{\beta} \sim N\left(\mathbf{0}, \sum_{g=1}^{G} \lambda_{R,g}^{-1} X_g X_g^T\right)$$

with $X_g$ the observed data for variable group $g$, $X_g X_g^T \in \mathbb{R}^{n\times n}$

- Approximate the low-dimensional integral with a Laplace approximation and optimise to obtain $\hat{\boldsymbol{\lambda}}_R$

## Step 2: act normal

- Problem: for $\alpha \neq 0$ we do not know $\pi(\boldsymbol{\eta}|\boldsymbol{\lambda}, \alpha)$
- Solution: approximate by a multivariate normal:

$$\pi(\boldsymbol{\eta}|\alpha, \boldsymbol{\lambda}) \approx N\left(0, \sum_{g=1}^{G} v_g X_g X_g^T\right)$$

  with prior variances $v_g = \mathrm{Var}^{\mathsf{EN}}_{\beta_k|\alpha,\lambda_{g_k}}[\beta_k]$.
- $\hat{v}_g = \hat{\lambda}_{R,g}^{-1}$

# Why would this approximation work?

- The linear predictor for each sample $i$ is a *weighted* average:

$$\eta_i = X_{i,:}\boldsymbol{\beta} = \sum_{k=1}^{p} X_{i,k}\beta_k$$

# Why would this approximation work?

- The linear predictor for each sample $i$ is a *weighted* average:

$$\eta_i = X_{i,:}\boldsymbol{\beta} = \sum_{k=1}^{p} X_{i,k}\beta_k$$

- Multivariate CLT (Eicker 1966): $\boldsymbol{\eta}$ is asymptotically multivariate normal

## Why would this approximation work?

- The linear predictor for each sample $i$ is a *weighted* average:

$$\eta_i = X_{i,:}\boldsymbol{\beta} = \sum_{k=1}^{p} X_{i,k}\beta_k$$

- Multivariate CLT (Eicker 1966): $\boldsymbol{\eta}$ is asymptotically multivariate normal
  - Assumptions on $\boldsymbol{\beta}$: each $\beta_k$ has finite group-specific prior mean and variance
  - Assumptions on $X$: $X \in \mathbb{R}^{n \times p}$ should be 'dense' in $p$

# Why would this approximation work?

- The linear predictor for each sample $i$ is a *weighted* average:

$$\eta_i = X_{i,:}\boldsymbol{\beta} = \sum_{k=1}^{p} X_{i,k}\beta_k$$

- Multivariate CLT (Eicker 1966): $\boldsymbol{\eta}$ is asymptotically multivariate normal
  - Assumptions on $\boldsymbol{\beta}$: each $\beta_k$ has finite group-specific prior mean and variance
  - Assumptions on $X$: $X \in \mathbb{R}^{n \times p}$ should be 'dense' in $p$

# Why would this approximation work?

- The linear predictor for each sample $i$ is a *weighted* average:

$$\eta_i = X_{i,:}\boldsymbol{\beta} = \sum_{k=1}^{p} X_{i,k}\beta_k$$

- Multivariate CLT (Eicker 1966): $\boldsymbol{\eta}$ is asymptotically multivariate normal
  - Assumptions on $\boldsymbol{\beta}$: each $\beta_k$ has finite group-specific prior mean and variance
  - Assumptions on $X$: $X \in \mathbb{R}^{n \times p}$ should be 'dense' in $p$

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 1.22 \\ 0 & 0 & \cdots & 0 & -1.60 \\ 0 & 0 & \cdots & 0 & -0.61 \end{bmatrix} \begin{bmatrix} 0.44 & 0.36 & \cdots & 1.65 & 1.22 \\ 0.18 & 1.00 & \cdots & -0.90 & -1.60 \\ -1.00 & 0.18 & \cdots & 0.22 & -0.61 \end{bmatrix} \begin{bmatrix} 0.44 & 0.36 & \cdots & 1.65 & 12200 \\ 0.18 & 1.00 & \cdots & -0.90 & 16000 \\ -1.00 & 0.18 & \cdots & 0.22 & -61000 \end{bmatrix}$$



# Step 3: transform ridge to elastic net penalties

$h(\lambda_g) := \text{Var}^{\mathsf{EN}}_{\beta_k|\alpha,\lambda_g}[\beta_k] \overset{\mathsf{set}}{=} \hat{v}_g = \hat{\lambda}_{R,g}^{-1}$. Root finding: $\hat{\boldsymbol{\lambda}} = h^{-1}\left(\hat{\boldsymbol{\lambda}}_R^{-1}\right)$

