How to learn from a lot: Empirical Bayes in high-dimensional prediction settings

Mark van de Wiel[1,2]

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

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

Contributions by: **Putri Novianti** (VUmc), **Magnus Münch** (Leiden/VUmc), **Dennis te Beest** (VUmc)

Our group: `www.bigstatistics.nl`

Overview

1. Motivating example (ridge regression)
2. Introduction Empirical Bayes (EB)

Overview

1. Motivating example (ridge regression)
2. Introduction Empirical Bayes (EB)

3. EB methods, hard: Maximization marginal likelihood
4. Intermezzo: Co-data
5. EB methods, easier: Method of Moments
   - Group-regularized ridge and elastic net
   - Example: Cervical cancer diagnostics
6. EB methods, easy
   - Random forest
   - Example: predicting metastasis for oral cancer

Overview

1. Motivating example (ridge regression)
2. Introduction Empirical Bayes (EB)

3. EB methods, hard: Maximization marginal likelihood
4. Intermezzo: Co-data
5. EB methods, easier: Method of Moments
   - Group-regularized ridge and elastic net
   - Example: Cervical cancer diagnostics
6. EB methods, easy
   - Random forest
   - Example: predicting metastasis for oral cancer

7. Discussion (Full Bayes, Cross-validation)

Motivating example, Simulated

- Suppose $p = 50$ covariates
- $1, \ldots, 25$ associated with response $\mathbf{Y}$; $26, \ldots, 50$ not
- Sample size $n = 40$

Motivating example, Simulated

- Suppose $p = 50$ covariates
- $1, \ldots, 25$ associated with response $\mathbf{Y}$; $26, \ldots, 50$ not
- Sample size $n = 40$
- Ordinary ridge regression:

$$\text{argmax}_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) - \lambda \sum_{i=1}^{50} \beta_j^2$$

