Empirical Bayes learning from co-data in high-dimensional prediction settings

Mark van de Wiel[1,2]

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

[2]Dep of Mathematics, VU University, Amsterdam, The Netherlands

Contributions by: **Putri Novianti** (VUmc), **Magnus Münch** (Leiden, VUmc)

Our group: `www.bigstatistics.nl`

Setting

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

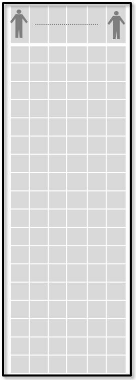
Co-data

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

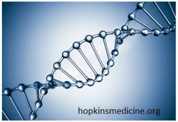
Co-data

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

Databases

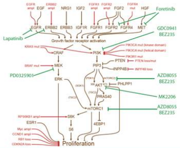


Primary Data

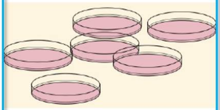


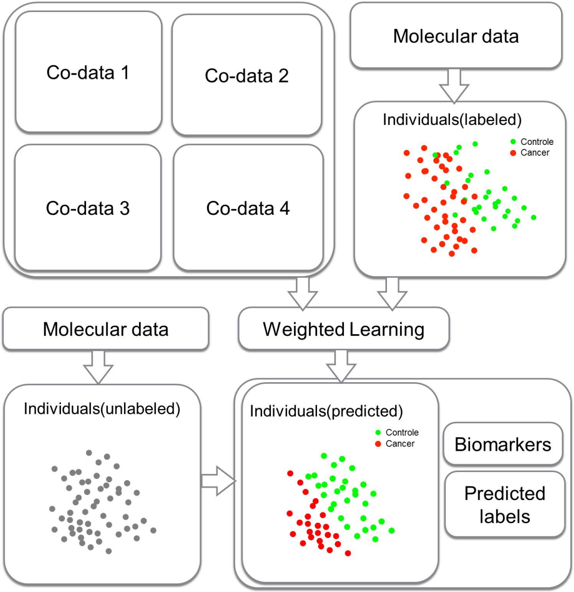
Related bio-molecules

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. E.g. in ridge:

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

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. E.g. in ridge:

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

Challenge: Estimation of hyperparameters λ_g

CV not attractive

Empirical Bayes (EB)

Definition[*]: EB estimates the prior from data

$\rightarrow$ Parametric form: estimate prior parameters
$\rightarrow$ Penalized regression: estimate penalty parameters; via link with prior

[*] Excellent discussions: Carlin & Louis (2000), Efron (2010), Van Houwelingen (2014)

Empirical Bayes (EB)

Definition[*]: EB estimates the prior from data

$\rightarrow$ Parametric form: estimate prior parameters
$\rightarrow$ Penalized regression: estimate penalty parameters; via link with prior

Why Empirical Bayes (EB)?

- EB estimators tend to improve for increasing p

[*] Excellent discussions: Carlin & Louis (2000), Efron (2010), Van Houwelingen (2014)

Empirical Bayes (EB)

Definition*: EB estimates the prior from data

→ Parametric form: estimate prior parameters
→ Penalized regression: estimate penalty parameters; via link with prior

Why Empirical Bayes (EB)?

- EB estimators tend to improve for increasing p
- EB fits well with allowing for prior information: can improve predictions

---

*Excellent discussions: Carlin & Louis (2000), Efron (2010), Van Houwelingen (2014)

Empirical Bayes (EB)

Definition*: EB estimates the prior from data

$\rightarrow$ Parametric form: estimate prior parameters
$\rightarrow$ Penalized regression: estimate penalty parameters; via link with prior

Why Empirical Bayes (EB)?

- EB estimators tend to improve for increasing p
- EB fits well with allowing for prior information: can improve predictions
- Computationally nicer than Full Bayes and CV

---

*Excellent discussions: Carlin & Louis (2000), Efron (2010), Van Houwelingen (2014)

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:

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

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:

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

High-dimensional integral $\rightarrow$ optimization hard

High-dimensional integral

Solutions:

- Laplace approximation (Shun & McCullagh, 1995)
- EM on Gibbs samples (Casella, 2001). Conceptually easy, but computationally very intensive.
- EM on Variational Bayes approximation (Bernardo et al., 2003). Fast, but dedicated approximations†.

