Supplementary Material: Automated calibration for stability selection in penalised regression and graphical models

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.*
1. **Supplementary Methods: existing calibration strategies**

In both LASSO-regularised regression and graphical modelling, the calibration of the hyper-parameter $\lambda$ is critical as it regulates the size of the set of selected features. State-of-the-art approaches for the choice of $\lambda$ are based on M-fold cross-validation minimising some error metric (e.g. Mean Squared Error in Prediction). For graphical models, information theory metrics are commonly used, including the Akaike, Bayesian, and Extended Bayesian Information Criterion (Akaike, 1998; Schwarz, 1978; Foygel and Drton, 2010; Chiquet et al., 2009):

$$AIC_\lambda = -2\ell(\hat{\Omega}_\lambda) + 2|E_\lambda|$$

$$BIC_\lambda = -2\ell(\hat{\Omega}_\lambda) + \log(n)|E_\lambda|$$

$$EBIC_\lambda = -2\ell(\hat{\Omega}_\lambda) + \log(n)\left(|E_\lambda| + 4\gamma \log(p)\right)$$

where $\ell(\hat{\Omega}_\lambda) = \frac{n}{2} \log \det (\hat{\Omega}_\lambda) - \text{tr}(\hat{\Omega}_\lambda S)$ is the penalised likelihood, $|E_\lambda|$ is the degrees of freedom (i.e. number of edges in the graph), and $\gamma$ is a hyper-parameter specific to the EBIC.

**References**

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

Chiquet, J., A. Smith, G. Grasseau, C. Matias, and C. Ambroise (2009, February). SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics -Oxford- 25*(3), 417–8.

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.