- Equivalent to $\beta_j \sim N(0, \sigma^2), j = 1, \ldots, 50$

Coefficients

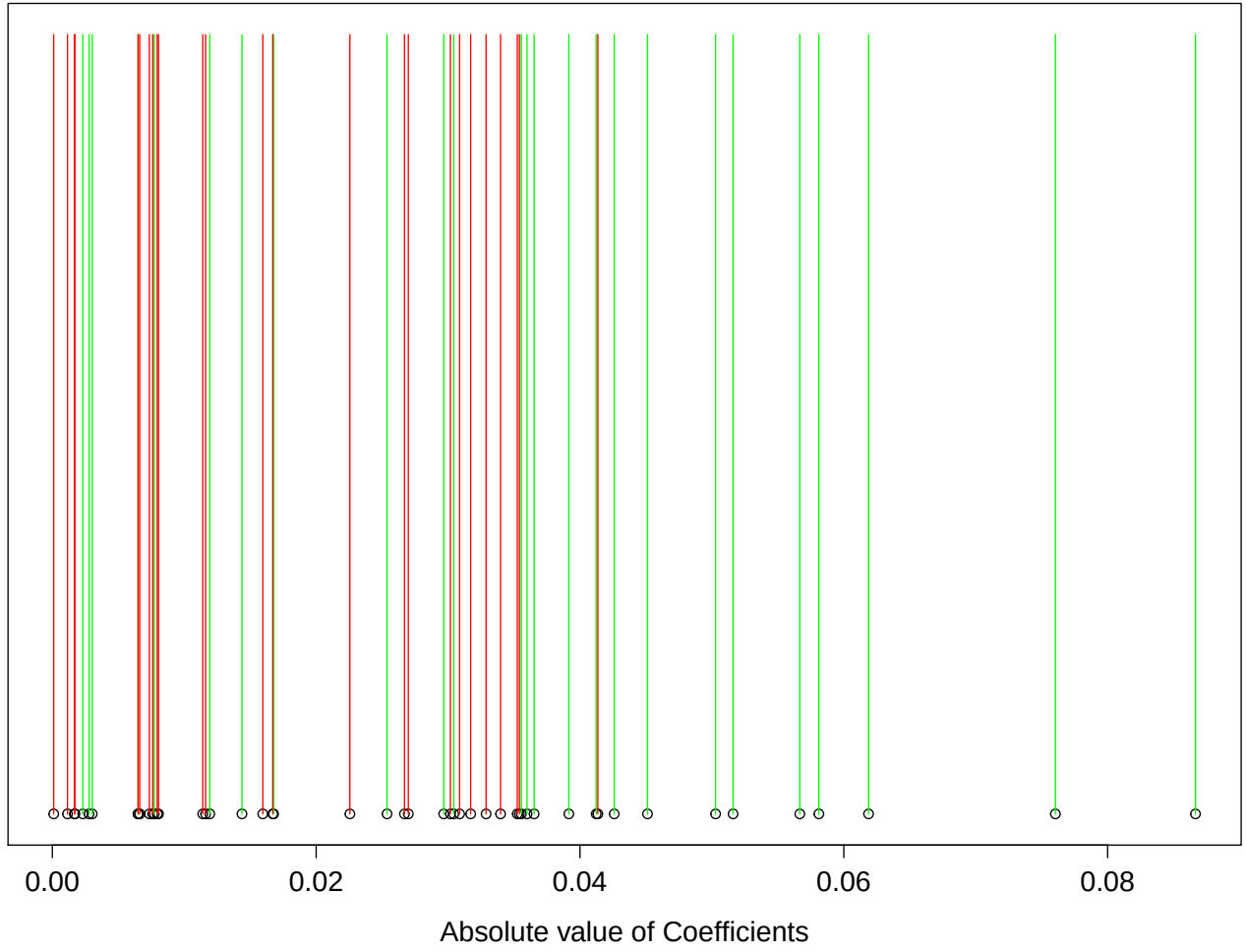


Figure: Ridge regression coefficients, Group 1, Group2

Sums of squares, Coefficients

```
> mean(coefs[1:25]^2)
[1] 0.001723317

> mean(coefs[-(1:25)]^2)
[1] 0.0004957746
```

Sums of squares, Coefficients

```
> mean(coefs[1:25]^2)
[1] 0.001723317

> mean(coefs[-(1:25)]^2)
[1] 0.0004957746
```

Better priors (ad hoc): $\beta_j \sim N(0, \sigma_1^2), j \in$ group 1,
$\beta_j \sim N(0, \sigma_2^2), j \in$ group 2 with $\sigma_1^2 = \mathbf{3}\sigma_2^2$.

Equivalently, $\lambda_1 = \frac{\mathbf{1}}{\mathbf{3}}\lambda_2$

Refitting reduces CV-MSE by 10-20%; Rank correlation prediction with response increases by 10-40%.

Prelude to variable selection

10 strongest covariates [Should be all from group 1]:

```
> top10_ridge
 [1] 1 1 1 1 2 1 1 1 2 2
> top10_groupridge
 [1] 1 1 1 1 1 1 1 1 1 1
```

To be continued...

Setting

- **Prediction or Diagnosis**
- **Main study**
  - 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

Empirical Bayes (EB)

- Regularization by informative prior (ridge: $\beta_i \sim N(0, \sigma^2)$)
- Empirical Bayes: estimate prior parameters from data
- EB also applicable in frequentist settings. Example: Logistic ridge, $\lambda = 1/(2\sigma^2)$:

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

- References
  - Books: Carlin & Louis, 2000; Efron, 2010
  - Review: Van Houwelingen, *Biom J*, 2014

Hard EB: Maximum marginal Likelihood

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

Marginal likelihood maximization:

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

Hard EB: Maximum marginal Likelihood

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

Marginal likelihood maximization:

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

Requires a likelihood. Optimization is hard, because

1. High-dimensional integral
2. Competitive prior parameters

Problem 1: High-dimensional integral

Solutions:

- Laplace approximation; may work well for sparse settings (Shun & McCullagh, *JRSSB*, 1995)
- EM on Gibbs samples (Casella, *Biostatistics*, 2001). Conceptually easy, computationally (often) terrible.
- EM on Variational Bayes approximation (Bernardo et al., *Bayesian analysis*, 2003). Fast, but requires dedicated approximations.

Problem 2: competitive prior parameters

- Elastic net:

$$\text{argmax}_{\boldsymbol{\beta}} \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) - \lambda_1 ||\boldsymbol{\beta}||_1 - \lambda_2 ||\boldsymbol{\beta}||_2,$$

- Equivalent Bayesian formulation, prior for β_j:

$$\pi(\beta_j) \propto \pi_\lambda(\beta_j) \propto \text{exp}[-\lambda_1 |\beta_j| - \lambda_2 \beta_j^2\}],$$

- λ_1 and λ_2 are competitive, also for CV (Waldron et al., 2011, *Bioinf.*)

- Small simulation study, linear model: $p = 200, n = 100, (\lambda_1, \lambda_2) = (2, 2)$

