SUPPLEMENTARY MATERIALS for Quantifying predictive uncertainty of aphasia severity in stroke patients with sparse heteroscedastic Bayesian high-dimensional regression

Anja Zgodic[1], Ray Bai[2], Jiajia Zhang[1], Yuan Wang[1], Christopher Rorden[3], Alexander C. McLain[1*]

[1*]Department of Epidemiology and Biostatistics, University of South Carolina, 915 Greene Street, Columbia, South Carolina, 29208, U.S.
[2]Department of Statistics, University of South Carolina, 1523 Greene Street, Columbia, South Carolina, 29208, U.S.
[2]Department of Psychology, University of South Carolina, 915 Greene Street, Columbia, South Carolina, 29208, U.S.

*Corresponding author(s). E-mail(s): mclaina@mailbox.sc.edu;

Keywords: Bayesian variable selection, ECM algorithm, Empirical Bayes, Heteroscedasticity, High-dimensional linear regression

A Expanded Methods Section

A.1 Model framework and posteriors

The complete model framework is presented in the main text. We briefly review the models on the mean and the variance. Our model on the mean allows for sparse and non-sparse predictors, though it does not mandate including non-sparse predictors. We write the model on the mean as

$$\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\gamma\beta}) + \boldsymbol{Z\varphi} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent the $n \times p$ sparse and $n \times z$ non-sparse predictor matrices, $\boldsymbol{\gamma\beta}$ is a Hadamard product, $\boldsymbol{\gamma} \in \{0,1\}^p$, and $\boldsymbol{\varphi} \in \mathcal{R}^z$. Let $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ with $\mathcal{D}_i = (Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{V}_i)$ denote the observed data. We add the following parametric model on the diagonal variance matrix $\boldsymbol{\Sigma}$,

$$-\log\{diag(\boldsymbol{\Sigma})\} = \boldsymbol{V\omega}, \tag{2}$$

where $\boldsymbol{\omega} \in \mathcal{R}^v$.

We use a Parameter-Expanded Expectation-Conditional-Maximization (PX-ECM) algorithm for estimation. Within the algorithm, the parameter estimates at iteration t are used to obtain the expected complete-data log-posterior distribution with respect to $\boldsymbol{\gamma}$,

$$E_{\boldsymbol{\gamma}}\left\{\log\ p(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\omega}|\boldsymbol{\mathcal{D}}, \boldsymbol{\gamma})|\boldsymbol{\mathcal{D}}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\varphi}^{(t)}, \boldsymbol{\omega}^{(t)}\right\}, \tag{3}$$

which is sequentially maximized to obtain maximum *a posteriori* (MAP) estimates $\boldsymbol{\beta}^{(t+1)}$, $\boldsymbol{\varphi}^{(t+1)}$, and $\boldsymbol{\omega}^{(t+1)}$ (i.e., $\boldsymbol{\Sigma}^{(t+1)}$) (discussed in Sections A.3 and A.4). For γ_k, plug-in empirical Bayes estimators are used to estimate the posterior expectation $\boldsymbol{p} = (p_1, \ldots, p_p)$ where $p_k = P(\gamma_k = 1|\boldsymbol{\mathcal{D}}, \pi)$ (discussed in Section A.4).

To show the conditional posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, we split $(\boldsymbol{X}, \boldsymbol{\beta})$ into $(\boldsymbol{X}_\gamma, \boldsymbol{\beta}_\gamma)$ when $\gamma_k = 1$ and $(\boldsymbol{X}_{\bar{\gamma}}, \boldsymbol{\beta}_{\bar{\gamma}})$ when $\gamma_k = 0$. Conditional on $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ (i.e., $\boldsymbol{\Sigma}$), the posterior distribution of $\boldsymbol{\gamma\beta}$ and $\boldsymbol{\varphi}$ is

$$\begin{pmatrix}\boldsymbol{\beta}_\gamma \\ \boldsymbol{\varphi}\end{pmatrix}\Bigg|(\boldsymbol{\mathcal{D}}, \boldsymbol{\omega}, \boldsymbol{\gamma}) \sim N\left\{(\boldsymbol{C}_\gamma'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma)^{-1}\boldsymbol{C}_\gamma'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, (\boldsymbol{C}_\gamma'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_\gamma)^{-1}\right\}$$

where $\boldsymbol{C}_\gamma = (\boldsymbol{X}_\gamma\ \boldsymbol{Z})$, while $\boldsymbol{\beta}_{\bar{\gamma}}|(\boldsymbol{\mathcal{D}}, \boldsymbol{\omega}, \boldsymbol{\gamma}) \sim \delta_0(\cdot)$ a point mass at zero. Given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\varphi}$ the posterior for $\boldsymbol{\omega}$ has density

$$f_\omega(\boldsymbol{\omega}) \propto \exp\left\{\boldsymbol{c}_\omega'\boldsymbol{H}_\omega\boldsymbol{\omega} - \boldsymbol{\kappa}_\omega'\exp\left(\boldsymbol{H}_\omega\boldsymbol{\omega}\right)\right\}, \tag{4}$$

which is proportional to an $MLG(\boldsymbol{0}, \boldsymbol{H}_\omega, \boldsymbol{c}_\omega, \boldsymbol{\kappa}_\omega)$ distribution with

$$\begin{aligned}\boldsymbol{H}_\omega &= \begin{bmatrix}\boldsymbol{V} \\ c^{-1/2}\sigma_\omega^{-1}\boldsymbol{I}_v\end{bmatrix},\ \boldsymbol{c}_\omega = \left(\frac{1}{2}\boldsymbol{1}_n', c\boldsymbol{1}_v'\right)',\ \text{and} \\ \boldsymbol{\kappa}_\omega &= \left(\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{\gamma\beta}) - \boldsymbol{Z\varphi}||^2, c\boldsymbol{1}_v'\right)',\end{aligned} \tag{5}$$

where $\boldsymbol{I}_v$ denotes a $v \times v$ identity matrix, $\boldsymbol{1}_v$ a $v \times 1$ vector of ones, and $||\boldsymbol{x}||^2 = \boldsymbol{xx}$ is a Hadamard product.

A.2 Estimation and notation

We begin this Section with an overview of ECM and PX-EM algorithms. The EM algorithm requires parameterizing a model by including latent parameters. Both latent and unknown parameters are estimated through an iterative process where the expectation is taken over latent parameters (E-step), which is used to maximize the expected

log-likelihood (M-step) to obtain estimates for unknown parameters (Dempster, Laird, & Rubin, 1977). In the ECM algorithm, the single M-step is replaced by multiple computationally simpler CM-steps, helping optimization (Meng & Rubin, 1993). The PX-EM is another extension of the EM approach, where the model is rewritten with auxiliary terms (i.e., expanded parameters) to help with stability and convergence (Liu, Rubin, & Wu, 1998).

We now describe the PROBE algorithm (McLain, Zgodic, & Bondell, 2025). The aim is to perform MAP estimation for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varphi}$ in a high-dimensional setting. PROBE is based on a PX-ECM algorithm, which is a combination of the ECM and PX-EM algorithms (Liu et al., 1998; Meng & Rubin, 1993). The CM-step of the PX-ECM results in coordinate-wise (predictor-specific) optimization, where the remaining parameters are restricted to their values at the previous iteration. As a result, both PROBE and H-PROBE partition the mean of the model (1) by predictor k. The notation $\boldsymbol{A}_{\setminus k}$ indicates the matrix, vector, or collection without predictor or element k. Define $\boldsymbol{W}_k = \boldsymbol{X}_{\setminus k}(\boldsymbol{\gamma}_{\setminus k}\boldsymbol{\beta}_{\setminus k}) + \boldsymbol{Z}\boldsymbol{\varphi} = (W_{1k}, \ldots, W_{nk})'$ and thus $E(\boldsymbol{Y}|\boldsymbol{W}_k) = \boldsymbol{X}_k\beta_k + \boldsymbol{W}_k$, where $\boldsymbol{W}_k$ encompasses the impact of all predictors except $\boldsymbol{X}_k$.

Note that $\boldsymbol{W}_k$ is unknown as it is a function of parameters $(\boldsymbol{\beta}_{\setminus k}, \boldsymbol{\varphi})$ and the "missing data" $\boldsymbol{\gamma}_{\setminus k}$. Since $\boldsymbol{W}_k$ is estimated as part of the algorithm, we use parameter-expansion (Liu et al., 1998) to include parameters α_k, which adjust for the impact of $\boldsymbol{W}_k$ when updating β_k for all $k = 1, \ldots, p$. This results in

$$E(\boldsymbol{Y}|\boldsymbol{W}_k) = \boldsymbol{X}_k\beta_k + \boldsymbol{W}_k\alpha_k, \tag{6}$$

for $k = 1, \ldots, p$. The α_k parameter helps estimate the posterior variance of $\beta_k|(\gamma_k = 1)$ more accurately since it accounts for the dependence between $\boldsymbol{W}_k$ and $\boldsymbol{X}_k$ (discussed further in Section A.3). This posterior variance is required in our implementation of the E-step and used to create prediction intervals.

The core difference between PROBE and H-PROBE is that the latter approach estimates the MAP of $\boldsymbol{\omega}$ in addition to the MAP of $\boldsymbol{\varphi}$ and $\beta_k|(\gamma_k = 1)$ for all $k = 1, \ldots, p$. Table A.1 provides a summary of the parameters estimated by H-PROBE and described in Sections A.3 and A.4.

H-PROBE generally consists of four steps. First, in the CM-steps discussed in Section A.3, $\boldsymbol{\varphi}$ and $\boldsymbol{\beta}_{\setminus k}$ are fixed to their values from the previous iteration when estimating β_k, for all k (Meng & Rubin, 1993). Then, the E-step discussed in Section A.4 updates $E(\boldsymbol{W}_k)$ and $E(\boldsymbol{W}_k^2)$, where the expectations are over $\boldsymbol{\gamma}_{\setminus k}$. Third, after the moments of the $\boldsymbol{W}_k$'s are updated, they are used to perform the subsequent CM-steps. Finally, the MAP estimator of $\boldsymbol{\omega}$ does not have a closed form. As a result, we perform the maximization for this parameter via quasi-Newton optimization (Fletcher, 1987). Algorithm 1 and subsequent Sections describe the H-PROBE steps in sequence.

