Hierarchical models and structured penalties

Manuela Zucknick and Theophilus Quachie Asenso

Oslo Centre for Biostatistics and Epidemiology

Contents

- Motivation and reason for hierarchical modeling
- Structure within responses
- Structure withing the covariates (Interaction models with hierarchical properties)
- Example with MADMMplasso

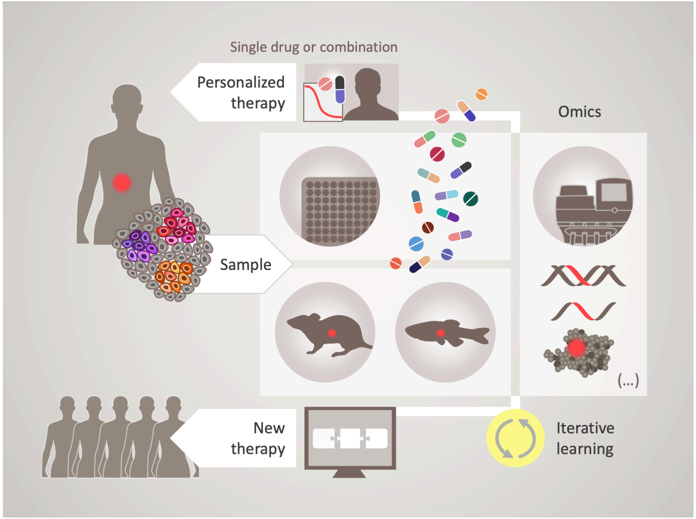
Motivation and reason for hierarchical modeling

$$\text{N cell lines} \; \overbrace{\begin{bmatrix} | & & | \\ y_{\cdot 1} & \dots & y_{\cdot D} \\ | & & | \end{bmatrix}}^{\substack{\text{Drug dose response} \\ \text{drug sensitivity}}} = Y$$

$$\text{N cell lines} \; \overbrace{\begin{bmatrix} | & & | \\ X_{\cdot 1} & \dots & X_{\cdot p} \\ | & & | \end{bmatrix}}^{\substack{\text{Genetic features} \\ \text{gene expression}}} = X$$

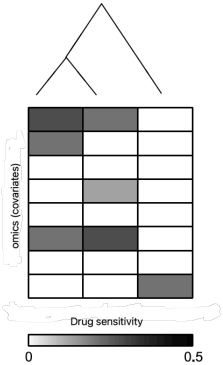
$$\text{N cell lines} \; \overbrace{\begin{bmatrix} | & & | \\ Z_{\cdot 1} & \dots & Z_{\cdot K} \\ | & & | \end{bmatrix}}^{\substack{\text{Interactions} \\ \text{Cancer type}}} = Z$$

Motivation and reason for hierarchical modeling



slide by Kjetil Taskén

Motivation and reason for hierarchical modeling



- Structures in the response matrix ([Kim and Xing, 2012], [Li et al., 2015]) for example correlations between drug responses due to similar chemical properties, drug target, drug functions, etc
- Structures within the covariates or with a set of modifying variables ([Li et al., 2015], [Tibshirani and Friedman, 2020]) for example gene-to-gene interactions, gene-to-cancer type interactions, correlated genes, etc

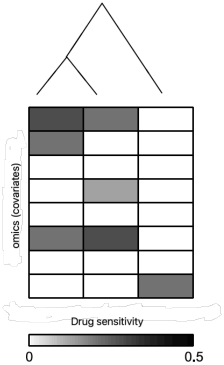
Motivation and reason for hierarchical modeling

How do we handle such problem?

- The response cannot be explained by only additive functions of the variables.
- There is the need to consider interactions
- We also need a model that captures the correlational structure in the response and not treat each response separately.

