# Inferring multiple graph structures

Julien Chiquet,
jointly with Christophe Ambroise, Camille Charbonnier,
Yves Grandvalet, Catherine Matias. . .

Laboratoire Statistique et Génome
UMR CNRS 8071, Université d'Évry Val d'Essonne & USC INRA

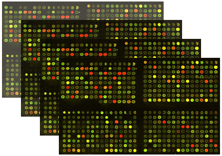
INRA Toulouse – 21 Janvier 2011

Chiquet, Grandvalet, Ambroise, *Statistics and Computing*, 2010.
Inferring multiple graphical structures.

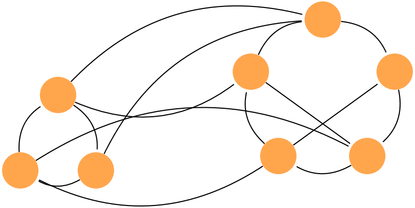
Chiquet, Grasseau, Charbonnier and Ambroise,
New release of R-package SIMoNe.

http://stat.genopole.cnrs.fr/softwares/simone

# Problem



Inference



few arrays ⇔ few examples
lots of genes ⇔ high dimension
interactions ⇔ very high dimension

Which interactions?

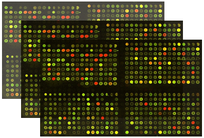
The main trouble is the low sample size and high dimensional setting

Our main hope is to benefit from sparsity: few genes interact

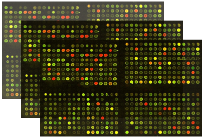
# Handling the scarcity of data

Merge several experimental conditions

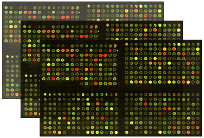
experiment 1



experiment 2

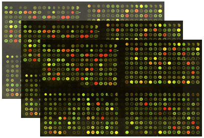


experiment 3



Inferring each graph independently does not help

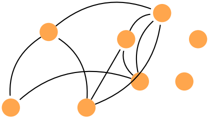
experiment 1



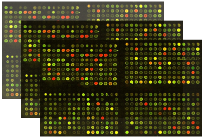
↓

$(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$

inference



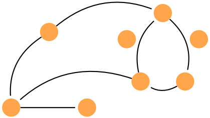
experiment 2



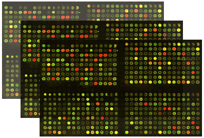
↓

$(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$

inference



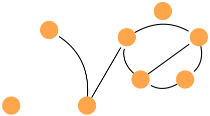
experiment 3



↓

$(X_1^{(3)}, \ldots, X_{n_3}^{(3)})$

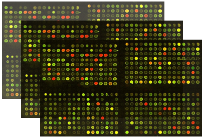
inference



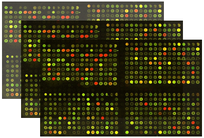
# Handling the scarcity of data

By pooling all the available data

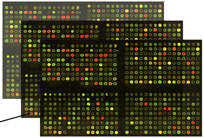
experiment 1



experiment 2

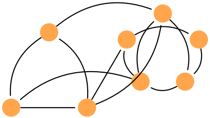


experiment 3



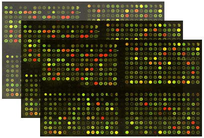
$$(X_1, \ldots, X_n), n = n_1 + n_2 + n_3.$$

inference



# Handling the scarcity of data

experiment 1

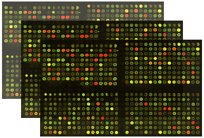


↓

$(X_1^{(1)}, \dots, X_{n_1}^{(1)})$

inference

experiment 2

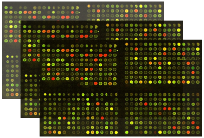


↓

$(X_1^{(2)}, \dots, X_{n_2}^{(2)})$

inference

experiment 3



↓

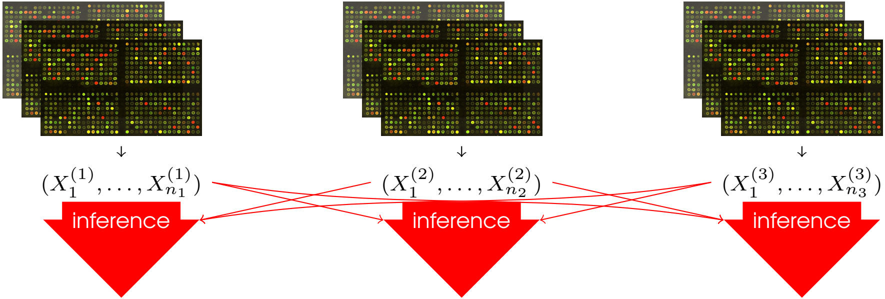
$(X_1^{(3)}, \dots, X_{n_3}^{(3)})$

inference

# Handling the scarcity of data

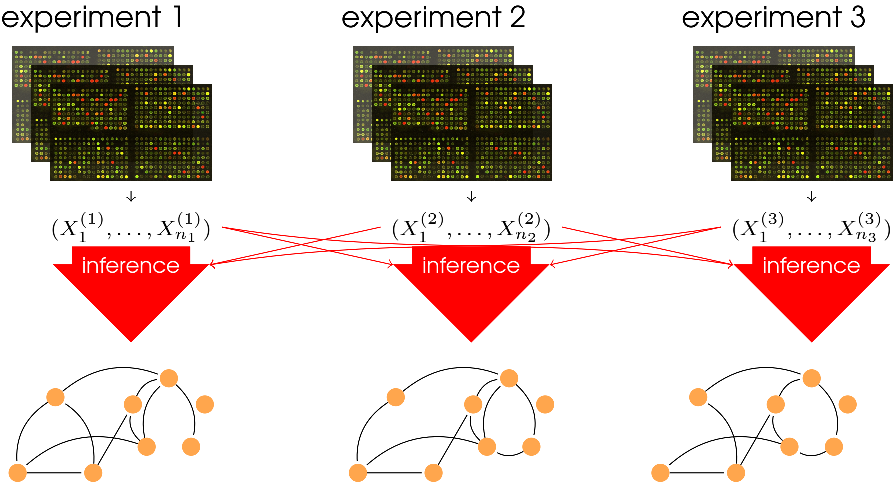
By breaking the separability

experiment 1 experiment 2 experiment 3



# Handling the scarcity of data

By breaking the separability



# Outline

Statistical model

Multi-task learning

Geometrical insights

Optimization strategy

Theoretical results

Experiments

# Outline

Statistical model

Multi-task learning

Geometrical insights

Optimization strategy

Theoretical results

Experiments

# Gaussian graphical modeling

Let

- $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$ and assume $n$ i.i.d. copies of $X$,
- $\mathbf{X}$ be the $n \times p$ matrix whose $k$th row is $X_k$,
- $\mathbf{\Theta} = (\theta_{ij})_{i,j \in \mathcal{P}} \triangleq \mathbf{\Sigma}^{-1}$ be the concentration matrix.

## Graphical interpretation

Since $\mathrm{cor}_{ij|\mathcal{P}\backslash\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$ for $i \neq j$,

$$X_i \perp\!\!\!\perp X_j | X_{\mathcal{P}\backslash\{i,j\}} \Leftrightarrow \begin{cases} \theta_{ij} = 0 \\ \text{or} \\ \text{edge } (i,j) \notin \text{ network.} \end{cases}$$

$\leadsto$ non zeroes in $\mathbf{\Theta}$ describes the graph structure.

# The model likelihood

Let $\mathbf{S} = n^{-1}\mathbf{X}^\intercal\mathbf{X}$ be the empirical variance-covariance matrix: $\mathbf{S}$ is a sufficient statistic for $\mathbf{X}$ $\Rightarrow \mathcal{L}(\boldsymbol{\Theta}; \mathbf{X}) = \mathcal{L}(\boldsymbol{\Theta}; \mathbf{S})$

## The log-likelihood

$$\mathcal{L}(\boldsymbol{\Theta}; \mathbf{S}) = \frac{n}{2} \log\det(\boldsymbol{\Theta}) - \frac{n}{2}\mathrm{trace}(\mathbf{S}\boldsymbol{\Theta}) - \frac{n}{2}\log(2\pi).$$

The MLE of $\boldsymbol{\Theta}$ is $\mathbf{S}^{-1}$

- not defined for $n < p$
- not sparse $\Rightarrow$ fully connected graph

# Penalized Approaches

## Penalized Likelihood (Banerjee *et al.*, 2008)

$$\underset{\boldsymbol{\Theta}\in\mathbb{S}_+}{\text{maximize}}\, \mathcal{L}(\boldsymbol{\Theta};\mathbf{S}) - \lambda\|\boldsymbol{\Theta}\|_1$$

- $\smile$ well defined for $n < p$
- $\smile$ sparse $\Rightarrow$ sensible graph
- $\frown$ SDP of size $\mathcal{O}(p^2)$ (solved by Friedman *et al.*, 2007)

## Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\arg\min} \frac{1}{n}\left\|\mathbf{X}_j - \mathbf{X}_{\backslash j}\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1$$

where $\mathbf{X}_j$ is the $j$th column of $\mathbf{X}$ and $\mathbf{X}_{\backslash j}$ is $\mathbf{X}$ deprived of $\mathbf{X}_j$