Along with the notation for $\boldsymbol{W}_k$ and $\boldsymbol{C}_k = (\boldsymbol{X}_k\ \boldsymbol{W}_k)$ described above, we define an 'overall' non-partitioned $\boldsymbol{W}$, denoted by $\boldsymbol{W}_{p+1} = \boldsymbol{X}(\boldsymbol{\gamma}\boldsymbol{\beta})$, and α_{p+1} as expanded parameters for $\boldsymbol{\varphi}$. We also define $\boldsymbol{C}_{p+1} = (\boldsymbol{Z}\ \boldsymbol{W}_{p+1})$. The calculations in Section A.4 require

$$W_{ik}^{(t-1)} = E(W_{ik}|\boldsymbol{\beta}_{\setminus k}^{(t-1)}, \boldsymbol{p}_{\setminus k}^{(t-1)}, \alpha_{p+1}^{(t-1)}, \boldsymbol{\omega}^{(t-1)}, \boldsymbol{\varphi}^{(t-1)}) \text{ for } k \geq 1$$

Table A.1 Parameters and estimates provided by H-PROBE. SM = Supplementary Materials.

Parameter	Estimate	Equation	Parameter Definition
$\boldsymbol{\beta}$	$\tilde{\boldsymbol{\beta}}$	(1), (6)	Vector of sparse regression coefficients for predictors in $\boldsymbol{X}$, conditional on $\boldsymbol{\gamma} = 1$, in the model on the conditional mean.
$\boldsymbol{S}^2$	$\tilde{\boldsymbol{S}}^2$	(11)	Vector of posterior variances of $\boldsymbol{\beta}\|(\boldsymbol{\gamma} = 1)$.
$\boldsymbol{\gamma}$	$\tilde{\boldsymbol{p}}$	(1), (6)	Vector of inclusion indicators for sparse coefficients $\boldsymbol{\beta}$ associated with predictors in $\boldsymbol{X}$.
$\boldsymbol{\varphi}$	$\tilde{\boldsymbol{\varphi}}$	(1)	Vector of non-sparse regression coefficients in the model on the conditional mean.
$\boldsymbol{W}_{p+1}$	$\tilde{\boldsymbol{W}}_{p+1}$	(6)	Overall (non-partitioned) latent parameter $\boldsymbol{W}_{p+1} = \boldsymbol{X}(\boldsymbol{\gamma}\boldsymbol{\beta})$.
α_{p+1}	$\tilde{\alpha}_{p+1}$	(6)	Overall (non-partitioned) coefficient adjusting for the impact of overall (non-partitioned) $\boldsymbol{W}_{p+1}$.
$\boldsymbol{\omega}$	$\tilde{\boldsymbol{\omega}}$	(2)	Vector of non-sparse regression coefficients in the model on the variance.
$\boldsymbol{\Sigma}$	$\tilde{\boldsymbol{\Sigma}}$	(2)	Diagonal variance matrix $\boldsymbol{\Sigma} = \exp\{-(\boldsymbol{V}\boldsymbol{\omega})\}$.

$$
\begin{aligned}
W_{i,p+1}^{(t-1)} &= E(W_{i,p+1}|\boldsymbol{\beta}^{(t-1)}, \boldsymbol{p}^{(t-1)}), \\
\boldsymbol{W}_{\ell}^{(t-1)} &= (W_{1\ell}^{(t-1)}, \ldots, W_{n\ell}^{(t-1)}) \text{ for } \ell \in (1, \ldots, p, p+1)
\end{aligned}
$$

with analogous notation for the second moments $W_{i,p+1}^{2(t-1)}$, $W_{ik}^{2(t-1)}$, and $\boldsymbol{W}_{\ell}^{2(t-1)}$. Some key quantities in the CM-step updates are

$$
\begin{aligned}
(\boldsymbol{W}_{\ell}^{'}\boldsymbol{\Sigma}^{-1})^{(t-1)} &= \sum_{i=1}^{n} W_{i\ell}^{(t-1)}/\sigma_i^{2(t-1)} \\
(\boldsymbol{W}_{\ell}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_{\ell})^{(t-1)} &= \sum_{i=1}^{n} W_{i\ell}^{2(t-1)}/\sigma_i^{2(t-1)} \text{ where } \sigma_i^{2(t-1)} = \exp(-\boldsymbol{V}_i\boldsymbol{\omega}^{(t-1)}).
\end{aligned}
$$

Further, let

$$
(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)} = \begin{pmatrix} \boldsymbol{X}_k'\boldsymbol{\Sigma}^{-1(t-1)}\boldsymbol{X}_k & \boldsymbol{X}_k'(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_k)^{(t-1)} \\ (\boldsymbol{W}_k^{'}\boldsymbol{\Sigma}^{-1})^{(t-1)}\boldsymbol{X}_k & (\boldsymbol{W}_k^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_k)^{(t-1)} \end{pmatrix} \quad \text{for } k = 1, \ldots, p. \tag{7}
$$

We define $(\boldsymbol{C}_{p+1}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{p+1})^{(t-1)}$ similarly to (7) where $\boldsymbol{X}_k$ and $\boldsymbol{W}_k$ are replaced with $\boldsymbol{Z}$ and $\boldsymbol{W}_{p+1}$, respectively.

A.3 CM-step

The CM-steps maximize the expected complete-data log-posterior distributions of (β_k, α_k) for partitions $k \in (1, \ldots, p)$, as well as $(\boldsymbol{\varphi}, \alpha_{p+1}, \boldsymbol{\omega})$ for partition $p+1$. The 'overall' model associated with the non-partitioned $\boldsymbol{W}_{p+1}$ is estimated via $E(\boldsymbol{Y}|\boldsymbol{W}_{p+1}) = \boldsymbol{Z}\boldsymbol{\varphi} + \alpha_{p+1}\boldsymbol{W}_{p+1}$ and focuses on the parameters $(\boldsymbol{\varphi}, \alpha_{p+1}, \boldsymbol{\omega})$. The remaining partitions focus on $(\beta_k, \alpha_k)'$ via (6).

For partitions $k = 1, \ldots, p$, the complete-data log-posterior distribution is denoted by $\log p(\beta_k, \alpha_k|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_k, \boldsymbol{\Gamma}_k)$ and its expectation $E_{\boldsymbol{\gamma}}\{\log p(\beta_k, \alpha_k|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_k, \boldsymbol{\Gamma}_k)|\boldsymbol{\mathcal{D}}, \boldsymbol{\beta}_{\setminus k}, \boldsymbol{\varphi}, \boldsymbol{\omega}, \boldsymbol{\Gamma}_{\setminus k})\}$ is conditional on $\boldsymbol{\beta}_{\setminus k}$, $\boldsymbol{p}_{\setminus k}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\omega}$. $\boldsymbol{\Gamma}_{\setminus k}$ represents the hyperparameters for $\boldsymbol{\beta}_{\setminus k}, \boldsymbol{\alpha}_{\setminus k}, \boldsymbol{\varphi}$, and $\boldsymbol{\omega}$. At iteration t for these partitions, the CM-step maximizes

$$(\hat{\beta}_k^{(t)}, \hat{\alpha}_k^{(t)})' = \text{argmax}_{(\beta_k, \alpha_k)} E_k^{(t-1)}\{\log p(\beta_k, \alpha_k)|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_k, \boldsymbol{\Gamma}_k)\} \tag{8}$$

where $E_k^{(t-1)}$ represents the expectation over $\boldsymbol{\gamma}_{\setminus k}$. For the kth partition of iteration t, the MAP estimator is

$$\begin{pmatrix} \hat{\beta}_k^{(t)} \\ \hat{\alpha}_k^{(t)} \end{pmatrix} = \left\{(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)}\right\}^{-1} \left(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\right)^{(t-1)} \boldsymbol{Y}. \tag{9}$$

For partition $p+1$, the complete-data log-posterior distribution is denoted by $\log p(\boldsymbol{\varphi}, \alpha_{p+1}|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_{p+1}, \boldsymbol{\Gamma})$ and its expectation $E_{\boldsymbol{\gamma}}\{\log p(\boldsymbol{\varphi}, \alpha_{p+1}|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_{p+1}, \boldsymbol{\Gamma})|\boldsymbol{\mathcal{D}}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Gamma})\}$ is conditional on $\boldsymbol{\beta}$, $\boldsymbol{\omega}$, and hyperparameters $\boldsymbol{\Gamma}$. For the overall model at partition $p+1$ and iteration t, the CM-step maximizes

$$(\hat{\boldsymbol{\varphi}}^{(t)}, \hat{\alpha}_{p+1}^{(t)})' = \text{argmax}_{(\boldsymbol{\varphi}, \alpha_{p+1})} E_{\boldsymbol{\gamma}}^{(t-1)}\{\log p(\boldsymbol{\varphi}, \alpha_{p+1})|\boldsymbol{\mathcal{D}}, \boldsymbol{W}_{p+1}, \boldsymbol{\Gamma})\}, \tag{10}$$

and the MAP values for $(\boldsymbol{\varphi}, \alpha_{p+1})$ are

$$\begin{pmatrix} \hat{\boldsymbol{\varphi}}^{(t)} \\ \hat{\alpha}_{p+1}^{(t)} \end{pmatrix} = \left\{(\boldsymbol{C}_{p+1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_{p+1})^{(t-1)}\right\}^{-1} \left(\boldsymbol{C}_{p+1}'\boldsymbol{\Sigma}^{-1}\right)^{(t-1)} \boldsymbol{Y}. \tag{11}$$

For $\boldsymbol{\omega}^{(t)}$, the MAP estimator has no closed form. As a result, we use a quasi-Newton optimizer (Fletcher, 1987) on the log posterior distribution of $\boldsymbol{\omega}$ to find the MAP estimate $\boldsymbol{\omega}^{(t)}$. Then, we obtain $\boldsymbol{\Sigma}^{-1(t)} = \exp\{-(\boldsymbol{V}'\boldsymbol{\omega}^{(t)})\}^{-1}$.

The E-step used in Section A.4 requires an estimate of the posterior variance of $\beta_k|(\gamma_k = 1)$. Here, the posterior covariance of $(\hat{\beta}_k^{(t)}, \hat{\alpha}_k^{(t)})$ is estimated by

$$\{(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)}\}^{-1}\{(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1})^{(t-1)}\boldsymbol{C}_k^{(t-1)}\}\{(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)}\}^{-1}, \tag{12}$$

where $\hat{S}_k^{2(t)}$ denotes the $(1,1)$ element. We use the $\hat{S}_k^{2(t)}$'s to create prediction intervals.

A.4 E-step

Before commencing with the E-step, to accelerate convergence we limit the step size using learning rates $q^{(t)}$ via

$$\begin{aligned} \beta_k^{(t)} &= (1-q^{(t)})\beta_k^{(t-1)} + q^{(t)}\hat{\beta}_k^{(t)}, \text{ and} \\ S_k^{2(t)} &= \{(1-q^{(t)})(S_k^{2(t-1)})^{-1} + q^{(t)}(\hat{S}_k^{2(t)})^{-1}\}^{-1}. \end{aligned} \tag{13}$$

We use $q^{(t)} = \frac{1}{t+1}$, which creates a moving average of β_k over iterations. McLain et al. (2025), Minka and Lafferty (2002), and Vehtari et al. (2020) provide a discussion on learning rates and their implications.