†Work in progress for elastic net and spike-and-slab

High-dimensional integral

Solutions:

- Laplace approximation (Shun & McCullagh, 1995)
- EM on Gibbs samples (Casella, 2001). Conceptually easy, but computationally very intensive.
- EM on Variational Bayes approximation (Bernardo et al., 2003). Fast, but dedicated approximations†.
- **Or** resort to alternative EB approach

† Work in progress for elastic net and spike-and-slab

Back to the ridge example

Empirical Bayes (EB) estimation of λ_g explores

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

when

$$j \in \text{Group } g : \beta_j \sim N(0, \tau_g^2), \tau_g^{-2} \propto \lambda_g$$

Back to the ridge example

Empirical Bayes (EB) estimation of λ_g explores

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

when

$$j \in \text{Group } g : \beta_j \sim N(0, \tau_g^2), \tau_g^{-2} \propto \lambda_g$$

$\rightarrow$ EB estimate of τ_g^2 renders estimate of λ_g.

EB for group-regularized ridge‡

Aim: $\hat{\tau}_g^2$ for group-regularized ridge: $\beta_i \sim N(0, \tau_g^2), i \in \mathcal{G}_g$

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

EB for group-regularized ridge[‡]

Aim: $\hat{\tau}_g^2$ for group-regularized ridge: $\beta_i \sim N(0, \tau_g^2), i \in \mathcal{G}_g$

Initial: $\hat{\beta}_i = \hat{\beta}_i^{\lambda_0}$. Moment equations $g = 1, \ldots, G$. $G = 2$:

$$\frac{1}{p_1} \sum_{i \in \mathcal{G}_1} \hat{\beta}_i^2 \approx \frac{1}{p_1} \sum_{i \in \mathcal{G}_1} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := h_1(\tau_1^2, \tau_2^2)$$

$$\frac{1}{p_2} \sum_{i \in \mathcal{G}_2} \hat{\beta}_i^2 \approx \frac{1}{p_2} \sum_{i \in \mathcal{G}_2} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := h_2(\tau_1^2, \tau_2^2),$$

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

EB for group-regularized ridge[‡]

Aim: $\hat{\tau}_g^2$ for group-regularized ridge: $\beta_i \sim N(0, \tau_g^2), i \in \mathcal{G}_g$

Initial: $\hat{\beta}_i = \hat{\beta}_i^{\lambda_0}$. Moment equations $g = 1, \ldots, G$. $G = 2$:

$$\frac{1}{p_1} \sum_{i \in \mathcal{G}_1} \hat{\beta}_i^2 \approx \frac{1}{p_1} \sum_{i \in \mathcal{G}_1} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := h_1(\tau_1^2, \tau_2^2)$$

$$\frac{1}{p_2} \sum_{i \in \mathcal{G}_2} \hat{\beta}_i^2 \approx \frac{1}{p_2} \sum_{i \in \mathcal{G}_2} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_i^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := h_2(\tau_1^2, \tau_2^2),$$

In general: System of G linear equations $\mathbf{b}_{\text{data}} = A\mathbf{t}$

Solution: $t = (\hat{\tau}_1^2, \ldots, \hat{\tau}_g^2)$.

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

Extension: Stability [§]

- Some co-data render *many* groups: e.g. pathways
- G large: system $\mathbf{b}_{\text{data}} = A\mathbf{t}$ becomes unstable
- Need to stabilize solution

