Joint estimation of multiple sparse regression functions with simple ℓ_1 penalties.

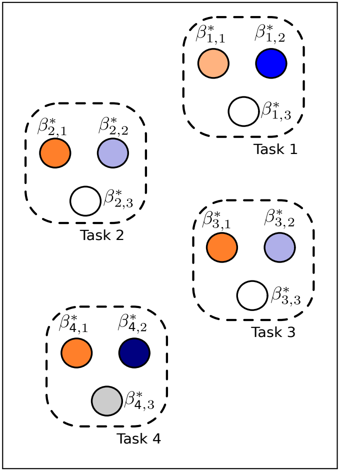
V. Viallon & E. Ollier

Univ. Lyon 1; IFSTTAR

Illustrative example

- Observations collected in several strata of a population
  - Cholesterol and gene expression levels in France, Italy, Spain and Germany for instance.
- "Naive" approaches
  - Estimate 4 independent models: does not account for the usual homogeneity: $\beta_{k,j} \simeq \beta_{k',j}$ for most genes j
  - Estimate 1 model, on all the data: masks the potential heterogeneity.

$\Rightarrow$ Need for dedicated approaches, that automatically adapts for the level of heterogeneity



Joint estimation of K linear regression models

- For each stratum k, data are supposed to follow the model

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\beta}_k^* + \boldsymbol{\xi}^{(k)}$$

where
  - $\mathbf{Y}^{(k)} = (Y_1^{(k)}, \ldots, Y_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$
  - $\boldsymbol{\xi}^{(k)} = (\xi_1^{(k)}, \ldots, \xi_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$
  - $\mathbf{X}^{(k)} = (x_1^{(k)\,T}, \ldots, x_{n_k}^{(k)\,T})^T \in \mathbb{R}^{n_k \times p}$.

- Introduce $n = \sum_{k=1}^{K} n_k$

$\Rightarrow$ K linear regression models on fixed design, with gaussian and homoscedastic errors (no intercept).

- Extensions : logistic regression, Cox model, etc.

Principle: decomposition of the $\boldsymbol{\beta}_k^*$'s

- We can always write $\boldsymbol{\beta}_k^* = \bar{\boldsymbol{\beta}}^* + \boldsymbol{\gamma}_k^*$
- $\bar{\beta}_j^*$ is the "common" effect of gene j over the strata:
  - $\bar{\beta}_j^* = \mathrm{mode}(\beta_{1,j}^*, \dots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{median}(\beta_{1,j}^*, \dots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{mean}(\beta_{1,j}^*, \dots, \beta_{K,j}^*)$

Principle: decomposition of the $\boldsymbol{\beta}_k^*$'s

- We can always write $\boldsymbol{\beta}_k^* = \bar{\boldsymbol{\beta}}^* + \boldsymbol{\gamma}_k^*$
- $\bar{\beta}_j^*$ is the "common" effect of gene j over the strata:
  - $\bar{\beta}_j^* = \mathrm{mode}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{median}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{mean}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
- Idea: Encourage solutions $(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ such that
  - $\widehat{\bar{\beta}}_j \approx \mathrm{median}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$
  - The common effect $\widehat{\bar{\boldsymbol{\beta}}}$ is sparse;
  - Vectors $\widehat{\boldsymbol{\beta}}_k - \widehat{\bar{\boldsymbol{\beta}}}$ are sparse: $\Rightarrow$ similarity of the $\widehat{\boldsymbol{\beta}}_k$'s

Principle: decomposition of the $\boldsymbol{\beta}_k^*$'s

- We can always write $\boldsymbol{\beta}_k^* = \bar{\boldsymbol{\beta}}^* + \boldsymbol{\gamma}_k^*$
- $\bar{\beta}_j^*$ is the "common" effect of gene j over the strata:
  - $\bar{\beta}_j^* = \mathrm{mode}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{median}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}_j^* = \mathrm{mean}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
- Idea: Encourage solutions $(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ such that
  - $\widehat{\bar{\beta}}_j \approx \mathrm{median}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j})$
  - The common effect $\widehat{\bar{\boldsymbol{\beta}}}$ is sparse;
  - Vectors $\widehat{\boldsymbol{\beta}}_k - \widehat{\bar{\boldsymbol{\beta}}}$ are sparse: $\Rightarrow$ similarity of the $\widehat{\boldsymbol{\beta}}_k$'s
- M_1

$$\sum_{k \geqslant 1} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\bar{\boldsymbol{\beta}} + \boldsymbol{\gamma}_k)\|_2^2}{2n} + \lambda_1 \|\bar{\boldsymbol{\beta}}\|_1 + \sum_{k \geqslant 1} \lambda_{2,k} \|\boldsymbol{\gamma}_k\|_1$$