Structure within responses

Structure within responses (with tree lasso)



$$\mathcal{G}_{m_5} = \{B_{j1}, B_{j2}, B_{j3}\}$$

$$\mathcal{G}_{m_4} = \{B_{j1}, B_{j2}\}$$

$$\mathcal{G}_{m_1} = \{B_{j1}\} \quad \mathcal{G}_{m_2} = \{B_{j2}\} \quad \mathcal{G}_{m_3} = \{B_{j3}\}$$

Structure within responses (with tree lasso)

- The set of internal and leaf nodes of the tree as M_{int}, M_{leaf} of size $|M_{\text{int}}|$ and $|M_{\text{leaf}}|$ respectively;
- The group of responses forming an internal node $m \in M_{\text{int}}$ as $\mathcal{G}_m$, where $\mathcal{G}_m \subseteq \{1, \ldots, D\}$ and let $B_j^{\mathcal{G}_m}$ denotes the j^{th} sub-vector of B, indexed by $\mathcal{G}_m$ with a group weight w_m.
- Each sub-vector $B_j^{\mathcal{G}_m}$ has elements $\{B_{jd}; d \in \mathcal{G}_m\}$.

Structure within responses (with tree lasso)

The simplified version of [Kim and Xing, 2012] is;

$$\min_B \frac{1}{2N}\|Y-\hat{Y}\|_F^2+\lambda\sum_{j=1}^{p}\sum_{m\in M_{\text{int}}} w_m\|B_j^{\mathcal{G}_m}\|_2+\lambda\sum_{j=1}^{p}\sum_{m\in M_{\text{leaf}}} w_m\|B_j^{\mathcal{G}_m}\|_2. \quad (1)$$

Structure withing the covariates

Interaction models with hierarchical properties

Interaction models with hierarchical properties

The hierNet model [Bien et al., 2013]

$$y = \beta_0 + \sum_{j}^{p} \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \epsilon, \tag{2}$$

where $\epsilon \sim \mathbb{N}(0, \sigma^2)$, $\beta \in \mathbb{R}^p$, $\Theta \in \mathbb{R}^{p \times p}$ and $\Theta_{jj} = 0$.

$$\min_{\beta_0 \in \mathbb{R}, \beta^{\pm} \in \mathbb{R}^p, \Theta \in \mathbb{R}^{p \times p}} \ell(\beta_0, \beta, \Theta) + \lambda \sum_{j} max\{|\beta_j|, \|\Theta_j\|_1\} + \frac{\lambda}{2} |\Theta\|_1 \tag{3}$$

Interaction models with hierarchical properties

Glinternet

Consider a dataset containing $\mathbf{y}$ response and two categorical variables F_1, F_2 with p_1, p_2 levels. Let $\mathbf{X}_1, \mathbf{X}_2$ be their corresponding indicator matrices with p_1, p_2 columns respectively.

Interaction models with hierarchical properties

The GLINTERNET model [Lim and Hastie, 2015]

$$\min_{\mu,\alpha,\tilde{\alpha}} \frac{1}{2} \left\| \mathbf{y} - \mathbf{1}\mu - \mathbf{X}_1\alpha_1 - \mathbf{X}_2\alpha_2 - [\mathbf{X}_1\mathbf{X}_2\mathbf{X}_{1:2}] \begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \alpha_{1:2} \end{bmatrix} \right\|_2^2 + \lambda(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{p_1\|\tilde{\alpha}_1\|_2^2 + p_2\|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}) \qquad (4)$$

$$\text{subject to} \quad \sum_{i=1}^{p_1} \alpha_1^i = 0, \quad \sum_{j=1}^{p_2} \alpha_2^j = 0, \quad , \sum_{i=1}^{p_1} \tilde{\alpha}_1^i = 0, \quad \sum_{j=1}^{p_2} \tilde{\alpha}_2^j = 0 \qquad (5)$$

$$\text{and} \quad \sum_{i=1}^{p_1} \alpha_{1:2}^{ij} = 0 \quad \text{for fixed} \quad j, \quad \sum_{j=1}^{p_2} \alpha_{1:2}^{ij} = 0 \quad \text{for fixed} \quad i, \qquad (6)$$

Interaction models with hierarchical properties

The GLINTERNET model [Lim and Hastie, 2015]

