SUPPORTING INFORMATION

Supporting Information for “Flexible co-data learning for high-dimensional prediction"

Mirrelijn M. van Nee*[1] | Lodewyk F.A. Wessels[2,3,4] | Mark A. van de Wiel[1,5]

[1]Epidemiology & Data Science | Amsterdam Public Health Research Institute, Amsterdam University Medical Centers, Noord-Holland, The Netherlands
[2]Molecular Carcinogenesis, Netherlands Cancer Institute, Noord-Holland, The Netherlands
[3]Computational Cancer Biology, Oncode Institute, Noord-Holland, The Netherlands
[4]Intelligent Systems, Delft University of Technology, Zuid-Holland, The Netherlands
[5]MRC Biostatistics Unit, University of Cambridge, Cambridgeshire, UK

Correspondence
*Mirrelijn M. van Nee, De Boelelaan 1089a, 1081HV, Amsterdam, The Netherlands.
Email: m.vannee@amsterdamumc.nl

Funding Information
The first author is supported by ZonMw TOP grant COMPUTE CANCER (40-00812-98-16012).

Supporting Information

This document contains supporting information for “Flexible co-data learning for high-dimensional prediction".

A | DETAILS MODEL ESTIMATION

The Method of Moments (MoM) may be used to obtain moment estimates for the prior parameters, e.g. for obtaining group prior variance estimates for linear and logistic regression[1]. Whereas the theoretical moments needed for MoM are analytical for linear regression, Taylor approximations[2] are used and generalised to derive approximations for other generalised linear models (GLMs) using first and second order derivatives for GLMs[3]. Besides, the approximation is extended to include moment estimations for group prior mean parameters as well. This could be used if one would want to shrink all β not to 0, but to a target[4] where the target itself now is estimated based on the data. By default, we use a ridge penalty on the group level to ensure stable group variance estimates that are automatically shrunk towards an ordinary ridge prior weight when co-data is non-informative. Differences in group sizes are taken into account when shrinking group variance estimates. The penalty matrices used will first be assumed to be of full rank, which doesn’t hold in particular when unpenalised covariates are to be included. However, we can show that the MoM estimating equations can be derived independently of unpenalised covariates.

Below, we derive moment-estimates for group prior means and variances $\boldsymbol{\mu}, \boldsymbol{\gamma} \in \mathbb{R}^G$, keeping notation similar to previous notation[3] in order to retrieve estimating equations general for all GLMs. We then fill in details for linear, logistic and Cox survival regression, and show how to use the same estimating equations to obtain co-data weights when combining multiple co-data sets. After showing how to handle unpenalised covariates, we give the details of the ridge hyperpenalty function, of

handling continous co-data and of implemented posterior covariate selection approaches. Lastly, we provide an interpretation of the hyperparameter estimates.

A.1 | Generalised linear models and derivatives

Consider one co-data set coded by the co-data matrix $Z \in \mathbb{R}^{p\times G}$, leaving out all superscripts $^{(d)}$ for notational convenience. Each β_k is a priori Gaussian distributed with some covariate-specific mean μ_k and variance τ_k^2 which are a function of the group specific prior mean vector $\boldsymbol{\mu}_{G\times 1} \in \mathbb{R}^G$ and overall and local prior variance $\tau^2_{overall}$, $\boldsymbol{\gamma} \in \mathbb{R}^G$:

$$\beta_k \overset{ind.}{\sim} N(\mu_k, \tau_k^2) := N(\boldsymbol{Z}_k\boldsymbol{\mu}, \tau^2_{global}\boldsymbol{Z}_k\boldsymbol{\gamma}),\ k = 1, .., p.$$

Reparameterise by $\boldsymbol{\tau}^2_{G\times 1} = \tau^2_{global}\boldsymbol{\gamma}$, assume (an estimate of) τ^2_{global} to be given. Denote the prior mean vector and precision matrix in p dimensions by

$$\boldsymbol{\mu}_{p\times 1} = Z\boldsymbol{\mu}_{G\times 1} \in \mathbb{R}^p,\ \Omega_{p\times p} = diag(Z\boldsymbol{\tau}^2_{G\times 1})^{-1} \in \mathbb{R}^{p\times p}, \tag{S.1}$$

with $diag(\boldsymbol{v})$ for a vector $\boldsymbol{v} \in \mathbb{R}^p$ denoting the diagonal matrix with elements $\boldsymbol{v}_k$ on the diagonal. Assume that $\Omega_{p\times p}$ is of full rank.

The penalised log likelihood, denoted by $\ell^\lambda(\boldsymbol{\beta})$ in [3], is, up to a constant c independent of $\boldsymbol{\beta}$, the same as the log of the joint distribution over Y and $\boldsymbol{\beta}$ given the penalty or prior parameters $\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}$: $\pi(Y, \boldsymbol{\beta}|\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1})$:

$$\begin{aligned}
\ell^\lambda(\boldsymbol{\beta}) &= \ell(\boldsymbol{\beta}) - \frac{1}{2}[\boldsymbol{\beta} - \boldsymbol{\mu}_{p\times 1}]^T\Omega_{p\times p}[\boldsymbol{\beta} - \boldsymbol{\mu}_{p\times 1}] + c \\
&= \log \pi(Y|\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}|\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) + \frac{p}{2}\log|2\pi\Omega_{p\times p}| \\
&= \log \pi(Y, \boldsymbol{\beta}|\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) + \frac{p}{2}\log|2\pi\Omega_{p\times p}|.
\end{aligned}$$

A.1.1 | Derivatives of penalised likelihood

Denote the first (partial) derivative of a function to a vector $\boldsymbol{\beta}$ by ∇_β and the second derivative by the Hessian H_β. As given in [3] and extended to including the target or prior mean vector $\boldsymbol{\mu}$, for a GLM with canonical link function, there exists a diagonal weight matrix $W(\boldsymbol{\beta}) = \mathrm{Var}_{Y|\boldsymbol{\beta}}(Y)$, which is usually a function of β, such that the first and second derivative of the penalised likelihood are given by:

$$\begin{aligned}
\frac{\partial \ell^\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} := \nabla_\beta \ell^\lambda(\boldsymbol{\beta}) &= \nabla_\beta \log \pi(Y, \boldsymbol{\beta}|\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) = \nabla_\beta \log \pi(\boldsymbol{\beta}|Y, \boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) \\
&= X^T[\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] - \Omega_{p\times p}[\boldsymbol{\beta} - \boldsymbol{\mu}_{p\times 1}]. \qquad \text{(S.2)} \\
\frac{\partial^2 \ell^\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} := H_\beta \ell^\lambda(\boldsymbol{\beta}) &= H_\beta \log \pi(Y, \boldsymbol{\beta}|\boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) = H_\beta \log \pi(\boldsymbol{\beta}|Y, \boldsymbol{\mu}_{G\times 1}, \boldsymbol{\tau}_{G\times 1}) \\
&= -X^T W(\boldsymbol{\beta})X - \Omega_{p\times p}. \qquad \text{(S.3)}
\end{aligned}$$

A.2 | Moment estimating equations

A.2.1 | Approximate mean and variance of penalised MLE

As used previously for logistic regression [2], one may use a first order Taylor approximation of the score function in $\tilde{\boldsymbol{\beta}}(\boldsymbol{y}, \tau^{overall})$ around $\boldsymbol{\beta}$ to find approximations for the mean and variance of the first smoothened estimate $\tilde{\boldsymbol{\beta}}$ using first estimates $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\tau}}^2, \tilde{\Omega}, \tilde{W} := W(\tilde{\boldsymbol{\beta}})$. Here we repeat some of the details, extended for GLMs with a target.

The first order Taylor approximation is given by

$$\begin{aligned}
\nabla_\beta \log \pi(\boldsymbol{y}, \tilde{\boldsymbol{\beta}}|\tilde{\boldsymbol{\mu}}_{G\times 1}, \tilde{\boldsymbol{\tau}}^2_{G\times 1}) = \nabla_\beta \log \pi(\boldsymbol{y}, \boldsymbol{\beta}|\tilde{\boldsymbol{\mu}}_{G\times 1}, \tilde{\boldsymbol{\tau}}^2_{G\times 1}) \\
+ H_\beta \log \pi(\boldsymbol{y}, \boldsymbol{\beta}|\tilde{\boldsymbol{\mu}}_{G\times 1}, \tilde{\boldsymbol{\tau}}^2_{G\times 1})[\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}] + O(||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2).
\end{aligned} \tag{S.4}$$

As the score function is equal to 0 in the penalised maximum likelihood estimate $\tilde{\boldsymbol{\beta}}$, we find the following first-order approximation for $\tilde{\boldsymbol{\beta}}$:

$$\tilde{\boldsymbol{\beta}} \approx \boldsymbol{\beta} - [H_\beta \log \pi(\boldsymbol{y}, \boldsymbol{\beta}|\tilde{\boldsymbol{\mu}}_{G\times 1}, \tilde{\boldsymbol{\tau}}^2_{G\times 1})]^{-1}\nabla_\beta \log \pi(\boldsymbol{y}, \boldsymbol{\beta}|\tilde{\boldsymbol{\mu}}_{G\times 1}, \tilde{\boldsymbol{\tau}}^2_{G\times 1}). \tag{S.5}$$

For GLMs, this equation can be rewritten as:

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &\approx [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[X^T[\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] - \tilde{\Omega}_{p\times p}[\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_{p\times 1}] + [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}]\boldsymbol{\beta}] \\ &= [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[X^T[\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] + \tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T W(\boldsymbol{\beta}) X\boldsymbol{\beta}].\end{aligned}$$

The mean with respect to the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\beta})$ is then given by:

$$\begin{aligned}E_{\boldsymbol{y}|\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}} &\approx E_{\boldsymbol{y}|\boldsymbol{\beta}}\left[[X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[X^T[\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] + \tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T W(\boldsymbol{\beta}) X\boldsymbol{\beta}]\right] \\ &= [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[X^T[E_{\boldsymbol{y}|\boldsymbol{\beta}}(\boldsymbol{y}) - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] + \tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T W(\boldsymbol{\beta}) X\boldsymbol{\beta}] \\ &= [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[\tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T W(\boldsymbol{\beta}) X\boldsymbol{\beta}] \\ &= \tilde{\boldsymbol{\mu}}_{p\times 1} + [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1} X^T W(\boldsymbol{\beta}) X[\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_{p\times 1}] \\ &\approx \tilde{\boldsymbol{\mu}}_{p\times 1} + [X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1} X^T \tilde{W} X[\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_{p\times 1}],\end{aligned} \tag{S.6}$$

and the variance is given by the diagonal of the covariance matrix:

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{y}|\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}} \approx \mathrm{Cov}_{\boldsymbol{y}|\boldsymbol{\beta}}\Big[&[X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1}[X^T[\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}(y)] \\ &+ \tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T W(\boldsymbol{\beta}) X\boldsymbol{\beta}]\Big] \\ = [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1} X^T &\mathrm{Cov}_{\boldsymbol{y}|\boldsymbol{\beta}}[\boldsymbol{y}]\, X[X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1} \\ = [X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1} X^T &W(\boldsymbol{\beta}) X[X^T W(\boldsymbol{\beta}) X + \tilde{\Omega}_{p\times p}]^{-1} \\ \approx [X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1} X^T &\tilde{W} X[X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1}.\end{aligned} \tag{S.7}$$

Note that we approximate the sample variance matrix W, which is still a function of $\boldsymbol{\beta}$, by $\tilde{W}$. For linear regression this approximation is in fact exact since W does not depend on $\boldsymbol{\beta}$.

A.2.2 | Moment equations for prior mean

The prior mean vector $\boldsymbol{\mu}_{G\times 1}$ can be computed by using the first moment. Denote $P_{G\leftarrow p} \in \mathbb{R}^{G\times p}$ as the matrix that averages the moments over each group, i.e. $[P_{G\leftarrow p}]_{gk} := |\mathcal{G}_g|^{-1}\mathbb{1}_{k\in\mathcal{G}_g}$. The system of moment estimating equations is given by:

$$\begin{cases} \frac{1}{|\mathcal{G}_1|}\sum_{k\in\mathcal{G}_1}\tilde{\beta}_k = \frac{1}{|\mathcal{G}_1|}\sum_{k\in\mathcal{G}_1} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\boldsymbol{\beta}}\left[\tilde{\beta}_k\right]\right], \\ \vdots \\ \frac{1}{|\mathcal{G}_G|}\sum_{k\in\mathcal{G}_G}\tilde{\beta}_k = \frac{1}{|\mathcal{G}_G|}\sum_{k\in\mathcal{G}_g} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\boldsymbol{\beta}}\left[\tilde{\beta}_k\right]\right], \end{cases} \tag{S.8}$$

$$\Leftrightarrow$$

$$P_{G\leftarrow p}\tilde{\boldsymbol{\beta}} = P_{G\leftarrow p} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\boldsymbol{\beta}}\left[\tilde{\boldsymbol{\beta}}\right]\right]. \tag{S.9}$$

Plugging in the mean of Equation (S.6) and further rewriting gives:

$$\begin{aligned}P_{G\leftarrow p}\tilde{\boldsymbol{\beta}} &= P_{G\leftarrow p} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\boldsymbol{\beta}}\left[\tilde{\boldsymbol{\beta}}\right]\right] \\ &\approx P_{G\leftarrow p} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[\tilde{\boldsymbol{\mu}}_{p\times 1} + [X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1} X^T \tilde{W} X[\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}_{p\times 1}]\right] \\ &= P_{G\leftarrow p}\left[\tilde{\boldsymbol{\mu}}_{p\times 1} + [X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1} X^T \tilde{W} X[Z\boldsymbol{\mu}_{G\times 1} - \tilde{\boldsymbol{\mu}}_{p\times 1}]\right].\end{aligned}$$

If we define a matrix C as follows then we can write the above as follows:

$$C := [X^T \tilde{W} X + \tilde{\Omega}_{p\times p}]^{-1} X^T \tilde{W} X, \tag{S.10}$$

$$P_{G\leftarrow p}[\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\mu}}_{p\times 1}] = P_{G\leftarrow p} C Z[\boldsymbol{\mu}_{G\times 1} - \tilde{\boldsymbol{\mu}}_{G\times 1}]. \tag{S.11}$$

So we find the following linear system:

$$A_\mu \boldsymbol{\mu}_{G\times 1} = \boldsymbol{b}_\mu, \tag{S.12}$$

$$A_\mu := P_{G\leftarrow p} C Z, \tag{S.13}$$

$$\boldsymbol{b}_\mu := A_\mu \tilde{\boldsymbol{\mu}}_{G\times 1} + P_{G\leftarrow p}[\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\mu}}_{p\times 1}] = P_{G\leftarrow p}[\tilde{\boldsymbol{\beta}} - [I_{p\times p} - C]\tilde{\boldsymbol{\mu}}_{p\times 1}]. \tag{S.14}$$

Note that $A_\mu \in \mathbb{R}^{G\times G}$ for $G \ll p$. Lastly, we can write each element of A_μ and $\boldsymbol{b}_\mu$ in the format of summing over groups as:

$$[A_\mu]_{g,h} = \frac{1}{|\mathcal{G}_g|}\sum_{k\in\mathcal{G}_g}\sum_{l\in\mathcal{G}_h}\frac{[C]_{k,l}}{|\mathcal{I}_l|}, \tag{S.15}$$

$$[b_\mu]_g = \frac{1}{|\mathcal{G}_g|}\sum_{k\in\mathcal{G}_g}[\tilde{\boldsymbol{\beta}} - [I_{p\times p} - C]\tilde{\boldsymbol{\mu}}_{p\times 1}]_k. \tag{S.16}$$

Remark 1. In high-dimensional data, by default we will shrink to 0, so $\tilde{\boldsymbol{\mu}}_{G\times 1} = \mathbf{0} = \boldsymbol{\mu}_{G\times 1}$.

A.2.3 | Moment equations for prior variance

The prior variance vector $\boldsymbol{\tau}_{G\times 1}$ can be computed by using the second moment equations and the estimate for $\boldsymbol{\mu}_{G\times 1}$. Use the same notation as above to denote $P_{G\leftarrow p} \in \mathbb{R}^{G\times p}$ as the matrix that averages the moments over each group, where $.^2$ denotes element-wise squaring:

$$\begin{cases} \frac{1}{|\mathcal{G}_1|}\sum_{k\in\mathcal{G}_1}\tilde{\beta}_k^2 = \frac{1}{|\mathcal{G}_1|}\sum_{k\in\mathcal{G}_1} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\beta}_k^2\right]\right], \\ \vdots \\ \frac{1}{|\mathcal{G}_G|}\sum_{k\in\mathcal{G}_G}\tilde{\beta}_k^2 = \frac{1}{|\mathcal{G}_G|}\sum_{k\in\mathcal{G}_g} E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\beta}_k^2\right]\right], \end{cases} \tag{S.17}$$

$$\Leftrightarrow$$

$$P_{G\leftarrow p}\tilde{\boldsymbol{\beta}}.^2 = P_{G\leftarrow p}E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}.^2\right]\right]. \tag{S.18}$$

Use $diag(M) := ([M]_{11}, [M]_{22}, .., [M]_{pp})^T$ to denote the diagonal vector of some matrix $M \in \mathbb{R}^{p\times p}$. Then we can derive, plugging in expressions of Equations (S.6) and (S.7):

$$\begin{aligned} P_{G\leftarrow p}\tilde{\boldsymbol{\beta}}.^2 &= P_{G\leftarrow p}E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[\mathrm{Var}_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}\right] + \left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}\right]\right].^2\right] \\ &= P_{G\leftarrow p}\Big\{E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[\mathrm{Var}_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}\right]\right] \\ &\quad +\mathrm{Var}_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}\right]\right] + E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[E_{\boldsymbol{Y}|\beta}\left[\tilde{\boldsymbol{\beta}}\right]\right].^2\Big\} \\ &= P_{G\leftarrow p}\Big\{E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[diag\left([X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}X^T\tilde{W}X[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}\right)\right] \\ &\quad +\mathrm{Var}_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}[\tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T\tilde{W}X\boldsymbol{\beta}]\right] \\ &\quad +E_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}\left[[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}[\tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T\tilde{W}X\boldsymbol{\beta}]\right].^2\Big\} \\ &= P_{G\leftarrow p}\Big\{diag\left([X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}X^T\tilde{W}X[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}\right) \\ &\quad + diag\Big([X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}X^T\tilde{W}X\mathrm{Cov}_{\beta|\boldsymbol{\mu}_{G\times 1},\boldsymbol{\tau}_{G\times 1}}[\boldsymbol{\beta}] \\ &\quad\quad \cdot X^T\tilde{W}X[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}\Big) \\ &\quad + \left[[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}[\tilde{\Omega}_{p\times p}\tilde{\boldsymbol{\mu}}_{p\times 1} + X^T\tilde{W}XZ\boldsymbol{\mu}_{G\times 1}]\right].^2\Big\}. \end{aligned}$$

Again using the matrix C as above, and $\tilde{\boldsymbol{v}}$ as vector for the variance, we can write:

$$C := [X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}X^T\tilde{W}X, \tag{S.19}$$

$$\tilde{\boldsymbol{v}} := diag\left([X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}X^T\tilde{W}X[X^T\tilde{W}X + \tilde{\Omega}_{p\times p}]^{-1}\right), \tag{S.20}$$

$$P_{G\leftarrow p}\tilde{\boldsymbol{\beta}}.^2 = P_{G\leftarrow p}[\tilde{\boldsymbol{v}} + C.^2Z\boldsymbol{\tau}_{G\times 1} + [[I - C]\tilde{\boldsymbol{\mu}}_{p\times 1} + CZ\boldsymbol{\mu}_{G\times 1}].^2], \tag{S.21}$$

and then we find the linear system

$$A_\tau\boldsymbol{\tau}_{G\times 1} = \boldsymbol{b}_\tau, \tag{S.22}$$

$$A_\tau := P_{G\leftarrow p}C.^2Z, \tag{S.23}$$

$$\boldsymbol{b}_\tau := P_{G\leftarrow p}[\tilde{\boldsymbol{\beta}}.^2 - [[I - C]\tilde{\boldsymbol{\mu}}_{p\times 1} + CZ\boldsymbol{\mu}_{G\times 1}].^2 - \tilde{\boldsymbol{v}}]. \tag{S.24}$$

Note that $A_\tau \in \mathbb{R}^{G\times G}$ for $G \ll p$. Again, we can write each element of A_τ and $\boldsymbol{b}_\tau$ in the format of summing over groups as:

$$[A_\tau]_{g,h} = \frac{1}{|\mathcal{G}_g|}\sum_{k\in\mathcal{G}_g}\sum_{l\in\mathcal{G}_h}\frac{[C].^2_{k,l}}{|\mathcal{I}_l|}, \tag{S.25}$$

$$[b_\tau]_g = \frac{1}{|\mathcal{G}_g|}\sum_{k\in\mathcal{G}_g}[\tilde{\boldsymbol{\beta}}.^2 - [[I-C]\tilde{\boldsymbol{\mu}}_{p\times 1} + CZ\boldsymbol{\mu}_{G\times 1}].^2 - \tilde{\boldsymbol{v}}]_k. \tag{S.26}$$

Remark 2. In high-dimensional data, most of the times we will shrink to 0, so $\tilde{\boldsymbol{\mu}}_{G\times 1} = \mathbf{0} = \boldsymbol{\mu}_{G\times 1}$.

