ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Automated calibration for stability selection in penalised regression and graphical models: a multi-OMICs network application exploring the molecular response to tobacco smoking

Journal:	*Journal of the Royal Statistical Society: Series C*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Bodinier, Barbara; Imperial College London, School of Public Health Filippi, Sarah; Imperial College London, Department of Mathematics Haugdahl Nøst, Therese; UiT The Arctic University of Norway, Department of Community Medicine Chiquet, Julien; Université Paris-Saclay, UMR MIA, AgroParisTech INRAE Chadeau-Hyam, Marc; Imperial College London, School of Public Health

SCHOLARONE™
Manuscripts

Automated calibration for stability selection in penalised regression and graphical models: a multi-OMICs network application exploring the molecular response to tobacco smoking

Barbara Bodinier

MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom.

E-mail: b.bodinier@imperial.ac.uk

Sarah Filippi

Department of Mathematics, Imperial College London, London, UK.

Therese Haugdahl Nøst

Systemsepidemiology, Department of Community Medicine, UiT The Arctic university of Norway, Tromsø, Norway.

Julien Chiquet

Université Paris-Saclay, AgroParisTech INRAE, UMR MIA, Paris, France.

Marc Chadeau-Hyam

MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom.

Summary. Stability selection represents an attractive approach to identify sparse sets of features jointly associated with an outcome in high-dimensional contexts. We introduce an automated calibration procedure via maximisation of an in-house stability score and accommodating *a priori*-known block structure (e.g. multi-OMIC) data. It applies to (LASSO) penalised regression and graphical models. Simulations show our approach outperforms non-stability-based and stability selection approaches using the original calibration. Application to multi-block graphical LASSO on real (epigenetic and transcriptomic) data from the Norwegian Women and Cancer study reveals a central/credible and novel cross-OMIC role of the LRRN3 in the biological response to smoking.

2   *B Bodinier et al.*

1.   Introduction

Tobacco smoking has long been established as a dangerous exposure causally linked to several severe chronic conditions. It has been estimated that one in five deaths in the United States was due to smoking (National Center for Chronic Disease et al., 2014). Nevertheless, the molecular mechanisms triggered and dysregulated by the exposure to tobacco smoking remain poorly understood. Over the past two decades, OMICs technologies have developed as valuable tools to explore molecular alterations due to external stressors or exposures (Niedzwiecki et al., 2019). Statistical analysis of OMICs data has enabled the identification of molecular markers of exposure at a single molecular level (Joehanes et al., 2016; Huan et al., 2016) and are progressively moving towards the integration of data arising from different platforms (Guida et al., 2015; Noor et al., 2019). There is an increasing need for efficient multivariate approaches accommodating high-dimensional and heterogeneous data typically exhibiting block-correlation structures. In particular, variable selection models can identify sparse sets of predictors and have proved useful for signal prioritisation in this context (Chadeau-Hyam et al., 2013; Vermeulen et al., 2018). Of these, the Least Absolute Shrinkage Selection Operator (LASSO) uses the ℓ_1-penalisation of the coefficients to achieve variable selection (Tibshirani, 1996). Extensions of these penalised regression models have been proposed for the estimation of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Friedman et al., 2007). By applying a ℓ_1-penalisation to the precision matrix (as defined by the inverse of the covariance matrix), the graphical LASSO identifies non-zero entries of the partial correlation matrix. The evaluation (and subsequent selection) of pairwise relationships between molecular features in graphical models can guide biological interpretation of the results, under the assumption that statistical correlations reflect molecular interactions (Barabási and Oltvai, 2004; Valcárcel et al., 2011).

We focus in the present paper on the calibration of feature selection models, where feature denotes interchangeably a variable (in the context of regression) or an edge (graphical model). We illustrate our approach with regularised models, in which the model size (number of selected features) is controlled by the penalty parameter. The choice of parameter has strong implications on the generated results. Calibration procedures using cross-validation (Friedman et al., 2010; Leng et al., 2006) or maximisation of information theory metrics, including the Bayesian (BIC) or Akaike (AIC) Information Criterion (Akaike, 1998; Schwarz, 1978; Foygel and Drton, 2010; Giraud, 2008) have been proposed.

These models can be complemented by stability approaches to enhance reliability of the findings (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013; Liu et al., 2010). In stability selection, the selection algorithm is combined with resampling techniques to identify the most stable signals. The model relies on the introduction of a second parameter: a threshold in selection proportion above which the corresponding feature is considered stable. A formula providing the upper-bound of the expected number of falsely selected features, or Per-Family Error Rate (PFER), as a function of the two parameters has been derived and is currently used to guide calibration (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). However, this calibration relies on the arbitrary choice of one of the two parameters, which can sometimes be difficult to justify.

We introduce a score measuring the overall stability of the set of selected features, and use it to propose a new calibration strategy for stability selection. Our intuition is that selection proportions in an unstable model would be uniformly distributed. Our calibration procedure does not rely on the arbitrary choice of any parameter. Optionally, the problem can be constrained on the expected number of falsely selected variables and generate sparser results with error control.

We also extend our calibration procedure to accommodate multiple blocks of data. This extension was motivated by the practical example on integration of data from different OMICs platforms. In this setting, block patterns arise, typically with higher (partial) correlations within a platform than between (Canzler et al., 2020). We propose here an extension of stability selection combined with the graphical LASSO accommodating data with a known block structure. For this approach, each block is tuned using a block-specific pair of parameters (penalty and selection proportion threshold) (Ambroise et al., 2009).

We conduct an extensive simulation study to evaluate the performances of our calibrated stability selection models and compare them to state-of-the-art approaches. Our multi-OMICs stability-enhanced graphical models are applied to targeted methylation and gene expression data from an existing cohort. These datasets are intergated in order to characterise the molecular response to tobacco smoking at multiple molecular levels. The transcript of the LRRN3 gene, and its closest CpG site were found to play a central role in the generated graph. These two variables have the largest numbers of cross-OMICs edges and appear to be linking two largely uni-OMICs modules. LRRN3 methylation and gene expression therefore appear as pivotal molecular signals driving the biological response to tobacco smoking.

2. Methods

2.1. Stability selection

Stability-enhanced procedures for feature selection proposed in the literature include stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) and the Stability Approach to Regularization Selection (StARS) (Liu et al., 2010). Both use an existing selection algorithm and complement it with resampling techniques to estimate the probability of selection of each feature using its selection proportion over the resampling iterations. Stability selection ensures reliability of the findings through error control.

The feature selection algorithms we use are (a) the LASSO in a regression framework (Tibshirani, 1996; Friedman et al., 2010), and (b) the graphical LASSO for the estimation of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008; Sustik M.A., 2012) (see Supplementary Methods, section 1.1 and 1.2 for more details on the algorithms). The latter aims at the construction of a conditional independence graph. In a graph with p nodes, for each pair of variables X, Y and Gaussian vector Z compiling the $(p-2)$ other variables, an edge is included if the conditional covariance $\text{cov}(X, Y|Z)$ is different from zero (see Supplementary Methods, section 1.3 for more details on model calibration).

Let $F_\lambda(j)$ denote the binary selection status of feature $j \in \{1, \ldots, N\}$ by the pe-

nalised model with parameter λ. Under the assumption that the selection of feature j is independent from the selection of any other feature $i \neq j$, the selection status follows a Bernouilli distribution with parameter $p_\lambda(j)$, the selection probability of feature j. The stability selection model is then defined as the set $V_{\lambda,\pi}$ of features with selection probability above a threshold π:

$$V_{\lambda,\pi} = \{j : p_\lambda(j) \geq \pi\} \tag{1}$$

For each feature j, the selection probability is estimated as the selection proportion across models with penalty parameter λ applied on K subsamples of the data.

