Pleiotropic mapping for genome-wide association studies using group variable selection.

B. Liquet[1,2] and K. Mengersen[2] and A. N. Pettitt[2] and M. Sutton[2]

[1] LMAP, Université de Pau et des Pays de L'Adour'

[2] ACEMS, QUT, Australia

Pleotropic mapping for genome-wide association studies using group variable selection

- Pleiotropy: genetic variants which affect multiple different complex diseases
- Example: genetic variants which affect both Breast and Thyroid cancer.
- Results from GWAS suggest that complex diseases are often affected by many variants with small effects (known as polygenicity)
- Aims:
  - statistical method to leverage pleiotropic effects
  - incoporate prior pathway knowledge to increase statistical power and identify important risk variants.

Genomics Data: Wide Data, High Dimensional Data

We have too many variables, prone to overfitting. Need to remove variable, or regularize, or both

- Main constraint: situation with $p > n$
- Strong colinearity among the variables.

Group structures within the data

- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

Group structures within the data

- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

We consider variables are divided into groups:

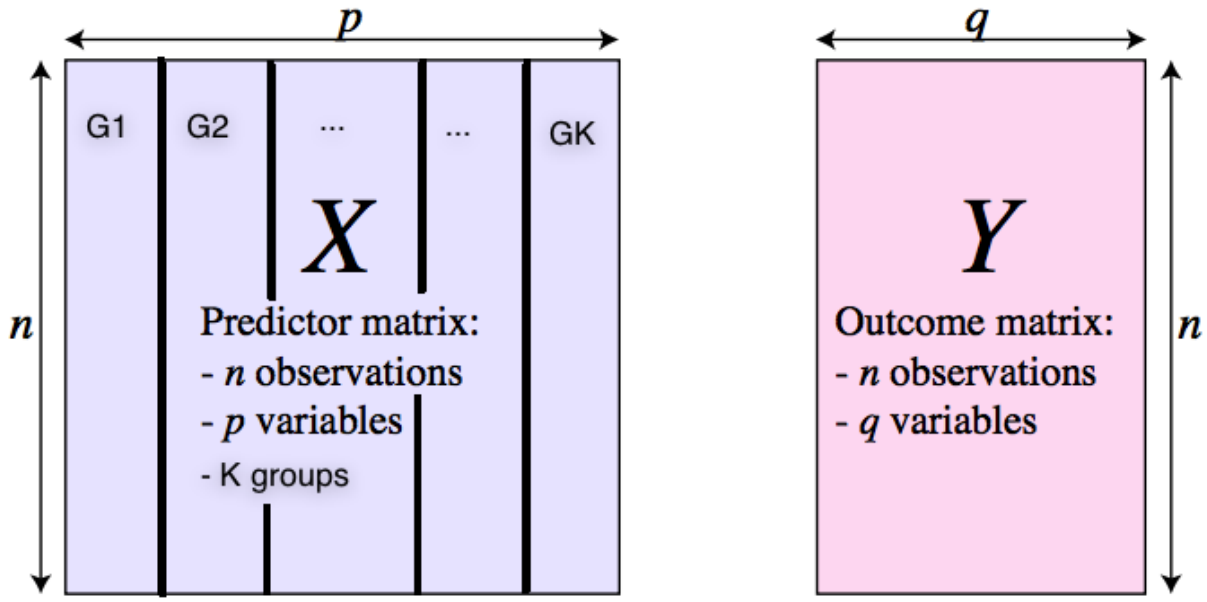
- Example p: SNPs grouped into K genes

$$\mathbf{X} = [\underbrace{SNP_1, \ldots + SNP_k}_{gene_1} \mid \underbrace{SNP_{k+1}, SNP_{k+2}, \ldots, SNP_h}_{gene_2} \mid \ldots \mid \underbrace{SNP_{l+1}, \ldots, SNP_p}_{gene_K}]$$

- Example p: genes grouped into K pathways/modules ($X_j = \text{gene}_j$)

$$\mathbf{X} = [\underbrace{X_1, X_2, \ldots, X_k}_{M_1} \mid \underbrace{X_{k+1}, X_{k+2}, \ldots, X_h}_{M_2} \mid \ldots \mid \underbrace{X_{l+1}, X_{l+2}, \ldots, X_p}_{M_K}]$$

Our contribution for Multivariate phenotypes



- Select group variables taking into account the data structures; all the variables within a group are selected otherwise none of them are selected
- Combine both sparsity of groups and within each group; only relevant variables within a group are selected

Our contribution for Multivariate phenotypes

Frequentist Approaches: Partial Least Squares (PLS)

- Sparse Group PLS : SNP $\subset$ Gene or Gene $\subset$ Pathways

  Liquet B., Lafaye de Micheaux P., Hejblum B. and Thiebaut R., (2016) *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* **Bioinformatics**, 32(1), 35–42.