GLINTERNET can be solved as an unconstrained group lasso problem by using the following equivalent objective function;

$$\underset{\mu,\beta}{argmin} \frac{1}{2} \|\mathbf{y} - \mathbf{1}\mu - \mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2 - \mathbf{X}_{1:2}\beta_{1:2}\|_2^2 + \lambda(\|\beta_1\|_2 + \|\beta_2\|_2 + \|\beta_{1:2}\|_2) \quad (7)$$

Interaction models with hierarchical properties (Pliable lasso)

$y \in \mathbb{R}^N$, $X \in \mathbb{R}^{N\times p}$ and $Z \in \mathbb{R}^{N\times K}$.The pliable lasso [Tibshirani and Friedman, 2020] model is given as;

$$
\begin{aligned}
\hat{y} &= \beta_0 \mathbf{1} + Z\theta_0 + \sum_{j=1}^{p} X_j(\beta_j \mathbf{1} + Z\theta_j) \\
&= \beta_0 + Z\theta_0 + X\beta + \sum_{j=1}^{p} (X_j \odot Z)\theta_j,
\end{aligned}
\tag{8}
$$

where $(X_j \odot Z)$ denoting the $N \times K$ matrix formed by multiplying each column of Z component-wise by the column vector X_j.

Interaction models with hierarchical properties (Pliable lasso)

The pliable lasso objective function

$$M(\beta_0, \theta_0, \beta, \theta) = \frac{1}{2N}\sum_i (y_i - \hat{y}_i)^2$$
$$+ (1-\alpha)\lambda \sum_{j=1}^{p} \overbrace{(\|(\beta_j, \theta_j)\|_2 + \|\theta_j\|_2)}^{\text{Overlapping group}} + \alpha\lambda \sum_{j,k} |\theta_{j,k}| \quad (9)$$

- y_i is the element of the fitted model $\beta_0\mathbf{1} + Z\theta_0 + \sum_{j=1}^{p} X_j(\beta_j\mathbf{1} + Z\theta_j)$.
- Overlapping group ensures **(asymmetric) weak hierarchy constraint**.

Interaction models with hierarchical properties

Table: Hierarchical Sparse modeling (HSM) methods

Penalty	Input dataset	Method	Type of hierarchy
hiernet [Bien et al., 2013]	(x, y)	Group lasso	$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0$ and $\hat{\beta}_k \neq 0$ $\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0$ or $\hat{\beta}_k \neq 0$
glinternet [Lim and Hastie, 2015]	(x, y)	Latent overlapping group lasso	$\hat{\Theta}_{jk} \neq 0 \Rightarrow \hat{\beta}_j \neq 0$ and $\hat{\beta}_k \neq 0$
plasso [Tibshirani and Friedman, 2020]	(x, y, z)	group lasso with overlapping groups	$\hat{\Theta}_{jk}$ can be non zero only if $\hat{\beta}_j \neq 0$. Converse not true

Example with MADMMplasso

Example with MADMMplasso

- Let $B \in \mathbb{R}^{D\times p\times(K+1)}$.
- The j^{th} row of B_d defined as $B_{jd} = [\boldsymbol{\beta}_{jd}, \boldsymbol{\theta}_{jd}] \in \mathbb{R}^{K+1}$.
- Let W be an $N \times p \times (1+K)$

$$W_{i,j,k} = \begin{cases} X_{ij}Z_{ik} & \text{for} \quad k \neq 1 \\ X_{ij} & \text{for} \quad k = 1, \end{cases} \tag{10}$$

$k = 1, 2, \ldots, K+1.$

$$\hat{Y} = \mathbf{1}\beta_0^T + Z\boldsymbol{\theta} + W * B, \tag{11}$$

where $W * B = [W * B_1 : W * B_2 : \ldots : W * B_D]$ to denote $N \times D$ matrix whose i, d element takes the form

$$(W * B)_{id} = \sum_{j=1}^{p} \sum_{k=1}^{K+1} W_{i,j,k} B_{jkd}, \quad i = 1, 2, \ldots N, \quad d = 1, 2, \ldots, D. \tag{12}$$

Example with MADMMplasso

- $B \in \mathbb{R}^{D \times p \times (K+1)}$.