Schwarz, G. (1978, 03). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.
Supplementary Figure S1: Simulation of data with a block correlation structure. The heatmaps show Pearson’s correlation (A) and partial correlation (B) matrices estimated on real data from 50 randomly chosen DNA methylation and gene expression markers. The bottom panel shows Pearson’s correlation (C) and partial correlation (D) matrices estimated on simulated data with $n = 200$ observations and a block structure (50 variables in each group). The simulated conditional independence structure between the $p = 100$ variables is that of a random graph ($\nu = 0.04$, $v_b = 0.2$). All partial correlations are estimated without penalisation.
Supplementary Figure S2: Choice of the value of parameter \( u \) for simulation of the precision matrix. The contrast of the simulated correlation matrix for a scale-free graphical model with \( p = 50 \) nodes and \( n = 100 \) observations is represented as a function of the parameter \( u \) on the log-scale. The chosen value for \( u \) is the one maximising the contrast (indicated by a red dashed line). The heatmaps of correlation matrices with extreme and calibrated values of the parameter \( u \) are showed.
Supplementary Figure S3: Visualisation of the PFER constraint in calibration of stability selection models. The calibration heatmap shows the stability score (colour-coded) as a function of $\lambda$ (or the corresponding average number of selected variables $q$) and $\pi$ (A). The white area (left) represents models for which the PFER computed using the Meinshausen and Bühlmann approach would exceed the threshold ($PFER_{MB} > 5$). The highest stability score obtained for a given penalty parameter $\lambda$ is represented for the unconstrained (blue) and constrained (red dotted line) approaches (B). Ordered selection proportions obtained from constrained calibration are reported (C). Stability selection is applied on simulated data with $n = 100$ observations for $p = 50$ variables, of which 10 contribute to the definition of the outcome with effect sizes in $\{-1, -0.5 \cup [0.5, 1]\}$ and an expected proportion of explained variance of 70%.
Supplementary Figure S4: Selection performances of state-of-the-art approaches and proposed calibrated stability selection graphical LASSO models applied on simulated data with scale-free underlying graph structure. We show the 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 ($n = 2p = 200$), intermediate ($n = p = 100$), and high ($n = p/2 = 50$) dimensions.
Supplementary Figure S5: Selection performances of state-of-the-art approaches and proposed calibrated stability selection graphical LASSO models with different thresholds in PFER. We show the 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). The threshold in PFER for stability selection models was set to 10 (A), 30 (B) or 50 (C). 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 ($n = 2p = 200$), intermediate ($n = p = 100$), and high ($n = p/2 = 50$) dimensions.
Supplementary Figure S6: Selection performances of state-of-the-art approaches and proposed calibrated stability selection LASSO models applied on simulated data. Models are applied on 1,000 simulated datasets with \( n = 500 \) observations and \( p = 1,000 \) predictor variables, of which an expected proportion \( \nu_Y = 0.02 \) contributes to the definition of the outcome with effect sizes in \{−1, 1\} and an expected proportion of explained variance of 40%. We simulate independent predictors, conditionally on the outcome (A) or blocks of correlated predictors where the conditional independence structure within blocks is that of a random network of density \( \nu = 0.02 \) (B). For both settings, we represent a heatmap of Pearson’s correlations between predictors in a typical simulation. We show the median, quartiles, minimum and maximum \( F_1 \)-score, precision and recall of LASSO models calibrated by 10-fold cross validation minimising the Mean Squared Error of Prediction (\( \lambda_{\text{min}} \)) or one standard error away from the minimum (\( \lambda_{\text{se}} \)), and stability selection LASSO models calibrated via error control (MB in blue, SS in green) or using the proposed stability score (red).
Supplementary Figure S7: Effect of the number of subsampling iterations $K$ on the selection performance and computation time. The median, 5$^{th}$ and 95$^{th}$ quantile of the $F_1$-score and computation time are reported for graphical LASSO stability selection models calibrated using the unconstrained approach and with different numbers of iterations $K$ (10, 20, 50, 100, 500, 1,000, 2,000, and 5,000). The models are applied on simulated data ($p = 100$) with underlying random graph structure ($\nu = 0.02$). The computation time in seconds is reported on the log-scale (X-axis). Performances are evaluated in low ($n = 2p = 200$), intermediate ($n = p = 100$), and high ($n = p/2 = 50$) dimensions.
Supplementary Figure S8: Effect of the choice of resampling technique on the selection performance. The median, 5th and 95th quantile of the $F_1$-score are reported for stability selection models calibrated using the unconstrained approach with different resampling approaches: subsampling with different subsample sizes $\tau$ between 0.1 and 0.9 (red), simultaneous selection in complementary pairs (CPSS, in dark blue) and bootstrapping (resampling with replacement, dark green). Performances are evaluated in low ($n = 2p = 200$), intermediate ($n = p = 100$), and high ($n = p/2 = 50$) dimensions.
Supplementary Figure S9: Single-block graphical model of DNA methylation markers of exposure to tobacco smoking. Calibration is done by maximising the stability score while ensuring that $\text{PFER}_{MB} < 70$ (A). CpG sites with at least one edge are represented in the graph (B).
Supplementary Figure S10: Heatmap of Pearson’s correlations estimated from measured levels of the 159 DNA methylation markers and 208 gene expression markers.
Supplementary Figure S11: Graphical LASSO model of smoking-related methylation (blue square) and gene expression (red circle) markers calibrated using the Bayesian Information Criterion (BIC). The BIC is represented as a function of the penalty parameter $\lambda$ (A). The graphical model generating the smallest BIC is showed (B).
Supplementary Figure S12: Multi-OMICs graphical model integrating DNA methylation (square) and gene expression (circle) markers of tobacco smoking with nodes coloured by chromosome.
**Supplementary Table S1: Median and inter-quartile range of the selection performance metrics and computation times obtained with different graphical models.** Models are applied on 1,000 simulated datasets with \( p = 100 \) variables following a multivariate Normal distribution corresponding to a random graph structure (\( \nu = 0.02 \) in low \((L, n = 2p = 200)\), intermediate \((I, n = p = 100)\), and high \((H, n = p/2 = 50)\) dimensions.

|             | \( n \) | TP | FP | FN | Precision | Recall | \( F_1 \)-score | Time (s) |
|-------------|--------|----|----|----|-----------|--------|----------------|----------|
| **AIC**     | 99.14  | 1990 | 148 | 10 | 0.00        | 0.00   | 0.00           | 0.00     |
|             | 99.13  | 564 | 205 | 0  | 0.00        | 0.00   | 0.00           | 0.00     |
|             | 98.13  | 264 | 148 | 0  | 0.00        | 0.00   | 0.00           | 0.00     |
| **EBC**     | 98.12  | 162 | 84  | 6  | 0.00        | 0.00   | 0.00           | 0.00     |
| **SARS**    | 95.12  | 61  | 41  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |

**Supplementary:** Automated calibration for stability selection in penalised regression and graphical models. Models are applied on 1,000 simulated datasets with \( n = 100 \) variables following a multivariate Normal distribution corresponding to a random graph structure (\( \nu = 0.02 \) in low \((L, n = 2p = 200)\), intermediate \((I, n = p = 100)\), and high \((H, n = p/2 = 50)\) dimensions.