Problem 2: competitive prior parameters

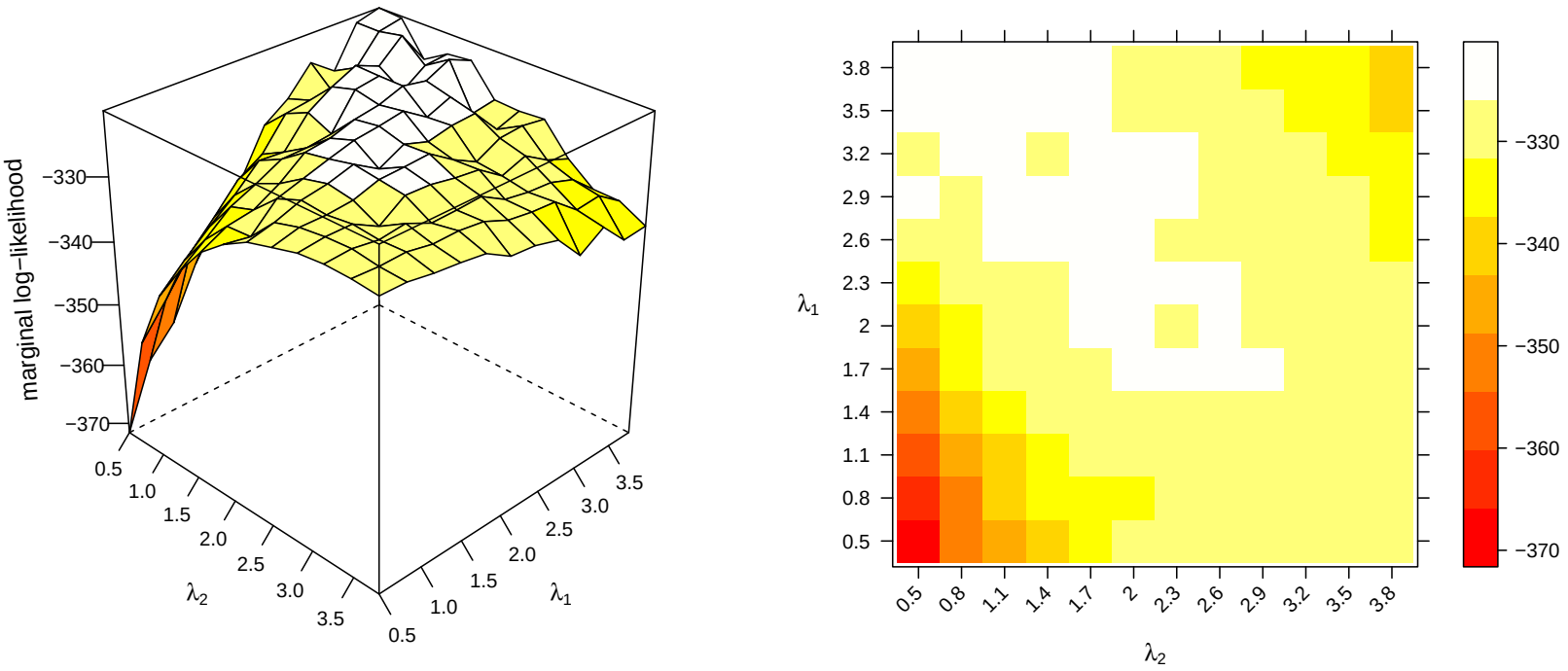
Simulation, linear model: $p = 200, n = 100, (\lambda_1, \lambda_2) = (2, 2)$

Bayesian elastic net: Li & Nin, *Bayesian Analysis*, 2010
Marginal likelihood from Gibbs: Chib, *JASA*, 1995

Problem 2: competitive prior parameters

Simulation, linear model: $p = 200, n = 100, (\lambda_1, \lambda_2) = (2, 2)$

Bayesian elastic net: Li & Nin, *Bayesian Analysis*, 2010
Marginal likelihood from Gibbs: Chib, *JASA*, 1995



Marginal likelihood as a function of λ_1 and λ_2

Intermezzo: Prior info from co-data

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

Examples of co-data

1. Published gene signature. Two groups of variables
2. Chromosome. Results in 24 groups
3. *p*-values from external study

Intermezzo: Prior info from co-data

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

Examples of co-data

1. Published gene signature. Two groups of variables
2. Chromosome. Results in 24 groups
3. *p*-values from external study

Idea: Use different tuning parameters $\lambda_1, \ldots, \lambda_G$ across G co-data-based groups. E.g. in ridge:

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

EB, (somewhat) easier: Moment estimation*

Motivating example: estimate σ_1^2, σ_2^2 for (group) ridge:

$\beta_j \sim N(0, \sigma_1^2), j \in$ group 1, $\beta_j \sim N(0, \sigma_2^2), j \in$ group 2

Idea: equate empirical moment(s) to theoretical ones

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

EB, (somewhat) easier: Moment estimation*

Motivating example: estimate σ_1^2, σ_2^2 for (group) ridge:

$\beta_j \sim N(0, \sigma_1^2), j \in$ group 1, $\beta_j \sim N(0, \sigma_2^2), j \in$ group 2

Idea: equate empirical moment(s) to theoretical ones

$$\frac{1}{p_1} \sum_{j \in \text{group 1}} \hat{\beta}_j^2 \approx \frac{1}{p_1} \sum_{j \in \text{group 1}} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_j^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := g_1(\sigma_1, \sigma_2)$$

$$\frac{1}{p_2} \sum_{j \in \text{group 2}} \hat{\beta}_j^2 \approx \frac{1}{p_2} \sum_{j \in \text{group 2}} E_{\boldsymbol{\beta}} \left[E[\hat{\beta}_j^2(\mathbf{Y}) | \boldsymbol{\beta}] \right] := g_2(\sigma_1, \sigma_2),$$

where $E_{\boldsymbol{\beta}}$ denoted expectation w.r.t. the prior(s) of $\boldsymbol{\beta}$.