Remarks

- M_1

$$\sum_{k\geqslant 1}\frac{\|\mathbf{Y}^{(k)}-\mathbf{X}^{(k)}(\bar{\boldsymbol{\beta}}+\boldsymbol{\gamma}_k)\|_2^2}{2n}+\lambda_1\|\bar{\boldsymbol{\beta}}\|_1+\sum_{k\geqslant 1}\lambda_{2,k}\|\boldsymbol{\gamma}_k\|_1$$

- At optimum, $\widehat{\bar{\beta}}_j$: shrunk version of median$(\widehat{\beta}_{1,j},\dots,\widehat{\boldsymbol{\beta}}_{K,j})$: penalization yields “identifiability”.

- If λ_1 is large enough, $\widehat{\bar{\boldsymbol{\beta}}}=\mathbf{0}_p$ and $\widehat{\boldsymbol{\beta}}_k=\widehat{\boldsymbol{\gamma}}_k$.

- If the $\lambda_{2,k}$'s are large enough, $\widehat{\boldsymbol{\gamma}}_k=\mathbf{0}_p$ and $\widehat{\boldsymbol{\beta}}_k=\widehat{\bar{\boldsymbol{\beta}}}$.

Adaptive version of M1 when $n_k \gg p$ and K is odd

- Denote by $\tilde{\boldsymbol{\beta}}_k$ initial OLS estimates of the $\boldsymbol{\beta}_k$'s.
- For all j, further define ℓ_j such that $\tilde{\beta}_{\ell_j,j} = \text{median}(\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{K,j})$.
- And set $\bar{\beta}_j^{\text{init}} = \tilde{\beta}_{\ell_j,j}$.

Adaptive version of M1 when $n_k \gg p$ and K is odd

- Denote by $\tilde{\boldsymbol{\beta}}_k$ initial OLS estimates of the $\boldsymbol{\beta}_k$'s.
- For all j, further define ℓ_j such that $\tilde{\beta}_{\ell_j,j} = \text{median}(\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{K,j})$.
- And set $\bar{\beta}_j^{\text{init}} = \tilde{\beta}_{\ell_j,j}$.
- Then setting $w_{k,j} = 1/|\tilde{\beta}_{k,j}|$ and $\nu_{k,j} = 1/|\tilde{\beta}_{k,j} - \bar{\beta}_j^{\text{init}}|$ consider the problem

$$\sum_{k\geqslant 1} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \sum_j w_{\ell_j,j}|\bar{\beta}_j| + \sum_{k\geqslant 1} \lambda_{2,k} \sum_j \nu_{k,j}|\gamma_{k,j}|\Big\}.$$

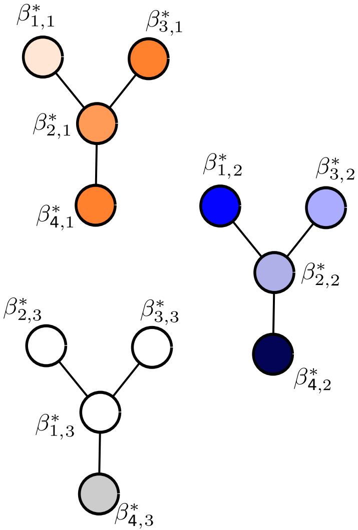
- This problem can be rewritten as

$$\sum_{k\geqslant 1} \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}_k\|_2^2}{2n} + \lambda_1 \sum_j w_{\ell_j,j}|\beta_{\ell_j,j}| + \sum_{k\geqslant 1} \lambda_{2,k} \sum_j \nu_{k,j}|\beta_{k,j} - \beta_{\ell_j,j}|$$

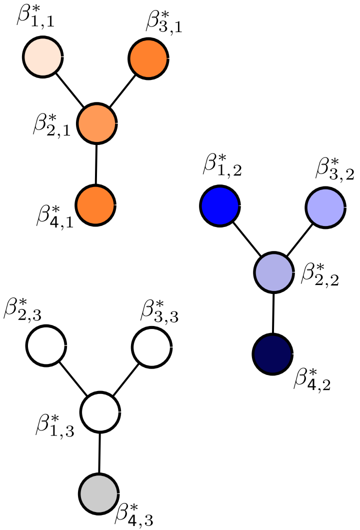
Connection with the "reference stratum" strategy:

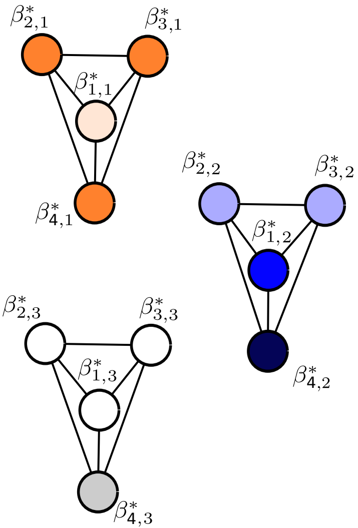
- This corresponds to the decomposition: $\beta_{k,j} = \beta_{\ell_j,j} + \gamma_{k,j}$, for $k \neq \ell_j$.
- Can be seen as a refinement of the following "standard" approach:
  - Select a reference stratum ℓ
  - Write $\boldsymbol{\beta}_k = \boldsymbol{\beta}_\ell + \boldsymbol{\delta}_k$
  - Select the non-zero components in $\boldsymbol{\beta}_\ell$ and $\boldsymbol{\delta}_k$ (*e.g.*, with the lasso).
- Advantages of our approach:
  - the reference ℓ_j stratum is automatically selected from the initial OLS estimates;
  - it does not have to be the same for all gene j.

M_1 vs GenFused



M_1 vs GenFused





Implementation of M_1: weighted lasso on "augmented" data.

- Set

$$\mathcal{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \vdots \\ \mathbf{Y}^{(K)} \end{pmatrix} \quad \mathcal{X} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \mathbf{X}^{(K)} \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \bar{\boldsymbol{\beta}} \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_K \end{pmatrix}$$

which belong to $\mathbb{R}^n$, $\mathbb{R}^{n\times(K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.

Implementation of M_1: weighted lasso on "augmented" data.

- Set

$$\mathcal{Y} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \vdots \\ \mathbf{Y}^{(K)} \end{pmatrix} \quad \mathcal{X} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & \mathbf{0} & \dots & \mathbf{X}^{(K)} \end{pmatrix} \quad \theta = \begin{pmatrix} \bar{\beta} \\ \gamma_1 \\ \vdots \\ \gamma_K \end{pmatrix}$$

which belong to $\mathbb{R}^n$, $\mathbb{R}^{n\times(K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.

- Setting $\nu = (1, \dots, \lambda_{2,1}/\lambda_1, \dots, \lambda_{2,K}/\lambda_1)$, the objective function minimized under M_1 can be written as

$$\frac{\|\mathcal{Y} - \mathcal{X}\theta\|_2^2}{2n} + \lambda_1 \|\theta\|_{\nu(1)},$$

where $\|\theta\|_{\nu(1)} = \sum_{j=1}^{(K+1)p} \nu_j |\theta_j|$: $\Rightarrow$ Weighted Lasso

- Asymptotic oracle properties of the adaptive versions
  - p and K are held fixed, while $n \to \infty$
  - General assumptions:
    - $n_k / n \kappa$, with $0 < \kappa < 1$
    - $(\mathbf{X}^{(k)\,T} \mathbf{X}^{(k)}) / n_k \to \mathbf{C}^{(k)}$ with $\lambda_{\min}(\mathbf{C}^{(k)}) > 0$
    - ...
  - Easy to derive from results obtained for GenFused (*e.g.*, V. et al., 2014)

- Non-asymptotic properties of M_1
  - Generally, $\mathcal{X}$ does not enjoy RIP, RE, or the Irrepresentability condition.
  - In some cases though, it does !
    - If the $\boldsymbol{\beta}_k^*$'s are all equal
    - If $K > 49\alpha^2 s^2$, where $\alpha > 1$ and $s \geqslant |supp(\theta^*)|$.