The stability selection model has two parameters (λ, π) that need to be calibrated. In the original paper, Meinshausen and Bühlmann use random subsamples of 50% of the observations. They introduce q_Λ, the average number of features selected by the underlying algorithm (e.g. LASSO) for a range of values $\lambda \in \Lambda$. Under the assumptions of (a) exchangeability between selected features, and (b) that the selection algorithm is not performing worse than random guessing, they derived an upper-bound of the PFER, denoted by PFER_{MB}, as a function of the number of selected features q_Λ and threshold in selection proportion π:

$$\text{PFER}_{MB}(\Lambda, \pi) = \frac{1}{2\pi - 1} \frac{q_\Lambda^2}{N}$$

With simultaneous selection in complementary pairs (CPSS), the selection proportions are obtained by counting the number of times the feature is selected in both the models fitted on a subsample of 50% of the observations and its complementary subsample made of the remaining 50% of observations (Shah and Samworth, 2013). Using this subsampling procedure, the exchangeability assumption is not required for the upper-bound PFER_{MB} to be valid. Under the assumption of unimodality of the distribution of selection proportions obtained with CPSS, Shah and Samworth also proposed a stricter upper-bound on the expected number of variables with low selection probabilities, denoted here by PFER_{SS}:

$$\text{PFER}_{SS}(\Lambda, \pi) = \begin{cases} \frac{1}{2 \times (2\pi - 1 - 1/K)} \frac{q_\Lambda^2}{N} & \text{if } \pi \leq 0.75 \\ \frac{4 \times (1 - \pi + 1/K)}{1 + 2/K} \frac{q_\Lambda^2}{N} & \text{otherwise.} \end{cases}$$

For simplicity, we consider here point-wise control (Λ reduces to a single value λ) with no effects on the validity of the formulas. Both approaches provide a relationship between λ (via q_λ), π and the upper-bound of the PFER such that if two of them are fixed, the third one can be calculated. The authors of both papers proposed to guide calibration based on the arbitrary choice of two of these three quantities. For example, the penalty parameter λ can be calibrated for a combination of fixed values of the selection proportion π and threshold in PFER.

To avoid the arbitrary choice of the selection proportion π or penalty λ, we introduce here a score measuring the overall stability of the model and use it to jointly calibrate these two parameters. We also consider the use of a user-defined threshold in PFER to limit the set of parameter values for λ and π to explore.

2.2. Stability score

The selection probability is estimated as the selection proportion of feature i over K resampling iterations Meinshausen and Bühlmann (2010). Under the assumption that the subsamples are independent, the selection count $H_\lambda(j)$ of feature $j \in \{1, \ldots, N\}$ over the K resampling iterations follows a binomial distribution:

$$H_\lambda(j) \sim B\big(K, p_\lambda(j)\big)$$

For a given penalty parameter λ, each feature j is classified as (a) stably selected if $H_\lambda(j) \geq K\pi$, (b) stably excluded if $H_\lambda(j) \leq K(1 - \pi)$, or (c) unstable if $(1 - \pi)K < H_\lambda(j) < K\pi$.

By considering the N features as observations, we can derive the likelihood of observing this classification, given λ and π:

$$
\begin{aligned}
L_{\lambda,\pi} = \prod_{j=1}^{N} \Big[&\mathbb{P}\big(H_\lambda(j) \geq K\pi\big)^{\mathbb{1}_{\{H_\lambda(j) \geq K\pi\}}} \\
&\times \mathbb{P}\big((1 - \pi)K < H_\lambda(j) < K\pi\big)^{\mathbb{1}_{\{(1-\pi)K < H_\lambda(j) < K\pi\}}} \\
&\times \mathbb{P}\big(H_\lambda(j) \leq K(1 - \pi)\big)^{\mathbb{1}_{\{H_\lambda(j) \leq K(1-\pi)\}}} \Big]
\end{aligned}
$$

Intuitively, we assume that the results are more reliable when the model is confident in its choice of features to include, i.e. when the selection procedure is not uniform. We then formulate that under the null hypothesis of no informative features, all features would have the same probability q_λ/N to be selected in the model, where q_λ is the average number of selected features across the K models fitted with penalty λ on the different subsamples of the data. Our stability score $S_{\lambda,\pi}$ is defined as minus the log-likelihood of the model under the uniform assumption:

$$\hat{S}_{\lambda,\pi} = -\log(L_{\lambda,\pi})$$

The score measures how unlikely a given model is to arise from the null hypothesis, for a given set of λ and π. As such, the higher the score, the more stable the set of selected features. By construction, this formula is accounting for (a) the total number of features N, (b) the number of iterations K, (c) the density of selected sets by the original procedure via λ, and (d) the level of stringency as measured by threshold π. The calibration approach we develop aims at identifying sets of parameters λ and π maximising our score:

$$\max_{\lambda,\pi} \; S_{\lambda,\pi} \tag{2}$$

Furthermore, this calibration technique can be extended to incorporate some error control via a constraint ensuring that the expected number of false positives is below an a priori fixed threshold in PFER η:

$$\max_{\lambda,\pi} \; S_{\lambda,\pi} \quad \text{such that } U_{\lambda,\pi} \leq \eta, \text{where} \tag{3}$$

$U_{\lambda,\pi}$ is the upper-bound used for error control in existing strategies (i.e. PFER_{MB} or PFER_{SS}) (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013).

In the following sections, the use of Equation (2) is referred to as unconstrained calibration, and that of Equation (3) as constrained calibration.

6   *B Bodinier et al.*

2.3.   Multi-block graphical models

The combination of heterogeneous groups of variables can create technically-induced patterns in the estimated (partial) correlation matrix, subsequently inducing bias in the generated graphical models. This can be observed, for example, when integrating the measured levels of features from different OMICs platforms. The between-platform (partial) correlations are overall weaker than within platforms (Supplementary Figure S1). This makes the detection of bipartite edges more difficult. This structure is known a priori and does not need to be inferred from the data. Indeed, the integration of data arising from G homogeneous groups of variables generates $B = \frac{G \times (G+1)}{2}$ two-dimensional blocks in the (partial) correlation matrix where variables are ordered by group (Ambroise et al., 2009).

To tackle this scaling issue, we propose to use and calibrate block-specific pairs of parameters, λ_b and π_b controlling the level of sparsity in block b. Let $E_b, b \in \{1, \dots, B\}$ denote the sets of edges belonging to each of the blocks, such that:

$$\bigcup_{b=1}^{B} E_b = \{1, \dots, N\}$$

The stability selection model can be defined more generally as:

$$V_{\lambda_1,\dots,\lambda_B,\pi_1,\dots,\pi_B} = \bigcup_{b=1}^{B} \{j \in E_b : p_{\lambda_1,\dots,\lambda_B}(j) \geq \pi_b\}, \text{ where} \tag{4}$$

The probabilities $p_{\lambda_1,\dots,\lambda_B}(j), j \in \{1, \dots, N\}$ are estimated as selection proportions of the edges obtained from graphical LASSO models fitted on K subsamples of the data with a block penalty matrix such that edge $j \in E_b$ is penalised with λ_b.