- Sparse Group subgroup PLS : SNP $\subset$ Gene $\subset$ Pathways

  M. Sutton, R. Thiebaut, and B. Liquet. (2018) *Sparse group subgroup Partial Least Squares with application to genomics data*. Statistics in Medicine.

Our contribution for Multivariate phenotypes

Frequentist Approaches: Partial Least Squares (PLS)

- Sparse Group PLS : SNP $\subset$ Gene or Gene $\subset$ Pathways

  Liquet B., Lafaye de Micheaux P., Hejblum B. and Thiebaut R., (2016) *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* **Bioinformatics**, 32(1), 35–42.

- Sparse Group subgroup PLS : SNP $\subset$ Gene $\subset$ Pathways

  M. Sutton, R. Thiebaut, and B. Liquet. (2018) *Sparse group subgroup Partial Least Squares with application to genomics data*. Statistics in Medicine.

Main ideas:

- combining L_1 and L_2 penalties into the optimization function
- **Sparse Group Penalties:**

$$\lambda_1 \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 + \lambda_2 \|\boldsymbol{\beta}\|_1$$

Our contribution for Multivariate phenotypes

Bayesian Approaches: Multivariate regression model

- Bayesian group lasso model with spike and slab priors

  Liquet, B., Mengersen, K., Pettitt, A. N. and Sutton, M. (2016). *Bayesian Variable Selection Regression Of Multivariate Responses For Group Data* Bayesian Analysis. Volume 12, Number 4 (2017), 1039-1067.

Main ideas:

- **spike and slab** priors providing variable selection at the group level.
- hierarchical spike and slab prior structure to select variables both at the group level and within each group.

Extention for Pleiotropy: Model for multiple GWAS

- Suppose we have data from K independent GWAS datasets, $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$, where $\mathcal{D}_k = (\{y_1, x_1\}, \ldots, \{y_{n_k}, x_{n_k}\})$
- $y_{ik} \in \{0, 1\}$ denotes the phenotype of the kth study
- $x_{ik} \in \mathbb{R}^p$ is the vector with corresponding p SNPs.
- Logistic regression model

$$Logit(P(y_{ik} = 1|x_{ik}) = x_{ik}^T \boldsymbol{\beta}_{.k} \text{ for } k = 1, \ldots, K,$$

- $\boldsymbol{\beta}_{.k} \in \mathbb{R}^p$ the regression coefficients for the kth GWAS.
- Let $\boldsymbol{\beta}_{j.} \in \mathbb{R}^K$, $j = 1, \ldots, p$, the vector of K regression coefficients corresponding to the jth SNP over the K GWAS.

Group Structure

- SNPs can be partitioned into G groups (genes or Pathways)
- Let $\pi_g, g = 1, \ldots, G$ the set of SNPs contained in the gth group with $p_g = |\pi_g|$.
- Matrix of all regression coefficients as $\mathbf{B} = (\boldsymbol{\beta}_{\cdot 1}, \ldots, \boldsymbol{\beta}_{\cdot K}) = (\boldsymbol{\beta}_{1\cdot}, \ldots, \boldsymbol{\beta}_{p\cdot})^T$.

Frequentist Approach

- The log likelihood for the combined datasets:

$$p(\mathcal{D} \mid \mathbf{B}) = \sum_{k=1}^{K} L_k \quad \text{where} \quad L_k \quad \text{Log-Likelihood of study k}$$

- The penalised likelihood estimate

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathbb{R}^{p \times K}}{\operatorname{argmin}} \left\{ - \sum_{k=1}^{K} L_k + \lambda_1 \|\mathbf{B}\|_{G_{2,1}} + \lambda_2 \|\mathbf{B}\|_{\ell_{2,1}} \right\} \tag{1}$$

- $G_{2,1}$-norm penalty $\|\mathbf{B}\|_{G_{2,1}} = \sum_{g=1}^{G} \sqrt{\sum_{i \in \pi_g} \sum_{j=1}^{K} \beta_{ik}^2}$
- $\ell_{2,1}$-norm penalty $\|\mathbf{B}\|_{\ell_{2,1}} = \sum_{i=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{ik}^2}$ respectively.
- The $G_{2,1}$-norm fixes the group structure across studies and encourages sparsity at a gene level.
- The $\ell_{2,1}$-norm which allows sparsity within a group.