- $\frown$ not symmetric, not positive-definite
- $\smile$ $p$ independent LASSO problems of size $(p-1)$

# Penalized Approaches

## Penalized Likelihood (Banerjee *et al.*, 2008)

$$\underset{\boldsymbol{\Theta}\in\mathbb{S}_+}{\text{maximize}}\, \mathcal{L}(\boldsymbol{\Theta};\mathbf{S}) - \lambda\|\boldsymbol{\Theta}\|_1$$

- ◡ well defined for $n < p$
- ◡ sparse $\Rightarrow$ sensible graph
- ⌒ SDP of size $\mathcal{O}(p^2)$ (solved by Friedman *et al.*, 2007)

## Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \frac{1}{n}\left\|\mathbf{X}_j - \mathbf{X}_{\backslash j}\,\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1$$

where $\mathbf{X}_j$ is the $j$th column of $\mathbf{X}$ and $\mathbf{X}_{\backslash j}$ is $\mathbf{X}$ deprived of $\mathbf{X}_j$

- ⌒ not symmetric, not positive-definite
- ◡ $p$ independent LASSO problems of size $(p-1)$

# Neighborhood *vs.* Likelihood

## Pseudo-likelihood (Besag, 1975)

$$\mathbb{P}(X_1, \dots, X_p) \simeq \prod_{j=1}^{p} \mathbb{P}(X_j | \{X_k\}_{k \neq j})$$

$$\widetilde{\mathcal{L}}(\mathbf{\Theta}; \mathbf{S}) = \frac{n}{2} \log \det(\mathbf{D}) - \frac{n}{2} \mathrm{trace}\left(\mathbf{S}\mathbf{D}^{-1}\mathbf{\Theta}^2\right) - \frac{n}{2}\log(2\pi)$$

$$\mathcal{L}(\mathbf{\Theta}; \mathbf{S}) = \frac{n}{2} \log \det(\mathbf{\Theta}) - \frac{n}{2} \mathrm{trace}(\mathbf{S}\mathbf{\Theta}) - \frac{n}{2}\log(2\pi)$$

with $\mathbf{D} = \mathrm{diag}(\mathbf{\Theta})$.

## Proposition (Ambroise, Chiquet, Matias, 2008)

*Neighborhood selection leads to the graph maximizing the penalized pseudo-log-likelihood*

Proof: $\hat{\beta}_i = -\frac{\widehat{\theta}_{ij}}{\widehat{\theta}_{jj}}$, where $\widehat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \widetilde{\mathcal{L}}(\mathbf{\Theta}; \mathbf{S}) - \lambda \|\mathbf{\Theta}\|_1$

# Neighborhood *vs.* Likelihood

## Pseudo-likelihood (Besag, 1975)

$$\mathbb{P}(X_1, \dots, X_p) \simeq \prod_{j=1}^{p} \mathbb{P}(X_j | \{X_k\}_{k \neq j})$$

$$\widetilde{\mathcal{L}}(\boldsymbol\Theta; \mathbf{S}) = \frac{n}{2} \log\det(\mathbf{D}) - \frac{n}{2} \mathrm{trace}\left(\mathbf{S}\mathbf{D}^{-1}\boldsymbol\Theta^2\right) - \frac{n}{2}\log(2\pi)$$

$$\mathcal{L}(\boldsymbol\Theta; \mathbf{S}) = \frac{n}{2} \log\det(\boldsymbol\Theta) - \frac{n}{2} \mathrm{trace}(\mathbf{S}\boldsymbol\Theta) - \frac{n}{2}\log(2\pi)$$

with $\mathbf{D} = \mathrm{diag}(\boldsymbol\Theta)$.

## Proposition (Ambroise, Chiquet, Matias, 2008)

*Neighborhood selection leads to the graph maximizing the penalized pseudo-log-likelihood*

Proof: $\hat\beta_i = -\frac{\widehat\theta_{ij}}{\widehat\theta_{jj}}$, where $\widehat{\boldsymbol\Theta} = \arg\max_{\boldsymbol\Theta} \tilde{\mathcal{L}}(\boldsymbol\Theta; \mathbf{S}) - \lambda\|\boldsymbol\Theta\|_1$

# Outline

# Multi-task learning

We have $T$ samples (experimental cond.) of the same variables

- $\mathbf{X}^{(t)}$ is the $t^{\text{th}}$ data matrix, $\mathbf{S}^{(t)}$ is the empirical covariance
- examples are assumed to be drawn from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(t)})$

Ignoring the relationships between the tasks leads to separable objectives

$$\underset{\mathbf{\Theta}^{(t)}\in\mathbb{R}^{p\times p}, t=1\dots,T}{\text{maximize}} \ \widetilde{\mathcal{L}}(\mathbf{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda\|\mathbf{\Theta}^{(t)}\|_1$$

Multi-task learning = solving the $T$ tasks jointly

We may **couple** the objectives

- through the fitting term term,
- through the penalty term.

# Multi-task learning

We have $T$ samples (experimental cond.) of the same variables

- $\mathbf{X}^{(t)}$ is the $t^{\text{th}}$ data matrix, $\mathbf{S}^{(t)}$ is the empirical covariance
- examples are assumed to be drawn from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(t)})$

Ignoring the relationships between the tasks leads to separable objectives

$$\underset{\mathbf{\Theta}^{(t)} \in \mathbb{R}^{p \times p}, t=1\dots,T}{\text{maximize}} \quad \widetilde{\mathcal{L}}(\mathbf{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda \|\mathbf{\Theta}^{(t)}\|_1$$

Multi-task learning $=$ solving the $T$ tasks jointly

We may **couple** the objectives

- through the fitting term term,
- through the penalty term.

# Coupling through the fitting term

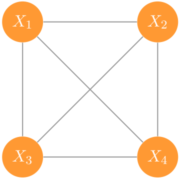
## Intertwined LASSO

$$\underset{\boldsymbol{\Theta}^{(t)},t\ldots,T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)}; \widetilde{\mathbf{S}}^{(t)}) - \lambda \|\boldsymbol{\Theta}^{(t)}\|_1$$

- $\overline{\mathbf{S}} = \frac{1}{n}\sum_{t=1}^{T} n_t \mathbf{S}^{(t)}$ is the "pooled-tasks" covariance matrix.
- $\widetilde{\mathbf{S}}^{(t)} = \alpha \mathbf{S}^{(t)} + (1-\alpha)\overline{\mathbf{S}}$ is a mixture between specific and pooled covariance matrices.

- $\alpha = 0$ pools the data sets and infers a single graph
- $\alpha = 1$ separates the data sets and infers $T$ graphs independently
- $\alpha = 1/2$ in all our experiments

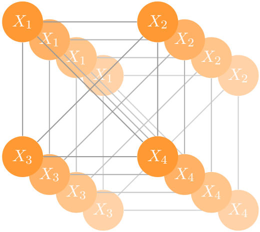
We group parameters by sets of corresponding edges across graphs:




## Graphical group-LASSO

$$\underset{\boldsymbol{\Theta}^{(t)}, t\dots,T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}\left(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}\right) - \lambda \sum_{i \neq j} \left( \sum_{t=1}^{T} \left(\theta_{ij}^{(t)}\right)^2 \right)^{1/2}$$

- ◡ Sparsity pattern shared between graphs
- ◠ Identical graphs across tasks




We group parameters by sets of corresponding edges across graphs:

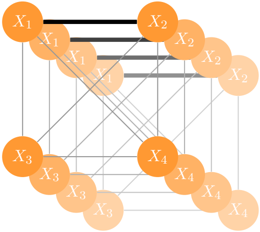
## Graphical group-LASSO

$$\underset{\boldsymbol{\Theta}^{(t)},t\dots,T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}\left(\boldsymbol{\Theta}^{(t)};\mathbf{S}^{(t)}\right) - \lambda \sum_{i\neq j} \left( \sum_{t=1}^{T} \left(\theta_{ij}^{(t)}\right)^2 \right)^{1/2}$$

- ⌣ Sparsity pattern shared between graphs
- ⌢ Identical graphs across tasks

# Coupling through penalties: group-LASSO

We group parameters by sets of corresponding edges across graphs:

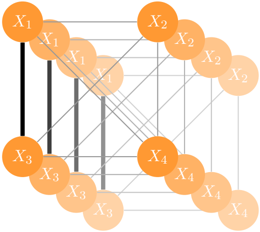



## Graphical group-LASSO

$$\underset{\boldsymbol{\Theta}^{(t)},t\ldots,T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}\left(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}\right) - \lambda \sum_{i \neq j} \left( \sum_{t=1}^{T} \left(\theta_{ij}^{(t)}\right)^2 \right)^{1/2}$$

- ◡ Sparsity pattern shared between graphs
- ◠ Identical graphs across tasks

We group parameters by sets of corresponding edges across graphs:



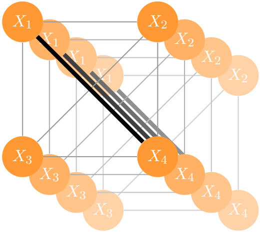

## Graphical group-LASSO

$$\underset{\boldsymbol{\Theta}^{(t)}, t \dots, T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}\left(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}\right) - \lambda \sum_{i \neq j} \left( \sum_{t=1}^{T} \left(\theta_{ij}^{(t)}\right)^2 \right)^{1/2}$$