Our stability score is then defined, by block, as:

$$S_{\lambda_1,\dots,\lambda_B,\pi_1,\dots,\pi_B} = -\log \Big(\prod_{b=1}^{B} \prod_{j \in E_b} \Big[\mathbb{P}\big(H_{\lambda_1,\dots,\lambda_B}^b(j) \geq K\pi_b\big)^{\mathbb{1}_{\{H_{\lambda_1,\dots,\lambda_B}^b(j) \geq K\pi_b\}}}$$

$$\times \mathbb{P}\big((1-\pi_b)K < H_{\lambda_1,\dots,\lambda_B}^b(j) < K\pi_b\big)^{\mathbb{1}_{\{(1-\pi_b)K < H_{\lambda_1,\dots,\lambda_B}^b(j) < K\pi_b\}}}$$

$$\times \mathbb{P}\big(H_{\lambda_1,\dots,\lambda_B}^b(j) \leq K(1-\pi_b)\big)^{\mathbb{1}_{\{H_{\lambda_1,\dots,\lambda_B}^b(j) \leq K(1-\pi_b)\}}} \Big] \Big)$$

Alternatively, we propose a block-wise decomposition, as described in Equation (5). To ensure that pairwise partial correlations in each block are estimated conditionally on all other $(p-2)$ nodes, we propose to estimate them from graphical LASSO models where the other blocks are weakly penalised (i.e. with small penalty λ_0). We introduce $p_{\lambda_1,\lambda_0}^b(j)$ and $H_{\lambda_1,\lambda_0}^b(j)$, the selection probability and count of edge $j \in E_b$ as obtained from graphical LASSO models fitted with a block penalty matrix such that edges $j \in E_b$ are penalised with λ_b and edges $i \in E_\ell, \ell \neq b$ are penalised with λ_0. We define the multi-block stability selection graphical model as the union of the sets of block-specific stable edges:

$$V_{\lambda_1,\dots,\lambda_B,\lambda_0,\pi_1,\dots,\pi_B} = \bigcup_{b=1}^{B} \{j \in E_b : p_{\lambda_1,\lambda_0}^b(j) \geq \pi_b\} \tag{5}$$

The pair of parameters is calibrated for each of the blocks separately using a block-specific stability score defined by:

$$
\begin{aligned}
S^b_{\lambda_b,\lambda_0,\pi_b} = -\log\Big(\prod_{j\in E_b}\Big[&\mathbb{P}\big(H^b_{\lambda_b,\lambda_0}(j)\geq K\pi_b\big)^{\mathbb{1}_{\{H^b_{\lambda_b,\lambda_0}(j)\geq K\pi_b\}}} \\
&\times \mathbb{P}\big((1-\pi_b)K < H^b_{\lambda_b,\lambda_0}(j) < K\pi_b\big)^{\mathbb{1}_{\{(1-\pi_b)K<H^b_{\lambda_b,\lambda_0}(j)<K\pi_b\}}} \\
&\times \mathbb{P}\big(H^b_{\lambda_b,\lambda_0}(j)\leq K(1-\pi_b)\big)^{\mathbb{1}_{\{H^b_{\lambda_b,\lambda_0}(j)\leq K(1-\pi_b)\}}}\Big]\Big)
\end{aligned}
$$

The implication of these assumptions are evaluated by comparing the two approaches described in Equations (5) and (4) in a simulation study.

2.4. Implementation
The stability selection procedure is applied for different values of λ and π and the stability score is computed for all visited pairs of parameters. The grid of λ values is chosen so that the underlying selection algorithm visits a range of models from empty to dense (up to 50% of edges selected by the graphical LASSO) (Friedman et al., 2010; Müller et al., 2016). Values of the threshold π vary between 0.6 and 0.9, as proposed previously (Meinshausen and Bühlmann, 2010).

2.5. Simulation models
In order to evaluate the performances of our approach and compare to other established calibration procedures, we simulated several datasets according to the models described below.

2.5.1. Linear regression
We simulated a dataset with $n=100$ observations of $p=50$ independent and Normally distributed predictors X, such that $X_j \sim \mathcal{N}(0,1), \forall j \in \{1,\ldots,p\}$. Among the $p=50$ predictors, we considered 10 signal variables and 40 variables not contributing to the outcome. The effect sizes β were uniformly sampled in $[-1,1]$ for signal variables and set to 0 for un-informative variables. The continuous outcome Y was sampled from a Gaussian distribution centered around $X\beta$ with standard deviation σ, such that $X\beta$ explains 60% of the variance of Y.

2.5.2. Graphical models
We build upon previously proposed models to simulate multivariate Normal data with an underlying graph structure (Zhao et al., 2012). Our contributions include (a) a procedure for the automated choice of the parameter ensuring that the generated correlation matrix has contrast, and (b) the simulation of block-structured data.

First, we simulate the binary adjacency matrix Θ of size $(p \times p)$ of a random graph with density ν using the Erdös-Rényi model (Erdös and Rényi, 1959) or a scale-free graph using the Barabási-Albert preferential attachment algorithm (Albert and Barabási, 2002;

8   *B Bodinier et al.*

Zhao et al., 2012). To introduce a block structure in the generate data, the non-diagonal entries of the precision matrix Ω are simulated such that:

$$\Omega_{ij} = \begin{cases} 0 & \text{if } \Theta_{ij} = 0 \\ \alpha_{ij} & \text{if } \Theta_{ij} = 1 \text{ and } i \text{ and } j \text{ belong to the same platform} \\ \alpha_{ij}v_b & \text{if } \Theta_{ij} = 1 \text{ and } i \text{ and } j \text{ belong to different platforms.} \end{cases}, \quad i \neq j \text{ where}$$

$\alpha_{ij} \sim U(\{-1, 1\})$ and $v_b \in [0, 1]$ is a user-defined parameter.

We ensure that the generated precision matrix is positive definite via diagonal dominance:

$$\Omega_{ii} = \sum_{j=1}^{p} |\Omega_{ij}| + u, \text{ where}$$

$u > 0$ is a parameter to be tuned.
The data is simulated from the centered multivariate Normal distribution with covariance Ω^{-1}.

The simulation model is controlled by five parameters:

(a) number of observations n,
(b) number of nodes p,
(c) density of the underlying graph $\nu \in [0, 1]$,
(d) scaling factor $v_b \in [0, 1]$ controlling the level of heterogeneity between blocks,
(e) constant $u > 0$ ensuring positive definiteness.

We propose to choose u so that the generated correlation matrix has a high contrast, as defined by the number of unique truncated correlation coefficients with three digits (Supplementary Figure S2). The parameter $v_b \in [0, 1]$ is set to 1 (no block structure) for single-block simulations and chosen to generate data with a visible block structure for multi-block simulations ($v_b = 0.2$). These models generate realistic correlation matrices (Supplementary Figure S1).

2.5.3. Performance metrics

Selection performances of the investigated models are measured in terms of precision p and recall r:

$$p = \frac{TP}{TP + FP} \text{ and } r = \frac{TP}{TP + FN}, \text{ where}$$

TP and FP are the numbers of true and false positives, respectively, and FN is the number of false negatives.

The F_1-score quantifies the overall selection performance based on a single metric:

$$F_1 = \frac{2 \times p \times r}{p + r}$$

Automated calibration for stability selection in penalised regression and graphical models   9

3. Simulation study

We use a simulation study to demonstrate the relevance of stability selection calibrated with our approach:

(a) in a linear regression context for the LASSO model,
(b) for graphical model using the graphical LASSO,
(c) for multi-block graphical models.

From these we evaluate the relevance of our stability score for calibration purposes, and compare our score to a range of existing calibration approaches including information theory criteria, StARS, and stability selection models using the previously proposed error control for different values of the threshold in selection proportion π. As sensitivity analyses, we evaluate the performances of stability selection for graphical models using different resampling approaches, different numbers of iterations K, and compare the two proposed approaches for multi-block calibration.

3.1. Simulation parameters

All simulation parameters were chosen in an attempt to generate realistic data with many strong signals and some more difficult to detect (weaker partial correlation).

For graphical models, we used $p = 100$ nodes with an underlying random graph structure of density $\nu = 0.02$ (99 edges on average, as would be obtained in a scale-free graph with the same number of nodes). For multi-block graphical models, we considered two homogeneous groups of 50 nodes each. Reported distributions of selection metrics were computed over 1,000 simulated datasets.

