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

Mark van de Wiel\textsuperscript{1,2,3}, Magnus Münch\textsuperscript{1,4}

\textsuperscript{1}Dep of Epidemiology and Biostatistics, VU University medical center
\textsuperscript{2}Dep of Mathematics, VU University, Amsterdam, NL
\textsuperscript{3}MRC Biostatistics unit, Cambridge University, UK (Visiting Fellow)
\textsuperscript{4}Mathematical Institute, Leiden University, NL

Our group: www.bigstatistics.nl
**Setting**

- **Prediction or Diagnosis**

- **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 $X_j = (X_{1j}, \ldots, X_{pj})$
  - Goal: find $f$ such that $Y_j \approx f(X_j)$
  - 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
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: $\arg\max_\beta \{ \mathcal{L}(Y; \beta) - \sum_{g=1}^{G} \lambda_g ||\beta_g||_2 \}$

$\rightarrow$ 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 ↔ hyper-parameters (prior)

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

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

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

Formal EB: Maximum marginal Likelihood

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

Marginal likelihood maximization:

$\hat{\alpha} = \arg\max_\alpha ML(\alpha)$, with $ML(\alpha) = \int_\beta L(Y; \beta) \pi_\alpha(\beta) d\beta$, 

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
Formal EB: Maximum marginal Likelihood

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

Marginal likelihood maximization:

$$\hat{\alpha} = \arg\max_{\alpha} ML(\alpha), \text{ with } ML(\alpha) = \int_{\beta} L(Y; \beta) \pi_\alpha(\beta) d\beta,$$

Optimization hard, because of the high-dimensional integral

- 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 $\sigma_g^2 (\lambda_g \propto \sigma_g^{-2})$, for ridge: $\beta_i \sim N(0, \sigma_g^2)$, $i \in \text{group } g$
EB using moments: group-regularized ridge

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

Intuitive Idea:

1. Run an initial ridge regression with one $\lambda$
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 $\sigma_g^2$ should be large (hence $\lambda_g$ small)
EB using moments: group-regularized ridge

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

Intuitive Idea:

1. Run an initial ridge regression with one $\lambda$
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 $\sigma_g^2$ should be large (hence $\lambda_g$ 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 $\sigma_1^2$, $\sigma_2^2$ ($\lambda_g \propto \sigma_g^{-2}$), for ridge:

$\beta_i \sim N(0, \sigma_i^2)$, $i \in$ group 1, $\beta_i \sim N(0, \sigma_i^2)$, $i \in$ 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 $\sigma_1^2, \sigma_2^2$ ($\lambda_g \propto \sigma_g^{-2}$), for ridge:

$\beta_i \sim N(0, \sigma_i^2), i \in \text{group } 1, \beta_i \sim N(0, \sigma_i^2), i \in \text{group } 2$

Idea: equate empirical moment(s) to theoretical ones

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

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

Result: System of equations $b_{\text{data}} = Ax$, $\lambda_g^{-1} \propto \hat{\sigma}_g^2 = x_g$.

†Details: Van de Wiel et al., Stat Med, 2016
Shrink the shrinkage parameters‡

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

\[ \hat{\sigma}^2 = A^{-1}b_{\text{data}} \text{ instable} \rightarrow \text{over-fitting.} \]

‡Details: Novianti et al., Bioinformatics, 2017
Shrink the shrinkage parameters\(^\dagger\)

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

\[ \sigma^2 = A^{-1}b_{data} \] instable \( \rightarrow \) over-fitting.

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

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

\(^\dagger\)Details: Novianti et al., *Bioinformatics*, 2017
Effect of shrinkage

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

$\lambda'_g$: 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
R-package GRridge, Github + Bioconductor

To be discussed during course
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 $j$: $g_j$.