Inference

- Inference using the alternating direction method of multipliers algorithm (ADMM).
- Novel approach for identifying pleiotropic effects as it accounts for gene specific and SNP specific effects using a variable selection approach.
- The method is only capable of producing a point estimate of **B** and acurrate estimation of the variance for these parameters is not easily given.

Bayesian Logistic regression with multivariate spike and slab prior: LogitMBGL-SS

- Let $\gamma = (\gamma_1, \ldots, \gamma_p)^T$ indicate the association status for SNPs where $\gamma_j = 1$ indicates that the jth SNP is associated to all K traits.
- Spike and slab prior for the jth SNP $\boldsymbol{\beta}_{j\cdot} \in \mathbb{R}^K$,

$$
\begin{aligned}
\boldsymbol{\beta}_{j\cdot} &\sim (1-\gamma_j)\mathcal{N}_K(0, \tau_j^2 \mathbf{V}) + \gamma_j \delta_0(\boldsymbol{\beta}_{j\cdot}) \\
\tau_j^2 &\sim \text{Gamma}\left(\frac{K+1}{2}, \frac{\lambda}{2}\right), \\
\mathbf{V} &\sim IW(d, Q), \\
\gamma_j &\sim \text{Bernolli}(\alpha_0) \\
\alpha_0 &\sim \text{Beta}(a, b)
\end{aligned}
$$

  for $j = 1, \ldots, p$, where $\delta_0(\boldsymbol{\beta}_j)$ denotes a point mass at $\mathbf{0} \in R^K$.
- Here, $\mathbf{V} \in \mathbb{R}^{K \times K}$ is a covariance matrix modeling the covariance of the SNP effect on the traits.

Extension

- Should perform well when the SNPs are independent.
- GWAS datasets: strong correlations that can occur between SNPs within the same gene.
- Solution: reparameterise the coefficeints to handle the sparsity at a gene grouping level and individual feature level separately.
- $\tau \in \mathbb{R}^p$ to model individual sparsity
- $\mathbf{b}^{(g)} \in \mathbb{R}^{p_g K}$ with $\mathbf{b}^{(g)} = (\mathbf{b}_1^{(g)^T}, \ldots, \mathbf{b}_{p_g}^{(g)^T})$ where $\mathbf{b}_j^{(g)} \in \mathbb{R}^K$ for group sparsity.

$$\boldsymbol{\beta}_{j\cdot} = \tau_j \mathbf{b}_j^{(g)}, \qquad \text{where } \tau_j \geqslant 0, \quad \text{for all } j \in \pi_g.$$

Bayesian Logistic regression using multivariate sparse group selection with spike and slab priors

$$\boldsymbol{\beta}_{j\cdot} = \tau_j \mathbf{b}_j^{(g)}, \qquad \text{where } \tau_j \geqslant 0, \quad \text{for all } j \in \pi_g.$$

We assume the following multivariate spike and slab

$$\begin{aligned}
\mathbf{b}^{(g)} &\sim (1-\alpha_0)\mathcal{N}_{p_g K}(0, \mathbb{I}_{p_g} \otimes \mathbf{V}) + \alpha_0 \delta_0(\mathbf{b}^{(g)}) \\
\tau_j &\sim (1-\alpha_1)\mathcal{N}^{+}\left(0, s^2\right) + \alpha_1 \delta_0(\tau_j), \\
\alpha_0 &\sim \text{Beta}(a_1, a_2) \\
\alpha_1 &\sim \text{Beta}(c_1, c_2) \\
s^2 &\sim \text{InvGamma}(1, t)
\end{aligned}$$

for $j \in \pi_g$ and $g = 1, \ldots, G$

Signal recovery:

(i) GMT: Grouped multi-task penalised logistic regression ($\lambda_1 > 0, \lambda_2 = 0$) using $G_{2,1}$-norm