---

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

EB: Moment estimation

$$\frac{1}{p_1}\sum_{j\in\text{group 1}}\hat{\beta}_j^2 \approx \frac{1}{p_1}\sum_{j\in\text{group 1}} E_{\boldsymbol{\beta}}\left[E[\hat{\beta}_j^2(\mathbf{Y})|\boldsymbol{\beta}]\right] := g_1(\sigma_1,\sigma_2)$$

- $E[\hat{\beta}_j^2(\mathbf{Y})|\boldsymbol{\beta}] = V[\hat{\beta}_j(\mathbf{Y})] + E[\hat{\beta}_j(\mathbf{Y})|\boldsymbol{\beta}]^2 = v_j + e_j^2$.
- v_j: known and constant in β_j.
- $e_j = \sum_k c_{jk}\beta_k$, c_{jk} known[†]. Penalty causes bias!

[†]see Le Cessie & Van Houwelingen, *Appl Statist*, 1992

EB: Moment estimation

$$\frac{1}{p_1}\sum_{j\in\text{group 1}}\hat{\beta}_j^2 \approx \frac{1}{p_1}\sum_{j\in\text{group 1}} E_{\boldsymbol{\beta}}\left[E[\hat{\beta}_j^2(\mathbf{Y})|\boldsymbol{\beta}]\right] := g_1(\sigma_1,\sigma_2)$$

- $E[\hat{\beta}_j^2(\mathbf{Y})|\boldsymbol{\beta}] = V[\hat{\beta}_j(\mathbf{Y})] + E[\hat{\beta}_j(\mathbf{Y})|\boldsymbol{\beta}]^2 = v_j + e_j^2$.
- v_j: known and constant in β_j.
- $e_j = \sum_k c_{jk}\beta_k$, c_{jk} known[†]. Penalty causes bias!
- For $E_{\boldsymbol{\beta}}[e_j^2]$: $E_{\boldsymbol{\beta}}[\beta_j\beta_k] = 0$, $E_{\boldsymbol{\beta}}[\beta_j^2] = \sigma_1^2$ and $E_{\boldsymbol{\beta}}[\beta_j^2] = \sigma_2^2$ $\Longrightarrow$ linear equation in (σ_1^2, σ_2^2)

[†]see Le Cessie & Van Houwelingen, *Appl Statist*, 1992

Suppose we want variable selection...

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

[‡] work in progress with Magnus Münch

Suppose we want variable selection...

Nicest solution: A coherent framework for EB estimation in a group 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

[‡] work in progress with Magnus Münch

Example: Diagnostics for cervical cancer

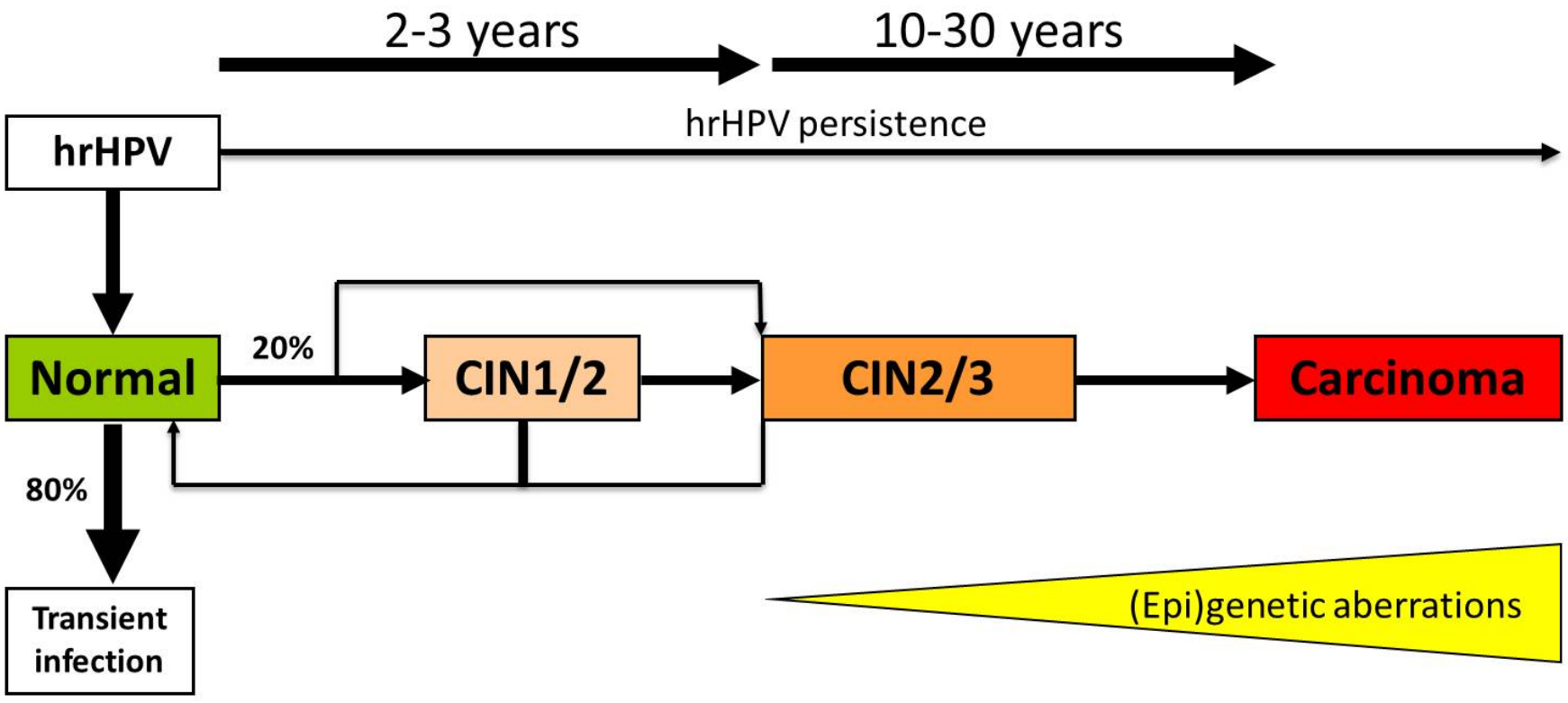
Current tests: Based on HPV (sometimes) i.c.w. cytology
$\Longrightarrow$ accurate, but requiring high standards of cytological training