- ◡ Sparsity pattern shared between graphs
- ◠ Identical graphs across tasks

# Coupling through penalties: group-LASSO

We group parameters by sets of corresponding edges across graphs:



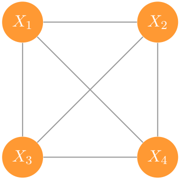

## Graphical group-LASSO

$$\underset{\boldsymbol{\Theta}^{(t)},t\dots,T}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}\left(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}\right) - \lambda \sum_{i\neq j} \left( \sum_{t=1}^{T} \left(\theta_{ij}^{(t)}\right)^2 \right)^{1/2}$$

- ◡ Sparsity pattern shared between graphs
- ◠ Identical graphs across tasks

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

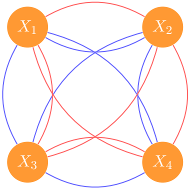
$$\underset{\substack{\boldsymbol{\Theta}^{(t)} \\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_+^2 \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_-^2 \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- Plausible in many other situations
- Sparsity pattern shared between graphs, which may differ

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

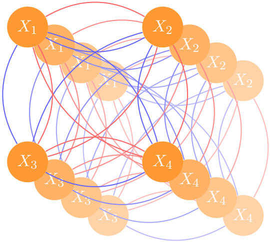
$$\underset{\substack{\boldsymbol{\Theta}^{(t)}\\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{+}^{2} \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{-}^{2} \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- ◡ Plausible in many other situations
- ◡ Sparsity pattern shared between graphs, which may differ

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

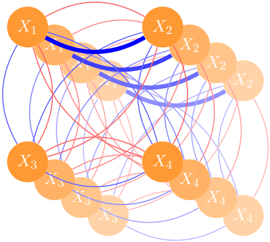
$$\underset{\substack{\boldsymbol{\Theta}^{(t)}\\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)};\boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i\neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_+^2 \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_-^2 \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0,u)$ and $[u]_- = \min(0,u)$.

- Plausible in many other situations
- Sparsity pattern shared between graphs, which may differ

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays



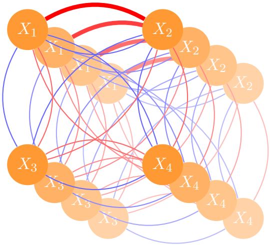

## Graphical cooperative-LASSO

$$\underset{\substack{\boldsymbol{\Theta}^{(t)} \\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_+^2 \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_-^2 \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- Plausible in many other situations
- Sparsity pattern shared between graphs, which may differ

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

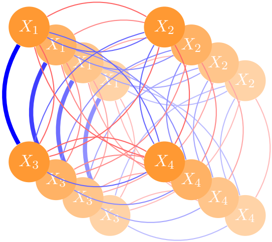
$$\underset{\substack{\boldsymbol{\Theta}^{(t)} \\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{+}^{2} \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{-}^{2} \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- Plausible in many other situations
- Sparsity pattern shared between graphs, which may differ

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

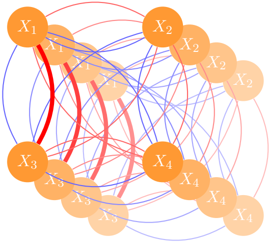
$$\underset{\substack{\boldsymbol{\Theta}^{(t)} \\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_+^2 \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_-^2 \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- ⌣ Plausible in many other situations
- ⌣ Sparsity pattern shared between graphs, which may differ

# Coupling through penalties: cooperative-LASSO

- Same grouping, and bet that correlations are likely to be sign consistent
- Gene interactions are either inhibitory or activating across assays




## Graphical cooperative-LASSO

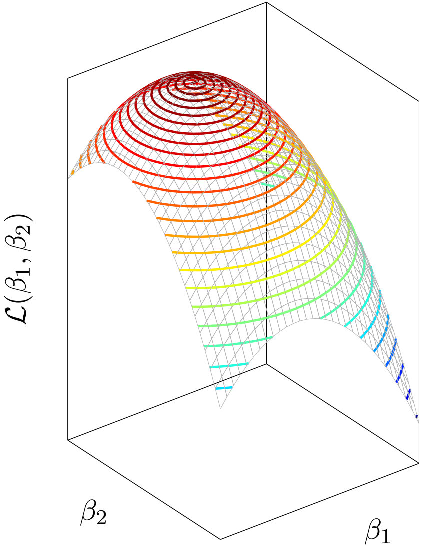
$$\underset{\substack{\boldsymbol{\Theta}^{(t)} \\ t=1,\dots,T}}{\text{maximize}} \sum_{t=1}^{T} \widetilde{\mathcal{L}}(\mathbf{S}^{(t)}; \boldsymbol{\Theta}^{(t)}) - \lambda \sum_{i \neq j} \left\{ \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{+}^{2} \right)^{\frac{1}{2}} + \left( \sum_{t=1}^{T} \left[\theta_{ij}^{(t)}\right]_{-}^{2} \right)^{\frac{1}{2}} \right\}$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

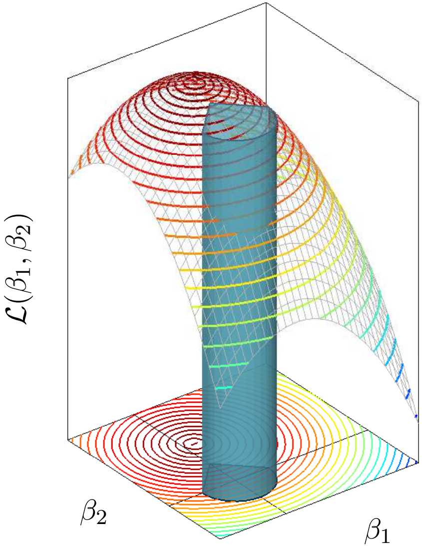
- ☺ Plausible in many other situations
- ☺ Sparsity pattern shared between graphs, which may differ

# Outline

# A Geometric View of Sparsity
## Constrained Optimization

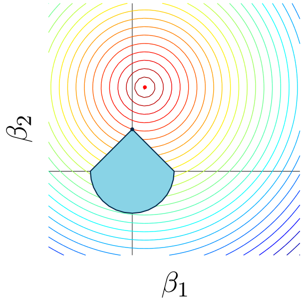



$$\max_{\beta_1, \beta_2} \mathcal{L}(\beta_1, \beta_2) - \lambda \Omega(\beta_1, \beta_2)$$




$$\max_{\beta_1,\beta_2} \mathcal{L}(\beta_1, \beta_2) - \lambda\Omega(\beta_1, \beta_2)$$

$$\Leftrightarrow \begin{cases} \max_{\beta_1,\beta_2} & \mathcal{L}(\beta_1, \beta_2) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases}$$

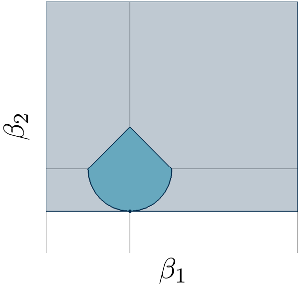



$$\max_{\beta_1,\beta_2} \mathcal{L}(\beta_1,\beta_2) - \lambda\Omega(\beta_1,\beta_2)$$

$$\Leftrightarrow \quad \begin{cases} \max_{\beta_1,\beta_2} & \mathcal{L}(\beta_1,\beta_2) \\ \text{s.t.} & \Omega(\beta_1,\beta_2) \leq c \end{cases}$$

An hyperplane supports a set iff

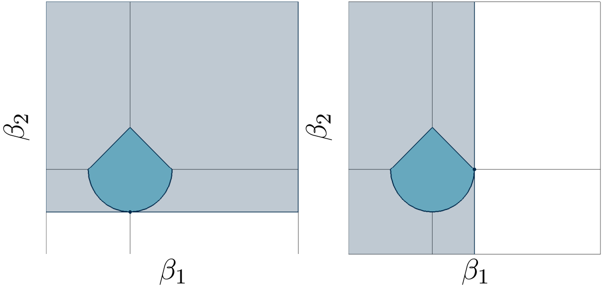
- the set is contained in one half-space
- the set has at least one point on the hyperplane



# A Geometric View of Sparsity
## Supporting Hyperplane

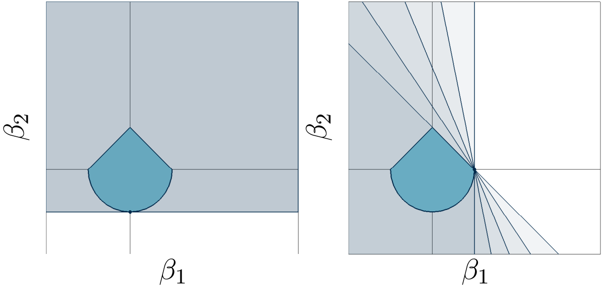
An hyperplane supports a set iff

- the set is contained in one half-space
- the set has at least one point on the hyperplane