A.3 | Moment equations for multiple co-data sets

For multiple co-data sets, each β_k is a priori distributed as:

$$\beta_k \overset{ind.}{\sim} N(\mu_k, \tau_k^2) := N\left(\sum_{d=1}^{D} w^{(d)}\boldsymbol{Z}_k^{(d)}\boldsymbol{\mu}^{(d)}, \tau^2_{global}\sum_{d=1}^{D} w^{(d)}\boldsymbol{Z}_k^{(d)}\boldsymbol{\gamma}^{(d)}\right),\ k = 1,..,p.$$

We can pool all $G_{total} := \sum_{d=1}^{D} G^{(d)}$ groups of all co-data sets together and use the same method of moment equations as above to derive moment estimates for the co-data weights. In what follows, assume that we shrink all β_k to 0, i.e. $\boldsymbol{\mu}^{(d)} = 0$ for all $d = 1,..,D$. A similar argument using the first moments only can be used if non-zero targets are to be used. To be able to use the same notation as above, define:

$$Z = \left[Z^{(1)}\ \cdots\ Z^{(D)}\right], \tag{S.27}$$

$$\boldsymbol{\tau}_{G_{total}\times 1} := \tau^2_{overall}[(w^{(1)}\boldsymbol{\gamma}^{(1)})^T\ \cdots\ (w^{(D)}\boldsymbol{\gamma}^{(D)})^T]^T, \tag{S.28}$$

$$\boldsymbol{\tau}_{p\times 1} = \tau^2_{overall}\sum_{d=1}^{D} w^{(d)}\boldsymbol{Z}^{(d)}\boldsymbol{\gamma}^{(d)} = Z\boldsymbol{\tau}_{G_{total}\times 1}. \tag{S.29}$$

Then we can follow the reasoning similar to above to arrive at the linear system as in Equation (S.22), where we have used that $\tilde{\boldsymbol{\mu}}_{p\times 1} = 0 = \boldsymbol{\mu}_{p\times 1}$:

$$\begin{aligned} A_w\boldsymbol{\tau}_{G_{total}\times 1} &= \boldsymbol{b}_w,\\ A_w &:= P_{G_{total}\leftarrow p}C.^2Z,\\ \boldsymbol{b}_w &:= P_{G_{total}\leftarrow p}[\tilde{\boldsymbol{\beta}}.^2 - \tilde{\boldsymbol{v}}], \end{aligned}$$

but now for $A_w \in \mathbb{R}^{G_{total}\times G_{total}}$ and $\boldsymbol{b}_w \in \mathbb{R}^{G_{total}}$. Plugging in the estimates for $\hat{\tau}^2_{overall}$ and $\hat{\boldsymbol{\gamma}}^{(d)}$, $d = 1,..,D$, we find the linear system for the vector of D unknown co-data weights $\boldsymbol{w} = (w^{(1)},..,w^{(D)})^T$:

$$\tilde{A}_w\boldsymbol{w} = \boldsymbol{b}_w,$$

with $\tilde{A}_w \in \mathbb{R}^{G_{total}\times D}$, and each column $[\tilde{A}]_{*,d}$ given by:

$$[\tilde{A}]_{*,d} = \hat{\tau}^2_{overall}\left[A_w\right]_{*,(1+\sum_{d'=1}^{d-1}G^{(d')}):(\sum_{d'=1}^{d}G^{(d')})}\hat{\boldsymbol{\gamma}}^{(d)}.$$

The group set weights estimate $\hat{\boldsymbol{w}}$ is the ordinary least squares estimate truncated at 0:

$$\hat{\boldsymbol{w}} = (\tilde{\boldsymbol{w}})_+,\ \tilde{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\ ||\tilde{A}\boldsymbol{w} - \boldsymbol{b}_w||_2^2. \tag{S.30}$$

Note that, since $D < G_{total}$, the least squares solution leads to stable solutions. For highly correlated group sets, the group set weights are correlated too, possibly leading to high variance in the group set weight estimates. One should take care in interpreting group set weights of highly correlated group sets.

A.4 | Details for specific examples

The moment equations boil down to a linear system for $\boldsymbol{\mu}$ as given in Equations (S.12) and (S.15) and one for $\boldsymbol{\tau}$ as given in Equations (S.22) and (S.25). These equations use the matrix $C \in \mathbb{R}^{p\times p}$ and vector $v \in \mathbb{R}^p$ as defined in Equation (S.19). To retrieve the moment equations for a specific GLM with link function $g^{-1}(\cdot)$, we only need an expression for the GLM-specific variance matrix $W(\boldsymbol{\beta}) = \mathrm{Var}_{Y|\boldsymbol{\beta}}(Y)$.

Below we give the details for linear, logistic and Cox survival regression.

A.4.1 | Linear regression

For linear regression, the response Y is gaussian distributed around the mean $X\boldsymbol{\beta}$ with variance σ^2 and following link function:

$$y_i \overset{ind.}{\sim} N\left(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\right),\ g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{X}_i\boldsymbol{\beta},\ i = 1, .., n. \tag{S.31}$$

The matrix $\tilde{W} := W(\tilde{\boldsymbol{\beta}})$ is given by:

$$W(\tilde{\boldsymbol{\beta}}) = \sigma^2 I_{n\times n}. \tag{S.32}$$

The approximations for the mean and variance in Equations (S.6) and (S.7) are in fact exact for the linear regression case.

A.4.2 | Logistic regression

For linear regression, the response Y follows a Bernoulli distribution with the vector of probabilities denoted by $\boldsymbol{p} = (p_1, .., p_n)^T$, and with the following link function:

$$y_i \overset{ind.}{\sim} Ber\left(p_i\right),\ g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}) = p_i := \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})},\ i = 1, .., n. \tag{S.33}$$

The matrix $\tilde{W} := W(\tilde{\boldsymbol{\beta}})$ is the diagonal matrix with diagonal elements given by:

$$[W(\tilde{\boldsymbol{\beta}})]_{ii} = \tilde{p}_i(1-\tilde{p}_i) = \frac{\exp(\boldsymbol{X}_i\tilde{\boldsymbol{\beta}})}{(1+\exp(\boldsymbol{X}_i\tilde{\boldsymbol{\beta}}))^2}. \tag{S.34}$$

A.4.3 | Cox survival regression

In Cox survival regression, the outcome $y_i = (t_i, d_i)$ denotes at which time t_i an event occurred, $d_i = 1$, or was censored, $d_i = 0$. Details for Cox survival regression are given in for example[3]. The hazard function $h_i(t)$ is proportional to a baseline hazard $h_0(t)$ with cumulative hazard $H_0(t)$:

$$h_i(t) = h_0(t)\exp(\boldsymbol{X}_i\boldsymbol{\beta}),\ i = 1, .., n,\ H_0(t) = \int_{s=0}^{t} h_0(s)\mathrm{d}s. \tag{S.35}$$

Similarly as in[3], the vector $\boldsymbol{y} - E_{\boldsymbol{y}|\boldsymbol{\beta}}[\boldsymbol{y}]$ in Equation (S.6) is replaced by the vector of martingale residuals:

$$\Delta_i := d_i - H_0(t_i)\exp(\boldsymbol{X}_i\tilde{\boldsymbol{\beta}}),\ i = 1, .., n. \tag{S.36}$$

The W matrix (denoted by D in[3]) is given by the following diagonal matrix:

$$\left[W(\tilde{\boldsymbol{\beta}})\right]_{ii} := H_0(t_i)\exp(\boldsymbol{X}_i\tilde{\boldsymbol{\beta}}),\ i = 1, .., n. \tag{S.37}$$

We use the well-known Breslow estimator to estimate H_0, which is based on the times of observed events, i.e. t_i for which $d_i = 1$:

$$\hat{H}_0(t) = \sum_{i:\ t_i\leq t} \hat{h}_0(t_i),\ \hat{h}_0(t_i) = d_i\left(\sum_{j:\ t_j\geq t_i} \exp(\boldsymbol{X}_j\tilde{\boldsymbol{\beta}})\right)^{-1}. \tag{S.38}$$

A.5 | Hypershrinkage ridge penalty

Consider the prior model for the regression coefficients for one co-data set matrix Z:

$$\beta_k \overset{ind.}{\sim} N\left(0, \tau^2_{global}\boldsymbol{Z}_k\boldsymbol{\gamma}\right). \tag{S.39}$$

The goal is to shrink the group parameter estimates $\boldsymbol{\gamma}$ in such a way that if the co-data is not informative, we shrink towards the ordinary ridge prior as a target prior distribution, i.e. all local variances are set to 1. Furthermore, the variance of the local variance estimates should then be the same for all p covariates and should not depend on the co-data matrix Z. These two assumptions can be expressed as follows:

$$E(\boldsymbol{\tau}^2_{local}) = E(Z\boldsymbol{\gamma}) = \mathbb{1}_{p\times 1},\ \mathrm{Var}(\boldsymbol{\tau}^2_{local}) = \mathrm{Var}(Z\boldsymbol{\gamma}) = \sigma^2_\gamma I_{p\times p}, \tag{S.40}$$

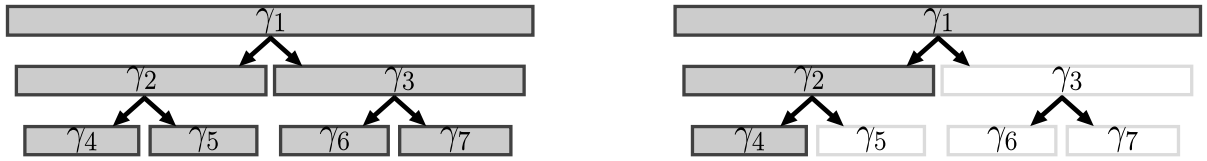

FIGURE S1 Left: discretise continuous scale in increasingly smaller groups by splitting at the median of the continuous co-data in that group. Right: use hierarchical lasso[5,6] to potentially select group weights only if all its parents in the hierarchy (e.g. γ_1 is the parent of γ_2 and γ_3) are selected. White groups are not selected.

for some variance $\sigma_\gamma^2 \geq 0$. Rewriting the expression above gives expressions for the mean and variance of $\boldsymbol{\gamma}$:

$$E(\boldsymbol{\gamma}) = E((Z^TZ)^{-1}Z^TZ\boldsymbol{\gamma}) = (Z^TZ)^{-1}Z^T\mathbb{1}_{p\times 1} := (Z^TZ)^{-1}Z^TZ\mathbb{1}_{G\times 1} = \mathbb{1}_{G\times 1}, \tag{S.41}$$

$$\text{Var}(\boldsymbol{\gamma}) = \text{Var}((Z^TZ)^{-1}Z^TZ\boldsymbol{\gamma}) = \sigma_\gamma^2(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1} = \sigma_\gamma^2(Z^TZ)^{-1}. \tag{S.42}$$

For disjunct groups, this latter expression reduces to

$$\text{Var}(\boldsymbol{\gamma}) = \sigma_\gamma^2\begin{bmatrix} |\mathcal{G}_1| & & \emptyset \\ & \ddots & \\ \emptyset & & |\mathcal{G}_G| \end{bmatrix}^{-1} := \sigma_\gamma^2 W_\gamma^{-1}. \tag{S.43}$$

We rescale $\boldsymbol{\gamma}$ such that all variances are on the same scale:

$$\boldsymbol{\gamma}' = W_\gamma^{1/2}\boldsymbol{\gamma},\ E(\boldsymbol{\gamma}') = W_\gamma^{1/2}\mathbb{1}_{G\times 1},\ \text{Var}(\boldsymbol{\gamma}') = \sigma_\gamma^2 I_{G\times G}. \tag{S.44}$$

We use a ridge penalty for $\boldsymbol{\gamma}'$ corresponding to the normal distribution with mean and variance given above, with hyperpenalty λ_γ inversely proportional to the variance σ_γ^2. Finally, given an estimate $\hat{\lambda}_\gamma$ we solve the optimisation problem given in Equation (8) for the rescaled $\boldsymbol{\gamma}'$ and scale back to obtain the parameter estimates for $\boldsymbol{\gamma}$:

$$W_\gamma^{1/2}\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}' = \underset{\boldsymbol{\gamma}'}{\text{argmin}} \left\{ ||AW_\gamma^{-1/2}\boldsymbol{\gamma}' - \boldsymbol{b}||_2^2 + \hat{\lambda}_\gamma \sum_{g=1}^{G} \left(\gamma_g' - \left[W_\gamma^{1/2} \right]_{gg} \right)^2 \right\}. \tag{S.45}$$

A.6 | Continuous co-data

Figure S1 illustrates the approach to adaptively discretise continuous co-data, described in Section 3.4.1.

A.7 | Covariate selection for prediction

Below we give the technical details needed for implementation of the options for post-hoc variable selection using three different approaches[7,8,9], using an elastic net penalty, DSS criterion and marginal penalised credible intervals respectively.

A.7.1 | Using elastic net

As is widely known, the lasso penalty is known to be able to automatically select variables, but is not stable when covariates are correlated. The elastic net penalty, a combination of the ridge and lasso penalty, can be seen as a stabilised lasso, in the sense that the added ridge penalty stabilises the covariate selection. In a similar manner, the elastic net penalty may be used[7], by rescaling the covariates with the weighted ridge penalty and adding a lasso penalty to perform selection. The procedure can be summarised as follows.

First rescale X and $\boldsymbol{\beta}$ to X' and $\boldsymbol{\beta}'$:

$$\Delta := \begin{bmatrix} \frac{1}{\hat{\tau}_{1,local}^2} & & \emptyset \\ & \ddots & \\ \emptyset & & \frac{1}{\hat{\tau}_{p,local}^2} \end{bmatrix},\ X' := X\Delta^{-\frac{1}{2}},\ \boldsymbol{\beta}' := \Delta^{\frac{1}{2}}\boldsymbol{\beta}. \tag{S.46}$$

Note that $X'\boldsymbol{\beta}' = X\boldsymbol{\beta}$, and $\beta'_k \sim N(0, \hat{\tau}^2_{global})$, $k = 1, .., p$. Then find the penalised maximum likelihood estimate for $\boldsymbol{\beta}'$ such that the desired number of covariates s is selected:

$$\hat{\boldsymbol{\beta}}' = \underset{\boldsymbol{\beta}'}{\operatorname{argmax}} \left\{ \log \pi \left(Y | X', \boldsymbol{\beta}' \right) + \frac{1}{\hat{\tau}^2_{global}} ||\boldsymbol{\beta}'||^2_2 + \lambda_1 ||\boldsymbol{\beta}'||_1 \right\}, \\ \lambda_1 \in \{ \lambda_1 \in \mathbb{R} : \ |\{k : \beta'_k \neq 0\}| = s \}. \tag{S.47}$$

Define $\mathcal{I}_s = \{k \in \{1, .., p\} : \hat{\beta}'_k \neq 0\}$ as the set of indices of selected covariates. Denote by $\boldsymbol{\beta}_s$ the regression coefficients of the selected covariates and $\boldsymbol{\beta}_{-s}$ the remaining regression coefficients. Lastly, refit the selected covariates to obtain the sparsified predictor $\hat{\boldsymbol{\beta}}_{sp.}$ on the right scale.

$$\hat{\boldsymbol{\beta}}_{sp.} = \underset{\boldsymbol{\beta}:\ \boldsymbol{\beta}_{-s}=\mathbf{0}}{\operatorname{argmax}} \left\{ \log \pi \left(Y | X, \boldsymbol{\beta} \right) + \frac{1}{\hat{\tau}^2_{global}} \sum_{k \in \mathcal{I}_s} \frac{1}{\hat{\tau}^2_{k,local}} \beta^2_k \right\}. \tag{S.48}$$

We propose to use either the previous weighted ridge estimates for $\hat{\tau}_{global}$ and $\hat{\boldsymbol{\tau}}_{local}$ to prevent overestimating in dense models, or set the local weights to 1 and refit $\hat{\tau}_{global}$ using maximum marginal likelihood or cross-validation to undo overshrinkage in sparse models.

A.7.2 | Using DSS

Hahn and Carvalho[10] propose to decouple shrinkage and selection (DSS). Decoupling here means that inference is done first using any prior, and selection is done afterwards based on the posterior, resulting in a sequence of sparse linear models. The posterior summary variable selection approach they propose is based on a loss function which balances the prediction error and sparseness of the point estimate of the regression coefficients $\boldsymbol{\beta}$. Given the posterior mean $\hat{\boldsymbol{\beta}}$, they first propose to use the following sparsified point estimate $\hat{\boldsymbol{\beta}}_{sp.}$:

$$\hat{\boldsymbol{\beta}}_{sp.} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \ \lambda ||\boldsymbol{\gamma}||_0 + \frac{1}{n} ||X\hat{\boldsymbol{\beta}} - X\boldsymbol{\gamma}||^2_2. \tag{S.49}$$

As the optimisation problem corresponding to the L_0-penalty is intractable, they propose to approximate the loss function by a local linear approximation with a weighted L_1-penalty:

$$\hat{\boldsymbol{\beta}}_{sp.} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \ \sum_j \frac{\lambda}{|w_j|} |\boldsymbol{\gamma}| + \frac{1}{n} ||X\hat{\boldsymbol{\beta}} - X\boldsymbol{\gamma}||^2_2, \tag{S.50}$$

where they use $w_j = \hat{\beta}_j$. This optimisation problem can be solved with existing software like `glmnet`.