To estimate $p_k = P(\gamma_k = 1|\boldsymbol{Y}, \pi_0)$ while maintaining uninformative priors, we use a plug-in empirical Bayes estimator, which is motivated by two-groups approach to multiple testing where many test statistics with zero and non-zero expectations are available (Castillo & Roquain, 2020; Efron, 2008; Liang, Paulo, Molina, Clyde, & Berger, 2008). We define test statistics as $\mathcal{T}_k^{(t)} = \beta_k^{(t)}/S_k^{(t)}$ with distribution $(1-\gamma_k)f_0(\cdot) + \gamma_k f_1(\cdot)$, where $f_0(\cdot) \sim N(0,1)$ and $f_1(\cdot)$ is unknown and related to f_β. We also require π_0, the proportion of null hypotheses. This yields the plug-in empirical Bayes estimator of the posterior expectation of γ_k as

$$p_k^{(t)} = 1 - \frac{\hat{\pi}_0^{(t)} f_0(\mathcal{T}_k^{(t)})}{\hat{f}^{(t)}(\mathcal{T}_k^{(t)})}, \tag{14}$$

where $\hat{\pi}_0^{(t)}$ and $\hat{f}^{(t)}$ are empirical Bayes estimates of π_0 and f based on the observed $\mathcal{T}_k^{(t)}$'s. In our simulations and data analyses, we use $\hat{\pi}^{(t)} = \sum_k I(P_k^{(t)} \geq \lambda)/\{p \times (1-\lambda)\}$, based on Storey (2007), where $P_k^{(t)}$ is a two-sided p-value for $\mathcal{T}_k^{(t)}$ and $\lambda = 0.1$ (Blanchard & Roquain, 2009), and Gaussian kernel density estimation on $\boldsymbol{\mathcal{T}}^{(t)} = (\mathcal{T}_1^{(t)}, \ldots, \mathcal{T}_p^{(t)})$ to obtain the estimated marginal distribution $\hat{f}^{(t)}$ (Silverman, 1986).

Finally, we estimate the first and second moments of $\boldsymbol{W}_\ell$. These moments are expectations of $\boldsymbol{W}_\ell$ over the unknown $\boldsymbol{\gamma}$. Through the use of the ECM, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ are fixed at their estimates from the previous iteration. Independence between γ_k's allows effective computation to be performed at the observation i level through

$$W_{i,p+1}^{(t)} = E\{\boldsymbol{X}_i(\boldsymbol{\gamma}\boldsymbol{\beta})|\boldsymbol{\beta}^{(t)}, \boldsymbol{p}^{(t)}\} = \boldsymbol{X}_i(\boldsymbol{\beta}^{(t)}\boldsymbol{p}^{(t)}), \tag{15}$$

and $W_{i,p+1}^{2(t)} = E(W_{i,p+1}^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{p}^{(t)}) = Var(W_{i,p+1}|\boldsymbol{\beta}^{(t)}, \boldsymbol{p}^{(t)}) + (W_{i,p+1}^{(t)})^2$ where

$$Var(W_{i,p+1}|\boldsymbol{\beta}^{(t)}, \boldsymbol{p}^{(t)}) = \boldsymbol{X}_i^2 \left\{\boldsymbol{\beta}^{(t)2}\boldsymbol{p}^{(t)}(1-\boldsymbol{p}^{(t)})\right\}. \tag{16}$$

Algorithm 1 The H-PROBE algorithm

Initialize $\boldsymbol{W}^{(0)}$, $\boldsymbol{W}^{2(0)}$, and $\boldsymbol{\Sigma}^{(0)}$
while $CC^{(t)} \geq \chi^2_{1,\varepsilon}$ and $\max(\boldsymbol{p}^{(t)}) > 0$ **do**
CM-step
Use $\boldsymbol{W}_\ell^{(t-1)}$ and $\boldsymbol{W}_\ell^{2(t-1)}$ to estimate $\boldsymbol{\xi}_\ell^{(t)}$ for $\ell = 1, \ldots, p, p{+}1$ via (9)–(11).
E-step
(a) Calculate $\beta_k^{(t)}$ and $S_k^{2(t)}$ using (13) for all k.
(b) Estimate $\hat{f}^{(t)}$ and $\hat{\pi}_0^{(t)}$ and use them to calculate $\boldsymbol{p}^{(t)}$ via (14).
(c) Calculate $\boldsymbol{W}^{(t)}$ and $\boldsymbol{W}^{2(t)}$ via (15) and (16).
Calculate $CC^{(t)}$ and check convergence.

Online calculations of $\boldsymbol{W}_k^{(t)}$ and $\boldsymbol{W}_k^{2(t)}$ are made by subtracting the contributions of the kth predictor from $\boldsymbol{W}_{p+1}^{(t)}$ and $\boldsymbol{W}_{p+1}^{2(t)}$, respectively. As a result, the high-dimensional matrix computations in (15) and (16) only need to be made once per iteration.

A.5 Model checks

Algorithm 1 shows H-PROBE steps in sequence. Upon convergence, H-PROBE provides MAP estimates $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{\varphi}}$, $\tilde{\alpha}_{p+1}$, $\tilde{\boldsymbol{\omega}}$, $\tilde{\boldsymbol{\Sigma}}$ as well as $\tilde{S}_k^2$, the posterior variance of $\tilde{\beta}_k|(\gamma_k = 1)$, for all k. While the properties of $\tilde{\boldsymbol{\varphi}}$ are not the focus of this research, we do wish to account for the uncertainty it contributes to the MAP estimates in prediction intervals. To this end, let
$\tilde{\Psi} = \{(\boldsymbol{C}'_{p+1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}_{p+1})^{(t-1)}\}^{-1}\{(\boldsymbol{C}'_{p+1}\tilde{\boldsymbol{\Sigma}}^{-1})^{(t-1)}\boldsymbol{C}_{p+1}^{(t-1)}\}\{(\boldsymbol{C}'_{p+1}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{C}_{p+1})^{(t-1)}\}^{-1}$
denote the estimated posterior covariance of $(\tilde{\boldsymbol{\varphi}}, \tilde{\alpha}_{p+1})$. Prediction intervals for a future observation are presented in detail in the main text.

In addition to the guidelines we described at the end of Section 2.3 of the main article, plots of the residuals or the log-squared residuals can also be used to determine if transformations of variables are necessary. Figure A.1 shows an example of residual plots used to validate model assumptions. Both panels are from homoscedastic LASSO models with 50 observations of 100 sparse predictors. The data used in each model was generated such that the residual variance is associated with a linear covariate (Panel (a)) and a non-linear covariate ($variable + variable^2$, Panel (b)). In both plots, the covariate has a relationship with the residual variance. This association is weaker in Panel (a), while it is stronger in Panel (b). The residuals in Panel (a) show that the linear form of the heterogeneity variable fulfills the parametric assumptions of the variance model in (2), whereas Panel (b) shows that additional non-linear transformations of the heterogeneity variable are needed to fulfill the assumptions of the variance model.

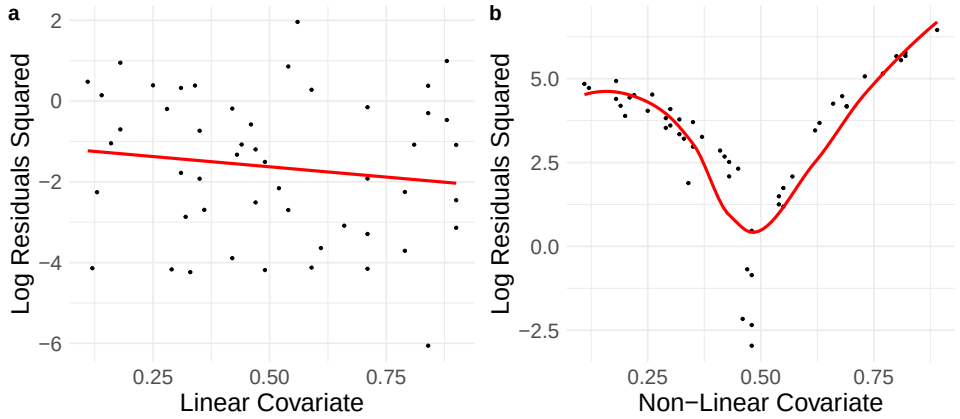

Fig. A.1 This figure shows residuals from a high-dimensional homoscedastic LASSO model. The Log Residuals Squared are plotted against a linear and non-linear heterogeneity variable. Plots of the residuals can be used to determine if transformations of variables are necessary to address heteroscedasticity. The residuals in Panel (a) show that the linear form of the heterogeneity variable fulfills the assumptions of the variance model in (2), whereas Panel (b) shows that additional transformations of the heterogeneity variable are needed to fulfill the assumptions of the variance model.

B Additional Simulation Results

B.1 Model and predictive performance

We examined the performance of H-PROBE by focusing on the Root Mean Squared Error (RMSE, main text) and Median Absolute Deviation (MAD) of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$, where data $\boldsymbol{X}$ consists of new observations not used during estimation (test set, $n = 400$). Figure B.2 shows that for all simulation settings except one, H-PROBE had the lowest MAD, especially when the proportion of signals, the number of predictors on the mean, or the effect size of $\boldsymbol{\beta}$ were higher. Results are shown for $\eta_\beta = 0.8$, binary $\boldsymbol{X}$, and $\Sigma_{cor} = 20$ for brevity. In the setting where H-PROBE slightly underperformed compared to PROBE, the MAD was 6% higher for H-PROBE.

To capture the overall joint uncertainty of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, we performed a Markov Chain Monte Carlo sensitivity check based on credible intervals for each parameter. Within each simulation setting, we examined the average Empirical Coverage Probability (ECP) of each parameter's credible intervals across the simulation iterations. Figure B.3 shows that overall, parameter $\boldsymbol{\beta}$ had lower error and credible interval ECP consistently around 95%. The spread of the ECPs for $\boldsymbol{\omega}$ were markedly wider than for $\boldsymbol{\beta}$ and did not cover the nominal coverage level across many settings. There were no consistent trends across simulation settings regarding overall joint uncertainty. For example, in some settings with lower proportions of signals ($\pi = 1\%$), the average credible interval ECP for parameter $\boldsymbol{\omega}$ covered the nominal level, more than settings with higher proportions of signals ($\pi = 5\%$), but only when the signal-to-noise (SNR) ratio was $SNR = 1$. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$ were very similar and are omitted.