Additional problem: Some women do not show up for screening

Molecular tests: Easy to implement, objective and potentially cost-effective
$+$ can be applied to self samples.

Challenging: Because self samples are of lower quality

Cervical carcinogenesis



Goal: Detect CIN3 lesions, to be removed surgically

Example: Diagnostics for cervical cancer

Example: Diagnostics for cervical cancer[¶]

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

Data:

- miRNA sequencing data
- $n = 56$: 32 Normal, 24 CIN3
- $p = 772$ (after filtering lowly abundant ones).
- Sqrt-transformed to quasi-Gaussian scale
- Standardized for penalty to have the same effect[§].

[§] Discussion on standardization: Van de Wiel et al., *Stat Med*, 2016

[¶] by Putri Novianti

Example: Diagnostics for cervical cancer

Co-data

- Conservation status:
  1. Non-conserved (552)
  2. Conserved across mammals (72)
  3. Broadly conserved, across most vertebrates (148)

- Standard deviation per variable
  - 10 groups of variable with decreasing s.d.
  - Allows natural variability to impact the classifier via penalty weights

Co-data results

$\lambda_g \propto \sigma_g^{-2}$; Penalty multipliers λ'_g: $\lambda_g = \lambda'_g \lambda, g = 1, \ldots, G$

Co-data results

$\lambda_g \propto \sigma_g^{-2}$; Penalty multipliers λ'_g: $\lambda_g = \lambda'_g \lambda, g = 1, \ldots, G$

Conservation status:

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

Co-data results

$\lambda_g \propto \sigma_g^{-2}$; Penalty multipliers λ'_g: $\lambda_g = \lambda'_g \lambda, g = 1, \ldots, G$

Conservation status:

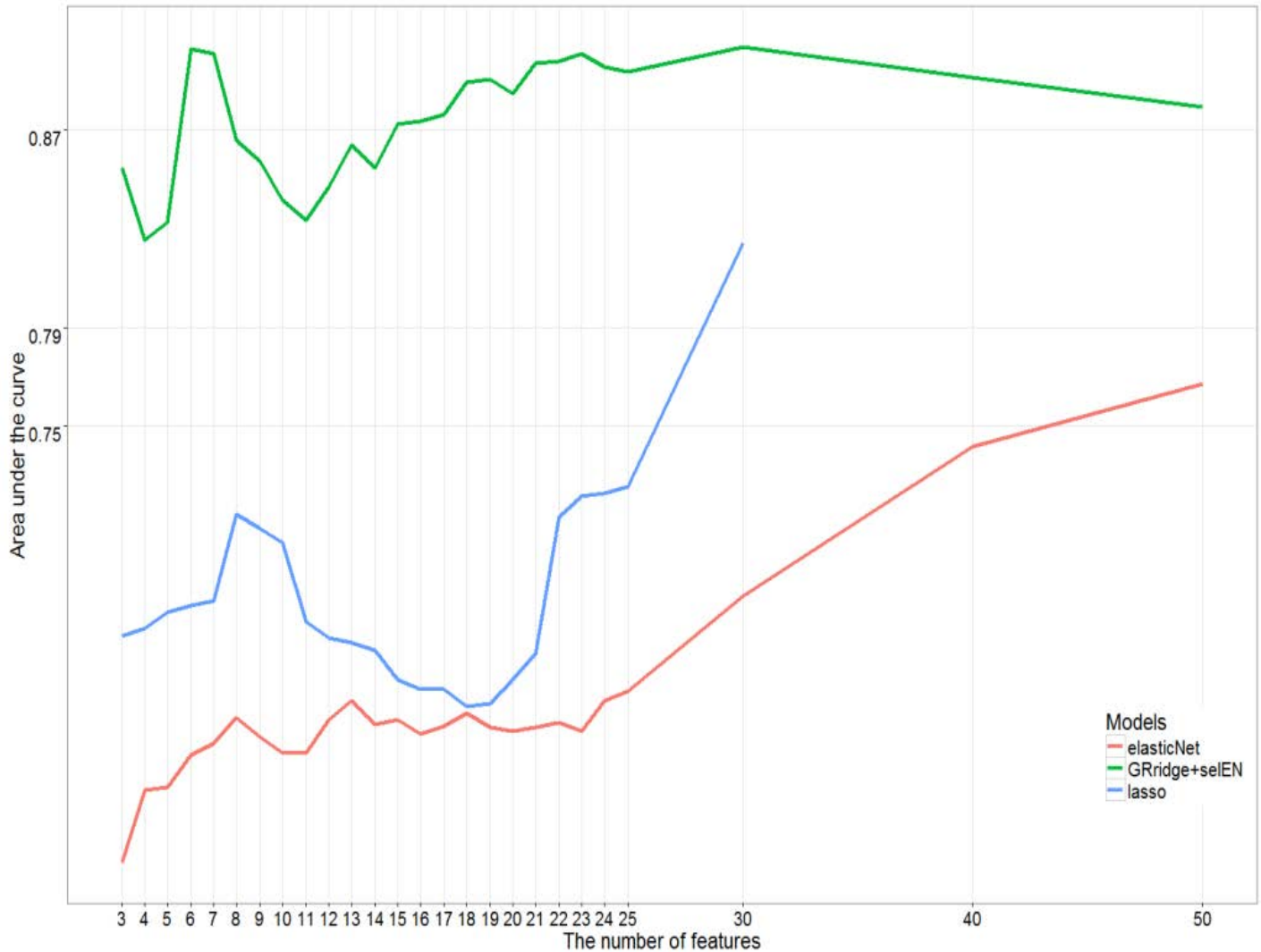
1. Non-conserved (552): $\lambda'_1 = 1.84$
2. Conserved across mammals (72): $\lambda'_1 = 0.61$
3. Broadly conserved across vertebrates (148): $\lambda'_3 = 0.30$