The general multi-response pliable lasso model can be written as

$$\min_{B \in \mathbb{R}^{D \times p \times (1+K)}} \quad \frac{1}{2N}\|Y - \hat{Y}\|_F^2 + \sum_{d=1}^{D}\left[(1-\alpha)\lambda\sum_{j=1}^{p}(\|B_{jd}\|_2 + \|B_{j(-1)d}\|_2) + \alpha\lambda\sum_{j=1}^{p}\|B_{j(-1)d}\|_1\right] \tag{13}$$

Example with MADMMplasso

Combining (13) and (1);

$$\min_{B \in \mathbb{R}^{D\times p \times (1+K)}} \frac{1}{2N}\|Y - \hat{Y}\|_F^2 + \lambda_1 \sum_{j=1}^{p} \sum_{m \in M_{\text{int}}} w_m \|B_j^{\mathcal{G}_m}\|_2 + \lambda_1 \sum_{j=1}^{p} \sum_{m \in M_{\text{leaf}}} w_m \|B_j^{\mathcal{G}_m}\|_2$$
$$+ \sum_{d}^{D} \left[(1-\alpha)\lambda_2 \sum_{j=1}^{p} (\|B_{jd}\|_2 + \|B_{j(-1)d}\|_2) + \alpha\lambda_2 \sum_{j=1}^{p} \|B_{j(-1)d}\|_1 \right]. \quad (14)$$

- We use **ADMM** [Boyd et al., 2011]: "The **alternating direction method of multipliers (ADMM)** is an algorithm that solves convex optimization problems by breaking them into smaller pieces, each of which are then easier to handle. It has recently found wide application in a number of areas." (https://stanford.edu/ boyd/admm.html)

Example with MADMMplasso: Introduction to ADMM

Given a separable objective function

$$\min_{\beta} f(\beta) + h(\beta), \tag{15}$$

- Introduce auxiliary variable ω to solve (15) as

$$\min_{\beta,\omega} f(\beta) + h(\omega) \quad \text{s.t} \quad \beta = \omega. \tag{16}$$

The problem in (16) can have a corresponding augmented Lagrangian in the form

$$L(\beta, \omega, \gamma) = f(\beta) + h(\omega) + \gamma^T(\beta - \omega) + (\rho/2)\|\beta - \omega\|_2^2. \tag{17}$$

Example with MADMMplasso : Introduction to ADMM

The ADMM algorithm updates β and ω in an alternating or sequential manner in the following way until convergence condition is met.

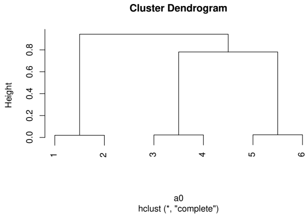
$$\begin{aligned}
\beta^{t+1} &= \arg\min_{\beta} \quad L(\beta, \omega^{t}, \gamma^{t}) \\
\omega^{t+1} &= \arg\min_{\omega} \quad L(\beta^{t+1}, \omega, \gamma^{t}) \\
\gamma^{t+1} &= \gamma^{t} + \rho(\beta^{t+1} - \omega^{t+1}).
\end{aligned} \tag{18}$$

Example with MADMMplasso

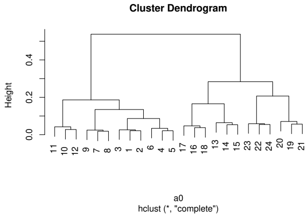
$$
\begin{aligned}
L(B, E, \tilde{E}, V, Q, H, \tilde{H}, O, P) &= \frac{1}{2N}\|Y - \hat{Y}\|_F^2 + \\
&\lambda_1 \sum_{j=1}^{p} \sum_{m \in M_{\text{int}}} w_m \|E_j^{\mathcal{G}_m}\|_2 + \lambda_2 \sum_d \sum_{j=1}^{p} w_d \|\tilde{E}_{jd}\|_2 \\
&+ \sum_d (1-\alpha)\lambda_3 \sum_{j=1}^{p} \sum_s \|V_{jd}^s\|_2 + \alpha\lambda_3 \sum_{j=1}^{p} \|Q_{jd}\|_1 + \sum_j H_j^T(\tilde{\tilde{B}}_j - E_j) + \sum_d \langle \tilde{H}_d, B_d - \tilde{E}_d \rangle \\
&+ \sum_d \sum_j O_{jd}^T(\tilde{B}_{jd} - V_{jd}) + \sum_d \langle P_d, B_d - Q_d \rangle \\
&+ \frac{\rho}{2} \sum_j \|\tilde{\tilde{B}}_j - E_j\|_2^2 + \frac{\rho}{2} \sum_d \|B_d - \tilde{E}_d\|_2^2 + \frac{\rho}{2} \sum_d \sum_j \sum_s \|\tilde{B}_{jd}^s - V_{jd}^s\|_2^2 + \frac{\rho}{2} \sum_d \|B_d - Q_d\|_2^2.
\end{aligned}
\tag{19}
$$