$g_j = 1$

$g_j = 2$

\vdots

$g_j = G$
Group-regularized elastic net

Model

\[ Y_i | \beta \sim \text{Bern}(\expit(X_i^T \beta)), \]

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

- Shrinks estimates towards zero
- ‘Global’ \( \alpha \) and \( \lambda \) 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 $\alpha$ and estimate $\lambda$ by CV for global shrinkage
  • Empirical Bayes estimation of $\lambda'$ by MML

Maximum marginal likelihood (MML)

$$\hat{\lambda}' = \arg\max_{\lambda'} \int_{\beta} L(Y; \beta) \pi_{\lambda'}(\beta) d\beta$$
Latent variables

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

- $\omega \mid \beta \sim \prod_{i=1}^{n} \mathcal{P}G(1, |X_i^T\beta|)$, independent of $Y_i$
- $\beta \mid \tau \sim \prod_{j=1}^{p} \mathcal{N}(0, \frac{\tau_j^{-1}}{\lambda'_{g(j)}(1-\alpha)\lambda\tau_j})$ and
  $\tau \sim \prod_{j=1}^{p} \mathcal{T}G\left(\frac{1}{2}, \frac{8(1-\alpha)}{\alpha^2\lambda}, (1, \infty)\right)$
Latent variables

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

- $\omega \mid \beta \sim \prod_{i=1}^{n} PG(1, |X_i^T \beta|)$, independent of $Y_i$
- $\beta \mid \tau \sim \prod_{j=1}^{p} N \left(0, \frac{\tau_j^{-1}}{\lambda_{g(j)}(1-\alpha)\lambda_{\tau_j}} \right)$ and
  $\tau \sim \prod_{j=1}^{p} TG \left(\frac{1}{2}, \frac{8(1-\alpha)}{\alpha^2\lambda}, (1, \infty) \right)$

**Computational reasons**

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

Recap Casella (2001):

\[ \lambda'^{(k+1)} = \arg\max_{\lambda'} \mathbb{E}_{\omega, \beta, \tau \mid Y} \left[ \log \mathcal{L}_{\lambda'}(Y, \omega, \beta, \tau; \lambda'^{(k)}) \right]. \]
EM algorithm

Recap Casella (2001):

$$\lambda^{(k+1)} = \arg\max_{\lambda'} \mathbb{E}_{\omega, \beta, \tau \mid Y} \left[ \log \mathcal{L}_{\lambda'}(Y, \omega, \beta, \tau; \lambda^{(k)}) \right].$$

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:

\[ p(\omega, \beta, \tau | Y) \approx q(\omega)q(\beta)q(\tau) =: Q \]

\[ \mathbb{E}_{p(\omega, \beta, \tau | Y)} [\log \mathcal{L}_{\lambda'}(\cdot)] \approx \mathbb{E}_{Q} [\log \mathcal{L}_{\lambda'}(\cdot)] =: f(\lambda') \]
Empirical-variational Bayes

Variational Bayes
Approximate posterior factorizes:

\[ p(\omega, \beta, \tau | Y) \approx q(\omega)q(\beta)q(\tau) =: Q \]

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

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

Feature selection

1. Plug estimated penalty parameters into frequentist elastic net:

\[ \hat{\beta} := \arg\max_{\beta} \log L(Y; \beta) + \frac{\alpha\lambda}{2} \sum_{j=1}^{p} \sqrt{\lambda'_{g(j)}|\beta_j|} + \frac{(1 - \alpha)\lambda}{2} \sum_{j=1}^{p} \lambda'_{g(j)}\beta_j^2 \]

2. Adjust \( \lambda \) 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
  → 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

**Conservation status**

- GRridge
- gren, $\alpha = 0.05$
- gren, $\alpha = 0.5$
- gren, $\alpha = 0.95$
- not group-regularized

<table>
<thead>
<tr>
<th>$\lambda_g$</th>
<th>Not conserved</th>
<th>Broadly conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend**

- **GRridge**
- **gren, $\alpha = 0.05$**
- **gren, $\alpha = 0.5$**
- **gren, $\alpha = 0.95$**
- **not group-regularized**
Clinician:

“That’s all nice, but does the predictive accuracy improve?”
Performance under variable selection

AUC assessed by LOOCV

![Graph showing performance under variable selection with AUC assessed by LOOCV. The graph plots AUC against the number of selected features for different regularization methods and parameters.](image-url)
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, $\lambda'_g$ fixed. Example in the Review.

---

**Te Beest, et al., BMC Bioinf, 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 GRidge, gren and dependencies.

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

††Slides available via: www.bigstatistics.nl