Joint estimation of multiple sparse regression functions with simple $\ell_1$ 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

$$Y^{(k)} = X^{(k)} \beta_k^* + \xi^{(k)}$$

where

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

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

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

- Extensions: logistic regression, Cox model, etc.
Principle: decomposition of the $\beta_k$’s

- We can always write $\beta_k^* = \bar{\beta}^* + \gamma_k$
- $\bar{\beta}^*_j$ is the “common” effect of gene $j$ over the strata:
  - $\bar{\beta}^*_j = \text{mode}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}^*_j = \text{median}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}^*_j = \text{mean}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
Principle: decomposition of the $\beta^*_k$’s

- We can always write $\beta^*_k = \bar{\beta}^* + \gamma^*_k$
- $\bar{\beta}^*_j$ is the “common” effect of gene $j$ over the strata:
  - $\bar{\beta}^*_j = \text{mode}(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$
  - $\bar{\beta}^*_j = \text{median}(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$
  - $\bar{\beta}^*_j = \text{mean}(\beta^*_{1,j}, \ldots, \beta^*_{K,j})$

- Idea: Encourage solutions $(\hat{\beta}_1, \ldots, \hat{\beta}_K)$ such that
  - $\hat{\beta}_j \approx \text{median}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j})$
  - The common effect $\hat{\beta}$ is sparse;
  - Vectors $\hat{\beta}_k - \hat{\beta}$ are sparse: $\Rightarrow$ similarity of the $\hat{\beta}_k$’s
Principle: decomposition of the $\beta_k^*$’s

- We can always write $\beta_k^* = \bar{\beta}^* + \gamma_k^*$
- $\bar{\beta}^*_j$ is the “common” effect of gene $j$ over the strata:
  - $\bar{\beta}^*_j = \text{mode}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}^*_j = \text{median}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
  - $\bar{\beta}^*_j = \text{mean}(\beta_{1,j}^*, \ldots, \beta_{K,j}^*)$
- Idea: Encourage solutions $(\hat{\beta}_1, \ldots, \hat{\beta}_K)$ such that
  - $\hat{\beta}_j \approx \text{median}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j})$
- The common effect $\hat{\beta}$ is sparse;
- Vectors $\hat{\beta}_k - \hat{\beta}$ are sparse: $\Rightarrow$ similarity of the $\hat{\beta}_k$’s
- $\mathcal{M}_1$
  \[
  \sum_{k \geq 1} \frac{\| Y^{(k)} - X^{(k)} (\bar{\beta} + \gamma_k) \|^2}{2n} + \lambda_1 \| \bar{\beta} \|_1 + \sum_{k \geq 1} \lambda_{2,k} \| \gamma_k \|_1
  \]
Remarks

- \( M_1 \)
  \[
  \sum_{k \geq 1} \frac{\| Y^{(k)} - X^{(k)}(\bar{\beta} + \gamma_k) \|_2^2}{2n} + \lambda_1 \| \bar{\beta} \|_1 + \sum_{k \geq 1} \lambda_{2,k} \| \gamma_k \|_1
  \]

- At optimum, \( \hat{\beta}_j \): shrunken version of \( \text{median}(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{K,j}) \): penalization yields “identifiability”.

- If \( \lambda_1 \) is large enough, \( \hat{\beta} = 0_p \) and \( \hat{\beta}_k = \hat{\gamma}_k \).

- If the \( \lambda_{2,k} \)'s are large enough, \( \hat{\gamma}_k = 0_p \) and \( \hat{\beta}_k = \hat{\beta} \).
Adaptive version of M1 when $n_k \gg p$ and $K$ is odd

- Denote by $\tilde{\beta}_k$ initial OLS estimates of the $\beta_k$'s.
- For all $j$, further define $\ell_j$ such that $\tilde{\beta}_{\ell_j,j} = \text{median}(\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{K,j})$.
- And set $\bar{\beta}^{\text{init}}_j = \tilde{\beta}_{\ell_j,j}$.
Adaptive version of M1 when $n_k \gg p$ and $K$ is odd