Example with MADMMplasso

$D = 7, p = 500, K = 4, N = 100$



$D = 24, p = 150, 500, K = 4, N = 100$



Simulated correlation structure of D drug response variables across N cell lines for simulated data set 1 (left) and 2 (right)."

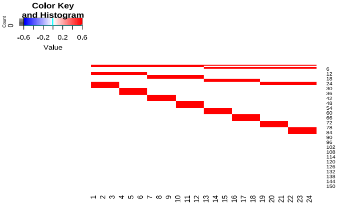
Example with MADMMplasso: Results for simulated data set 1

Table: Results from the multi-response simulation 1 with weak hierarchical structure in the response.

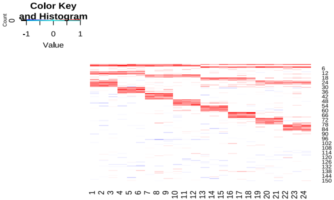
Model	$(1/Dp)\|\hat{\beta}-\beta\|_1$	Sensitivity[1]	Specificity[2]	Non-zero[3]	Test error (SD)[4]
Plasso	0.021	1	0.763	733	19.693 (2.408)
Tree lasso	0.066	1	0.142	2577	34.045 (1.802)
MADMMplasso	0.006	1	0.991	237	5.050 (0.681)

[1] Sensitivity is the proportion of non-zero coefficients estimated as non-zeros.
[2] Specificity is the proportion of zero-coefficients estimated as zeros.
[3] The total number of non-zero coefficients in the model. We counted the coefficients with at least two non-zero values across the 10 simulations.
Number of non-zero coefficients $= \sum_{j=1}^{p}\sum_{d=1}^{D}\{(\sum_{r=1}^{10}\mathbf{1}_{\{\beta_{jd}^{r}\neq 0\}}) \geq 2\}$. Note that the selection is out of $p \times D = 3000$ features in total.
[4] The MSE on an independent test dataset. We include the standard deviation (SD) across the 10 simulations.

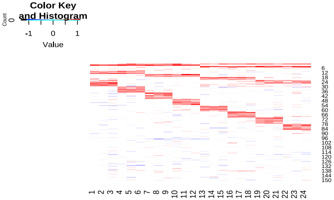
Example with MADMMplasso: Results for simulated data set 2



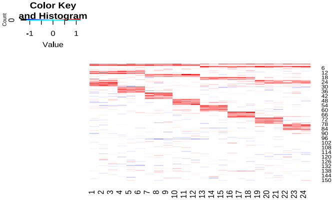
a True structure



b MADMMplasso



c plasso



d Tree lasso

Example with MADMMplasso: Results for simulated data set 2

Table: Results from the multi-response simulation 2 with strong hierarchical structure in the responses.