# Step 3: transform ridge to elastic net penalties

$$h(\lambda_g) := \text{Var}^{\text{EN}}_{\beta_k|\alpha,\lambda_g}[\beta_k] \overset{\text{set}}{=} \hat{v}_g = \hat{\lambda}^{-1}_{R,g}.\ \text{Root finding: } \hat{\boldsymbol{\lambda}} = h^{-1}\left(\hat{\boldsymbol{\lambda}}^{-1}_R\right)$$

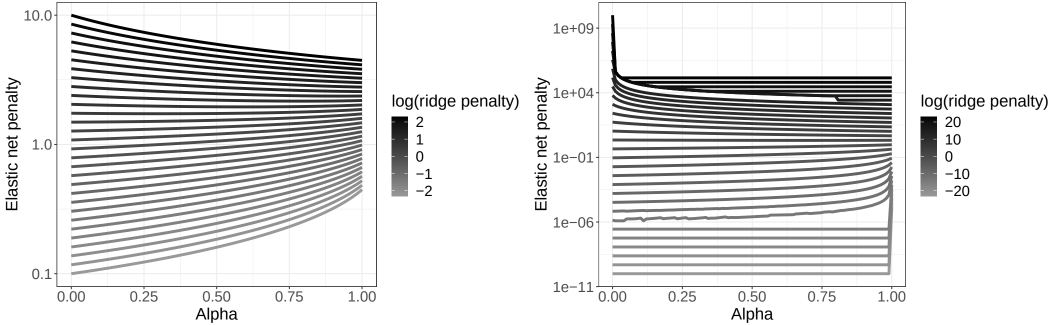



Figure: Transformation ridge penalty to elastic net penalty for different $\alpha$

# Application to classifying treatment response

- Main data ($n = 88, p = 2114$): miRNA expression
- Co-data: 8 variable groups, reflecting difference tumor vs normal tissue (other samples)
- Predict therapy response in colon cancer: clinical benefit vs disease progression

# Application to classifying treatment response

- Main data ($n = 88, p = 2114$): miRNA expression
- Co-data: 8 variable groups, reflecting difference tumor vs normal tissue (other samples)
- Predict therapy response in colon cancer: clinical benefit vs disease progression
- Assess computing time and performance in 10-fold cross-validation of the following methods:

| **Method** | **Group-adaptive** | **Generic in response** |
|---|---|---|
| elastic net (baseline) | x | v |
| fwEN (groups) | x | v |
| fwEN (continuous) | x | v |
| ipflasso | v/x | v |
| gren | v | x |
| squeezy | v | v |

# Results

**Group-penalties:**

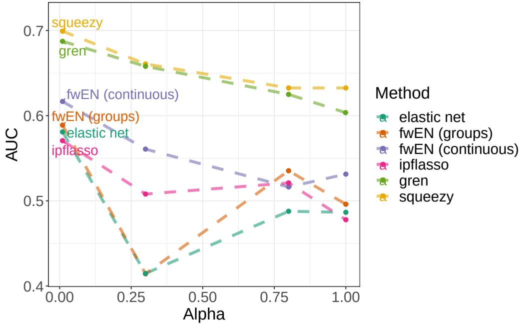
| | Mixing parameter | | |
|---|---|---|---|
| penalty | $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
| $\hat{\lambda}_1^{\text{EN}}$ | 100 | 24 | 14 |
| $\hat{\lambda}_2^{\text{EN}}, \dots, \hat{\lambda}_8^{\text{EN}}$ | $\approx 10^8$ | $\approx 14 * 10^3$ | $\approx 14 * 10^3$ |

# Results

## Group-penalties:

| penalty | Mixing parameter $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
|---|---|---|---|
| $\hat{\lambda}_1^{\text{EN}}$ | 100 | 24 | 14 |
| $\hat{\lambda}_2^{\text{EN}}, \dots, \hat{\lambda}_8^{\text{EN}}$ | $\approx 10^8$ | $\approx 14 * 10^3$ | $\approx 14 * 10^3$ |

## Predictive performance:

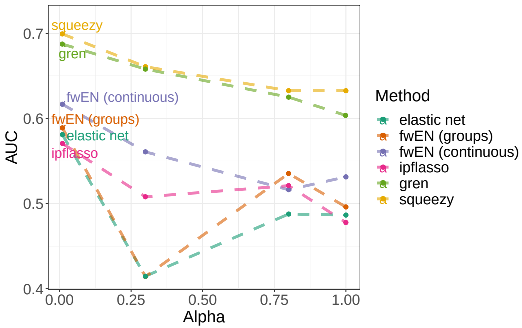


# Results

## Group-penalties:

| | Mixing parameter | | |
|---|---|---|---|
| penalty | $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
| $\hat{\lambda}_1^{\text{EN}}$ | 100 | 24 | 14 |
| $\hat{\lambda}_2^{\text{EN}}, \dots, \hat{\lambda}_8^{\text{EN}}$ | $\approx 10^8$ | $\approx 14 * 10^3$ | $\approx 14 * 10^3$ |