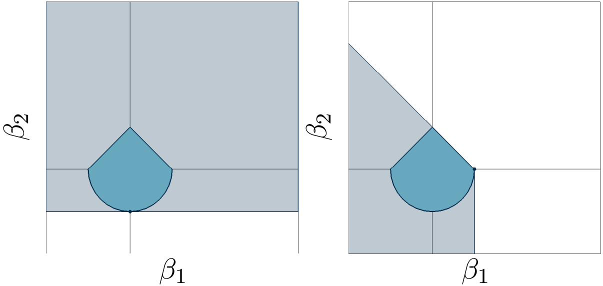
An hyperplane supports a set iff

- the set is contained in one half-space
- the set has at least one point on the hyperplane



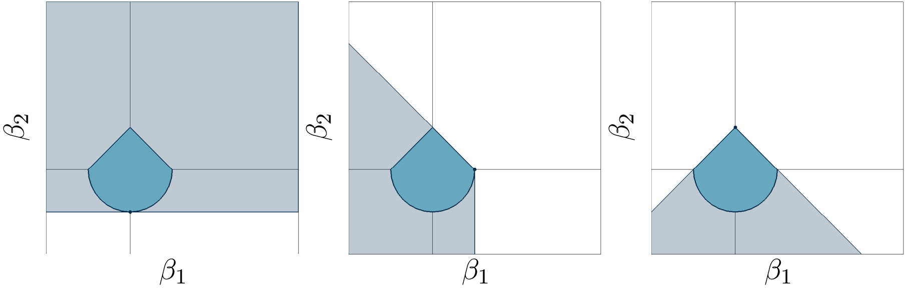
An hyperplane supports a set iff

- the set is contained in one half-space
- the set has at least one point on the hyperplane



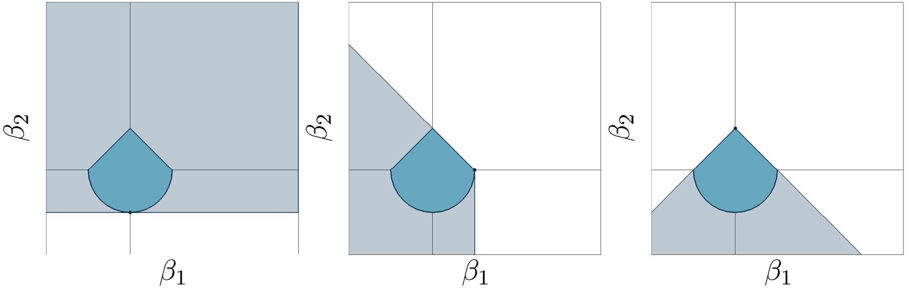
An hyperplane supports a set iff

- the set is contained in one half-space
- the set has at least one point on the hyperplane

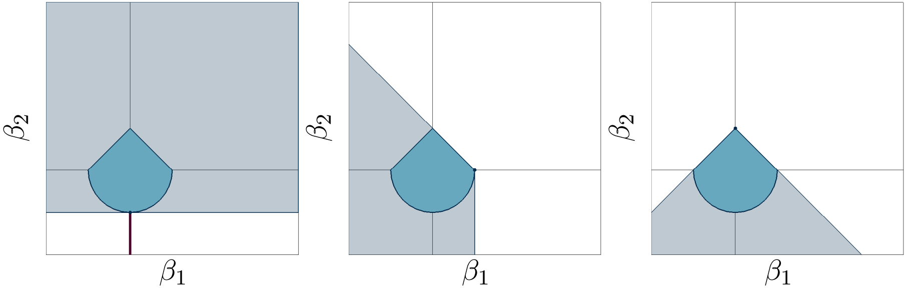



There are Supporting Hyperplane at all points of convex sets: Generalize tangents