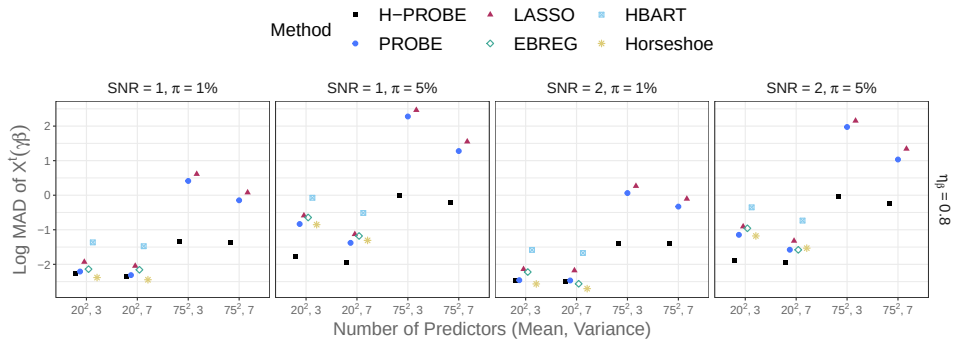

Fig. B.2 This figure shows model performance results from the numerical studies, via the log Median Absolute Deviation (MAD) of $\boldsymbol{X}'(\gamma\boldsymbol{\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set) for six methods. The methods we compare are: H-PROBE (black squares), PROBE (blue circles), LASSO (maroon triangles), EBREG (green diamonds), HBART (blue squares with inner cross), or a Bayesian model with a horseshoe prior (yellow stars), for selected simulation settings.

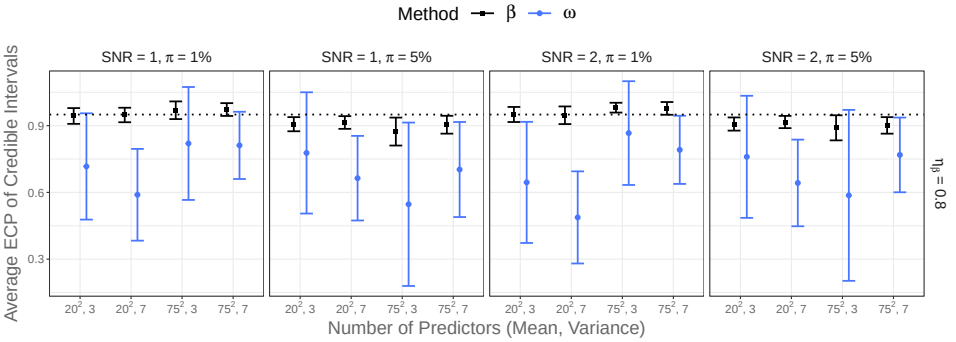

Fig. B.3 This figure shows an approximate measure of the joint estimation error of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ for selected simulation settings. The approximate measure is the average Empirical Coverage Probability (ECP) of the parameters' Monte Carlo Markov chain credible intervals, across the simulation iterations. Black squares indicate the average credible interval ECP for $\boldsymbol{\beta}$, while blue circles represent the ECP for $\boldsymbol{\omega}$.

We also examined the bias and standard deviation of the $\tilde{\boldsymbol{\omega}}$ coefficient estimates, from the model on the variance. Results are shown for binary $\boldsymbol{X}$, $\Sigma_{cor} = 20$, and $n = 400$ for brevity. Figure B.4 shows the average bias for the intercept ω_1, the first continuous coefficient ω_2, and the first binary coefficient ω_3, with the vertical bars indicating the minimum and maximum bias across v and SNR settings. The results are displayed by p, π, and η_β settings. In all settings, the bias was lower for first continuous and binary coefficients ω_2 and ω_3, respectively, compared to intercept ω_1. Generally, bias was more pronounced overall in the ultra high-dimensional setting ($p = 75^2$) with more true signals among the available predictors ($\pi = 5\%$). The standard deviation for intercept ω_1 and first binary coefficient ω_3 was higher than for continuous coefficient ω_2 in all settings.

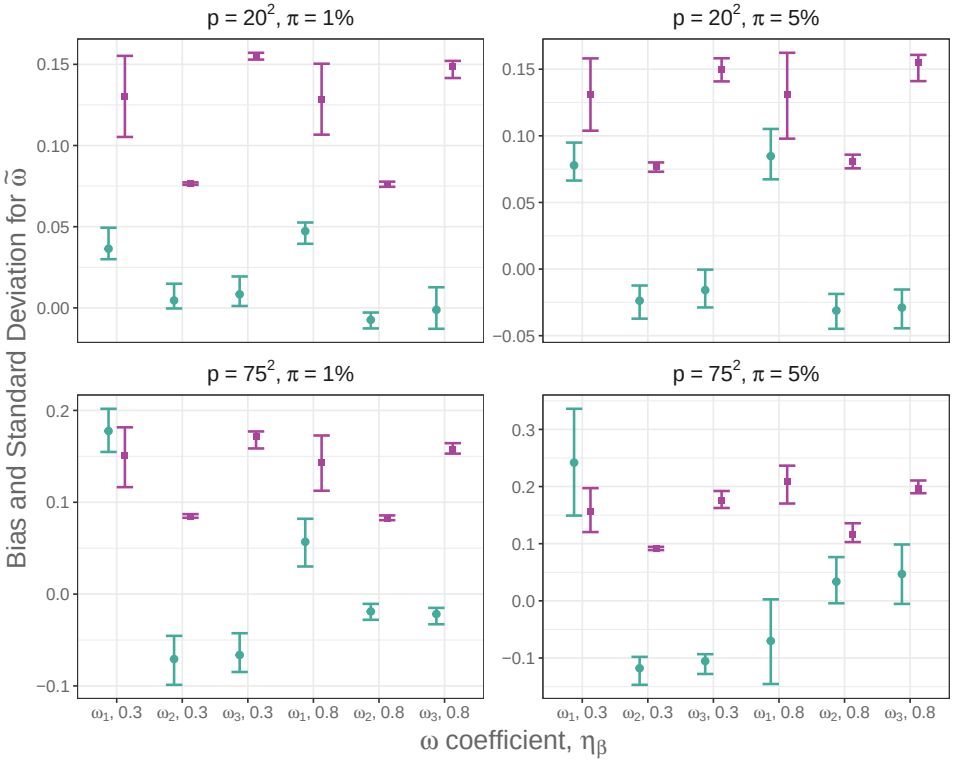

Fig. B.4 This figure show the average bias (green circle) and standard deviation (purple squares) of the $\tilde{\omega}$ estimates from H-PROBE simulations. Vertical lines represent the minimum and maximum bias and standard deviations averaged across simulation settings displayed.

To investigate performance in simulation settings closer to our application, we focused on settings where $n = 200$ and also varied the degree of correlation between the (binary) predictors we generated, with $\Sigma_{cor} = (10, 20, 30)$ where a higher number indicates more correlation. Figure B.5 mirrors Figure 3 from the main text, and shows Log Root Mean Squared Errors (RMSE) when $n = 200$ across various levels of predictor correlation. The results are highly similar to those when $n = 400$ and $\Sigma_{cor} = 20$. In nearly all simulation settings, H-PROBE had the lowest RMSE and this finding varied little whether predictors were more or less correlated.

B.2 Prediction interval performance

Figure B.6 mirrors Figure 4 of the main text and shows ECPs of PI for H-PROBE, PROBE, and Conformal Inference. Settings with $p = 75^2$ and $\pi = 5\%$ are not shown for PROBE and H-PROBE as they fall outside of the sparsity assumptions (more signals than observations) of these methods and inference may therefore not be performed accurately. Generally, the same ECP and PI trends are observed with $n = 200$ as with $n = 400$ (main text) and patterns were similar across $\Sigma_{cor} = (10, 20, 30)$. In all

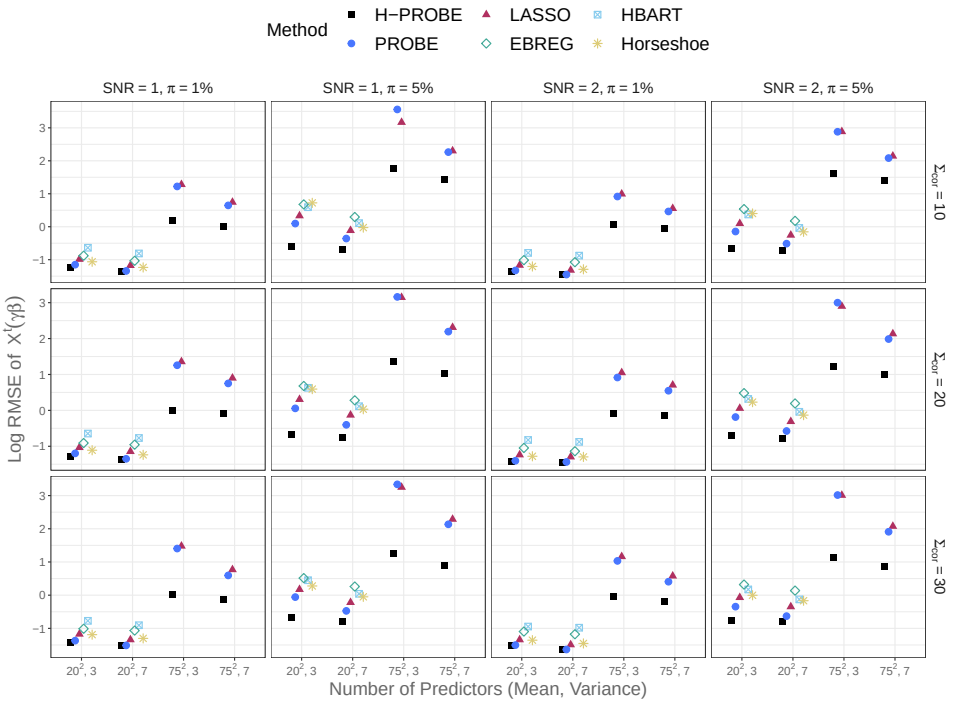

Fig. B.5 This figure shows the model performance results from the numerical studies, for selected settings when the number of observations is closer to our real-world application, $n = 200$, and where the degree of dependence between the predictors varies ($\Sigma_{cor} = (10, 20, 30)$). The figure presents the log Root Mean Squared Errors (RMSE) of $\boldsymbol{X}'(\gamma\boldsymbol{\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set) for six methods. The methods we compare are: H-PROBE (black squares), PROBE (blue circles), LASSO (maroon triangles), EBREG (green diamonds), HBART (blue squares with inner cross), or a Bayesian model with a horseshoe prior (yellow stars).