Model	$(1/Dp)\|\hat{\beta}-\beta\|_1$	Sensitivity[1]	Specificity[2]	Non-zero[3]	Test error (SD)[4]
$p=150$					
Plasso	0.034	1	0.446	2155	2.512 (0.181)
Tree lasso	0.036	1	0.345	2483	2.072 (0.095)
MADMMplasso	0.0299	1	0.727	2014	1.972 (0.112)
$p=500$					
Plasso	0.014	0.994	0.814	2514	4.57 (1.038)
Tree lasso	0.023	1	0.360	7826	2.927 (0.163)
MADMMplasso	0.010	1	0.912	1891	2.230 (0.116)

[1] Sensitivity is the proportion of non-zero coefficients estimated as non-zeros.
[2] Specificity is the proportion of zero-coefficients estimated as zeros.
[3] Number of non-zero coefficients $=\sum_{j=1}^{p}\sum_{d=1}^{D}\{(\sum_{r=1}^{10}\mathbf{1}_{\{\beta^r_{jd}\neq 0\}})\geq 2\}$. Note that the selection is out of $p\times D=3600$ (for $p=150$) or 12000 (for $p=500$) features in total.
[4] The MSE on an independent test dataset. We included the standard deviation (SD) across the 10 simulations.

Example with MADMMplasso: Real data

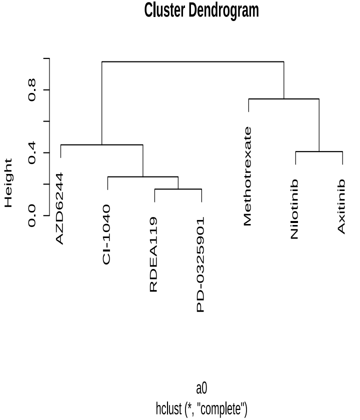
'Genomics of drug sensitivity in cancer' [Garnett et al., 2012]

- Large-scale pharmacogenomic study with $N = 498$ cell lines and $D = 97$ drugs (we used 7 drugs).
- Outcome data: $\log(IC_{50})$ from dose-response experiments
- Random draws of 80% cell lines as training data and 20% as validation data.
- Input data: Z as cancer types (13 cancer types, $K = 12$), X as mRNA expression (p=2602)

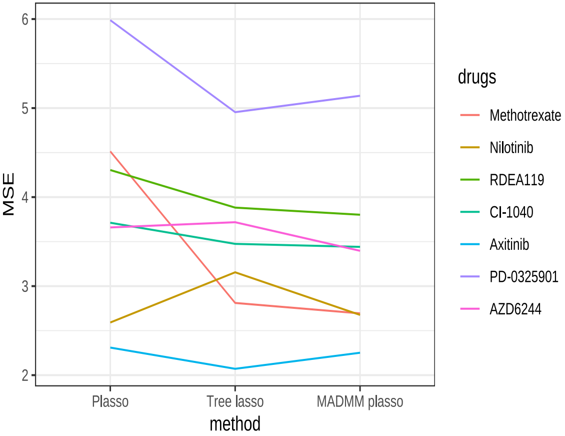
Example with MADMMplasso: Real data: Drug information

- **PD-0325901, RDEA119, CI-1040, AZD6244**: MEK1 inhibitors with highly correlated IC50 values.
- **Methotrexate:** general cytotoxic drug not targeted to specific genes/pathways
- **Nilotinib:** inhibits the BCR-ABL fusion gene characteristic for chronic myeloid leukemia. Related to Axitinib (smaller effect)

Example with MADMMplasso: Real data



e Correlation structure of 7 drug response variables across 400 cell lines



f Test error

Example with MADMMplasso: Real data

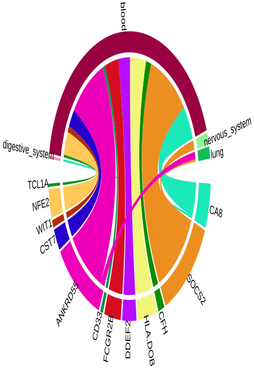
GDSC [Garnett et al., 2012]

Table: Results from the GDSC data.

Model	Non-zero coefficients[1]	Test error (SD)[2]
Plasso	724	3.648 (0.270)
Tree lasso	1016	3.404 (0.268)
MADMMplasso	1424	3.227 (0.267)

[1] The number of non-zero coefficients in the model. We counted the coefficients with at least two non-zero values across the 10 repeated data splits. Number of non-zero coefficients $= \sum_{j=1}^{p} \sum_{d=1}^{D} \{ (\sum_{r=1}^{10} \mathbf{1}_{\{\beta_{jd}^{r} \neq 0\}}) \geq 2 \}$
Note that the selection is out of $p \times D = 18844$ features in total.
[2] The MSE on an independent test data. We included the standard deviation (SD) across the 10 repeated data splits.