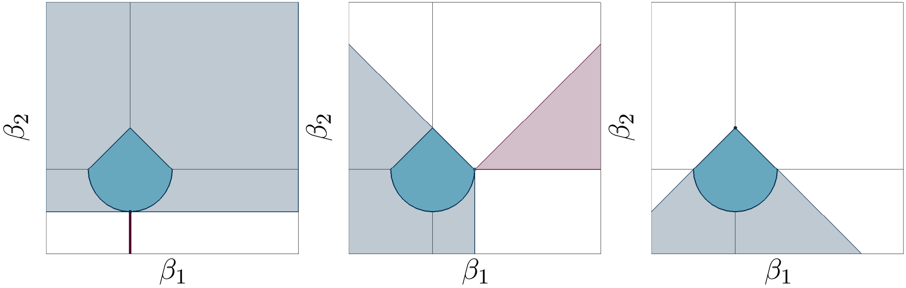
Generalizes normals



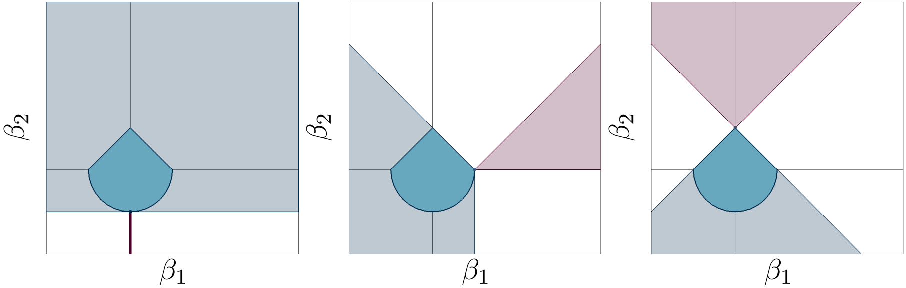
Generalizes normals



Generalizes normals



Generalizes normals




Shape of dual cones $\Rightarrow$ sparsity pattern

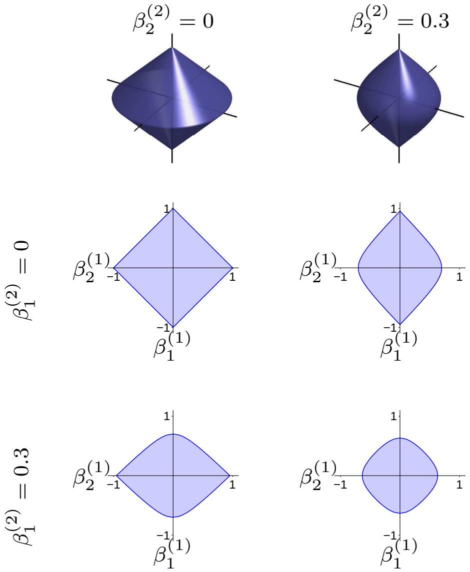
# Group-LASSO balls

**Admissible set**

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

**Unit ball**

$$\sum_{i=1}^{2} \left( \sum_{t=1}^{2} {\beta_i^{(t)}}^2 \right)^{1/2} \leq 1$$



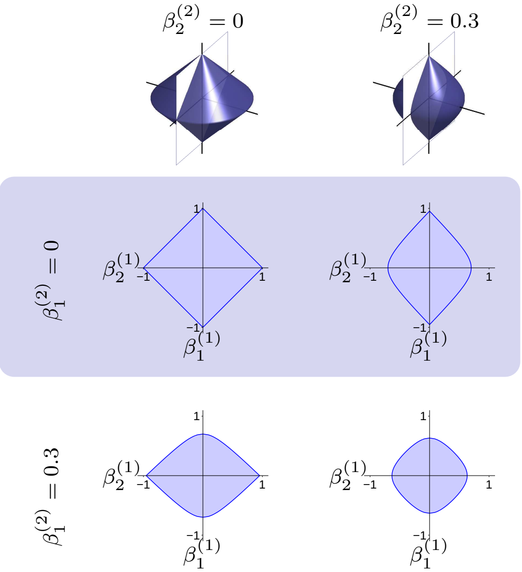
# Group-LASSO balls

**Admissible set**

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

**Unit ball**

$$\sum_{i=1}^{2} \left( \sum_{t=1}^{2} {\beta_i^{(t)}}^2 \right)^{1/2} \leq 1$$



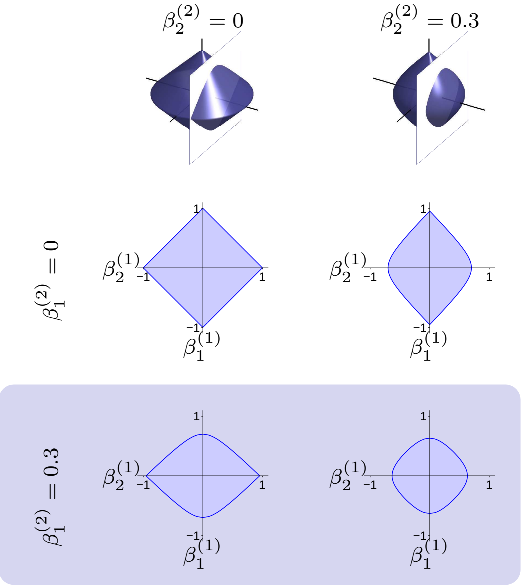
# Group-LASSO balls

**Admissible set**

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

**Unit ball**

$$\sum_{i=1}^{2} \left( \sum_{t=1}^{2} {\beta_i^{(t)}}^2 \right)^{1/2} \leq 1$$



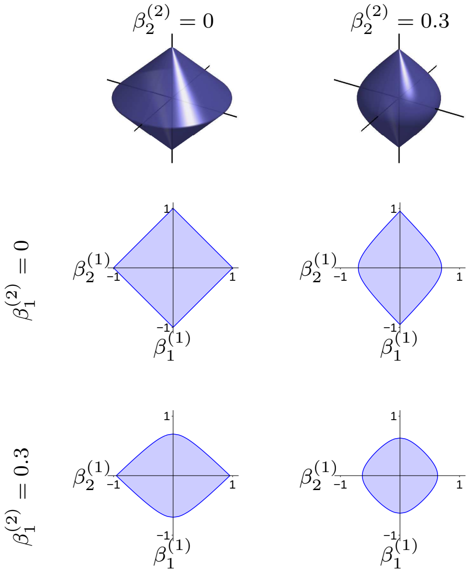
# Group-LASSO balls

Admissible set

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

Unit ball

$$\sum_{i=1}^{2} \left( \sum_{t=1}^{2} \beta_i^{(t)^2} \right)^{1/2} \leq 1$$



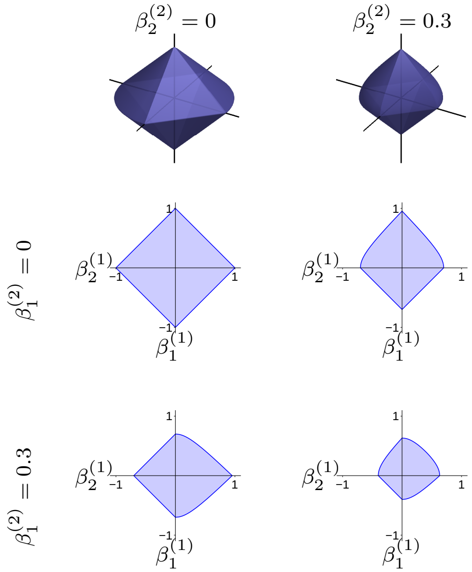
# Cooperative-LASSO balls

**Admissible set**

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

**Unit ball**

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$



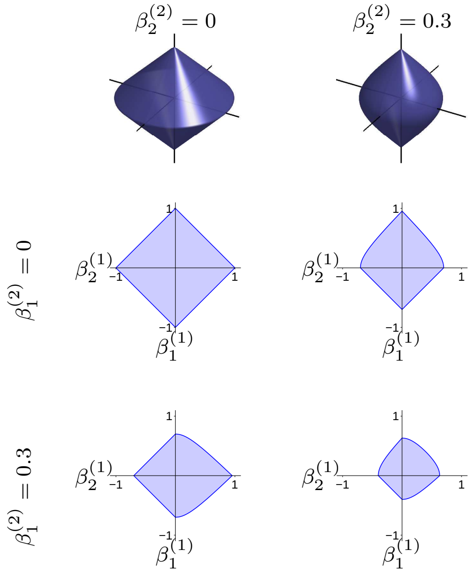
# Cooperative-LASSO balls

## Admissible set

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

## Unit ball

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$



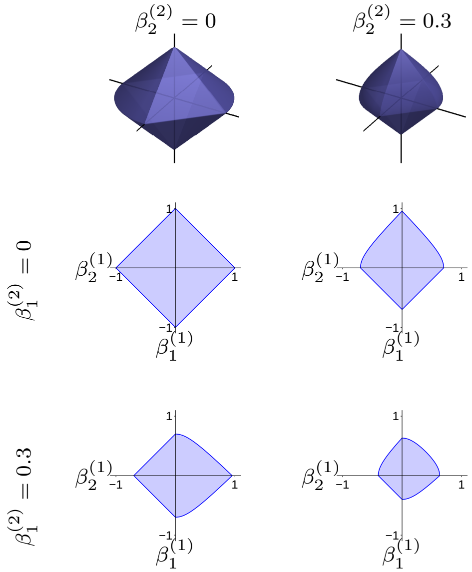
# Cooperative-LASSO balls

**Admissible set**

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

**Unit ball**

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$

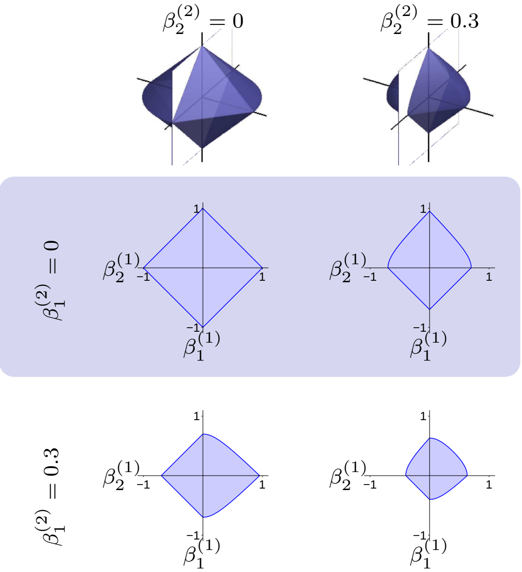


## Admissible set

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

## Unit ball

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$

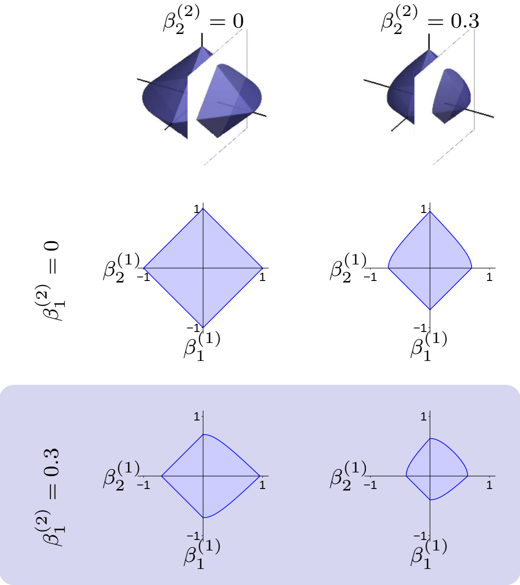


## Admissible set

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

## Unit ball

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$

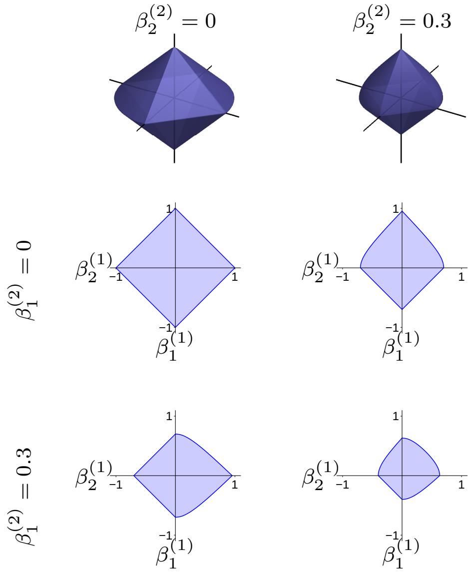


## Admissible set

- 2 tasks ($T = 2$)
- 2 coefficients ($p = 2$)

## Unit ball

$$\sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(\beta_j^{(t)}\right)_+^2\right)^{1/2} + \sum_{j=1}^{2}\left(\sum_{t=1}^{2}\left(-\beta_j^{(t)}\right)_+^2\right)^{1/2} \leq 1$$



# Outline

Statistical model

Multi-task learning

Geometrical insights

Optimization strategy

Theoretical results

Experiments

# Decomposition strategy

## Estimate the $j^{\text{th}}$ neighborhood of the $T$ graphs

$$\underset{\boldsymbol{\Theta}^{(t)},t=1\dots,T}{\text{maximize}} \sum_{t=1}^{T} \tilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda\, \Omega(\mathbf{K}^{(t)})$$

decomposes into $p$ convex optimization problems of size

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{T\times(p-1)}}{\arg\min}\ f_j(\boldsymbol{\beta}) + \lambda\, \Omega(\boldsymbol{\beta})$$

where $\widehat{\boldsymbol{\beta}}_j$ is a minimizer iff $0 \in \nabla_{\boldsymbol{\beta}} f_j(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta})$

# Decomposition strategy

## Estimate the $j^{\text{th}}$ neighborhood of the $T$ graphs

$$\underset{\boldsymbol{\Theta}^{(t)},t=1\dots,T}{\text{maximize}} \sum_{t=1}^{T} \tilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda\, \Omega(\mathbf{K}^{(t)})$$

decomposes into $p$ convex optimization problems of size

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{T\times(p-1)}}{\arg\min}\ f_j(\boldsymbol{\beta}) + \lambda\, \Omega(\boldsymbol{\beta})$$

where $\widehat{\boldsymbol{\beta}}_j$ is a minimizer iff $0 \in \nabla_{\boldsymbol{\beta}} f_j(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta})$
Intertwined LASSO:

$$\Omega(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left\| \boldsymbol{\beta}^{(t)} \right\|_1 ,$$

where $\boldsymbol{\beta} = \left( \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(T)} \right)^{\intercal}$, $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^{p-1}$