A.7.3 | Using marginal penalised credible regions

Bondell and Reich[9] show that variable selection can be done consistently via penalised credible regions. They prove that their proposed approach using marginal posterior credible sets is consistent in variable selection even when p grows exponentially fast relative to the sample size, useful for high-dimensional data where $p \gg n$.

They propose to use the following set A_n of selected variables based on a thresholding selection rule:

$$A_n = \{ j : \ |\beta_j| > t_{n,j} \}, \tag{S.51}$$

where the threshold $t_{n,j}$ determines the size of A_n, or equivalently, the number of variables that is selected. They propose to use the following threshold:

$$t_{n,j} = s_j t_n, \ s_j = \frac{\sqrt{\operatorname{Var}_{\boldsymbol{\beta}|\boldsymbol{Y}}(\beta_j)}}{\min_j \sqrt{\operatorname{Var}_{\boldsymbol{\beta}|\boldsymbol{Y}}(\beta_j)}}. \tag{S.52}$$

Note that whereas the selection procedure is done marginally, the threshold depends on the full posterior.

We approximate the marginal posterior standard deviation in Equation (S.52) for GLMs penalised with a weighted ridge prior, using a Laplace approximation around the posterior mode $\hat{\boldsymbol{\beta}}$.

Result. Consider a GLM with diagonal weight matrix $W = Var_{\boldsymbol{Y}|\boldsymbol{\beta}}(\boldsymbol{Y})$, that is penalised by a weighted ridge penalty, denoted by the diagonal penalty matrix Δ and corresponding prior variance $\boldsymbol{\tau}^2_{global}$. Define $\tilde{X} = W^{1/2} X \Delta^{-1/2}$ and denote the SVD of $\tilde{X}$

as $\tilde{X} = UDV^T$. The posterior standard deviation of β_j can be approximated by:

$$\sqrt{\text{Var}_{\beta|\boldsymbol{Y},\boldsymbol{\tau}}(\beta_j)} \approx \Delta_{jj}^{-1/2}\sqrt{1 - [VD^2(D^2+I)^{-1}V^T]_{jj}}. \tag{S.53}$$

For the linear regression case, this approximation is in fact an equality.

Derivation. Denote the maximum penalised likelihood estimate, and equivalently the posterior mode, by $\hat{\boldsymbol{\beta}}$. We can approximate the posterior by a Laplace approximation using a Taylor expansion of the log posterior around the mode. The Taylor expansion is given by:

$$\begin{aligned}\log \pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\tau}) &\approx \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\nabla_\beta \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau}) \\ &\quad + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\nabla_\beta^2 \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \\ &= \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau}) + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\nabla_\beta^2 \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),\end{aligned}$$

where the approximation is in fact an equality when linear regression is considered. Taking the exponential on both sides leads to:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\tau}) \dot{\propto} \exp\left(\frac{-1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\left[-\nabla_\beta^2 \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau})\right](\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right),$$

where we use $\dot{\propto}$ to denote "approximately proportional to". So we can approximate the posterior with the following multivariate gaussian:

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\tau} \dot{\sim} N\left(\hat{\boldsymbol{\beta}}, \left[-\nabla_\beta^2 \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau})\right]^{-1}\right),$$

where we use $\dot{\sim}$ to denote "approximately distributed as". The posterior covariance matrix for a GLM is approximated by:

$$\text{Cov}_{\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\tau}}(\boldsymbol{\beta}) \approx \left[-\nabla_\beta^2 \log \pi(\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{\tau})\right]^{-1} = \left[X^T W(\hat{\boldsymbol{\beta}})X + \Delta\right]^{-1}.$$

which in turn we can write as, using Woodbury's matrix inversion identity, substituting $\tilde{X} = W^{-1/2}X\Delta^{-1/2}$ and the SVD of $\tilde{X}$:

$$\begin{aligned}\left[X^TWX+\Delta\right]^{-1} &= \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}X^TW^{1/2}\left(I_{n\times n} + W^{1/2}X\boldsymbol{\Delta}^{-1}X^TW^{1/2}\right)^{-1}W^{1/2}X\boldsymbol{\Delta}^{-1} \\ &= \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1/2}\tilde{X}^T\left(I_{n\times n} + \tilde{X}\tilde{X}^T\right)^{-1}\tilde{X}\boldsymbol{\Delta}^{-1/2} \\ &= \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1/2}VDU^T\left(I_{n\times n} + UDV^TVDU^T\right)^{-1}UDV^T\boldsymbol{\Delta}^{-1/2} \\ &= \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1/2}VD^2\left(I_{n\times n} + D^2\right)^{-1}V^T\boldsymbol{\Delta}^{-1/2}.\end{aligned}$$

The marginal posterior standard deviations are given by the square root of the diagonal elements:

$$\sqrt{\text{Var}_{\beta|\boldsymbol{Y},\boldsymbol{\tau}}(\beta_j)} \approx \Delta_{jj}^{-1/2}\sqrt{1 - [VD^2(D^2+I)^{-1}V^T]_{jj}}.$$

A.8 | Unpenalised covariates

We may group covariates that we do not want to penalise (e.g. an intercept) in a group, say group $\mathcal{G}_0$. Not penalising corresponds to a Bayesian prior with mean $\mu_0^\beta = 0$ and $\tau_0^2 = \infty$, and penalty 0. Furthermore, for the matrix C as defined in Equation (S.19), $[C]_{kl} = 0$ for every $l \in \mathcal{G}_0$, $k \neq l$:

Lemma 1. Let $l \in \mathcal{G}_0$ be an unpenalised covariate without correlation with other covariates. Then, for $k \neq l$:

$$[C]_{kl} = \left[(X^T\tilde{W}X + \tilde{\Omega})^{-1}X^T\tilde{W}X\right]_{kl} = 0, \tag{S.54}$$

and therefore also $[C]_{lk} = [C]_{kl} = 0$.

Proof. First, note that the matrix C is equal to:

$$\begin{aligned}C &= (X^T\tilde{W}X + \tilde{\Omega})^{-1}X^T\tilde{W}X = (X^T\tilde{W}X + \tilde{\Omega})^{-1}(X^T\tilde{W}X + \tilde{\Omega} - \tilde{\Omega}) \\ &= I - (X^T\tilde{W}X + \tilde{\Omega})^{-1}\tilde{\Omega}.\end{aligned}$$

So, for $k \neq l$:

$$\begin{aligned}[C]_{kl} &= -\left[(X^T\tilde{W}X + \tilde{\Omega})^{-1}\tilde{\Omega}\right]_{kl} \\ &= -\sum_{i=1}^{p}\left[(X^T\tilde{W}X + \tilde{\Omega})^{-1}\right]_{ki}\left[\tilde{\Omega}\right]_{il} \\ &= 0,\end{aligned}$$

where the latter equation holds since the l^{th} column of the precision matrix corresponding to an unpenalised variable contains only 0. Note that $\boldsymbol{C}$ is symmetric since it is a product of sums of symmetrix matrices. Therefore we can conclude that $[C]_{lk} = [C]_{kl} = 0$. □

As a result from this lemma and Equations (S.15),(S.25), we see that:

$$[A_\mu]_{g0} = [A_\mu]_{0g} = 0,\ [A_\tau]_{g0} = [A_\tau]_{0g} = 0,\ \forall g = 1, .., G. \tag{S.55}$$

Therefore we can compute the moment estimates using the block matrix of A_μ and A_τ corresponding to the penalised groups only. So, after we have computed $\boldsymbol{C}$ using both penalised and unpenalised covariates, we only need the rows and columns of $\boldsymbol{C}$ corresponding to penalised covariates to obtain the moment estimates.

A.9 | Interpretation hyperparameter estimates

The prior variance parameters model the scale of covariates in a group. The empirical Bayes estimates allow for interpretation on covariate level, group level, group set level and global level.

First consider the covariate-specific prior variance of β_k. It may be written as a sum over groups and group sets, such that the prior is given by:

$$\beta_k|\tau^2_{global}, \tau^2_{local,k} \sim N\left(0, \sum_{d=1}^{D}\sum_{g\in\mathcal{G}^{(d)}} \tau^2_{global} w^{(d)} Z^{(d)}_{kg}\gamma^{(d)}_g\right).$$

The a priori expected magnitude of β_k is a function of the prior standard deviation, and may be written as sum over groups and group sets as well:

$$\begin{aligned}E_{\beta_k|\tau^2_{global},\tau^2_{local,k}}(|\beta_k|) &= \sqrt{\frac{2}{\pi}\text{Var}_{\beta_k|\tau^2_{global},\tau^2_{local,k}}(\beta_k)} = \sqrt{\frac{2}{\pi}}\frac{\text{Var}_{\beta_k|\tau^2_{global},\tau^2_{local,k}}(\beta_k)}{\sqrt{\text{Var}_{\beta_k|\tau^2_{global},\tau^2_{local,k}}(\beta_k)}} \\ &= \sum_{d=1}^{D}\sum_{g\in\mathcal{G}^{(d)}}\left[\sqrt{\frac{2}{\pi}}\frac{\tau^2_{global} w^{(d)} Z^{(d)}_{kg}\gamma^{(d)}_g}{\sqrt{\sum_e\sum_h \tau^2_{global} w^{(e)} Z^{(e)}_{kh}\gamma^{(e)}_h}}\right].\end{aligned} \tag{S.56}$$

The last expression may be used to visualise which groups and group sets contribute most to the prior variance of a specific covariate.

Then, recall that the global, group and group set hyperparameters are estimated in a hierarchical fashion. This allows for the following interpretations, using the expression above:

1. τ_{global} quantifies how much signal the data contain overall in terms of average effect size;

   the overall level of regularisation, τ^2_{global}, is first estimated, ignoring the groups and group sets. So, each β_k is then normally distributed with the same prior variance, and the scale parameter estimate says something about the expected or average magnitude of β_k over all p covariates:

   $$\tau_{global} = \sqrt{\frac{\pi}{2}}E_{\beta_k|\tau^2_{global}}(|\beta_k|) \approx \sqrt{\frac{\pi}{2}}\frac{1}{p}\sum_{k=1}^{p}|\beta_k|. \tag{S.57}$$

   So τ_{global} quantifies how much signal the data contain in terms of average effect size.

2. $\sqrt{\gamma^{(d)}_g}$ quantifies relatively how much signal may be attributed to group g of group set d in terms of expected or average effect size in that group, relative to the global level or average;

the group weights $\boldsymbol{\gamma}^{(d)}$ are estimated given τ^2_{global} for each co-data source separately. For non-overlapping groups, we can write:

$$\sqrt{\gamma_g^{(d)}} = \frac{\sqrt{\frac{\pi}{2}} E_{\beta_k|\tau^2_{global},\gamma_g^{(d)}}\left(|\beta_k|\right)}{\tau_{global}} = \frac{E_{\beta_k|\tau^2_{global},\gamma_g^{(d)}}\left(|\beta_k|\right)}{E_{\beta_k|\tau^2_{global}}(|\beta_k|)} \approx \frac{\frac{1}{G}\sum_{k\in\mathcal{G}_g^{(d)}}|\beta_k|}{\frac{1}{p}\sum_{k=1}^{p}|\beta_k|}. \tag{S.58}$$

The interpretation is the same for overlapping groups, for which one may derive a similar expression in which effect sizes are scaled to correct for the fact that different covariates may belong to differently many groups.

3. $\sqrt{w^{(d)}}$ quantifies relatively how much signal may be attributed to group set d in terms of contribution to the expected or average effect size, relative to the global level or average;

   the co-data weights are estimated given the estimated group weights and global level of regularisation. The interpretation is the same as that for groups of covariates, but then on group set level.

B | SIMULATION STUDY

This simulation study illustrates the benefits of using hypershrinkage, i.e. an extra level of shrinkage on the group weights. First, we demonstrate that when the co-data is not informative, the group weights and therefore local variances are shrunk to 1, retrieving prediction and group prior variance estimation errors similar to ordinary ridge. When the co-data is informative, the group weight estimates are shrunk little, improving the predictions and estimations compared to ordinary ridge. When both informative and random co-data sets are combined, `ecpc` correctly places relatively more weight on the informative co-data. Second, we compare performance to a full Bayes model with vague hyperprior. We demonstrate that the full Bayes method does not enjoy the benefits of hypershrinkage; it learns from informative co-data, but overfits when co-data is not informative. Lastly, we illustrate how the combination of hierarchical lasso and ridge hypershrinkage may be used for hierarchical, overlapping groups, in which some form of hypershrinkage is necessary to obtain a unique solution in the linear system of moment equations given in Equation (8).

B.1 | Simulation set-up

We consider linear regression for some fixed vector of regression coefficients $\boldsymbol{\beta}^0$. We simulate 100 pairs of training and test sets with the number of samples $n = 100$ and the number of covariates $p = 300$. We simulate for each pair of training and test sets, for variance parameters $\sigma^2 = 1, \tau^2 = 0.1$:

$$\begin{aligned} &\boldsymbol{\beta}^0 \sim N\left(0, \tau^2 I_{p\times p}\right),\ \left[X_{train}\right]_{ij}, \left[X_{test}\right]_{ij} \overset{i.i.d.}{\sim} N(0,1),\ i = 1,..,n,\ j = 1,..,p, \\ &\boldsymbol{Y}_{train} \sim N\left(X_{train}\boldsymbol{\beta}^0, \sigma^2 I_{n\times n}\right),\ \boldsymbol{Y}_{test} \sim N\left(X_{test}\boldsymbol{\beta}^0, \sigma^2 I_{n\times n}\right). \end{aligned} \tag{S.59}$$

Consider the following non-informative and informative co-data:

1. `Random`: randomly assign the 300 covariates to G approximately equally sized groups, with G in the range of $1 - 30$.
2. `Informative`: assign the covariates to G approximately equally sized groups based on the ranking of the size of each regression coefficient, $|\beta_k^0|$, $k = 1,..,p$. So there exists an ordering of the groups such that for each pair of two groups $\mathcal{G}_i, \mathcal{G}_j$, $1 \leq i < j \leq G$, and for all $k \in \mathcal{G}_i$, $l \in \mathcal{G}_j$: $|\beta_k^0| < |\beta_l^0|$.

B.2 | Benefits of hypershrinkage

We use the default ridge penalty as hypershrinkage for the group weights with 1 as target, such that the global-local prior variances $\tau^2_{global}\tau^2_{k,local}$ are shrunk to the global prior variance, corresponding to an ordinary ridge prior on the covariate level. We train the following models on the training data for both types of co-data and an increasing number of groups G: 1) `ecpc` with hypershrinkage; 2) `ecpc` without hypershrinkage, i.e. optimise the objective in Equation (8) without any added penalty function; 3) `GRridge`[1], which uses a regularisation on the group level based on permutations of the covariates' group indices, and 4) `ordinary ridge`, a ridge model that uses one overall penalty.

First, we compare the estimated prior group variances in the co-data setting with $G = 5$ groups. The estimates are compared to the 'true' prior group variance in each simulated data set, the values that maximise the prior distribution given the true,

simulated regression coefficients $\boldsymbol{\beta}^0$. Figure S2 shows the estimated prior group variances and the mean squared error (MSE) of those estimates per group. The estimates of `ecpc` lie around the line $y = x$, indicating that the estimates are on average approximately equal to the prior maximiser. The estimates of `ordinary ridge` do not depend on the groups. The estimates of `GRridge` show high variance for the groups with largest effect sizes, possibly because of the ad-hoc regularisation on the group level using permutations. The MSE of the prior variances estimated with `ecpc` is larger for groups with larger effect sizes.

Figure S3 shows the mean squared error (MSE) of the group prior variance estimates averaged over g groups and the MSE of the predictions on the test data as performance measure, for g ranging from $1 - 30$. When the co-data is non-informative, `ecpc` with hypershrinkage performs similarly to `ordinary ridge`, as the group weights of the random groups are shrunk towards 1. Besides, `ecpc` with hypershrinkage outperforms both `ecpc` without hypershrinkage, as it is not able to shrink the group weights, and `GRridge`, which uses the more ad-hoc type of regularisation described above. When the co-data is informative, `ecpc` with hypershrinkage shrinks little, performing similarly to `ecpc` without hypershrinkage, and outperforming `GRridge`. Moreover, all three methods outperform `ordinary ridge` as they benefit from the co-data.

Lastly, we combine the random and informative co-data and train `ecpc` with and without hypershrinkage on the two co-data sources. We expect that the random co-data set obtains a group set weight of 0, while the informative co-data set obtains group set weight of 1. Figures S4a and S4b show that the group set weights estimated with `ecpc` with hypershrinkage are better than those estimated with `ecpc` without hypershrinkage in the setting with five random and five informative co-data groups. Figure S4c shows similar results for the average MSE of group set weights for various number of groups. Figure S4d shows that both methods perform similarly in terms of prediction, as both place relatively more weight on the informative co-data. The predictive performance of `ecpc` with hypershrinkage is slightly lower, possibly because of the truncation of the group set weight at 0.

B.3 | Comparison hypershrinkage to full Bayes

Consider the same simulation set-up as above. We compare performance of `ecpc` with the full Bayes method `graper`[11], trained and tested on the same data sets. The method `graper` imposes a vague hyperprior on the prior parameters and uses a variational Bayes approach to approximate the posterior for $\boldsymbol{\beta}$ as multivariate distribution (`graper (dense, multiv.)`) or factorised over all covariates (`graper (dense)`). The posterior mean of $\boldsymbol{\beta}$ is then used for prediction. Figure S5 illustrates how performance of `graper` and `ecpc` is affected by the number of groups and (non-)informativeness of co-data. Whereas `ecpc` with hypershrinkage is able to adapt the degree of hypershrinkage, `graper` is not able to adapt the fixed, vague hyperprior. Consequently, it shows similar behaviour as `ecpc` without hypershrinkage; it performs better than ordinary ridge when co-data are informative, but overfits when co-data are random.

B.4 | Hypershrinkage for hierarchical, overlapping groups

Consider the same simulation set-up as above, but now with hierarchical, overlapping groups as co-data. We obtain these groups with the approach described in Section 3.4.1, using either a random ordering of the regression coefficients or the true order of the regression coefficients as continuous co-data. So, the group with all covariates is split into two groups, which in turn are both split into two groups, and so on. We stop this recursive splitting at eight groups, called the leaf groups. The average effect size of covariates in the leaf groups is approximately the same for the random hierarchical group set, and it changes gradually for the informative hierarchical group set, as illustrated in Figure S6. We fit `ecpc` with the proposed combination of hierarchical lasso and ridge hypershrinkage to select and estimate the hierarchical group prior variance weights. Note that it is essential to use some form of regularisation on the overlapping, hierarchical groups, as the linear system in Equation (8) is singular. Therefore, it is not possible to fit `ecpc` without hypershrinkage on the hierarchical, overlapping groups. Instead, we compare the performance in estimation of the group weights and prediction of the response in the test data with `ordinary ridge` and `ecpc` fit without hypershrinkage on the leaf groups only.

Table S1 shows how many times the random and informative hierarchical groups are selected by `ecpc` in the 100 simulated training data sets. The group with all covariates is always selected, as it is on top of the hierarchy. Smaller groups lower in the hierarchy are selected fewer times. In the random hierarchical group set, the selection percentages are similar across leaf groups, reflecting similarity of the average effect size in the leaf groups. In the informative group set, the first and last group are selected most of the times in each hierarchical layer, while the groups in between are rarely selected. This is expected, as the first and last group in the informative group set have a larger average effect size (Figure S6).

