Improving high-dimensional prediction by empirical Bayes learning from co-data

Mark van de Wiel[1,2,3], Magnus Münch[1,4]

[1]Dep of Epidemiology and Biostatistics, VU University medical center

[2]Dep of Mathematics, VU University, Amsterdam, NL

[3]MRC Biostatistics unit, Cambridge University, UK (Visiting Fellow)

[4]Mathematical Institute, Leiden University, NL

Our group: `www.bigstatistics.nl`

Setting

- **Prediction or Diagnosis**
- **Primary data**
  - Variables $i = 1, \dots, p$; Individuals $j = 1, \dots, n$; $p > n$
  - Focus on binary response Y_j (e.g. case vs control)
  - Measurements $\mathbf{X}_j = (X_{1j}, \dots, X_{pj})$
  - Goal: find f such that $Y_j \approx f(\mathbf{X}_j)$
  - Here, f: *logistic regression*
  - Some form of regularization required
- **Focus**
  - Differential regularization based on prior information: **co-data**

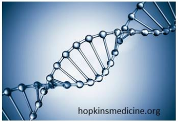
Co-data

Definition Co-data: any information on the *variables* that does not use the response labels of the primary data

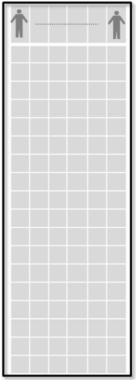
Co-data

Definition Co-data: any information on the *variables* that does not use the response labels of the primary data

Databases

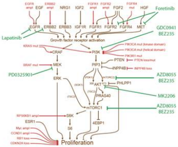


Related bio-molecules

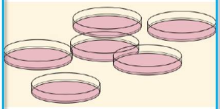


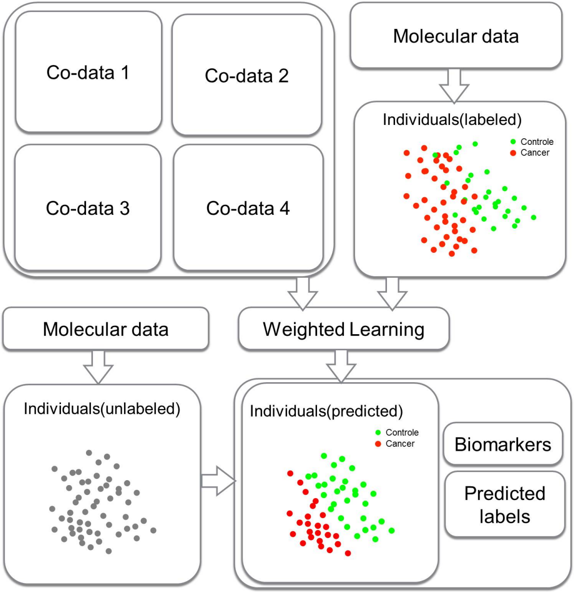
Primary Data

Pathways



Cell lines




Co-data 1
Co-data 2
Co-data 3
Co-data 4
Molecular data
Individuals(labeled)
Controle
Cancer
Molecular data
Weighted Learning
Individuals(unlabeled)
Individuals(predicted)
Controle
Cancer
Biomarkers
Predicted labels

Use of co-data

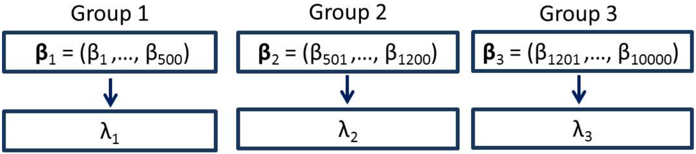
Groups: Co-data determine G prior groups of variables

Idea: Use different penalty weights $\lambda_1, \ldots, \lambda_G$ across G co-data-based groups.