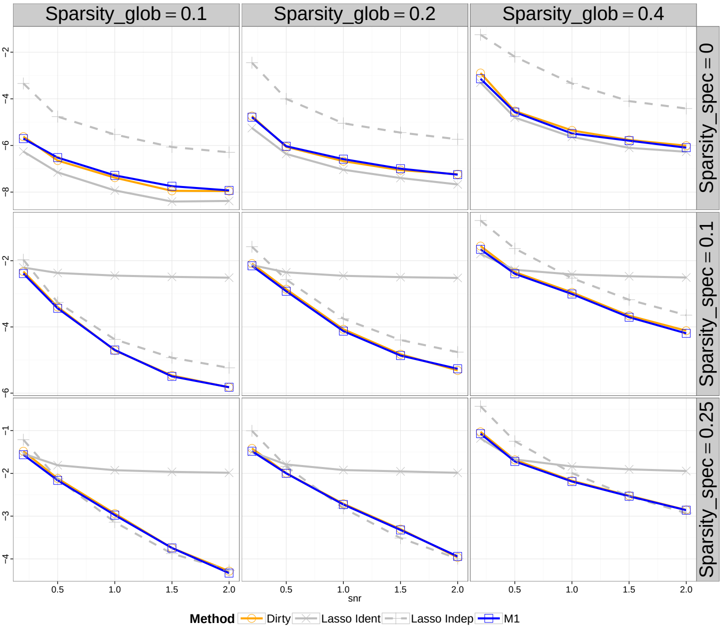
The Dirty Model of Jalali et al., 2013

- Jalali et al. consider the following objective function

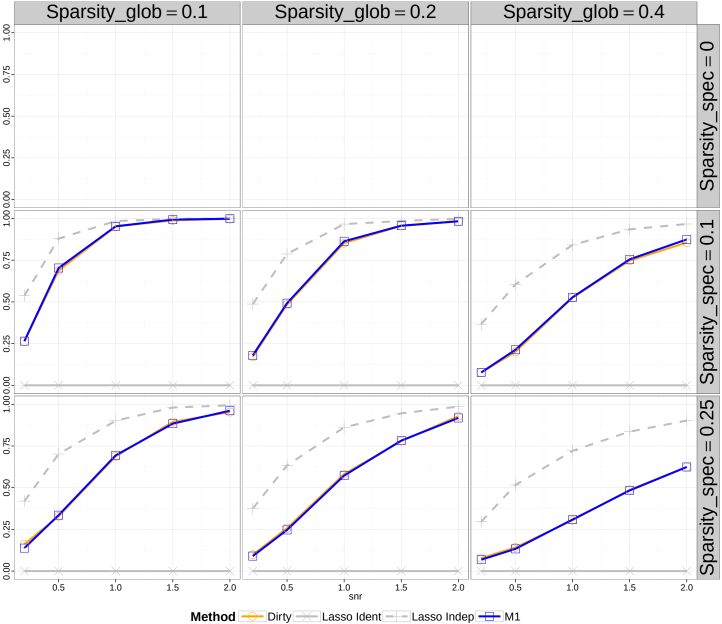
$$\sum_k \left\{ \frac{\|\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\mathbf{R}^{(k)} + \mathbf{S}^{(k)})\|_2^2}{2n_k} + \lambda_S \|\mathbf{S}^{(k)}\|_1 \right\} + \lambda_R \sum_{j=1}^{p} \|R_j\|_\infty$$

  avec $R_j = (R_j^{(1)}, \dots, R_j^{(K)})^T$.
- The optimal solution $\widehat{\mathbf{B}} = \widehat{\mathbf{R}} + \widehat{\mathbf{S}}$ is then the sum of a row-sparse and a sparse matrix.
- Our approaches can be seen as simplified versions of Dirty: in particular, M_1
  - is much faster to run for linear regression models (many fast functions for the lasso)
  - is ready to use under many other regression models:
    - GLMs (logistic, Poisson, multinomial, ...), Cox models
    - conditional logistic regression, Cox models with competing risks, ...
- When $K = 2$, and if $\beta^*_{1j}\beta^*_{2j} \geqslant 0$, $M_1 \geqslant$ Dirty for support recovery!

Log-Prediction Error $(p = 15, K = 5, n_k = 225)$



Heterogeneity detection ($p = 15, K = 5, n_k = 225$)



Conclusion/ Discussion

- Our approach has deep connections with Ref^{ce}. Stratum, GenFused, and Dirty:
  - Refined version of the Ref^{ce}. Stratum strategy: the selection of the reference stratum is automatic, and "gene"-dependent.
  - Simplified version of the Dirty; similar empirical performance
  - Ready to use under a variety of models: linear, logistic, Cox, conditional logistic, ...

- Perspectives:
  - Extend our non-asymptotic results (lasso on correlated designs, Dalalyan et al. 2014).
  - Apply the approach on real data sets: Breast Cancer (IARC, WHO), Drivers Responsibility (IFSTTAR), ...
  - Extend the method to graphical models, ...