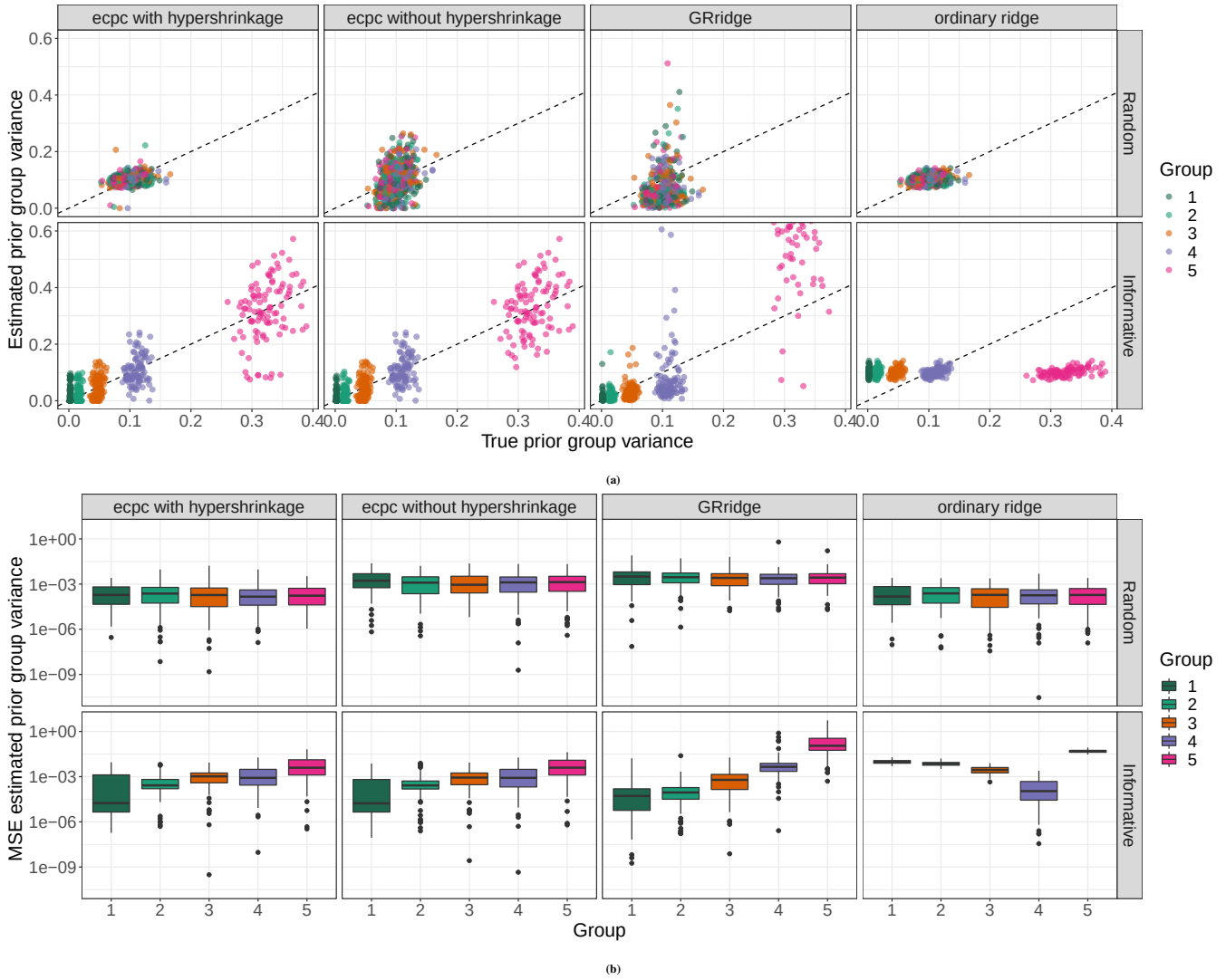

FIGURE S2 Simulation study based on 100 training and test sets and random or informative co-data of five groups. a) The estimated prior group variance versus the 'true' prior group variance for several methods. The 'true' prior group variances are given by the prior group variances that maximise the prior distribution given the simulated regression coefficients. Note that some of the large point estimates of `GRridge` are cut off in the informative co-data setting. The dotted line indicates the line $y = x$; b) MSE of the group variance estimates per group for several methods.

Figure S7 shows the estimated prior group variances in the leaf groups and MSE. The 'true' prior group variance is again given by the values that maximise the prior distribution given the true, simulated regression coefficients $\boldsymbol{\beta}^0$. The estimates of `ecpc` resulting from the hierarchical group set tend to be slightly larger than those resulting from the leaf groups only. This difference is larger for the groups that have small average effect in the informative hierarchical group set, as the group with all covariates is always selected (Table S1).

Figure S8 shows the MSE of the prior group variance estimates averaged over the leaf groups and the MSE of the predictions on the test data. In terms of estimation performance, `ecpc` estimated on the hierarchical groups is outperformed by `ordinary ridge` in the random hierarchical co-data and slightly outperformed by `ecpc` without hypershrinkage on the leaf groups only in the informative hierarchical co-data. This may be expected as `ecpc` uses more group parameters to model the prior variance in the leaf groups in the hierarchical group set than in the group set with leaf groups only. In terms of prediction performance, however, `ecpc` with the hierarchical groups is competitive to `ordinary ridge` in the random hierarchical group set and slightly outperforms `ecpc` on the leaf groups only in the informative hierarchical co-data.

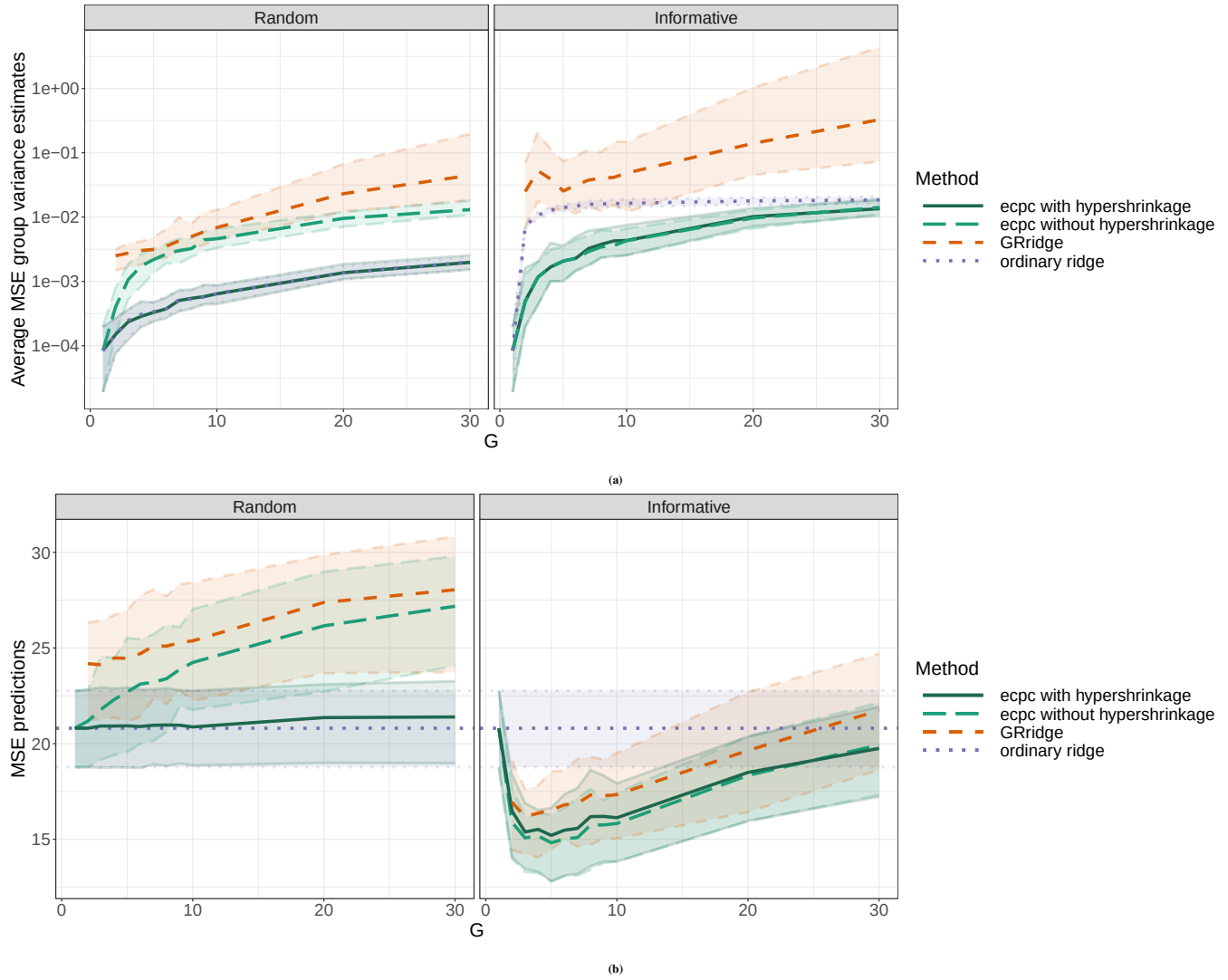

FIGURE S3 Simulation study based on 100 training and test sets and random co-data (left) or informative co-data (right) for various number of groups and several methods. The lines indicate the mean and the shaded bands indicate the 25%, 75% quantiles. a) MSE of the group prior variance estimates averaged over the groups; b) MSE of the predictions on the test sets.

C | DATA APPLICATIONS

C.1 | Predicting therapy response in colorectal cancer

The results of the first data application using miRNA expression are discussed in Section 5.1. Here we provide mentioned additional figures.

Interpretation of estimated hyperparameters. The group weights of the other group sets are shown in Figure S9. Figure S10 displays composition of the prior for several miRNAs in terms of group sets and groups, and its impact on regression parameter estimates.

Performance. Figure S11 shows the performance of `ecpc` in the dense setting and covariate sparse setting when abundance and standard deviation are discretised in 5, 10 or 20 groups. The performances are comparable, with the model based on 20 groups in abundance and standard deviation performing slightly better in the dense setting, and slightly worse in the sparse setting. The performance of the group sparse models is shown in Figure S12. Here, `ecpc` is combined with a lasso penalty on the group level on all groups of the five group sets to obtain a group sparse model. Group lasso uses a latent overlapping group (LOG) penalty[6,5] on all groups of the first three co-data sources and the leaf groups in the tree of the `lfdr1` and `lfdr2`

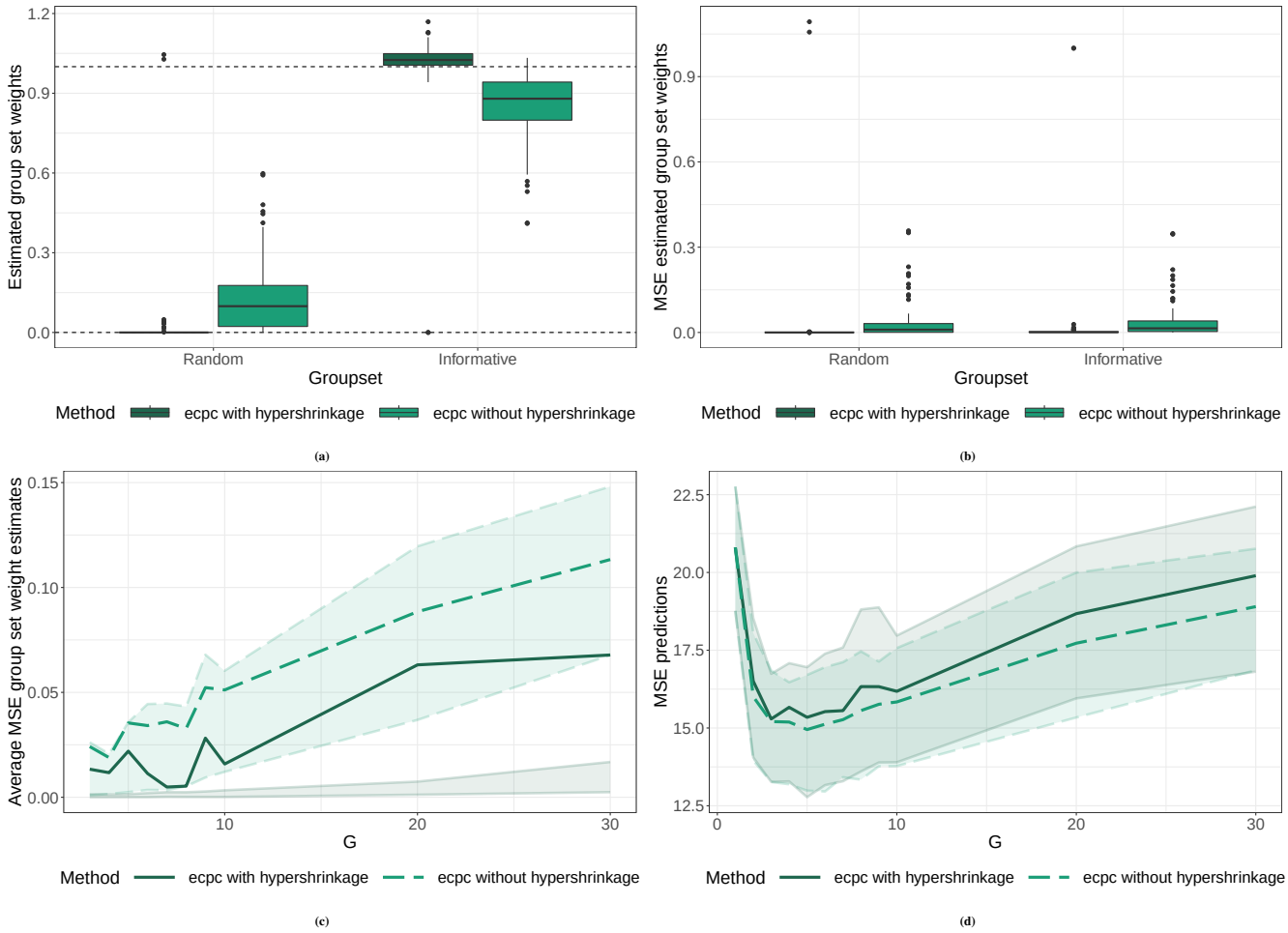

FIGURE S4 Simulation study based on 100 training and test sets and both random and informative co-data of five groups (top) or of a varying number of groups (bottom). a) The estimated prior group set weights. The horizontal lines represent the 'true' group set weights at 1 for the informative co-data set and at 0 for the random co-data set; b) MSE of the group set weight estimates; c) MSE of the group set weight estimates averaged over the two group sets. The lines in Figures c) and d) indicate the mean and the shaded bands indicate the 25%, 75% quantiles; d) MSE of the predictions on the test sets.

group sets, without distinguishing between co-data sources. Hierarchical lasso uses a LOG penalty on all groups of all co-data sources. For the lfdr group sets, the implied hierarchical constraints are that covariates in an lfdr group can be included only when all covariates in the groups with lower lfdrs are included as well[5]. `ecpc` adequately learns from co-data and outperforms group lasso and hierarchical lasso. Then, Figure S13 shows the AUC performance of various post-hoc selection methods on the cross-validation folds. Figure S14 shows the predicted values for several models with 25 selected markers and corresponding ROC curves. A predictive model is more likely to be accepted for practical use when it is limited to a selection of few covariates and performs well, both in terms of ranking (as assessed by ROC/AUC) and in terms of good separation of the two groups in absolute prediction scores. The parsimonious predictor with 25 covariates selected by `ecpc` obtains overall highest performance in terms of AUC (Figure 5d in the main article) and separates the two groups much better than the other methods (Figure S14).

Covariate selection. Figure S15 shows the absolute values of the estimated regression coefficients for `ecpc` and `ordinary ridge`. The density plot is more heavy-tailed for `ecpc`, which facilitates posterior selection. We fit `ecpc` and `elastic net` for $\alpha = 0.3$ and $\alpha = 0.8$ on subsamples of size $\approx \frac{2}{3}n$, stratified for response, to assess stability of covariate selection. We use leave-one-out cross validation to estimate the global prior variance in `ecpc`, use the default post-hoc selection procedure to select 25 covariates for each subsample and count the number of overlapping miRNAs in each pairwise comparison of selected sets. For `elastic net`, we keep the value of α fixed and tune λ to select 25 covariates. We repeat the analysis for a selection

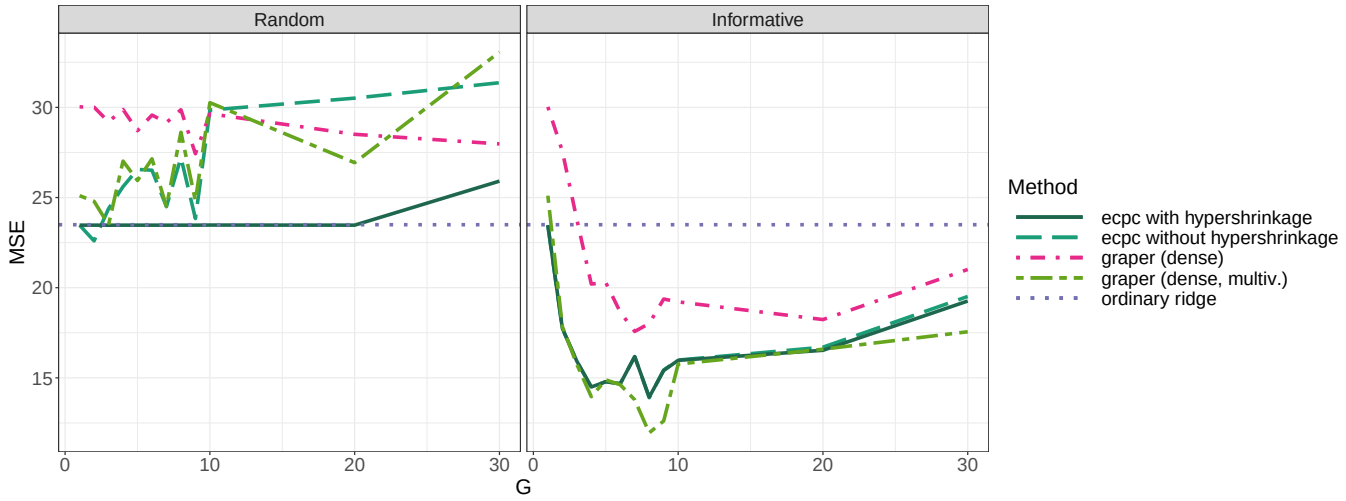

FIGURE S5 MSE of the predictions on a test set for various number of groups, for various methods and for random co-data (left) or informative co-data (right). The full Bayes method `graper` imposes a vague hyperprior on the hyperparameters and overfits on random co-data.