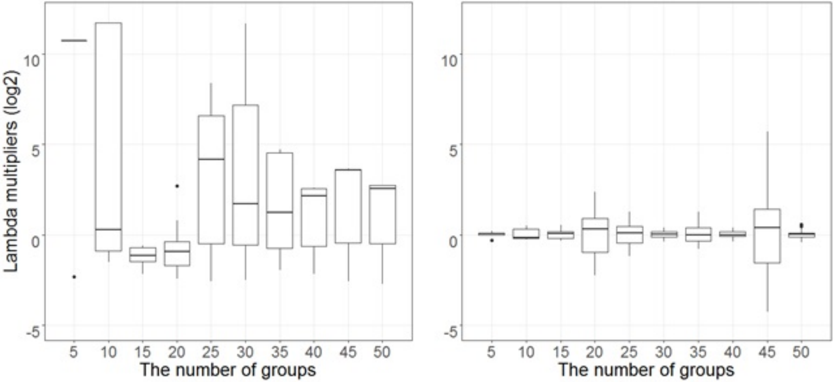
Solutions

1. Enforce monotony when grouping based on continuous co-data (e.g. external p-values)
2. Shrink A to a stable target T: $\tilde{A}_q = qA + (1-q)T$.

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

Effect of shrinkage of *A*

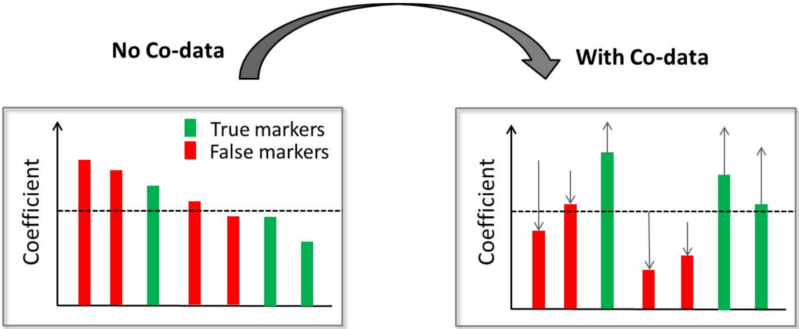
Real data, *random* groups of variables



Left: No Shrinkage; Right: Shrinkage

Variable selection

Why can co-data help?



Variable selection

Current solution:

1. Estimate group penalties from ridge regression, possibly for multiple groupings

2. Select k variables by introducing non-grouped L_1 penalty (i.e. thresholding)

3. Refit the model using the selected variables and their respective L_2 penalties

Variable selection

Current solution:

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

“Bet on sparsity”: yes, but *after* penalty weighting

Software[¶]

R-package `GRridge`, Github + Bioconductor:

- Logistic, linear and survival
- Auxiliary functions for co-data processing (from TCGA etc.)
- Allows unpenalized covariates
- Built-in CV for comparison with ridge & lasso

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

Software¶

R-package `GRridge`, Github + Bioconductor:

- Logistic, linear and survival
- Auxiliary functions for co-data processing (from TCGA etc.)
- Allows unpenalized covariates
- Built-in CV for comparison with ridge & lasso

Comparison (one grouping only): Sparse group lasso, `SGL` (Simon et al., *J Comp Graph Stat*, 2013).

¶Details: Novianti et al., *Bioinformatics*, 2017

Example: Diagnostics for cervical cancer

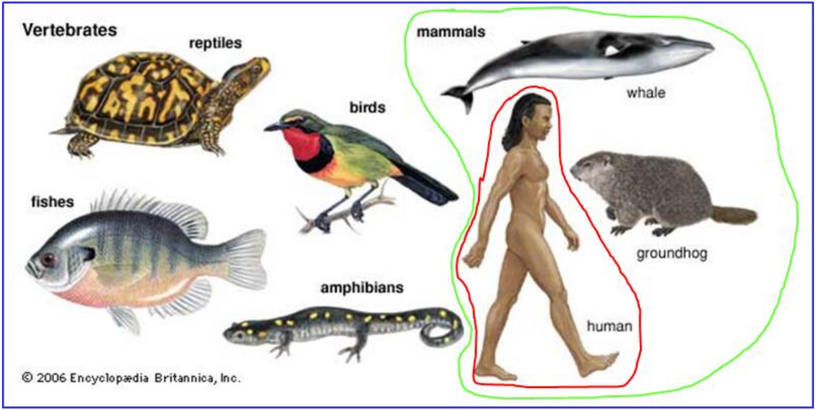
Goal: Select markers for classifying Normal vs CIN3
$\rightarrow$ final goal is a cheap PCR assay

Data:

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

Co-data 1: Conservation status

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



Co-data 2: Standard deviation

- Current practice: standardize variable j by sd: s_j
  $\rightarrow$ effective penalty λs_j^2 (Zwiener et al, 2014)
  $\rightarrow$ too large advantage for small s_j's?