# Decomposition strategy

## Estimate the $j^{\text{th}}$ neighborhood of the $T$ graphs

$$\underset{\boldsymbol{\Theta}^{(t)},t=1\dots,T}{\text{maximize}} \sum_{t=1}^{T} \tilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda\, \Omega(\mathbf{K}^{(t)})$$

decomposes into $p$ convex optimization problems of size

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{T\times(p-1)}}{\text{argmin}}\ f_j(\boldsymbol{\beta}) + \lambda\, \Omega(\boldsymbol{\beta})$$

where $\widehat{\boldsymbol{\beta}}_j$ is a minimizer iff $0 \in \nabla_{\boldsymbol{\beta}} f_j(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta})$
Group-LASSO:

$$\Omega(\boldsymbol{\beta}) = \sum_{i=1}^{p-1} \left\| \boldsymbol{\beta}_i^{[1:T]} \right\|_2$$

where $\boldsymbol{\beta}_i^{[1:T]}$ is the vector corresponding to the edges $(i, j)$ across graphs

$$\underset{\boldsymbol{\Theta}^{(t)},t=1\ldots,T}{\text{maximize}} \sum_{t=1}^{T} \tilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)};\mathbf{S}^{(t)}) - \lambda\,\Omega(\mathbf{K}^{(t)})$$

decomposes into $p$ convex optimization problems of size

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{T\times(p-1)}}{\arg\min}\ f_j(\boldsymbol{\beta}) + \lambda\,\Omega(\boldsymbol{\beta})$$

where $\widehat{\boldsymbol{\beta}}_j$ is a minimizer iff $0 \in \nabla_{\boldsymbol{\beta}} f_j(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta})$
Coop-Lasso:

$$\Omega(\boldsymbol{\beta}) = \sum_{i=1}^{p-1} \left( \left\| \left( \boldsymbol{\beta}_i^{[1:T]} \right)_+ \right\|_2 + \left\| \left( -\boldsymbol{\beta}_i^{[1:T]} \right)_+ \right\|_2 \right)$$

where $\boldsymbol{\beta}_i^{[1:T]}$ is the vector corresponding to the edges $(i,j)$ across graphs

# Active set algorithm: ☺ yellow belt

// 0. INITIALIZATION $\boldsymbol{\beta} \leftarrow \mathbf{0}, \mathcal{A} \leftarrow \emptyset$

**while** $\mathbf{0} \notin \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ **do**

// 1. MASTER PROBLEM: OPTIMIZATION WITH RESPECT TO $\boldsymbol{\beta}_{\mathcal{A}}$

Find a solution $\mathbf{h}$ to the smooth problem

$$\nabla_{\mathbf{h}} f(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) + \lambda \partial_{\mathbf{h}} \Omega(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) = 0, \quad \text{where } \partial_{\mathbf{h}} \Omega = \{\nabla_{\mathbf{h}} \Omega\} .$$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}$

// 2. IDENTIFY NEWLY ZEROED VARIABLES;

$\mathcal{A} \leftarrow \mathcal{A} \backslash \{i\}$

// 3. IDENTIFY NEW NON-ZERO VARIABLES;

// Select a candidate $i \in \mathcal{A}^c$

$$i \leftarrow \arg\max_{j \in \mathcal{A}^c} v_j, \text{ where } v_j = \min_{\nu \in \partial_{\beta_j} \Omega} \left| \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \nu \right|$$

**end**

# Active set algorithm: ☹ orange belt

// 0. INITIALIZATION $\boldsymbol{\beta} \leftarrow \mathbf{0}, \mathcal{A} \leftarrow \emptyset$

**while** $\mathbf{0} \notin \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ **do**

// 1. MASTER PROBLEM: OPTIMIZATION WITH RESPECT TO $\boldsymbol{\beta}_{\mathcal{A}}$

Find a solution $\mathbf{h}$ to the smooth problem

$$\nabla_{\mathbf{h}} f(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) + \lambda \partial_{\mathbf{h}} \Omega(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) = 0, \quad \text{where } \partial_{\mathbf{h}} \Omega = \{\nabla_{\mathbf{h}} \Omega\} .$$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}$

// 2. IDENTIFY NEWLY ZEROED VARIABLES;

$\mathcal{A} \leftarrow \mathcal{A} \backslash \{i\}$

// 3. IDENTIFY NEW NON-ZERO VARIABLES;

// Select a candidate $i \in \mathcal{A}^c$ which violates the more the optimality conditions

$$i \leftarrow \arg\max_{j \in \mathcal{A}^c} v_j, \text{ where } v_j = \min_{\nu \in \partial_{\beta_j} \Omega} \left| \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \nu \right|$$

**if** *it exists such an* $i$ **then**

$\mathcal{A} \leftarrow \mathcal{A} \cup \{i\}$

**else**

Stop and return $\boldsymbol{\beta}$, which is optimal

**end**

**end**

# Active set algorithm: ☹ green belt

// 0. INITIALIZATION $\boldsymbol{\beta} \leftarrow \mathbf{0}, \mathcal{A} \leftarrow \emptyset$

**while** $0 \notin \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ **do**

// 1. MASTER PROBLEM: OPTIMIZATION WITH RESPECT TO $\boldsymbol{\beta}_{\mathcal{A}}$

Find a solution $\mathbf{h}$ to the smooth problem

$$\nabla_{\mathbf{h}} f(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) + \lambda \partial_{\mathbf{h}} \Omega(\boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}) = 0, \quad \text{where } \partial_{\mathbf{h}} \Omega = \{\nabla_{\mathbf{h}} \Omega\} \ .$$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{h}$

// 2. IDENTIFY NEWLY ZEROED VARIABLES;

**while** $\exists i \in \mathcal{A} : \beta_i = 0$ *and* $\min_{\nu \in \partial_{\beta_i} \Omega} \left| \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_i} + \lambda \nu \right| = 0$ **do**

| $\mathcal{A} \leftarrow \mathcal{A} \backslash \{i\}$

**end**

// 3. IDENTIFY NEW NON-ZERO VARIABLES;

// Select a candidate $i \in \mathcal{A}^c$ such that an infinitesimal change of $\beta_i$ provides the highest reduction of $L$

$i \leftarrow \arg\max_{j \in \mathcal{A}^c} v_j$, where $v_j = \min_{\nu \in \partial_{\beta_j} \Omega} \left| \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_j} + \lambda \nu \right|$

**if** $v_i \neq 0$ **then**

| $\mathcal{A} \leftarrow \mathcal{A} \cup \{i\}$

**else**

| Stop and return $\boldsymbol{\beta}$, which is optimal

**end**

**end**

# Outline

# (Sparse) linear regression setup

Let $Y$ be a response variable, $X = (X_1, \dots, X_p)$ a vector of $p$ features,

$$Y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon} = \sum_{j=1}^{p} X_j \beta_j^\star + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \ ,$$

- $\mathcal{S} = \{j, \boldsymbol{\beta}_j^\star \neq 0\}$ is the true support,
- $\boldsymbol{\beta}^\star$ has a group structure $\{\mathcal{G}_k\}_{k=1,\dots,K}$.

## Cooperative-Lasso estimate of $\beta^\star$

Given the training vector $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ and the $n \times p$ design matrix $\mathbf{X}$ whose $j$th column $\mathbf{x}_j = (x_j^1, \dots, x_j^n)^\intercal$,

$$\hat{\beta}^{\mathrm{coop}} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_n^2 + \lambda_n \sum_{k=1}^{K} \|[\boldsymbol{\beta}_{\mathcal{G}_k}]_+\| + \|[\boldsymbol{\beta}_{\mathcal{G}_k}]_-\|. \ ,$$

# (Sparse) linear regression setup

Let $Y$ be a response variable, $X = (X_1, \dots, X_p)$ a vector of $p$ features,

$$Y = X\boldsymbol{\beta}^\star + \varepsilon = \sum_{j=1}^{p} X_j \beta_j^\star + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \ ,$$

- $\mathcal{S} = \{j, \boldsymbol{\beta}_j^\star \neq 0\}$ is the true support,
- $\boldsymbol{\beta}^\star$ has a group structure $\{\mathcal{G}_k\}_{k=1,\dots,K}$.

## Cooperative-Lasso estimate of $\boldsymbol{\beta}^\star$

Given the training vector $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ and the $n \times p$ design matrix $\mathbf{X}$ whose $j$th column $\mathbf{x}_j = (x_j^1, \dots, x_j^n)^\intercal$,

$$\hat{\boldsymbol{\beta}}^{\text{coop}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \sum_{k=1}^{K} \|[\boldsymbol{\beta}_{\mathcal{G}_k}]_+\| + \|[\boldsymbol{\beta}_{\mathcal{G}_k}]_-\|. \ ,$$

# Technical assumptions

Let $\Psi = \mathbb{E} X X^\intercal$ be the covariance matrix of $X$.