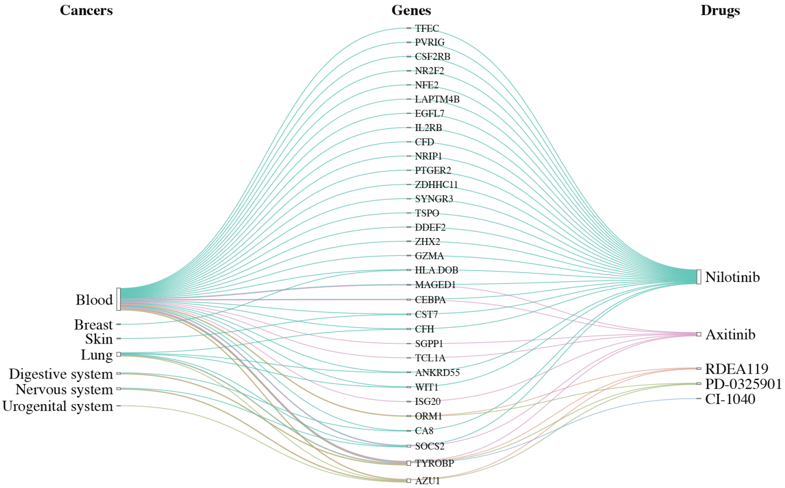
Example with MADMMplasso: Real data : Selected interaction effects for Nilotinib



Suppressor of cytokine signaling 2 (SOCS2) is involved in the signal transduction cascades in CML cells [Schultheis et al., 2002]

Example with MADMMplasso: Real data: Summary of all selected interaction effects

GDSC [Garnett et al., 2012]



Summary

- We have considered problems with hierarchical structures.
- The model involved main and interaction effects.
- The response cannot be explained by additive functions of the variables hence the need for hierarchical modeling.
- The procedure involved the implementation of the **pliable lasso penalty**.
- <u>Our extensions</u>
  - **Multi-response problem** with **tree-guided structure**.
  - The implementation of the **ADMM algorithm** made it possible to handle the overlapping groups in both the covariates and the responses.
  - The R package (**MADMMplasso**) is publicly available on https://github.com/ocbe-uio/MADMMplasso

Email: t.q.asenso@medisin.uio.no

This work received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie Actions Grant, agreement No. 80113 (Scientia fellowship)

References I

Bien, J., Taylor, J., and Tibshirani, R. (2013).
A lasso for hierarchical interactions.
The Annals of Statistics, 41(3):1111–1141.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011).
Distributed optimization and statistical learning via the alternating direction method of multipliers.
Foundations and Trends® in Machine learning, 3(1):1–122.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., et al. (2012).
Systematic identification of genomic markers of drug sensitivity in cancer cells.
Nature, 483(7391):570–575.

References II

Kim, S. and Xing, E. P. (2012).
Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping.
The Annals of Applied Statistics, 6(3):1095–1117.

Li, Y., Nan, B., and Zhu, J. (2015).
Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure.
Biometrics, 71(2):354–363.

Lim, M. and Hastie, T. (2015).
Learning interactions via hierarchical group-lasso regularization.
Journal of Computational and Graphical Statistics, 24(3):627–654.
PMID: 26759522.

References III

Schultheis, B., Carapeti-Marootian, M., Hochhaus, A., Weisser, A., Goldman, J. M., and Melo, J. V. (2002).

Overexpression of SOCS-2 in advanced stages of chronic myeloid leukemia: possible inadequacy of a negative feedback mechanism.

Blood, 99(5):1766–1775.

Tibshirani, R. and Friedman, J. (2020).

A pliable lasso.

Journal of Computational and Graphical Statistics, 29(1):215–225.

THANK YOU