Use of co-data

Groups: Co-data determine G prior groups of variables

Idea: Use different penalty weights $\lambda_1, \ldots, \lambda_G$ across G co-data-based groups. $G = 3$:



E.g. Ridge: $\text{argmax}_{\beta}\{\mathcal{L}(\mathbf{Y}; \beta) - \sum_{g=1}^{G} \lambda_g ||\boldsymbol{\beta}_g||_2\}$

$\rightarrow$ **CV** not attractive

Empirical Bayes (EB)

Empirical Bayes: estimate hyper-parameters from data

Relation penalty parameters $\leftrightarrow$ hyper-parameters (prior)

Empirical Bayes (EB)

Empirical Bayes: estimate hyper-parameters from data

Relation penalty parameters $\leftrightarrow$ hyper-parameters (prior)

E.g. logistic ridge: $\beta_i \sim N(0, \sigma_g^2), i \in \text{group}_g; \lambda_g = 1/(2\sigma_g^2)$:

$$\text{argmax}_{\boldsymbol{\beta}}\{\mathcal{L}(\mathbf{Y};\boldsymbol{\beta}) - \sum_{g=1}^{G} \lambda_g ||\boldsymbol{\beta}_g||_2\} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\sigma}}^{\text{MAP}} = \text{mode}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\beta}|\mathbf{Y}))$$

Previous work

- **EB**: Morris, Carlin & Louis, Efron, George, Casella, Van Houwelingen, etc.
- Blog: David Robinson: varianceexplained.org
- Review: EB for high-dimensional prediction*
  - High-dimensional vs low-dimensional
  - Theory on EB estimator ($p \uparrow$) for simple linear case
  - Various EB methodologies
  - Spike-and-slab

*VdW, Münch, arXiv, to appear: *Scand J Stat*

Previous work

- **EB**: Morris, Carlin & Louis, Efron, George, Casella, Van Houwelingen, etc.
- Blog: David Robinson: varianceexplained.org
- Review: EB for high-dimensional prediction*
  - High-dimensional vs low-dimensional
  - Theory on EB estimator ($p \uparrow$) for simple linear case
  - Various EB methodologies
  - Spike-and-slab
- **Groups**: group-lasso (Meier et al.) + many versions thereof

*VdW, Münch, arXiv, to appear: *Scand J Stat*

Formal EB: Maximum marginal Likelihood

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Prior(s): $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

Marginal likelihood maximization:

$$\hat{\boldsymbol{\alpha}} = \text{argmax}_{\alpha} \text{ML}(\boldsymbol{\alpha}), \text{ with } \text{ML}(\boldsymbol{\alpha}) = \int_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) \pi_{\boldsymbol{\alpha}}(\boldsymbol{\beta}) d\boldsymbol{\beta},$$

Formal EB: Maximum marginal Likelihood

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Prior(s): $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\beta})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

Marginal likelihood maximization:

$$\hat{\boldsymbol{\alpha}} = \text{argmax}_{\boldsymbol{\alpha}} \text{ML}(\boldsymbol{\alpha}), \text{ with } \text{ML}(\boldsymbol{\alpha}) = \int_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) \pi_{\boldsymbol{\alpha}}(\boldsymbol{\beta}) d\boldsymbol{\beta},$$

Optimization hard, because of the high-dimensional integral

- Laplace approximation (Shun & McCullagh, *JRSSB*, 1995)
- EM on Gibbs samples (Casella, *Biostatistics*, 2001) or on Variational Bayes approximation (Part II: Elastic Net).
- Moment estimation

EB using moments: group-regularized ridge

Estimate σ_g^2 ($\lambda_g \propto \sigma_g^{-2}$), for ridge: $\beta_i \sim N(0, \sigma_g^2), i \in$ group g

EB using moments: group-regularized ridge

Estimate σ_g^2 ($\lambda_g \propto \sigma_g^{-2}$), for ridge: $\beta_i \sim N(0, \sigma_g^2), i \in$ group g

Intuitive Idea:

1. Run an initial ridge regression with one λ
2. For $g = 1, 2$, consider mean squares of coefficients:

$$MS_g = \frac{1}{p_g} \sum_{i \in \text{group } g} \hat{\beta}_i^2$$

3. If MS_g is large then σ_g^2 should be large (hence λ_g small)

EB using moments: group-regularized ridge

Estimate σ_g^2 ($\lambda_g \propto \sigma_g^{-2}$), for ridge: $\beta_i \sim N(0, \sigma_g^2), i \in$ group g

Intuitive Idea:

1. Run an initial ridge regression with one λ
2. For $g = 1, 2$, consider mean squares of coefficients:

$$MS_g = \frac{1}{p_g} \sum_{i \in \text{group } g} \hat{\beta}_i^2$$

3. If MS_g is large then σ_g^2 should be large (hence λ_g small)

More difficult, because $E(MS_g)$ depends also on variables *not in* group g (biased estimation)

EB using moment estimation[†]

Two-group example: estimate σ_1^2, σ_2^2 ($\lambda_g \propto \sigma_g^{-2}$), for ridge:

$\beta_i \sim N(0, \sigma_1^2), i \in$ group 1, $\beta_i \sim N(0, \sigma_2^2), i \in$ group 2

Idea: equate empirical moment(s) to theoretical ones

[†] Details: Van de Wiel et al., *Stat Med*, 2016

EB using moment estimation†

Two-group example: estimate σ_1^2, σ_2^2 ($\lambda_g \propto \sigma_g^{-2}$), for ridge:

$\beta_i \sim N(0, \sigma_1^2), i \in$ group 1, $\beta_i \sim N(0, \sigma_2^2), i \in$ group 2

Idea: equate empirical moment(s) to theoretical ones

$$\frac{1}{p_1} \sum_{i \in \text{group 1}} \hat{\beta}_i^2 \approx \frac{1}{p_1} \sum_{i \in \text{group 1}} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := f_1(\sigma_1^2, \sigma_2^2)$$
$$\frac{1}{p_2} \sum_{i \in \text{group 2}} \hat{\beta}_i^2 \approx \frac{1}{p_2} \sum_{i \in \text{group 2}} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := f_2(\sigma_1^2, \sigma_2^2),$$

Result: System of equations $\mathbf{b}_{\text{data}} = A\mathbf{x}$, $\lambda_g^{-1} \propto \hat{\sigma}_g^2 = x_g$.

---

†Details: Van de Wiel et al., *Stat Med*, 2016

Shrink the shrinkage parameters[‡]

Co-data may consist of many groups (e.g. pathways)

$\rightarrow \hat{\boldsymbol{\sigma}}^2 = A^{-1}\mathbf{b}_{\text{data}}$ instable $\rightarrow$ over-fitting.

[‡]Details: Novianti et al., *Bioinformatics*, 2017

Shrink the shrinkage parameters[‡]

Co-data may consist of many groups (e.g. pathways)

$\rightarrow \hat{\boldsymbol{\sigma}}^2 = A^{-1}\mathbf{b}_{\text{data}}$ instable $\rightarrow$ over-fitting.

Solution: shrink A to stable target matrix, e.g. $T = \text{diag}(A)$:

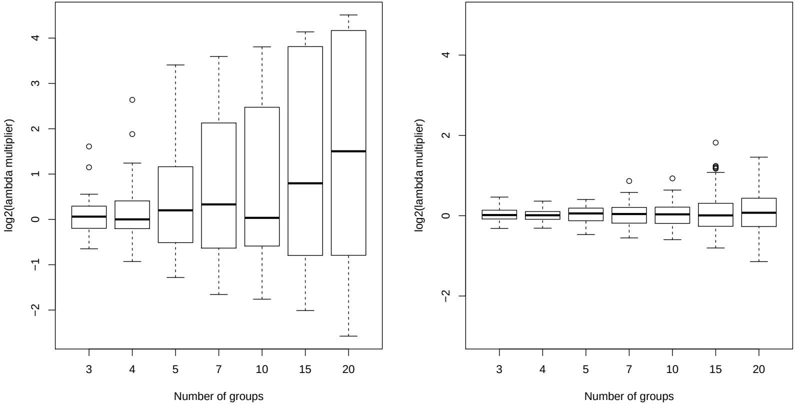
$$\tilde{A}_q = qA + (1-q)T$$

[‡]Details: Novianti et al., *Bioinformatics*, 2017

Effect of shrinkage

Real data, *random* groups of variables; Penalties: $\lambda_g = \lambda'_g \lambda$