|             | \( n \) | TP | FP | FN | Precision | Recall | \( F_1 \)-score | Time (s) |
|-------------|--------|----|----|----|-----------|--------|----------------|----------|
|             | 94.11  | 66  | 30  | 5  | 0.00        | 0.00   | 0.00           | 0.00     |
| **MB**      | 95.11  | 68  | 36  | 4  | 0.00        | 0.00   | 0.00           | 0.00     |
|             | 95.12  | 62  | 40  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |
| **SS**      | 95.10  | 56  | 41  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |

**Supplementary:** Automated calibration for stability selection in penalised regression and graphical models. Models are applied on 1,000 simulated datasets with \( n = 100 \) variables following a multivariate Normal distribution corresponding to a random graph structure (\( \nu = 0.02 \) in low \((L, n = 2p = 200)\), intermediate \((I, n = p = 100)\), and high \((H, n = p/2 = 50)\) dimensions.

|             | \( n \) | TP | FP | FN | Precision | Recall | \( F_1 \)-score | Time (s) |
|-------------|--------|----|----|----|-----------|--------|----------------|----------|
|             | 94.10  | 54  | 46  | 4  | 0.00        | 0.00   | 0.00           | 0.00     |
| **CPSS**    | 94.10  | 46  | 40  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |
|             | 94.10  | 48  | 40  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |
| **EBIC**    | 94.10  | 52  | 39  | 5  | 0.00        | 0.00   | 0.00           | 0.00     |
|             | 94.10  | 48  | 40  | 3  | 0.00        | 0.00   | 0.00           | 0.00     |
Supplementary Table S2: Median and inter-quartile range of the computation times (in seconds) of stability selection obtained with different numbers of variables $p$. Models are applied on 1,000 simulated datasets with $n = 500$ observations. For stability selection LASSO models, we use $p = 1,000, 2,500, 5,000, 7,500$ or 10,000 independent predictors, conditionally on the outcome. For stability selection graphical LASSO models, we use $p = 100, 250, 500, 750$ or 1,000 variables following a multivariate Normal distribution corresponding to a random graph structure ($\nu = 0.02$). For graphical models, we report computation times with or without 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 (argument "start" in the R package sharp). For LASSO models, we always use warm start as implemented in the R package glmnet.

| LASSO | Graphical LASSO |
|-------|----------------|
| $p$   | $p$            | Cold start | Warm start |
| 1,000 | 18 [5]        | 100        | 69 [33]    | 51 [22]    |
| 2,500 | 35 [9]        | 250        | 313 [104]  | 247 [73]   |
| 5,000 | 59 [20]       | 500        | 2,759 [1,163] | 1,796 [658] |
| 7,500 | 86 [35]       | 750        | 14,513 [5,983] | 7,402 [3,472] |
| 10,000| 124 [63]      | 1,000      | 99,108 [29,563] | 40,240 [10,787] |
**Supplementary: Automated calibration for stability selection in penalised regression and graphical models**

|       | TP [9] | FP [47] | FN [12] | Precision [143] | Recall [104] | F1-score [0.89] | Time (s) [37] |
|-------|--------|---------|---------|-----------------|-------------|-----------------|--------------|
| Single Overall | 84 | 79 | 13 | 0.521 | 0.863 | 0.643 | 96 |
| Within 1 | 24 | 24 | 0 | 0.500 | 1.000 | 0.663 | 0.130 |
| Between | 36 | 30 | 13 | 0.545 | 0.745 | 0.604 | 0.105 |

|       | TP [6] | FP [15] | FN [0] | Precision [149] | Recall [0.00] | F1-score [0.127] | Time (s) [37] |
|-------|--------|---------|--------|-----------------|-------------|-----------------|--------------|
| L Within 2 | 24 | 24 | 0 | 0.500 | 1.000 | 0.667 | 0.127 |

|       | TP [34] | FP [6] | FN [6] | Precision [125] | Recall [0.056] | F1-score [0.87] | Time (s) [37] |
|-------|---------|--------|--------|-----------------|-------------|-----------------|--------------|
| Multi Overall | 93 | 73 | 13 | 0.561 | 0.941 | 0.703 | 269 |
| Within 1 | 24 | 16 | 0 | 0.585 | 1.000 | 0.737 | 0.102 |
| Between | 44 | 38 | 5 | 0.543 | 0.893 | 0.667 | 0.110 |
| Within 2 | 24 | 16 | 0 | 0.587 | 1.000 | 0.737 | 0.106 |