(A1) $X$ and $Y$ have finite fourth order moments $\mathbb{E}\|X\|^4 < \infty$, $\mathbb{E}\|Y\|^4 < \infty$,

(A2) the covariance matrix $\Psi = \mathbb{E} X X^\intercal \in \mathbb{R}^{p \times p}$ is invertible,

(A3) for every $k = 1, \dots, K$, if $\|[\boldsymbol{\beta}^\star]_+\| > 0$ and $\|[\boldsymbol{\beta}^\star]_-\| > 0$ then for every $j \in \mathcal{G}_k$ $\boldsymbol{\beta}^\star_j \neq 0$. *(There should not be any zero in a group with positive and negative coefficients).*

# Irrepresentability condition

Define $\mathcal{S}_k = \mathcal{S} \cap \mathcal{G}_k$ the support within a group and

$$D(\boldsymbol{\beta})]_{jj} = \|[\mathrm{sign}(\beta_j)\boldsymbol{\beta}_{\mathcal{G}_k}]_+\|^{-1}.$$

Assume there exists $\eta > 0$ such that

- for every group $k$ to switch off (where $\mathcal{S}_k^c = \mathcal{G}_k$),

$$\max(\|[\Psi_{\mathcal{S}_k^c\mathcal{S}}\Psi_{\mathcal{S}\mathcal{S}}^{-1}D(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star]_+\|, \|[\Psi_{\mathcal{S}_k^c\mathcal{S}}\Psi_{\mathcal{S}\mathcal{S}}^{-1}D(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star]_-\|) \leq 1-\eta,$$

- for every group $k$ with zero coefficients and either positive or negative coefficients, define $\nu_k = 1$ if positive coefficients are activated, $\nu_k = -1$ otherwise, and require

$$\begin{cases} \nu_k \Psi_{\mathcal{S}_k^c\mathcal{S}}\Psi_{\mathcal{S}\mathcal{S}}^{-1}D(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star \leq 0 & \text{component-wise} \\ \|\Psi_{\mathcal{S}_k^c\mathcal{S}}\Psi_{\mathcal{S}\mathcal{S}}^{-1}D(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star\| \leq 1-\eta. & \end{cases}$$

# Consistency results

**Theorem** (Chiquet, Grandvalet, Charbonnier, in progress!)

*If assumptions (A1-3) are satisfied and if there exists $\eta > 0$, then for every sequence $\lambda_n$ such that $\lambda_n = \lambda_0 n^{-\gamma},\ \gamma \in ]0, 1/2[$,*

$$\hat{\boldsymbol{\beta}}^{\text{coop}} \xrightarrow{P} \boldsymbol{\beta}^\star \quad \textit{and} \quad \mathbb{P}(\mathcal{S}(\hat{\boldsymbol{\beta}}^{\text{coop}}) = \mathcal{S}) \to 1. \tag{1}$$

Asymptotically, the cooperative-Lasso is unbiased and enjoys exact support recovery (even when there are irrelevant variables within a group $\mathcal{G}_k$).

# Outline

Statistical model

Multi-task learning

Geometrical insights

Optimization strategy

Theoretical results

Experiments

# Data Generation

We set

- the number of nodes $p$
- the number of edges $K$
- the number of examples $n$

Process

1. Generate a random adjacency matrix with $2\,K$ off-diagonal terms
2. Compute the normalized Laplacian $\mathbf{L}$
3. Generate a symmetric matrix of random signs $\mathbf{R}$
4. Compute the concentration matrix $\Theta^{\star}_{ij} = L_{ij}\,R_{ij}$
5. compute $\mathbf{\Sigma}^{\star}$ by pseudo-inversion of $\mathbf{\Theta}^{\star}$
6. generate correlated Gaussian data $\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{\star})$

# Simulating Related Tasks

Generate

1. an "ancestor" with $p = 20$ nodes and $K = 20$ edges
2. $T = 4$ children by adding *and* deleting $\delta$ edges
3. $T = 4$ Gaussian samples

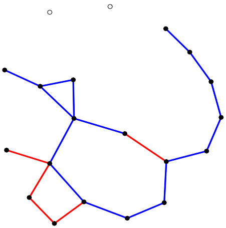


Figure: ancestor and children with $\delta = 2$ perturbations

Generate

1. an "ancestor" with $p = 20$ nodes and $K = 20$ edges
2. $T = 4$ children by adding *and* deleting $\delta$ edges
3. $T = 4$ Gaussian samples

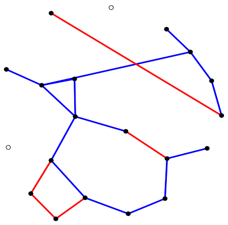


Figure: ancestor and children with $\delta = 2$ perturbations

Generate

1. an "ancestor" with $p = 20$ nodes and $K = 20$ edges
2. $T = 4$ children by adding *and* deleting $\delta$ edges
3. $T = 4$ Gaussian samples

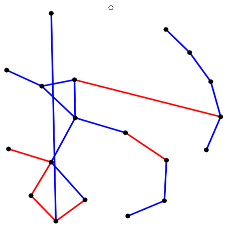


Figure: ancestor and children with $\delta = 2$ perturbations

# Simulation results

## Precision/Recall curve

precision = TP/(TP+FP)

recall = TP/P (power)