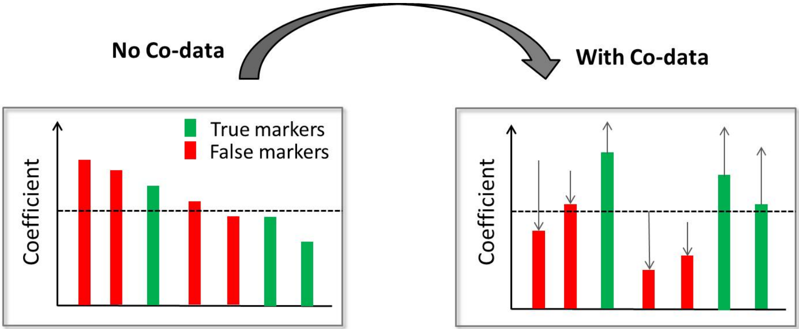
λ'_g: lambda multiplier; $\log_2(\lambda'_g)$ should $\approx \log_2(1) = 0$



Left: No Shrinkage; Right: Shrinkage

Suppose we want variable selection...

Why can co-data help?



Suppose we want variable selection...

Nicest solution: A coherent framework for EB estimation in a group-regularized elastic net setting[§]

[§] Part II

Suppose we want variable selection...

Nicest solution: A coherent framework for EB estimation in a group-regularized elastic net setting[§]

Ad-hoc solution:

1. Estimate group penalties from ridge regression, possibly for multiple groupings
2. Select k variables by introducing non-grouped L_1 penalty
3. Refit the model using the selected variables and their respective L_2 penalties

---

[§] Part II

Software[¶]

R-package `GRridge`, Github + Bioconductor

[¶] To be discussed during course

Software¶

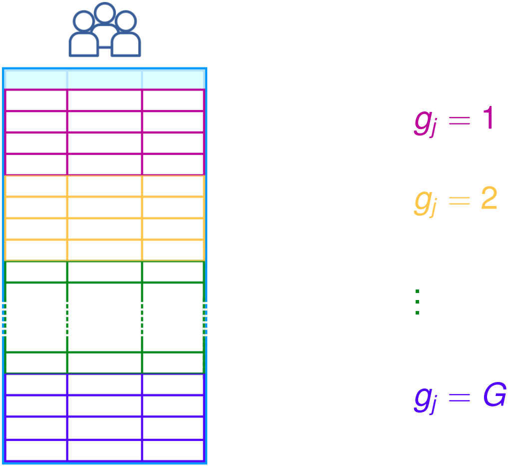
R-package `GRridge`, Github + Bioconductor

- Allows iteration; CVlik as stopping criterion
- Allows *multiple* sources of co-data, as groups
- Allows *overlapping* groups, e.g. pathways
- Auxiliary functions for co-data processing
- Built-in CV for comparison with ridge & lasso

¶ To be discussed during course

Part II: Group-regularized elastic net[‖]

Group of feature j: g_j.



[‖]Magnus Münch et al.