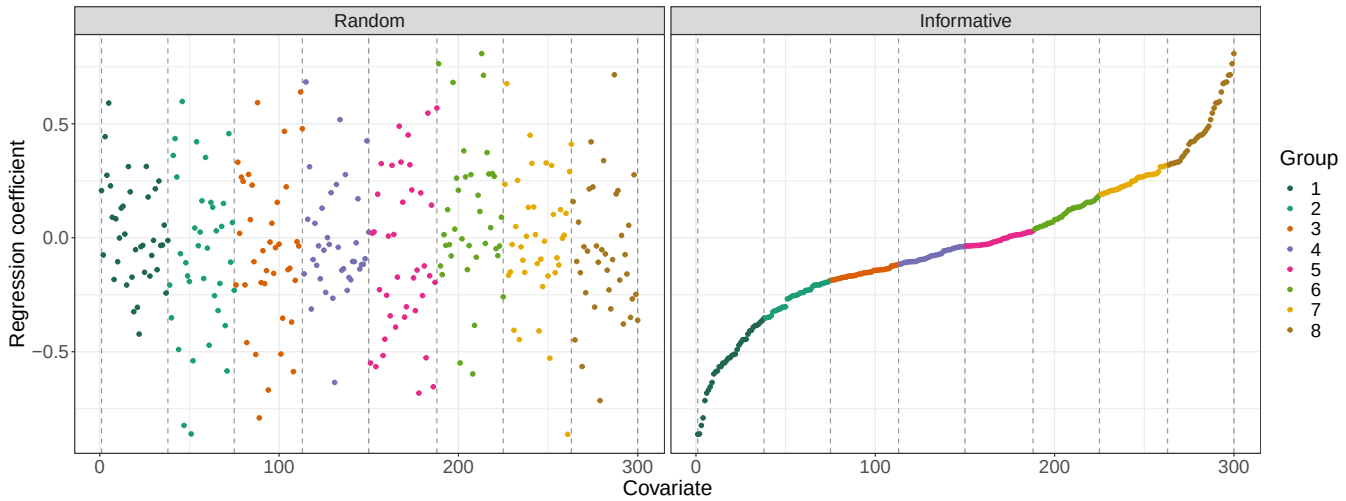

FIGURE S6 Illustration of the leaf groups used in the hierarchical group set in one of the simulated data sets. Either a random order or the order of the regression coefficients is used to make hierarchical groups with the procedure described in Section 3.4.1. The leaf groups are numbered from $\mathcal{G}_1$ to $\mathcal{G}_8$, resulting in the group set with overlapping groups: $\left\{\bigcup_{i=1}^{8}\mathcal{G}_i,\ \bigcup_{i=1}^{4}\mathcal{G}_i,\ \bigcup_{i=5}^{8}\mathcal{G}_i,\ \mathcal{G}_1\cup\mathcal{G}_2\ \mathcal{G}_3\cup\mathcal{G}_4,\ \mathcal{G}_5\cup\mathcal{G}_6,\ \mathcal{G}_7\cup\mathcal{G}_8,\ \mathcal{G}_1,..,\mathcal{G}_8\right\}$.

of 50 covariates. Figure S16 shows histograms of the overlap between selections of covariates, with the corresponding AUC performance given in Figure S17.

Validation. To assess the broader use of the four sets of 25 markers, selected by either `ecpc`, `GRridge`, or `elastic net` ($\alpha \in \{0.3, 0.8\}$), we study their association with overall survival (OS) as related outcome. First, on the same samples, then on an independent validation set. For the first, we dichotomize the 88 samples into a low and high risk group based on the clinical benefit prediction: the (cross-validated) linear predictors from the models with 25 markers were ranked, and the median was used to distinguish the two groups. The survival curves of the co-data learnt `ecpc` and `GRridge` clearly separate the two groups better than those of `elastic net` (Figure S18). Note that formal statistical testing is hampered here due to the cross-validation (and hence dependent) nature of the linear predictors. Second, we tested each of the four sets of 25 markers on a large independent set: The Cancer Genome Atlas (TCGA) colonadenoma (COAD) set, with miRNA and matched OS data for

Random								Informative							
$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$	$\mathcal{G}_7$	$\mathcal{G}_8$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$	$\mathcal{G}_7$	$\mathcal{G}_8$
100								100							
90				94				95				99			
61		51		59		64		95		17		18		97	
31	34	20	32	36	32	39	32	92	8	6	14	12	6	5	96

TABLE S1 Number of times each (union of) leaf group(s) is selected in the 100 simulated training data sets in the random or informative hierarchical co-data illustrated in Figure S6, when `ecpc` is combined with hierarchical lasso and ridge hypershrinkage.

420 individuals. The preprocessing of the TCGA [12] COAD data is described in [13]. For the four sets of 25 miRNAs, 17-21 were available in the TCGA set as well. The large sample size allowed straightforward likelihood ratio testing of each of the four sets with matched miRNAs in a Cox model. Tables S2 and S3 show the p-values of the multivariate Cox model fit on the TCGA COAD data. Here, the set selected by `ecpc` was the only significant one: $p = 0.01$, whereas for others $p > 0.30$, hence strongly non-significant (Table S2). Also, the Wald-test based p-values per marker in the multivariate Cox model (Table S3) support the better validation of the `ecpc` markers. Comparing the prior variance composition of `ecpc` and `GRridge` of the top five markers (Figure S19) suggests that flexibly handling multiple co-data enabled `ecpc` to focus on the `lfdr2` group set and small-valued lfdr groups, thereby selecting the top five markers, while `GRridge` selected only one.

Method (p-value overall likelihood ratio test)

ecpc (0.010)			**Grridge (0.888)**			**elastic net $\alpha = 0.3$ (0.317)**			**elastic net $\alpha = 0.8$ (0.562)**		
miRNA	**coefficient**	**p-value**	**miRNA**	**coefficient**	**p-value**	**miRNA**	**coefficient**	**p-value**	**miRNA**	**coefficient**	**p-value**
hsa.mir.31	2.73E-02	1.89E-03	hsa.mir.335	1.77E-02	3.91E-02	hsa.mir.2467	6.64E-01	8.26E-03	hsa.mir.2467	6.77E-01	6.48E-03
hsa.mir.2467	6.52E-01	9.06E-03	hsa.mir.552	-1.06E-02	6.46E-02	hsa.mir.31	2.20E-02	9.70E-03	hsa.mir.3617	2.31E-01	5.96E-02
hsa.mir.338	7.19E-03	9.70E-03	hsa.mir.29c	4.43E-03	2.17E-01	hsa.mir.3617	2.04E-01	1.02E-01	hsa.mir.146b	4.65E-03	2.81E-01
hsa.mir.503	4.40E-02	1.16E-02	hsa.mir.7974	-4.94E-02	3.38E-01	hsa.mir.4750	-9.00E-01	2.09E-01	hsa.mir.6873	-4.90E-01	2.86E-01
hsa.mir.335	1.87E-02	2.55E-02	hsa.mir.95	1.84E-02	3.47E-01	hsa.mir.146b	4.50E-03	3.00E-01	hsa.mir.892a	2.67E-01	3.20E-01
hsa.mir.3145	2.08E-01	5.19E-02	hsa.mir.30e	-2.14E-03	3.48E-01	hsa.mir.592	8.24E-03	3.46E-01	hsa.mir.3929	1.83E-01	5.40E-01
hsa.mir.17	-5.52E-03	1.73E-01	hsa.mir.181d	1.28E-02	4.74E-01	hsa.mir.380	5.16E-02	4.37E-01	hsa.mir.380	3.76E-02	5.67E-01
hsa.mir.181d	2.16E-02	2.14E-01	hsa.mir.135b	-4.81E-03	4.80E-01	hsa.mir.3200	2.03E-02	5.57E-01	hsa.mir.3622a	-1.20E-01	5.79E-01
hsa.mir.892a	3.03E-01	2.59E-01	hsa.mir.421	-2.97E-02	5.17E-01	hsa.mir.4659b	2.04E-01	6.96E-01	hsa.mir.6780a	1.94E-01	6.38E-01
hsa.mir.30e	-2.55E-03	2.60E-01	hsa.mir.224	6.74E-03	5.47E-01	hsa.mir.3622a	-7.86E-02	7.21E-01	hsa.mir.4659b	2.23E-01	6.74E-01
hsa.mir.552	-6.25E-03	2.63E-01	hsa.mir.548ar	-1.54E-01	5.51E-01	hsa.mir.6780a	-1.36E-01	7.54E-01	hsa.mir.6761	1.84E-02	7.44E-01
hsa.mir.135b	-7.49E-03	2.67E-01	hsa.mir.195	-1.23E-02	5.60E-01	hsa.mir.3929	9.96E-02	7.61E-01	hsa.mir.98	-5.16E-03	7.74E-01
hsa.mir.7974	-5.08E-02	3.20E-01	hsa.mir.17	-2.23E-03	6.28E-01	hsa.mir.4467	-2.99E-01	7.70E-01	hsa.mir.4467	-2.78E-01	7.85E-01
hsa.mir.29c	3.52E-03	3.33E-01	hsa.mir.3200	1.82E-02	6.35E-01	hsa.mir.6761	2.89E-03	9.59E-01	hsa.mir.592	-7.78E-04	9.26E-01
hsa.mir.183	-1.07E-03	3.45E-01	hsa.mir.183	-5.02E-04	6.48E-01	hsa.mir.98	5.11E-04	9.79E-01	hsa.mir.548g	-1.69E+01	9.96E-01
hsa.mir.431	-2.53E-02	3.89E-01	hsa.mir.4454	2.03E-02	7.39E-01	hsa.mir.548g	-1.55E+01	9.96E-01	hsa.mir.548ap	NA	1.00E+00
hsa.mir.548ar	-2.29E-01	3.98E-01	hsa.mir.18a	7.35E-04	9.73E-01	hsa.mir.6801	6.51E-04	9.98E-01			
hsa.mir.3200	1.05E-02	7.65E-01	hsa.mir.431	-6.57E-04	9.84E-01	hsa.mir.548ap	NA	1.00E+00			
hsa.mir.195	5.20E-03	8.11E-01	hsa.mir.549a	-4.20E-04	9.94E-01						

TABLE S2 A multivariate Cox survival model is fit on the TCGA COAD data on the 25 markers selected by several methods in the miRNA data. For each method, the table shows the p-value of the overall likelihood ratio test, estimated regression coefficients, corresponding p-values and names of the selected miRNAs, ordered in increasing p-value. Results are shown only for miRNAs that could be matched. Note that the regression coefficient for one miRNA is NA, as the measurements in TCGA were constant for this miRNA.

C.2 | Classifying cervical cancer stage

The main results of the second data application using methylation data are summarised in Section 5.2. Here we provide the full discussion of the results. We use methylation data from a study on cervical cancer extensively described in [14]. The goal is to find a classifier that best distinguishes normal tissue from CIN3 tissue, a stage with a high risk of progressing to cervical cancer,

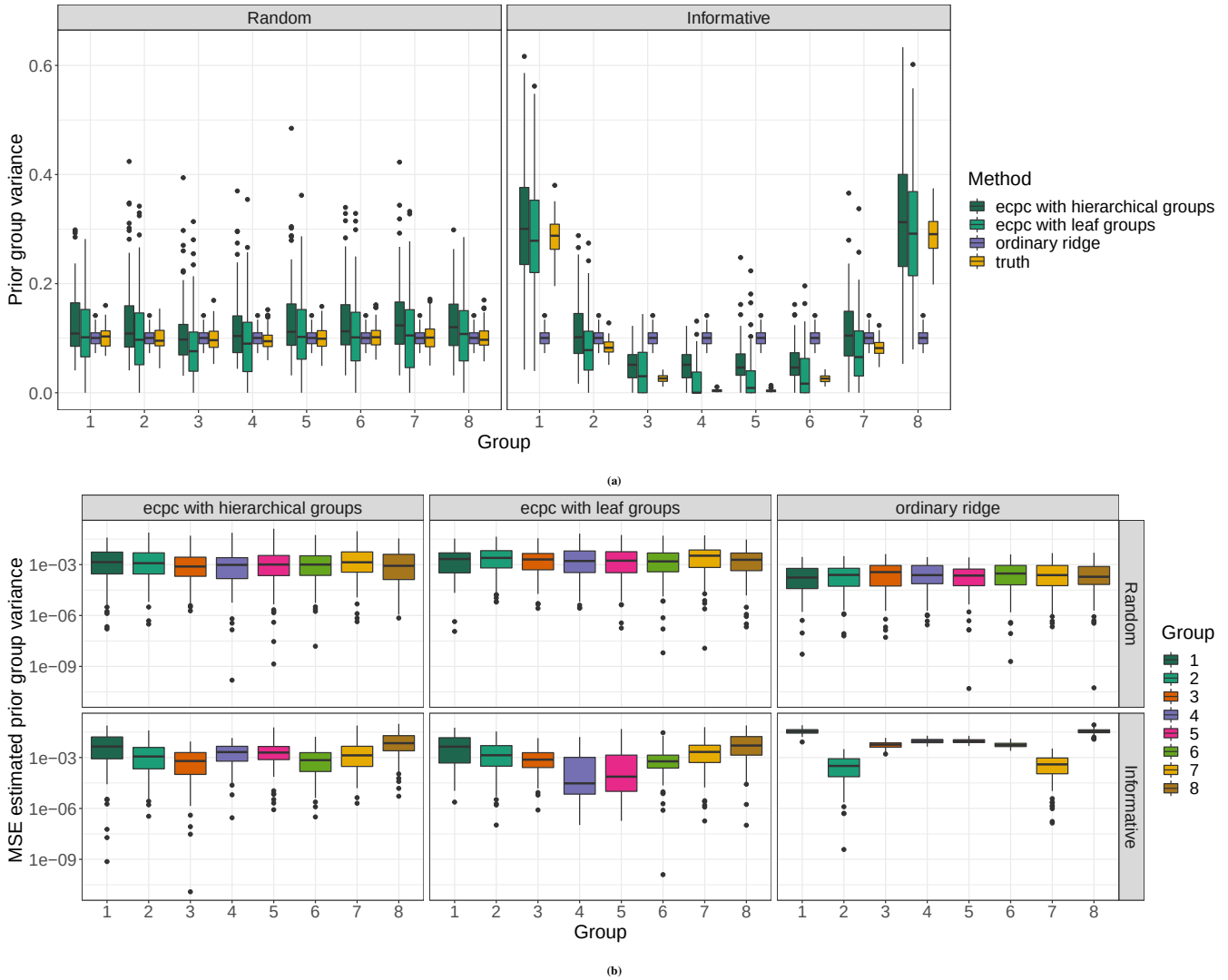

FIGURE S7 Simulation study based on 100 training and test sets and random or informative co-data of hierarchical groups or leaf groups only. a) The estimated prior group variance versus the 'true' prior group variance for several methods. The 'true' prior group variances are given by the prior group variances that maximise the prior distribution given the simulated regression coefficients; b) MSE of the group variance estimates per leaf group for several methods.

in self-taken samples of cervical tissue of women. The methylation levels are measured in $n = 64$ independent individuals with normal tissue (control) or CIN3 tissue (case). After prefiltering, the data consist of methylation levels of $p = 2720$ probes corresponding to unique locations in the DNA. We apply `ecpc` with and without post-hoc selection with the following two co-data sets, illustrated in Figure S20: 1) `CpG-islands`: five non-overlapping groups based on the genomic annotation of distance to the closest CpG-island. A CpG-location is a location on the DNA where a C base precedes a G base, with regions of a relatively high ratio of CpG locations called CpG-islands. DNA methylation is a molecular mechanism that is known to play a role in cancer development. The five groups are, ordered in increasing distance: CpG-island, North Shore, South Shore, North Shelf and South Shelf. We use the default ridge shrinkage as extra level of shrinkage on the group level; 2) `p-values`: continuous p-values for each probe are obtained from an external, similar study[15]. These data cannot be used directly for the classifier as the contamination by different cell types in these samples differs substantially from that of the primary data, the self-obtained samples. However, probes with lower p-values can be expected to be more important for the prediction than probes with high p-values. We adaptively discretise the p-values in a similar manner as the lfdrs in the first application.

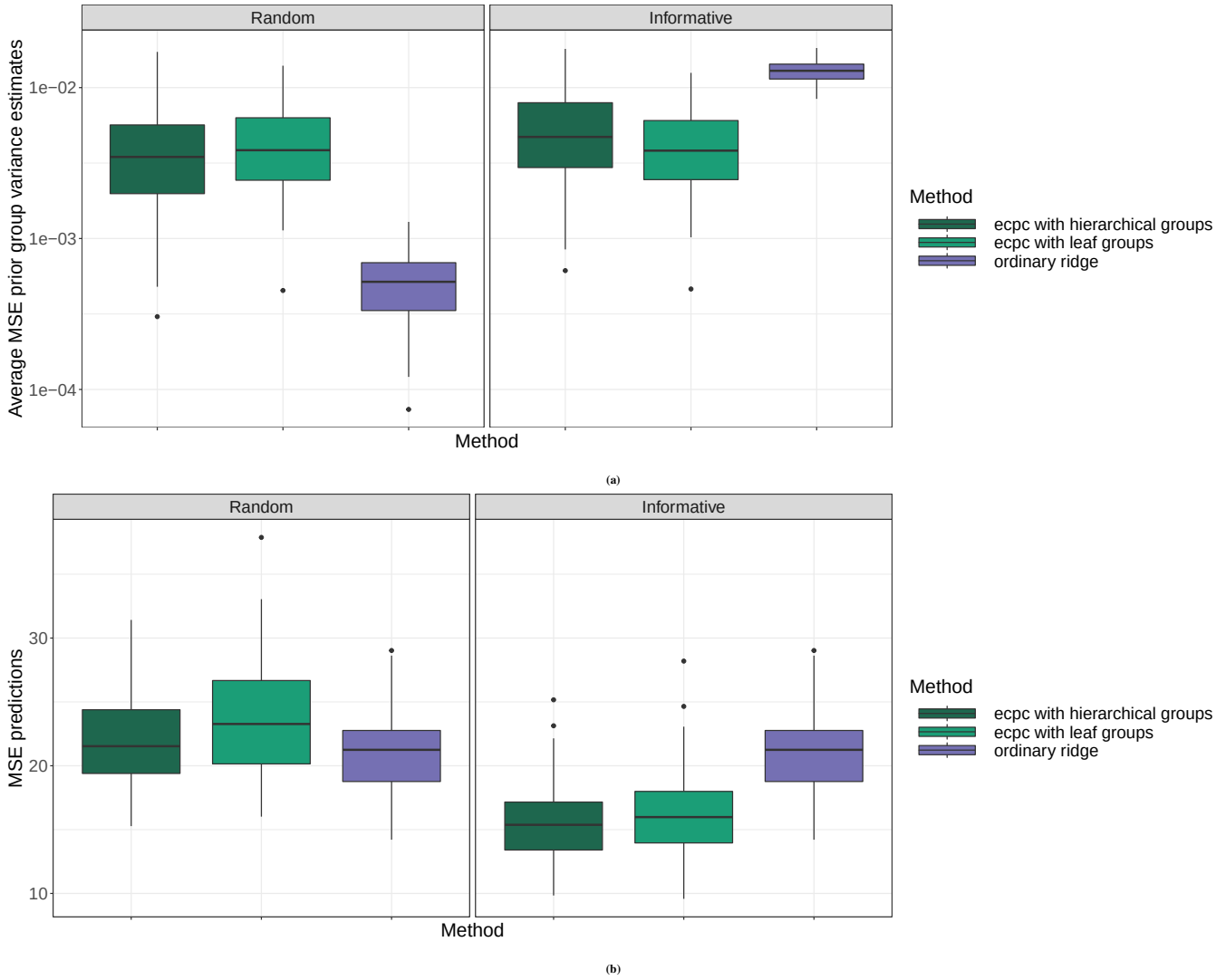

FIGURE S8 Simulation study based on 100 training and test sets and random co-data (left) or informative co-data (right) for various number of groups and several methods. The lines indicate the mean and the shaded bands indicate the 25%, 75% quantiles. a) MSE of the group prior variance estimates averaged over the groups; b) MSE of the predictions on the test sets.