Unless otherwise stated, stability selection models were applied on grids of 50 dataset-specific penalty parameter values and 31 values for the threshold in selection proportion between 0.6 and 0.9. The stability-enhanced models were based on $K = 100$ (complementary) subsamples of 50% of the observations.

3.2. Applications to simulated data

Our stability selection approach is first applied to the LASSO for the selection of variables jointly associated with a continuous outcome in simulated data (Figure 1).

The penalty parameter λ and threshold in selection proportion π are jointly calibrated to maximise the stability score (Figure 1-A). Stably selected variables are then identified as those with selection proportions greater than the calibrated parameter $\hat{\pi} = 0.81$ (dark red line) in LASSO models fitted on 50% of the data with calibrated penalty parameter $\hat{\lambda} = 0.28$ (Figure 1-B). The resulting set of stably selected variables includes 7 of the 10 'true' variables used to simulate the outcome and 1 'wrongly selected' variables we did not use in our simulation.

To limit the number of 'wrongly selected' features, we can restrict the values of λ and π visited so they ensure a control of the PFER (Supplementary Figure S3). In that constrained optimisation, the values of λ and π yielding a PFER exceeding the specified threshold are discarded and corresponding models are not evaluated (Supplementary Figure S3-A). The maximum stability score can be obtained for different pairs (λ, π)

10   *B Bodinier et al.*
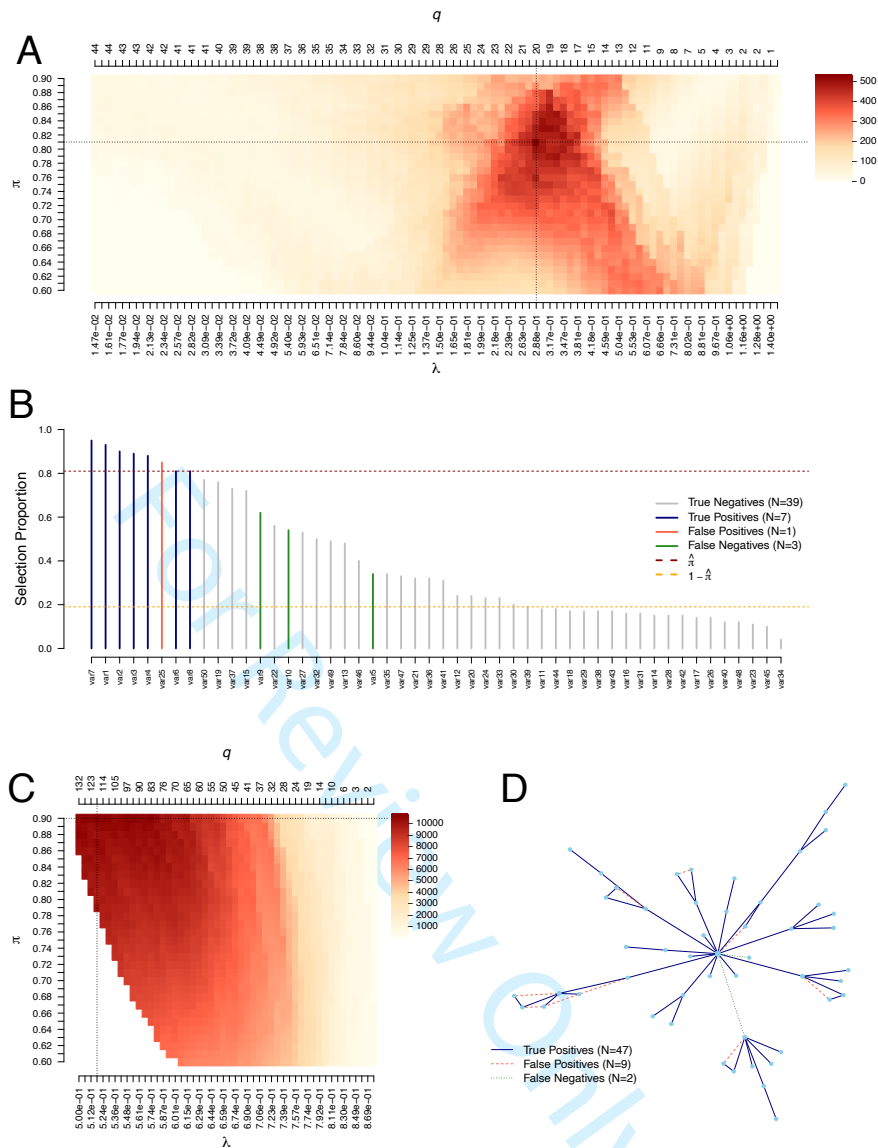
Fig. 1. Stability selection LASSO (A-B) and graphical LASSO (C-D) applied on simulated data. Calibration plots (A, C) show the stability score (colour-coded) as a function of the penalty parameter λ, corresponding number of features selected q, and threshold in selection proportion π. Results are presented as ordered selection proportions of the variables at the calibrated λ (B) and graph representation of the detected and missed edges (D). Selection performances are assessed in terms of numbers of True Positives (dark blue), False Positives (red dashed line) and False Negatives (green dotted line). Calibration of the stability selection graphical LASSO ensures that the expected number of False Positives (PFER) is below 20. The data is simulated for $p = 50$ variables and $n = 100$ observations, with $\nu = 0.2$ for the regression model, and an underlying scale-free structure for the graphical model.

depending on the constraint, but our simulation shows that differences in the maximal stability score (Supplementary Figure S3-B) and resulting selected variables are small (Supplementary Figure S3-C) if the constraint is not too stringent.

Our stability score is also used to calibrate the graphical LASSO for the estimation of a conditional independence graph, while controlling the expected number of falsely selected edges below 20 (Figure 1-C). The calibrated graph (Figure 1-D) included 56 (47 rightly, in plain dark blue and 9 wrongly, in dashed red lines) stably selected edges (i.e. with selection proportions $\geq \hat{\pi} = 0.90$), based on graphical LASSO models fitted on 50% of the data with penalty parameter $\lambda = 0.52$. The 9 wrongly selected edges tend to be between nodes that are otherwise connected in this example (marginal links). The 2 missed edges are connected to the central hub and thus correspond to smaller partial correlations, more difficult to detect.

3.3. Evaluation of model performance and comparison with existing approaches

Our simulations show that models with higher stability score yield higher selection performances (as measured by the F_1-score), making it a relevant metric for calibration (Figure 2-A)). We also find that irrespective of the value of λ and π, stability selection models outperform the original implementation of the graphical LASSO (Figure 2-B). Graphical LASSO calibrated using the BIC or EBIC (see Supplementary Methods, section 1.3) generate dense graphs resulting in high recall (1 and 0.97, respectively) and poor precision values (0.19 and 0.42). Our stability score instead yield sparser models, resulting is slightly lower recall values (0.91) which did not include many irrelevant edges, as captured by the far better precision value (0.78). Our simulation also shows that the constraint controlling the PFER further improves the overall performance through the generation of a sparser model.

Our calibrated stability selection models are compared with state-of-the-art graphical model estimation approaches on 1,000 simulated datasets in low, intermediate and high-dimension (Figure 3, Supplementary Table S1). Non stability-enhanced graphical LASSO models, calibrated using information theory criteria, are generally the worst performing models (median F_1-score < 0.6 across dimensionality settings). StARS models, applied with the same number of subsampling iterations and using default values for other parameters, have the highest median numbers of True Positives. However, they include more False Positives than stability selection models, making it less competitive in terms of F_1-score (best performing in high-dimension with a median F_1-score of 0.65). For stability selection models calibrated using error control (MB (Meinshausen and Bühlmann, 2010), SS (Shah and Samworth, 2013)), the optimal choice of π seems to depend on many parameters including the dimensionality and structure of the graph (Supplementary Figure S4). By jointly calibrating the two parameters, our models show generally better performances compared to models calibrated solely using error control on these simulations (median F_1-score ranging from 0.68 to 0.72 using $\text{PFER}_{SS} < 20$ only in high dimension, compared to 0.73 using constrained calibration maximising the stability score). Results were consistent when using different thresholds in PFER (Supplementary Figure S5). Computation times of all stability selection models are comparable and acceptable in practice (less than 3 minutes in these settings, Supplementary Table S1).