Co-data 2: Standard deviation

- Current practice: standardize variable j by sd: s_j
  $\rightarrow$ effective penalty λs_j^2 (Zwiener et al, 2014)
  $\rightarrow$ too large advantage for small s_j's?

- Our solution:

  1. Standardize by s_j
  2. G groups of variables with decreasing s_j
  3. Effective penalty $j \in \mathcal{G}_g$: $\lambda_j = \hat{\tau}_g^{-2} \lambda s_j^2$

Co-data 2: Standard deviation

- Current practice: standardize variable j by sd: s_j
  $\rightarrow$ effective penalty λs_j^2 (Zwiener et al, 2014)
  $\rightarrow$ too large advantage for small s_j's?

- Our solution:
  1. Standardize by s_j
  2. G groups of variables with decreasing s_j
  3. Effective penalty $j \in \mathcal{G}_g$: $\lambda_j = \hat{\tau}_g^{-2} \lambda s_j^2$

- Allows a more non-parametric link between s_j and λ_j

Co-data results

For $j \in \mathcal{G}_g$, penalty factor: $\lambda'_g \propto \tau_g^{-2}$

Co-data results

For $j \in \mathcal{G}_g$, penalty factor: $\lambda'_g \propto \tau_g^{-2}$

Conservation status:

1. Non-conserved (552): $\lambda'_1 = 1.84$
2. Conserved across mammals (72): $\lambda'_2 = 0.61$
3. Broadly conserved across vertebrates (148): $\lambda'_3 = 0.30$

Co-data results

For $j \in \mathcal{G}_g$, penalty factor: $\lambda'_g \propto \tau_g^{-2}$

Conservation status:

1. Non-conserved (552): $\lambda'_1 = 1.84$
2. Conserved across mammals (72): $\lambda'_2 = 0.61$
3. Broadly conserved across vertebrates (148): $\lambda'_3 = 0.30$

Standard deviation:
Range from $\lambda'_1 = 0.56$ (large s.d.) to $\lambda'_{10} = 1.80$ (small s.d.)

$\rightarrow$ Indeed, partly 'undoes' the effect of standardization (for $j \in \mathcal{G}_g$: $\lambda_j \propto \lambda'_g s_j^2$).

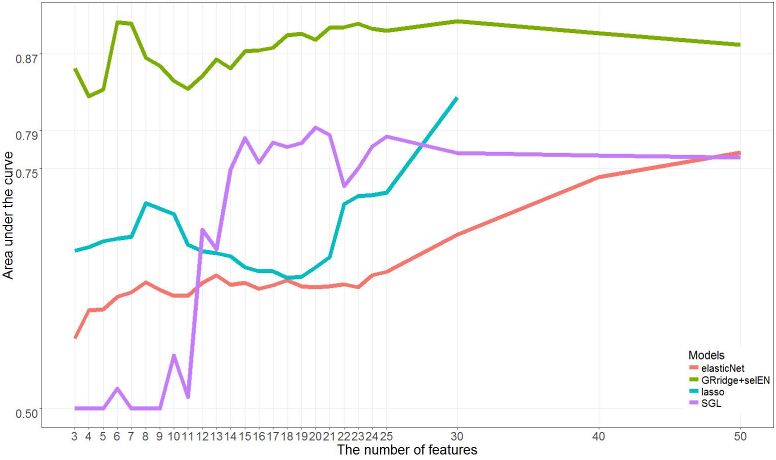
Clinician:

“That’s all nice, but does the predictive accuracy improve?”

“Do I get the good biomarkers?”