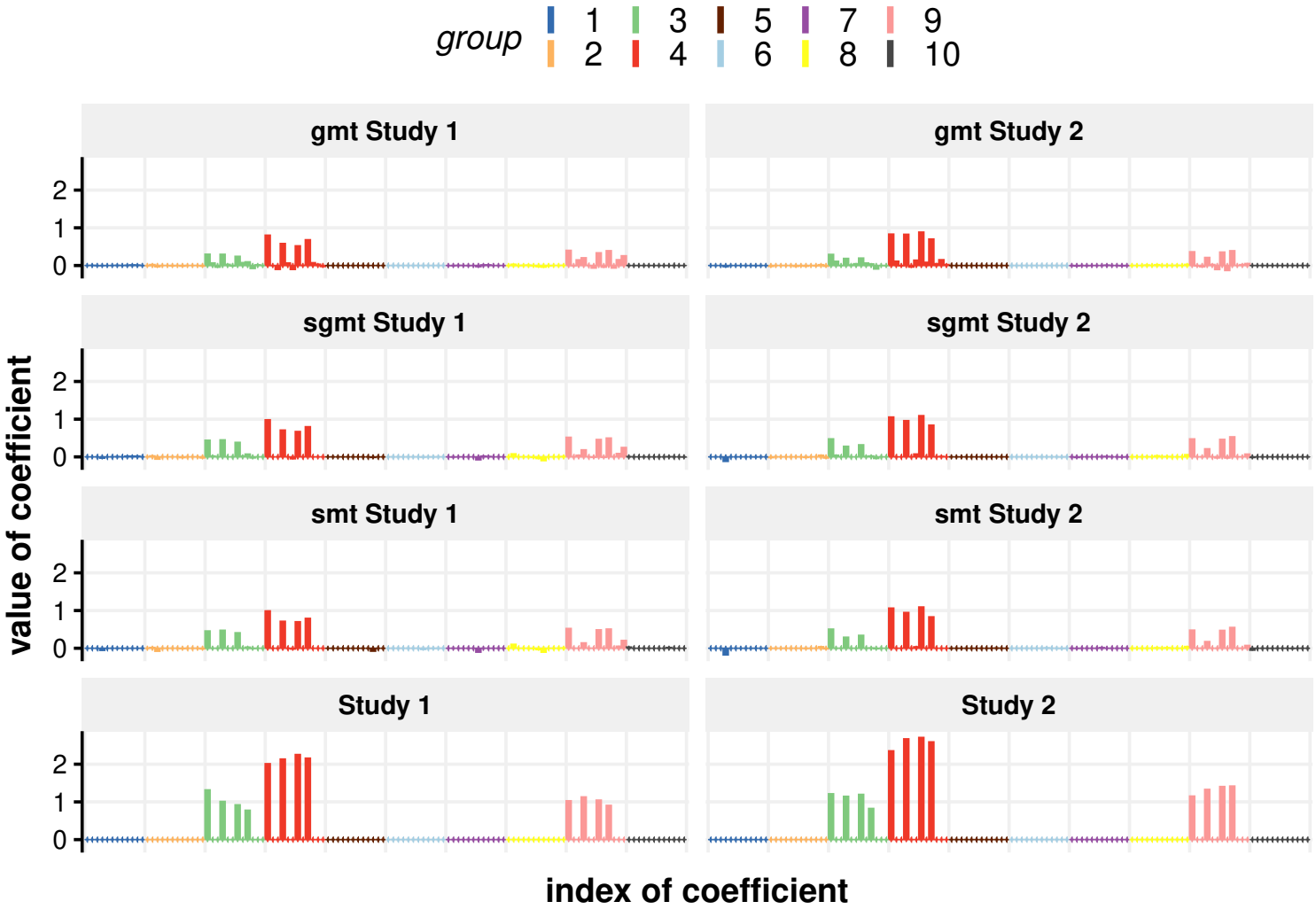
(ii) SMT: Sparse multi-task penalid penalised logistic regression ($\lambda_1 = 0, \lambda_2 > 0$) using $\ell_{2,1}$-norm

(iii) SGMT: Sparse group multi-task penalised logistic regression ($\lambda_1 > 0, \lambda_2 > 0$)