12   *B Bodinier et al.*
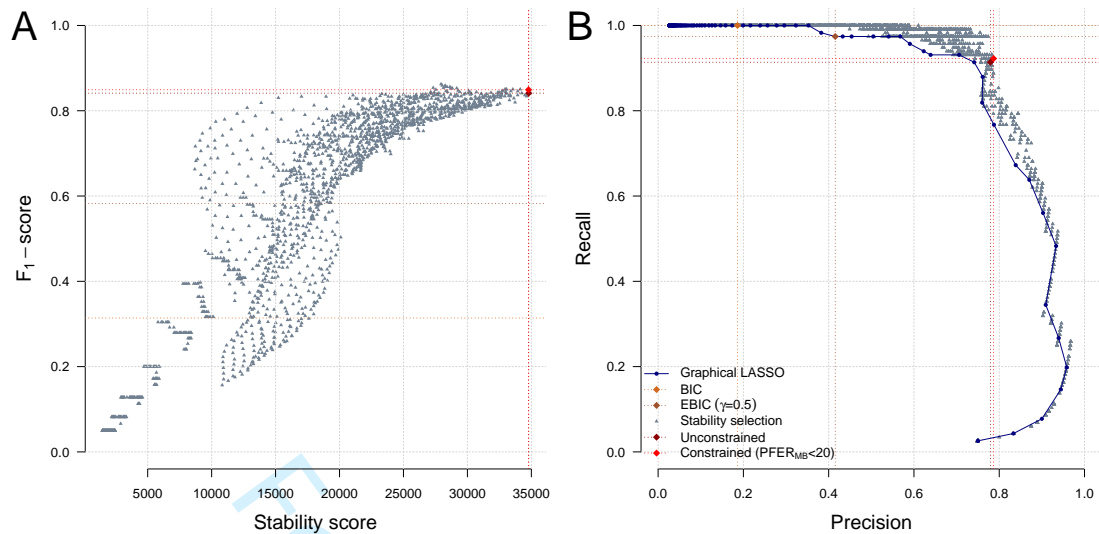
Fig. 2. Selection performance in stability selection and relevance of the stability score for calibration. The F_1-score of stability selection models fitted with a range of λ and π values is represented as a function of the stability score (A). Calibrated stability selection models using the unconstrained (dark red) and constrained (red) approaches are highlighted. The precision and recall of visited stability selection models (grey) and corresponding graphical LASSO models (dark blue) are reported (B). The calibrated models using the BIC (beige) or EBIC (brown) are also showed (B).

3.4. Sensitivity to the choice of resampling procedure

Stability selection can be implemented with different numbers of iterations K and resampling techniques (subsampling, bootstrap or CPSS approaches, and subsample size). We show in a simulation study with $p = 100$ nodes that (a) the effect of the number of iterations K reaches a plateau after 50 of iterations, and (b) that the best performances were obtained for bootstrap samples or subsamples of 50% of the observations (Supplementary Figures S6 and S7).

3.5. Multi-block extension for graphical models

Our single and multi-block calibration procedures are applied on simulated datasets with a block structure in different dimensionality settings. Block specific selection performances of both approaches can be visualised in precision-recall plots (Figure 4, Supplementary Table S2). Irrespective of the dimensionality, accounting for the block structure as proposed in Equation (5) with $\lambda_0 = 0.1$ generates an increase in selection performance in both within and between blocks (up to 7% in overall median F_1-score in low dimension). This gain in performance comes at the price of an increased computation time (from 2 to 6 minutes in low dimension).

Additionally, we show in Supplementary Table S3 that the choice of λ_0 has limited effects on the selection performances, as long as it is relatively small ($\lambda_0 \leq 0.1$). We choose $\lambda_0 = 0.1$ for a good balance between performance and computation time. We
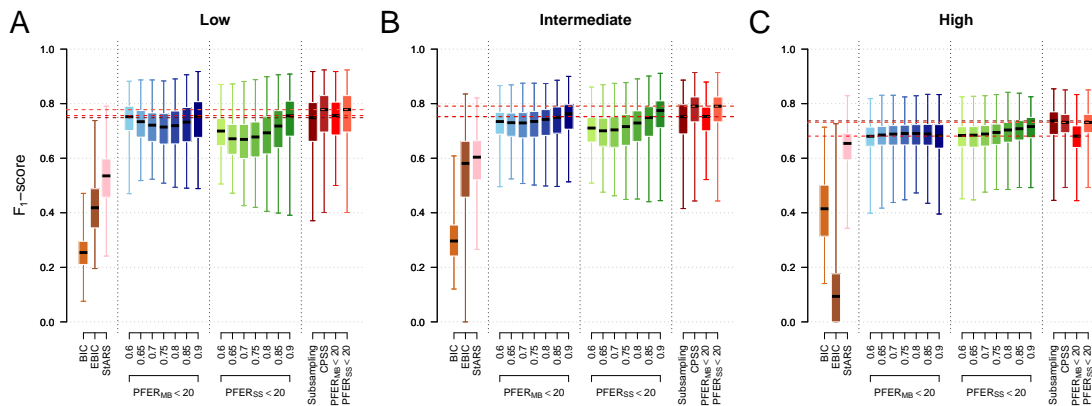
Fig. 3. Selection performances of state-of-the-art approaches and proposed calibrated stability selection models. Median, quartiles, minimum and maximum F_1-score of graphical LASSO models calibrated using the BIC, EBIC, StARS, and stability selection graphical LASSO models calibrated via error control (MB in blue, SS in green) or using the proposed stability score (red). Models are applied on 1,000 simulated datasets with $p = 100$ variables following a multivariate Normal distribution corresponding to a random graph structure ($\nu = 0.02$). Performances are estimated in low ($n = 2p$, A), intermediate ($n = p$, B), and high ($n = p/2$, C) dimensions.

also show that the use of Equation (5) gives better selection performances than that of Equation (4) (median F_1-score ≥ 0.71 compared to 0.57). In particular, it drastically reduces the numbers of False Positives in the off-diagonal block.

4. Application: molecular signature of smoking

4.1. Data overview

DNA methylation and gene expression data were previously measured in the same plasma samples from 251 women from the Norwegian Women and Cancer (NOWAC) study (Sandanger et al., 2018). This nested prospective case-control study of lung cancer includes 125 future cases (mean time-to-diagnosis of 4 years) and 126 healthy controls. The data was pre-processed as previously described (Guida et al., 2015). DNA methylation at each CpG site are originally expressed as a proportion of methylated sequences across all copies (β-values). We applied a logit$_2$-transformation of the methylation proportions to ensure these are real numbers (M-values). The gene expression data was normalised using log-transformation. Features missing in more than 30% of the samples were excluded, and the remaining data was imputed using the k-nearest neighbour algorithm. To remove technical confounding, the data was de-noised by extracting the residuals from linear mixed models with the OMIC feature as the outcome and modelling technical covariates (chip and position) as random intercepts (Sandanger et al., 2018).