- Denote by $\tilde{\beta}_k$ initial OLS estimates of the $\beta_k$'s.
- For all $j$, further define $\ell_j$ 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

$$\sum_{k\geq 1} \frac{\|Y^{(k)} - X^{(k)}\beta_k\|^2_2}{2n} + \lambda_1 \sum_j w_{\ell_j,j} |\tilde{\beta}_{\ell_j,j}| + \sum_{k\geq 1} \lambda_{2,k} \sum_j \nu_{k,j} |\gamma_{k,j}| \}.$$ 

- This problem can be rewritten as

$$\sum_{k\geq 1} \frac{\|Y^{(k)} - X^{(k)}\beta_k\|^2_2}{2n} + \lambda_1 \sum_j w_{\ell_j,j} |\beta_{\ell_j,j}| + \sum_{k\geq 1} \lambda_{2,k} \sum_j \nu_{k,j} |\beta_{k,j} - \beta_{\ell_j,j}|.$$
Connection with the “reference stratum” strategy:

- This corresponds to the decomposition: \( \beta_{k,j} = \beta_{\ell,j} + \gamma_{k,j} \), for \( k \neq \ell_j \).

- Can be seen as a refinement of the following “standard” approach:
  - Select a reference stratum \( \ell \)
  - Write \( \beta_k = \beta_\ell + \delta_k \)
  - Select the non-zero components in \( \beta_\ell \) and \( \delta_k \) (e.g., with the lasso).

- Advantages of our approach:
  - the reference \( \ell_j \) 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

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

which belong to $\mathbb{R}^n$, $\mathbb{R}^{n \times (K+1)p}$ and $\mathbb{R}^{(K+1)p}$, respectively.
Implementation of $M_1$: weighted lasso on “augmented” data.

1. Set

\[
\mathbf{y} = \begin{pmatrix}
\mathbf{y}^{(1)} \\
\vdots \\
\mathbf{y}^{(K)}
\end{pmatrix} \quad \mathbf{x} = \begin{pmatrix}
\mathbf{x}^{(1)} & \mathbf{x}^{(1)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{x}^{(K)} & 0 & \cdots & \mathbf{x}^{(K)}
\end{pmatrix} \quad \mathbf{\theta} = \begin{pmatrix}
\bar{\beta} \\
\gamma_1 \\
\vdots \\
\gamma_K
\end{pmatrix}
\]

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

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

\[
\frac{\|\mathbf{y} - \mathbf{x}\mathbf{\theta}\|_2^2}{2n} + \lambda_1 \|\mathbf{\theta}\|_{\nu(1)},
\]

where \( \|\mathbf{\theta}\|_{\nu(1)} = \sum_{j=1}^{(K+1)p} \nu_j|\theta_j| : \Rightarrow \text{Weighted Lasso} \)
Asymptotic oracle properties of the adaptive versions

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

Non-asymptotic properties of $M_1$

- Generally, $X$ does not enjoy RIP, RE, or the Irrepresentability condition.
- In some cases though, it does!
  - If the $\beta^*_k$'s are all equal
  - If $K > 49\alpha^2 s^2$, where $\alpha > 1$ and $s \geq |supp(\theta^*)|$. 
The Dirty Model of Jalali et al., 2013

- Jalali et al. consider the following objective function

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

avec \( R_j = (R_j^{(1)}, \ldots, R_j^{(K)})^T \).

- The optimal solution \( \hat{B} = \hat{R} + \hat{S} \) 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 \( \beta_{1j}^* \beta_{2j}^* \geq 0 \), \( M_1 \geq \) Dirty for support recovery!
Log-Prediction Error ($p = 15$, $K = 5$, $n_k = 225$)
Heterogeneity detection ($p = 15$, $K = 5$, $n_k = 225$)

<table>
<thead>
<tr>
<th>Sparsity_glob = 0.1</th>
<th>Sparsity_glob = 0.2</th>
<th>Sparsity_glob = 0.4</th>
</tr>
</thead>
</table>

Method: Dirty Lasso Ident Lasso Indep M1
Conclusion/ Discussion

- Our approach has deep connections with Ref\textsuperscript{ce}. Stratum, GenFused, and Dirty:
  - Refined version of the Ref\textsuperscript{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, ...