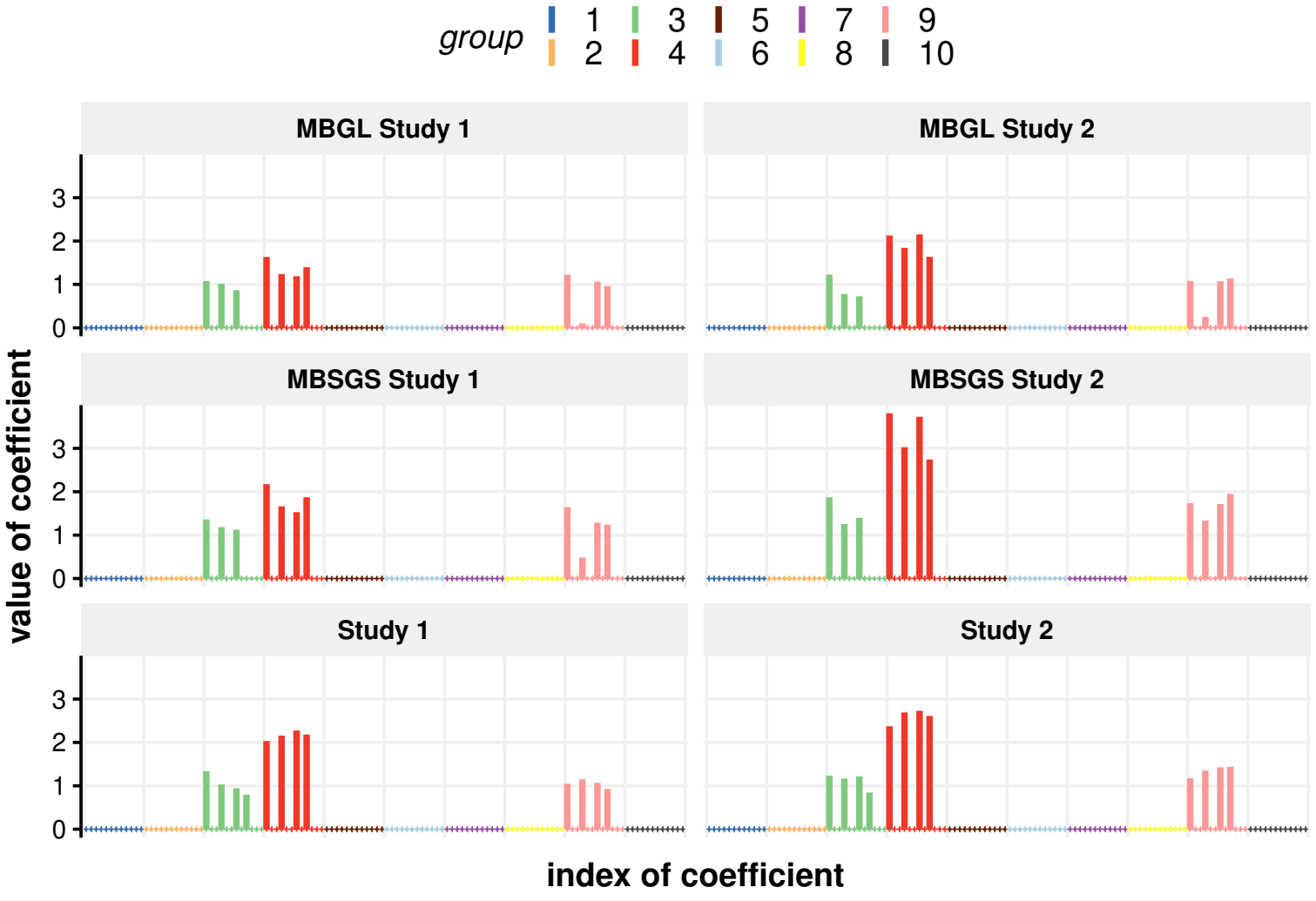
(iv) LOGITMBGL: Bayesian logistic regression using multivariate group lasso with spike and slab prior

(v) LOGITMBSGS: Bayesian logistic regression using multivariate sparse group selection with spike and slab prior

Results: Frequentist



Results: Bayesian



Main Conclusion on the simulation studies

- The penalised approaches perform reasonably well in variable selection but the reconstructed signal is underestimated.
- In general, the penalised likelihood methods suffer in terms of false negatives, selecting more variables to be nonzero than the Bayesian methods.
- The Bayesian methods perform the best in terms of signal recovery measured by the ℓ_1 error and variable selection performance metrics.
- The penalised likelihood approaches are computationally efficient using alternating direction method of multipliers algorithm
- Simulation results suggest that when computationally possible the Bayesian estimators should be used.
- The multivariate Bayesian sparse group selection with spike and slab prior performed the best in terms of signal recovery.
- The Bayesian method provides a natural method for quantifying the variability of the estimated coefficients.

What Next ?

- Application on real data: case/control studies
  - Breast Cancer and Thyroide Cancer
  - Thyroide Cancer (482 case, 463 control)
  - Breast Cancer (1172 case, 1125 control)
  - 6677 SNPs from 618 genes from 10 non-overlapping gene pathways.

Acknowledgement

This work is funded by “ La ligue contre le Cancer”

It is part of our current project on

“Cross Cancer Genomic Investigation of Pleiotropy effect and GxE interactions at pathway level: application to breast and thyroid cancers"