Group-regularized elastic net

Model

$$Y_i|\boldsymbol{\beta} \sim \text{Bern}(\text{expit}(\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\beta})),$$

$$\beta_j \overset{ind}{\sim} \exp\left[-\frac{1}{2}\left(\alpha\lambda \cdot \sqrt{\lambda'_{g(j)}}|\beta_j| + (1-\alpha)\lambda \cdot \lambda'_{g(j)}\beta_j^2\right)\right]$$

- Shrinks estimates towards zero
- 'Global' α and λ determine overall shrinkage
- Elastic net with penalty weights $w_{g(j)} = (\lambda'_{g(j)})^{1/2}$:
  $\alpha\lambda|w_{g(j)} \cdot \beta_j| + (1-\alpha)\lambda(w_{g(j)} \cdot \beta_j)^2$

Penalty parameter estimation

Cross-validation

- Prohibitively slow and unstable with even few groups

Hybrid CV and Empirical Bayes

- Fix α and estimate λ by CV for global shrinkage
- Empirical Bayes estimation of $\boldsymbol{\lambda}'$ by MML

Maximum marginal likelihood (MML)

$$\hat{\boldsymbol{\lambda}}' = \text{argmax}_{\boldsymbol{\lambda}'} \int_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) \pi_{\boldsymbol{\lambda}'}(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

Latent variables

Extra latent variables (Polson et al., 2013; Li & Nin, 2010)

- $\boldsymbol{\omega}|\boldsymbol{\beta} \sim \prod_{i=1}^{n} \mathcal{PG}(1, |\mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}|)$, independent of Y_i
- $\boldsymbol{\beta}|\boldsymbol{\tau} \sim \prod_{j=1}^{p} \mathcal{N}\left(0, \frac{\tau_j - 1}{\lambda'_{g(j)}(1-\alpha)\lambda\tau_j}\right)$ and
  $\boldsymbol{\tau} \sim \prod_{j=1}^{p} \mathcal{TG}\left(\frac{1}{2}, \frac{8(1-\alpha)}{\alpha^2\lambda}, (1, \infty)\right)$

Latent variables

Extra latent variables (Polson et al., 2013; Li & Nin, 2010)

- $\boldsymbol{\omega}|\boldsymbol{\beta} \sim \prod_{i=1}^{n} \mathcal{PG}(1, |\mathbf{X}_i^\mathsf{T}\boldsymbol{\beta}|)$, independent of Y_i
- $\boldsymbol{\beta}|\boldsymbol{\tau} \sim \prod_{j=1}^{p} \mathcal{N}\left(0, \frac{\tau_j - 1}{\lambda'_{g(j)}(1-\alpha)\lambda\tau_j}\right)$ and
  $\boldsymbol{\tau} \sim \prod_{j=1}^{p} \mathcal{TG}\left(\frac{1}{2}, \frac{8(1-\alpha)}{\alpha^2\lambda}, (1, \infty)\right)$

Computational reasons

- $\boldsymbol{\omega}$ renders logistic part ‘easy’: it disappears in the calculations
- $\boldsymbol{\tau}$ makes posterior calculations of $\boldsymbol{\beta}$ easier

EM algorithm

Recap Casella (2001):

$$\boldsymbol{\lambda}'^{(k+1)} = \text{argmax}_{\boldsymbol{\lambda}'} \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\tau}|\mathbf{Y}} \left[\log \mathcal{L}_{\boldsymbol{\lambda}'}(\mathbf{Y}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{\lambda}'^{(k)}) \right].$$

EM algorithm

Recap Casella (2001):

$$\boldsymbol{\lambda}'^{(k+1)} = \text{argmax}_{\boldsymbol{\lambda}'} \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\tau}|\mathbf{Y}} \left[\log \mathcal{L}_{\boldsymbol{\lambda}'}(\mathbf{Y}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{\lambda}'^{(k)}) \right].$$

Exact expectation is difficult, options:

- Monte Carlo approximation: slow
- Laplace approximation: not accurate in high dimensional space
- Variational Bayes: fast and accurate (for the posterior mean)