Σ_{cor} settings, H-PROBE had ECPs centered at the nominal level 0.95 despite having much shorter PI lengths than PROBE and Conformal Inference. As shown in Figure B.7, PROBE and Conformal Inference have substantially larger PI lengths and therefore can easily reach ECPs greater than 0.95. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$ were very similar and are omitted.

B.3 Variable selection performance

To evaluate the variable selection performance of the methods under $n = 200$ and $\Sigma_{cor} = (10, 20, 30)$, we display the True Positive Rate (TPR) and False Discovery Rate in Figures B.8 and B.9, respectively. As in Figure 5 of the main text, H-PROBE has the highest TPR in all settings displayed (Figure B.8). H-PROBE has a lower FDR than LASSO in all settings, while PROBE has the lowest FDR (Figure B.9).

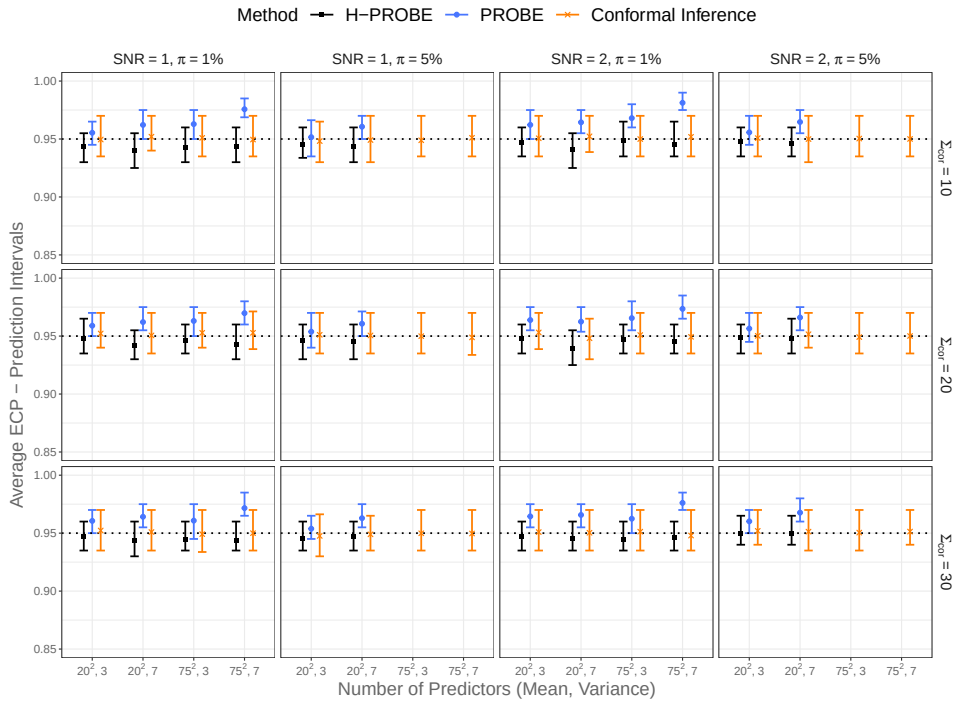

Fig. B.6 This figure shows the results related to Prediction Intervals (PIs) and Empirical Coverage Probabilities (ECP) from the numerical studies, under new settings $n = 200$ and $(\Sigma_{cor} = (10, 20, 30))$. The ECPs are defined as the proportion of PIs that contained the value $Y_{i,test}$ from the test set. The methods compared in this figure are H-PROBE (black squares), PROBE (blue circles), and Conformal Inference (orange crosses). Vertical lines represent the first and third quartiles of the distributions of ECPs for PIs. Settings with $p = 75^2$ and $\pi = 5\%$ are not shown for PROBE and H-PROBE as they fall outside of the sparsity assumptions.

B.4 Sensitivity to initialization

To evaluate the sensitivity of H-PROBE to the algorithm's initial values, we examined the impact of three initialization schemes on model performance and PI ECP for various settings. The first initialization scheme, Initialization A, is $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and $\boldsymbol{p}^{(0)} = \mathbf{0}$, which gives $\boldsymbol{W}_k^{(0)} = \mathbf{0}$ and $\boldsymbol{W}_k^{2(0)} = \mathbf{0}$ for all k as described in Section 2.3 of the main text. The second initialization scheme, Initialization B, uses the results of the PROBE algorithm to give starting values for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{p}^{(0)}$, $\boldsymbol{W}_k^{(0)}$, and $\boldsymbol{W}_k^{2(0)}$ for all k (McLain et al., 2025). Finally, the third initialization scheme, Initialization C, draws $\boldsymbol{\beta}^{(0)}$ values from a standard normal distribution, sets $\boldsymbol{p}^{(0)}$ equal to the absolute value of $\boldsymbol{\beta}^{(0)}$ truncating at 0 and 1, in turn giving values for $\boldsymbol{W}_k^{(0)}$ and $\boldsymbol{W}_k^{2(0)}$.

Figure B.10 shows simulation results for H-PROBE from settings initialized using each of the three initialization schemes. For brevity, we focus on the results for $\eta_\beta = 0.8$, $p = 20^2$, $\Sigma_{cor} = 20$, and binary $\boldsymbol{X}$ predictors. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$ were very similar and are omitted. Both the model performance (RMSE of $\boldsymbol{X}'(\boldsymbol{\gamma}\boldsymbol{\beta})$) and predictive inference performance (ECP of PIs) remain mostly unaffected by the

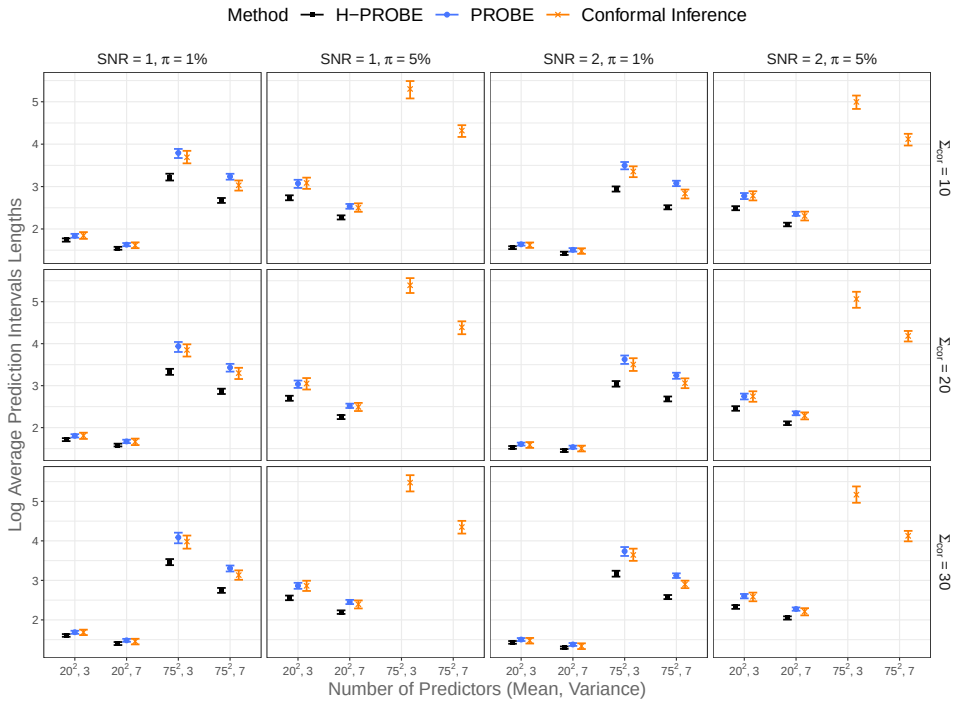

Fig. B.7 This figure shows the results related to Log Average Prediction Intervals (PI) lengths from the numerical studies, under new settings $n = 200$ and $(\Sigma_{cor} = (10, 20, 30))$. The methods compared in this figure are H-PROBE (black squares), PROBE (blue circles), and Conformal Inference (orange crosses). Vertical lines represent the first and third quartiles of the distributions of PI lengths. Settings with $p = 75^2$ and $\pi = 5\%$ are not shown for PROBE and H-PROBE as they fall outside of the sparsity assumptions.

initialization scheme, showing very similar average RMSEs and ECPs and interquartile ranges within simulation settings.

B.5 Robustness to misspecification

To evaluate the robustness of H-PROBE to misspecifications in the model on the variance, additional simulation studies were run. These studies examine the situation where additional (unnecessary) variables are included in $\boldsymbol{V}$ since the situation where important variables are missing from $\boldsymbol{V}$ can be inferred from the PROBE results (where there is no $\boldsymbol{V}$). In the simulations where the variance model was correctly specified, we generated $\boldsymbol{V}$ to include an intercept, along with an equal number of standard normal and Bernoulli(0.5) predictor variables, and modeled $\boldsymbol{V}$ accordingly. In the misspecified H-PROBE variance models, we modeled $\boldsymbol{V}$ with a superfluous squared standard normal predictor.

Figure B.11 shows simulation results for correctly and incorrectly specified H-PROBE variance models. For brevity, we focus on the results for $\eta_\beta = 0.8$, $p = 20^2$, $\Sigma_{cor} = 20$, and binary $\boldsymbol{X}$ predictors. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$

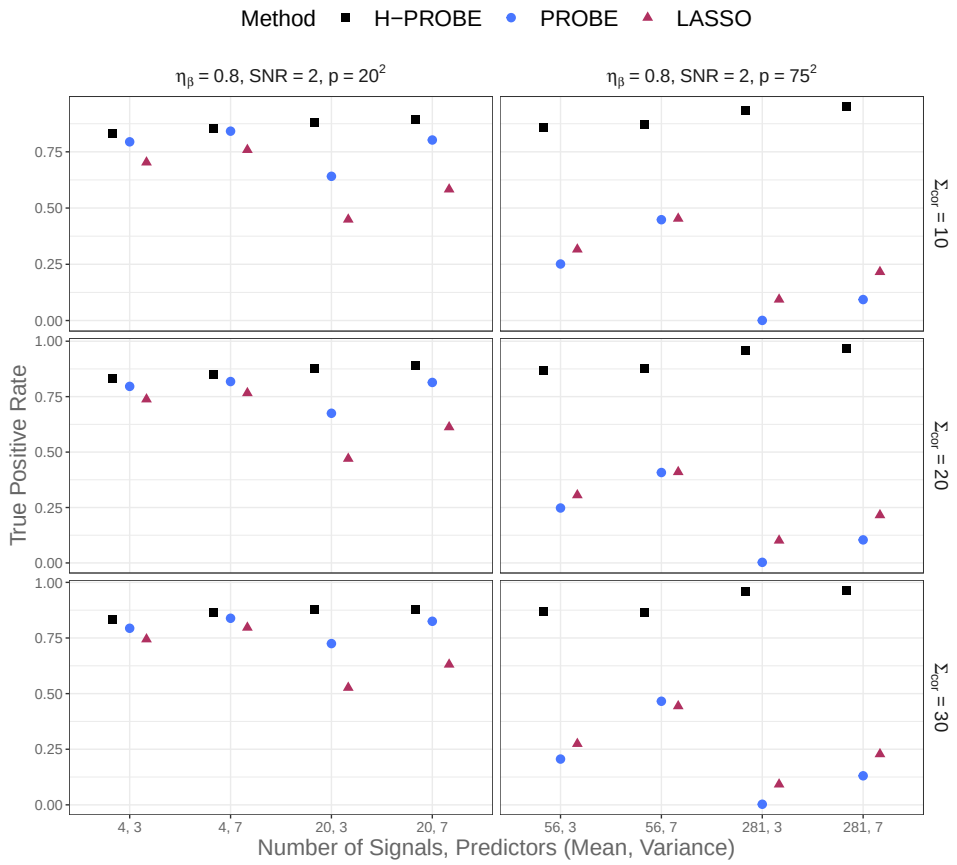

Fig. B.8 This figure shows True Positive Rate (TPR) results from the numerical studies, under new settings $n = 200$ and $(\Sigma_{cor} = (10, 20, 30))$. The TPR is displayed as a function of true signals ($|\boldsymbol{\gamma}| = p\pi$) and the number of predictors on the variance. H-PROBE is represented by black squares, PROBE by blue circles, and LASSO by maroon triangles.