## Predictive performance:




## Computing time (10 folds):

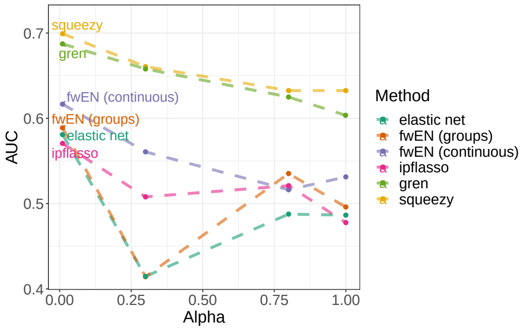
| Method | Time (s) |
|---|---|
| elastic net | 10.17 |
| fwEN (groups) | 120.20 |
| **squeezy** | **130.58** |
| fwEN (continuous) | 140.11 |
| **gren** | **3510.52** |
| ipflasso | 4220.39 |

# Results

## Group-penalties:

| penalty | Mixing parameter $\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 1$ |
|---|---|---|---|
| $\hat{\lambda}_1^{\text{EN}}$ | 100 | 24 | 14 |
| $\hat{\lambda}_2^{\text{EN}}, \ldots, \hat{\lambda}_8^{\text{EN}}$ | $\approx 10^8$ | $\approx 14 * 10^3$ | $\approx 14 * 10^3$ |

## Predictive performance:




## Computing time (10 folds):

| Method | Time (s) |
|---|---|
| elastic net | 10.17 |
| fwEN (groups) | 120.20 |
| **squeezy** | **130.58** |
| fwEN (continuous) | 140.11 |
| **gren** | **3510.52** |
| ipflasso | 4220.39 |

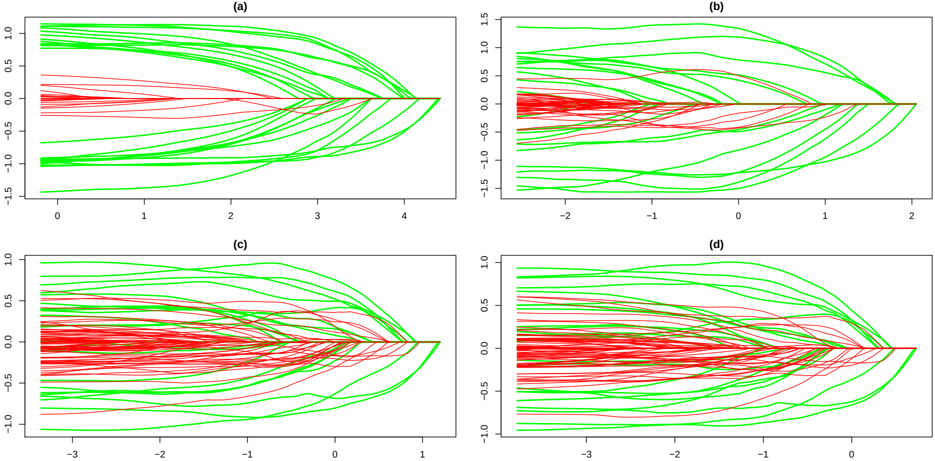
## Conclusion:

- Adaptive learning from co-data improves performance
- Squeezy performs as well as gren, but 25 times faster

# Better variable selection

**Simulation setting**

(a): Strongly informative co-data; (b) Weakly informative; (c) Non-informative; (d) No co-data



# Extensions

- Approximation works for other priors: spike-and-slab, bridge, but not horseshoe
- MVN can be visually checked a posteriori (Q-Q plot)
- Continuous co-data; overlapping groups

# Part II: Average treatment effect (ATE) estimation using adaptive EN

- Non-randomized setting
- Potential outcomes framework. ATE: $\theta = E[Y^{d=1}] - E[Y^{d=0}]$
- Standard doubly robust estimator
- Estimator for ATE in HD settings
- Link to Part I