14   *B Bodinier et al.*
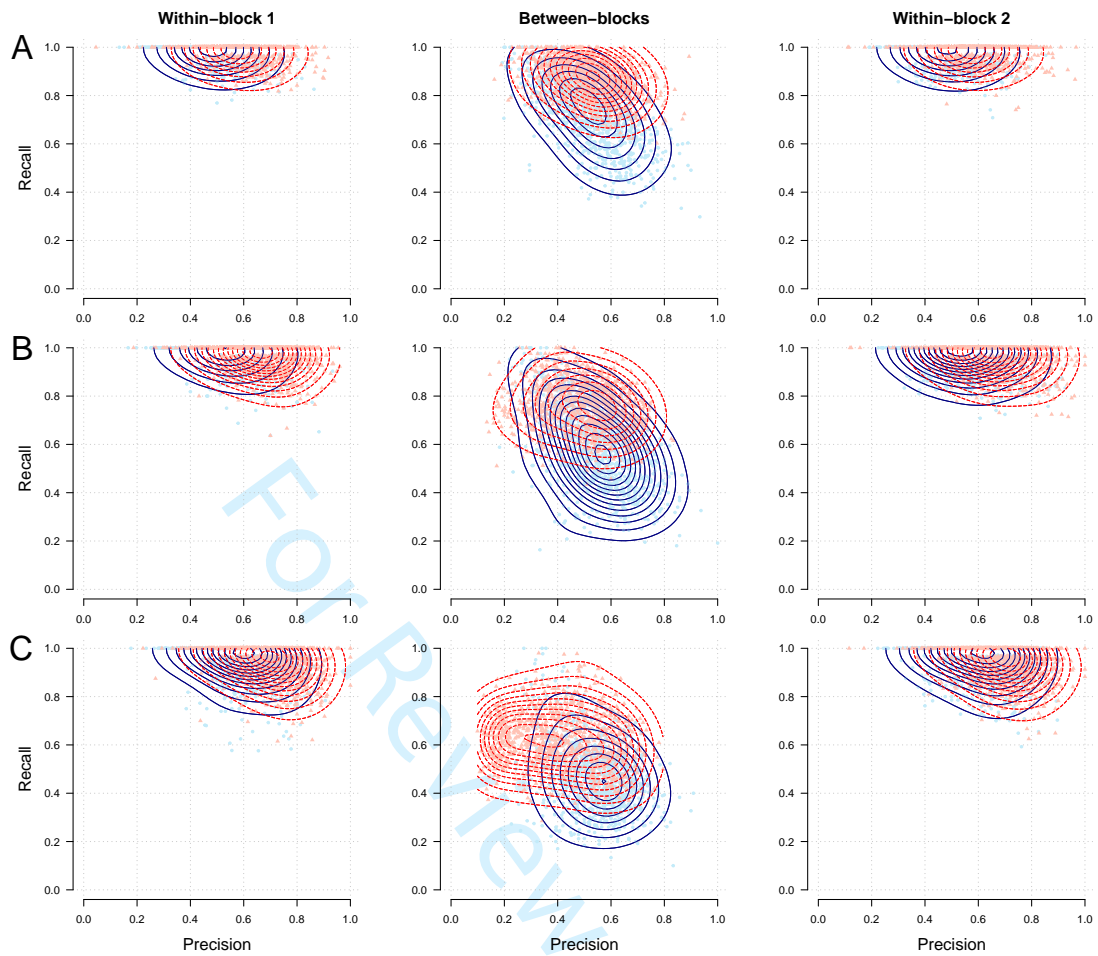
Fig. 4. Precision-recall showing single and multi-block stability selection graphical models applied on simulated data with a block structure. Models are applied on 1,000 simulated datasets (points) with $p = 100$ variables following a multivariate Normal distribution corresponding to a random graph ($\nu = 0.02$) and with known block structure (50 variables per group, using $v_b = 0.2$). The contour lines indicate estimated 2-dimensional density distributions. Performances are evaluated in low (A, $n = 2p$), intermediate (B, $n = p$), and high (C, $n = p/2$) dimensions.

4.2. *Targeted smoking-related markers*

We focused on a set of 160 CpG sites found differentially methylated in never *vs.* former smokers at a 0.05 Bonferroni corrected significance level in a large meta-analysis of 15,907 participants from 16 different cohorts (Joehanes et al., 2016). Similarly, we selected a set of 156 transcripts found differentially expressed in never and current smokers from a meta-analysis including 10,233 participants from 6 cohorts (Huan et al., 2016). Of these, 159 CpG sites and 142 transcripts were assayed in our dataset.

4.3. Multi-OMICs graph

In order to get a more comprehensive understanding of the biological response to smoking we integrate methylation data, known to reflect long-term exposure to tobacco smoking, and gene expression, which is functionally well characterised, and seek for correlation patterns across these smoking-related signals via the estimation of a multi-OMICs graph.

We accommodate the heterogeneous data structure (Supplementary Figure S8) by calibrating three pairs of block-specific parameters (λ, π) using our multi-block strategy (Figure 5-A). We found a total of 601 edges, including 150 in the within-methylation block, 425 in the within-gene expression block, and 26 cross-OMICs edges (Figure 5-B). The detected links reflect potential participation to common regulatory processes of both transcripts and CpG sites. As our analysis was limited to smoking-related markers, connected nodes can be hypothesised to jointly contribute to the biological response to tobacco smoking.

For comparison, we estimate the graphical LASSO model calibrated using the BIC on the same data (Supplementary Figure S9). Of the 601 edges included in the stability selection graph, 583 were also in the BIC-calibrated graph. The BIC-calibrated graph is more dense (N=6,744 edges), which makes it difficult to interpret. As this procedure does not account for the block structure in the data, two modules corresponding to the two platforms are clearly visible.

DNA methylation nodes are annotated with the symbol of its closest gene on the genome (Joehanes et al., 2016). Most sets of CpG sites annotated with the same gene symbol are interconnected in the graph (e.g. AHRR, GNG12-AS1, and ALPPL2 on chromosomes 5, 3, and 2, respectively). The data includes a CpG site and a transcript with the same annotation for two genes, but only found a cross-OMIC link for LRRN3 (Guida et al., 2015). Such strong correlations involving features that are closely located on the genome, or cis-effects, have been reported previously (Robinson et al., 2014).

Our approach also detects edges involving features that are not closely located on the genome, and are even on different chromosomes (Supplementary Figure S10). This suggests that complex long-range mechanisms could be at stake (Jones, 2012).

Closer examination of the graph highlights the LRRN3 transcript (Guida et al., 2015), encoding a transmembrane signal receptor, which is linked to 4 CpG sites, including AHRR, ALPPL2 and a CpG site annotated as LRRN3 (cg09837977). It is the only transcript with such a central position among methylation markers (Figure 5-B).

We incorporate functional information in the visualisation using Reactome pathways (Figure 5-B) (Langfelder and Horvath, 2008; Jassal et al., 2020). As previously reported, the immune system and in signal transduction (red) pathways were largely represented in the targeted set (Huan et al., 2016; Sandanger et al., 2018). Interestingly, the group of interconnected nodes around RPL4 (green) was involved in a range of pathways including the cellular response to stress, translation, and developmental biology. Similarly, the transcripts involved in the metabolism of lipids (yellow) are closely related in the graph. Altogether these results confirm the functional proximity of the nearby variables from our graph, hence lending biological plausibility of its topology.

