Setup:
- 6 single task lassos (on 6 different data sets) x 10 random seeds → 6 sets of 10 coefficients per gene
- 1 multitask lasso with 6 tasks x 10 random seeds → 6 sets of 10 coefficients per gene

In other words, 2 sets (one from a single task lasso, one from the multitask lasso) of 10 coefficients per gene and per data set.

Step 1:

Remove outliers in each list of 10 coefficients → 2 sets of *at most* 10 coefficients per gene and per data set. Let us call these sets S^{st}_{jt} (set of coefficients from the single task method, for gene j and data set t) and S^{mt}_{jt} (set of coefficients from the multitask method, for gene j and data set t)

Step 2:

The goal is to know whether the signs of the coefficients agree 1) within each set 2) across all sets

Given a set S of m coefficients, which can each be 0, positive, or negative, we call m_0, m_+, m_- the numbers of 0 coefficients, the number of positive coefficients, and the number of negative coefficients, respectively.

Our score for set S is

$$s(S) = \begin{cases} 0 & \text{if } (m_0 \geq (m-1)) \text{ or } (m_+ = m_-) \\ \frac{m_+}{(m_+ + m_-)} & \text{if } m_+ > m_- \\ -\frac{m_-}{(m_+ + m_-)} & \text{otherwise.} \end{cases}$$

Proposition: score each gene j according to

- whether the sign of its coefficients agree **within each** S^{st}_{jt} and S^{st}_{jt} for t=1, 2, ..., 6:
    - for this, we compute $s(S^{\mathrm{st}}_{jt})$ and $s(S^{\mathrm{mt}}_{jt})$
    - we can decide to keep considering gene j if $s(S^{\mathrm{st}}_{jt}) = 1$ and $s(S^{\mathrm{mt}}_{jt}) = 1$ for all t=1, 2, ..., 6 (stringent; corresponds to score2A=6 I think) → start here, only do the steps below if this is too stringent

      Intermediate step: for each data set t, report the proportion of genes for which $m_0 < (m-1)$ (ie at least 2 non-0 coefficients) and $s(S^{\mathrm{st}}_{jt}) = s(S^{\mathrm{mt}}_{jt}) = 1$ (ie full agreement of the coefficient signs) to help decide whether to keep this stringent threshold or not
    - or we can be more lenient and keep considering gene j
        - if $s(S^{\mathrm{st}}_{jt}) = 1$ and $s(S^{\mathrm{mt}}_{jt}) = 1$ for at least 5 data sets/tasks/values of t (allowing one dataset to give inconsistent results)
        - if $s(S^{\mathrm{st}}_{jt}) \geq 0.9$ and $s(S^{\mathrm{mt}}_{jt}) \geq 0.9$ for all t=1, 2, ...,6 (allowing one of the random seeds to be inconsistent with the others)
- whether the sign of its coefficients agree **across all** S^{st}_{jt} and S^{mt}_{jt}:
    - only for the genes j we have kept from the previous step
    - compute the score on the union of all coefficients: $s\left(\bigcup_{t=1}^{6} S^{\mathrm{st}}_{jt} \cup \bigcup_{t=1}^{6} S^{\mathrm{mt}}_{jt}\right)$

      Intermediate step: report the distribution (histogram) of those scores to help decide of thresholds defining strong and weak SIGs
    - in the end, report as
        - "strict SIGs" those for which $|s| = 1$ (you directly see whether they're + or - from the sign of s)

- "strong SIGs" those for which $\theta_{\text{strong}} \leq |s| < 1$, where θ_{strong} is for example 0.9 (90% of the at most 120 coefficients agree) → adjust the value of the threshold depending on your results
- "weak SIGs" those for which $\theta_{\text{weak}} \leq |s| < \theta_{\text{strong}}$, where θ_{strong} is for example 0.7 (70% of the at most 120 coefficients agree) → adjust the value of the threshold depending on your results

Question 2

Si on s'intéresse à l'intersection de deux méthodes seulement, je calculerai les scores uniquement sur les deux sets en question, mais je croiserai au final les résultats pour reporter 1) les gènes qui sont des SIGs pour ces deux méthodes et 2) parmi eux, les gènes qui sont aussi des SIGs pour l'ensemble des méthodes.