|       | TP [40] | FP [13] | FN [0] | Precision [149] | Recall [0.00] | F1-score [0.136] | Time (s) [37] |
|-------|---------|---------|--------|-----------------|-------------|-----------------|--------------|
| Single Overall | 24 | 24 | 0 | 0.550 | 1.000 | 0.704 | 107 |
| Within 1 | 24 | 19 | 0 | 0.566 | 0.582 | 0.553 | 0.107 |
| Between | 29 | 22 | 21 | 0.566 | 0.582 | 0.553 | 0.107 |
| Within 2 | 24 | 19 | 0 | 0.554 | 1.000 | 0.708 | 0.135 |

|       | TP [40] | FP [13] | FN [0] | Precision [152] | Recall [0.080] | F1-score [0.092] | Time (s) [37] |
|-------|---------|---------|--------|-----------------|-------------|-----------------|--------------|
| I Within 2 | 24 | 11 | 0 | 0.674 | 1.000 | 0.800 | 0.113 |
| Between | 35 | 11 | 8 | 0.526 | 0.769 | 0.614 | 0.141 |
| Within 2 | 24 | 11 | 0 | 0.676 | 1.000 | 0.800 | 0.115 |

|       | TP [40] | FP [13] | FN [10] | Precision [146] | Recall [0.109] | F1-score [0.072] | Time (s) [37] |
|-------|---------|---------|---------|-----------------|-------------|-----------------|--------------|
| Single Overall | 71 | 47 | 28 | 0.597 | 0.716 | 0.641 | 242 |
| Within 1 | 24 | 14 | 0 | 0.606 | 1.000 | 0.746 | 0.131 |
| Between | 23 | 18 | 26 | 0.559 | 0.465 | 0.496 | 0.110 |
| Within 2 | 24 | 11 | 0 | 0.676 | 1.000 | 0.800 | 0.115 |

|       | TP [75] | FP [13] | FN [11] | Precision [214] | Recall [0.093] | F1-score [0.143] | Time (s) [37] |
|-------|---------|---------|---------|-----------------|-------------|-----------------|--------------|
| H Within 2 | 77 | 71 | 21 | 0.519 | 0.578 | 0.627 | 534 |
| Between | 31 | 48 | 18 | 0.396 | 0.625 | 0.480 | 0.225 |
| Within 2 | 23 | 9 | 1 | 0.710 | 0.967 | 0.812 | 0.093 |

**Supplementary Table S3:** Median and inter-quartile range of the selection performance metrics and computation times obtained with single and multi-block stability selection applied on simulated data with a block structure. For each block, 50 different penalty parameter values are explored. Models are applied on 1,000 simulated datasets 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 $v_b = 0.2$). Performances are evaluated in low (L, $n = 2p = 200$), intermediate (I, $n = p = 100$), and high (H, $n = p/2 = 50$) dimensions.
### Supplementary Table S4: Median and inter-quartile range of the selection performance metrics and computation times obtained with different stability selection models on simulated data with a known block structure.

We compare stability selection not accounting for the block structure (Equation (1), denoted by S-B), using block-specific parameters (Equation (4), M-P) and combining block-specific models calibrated while using different penalties $\lambda_0$ for the other blocks (Equation (5), M-B). For each block, 30 different penalty parameter values are explored. Models are applied on 1,000 simulated datasets 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 $v_b = 0.2$). Performances are evaluated in low dimension ($n = 2p = 200$).