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

$\Longrightarrow$ Indeed, partly 'undoes' the effect of standardization.

Variable selection: Data example

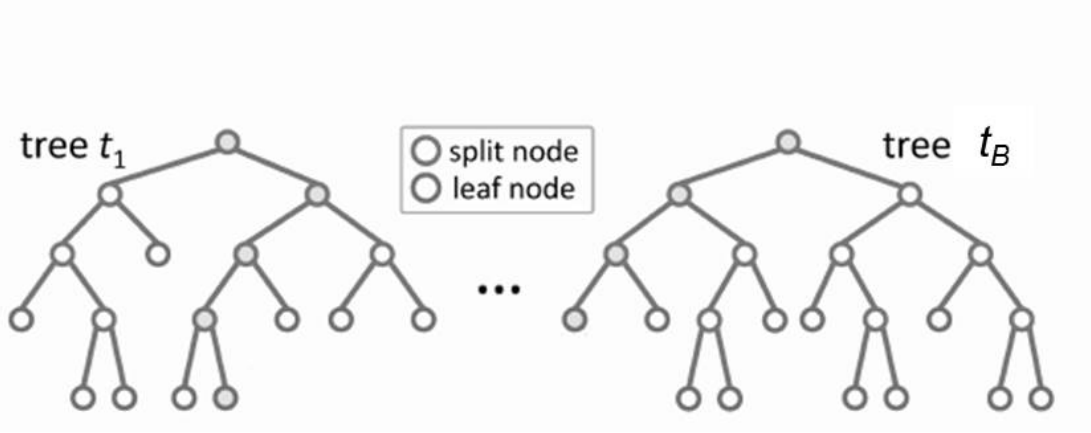
AUC assessed by LOOCV



GRridge + EN selection,Lasso,Elastic Net

EB, easy: Random Forest

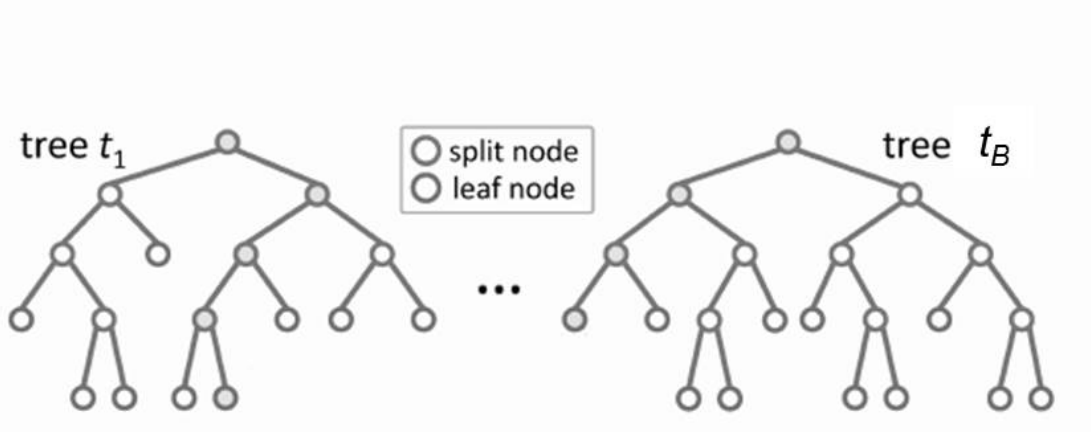
Random Forest Classifier



- 'Regularization' by **Uniform** sampling of $m_{\text{try}} = \sqrt{p}$ candidate variables per node split