# Simulation results
## large sample size

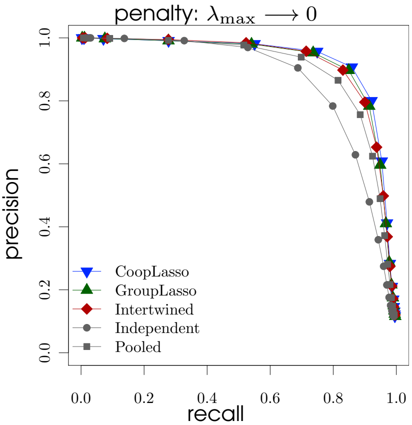



Figure: $n_t = 100, \delta = 1$

# Simulation results
## large sample size

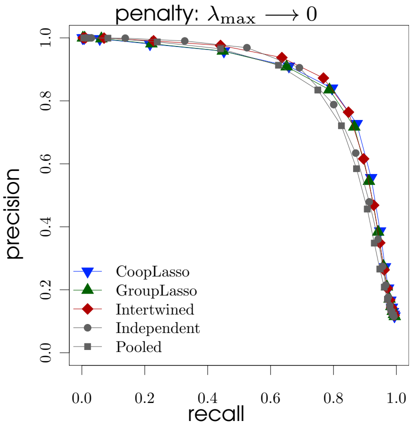



Figure: $n_t = 100, \delta = 3$

# Simulation results
## large sample size

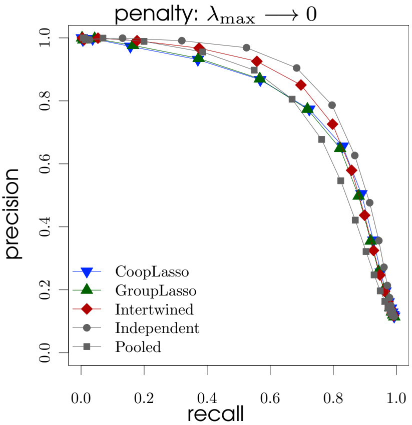



Figure: $n_t = 100$, $\delta = 5$

# Simulation results

## medium sample size

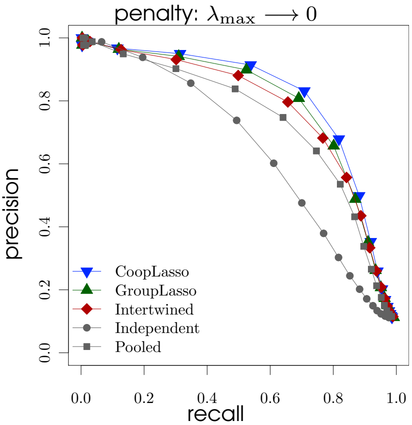



Figure: $n_t = 50, \delta = 1$

# Simulation results
## medium sample size

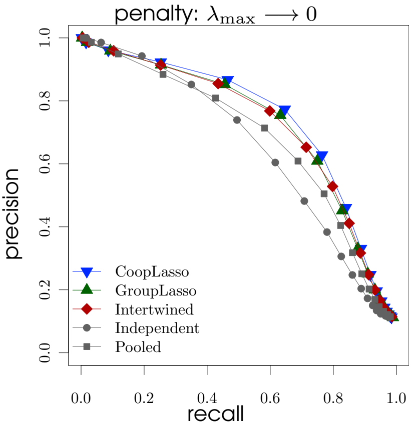



Figure: $n_t = 50$, $\delta = 3$

# Simulation results
## medium sample size

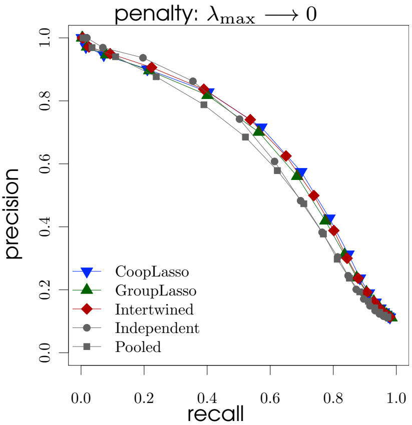



Figure: $n_t = 50, \delta = 5$

# Simulation results
## small sample size

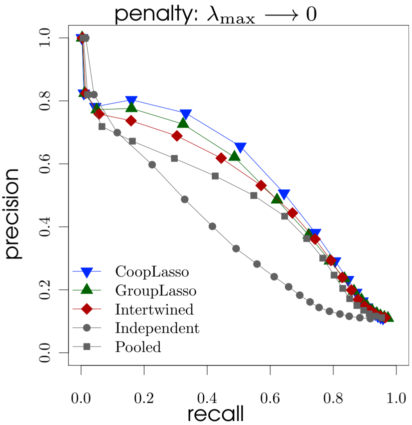



Figure: $n_t = 25, \delta = 1$

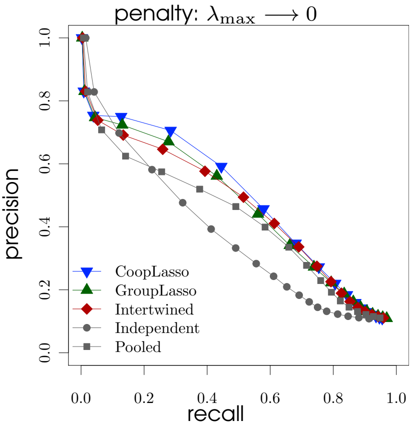



Figure: $n_t = 25, \delta = 3$

# Simulation results
## small sample size

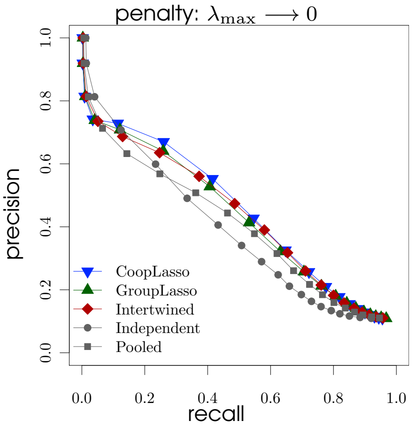



Figure: $n_t = 25, \delta = 5$

## Two types of patients

Patient response can be classified either as

1. pathologic complete response (PCR)
2. residual disease (not PCR)

## Gene expression data

- 133 patients (99 not PCR, 34 PCR)
- 26 identified genes (differential analysis)

# Package Demo

cancer data: Coop-Lasso

# Further investigations

## Introduce prior through existing pathways

Jeanmouin, Guedj, Ambroise (preprint `http://arxiv.org`)
Defining a robust biological prior from Pathway Analysis to drive Network Inference

Marine will speak at SMPGD '11 ☺

*"Due to the vast space of possible networks and the relatively small amount of data available, inferring genetic networks from gene expression data is one of the most challenging work in the post-genomic era. (...) We propose an original approach for inferring gene regulation network using a robust biological prior on structure in order to limit the set of candidate networks."*

# To sum-up

- Clarified links between neighborhood selection and graphical LASSO
- Identified the relevance of Multi-Task Learning in network inference
- First methods for inferring multiple Gaussian Graphical Models
- Consistent improvements upon the available baseline solutions
- Available in the R package SIMoNe

# Perspectives

## Issues

1. How can we choose for a unique network ? (should we ?)
   - Explore model-selection capabilities,
   - Network comparison.
2. Robustness
   - Test the validity of an edge ? Of a whole motif ?
   - Bootstrap greatly improves the inference but is computationally intensive,
   - Introduce more biological prior (semi-supervised learning).
3. Biological studies
   - Breast cancer (Marine),
   - Parkinson (with J.-C. Corvol, Pitié Salpétrière and Camille),
   - Bacillus subtilis and Staphylococcus aureus (ANR NOUGA déposée: heterogeneous data, RNAseq, new, prior etc.).

## Coop-Lasso

Theoretical analysis and other applications in genetics with penalized linear / logistic regression.

# Outline

Model selection

More details on optimisation

# Tuning the penalty parameter
## What does the literature say?

### Theory based penalty choices

1. Optimal order of penalty in the $p \gg n$ framework: $\sqrt{n \log p}$
   *Bunea et al. 2007, Bickel et al. 2009*
2. Control on the probability of connecting two distinct connectivity sets
   *Meinshausen et al. 2006, Banerjee et al. 2008, Ambroise et al. 2009*

$\rightsquigarrow$ practically much too conservative

### Cross-validation

- Optimal in terms of prediction, not in terms of selection
- Problematic with small samples:
  changes the sparsity constraint due to sample size

## Theorem (Zou et al. 2008)

$$\mathrm{df}(\hat{\beta}_\lambda^{\text{lasso}}) = \left\| \hat{\beta}_\lambda^{\text{lasso}} \right\|_0$$

## Straightforward extensions to the graphical framework

$$\mathrm{BIC}(\lambda) = \mathcal{L}(\hat{\mathbf{\Theta}}_\lambda; \mathbf{X}) - \mathrm{df}(\hat{\mathbf{\Theta}}_\lambda)\frac{\log n}{2}$$

$$\mathrm{AIC}(\lambda) = \mathcal{L}(\hat{\mathbf{\Theta}}_\lambda; \mathbf{X}) - \mathrm{df}(\hat{\mathbf{\Theta}}_\lambda)$$

Rely on asymptotic approximations, but still relevant for small data set

# Outline

Model selection

More details on optimisation

# Decomposition strategy (1)

Consider the $(pT) \times (pT)$ block-diagonal matrix $\mathbf{C}$ composed by the empirical covariance matrices of each tasks

$$\mathbf{C} = \begin{pmatrix} \mathbf{S}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathbf{S}^{(T)} \end{pmatrix},$$

and define

$$\mathbf{C}_{\backslash i \backslash i} = \begin{pmatrix} \mathbf{S}^{(1)}_{\backslash i \backslash i} & & 0 \\ & \ddots & \\ 0 & & \mathbf{S}^{(T)}_{\backslash i \backslash i} \end{pmatrix}, \quad \mathbf{C}_{i \backslash i} = \begin{pmatrix} \mathbf{S}^{(1)}_{i \backslash i} \\ \vdots \\ \mathbf{S}^{(T)}_{i \backslash i} \end{pmatrix}.$$

The $(p-1)T \times (p-1)T$ matrix $\mathbf{C}_{\backslash i \backslash i}$ is the matrix $\mathbf{C}$ where we removed each line and each column pertaining to variable $i$.

# Decomposition strategy (2)

## Estimate the $i^{\text{th}}$ neighborhood of the $T$ tasks bind together

$$\underset{\boldsymbol{\Theta}^{(t)}, t=1\dots,T}{\arg\max} \sum_{t=1}^{T} \tilde{\mathcal{L}}(\boldsymbol{\Theta}^{(t)}; \mathbf{S}^{(t)}) - \lambda\, \Omega(\boldsymbol{\Theta}^{(t)})$$

decomposes into $p$ convex optimization problems

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{T \times (p-1)}}{\arg\min} \; f(\boldsymbol{\beta}; \mathbf{C}) + \lambda\, \Omega(\boldsymbol{\beta}),$$

where we set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\Theta}_{i \setminus i}^{(t)}$ and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \vdots \\ \boldsymbol{\beta}^{(T)} \end{pmatrix} \in \mathbb{R}^{T \times (p-1)}.$$

# Solving the sub-problem

## Subdifferential approach

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{T \times (p-1)}} L(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Omega(\boldsymbol{\beta}) \ ,$$

$\boldsymbol{\beta}$ is a minimizer iif $\mathbf{0}_p \in \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$, with

$$\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta}).$$

# Solving the sub-problem

## Subdifferential approach

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{T \times (p-1)}} L(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Omega(\boldsymbol{\beta}) \ ,$$

$\boldsymbol{\beta}$ is a minimizer iif $\mathbf{0}_p \in \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$, with

$$\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta}).$$

For the graphical Intertwined LASSO

$$\Omega(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left\| \boldsymbol{\beta}^{(t)} \right\|_1 \ ,$$

where the grouping effect is managed by the function $f$.

# Solving the sub-problem

## Subdifferential approach

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{T\times(p-1)}} L(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Omega(\boldsymbol{\beta}) \ ,$$

$\boldsymbol{\beta}$ is a minimizer iif $\mathbf{0}_p \in \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$, with

$$\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta}).$$

For the graphical Group-LASSO

$$\Omega(\boldsymbol{\beta}) = \sum_{i=1}^{p-1} \left\| \boldsymbol{\beta}_i^{[1:T]} \right\|_2 \ ,$$

where $\boldsymbol{\beta}_i^{[1:T]} = \left(\beta_i^{(1)}, \dots, \beta_i^{(T)}\right)^\intercal \in \mathbb{R}^T$ is the vector of the $i$th component across tasks.

# Solving the sub-problem

## Subdifferential approach

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{T \times (p-1)}} L(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Omega(\boldsymbol{\beta}) \quad ,$$

$\boldsymbol{\beta}$ is a minimizer iif $\mathbf{0}_p \in \partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$, with

$$\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \partial_{\boldsymbol{\beta}} \Omega(\boldsymbol{\beta}).$$

For the graphical Coop-Lasso

$$\Omega(\boldsymbol{\beta}) = \sum_{i=1}^{p-1} \left( \left\| \left[ \boldsymbol{\beta}_i^{[1:T]} \right]_+ \right\|_2 + \left\| \left[ \boldsymbol{\beta}_i^{[1:T]} \right]_- \right\|_2 \right) \quad ,$$

where $\boldsymbol{\beta}_i^{[1:T]} = \left( \beta_i^{(1)}, \ldots, \beta_i^{(T)} \right)^\intercal \in \mathbb{R}^T$ is the vector of the $i$th component across tasks.