were very similar and are omitted. The Log RMSE of $\boldsymbol{X}'(\boldsymbol{\gamma}\boldsymbol{\beta})$ values were slightly elevated for the misspecified model, with some overlap in their ranges (Figure B.11 Panel A). Misspecification in the variance model again did not severely impact ECPs, which were all close to the nominal 0.95 level, and always included 0.95 in their range (Figure B.11 Panel B). The misspecified model led to larger interquartile ranges for the ECPs. However, adding superfluous variables to the variance model does not appear to hamper the overall error rate of the PIs.

B.6 Comparison with traditional heteroscedastic approaches in $p \ll n$ settings

To compare H-PROBE to established methods for heteroscedastic linear regression, we conduct simulations in the $p \ll n$ setting. The simulation settings are as described

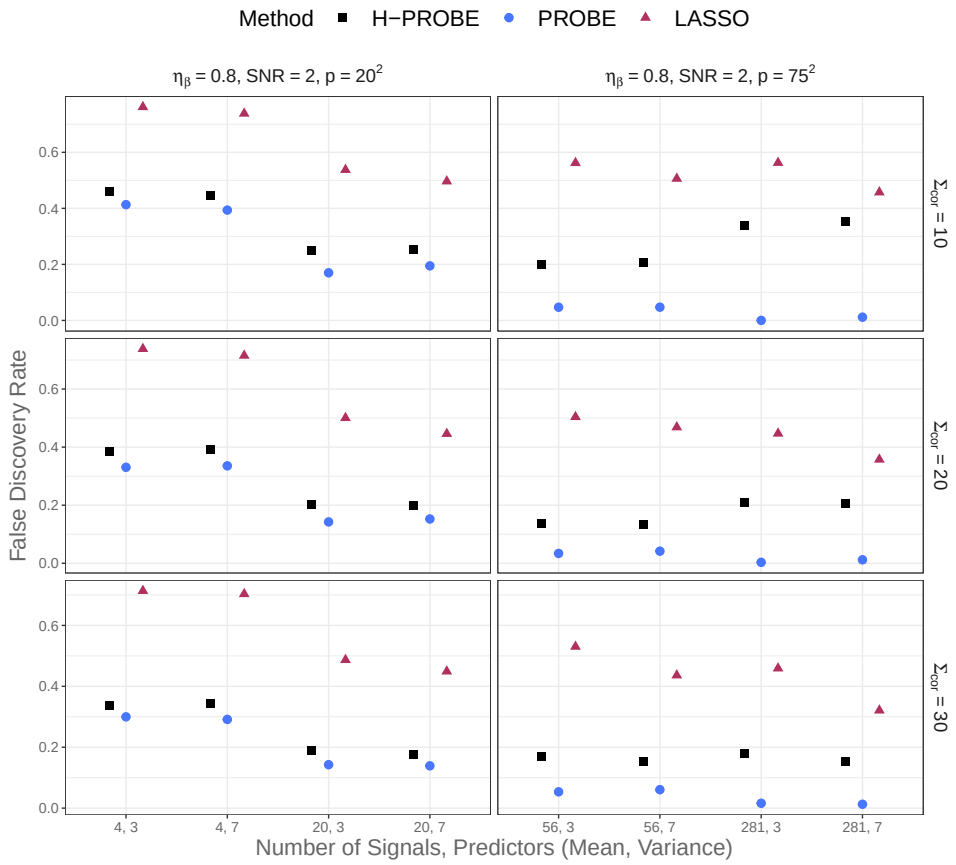

Fig. B.9 This figure shows False Discovery Rate (FDR) results from the numerical studies, under new settings $n = 200$ and $(\Sigma_{cor} = (10, 20, 30))$. The FDR is displayed as a function of true signals ($|\gamma| = p\pi$) and the number of predictors on the variance. H-PROBE is represented by black squares, PROBE by blue circles, and LASSO by maroon triangles.

in Section 3 of the main text, with the following differences: the number of predictors in the model is $p = 4^2$, and the degree of dependence between the predictors is $\Sigma_{cor} = 10$. Since $\pi = 1\%$ and $\pi = 5\%$ yield less than one signal, we instead use $\pi = 6\%$ and $\pi = 12\%$, which give one and two signals, respectively. In addition to the H-PROBE, PROBE, LASSO, EBREG, HBART, and Horseshoe methods, we incorporate linear regression with ordinary least squares (OLS), OLS with White standard errors (OLSW, Eicker, 1967; Huber, 1967; White, 1980), as well as Smyth (2002)'s heteroscedastic regression with non-sparse models on the mean and the variance using residual maximum likelihood (REML). Using lower values for p and Σ_{cor} and ($p = 4^2$, $\Sigma_{cor} = 10$) facilitated the matrix operations required in the REML approach. Higher values led to less-than-full rank matrices due to the nature of the application for which we developed H-PROBE, where predictors are correlated.

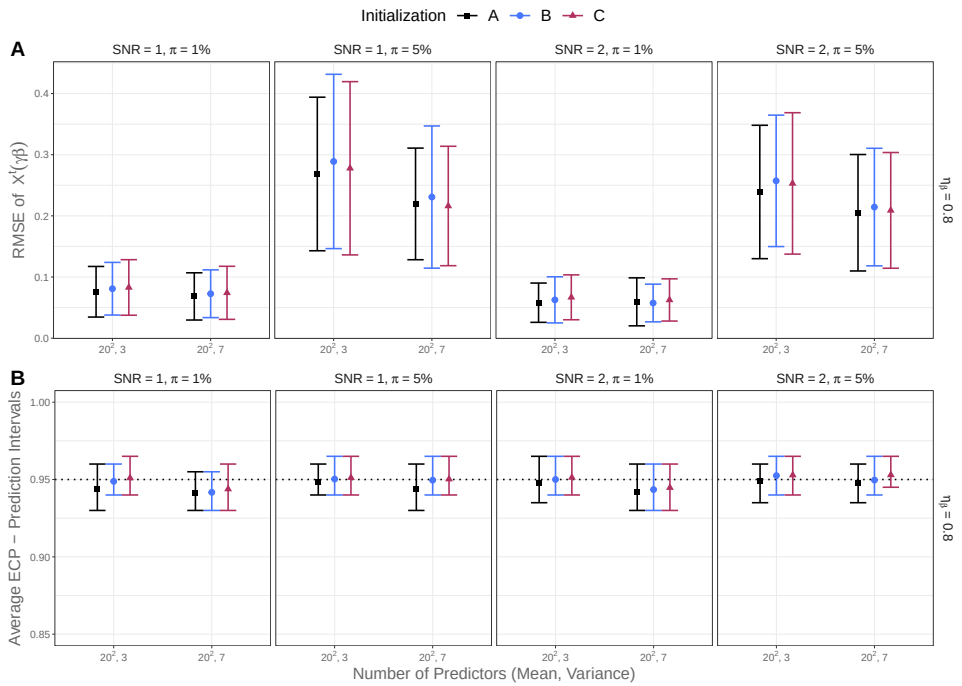

Fig. B.10 (A) Model performance and (B) Prediction Interval (PI) results from numerical studies using different initial values for H-PROBE only. Panel (A) shows the log Root Mean Squared Errors (RMSE) of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set). Panel (B) shows empirical coverage probabilities (ECPs) of PIs. The three initialization schemes are represented by the color legend. Vertical lines represent the first and third quartiles of the distributions of RMSE and ECPs of PIs.

Figure B.12 shows the model and predictive inference performance of the nine methods in low-dimensional settings. For brevity, we focus on the results for $\eta_\beta = 0.8$ and binary $\boldsymbol{X}$ predictors. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$ were very similar and are omitted. Figure B.12 Panel A shows that traditional approaches for heteroscedasticity, such as OLS, OLSW, and REML, had a higher Log RMSE of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$ than all other approaches. This is likely because they do not perform variable selection or penalization. H-PROBE either had the lowest Log RMSE or was among the methods with lower Log RMSEs. In Figure B.12 Panel B, OLS and OLSW had the lowest ECPs for PIs, as expected, while the other traditional heteroscedastic approach, REML, had larger ECPs than all other methods, exceeding the 95% nominal level. This is further evidenced by Figure B.12 Panel C, where REML has by far the largest average PI length of all methods.