EB, easy: Random Forest

Random Forest Classifier



- 'Regularization' by **Uniform** sampling of $m_{\text{try}} = \sqrt{p}$ candidate variables per node split
- **Idea**: Replace uniform 'prior' by one informed by co-data
- No likelihood: informal Empirical Bayes

Co-RF: Algorithm

1. Fit ordinary Random Forest (RF)
2. Calculate for each variable i how often selected: v_i
3. Determine S potentially relevant co-data sources, $c_{is}, i = 1, \ldots, p, s = 1, \ldots, S$

Co-RF: Algorithm

1. Fit ordinary Random Forest (RF)
2. Calculate for each variable i how often selected: v_i
3. Determine S potentially relevant co-data sources, $c_{is}, i = 1, \ldots, p, s = 1, \ldots, S$
4. **Robustly** regress v_i on co-data info C_i
5. Regression renders fitted selection frequency: f_i
6. Truncate f_i: $f_i' = (f_i - \gamma E[f_i^{\text{uni}}])_+$
7. Run new RF, with prior $p_i^{\text{new}} \propto f_i'$ per node split

Co-RF: the regression

- Variables are the 'samples'. Only interested in mean approximation:

$$v_i \approx g_{\boldsymbol{\alpha}}(C_i)$$

- Regression: parsimonious to avoid overfitting!
- Nominal co-data: cluster small groups of variables
- Continuous co-data:
  - Parameterize (e.g. $\alpha \log(p_i)$, p_i: external p-value)
  - Or (monotone), penalized spline

Example[‖]: Oral cancer

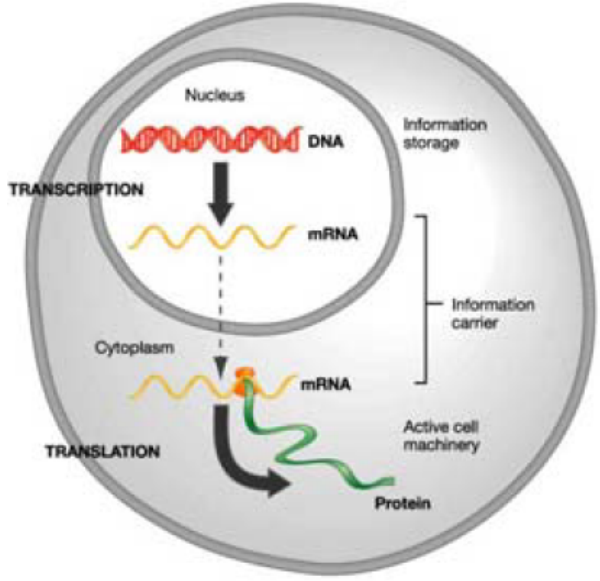
Setting

- TCGA data, oral cancer, $n = 262, p = 16.012$
- Response: Lymph node metastasis (Yes/No)
- Main data: normalized mRNA expression, RNAseq
- Co-data: Kendall correlation with matched DNA copy number data (gene-gene)

[‖] by Dennis te Beest

Why DNA as co-data?