16   *B Bodinier et al.*
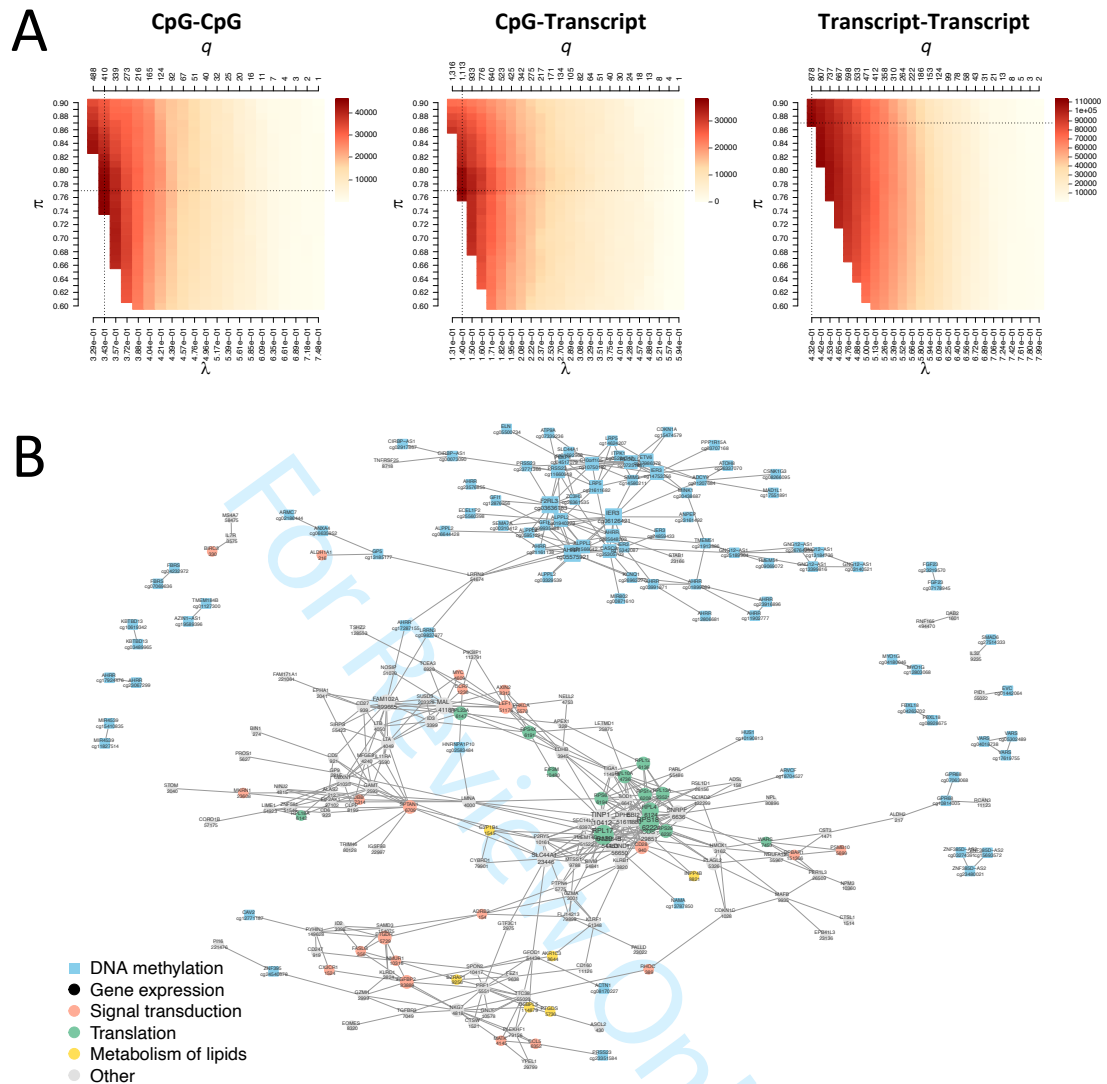
Fig. 5. Multi-OMICs graphical model integrating DNA methylation and gene expression markers of tobacco smoking. The three blocks are calibrated separately using models where the other blocks are weakly penalised (λ_0=0.1), while ensuring that PFER$_{MB} < 110$ overall (A). The stability selection model includes edges that are stably selected in each block (B).

5. Discussion

The development of stability-enhanced models accommodating data with a known block structure we proposed was triggered by the multi-OMICs application for the characterisation of the molecular signature of smoking. Their application to methylation and gene expression data gave further insights on the long-range correlations previously reported (Guida et al., 2015), and revealed a credible pivotal cross-OMICs role of the LRRN3

transcript (Huan et al., 2016). Annotation of the networks using biological information from the Reactome database identifies modules mostly composed of nodes belonging to the same pathways, suggesting that statistical correlations can reflect functional role in shared biological pathways.

The selection performances of our (multi-block) graphical models were evaluated in a simulation study. We showed that stability selection models yield higher F_1-score, to the cost of a (limited) increase in computation time. The computational efficiency of the proposed approaches can easily be improved through parallelisation and the use of warm start, where models are iteratively fitted over a path from larger to smaller penalty values and the estimate from the previous iteration is a starting point for the gradient descent algorithm (Friedman et al., 2010). We also demonstrated that the proposed calibration procedure is generally identifying the optimal threshold in selection proportion which leads to overall equivalent or better performances than previously proposed approaches based solely on error control. Our multi-block extension was successful in removing some of the technical bias through a more flexible modelling, but generated a ten-fold increase in computation time compared to single-block models on these simulations.

Our simulation studies indicate that, even when the assumptions of the model are verified (including the multivariate Normal distribution), the estimations of the graphical models are not perfect and need to be interpreted with care. In particular, some of the edges selected may correspond to marginal relationships (and not true conditional links). On the other hand, the absence of an edge does not necessarily indicate that there is no conditional association between the two nodes (especially for cross-group edges, for which the signal is diluted). Reassuringly, the overall topology of the graph seems relevant, as observed when applied on data with a scale-free graphical structure.

The stability selection approach and calibration procedure introduced here could also be used in combination with other variable selection algorithms, including penalised unsupervised models that cannot rely on the minimisation of an error term in cross-validation (Zou et al., 2006), or extensions modelling longitudinal (Charbonnier et al., 2010) or count data (Chiquet et al., 2018). The method comes with some level of flexibility and user-controlled choices. Depending on the application and its requirements, the models can be tailored to generate more or less conservative results using (a) the threshold in PFER controlling the sparsity of the selected sets, and (b) considering features with intermediate selection proportions (between $1 - \pi$ and π).

The main limitation of the stability-based approach is its scalability. So far, (multi-block) stability selection with the graphical LASSO can be applied on up to a few hundreds of features. All stability-based graphical models presented in this paper were estimated in a few minutes, however the complexity of the problem is rapidly increasing with the number of nodes. Despite the possible parallelisation and recent faster implementations of the models (Sustik M.A., 2012), generated graphs can become very dense and more efficient ways of visualising and summarising the results will be needed. Alternatively, as structures of redundant interconnected nodes (cliques) can be observed, summarising these in super-nodes could be valuable. This could be achieved using clustering or dimensionality reduction approaches, or by incorporating a priori biological knowledge in the model.

18   *B Bodinier et al.*

6.   Data Availability Statement

7.   Acknowledgements

References

Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213. New York, NY: Springer New York.

Albert, R. and A.-L. Barabási (2002, Jan). Statistical mechanics of complex networks. *Rev. Mod. Phys. 74*, 47–97.

Ambroise, C., J. Chiquet, and C. Matias (2009). Inferring sparse gaussian graphical models with latent structure. *Electron. J. Statist. 3*, 205–238.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008, June). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res. 9*, 485–516.

Barabási, A.-L. and Z. N. Oltvai (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics 5*(2), 101–113.

Canzler, S., J. Schor, W. Busch, K. Schubert, U. E. Rolle-Kampczyk, H. Seitz, H. Kamp, M. von Bergen, R. Buesen, and J. Hackermüller (2020). Prospects and challenges of multi-omics data integration in toxicology. *Archives of Toxicology 94*(2), 371–388.

Chadeau-Hyam, M., G. Campanella, T. Jombart, L. Bottolo, L. Portengen, P. Vineis, B. Liquet, and R. C. Vermeulen (2013). Deciphering the complex: Methodological overview of statistical models to derive omics-based biomarkers. *Environmental and Molecular Mutagenesis 54*(7), 542–557.