C Additional Data Analysis Results

This Section provides additional results and information regarding the AQ application and analysis. To ensure that the non-linear relationship between the error variance σ_i^2 and total brain damage could not be remedied by a transformation of

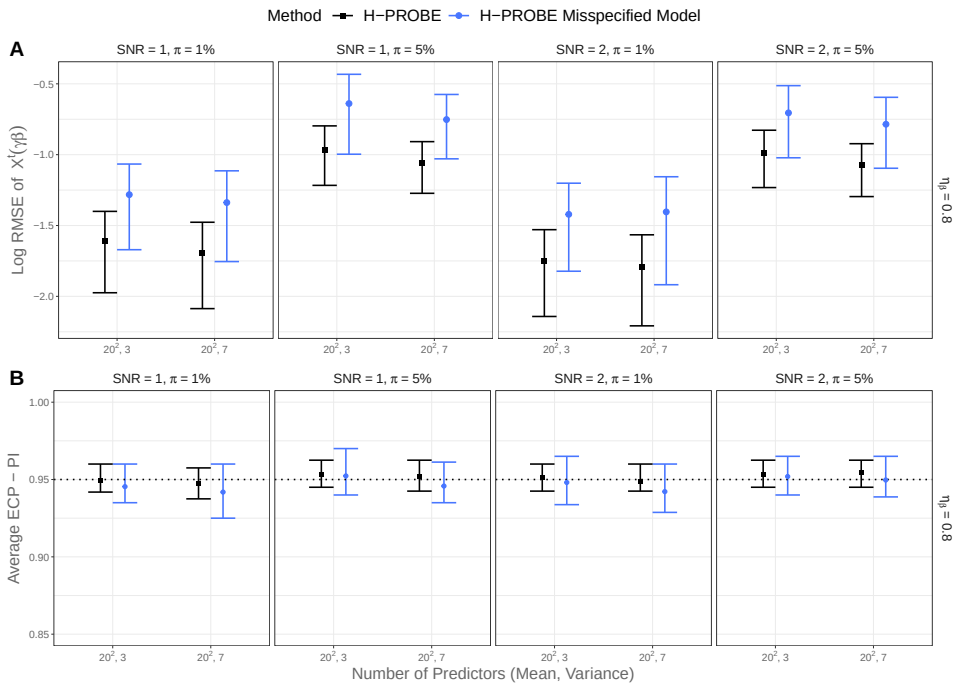

Fig. B.11 (A) Model performance and (B) Prediction Interval (PI) results from numerical studies using correctly specified (black squares) and misspecified (blue circles) variance models in the H-PROBE method. Panel (A) shows the log Root Mean Squared Errors (RMSE) of $\boldsymbol{X}'(\boldsymbol{\gamma}\boldsymbol{\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set). Panel (B) shows empirical coverage probabilities (ECPs) of PIs. Vertical lines represent the first and third quartiles of the distributions of RMSE and ECPs of PIs.

the Aphasia Quotient (AQ) outcome, we performed additional analysis examining transformations of AQ. We examined multiple transformations and provide figures for two of the transformations. We applied log and square root inverse transformations to AQ, $AQ_{log-inv} = log\left(\frac{100-AQ}{100}\right)$ and $AQ_{sqrt-inv} = \sqrt{\left(\frac{100-AQ}{100}\right)}$, and modeled $AQ_{log-inv}, AQ_{sqrt-inv}$ using the PROBE method, a homoscedastic approach. Figure C.13 shows that despite of both transformations, the non-linearity of the relationship between the error variance σ_i^2 and total brain damage remains. This result remained in other transformations we examined. The dashed blue line in Figure C.13.a represents the form we assumed for the variance model in our data analysis (Equation (6) of the main text). The dashed blue line effectively approximates the LOESS line, indicating that our choice of predictors in the H-PROBE model on the variance performs well. Figure C.14 shows that H-PROBE provides predictions $\hat{Y}_i$ that have a different range for distinct values of estimated $\tilde{\sigma}_i^2$. This important characteristic of heteroscedastic data is not detected by PROBE. Figure C.15 mirrors Figure 5 from the main text and shows PI lengths for each patient and each method by cross-validation fold, including the REML method. The Conformal Split and EBREG methods had similar PI lengths,

while REML had large PI lengths overall. H-PROBE displayed the widest range of PI lengths, reflected by the differing estimated $\tilde{\sigma}_i^2$ by subject from Figure C.14.

For both PROBE and H-PROBE, the voxels with large coefficients appear to be located in and around the Inferior Frontal Gyrus, which contains Broca's region, a vital area of the brain for speech production. The pathophysiology of strokes results in detrimental effects on cerebral structures and functions. As a result, negative $\boldsymbol{\beta}$ estimates, which suggest a protective or beneficial impact of strokes on the brain, contradicts established neuroscientific understanding. As shown in Figure C.16, most of the selected $\boldsymbol{\beta}$ estimates for H-PROBE and PROBE are positive. The LASSO model, which was markedly more sparse, resulted in only negative $\boldsymbol{\beta}$ estimates. Among voxels with larger $\tilde{p}_k$ ($\tilde{p}_k > 0.05$) all coefficients estimated by H-PROBE were positive, versus all negative for LASSO. Figure C.17 provides per-voxel statistical brain maps for the LASSO method. Since stroke injury is constrained by vasculature, and the presence of a brain injury is an inclusion criteria of the study data, negative $\boldsymbol{\beta}$ estimates for a given voxel indicate that the core brain modules related to speech have been spared and are omitted from Figure C.17. The LASSO model overwhelmingly resulted in negative $\boldsymbol{\beta}$ estimates, and only one voxel appears on the brain maps, in the fourth brain slide from the left.

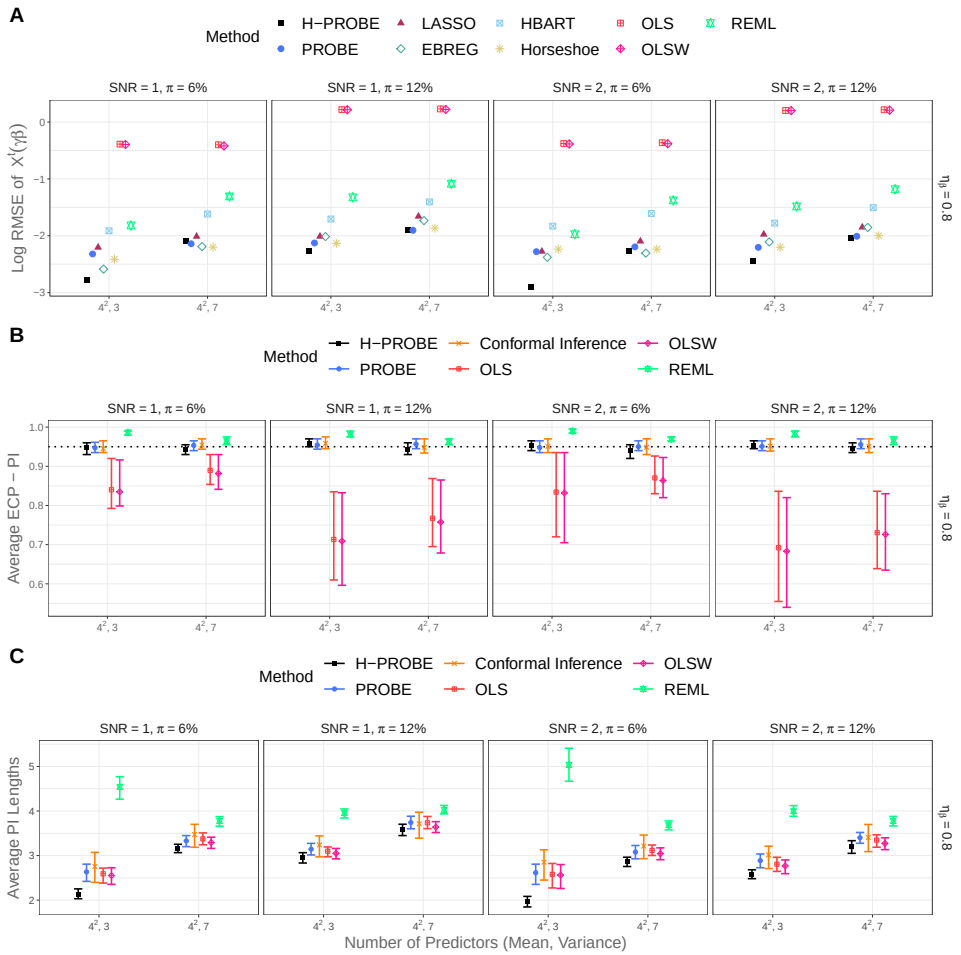

Fig. B.12 (A) Model performance and (B)–(C) prediction interval results from numerical studies using new low-dimensional simulation settings. The variable selection methods we compare are: H-PROBE (black squares), PROBE (blue circles), LASSO (maroon triangles), EBREG (green diamonds), HBART (blue squares with inner cross), Bayesian model with a horseshoe prior (yellow stars). The methods without variable selection are: OLS (red squares with inner cross), OLSW (pink diamonds with inner cross), and REML (green stars). Panel (A) shows the log Root Mean Squared Errors (RMSE) of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set) for the nine methods compared. Panel (B) shows empirical coverage probabilities (ECPs) of Prediction Intervals (PIs). The ECPs are defined as the proportion of PIs that contained the value $Y_{i,test}$ from the test set. Vertical lines represent the first and third quartiles of the distributions of ECPs for PIs and of PI lengths.

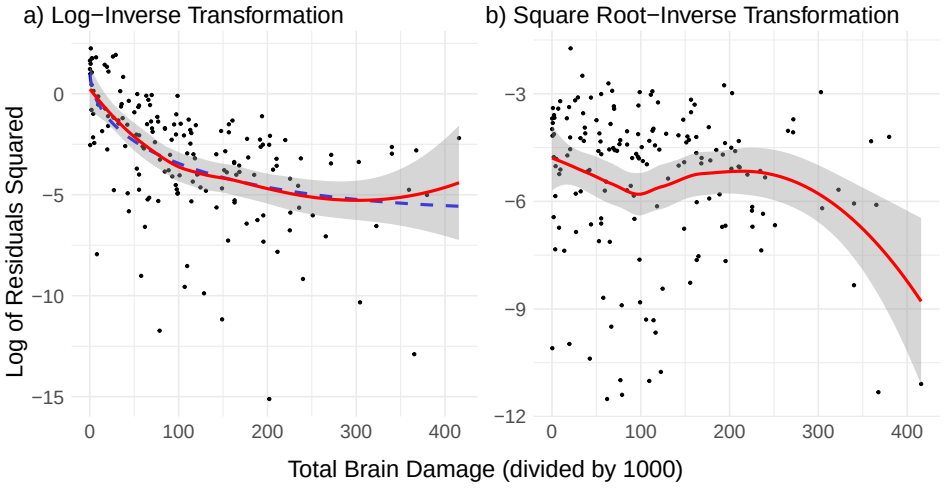

Fig. C.13 This figure shows data from our real-world application. In this data, the Aphasia Quotient (AQ) outcome is related to the total brain damage (TBD) covariate (defined as the number of brain voxels with lesions). Panels (a) and (b) show the log of squared residuals from homoscedastic high-dimensional linear regression performed via the PROBE method, using brain image data as predictors and AQ as the outcome. In Panel (a), a log-inverse transformation was used, $AQ_{log-inv} = \log AQ_{inv}$, while Panel (b) used a square-root-inverse transformation, $AQ_{sqrt-inv} = \sqrt{AQ_{inv}}$ where $AQ_{inv} = (100 - AQ)/100$. The red lines in Panels (a)–(b) represent a locally estimated scatterplot smoothing (LOESS) fit, along with its standard error in grey shading. The dashed blue line in Panel (a) represents the form we assumed for the variance model in our data analysis (Equation (6) of the main text).