Performance under variable selection

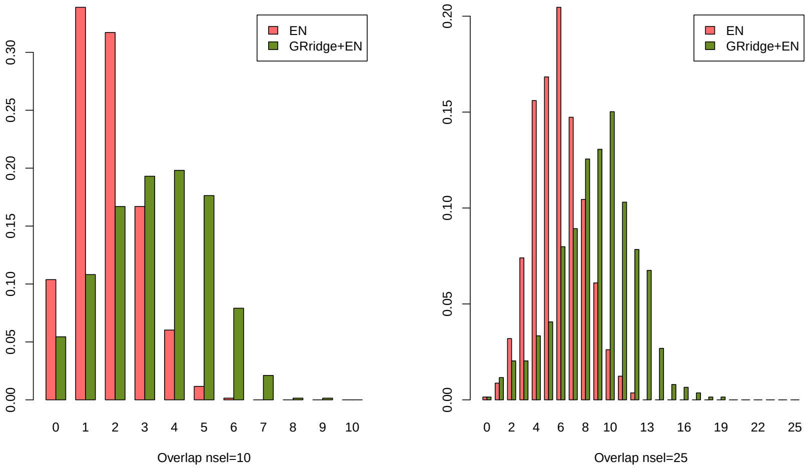
AUC assessed by LOOCV



GRridge + EN, Sparse group-lasso, Lasso, Elastic Net

Stability of selection

50 re-sampled versions of data set. Overlap in selected variables between pairs of re-samples



Other applications, extensions

Hybrid Bayes - Empirical Bayes

- $\lambda_g = \lambda'_g \lambda$. λ'_g: EB; Common λ: Full Bayes (prior)

[‖] *Ann Appl Stat*, 2016; *Biom J*, 2017

[**] arXiv, 2017

Other applications, extensions

Hybrid Bayes - Empirical Bayes

- $\lambda_g = \lambda'_g \lambda$. λ'_g: EB; Common λ: Full Bayes (prior)

Networks (Gwenaël Leday, Gino Kpogbezan et al.[‖])

- Bayesian SEM: Variational Bayes + EB + prior network

[‖] *Ann Appl Stat*, 2016; *Biom J*, 2017

[**] arXiv, 2017

Other applications, extensions

Hybrid Bayes - Empirical Bayes

- $\lambda_g = \lambda'_g \lambda$. λ'_g: EB; Common λ: Full Bayes (prior)

Networks (Gwenaël Leday, Gino Kpogbezan et al.[‖])

- Bayesian SEM: Variational Bayes + EB + prior network

Random Forest (Dennis te Beest[**])

- Co-data moderated Random Forest

[‖] *Ann Appl Stat*, 2016; *Biom J*, 2017
[**] arXiv, 2017

Take home

Empirical Bayes...

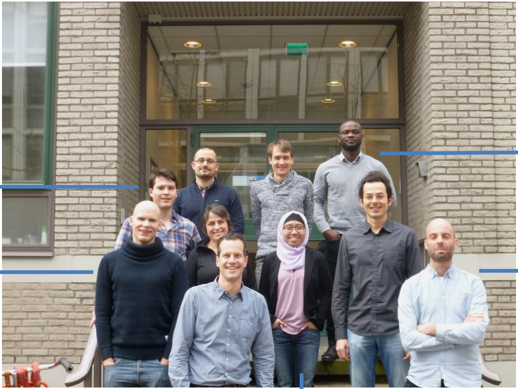
... is a versatile technique to learn

1. **from a lot...(many variables)**
2. **...and a lot more (co-data)**

Acknowledgements

Magnus Münch
(Bayes EN)

Dennis te Beest
(RF)



Gino Kpogbezan
(Networks)

Wessel "Ridge"
van Wieringen

Putri Novianti (GRridge)

+ miRNAseq data: Barbara Snoek, Saskia Wilting, Renske Steenbergen (VUmc)

+ Grant: NWO ZONMW-TOP

VUmc

Details

Method: Van de Wiel MA, Lien TG, Verlaat W, Van Wieringen WN, Wilting SM (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Stat Med.*, **35**, 368-381.

Software: Novianti PW, Snoek B, Wilting SM, van de Wiel MA (2017). Better diagnostic signatures from RNAseq data through use of auxiliary co-data. *Bioinformatics*, **33**, 1572-1574.

QUESTIONS?[††]

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