Charbonnier, C., J. Chiquet, and C. Ambroise (2010). Weighted-lasso for structured network inference from time course data. *Statistical Applications in Genetics and Molecular Biology 9*(1).

Chiquet, J., M. Mariadassou, and S. Robin (2018). Variational inference for sparse network reconstruction from count data.

Erdös, P. and A. Rényi (1959). On random graphs i. *Publicationes Mathematicae Debrecen 6*, 290–297.

Foygel, R. and M. Drton (2010). Extended bayesian information criteria for gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 604–612. Curran Associates, Inc.

Friedman, J., T. Hastie, and R. Tibshirani (2007, 12). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles 33*(1), 1–22.

Giraud, C. (2008). Estimation of gaussian graphs by model selection. *Electron. J. Statist. 2*, 542–563.

Guida, F., T. M. Sandanger, R. Castagné, G. Campanella, S. Polidoro, D. Palli, V. Krogh, R. Tumino, C. Sacerdote, S. Panico, G. Severi, S. A. Kyrtopoulos, P. Georgiadis, R. C. Vermeulen, E. Lund, P. Vineis, and M. Chadeau-Hyam (2015, 01). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human Molecular Genetics 24*(8), 2349–2359.

Huan, T., R. Joehanes, C. Schurmann, K. Schramm, L. C. Pilling, M. J. Peters, R. Mägi, D. DeMeo, G. T. O'Connor, L. Ferrucci, A. Teumer, G. Homuth, R. Biffar, U. Völker, C. Herder, M. Waldenberger, A. Peters, S. Zeilinger, A. Metspalu, A. Hofman, A. G. Uitterlinden, D. G. Hernandez, A. B. Singleton, S. Bandinelli, P. J. Munson, H. Lin, E. J. Benjamin, T. Esko, H. J. Grabe, H. Prokisch, J. B. van Meurs, D. Melzer, and D. Levy (2016, 08). A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. *Human Molecular Genetics 25*(21), 4611–4623.

Jassal, B., L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio (2020). The reactome pathway knowledgebase. *Nucleic Acids Res 48*(D1), D498–d503.

Joehanes, R., A. C. Just, R. E. Marioni, L. C. Pilling, L. M. Reynolds, P. R. Mandaviya, W. Guan, T. Xu, C. E. Elks, S. Aslibekyan, H. Moreno-Macias, J. A. Smith, J. A. Brody, R. Dhingra, P. Yousefi, J. S. Pankow, S. Kunze, S. H. Shah, A. F. McRae, K. Lohman, J. Sha, D. M. Absher, L. Ferrucci, W. Zhao, E. W. Demerath, J. Bressler, M. L. Grove, T. Huan, C. Liu, M. M. Mendelson, C. Yao, D. P. Kiel, A. Peters, R. Wang-Sattler, P. M. Visscher, N. R. Wray, J. M. Starr, J. Ding, C. J. Rodriguez, N. J. Wareham, M. R. Irvin, D. Zhi, M. Barrdahl, P. Vineis, S. Ambatipudi, A. G.

20   *B Bodinier et al.*

Uitterlinden, A. Hofman, J. Schwartz, E. Colicino, L. Hou, P. S. Vokonas, D. G. Hernandez, A. B. Singleton, S. Bandinelli, S. T. Turner, E. B. Ware, A. K. Smith, T. Klengel, E. B. Binder, B. M. Psaty, K. D. Taylor, S. A. Gharib, B. R. Swenson, L. Liang, D. L. DeMeo, G. T. O'Connor, Z. Herceg, K. J. Ressler, K. N. Conneely, N. Sotoodehnia, S. L. R. Kardia, D. Melzer, A. A. Baccarelli, J. B. J. van Meurs, I. Romieu, D. K. Arnett, K. K. Ong, Y. Liu, M. Waldenberger, I. J. Deary, M. Fornage, D. Levy, and S. J. London (2016). Epigenetic signatures of cigarette smoking. *Circulation: Cardiovascular Genetics 9*(5), 436–447.

Jones, P. A. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics 13*(7), 484–492.

Langfelder, P. and S. Horvath (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics 9*(1), 559.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica 16*(4), 1273–1284.

Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, Red Hook, NY, USA, pp. 1432–1440. Curran Associates Inc.

Meinshausen, N. and P. Bühlmann (2006, 06). High-dimensional graphs and variable selection with the lasso. *Ann. Statist. 34*(3), 1436–1462.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Müller, C. L., R. A. Bonneau, and Z. D. Kurtz (2016, May). Generalized stability approach for regularized graphical models. *pre-print*.

National Center for Chronic Disease, P., H. P. U. O. on Smoking, and Health (2014). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Atlanta (GA): Centers for Disease Control and Prevention (US).

Niedzwiecki, M. M., D. I. Walker, R. Vermeulen, M. Chadeau-Hyam, D. P. Jones, and G. W. Miller (2019). The exposome: Molecules to populations. *Annu Rev Pharmacol Toxicol 59*, 107–127.

Noor, E., S. Cherkaoui, and U. Sauer (2019). Biological insights through omics data integration. *Current Opinion in Systems Biology 15*, 39 – 47. Gene regulation.

Robinson, M. D., A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber, and X. Zhou (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics 5*, 324.

Sandanger, T. M., T. H. Nøst, F. Guida, C. Rylander, G. Campanella, D. C. Muller, J. van Dongen, D. I. Boomsma, M. Johansson, P. Vineis, R. Vermeulen, E. Lund, and M. Chadeau-Hyam (2018). Dna methylation and associated gene expression in

blood prior to lung cancer diagnosis in the norwegian women and cancer cohort. *Sci Rep 8*(1), 16714.

Schwarz, G. (1978, 03). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Shah, R. D. and R. J. Samworth (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 55–80.

Sustik M.A., C. B. (2012). Glassofast: An efficient glasso implementation. *UTCS Technical Report TR-12-29:1-3*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Valcárcel, B., P. Würtz, N.-K. Seich al Basatena, T. Tukiainen, A. J. Kangas, P. Soininen, M.-R. Järvelin, M. Ala-Korpela, T. M. Ebbels, and M. de Iorio (2011, 09). A differential network approach to exploring differences between biological states: An application to prediabetes. *PLOS ONE 6*(9), 1–9.

Vermeulen, R., F. Saberi Hosnijeh, B. Bodinier, L. Portengen, B. Liquet, J. Garrido-Manriquez, H. Lokhorst, I. A. Bergdahl, S. A. Kyrtopoulos, A.-S. Johansson, P. Georgiadis, B. Melin, D. Palli, V. Krogh, S. Panico, C. Sacerdote, R. Tumino, P. Vineis, R. Castagné, M. Chadeau-Hyam, on behalf of the EnviroGenoMarkers Consortium Consortium members, M. Botsivali, A. Chatziioannou, I. Valavanis, J. C. Kleinjans, T. M. de Kok, H. C. Keun, T. J. Athersuch, R. Kelly, P. Lenner, G. Hallmans, E. G. Stephanou, A. Myridakis, M. Kogevinas, L. Fazzo, M. De Santis, P. Comba, B. Bendinelli, H. Kiviranta, P. Rantakokko, R. Airaksinen, P. Ruokojarvi, M. Gilthorpe, S. Fleming, T. Fleming, Y.-K. Tu, T. Lundh, K.-L. Chien, W. J. Chen, W.-C. Lee, C. Kate Hsiao, P.-H. Kuo, H. Hung, and S.-F. Liao (2018). Pre-diagnostic blood immune markers, incidence and progression of b-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses. *International Journal of Cancer 143*(6), 1335–1347.

Zhao, Roeder, Lafferty, and Wasserman (2012). The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research 13*, 1059–1062.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.