We perform a 20-fold cross-validation to assess performance in terms of AUC for various dense, group sparse and covariate sparse methods. Different folds rendered similar results as shown below. Again, we show the results for the default posterior selection strategy, using an additional L1-penalty. This matched or outperformed other posterior selection strategies (Figure S24). Including standard deviations as another co-data group set as in the first application rendered similar results in terms of performance.

Interpretation of estimated hyperparameters. Figure S21 shows the estimated group set weights and group weights across the folds. The `p-value` group set is the only group set that is selected in all folds, indicating that this group set is more informative for the prediction than the `CpG-islands` group set. The Island and South shelf group obtain group weights higher than 1 and are deemed more important for the prediction. The magnitude of regression coefficients in the CpG-island group is on average around $\sqrt{20} \approx 4.5$ times as large as the global average. For covariates in the group with smallest p-values, the average magnitude of regression coefficients is around $\sqrt{30} \approx 5.5$ times as large as the global average. Groups with lower average p-value obtain a higher prior variance or equivalently, lower penalty.

Performance. Figure S22 shows the AUC versus the number of selected parameters for several dense and covariate sparse methods. First, compared to other dense models, `ecpc` performs similar to `GRridge` and `ordinary ridge`, and outperforms

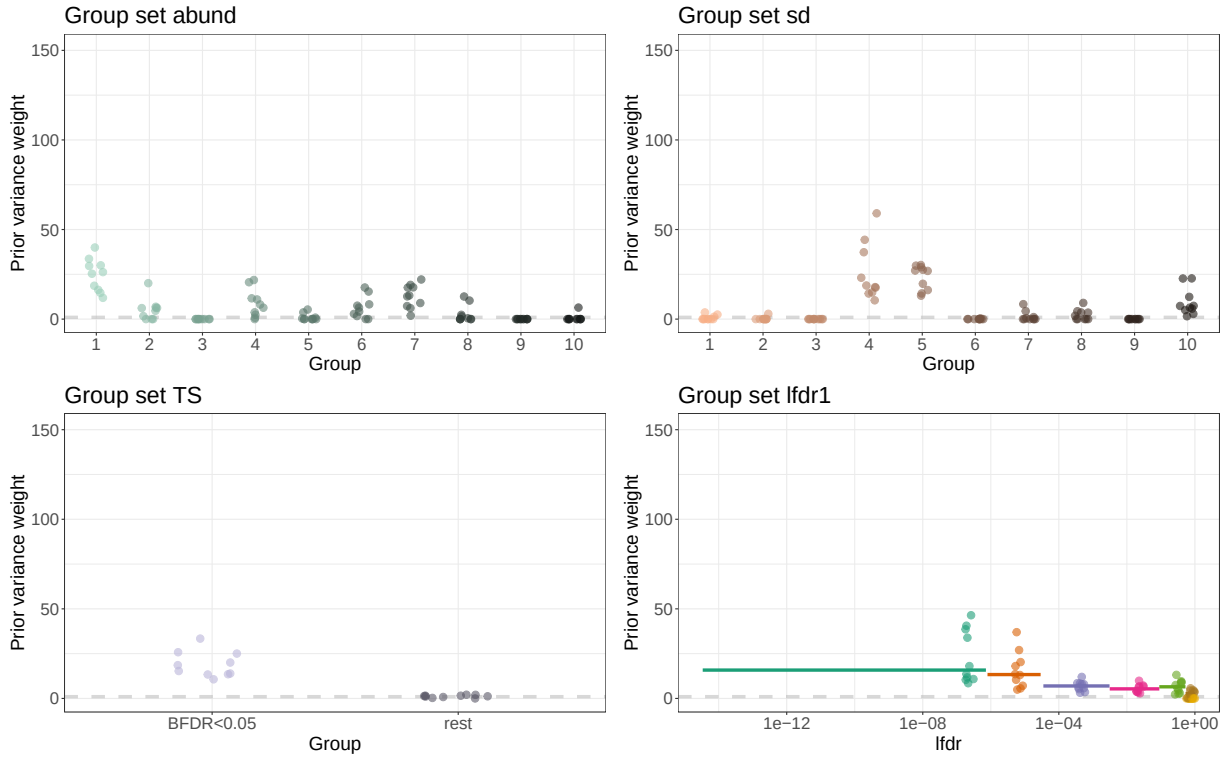

FIGURE S9 Results of 10-fold CV in miRNA data example. Estimated local variance for the first four group sets.

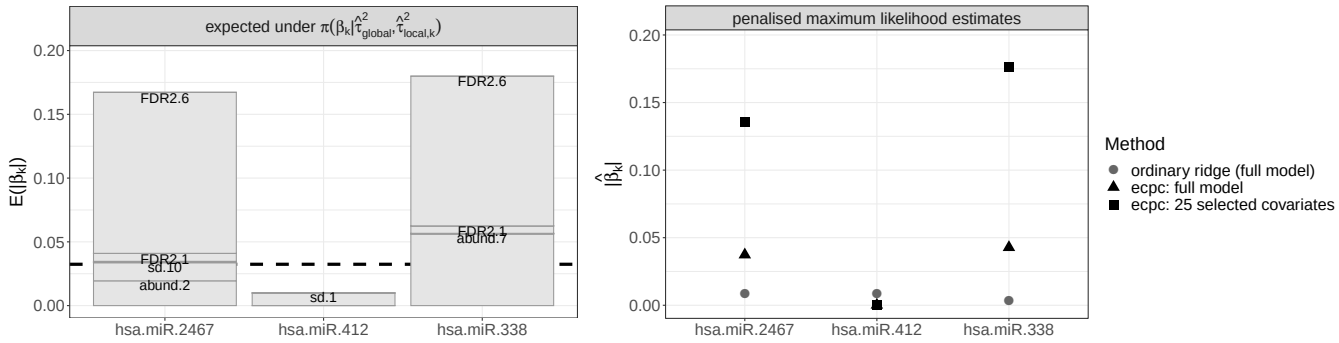

FIGURE S10 Effect of co-data on β_k estimates for several miRNAs. The prior expected magnitude of β_k is a sum of contributions of each group to which the covariate belongs (see Section A.9). Covariates in multiple important groups obtain higher weight than the globally expected average magnitude (dashed horizontal line). The values of $|\hat{\beta}_k|$ corresponding to the miRNAs hsa.miR.2467 and hsa.miR.412 are approximately equal when estimated with ordinary ridge, and larger than that of hsa.miR.338. For hsa.miR.2467 and hsa.miR.338, these coefficients are boosted in `ecpc`, as both belong to multiple important groups, whereas that of hsa.miR.412 is shrunk.

`random forest`. Then, compared to other covariate sparse models, `GRridge` outperforms the other methods for models with more than five selected covariates. `ecpc` results in a peak performance of an AUC= 0.73 at 4 parameters, outperforming the benchmark `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$. While `ecpc` is slightly superior to `GRridge` for very sparse models, its performance initially decreases when including more covariates, and then closes up on `GRridge` again when approaching 100 covariates. We conjecture that this is due to the extremer weights `ecpc` assigns to the smallest p-value group. Furthermore, we apply `graper` to the leaf groups of the hierarchical p-value group set, found to be most important by `ecpc` (Figure S21). Then, `graper` slightly outperforms `ecpc` in the dense setting, with an AUC of 0.71 and is competitive in the sparse setting, with an

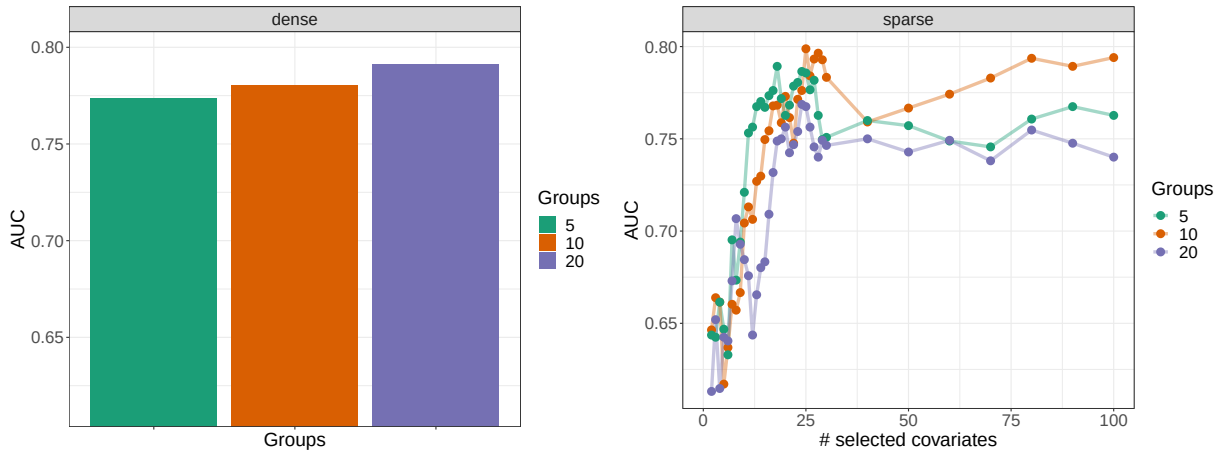

FIGURE S11 Results of 10-fold CV in miRNA data example. AUC performance when abundance and standard deviation are discretised in 5, 10 or 20 groups.

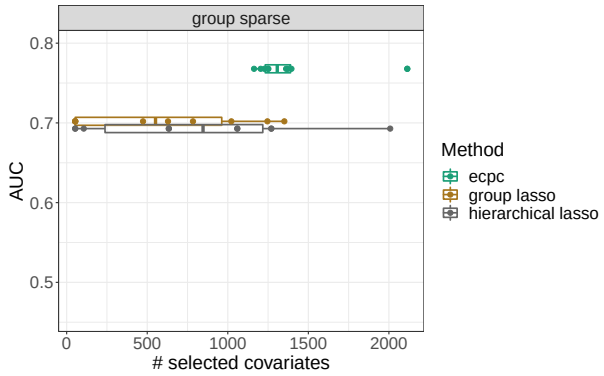

FIGURE S12 Results of 10-fold CV in miRNA data example. AUC in various group sparse models, with the boxplot and points illustrating the variance in selected number of variables in the folds.

AUC of 0.70. Note however that here `graper` uses information from `ecpc` on the most informative group set for these data, which may introduce a benefit for the former.

Besides, `ecpc` is combined with a lasso penalty on the group level to obtain a group sparse model. The group lasso and hierarchical lasso use a LOG penalty similar as used in the first data application described above. Hierarchical lasso selects only the one or two groups with lowest average p-value and slightly outperforms group lasso and the group sparse version of `ecpc` (Figure S23).

Covariate selection. Similarly as in the first data application, `ecpc` facilitates posterior selection, as the distribution of the estimated regression coefficients is more heavy-tailed as compared to when `ordinary ridge` is used (Figure S25). To assess covariate selection stability, we perform the same analysis based on subsamples of the data as used and described for the first data application above. Again, `ecpc` results in a larger overlap between selections when compared to `elastic net` for $\alpha = 0.3$ and $\alpha = 0.8$ (Figure S26) with similar performance in terms of AUC (Figure S27).

C.3 | Classifying lymph node metastasis

We apply `ecpc` to classify presence of lymph node metastasis (LNM). The data and three co-data sets are preprocessed and described by Te Beest et al. [16]. The data consist of RNAseqv2 gene expression profiles from $n = 133$ HPV negative samples for $p = 12838$ probes. The co-data are: 1) `signature`: two non-overlapping groups for a group of genes that have previously been identified as gene signature (group 2) and one with the rest (group 1). We use no hypershrinkage. 2) `p-values`: continuous

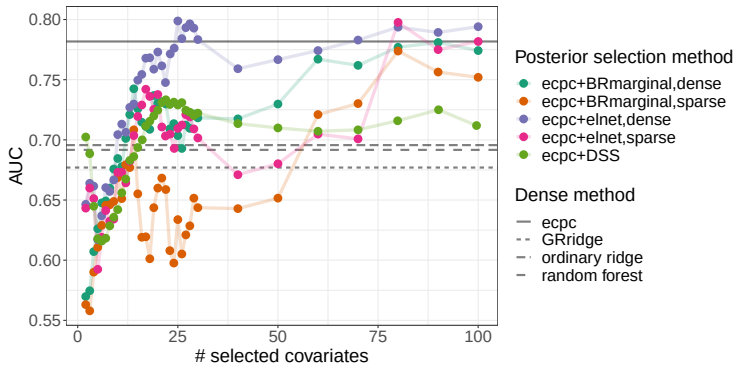

FIGURE S13 Results of 10-fold CV in miRNA data example. AUC for sparse models using various post-hoc selection methods.

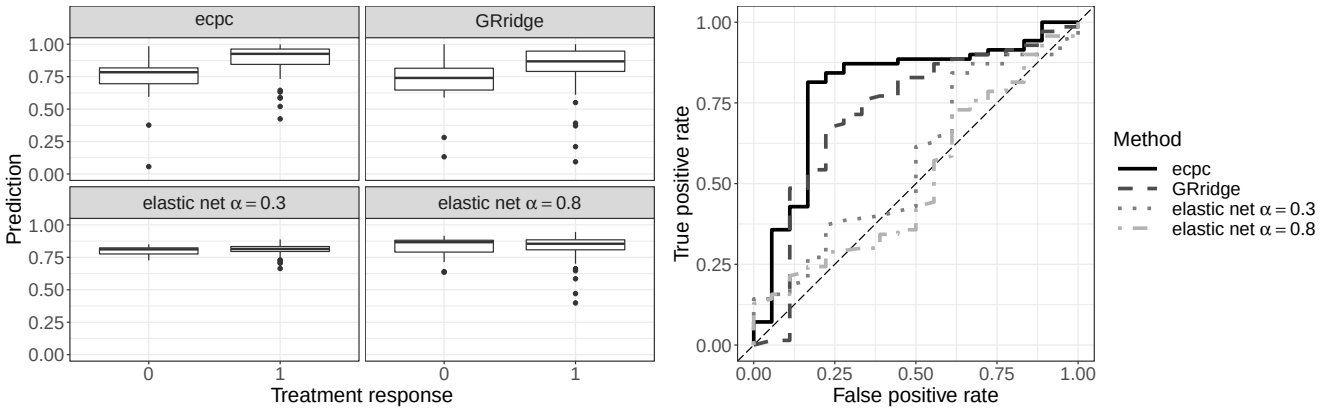

FIGURE S14 Results of 10-fold CV in miRNA data example for different models with 25 selected covariates. Left: out-of-bag predictions versus the observed treatment response. Right: ROC curves.

p-values are obtained from an external, similar study on microarray data. We adaptively discretise the p-values in a similar manner as the lfdrs in the first application. 3) `cis correlation`: continuous cis correlations between DNA copy number and RNAseqv2 data. The DNA copy number data is measured in the same patients. We adaptively discretise the correlations in a similar, but mirrored manner as the lfdrs in the first application; splitting only the groups with higher correlations. We test performance in terms of AUC for various dense and covariate sparse methods on independent test data with 97 samples [16].

Interpretation of estimated hyperparameters. Figure S28 shows the estimated group set weights and group weights. The `p-values` group set obtains most weight. The group with genes from the signature obtains higher weight than the group with the rest of the genes, corroborating the importance of the gene signature. However, this group set is less important than the other two group sets, and obtains zero group set weight. Covariates in the group with smallest p-values have on average $\sqrt{75} \approx 8.6$ times larger effect size than the global average effect size. Genes with higher cis correlation with DNA copy number obtain higher prior variance weight, with the group around a correlation of 0.5 obtaining the largest weight.

Performance. Figure S29 shows the AUC versus the number of selected parameters for several dense and covariate sparse methods. Our method `ecpc` outperforms other dense models, improving the AUC from the co-data agnostic `ordinary ridge` from 0.69 to 0.72. Note that `graper` is applied here to the leaf groups of the `p-values` group set only, found to be most important by `ecpc`. Compared to sparse models, `ecpc` is competitive for highly sparse models and outperforms other methods for models with more than 50 covariates. The method `graper` applied for the sparse setting using a spike-and-slab prior obtained a competitive performance of an AUC of 0.69, but does not select covariates.

Covariate selection. To assess covariate selection stability, we perform the same analysis based on subsamples of the data as used and described for the first data application above. We use 50 stratified subsamples from the LNM data and use the same,

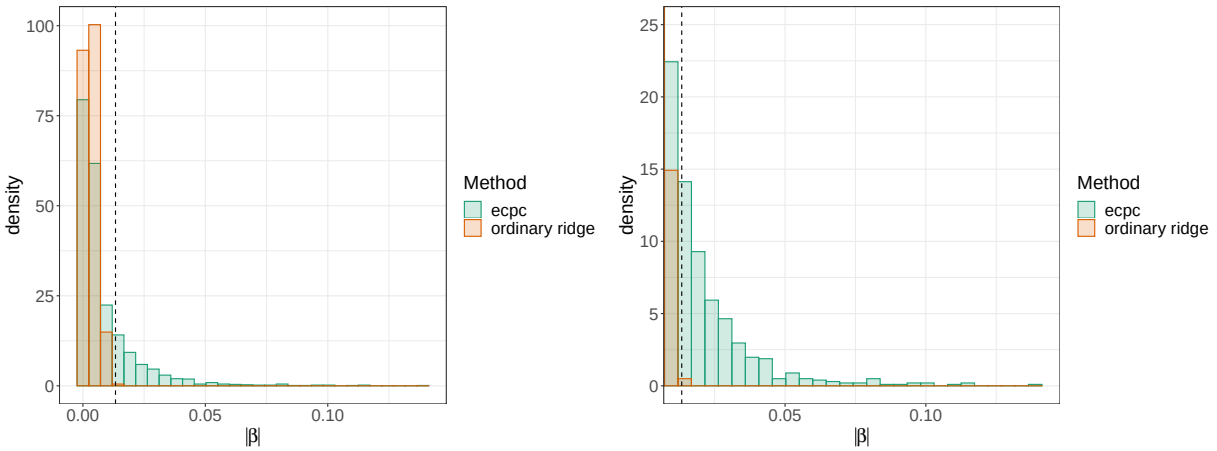

FIGURE S15 miRNA data example. Left: histogram and density plot of absolute value of estimated regression coefficients using `ecpc` or `ordinary ridge`. Right: histogram of highest 0.1 quantile of the absolute value of the regression coefficients. `ecpc` results in more heavy-tailed distributed estimates compared to `ordinary ridge`.