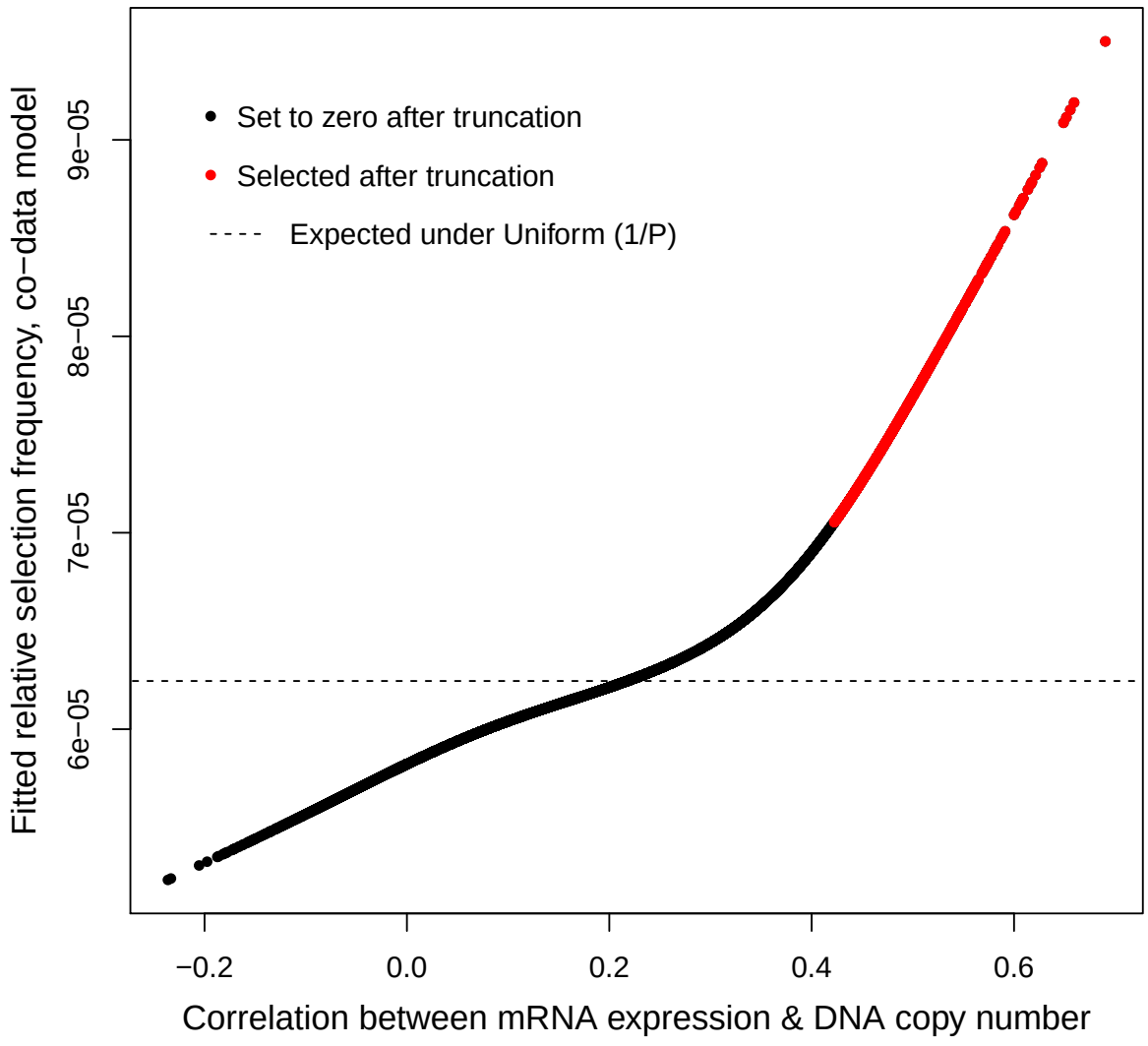
1. DNA copy number in tumor affects mRNA expression



2. DNA is more stable than mRNA

3. Co-data: DNA not required for future samples (as it would be for integrated classifiers)

Regression on co-data: monotone spline



Classification results

- Accuracy assessed by 10-fold CV
- Number of misclassifications drops from 112 (43%) to 88 (34%)
- PPV increases from 59% to 66%
- NPV increases from 53% to 67%

Software, handling co-data

- Group-regularized ridge: R-package `GRridge`, Github
  - **Multiple** sources of co-data, as groups
  - Elastic net-type variable selection
- Co-data Random Forest: `CoRF`. Under development.
  - Handles nominal, ordinal and continuous co-data
  - Computationally very efficient
- Alternatives: Group-lasso +: `grpreg` (Breheny, CRAN), Sparse version: `SGL` (Simon et al., CRAN).
  - Based on group penalties
  - **One** source of co-data represented as groups.

Discussion: CV versus EB

	Cross-Validation	Empirical Bayes
Tuned to Prediction	++	+
Easy to Implement	++	-/+/++
Multiple Penalties	-	++
Bayesian Models	-	+

Discussion: CV versus EB

	Cross-Validation	Empirical Bayes
Tuned to Prediction	++	+
Easy to Implement	++	-/+/++
Multiple Penalties	-	++
Bayesian Models	-	+

Hybrid methods:
a) CV for 'master-penalty' λ, EB for multipliers λ'_g, $\lambda_g = \lambda\lambda'_g$
b) CV-parameter tunes EB weights.

Discussion: CV versus EB

	Cross-Validation	Empirical Bayes	Hybrid Methods
Tuned to Prediction	++	+	++
Easy to Implement	++	-/+/++	-/+/++
Multiple Penalties	-	++	++
Bayesian Models	-	+	-/+

Discussion: Full Bayes versus EB

	Full Bayes	Empirical Bayes
Error Propagation	++	+/-
Coverage, Intervals	+	+
Computational	-	+

Discussion: Full Bayes versus EB

	Full Bayes	Empirical Bayes
Error Propagation	++	+/-
Coverage, Intervals	+	+
Computational	-	+

Hybrid method: FB for 'master-parameter', EB for multipliers:

Logistic group-ridge: $\beta_i \sim N(0, \tau_g^2)$,

$$\tau_g^{-2} = \tau^{-2}\tau_g'$$
$$\tau^{-2} \sim G(\alpha_1, \alpha_2),$$

Discussion: Full Bayes versus EB

	Full Bayes	Empirical Bayes	Hybrid Methods
Error Propagation	++	+/-	+
Coverage, Intervals	+	+	++
Computational	-	+	+

Main message

Empirical Bayes (EB) allows one to learn

1. **from a lot...(many variables)**
   Many flavors of EB in prediction, from hard to easy

Main message

Empirical Bayes (EB) allows one to learn

1. **from a lot...(many variables)**
   Many flavors of EB in prediction, from hard to easy

2. **...and a lot more (prior information)**
   EB particularly useful for differential regularization

QUESTIONS?**††

**These slides are available via www.bigstatistics.nl

††Review available on request