| $\lambda_0$ | TP | FP | FN | Precision | Recall | F1-score | Time (s) |
|-------------|----|----|----|-----------|--------|----------|-----------|
| S-B Overall | 85 [9] | 78 [44] | 14 [12] | 0.523 [0.134] | 0.859 [0.107] | 0.646 [0.081] | 70 [30] |
| Within 1 | 25 [7] | 24 [15] | 0 [6] | 0.500 [0.152] | 1.000 [0.000] | 0.667 [0.132] | |
| Between | 36 [11] | 29 [27] | 14 [11] | 0.547 [0.187] | 0.725 [0.209] | 0.603 [0.099] | |
| Within 2 | 24 [7] | 23 [14] | 0 [6] | 0.504 [0.139] | 1.000 [0.000] | 0.667 [0.123] | |
| M-P Overall | 92 [11] | 134 [55] | 6 [6] | 0.408 [0.109] | 0.935 [0.060] | 0.566 [0.098] | 2059 [870] |
| Within 1 | 24 [6] | 11 [1] | 0 [1] | 0.679 [0.212] | 1.000 [0.050] | 0.800 [0.138] | 8493 [15464] |
| Between | 45 [8] | 110 [51] | 5 [4] | 0.720 [0.473] | 0.737 [0.152] | 0.603 [0.099] | 82 |
| Within 2 | 23 [7] | 6 [5] | 1 [3] | 0.778 [0.152] | 0.944 [0.115] | 0.844 [0.091] | 82 |
| M-B 0 Overall | 82 [8] | 28 [48] | 17 [10] | 0.748 [0.105] | 0.944 [0.091] | 0.844 [0.091] | 82 |
| Within 1 | 23 [6] | 6 [5] | 1 [3] | 0.750 [0.178] | 0.958 [0.094] | 0.833 [0.104] | 82 |
| Between | 36 [7] | 114 [51] | 13 [11] | 0.720 [0.428] | 0.737 [0.152] | 0.603 [0.099] | 82 |
| Within 2 | 23 [7] | 6 [5] | 1 [3] | 0.778 [0.152] | 0.944 [0.115] | 0.844 [0.091] | 82 |
| 0.001 Overall | 83 [8] | 33 [65] | 15 [11] | 0.715 [0.125] | 0.817 [0.105] | 0.746 [0.119] | 8296 [1577] |
| Within 1 | 23 [6] | 6 [5] | 1 [3] | 0.750 [0.178] | 0.958 [0.094] | 0.833 [0.104] | 8296 [1577] |
| Between | 37 [7] | 16 [55] | 12 [9] | 0.705 [0.473] | 0.755 [0.152] | 0.689 [0.266] | 8296 [1577] |
| Within 2 | 23 [6] | 6 [5] | 1 [3] | 0.778 [0.152] | 0.944 [0.115] | 0.844 [0.091] | 8296 [1577] |
| 0.01 Overall | 85 [9] | 43 [99] | 11 [9] | 0.672 [0.201] | 0.839 [0.105] | 0.749 [0.146] | 83 [288] |
| Within 1 | 24 [6] | 10 [8] | 0 [1] | 0.699 [0.175] | 1.000 [0.045] | 0.812 [0.120] | 83 [288] |
| Between | 40 [8] | 20 [32] | 10 [8] | 0.671 [0.315] | 0.807 [0.139] | 0.714 [0.177] | 83 [288] |
| Within 2 | 23 [7] | 10 [8] | 0 [1] | 0.790 [0.165] | 1.000 [0.048] | 0.811 [0.113] | 83 [288] |
| 0.1 Overall | 95 [11] | 74 [33] | 9 [7] | 0.590 [0.124] | 0.938 [0.058] | 0.740 [0.084] | 152 [71] |
| Within 1 | 25 [6] | 17 [10] | 0 [0] | 0.591 [0.143] | 1.000 [0.000] | 0.741 [0.109] | 152 [71] |
| Between | 44 [8] | 37 [22] | 6 [6] | 0.550 [0.158] | 0.884 [0.105] | 0.671 [0.104] | 152 [71] |
| Within 2 | 24 [7] | 16 [10] | 0 [0] | 0.591 [0.133] | 1.000 [0.000] | 0.742 [0.101] | 152 [71] |
| 0.5 Overall | 95 [11] | 159 [96] | 4 [5] | 0.525 [0.107] | 0.955 [0.048] | 0.466 [0.115] | 84 [30] |
| Within 1 | 25 [7] | 30 [22] | 0 [0] | 0.451 [0.197] | 1.000 [0.000] | 0.620 [0.185] | 84 [30] |
| Between | 45 [9] | 137 [59] | 4 [5] | 0.247 [0.082] | 0.914 [0.091] | 0.389 [0.096] | 84 [30] |
| Within 2 | 24 [6] | 29 [21] | 0 [0] | 0.456 [0.188] | 1.000 [0.000] | 0.624 [0.177] | 84 [30] |
| 1 Overall | 95 [11] | 225 [99] | 4 [5] | 0.297 [0.102] | 0.964 [0.047] | 0.453 [0.114] | 75 [25] |
| Within 1 | 25 [7] | 30 [22] | 0 [0] | 0.447 [0.191] | 1.000 [0.000] | 0.618 [0.181] | 75 [25] |
| Between | 45 [9] | 165 [69] | 4 [5] | 0.218 [0.074] | 0.914 [0.087] | 0.350 [0.094] | 75 [25] |
| Within 2 | 24 [6] | 29 [21] | 0 [0] | 0.453 [0.187] | 1.000 [0.000] | 0.622 [0.176] | 75 [25] |