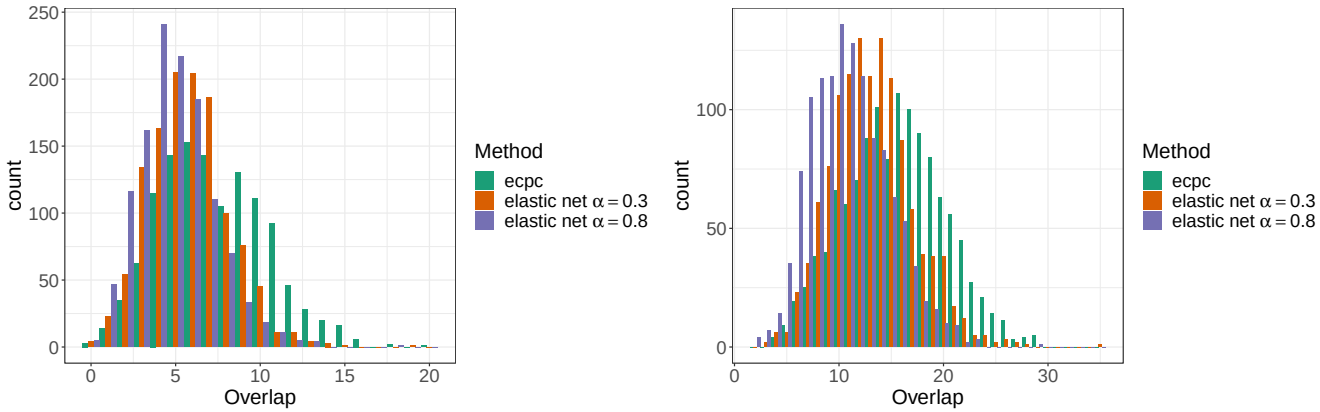

FIGURE S16 Results based on 50 stratified subsamples and corresponding test sets in miRNA data example. Histograms of overlap between selections of 25 (left) or 50 (right) markers for the methods `ecpc` and `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$.

independent test set to assess performance. Our method `ecpc` results in a larger overlap between selections when compared to `elastic net` for $\alpha = 0.3$ and $\alpha = 0.8$ (Figure S30), with better performance in terms of AUC (Figure S31).

References

1. van de Wiel M, Lien T, Verlaat W, Wieringen vW, Wilting S. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in medicine* 2016; 35: 368–381.

2. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C Applied Stat.* 1992; 41(1): 191–201.

3. Meijer RJ, Goeman JJ. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal* 2013; 55(2): 141–155.

4. Wieringen vWN. Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169* 2015.

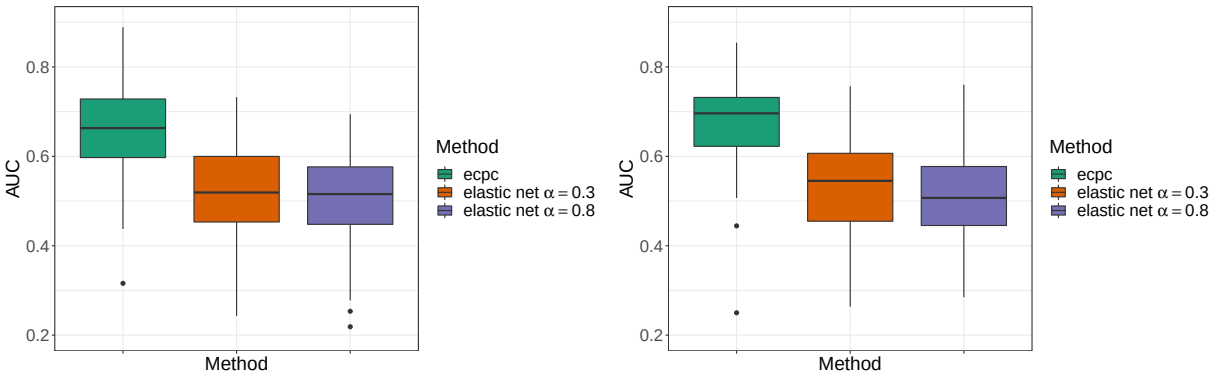

FIGURE S17 Results based on 50 stratified subsamples and corresponding test sets in miRNA data example. Boxplots of the AUC performance of `ecpc`, `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$ on the test set based on selections of 25 covariates (left) or 50 covariates (right) in each subsample.

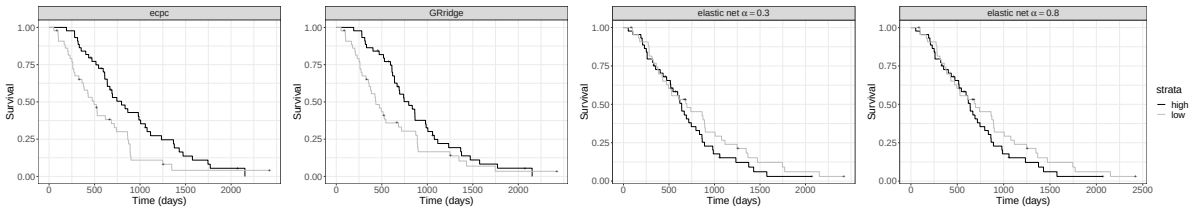

FIGURE S18 miRNA data. Kaplan Meier estimates of the survival function of overall survival (OS) for the samples with high (≥median) or low (<median) probability on clinical benefit for several models with 25 selected markers.

5. Yan X, Bien J, others . Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science* 2017; 32(4): 531–560.

6. Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In: ACM. ; 2009: 433–440.

7. Novianti PW, Snoek BC, Wilting SM, Wiel v. dMA. Better diagnostic signatures from RNAseq data through use of auxiliary co-data. *Bioinformatics* 2017; 33(10): 1572–1574.

8. Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. In: AISTATS. ; 2009: 73–80.

9. Bondell HD, Reich BJ. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association* 2012; 107(500): 1610–1624.

10. Hahn PR, Carvalho CM. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* 2015; 110(509): 435–448.

11. Velten B, Huber W. Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes. *Biostatistics* 2019. kxz034doi: 10.1093/biostatistics/kxz034

12. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* 2015; 19(1A): A68.

13. Rauschenberger A, Ciocănea-Teodorescu I, Jonker MA, Menezes RX, van de Wiel MA. Sparse classification with paired covariates. *Advances in Data Analysis and Classification* 2019: 1–18.

14. Verlaat W, Snoek BC, Heideman DA, et al. Identification and validation of a 3-gene methylation classifier for HPV-based cervical screening on self-samples. *Clinical Cancer Research* 2018: clincanres–3615.

miRNA	coefficient	p-value	method
hsa.mir.31	2.73E-02	1.89E-03	ecpc
hsa.mir.31	2.20E-02	9.70E-03	elastic net $\alpha = 0.3$
hsa.mir.2467	6.77E-01	6.48E-03	elastic net $\alpha = 0.8$
hsa.mir.2467	6.64E-01	8.26E-03	elastic net $\alpha = 0.3$
hsa.mir.2467	6.52E-01	9.06E-03	ecpc
hsa.mir.338	7.19E-03	9.70E-03	ecpc
hsa.mir.503	4.40E-02	1.16E-02	ecpc
hsa.mir.335	1.87E-02	2.55E-02	ecpc
hsa.mir.335	1.77E-02	3.91E-02	GRidge
hsa.mir.3145	2.08E-01	5.19E-02	ecpc
hsa.mir.3617	2.31E-01	5.96E-02	elastic net $\alpha = 0.8$
hsa.mir.3617	2.04E-01	1.02E-01	elastic net $\alpha = 0.3$
hsa.mir.552	-1.06E-02	6.46E-02	GRridge
hsa.mir.552	-6.25E-03	2.63E-01	ecpc
hsa.mir.4750	-9.00E-01	2.09E-01	elastic net $\alpha = 0.3$
hsa.mir.29c	4.43E-03	2.17E-01	GRridge
hsa.mir.29c	3.52E-03	3.33E-01	ecpc
hsa.mir.6873	-4.90E-01	2.86E-01	elastic net $\alpha = 0.8$
hsa.mir.892a	3.03E-01	2.59E-01	ecpc
hsa.mir.892a	2.67E-01	3.20E-01	elastic net $\alpha = 0.8$
hsa.mir.146b	4.65E-03	2.81E-01	elastic net $\alpha = 0.8$
hsa.mir.146b	4.50E-03	3.00E-01	elastic net $\alpha = 0.3$
hsa.mir.30e	-2.55E-03	2.60E-01	ecpc
hsa.mir.30e	-2.14E-03	3.48E-01	GRridge
hsa.mir.7974	-5.08E-02	3.20E-01	ecpc
hsa.mir.7974	-4.94E-02	3.38E-01	GRridge
hsa.mir.181d	2.16E-02	2.14E-01	ecpc
hsa.mir.181d	1.28E-02	4.74E-01	GRridge
hsa.mir.95	1.84E-02	3.47E-01	GRridge
hsa.mir.135b	-7.49E-03	2.67E-01	ecpc
hsa.mir.135b	-4.81E-03	4.80E-01	GRridge
hsa.mir.17	-5.52E-03	1.73E-01	ecpc
hsa.mir.17	-2.23E-03	6.28E-01	GRridge
hsa.mir.548ar	-2.29E-01	3.98E-01	ecpc
hsa.mir.548ar	-1.54E-01	5.51E-01	GRridge
hsa.mir.183	-1.07E-03	3.45E-01	ecpc
hsa.mir.183	-5.02E-04	6.48E-01	GRridge
hsa.mir.380	5.16E-02	4.37E-01	elastic net $\alpha = 0.3$
hsa.mir.380	3.76E-02	5.67E-01	elastic net $\alpha = 0.8$
hsa.mir.421	-2.97E-02	5.17E-01	GRridge
hsa.mir.224	6.74E-03	5.47E-01	GRridge
hsa.mir.592	8.24E-03	3.46E-01	elastic net $\alpha = 0.3$
hsa.mir.592	-7.78E-04	9.26E-01	elastic net $\alpha = 0.8$
hsa.mir.3622a	-1.20E-01	5.79E-01	elastic net $\alpha = 0.8$
hsa.mir.3622a	-7.86E-02	7.21E-01	elastic net $\alpha = 0.3$
hsa.mir.3929	1.83E-01	5.40E-01	elastic net $\alpha = 0.8$
hsa.mir.3929	9.96E-02	7.61E-01	elastic net $\alpha = 0.3$
hsa.mir.3200	2.03E-02	5.57E-01	elastic net $\alpha = 0.3$
hsa.mir.3200	1.82E-02	6.35E-01	GRridge
hsa.mir.3200	1.05E-02	7.65E-01	ecpc
hsa.mir.4659b	2.23E-01	6.74E-01	elastic net $\alpha = 0.8$
hsa.mir.4659b	2.04E-01	6.96E-01	elastic net $\alpha = 0.3$
hsa.mir.195	-1.23E-02	5.60E-01	GRridge
hsa.mir.195	5.20E-03	8.11E-01	ecpc
hsa.mir.431	-2.53E-02	3.89E-01	ecpc
hsa.mir.431	-6.57E-04	9.84E-01	GRridge
hsa.mir.6780a	1.94E-01	6.38E-01	elastic net $\alpha = 0.8$
hsa.mir.6780a	-1.36E-01	7.54E-01	elastic net $\alpha = 0.3$
hsa.mir.4454	2.03E-02	7.39E-01	GRridge
hsa.mir.4467	-2.99E-01	7.70E-01	elastic net $\alpha = 0.3$
hsa.mir.4467	-2.78E-01	7.85E-01	elastic net $\alpha = 0.8$
hsa.mir.6761	1.84E-02	7.44E-01	elastic net $\alpha = 0.8$
hsa.mir.6761	2.89E-03	9.59E-01	elastic net $\alpha = 0.3$
hsa.mir.98	-5.16E-03	7.74E-01	elastic net $\alpha = 0.8$
hsa.mir.98	5.11E-04	9.79E-01	elastic net $\alpha = 0.3$
hsa.mir.18a	7.35E-04	9.73E-01	GRridge
hsa.mir.549a	-4.20E-04	9.94E-01	GRridge
hsa.mir.548g	-1.69E+01	9.96E-01	elastic net $\alpha = 0.8$
hsa.mir.548g	-1.55E+01	9.96E-01	elastic net $\alpha = 0.3$
hsa.mir.6801	6.51E-04	9.98E-01	elastic net $\alpha = 0.3$
hsa.mir.548ap	NA	1.00E+00	elastic net $\alpha = 0.3$
hsa.mir.548ap	NA	1.00E+00	elastic net $\alpha = 0.8$

TABLE S3 As Table S2, but sorted per miRNA and increasing p-value.

15. Farkas SA, Milutin-Gašperov N, Grce M, Nilsson TK. Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer. *Epigenetics* 2013; 8(11): 1213–1225.

16. Te Beest DE, Mes SW, Wilting SM, Brakenhoff RH, Wiel v. dMA. Improved high-dimensional prediction with Random Forests by the use of co-data. *BMC bioinformatics* 2017; 18(1): 1–11.

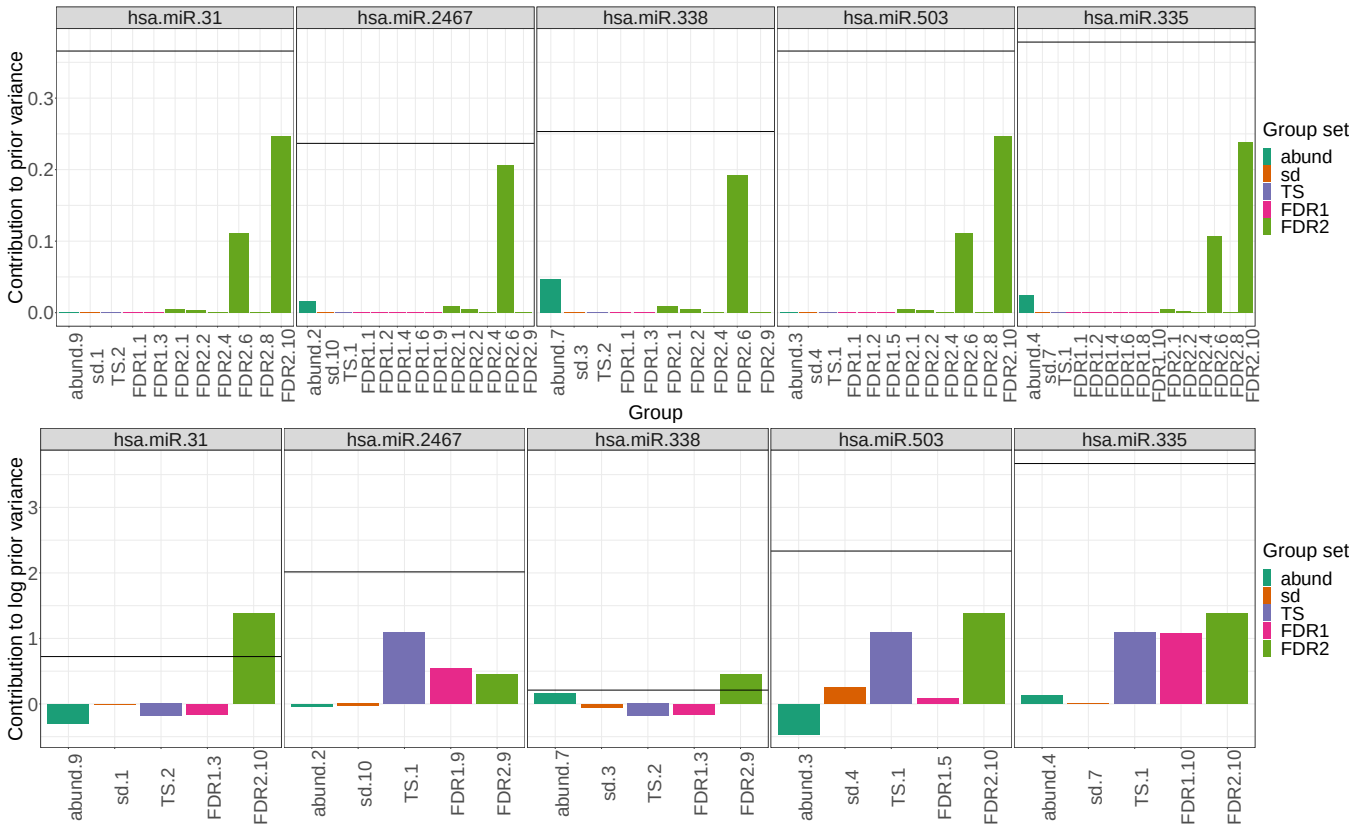

FIGURE S19 miRNA data. Composition of the covariate-specific prior variance of `ecpc` and `GRridge` for the top 5 covariates in Table S3. Top: in `ecpc`, the covariate-specific prior variance for β_k (horizontal line) is the sum over co-data weights and group weights. The y-axis shows the contribution to this sum of each group and group set to which the specific covariate belongs. Bottom: in `GRridge`, multiple group sets are handled implicitly and penalty multipliers of groups of different group sets are multiplied. Hence, contributions of each group set are summed on the logarithmic scale. The total log prior variance is the sum of these contributions (horizontal line) plus the log of the global prior variance.

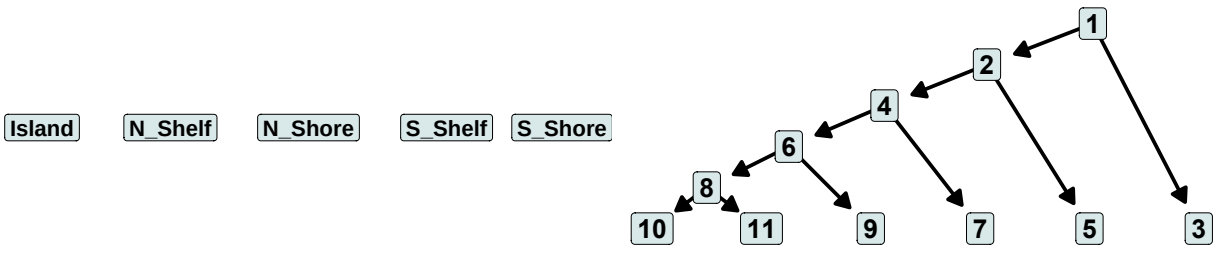

FIGURE S20 Illustration of the co-data group sets used the Verlaat data. Left: `CpG-islands`, five non-overlapping groups ordered in distance to the nearest CpG-island. Right: `p-values`, groups on the left correspond to lower p-values and are split recursively into two groups. The hierarchical structure on the groups is used in the extra level of shrinkage to find a discretisation that fits the data well as described in Section 3.4.1 in the main article.

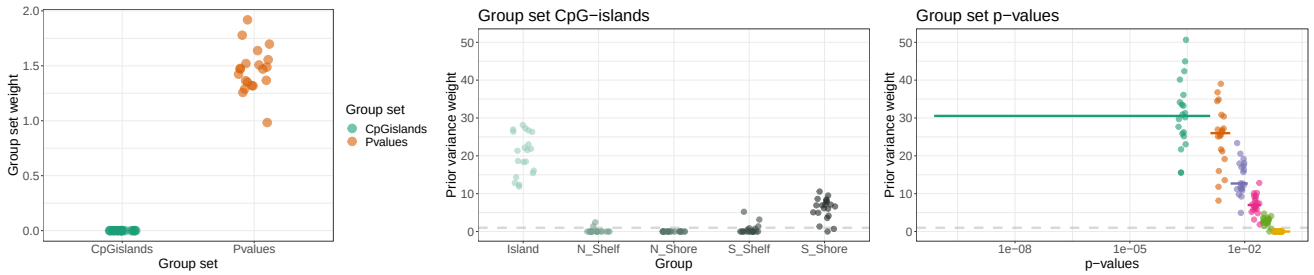

FIGURE S21 Results of 20-fold CV in Verlaat data example. Left: estimated co-data group set weights. Middle: estimated group weights in `CpG-islands` group set. Right: estimated local variance in `p-values` group set; the median is shown from covariates in the leaf groups of the hierarchical tree illustrated in Figure S20 (horizontal line ranging from the minimum to maximum p-value in that group), and the corresponding estimates in the folds are shown (points, jittered along the median p-value in that group). A larger prior variance corresponds to a smaller penalty.