# Why does a simple approach fail in HD settings?

- Standard doubly robust estimator:

$$\hat{E}[Y^{d=1}] = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i D_i}{\widehat{\mathsf{ps}}(X_i)} - \frac{\hat{Y}(1,X_i)(D_i - \widehat{\mathsf{ps}}(X_i))}{\widehat{\mathsf{ps}}(X_i)}\right\}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{Y}(1,X_i) + \frac{D_i(Y_i - \hat{Y}(1,X))}{\widehat{\mathsf{ps}}(X_i)}\right\}$$

$D_i = 0,1$ (treatment), $Y_i$ : response, $\hat{Y}(1,X_i)$: response prognosis, $\widehat{\mathsf{ps}}(X_i)$: propensity score (may depend on clinical confounders as well)

- Likewise, $\hat{E}[Y^{d=0}]$. ATE: $\hat{\theta} = \hat{E}[Y^{d=1}] - \hat{E}[Y^{d=0}]$

# Why does a simple approach fail in HD settings?

- Standard doubly robust estimator:

$$\hat{E}[Y^{d=1}] = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i D_i}{\widehat{\mathsf{ps}}(X_i)} - \frac{\hat{Y}(1,X_i)(D_i - \widehat{\mathsf{ps}}(X_i))}{\widehat{\mathsf{ps}}(X_i)}\right\}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{Y}(1,X_i) + \frac{D_i(Y_i - \hat{Y}(1,X))}{\widehat{\mathsf{ps}}(X_i)}\right\}$$

$D_i = 0, 1$ (treatment), $Y_i$ : response, $\hat{Y}(1, X_i)$: response prognosis, $\widehat{\mathsf{ps}}(X_i)$: propensity score (may depend on clinical confounders as well)

- Likewise, $\hat{E}[Y^{d=0}]$. ATE: $\hat{\theta} = \hat{E}[Y^{d=1}] - \hat{E}[Y^{d=0}]$
- Works when either $\hat{Y}(t, X_i), t = 0, 1$ or $\widehat{\mathsf{ps}}(X_i)$ is unbiased.
- In high-dimensional settings both are inevitably biased.

# Solution: sample splitting

Chernozhukov et al. 2018 show that sample splitting provides a solution:

$$\tilde{E}[Y^{d=1}] = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{Y}^{(-i)}(1, X_i) + \frac{D_i(Y_i - \tilde{Y}^{(-i)}(1, X_i))}{\widetilde{\mathsf{ps}}^{(-i)}(X_i)}\right\},$$

where $\tilde{Y}^{(-i)}(d, X_i), \widetilde{\mathsf{ps}}^{(-i)}(X_i)$ refer to (mean) predictions on models learnt without sample $i$.

Can be extended to *local* ATE ($=$ function of $X_i$) $\rightarrow$ relevant for treatment optimization.

## Links to Part I

- Trivial link: three prediction problems, $Y(0, X), Y(1, X), \mathsf{ps}(X)$.
  - Adaptive EN is a candidate learner for these
  - Sample splitting $\Rightarrow$ computational efficiency for penalty estimation crucial

# Links to Part I

- Trivial link: three prediction problems, $Y(0, X), Y(1, X), \text{ps}(X)$.
  - Adaptive EN is a candidate learner for these
  - Sample splitting $\Rightarrow$ computational efficiency for penalty estimation crucial

- Coupling predictions $Y(0, X)$ and $Y(1, X)$. Attractive to limit or shrink the number of parameters to estimate.

## Coupling predictions $Y(0,X)$ and $Y(1,X)$

1. Homogenous effect of variables.
   - $Y(D_i, X_i) = \beta_0 + D_i\gamma_0 + X_i\boldsymbol{\beta}$
   - Often too simple, not realistic.

# Coupling predictions $Y(0, X)$ and $Y(1, X)$

1. Homogenous effect of variables.
   - $Y(D_i, X_i) = \beta_0 + D_i\gamma_0 + X_i\boldsymbol{\beta}$
   - Often too simple, not realistic.

2. Interaction effect
   - Interaction effect. $Y(D_i, X_i) = \beta_0 + D_i\gamma_0 + X_i\boldsymbol{\beta} + D_iX_i\boldsymbol{\gamma}$
   - $\boldsymbol{\beta}$ : EN$(\alpha, \lambda_{\boldsymbol{\beta}})$ prior; $\boldsymbol{\gamma}$ : EN$(\alpha, \lambda_{\boldsymbol{\gamma}})$ prior
   - Shrinks differential effects ($\boldsymbol{\gamma}$) to 0. Squeezy applies to estimate $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$

# Coupling predictions $Y(0, X)$ and $Y(1, X)$

1. Homogenous effect of variables.
   - $Y(D_i, X_i) = \beta_0 + D_i\gamma_0 + X_i\boldsymbol{\beta}$
   - Often too simple, not realistic.