Empirical-variational Bayes

Variational Bayes

Approximate posterior factorizes:

$$p(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{Y}) \approx q(\boldsymbol{\omega}) q(\boldsymbol{\beta}) q(\boldsymbol{\tau}) =: Q$$

$$\downarrow$$

$$\mathbb{E}_{p(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{Y})} [\log \mathcal{L}_{\boldsymbol{\lambda}'}(\cdot)] \approx \mathbb{E}_Q [\log \mathcal{L}_{\boldsymbol{\lambda}'}(\cdot)] =: f(\boldsymbol{\lambda}')$$

Empirical-variational Bayes

Variational Bayes
Approximate posterior factorizes:

$$p(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{Y}) \approx q(\boldsymbol{\omega}) q(\boldsymbol{\beta}) q(\boldsymbol{\tau}) =: Q$$

$$\downarrow$$

$$\mathbb{E}_{p(\boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{Y})} \left[\log \mathcal{L}_{\boldsymbol{\lambda}'}(\cdot) \right] \approx \mathbb{E}_{Q} \left[\log \mathcal{L}_{\boldsymbol{\lambda}'}(\cdot) \right] =: f(\boldsymbol{\lambda}')$$

EM algorithm

- E-step is an iterative VB algorithm itself to find Q.
- M-step, $\text{argmax}_{\boldsymbol{\lambda}'} f(\boldsymbol{\lambda}')$, is now convex and easily solved.

Automatic feature selection

Feature selection

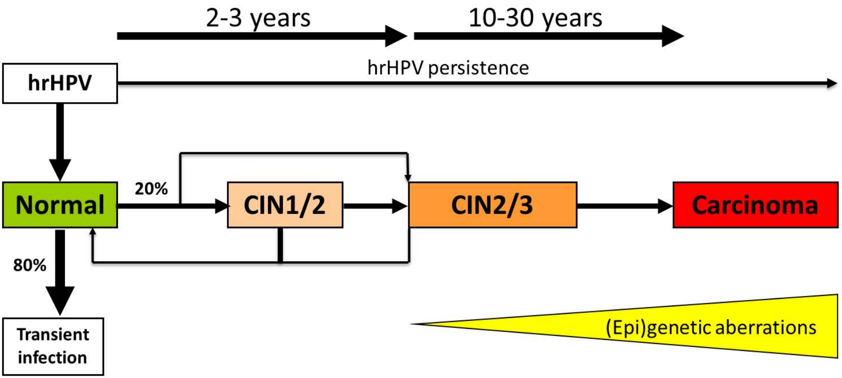
1. Plug estimated penalty parameters into frequentist elastic net:

$$\hat{\boldsymbol{\beta}} := \mathrm{argmax}_{\boldsymbol{\beta}} \log \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) + \frac{\alpha\lambda}{2} \sum_{j=1}^{p} \sqrt{\lambda'_{g(j)}} |\beta_j| + \frac{(1-\alpha)\lambda}{2} \sum_{j=1}^{p} \lambda'_{g(j)} \beta_j^2$$

2. Adjust λ until desired number of features selected

- The L_1-norm penalty term ensures automatic feature selection
- Estimated penalty multipliers may enhance predictive performance

Example: Cervical cancer



Goal: Detect CIN3 lesions, to be removed surgically

Example: Diagnostics for cervical cancer

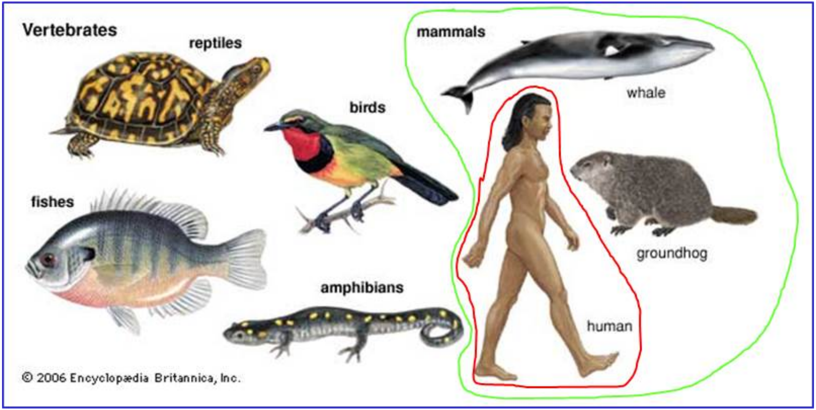
Goal: Select markers for classifying Normal vs CIN3
$\rightarrow$ final goal is a cheap PCR assay