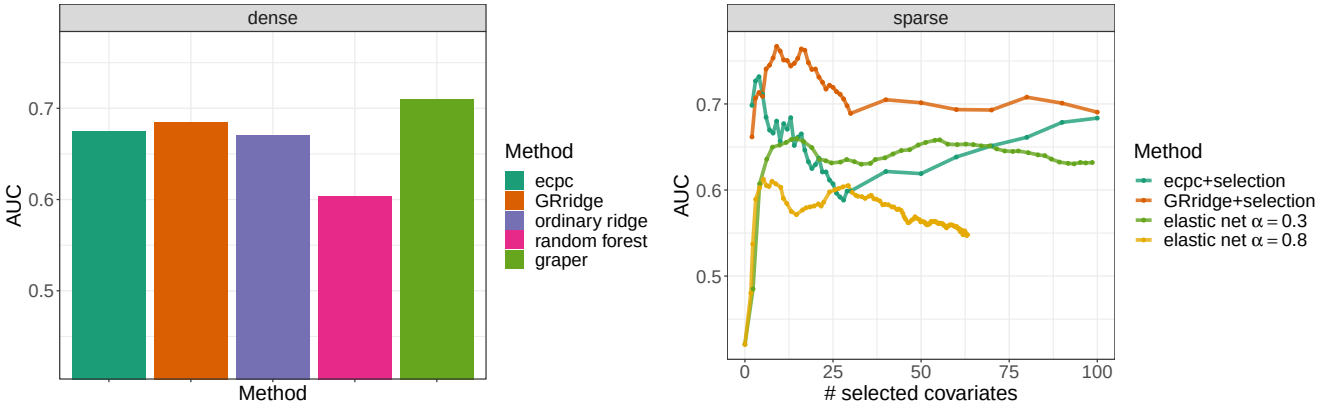

FIGURE S22 Results of 20-fold CV in Verlaat data example. AUC in various dense models (left) and sparse models (right).

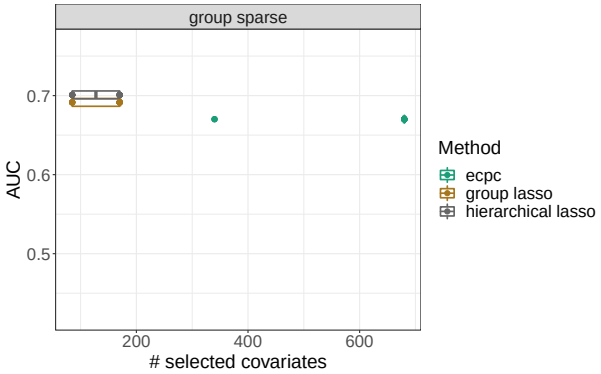

FIGURE S23 Results of 20-fold CV in Verlaat data example. AUC in various group sparse models

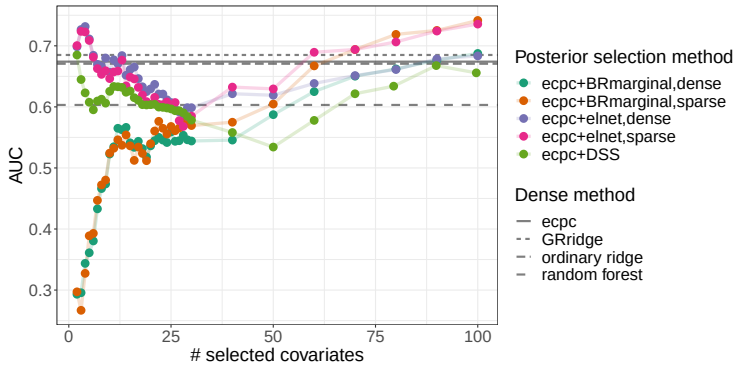

FIGURE S24 Results of 20-fold CV in Verlaat data example. AUC for sparse models using various post-hoc selection methods.

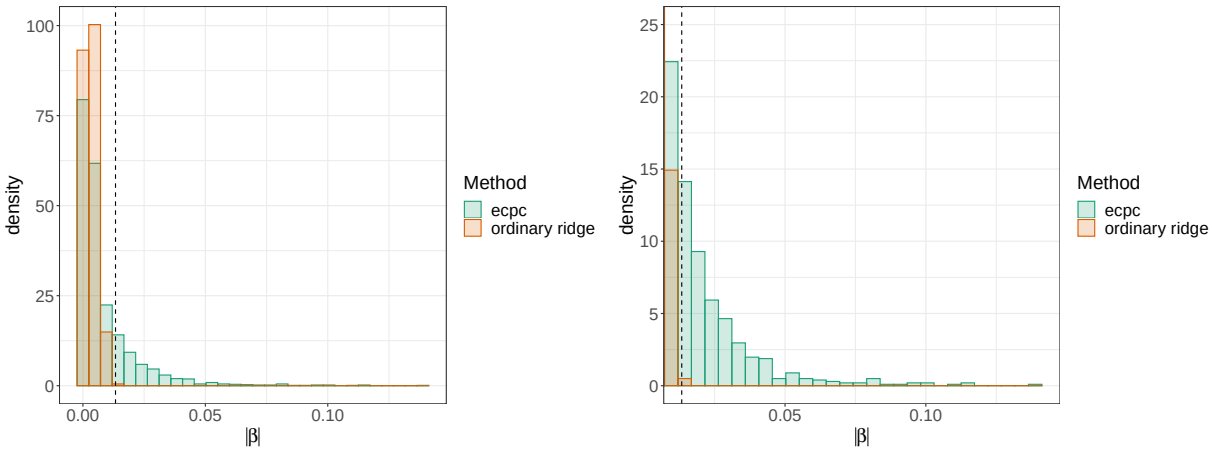

FIGURE S25 Verlaat data example. Left: histogram and density plot of absolute value of estimated regression coefficients using `ecpc` or `ordinary ridge`. Right: histogram of highest 0.1 quantile of the absolute value of the regression coefficients. `ecpc` results in more heavy-tailed distributed estimates compared to `ordinary ridge`.

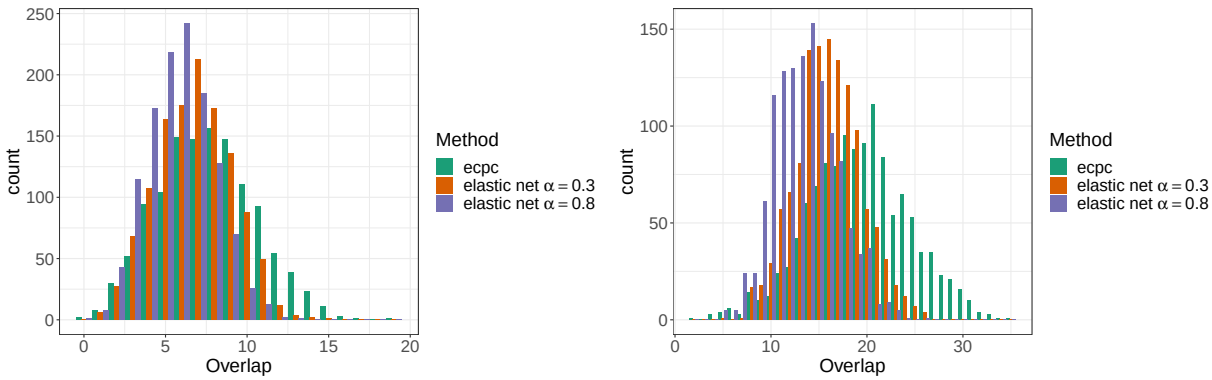

FIGURE S26 Results based on 50 stratified subsamples in Verlaat data example. Histogram of number of overlapping variables in pairwise comparisons of selections of 25 covariates (left) or 50 covariates (right) in each subsample, for the methods `ecpc`, `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$.

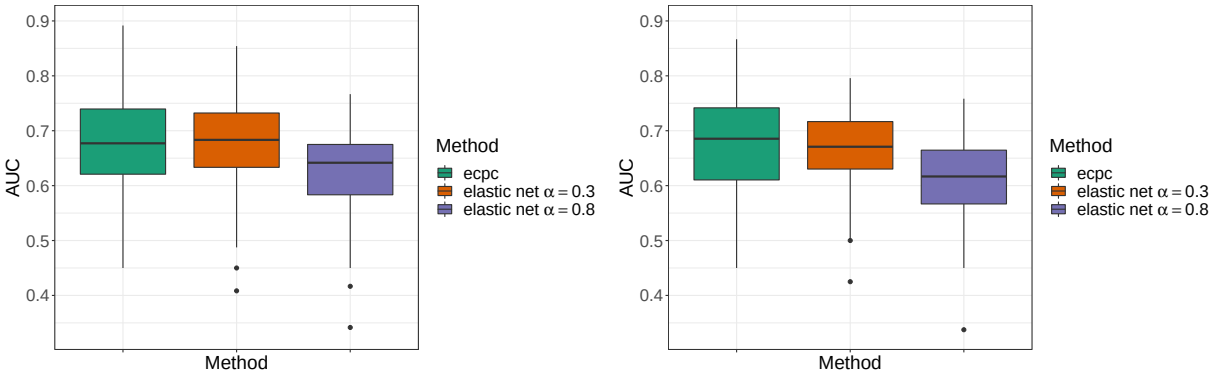

FIGURE S27 Results based on 50 stratified subsamples and corresponding test sets in Verlaat data example. Boxplot of the AUC performance of `ecpc`, `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$ on the test set based on selections of 25 covariates (left) or 50 covariates (right) in each subsample.

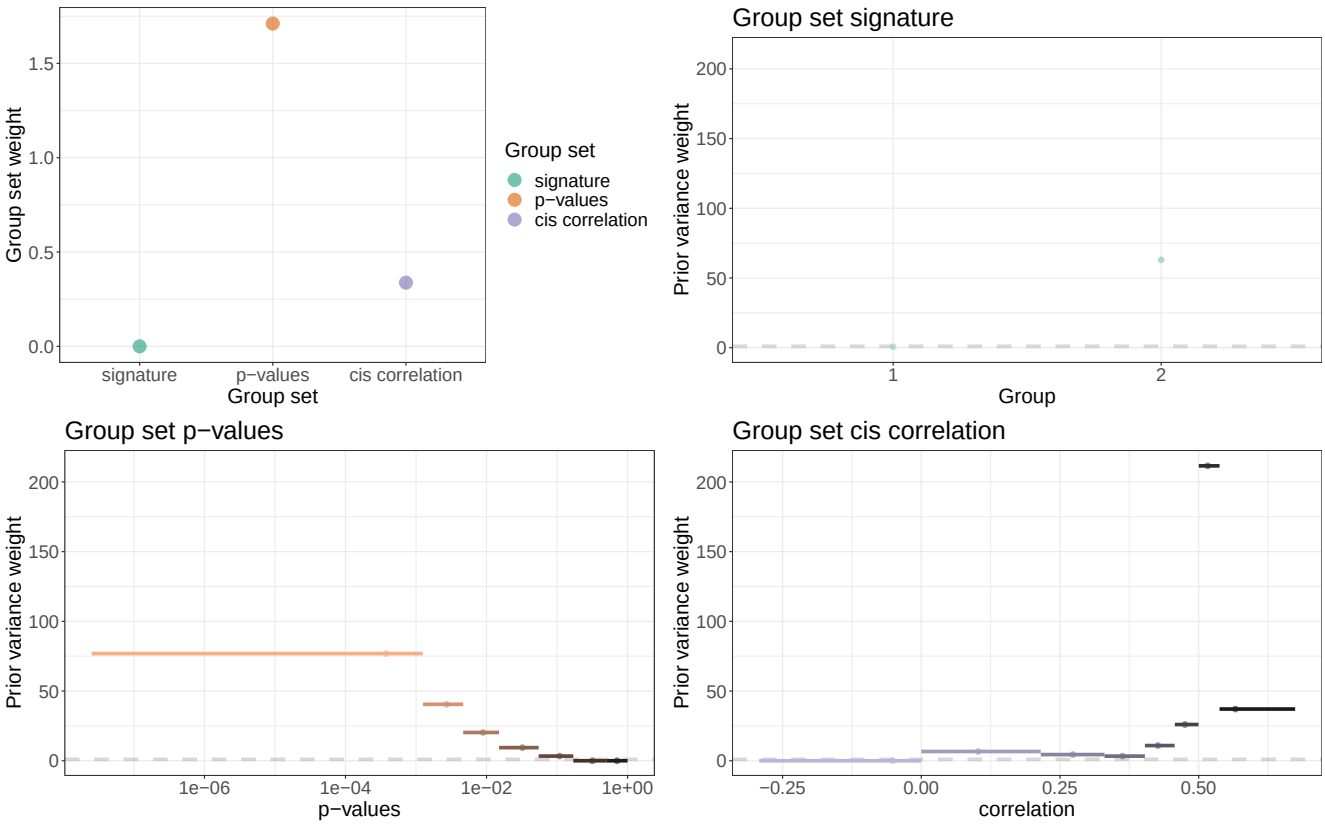

FIGURE S28 Results of LNM data example. Top left: estimated co-data group set weights. Top right: estimated group weights in `signature` group set. Bottom: estimated local variance in `p-values` group set (left) and `cis correlation` group set; the median is shown from covariates in the leaf groups of the hierarchical tree used in the adaptive discretisation (horizontal line ranging from the minimum to maximum p-value in that group). A larger prior variance corresponds to a smaller penalty.

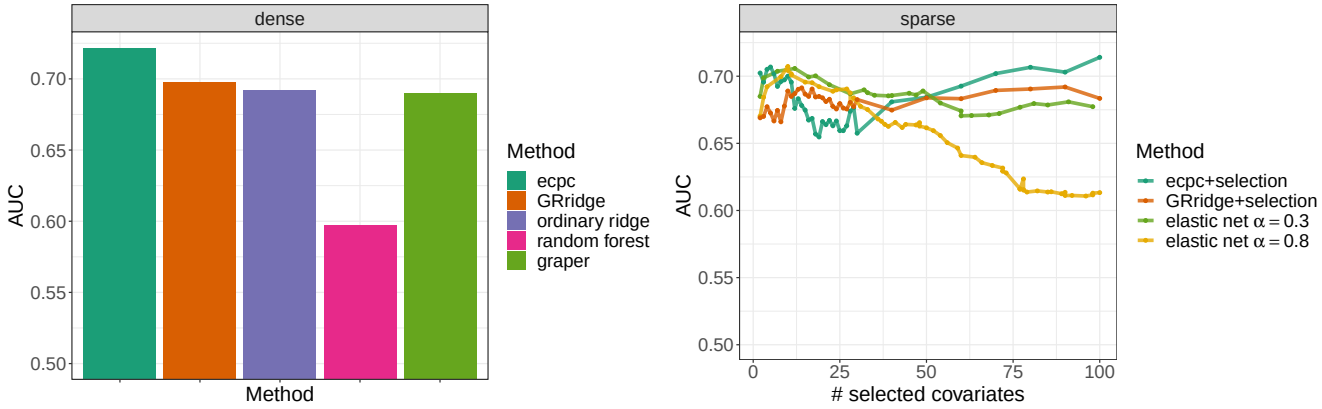

FIGURE S29 Results of LNM data example. AUC in various dense models (left) and sparse models (right).

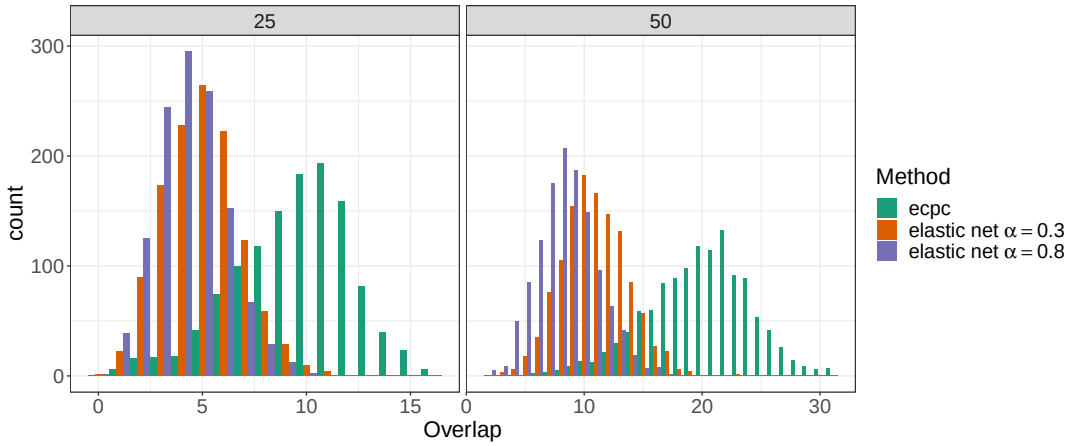

FIGURE S30 Results based on 50 stratified subsamples of the LNM data. Histogram of number of overlapping variables in pairwise comparisons of selections of 25 covariates (left) or 50 covariates (right) in each subsample, for the methods `ecpc`, `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$.

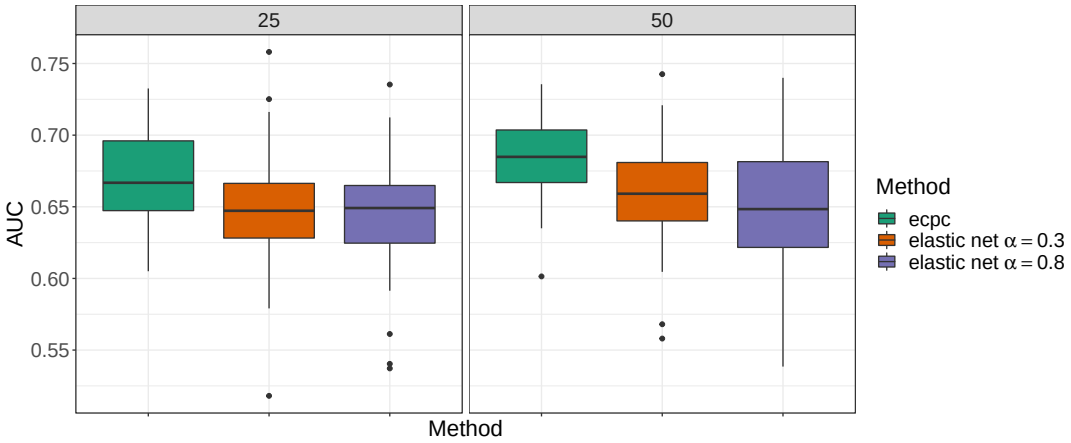

FIGURE S31 Results based on 50 stratified subsamples of the LNM data. Boxplot of the AUC performance of `ecpc`, `elastic net` with $\alpha = 0.3$ and $\alpha = 0.8$ on the test set based on selections of 25 covariates (left) or 50 covariates (right) in each subsample.