2. Interaction effect
   - Interaction effect. $Y(D_i, X_i) = \beta_0 + D_i\gamma_0 + X_i\boldsymbol{\beta} + D_iX_i\boldsymbol{\gamma}$
   - $\boldsymbol{\beta}$ : EN$(\alpha, \lambda_{\boldsymbol{\beta}})$ prior; $\boldsymbol{\gamma}$ : EN$(\alpha, \lambda_{\boldsymbol{\gamma}})$ prior
   - Shrinks differential effects ($\boldsymbol{\gamma}$) to 0. Squeezy applies to estimate $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$

3. Control group as prior
   - $Y(D_i = 0, X_i) = \beta_0 + X_i\boldsymbol{\beta}$; and $Y(D_i = 1, X_i) = \gamma_0 + X_i\boldsymbol{\gamma}$
   - $\gamma_k$ : EN$(\alpha, \lambda_{g_k(\hat{\beta}_k)})$ prior. E.g. $g_k(0) = 1; g_k(\hat{\beta}_k) = 2$, for $\hat{\beta}_k \neq 0$
   - Attractive when control group is large, (experimental) treatment group is small
   - Squeezy applies to estimate $\lambda$'s; extends efficiently to inclusion of variable groups.

- More information on the adaptive EN in (van Nee, van de Brug, and van de Wiel 2021)
- R-package squeezy available on CRAN


Boulesteix, Anne-Laure et al. (2017). "IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data". In: *Computational and mathematical methods in medicine* 2017.

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68.

Eicker, F (1966). "A multivariate central limit theorem for random linear vector forms". In: *The Annals of Mathematical Statistics*, pp. 1825–1828.

Fong, Edwin and Chris Holmes (2019). "On the marginal likelihood and cross-validation". In: *arXiv preprint arXiv:1905.08737*.

Münch, Magnus M et al. (Dec. 2019). "Adaptive group-regularized logistic elastic net regression". In: *Biostatistics*. kxz062. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxz062.

Tay, J Kenneth et al. (2020). "Feature-weighted elastic net: using" features of features" for better prediction". In: *arXiv preprint arXiv:2006.01395*.

van Nee, Mirrelijn M, Tim van de Brug, and Mark A van de Wiel (2021). "Fast marginal likelihood estimation of penalties for group-adaptive elastic net". In: *arXiv preprint arXiv:2101.03875*.

Veerman, Jurre R, Gwenaël GR Leday, and Mark A van de Wiel (2019). "Estimation of variance components, heritability and the ridge penalty in high-dimensional generalized linear models". In: *Communications in Statistics-Simulation and Computation*, pp. 1–19.

## Appendix: multivariate central limit theorem

**Multivariate central limit theorem for linear random vector forms (Eicker 1966)**

Suppose $\beta_j \overset{ind.}{\sim} \pi_{g_j}(\beta_j)$ for $j = 1, \ldots, p$ and group-specific prior $\pi_{g_j}(\cdot)$, with $E(\beta_j) = 0$ and $\mathrm{Var}(\beta_j) = \tau^2_{g_j} \in (0, \infty)$. For $g = 1, \ldots, G$, let $G_g$ be the group size and let $X_g \in \mathbb{R}^{n \times G_g}$ be the weights corresponding to group $g$. Let $X_{*j}$ denote the $j^{th}$ column of $X \in \mathbb{R}^{n \times p}$. Suppose $X_{*j} \neq \mathbf{0} \in \mathbb{R}^n$ for all $j$, $rank(X) = n$ for all $p$, and for $p \to \infty$,

$$\max_{j=1,\ldots,p} X^T_{*j} \left(XX^T\right)^{-1} X_{*j} \to 0. \tag{1}$$

Then, for fixed $G$, fixed $n$, and $p \to \infty$,

$$\left(\sum_g \tau^2_g X_g X^T_g\right)^{-1/2} X\boldsymbol{\beta} \xrightarrow{d} N(0, I_{n \times n}), \tag{2}$$

where $I_{n \times n}$ is the $(n \times n)$-dimensional identity matrix and $\left(\sum_g \tau^2_g X_g X^T_g\right)^{-1/2}$ is the inverse of the unique positive definite square root of $\sum_g \tau^2_g X_g X^T_g$.

If $n = 1$ then condition (1) is equivalent to $\max_{j=1,\ldots,p} x^2_{1j} \Big/ \sum_{j=1}^p x^2_{1j} \to 0$, for $p \to \infty$. Informally, condition (1) can be interpreted as each variable being asymptotically negligible in size compared to the full data set.