Data:

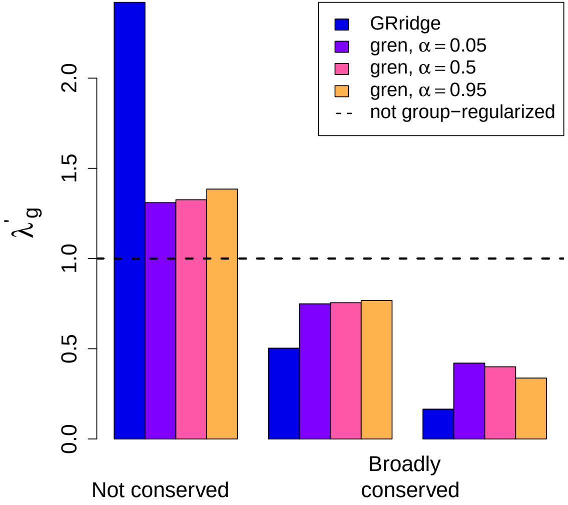
- microRNA sequencing data on *self-samples*
- $n = 56$: 32 Normal, 24 CIN3
- $p = 772$ (after filtering lowly abundant ones).
- Sqrt-transformed
- Standardized

Co-data: Conservation status

1. Non-conserved, human only (552)
2. Conserved across mammals (72)
3. Broadly conserved, across most vertebrates (148)



Co-data results

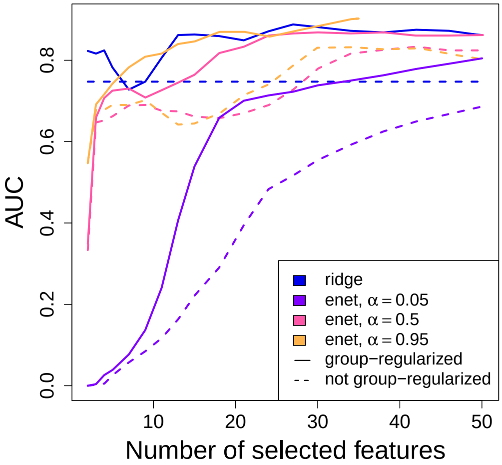


Clinician:

"That's all nice, but does the predictive accuracy improve?"

Performance under variable selection

AUC assessed by LOOCV



Extensions, other co-data applications

Generalized ridge: covariance structures (in progress)

Random Forest: Allows flexible co-data.[**]

Networks: Bayesian SEM: VB + EB + prior network[††]

Hybrid Bayes-Empirical Bayes: $\lambda_g = \lambda\lambda'_g$, $\lambda \sim$ hyper-prior, λ'_g fixed. Example in the Review.

[**] Te Beest, et al., *BMC Bioinf*, 2017

[††] Leday, Kpogbezan, et al., *Ann Appl Stat*, 2016; *Biom J*, 2017

Thanks

Magnus Münch (Leiden Univ / VUmc)

Thanks

Magnus Münch (Leiden Univ / VUmc)

Cervical cancer data: Saskia Wilting (Erasmus MC), Barbara Snoek (VUmc)

Co-data: Putri Novianti (VUmc)

Stats: Wessel van Wieringen, Carel Peeters (VUmc); Aad van der Vaart (Leiden Univ)

QUESTIONS?[‡‡]

COURSE: Please install `GRridge`, `gren` and dependencies.

See `https://magnusmunch.github.io/co-data_learning/`

[‡‡] Slides available via: www.bigstatistics.nl