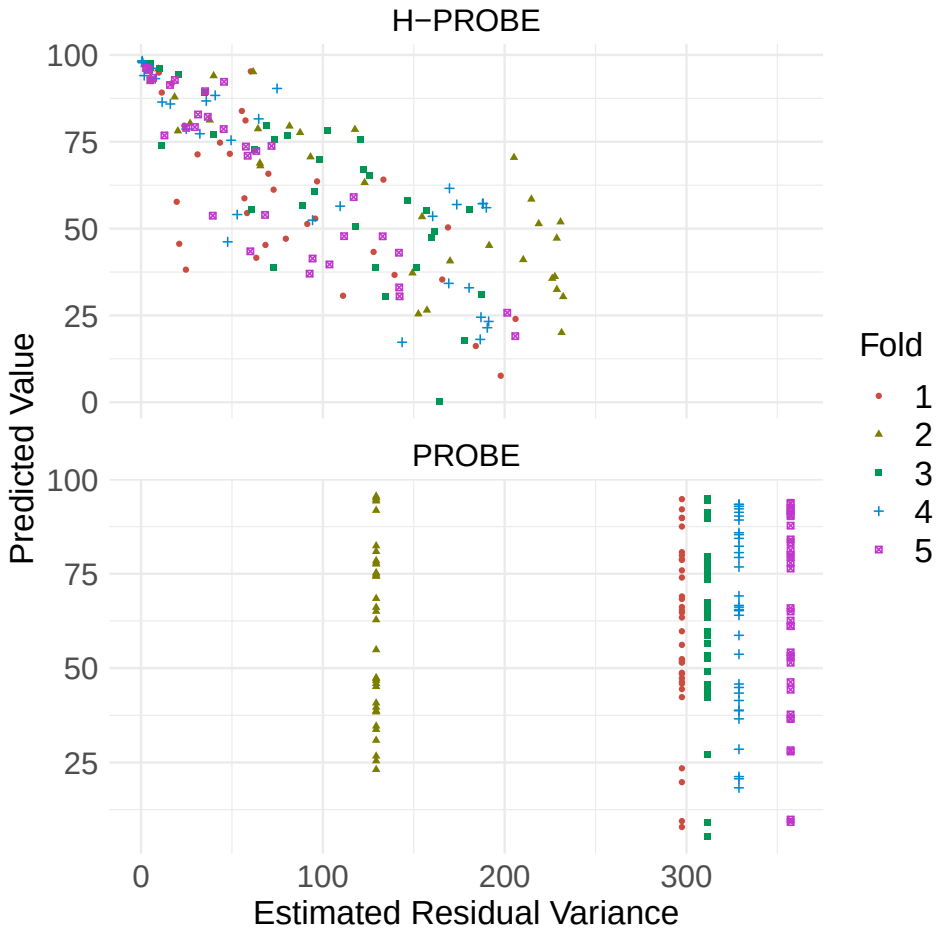

Fig. C.14 This figure presents predicted values of the Aphasia Quotient (AQ) for each patient i, plotted as a function of each patient's estimated error $\tilde{\sigma}_i^2$ using the H-PROBE and PROBE methods. The cross-validation fold for each patient is indicated in the color legend. Note that within a given fold, $\tilde{\sigma}_i^2$ estimates for observation i are the same when using PROBE.

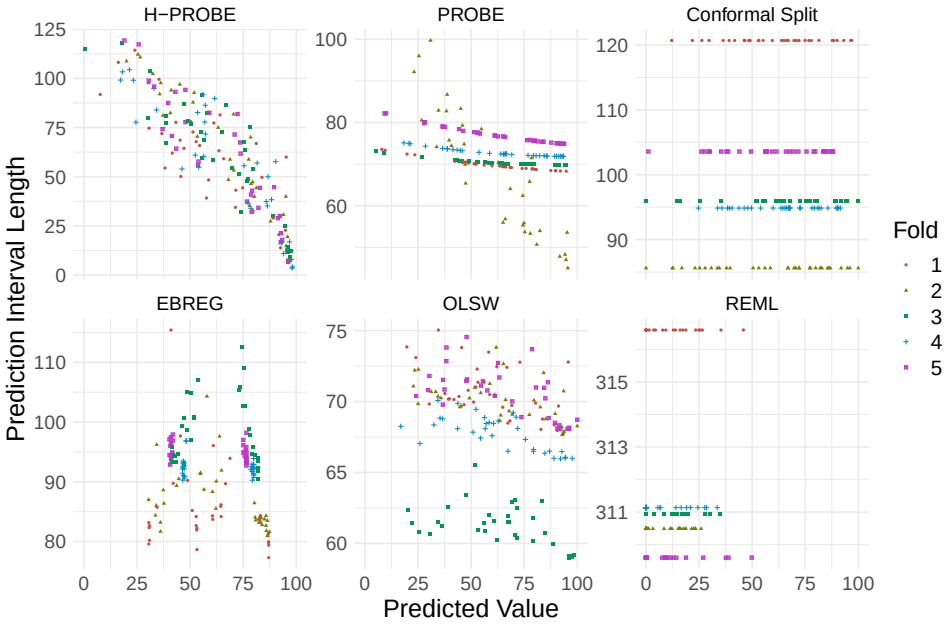

Fig. C.15 This figure presents Prediction Interval (PI) lengths as a function of each patient's predicted Aphasia Quotient (AQ). H-PROBE is compared to three high-dimensional but homoscedastic methods, PROBE, Conformal Inference, and EBREG, as well as two low-dimensional but heteroscedastic methods, OLSW and REML. The color legend represents the cross-validation fold in which predictions were obtained.

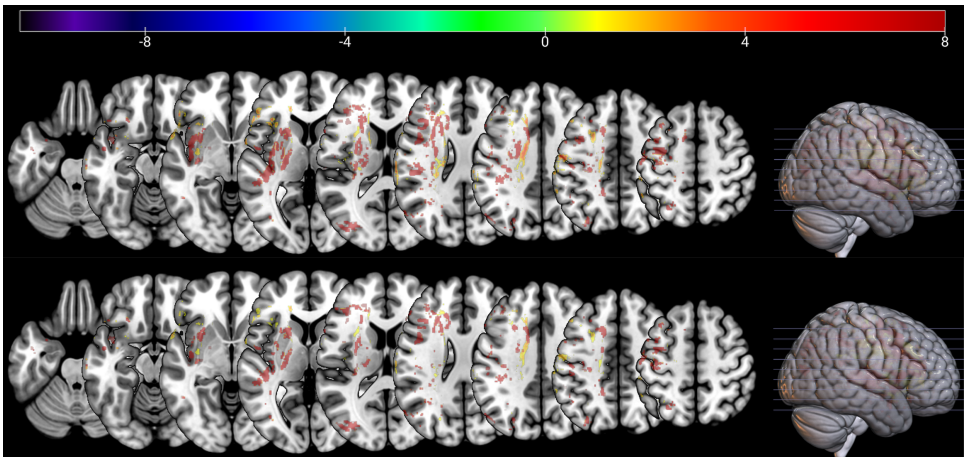

Fig. C.16 Brain maps showing position, direction, and magnitude of voxel-specific $\boldsymbol{\beta}$ coefficients across different slides for PROBE (top) and H-PROBE (bottom). The color legend represents the magnitude of $\boldsymbol{\beta}$ estimates where $|\boldsymbol{\beta}| < 0.1$ are omitted.

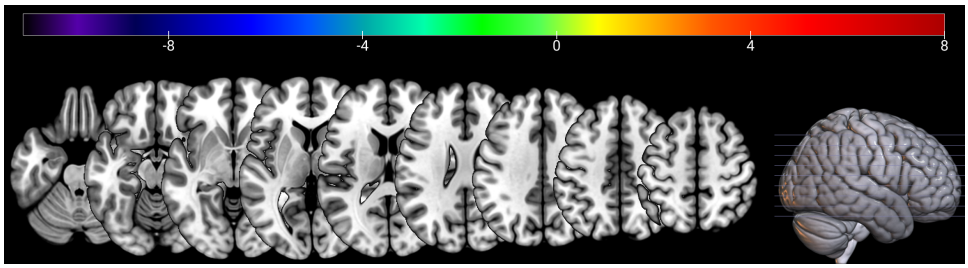

Fig. C.17 Brain maps showing position, direction, and magnitude of voxel-specific $\boldsymbol{\beta}$ coefficients across different slides for the LASSO model. The color legend represents the magnitude of $\boldsymbol{\beta}$ estimates.

References

Blanchard, G., & Roquain, E. (2009). Adaptive fdr control under independence and dependence. *Journal of Machine Learning Research*, *10*, 2837 –2831,

Castillo, I., & Roquain, É. (2020, 10). On spike and slab empirical Bayes multiple testing. *The Annals of Statistics*, *48*(5), 2548–2574, https://doi.org/10.1214/19-AOS1897

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–38, (With discussion)

Efron, B. (2008). Microarrays, empirical bayes and the two-group model. *Statistical Science*, *23*(1), 1–22,

Eicker, F. (1967). Limit Theorems for Regression with Unequal and Dependent Errors. *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 59–82).

Fletcher, R. (1987). *Practical methods of optimization.* New York: John Wiley & Sons.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 221–233).

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423,

Liu, C., Rubin, D.B., Wu, Y.N. (1998, 12). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85*(4), 755-770, https://doi.org/10.1093/biomet/85.4.755 Retrieved from https://doi.org/10.1093/biomet/85.4.755

McLain, A.C., Zgodic, A., Bondell, H. (2025). Sparse high-dimensional linear regression with a partitioned empirical bayes ECM algorithm. *Computational Statistics and Data Analysis*, *207*, 108146,

Meng, X.L., & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278, https://doi.org/10.1093/biomet/80.2.267

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. *Proceedings of the eighteenth conference on uncertainty in artificial intelligence* (pp. 352–359).

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London. Retrieved from https://doi.org/10.1007/978-1-4899-3324-9

Smyth, G. (2002). An efficient algorithm for reml in heteroscedastic regression. *Journal of Computational and Graphical Statistics*, *11*(4), 836–847,

Storey, J.D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *69*(3), 347–368, https://doi.org/10.1111/j.1467-9868.2007.005592.x Retrieved from http://dx.doi.org/10.1111/j.1467-9868.2007.005592.x

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., ... Robert, C.P. (2020). Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data. *Journal of Machine Learning Research*, *21*(17), 1–53, Retrieved 2021-05-20, from http://jmlr.org/papers/v21/18-817.html

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838,