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, George Mason University, 4400 University Drive, MS 4A7, Fairfax, Virginia, 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;

Abstract

Sparse linear regression methods for high-dimensional data commonly assume that errors have constant variance, which can be violated in practice. For example, Aphasia Quotient (AQ) is a critical measure of language impairment and informs treatment decisions, but it is challenging to measure in stroke patients. It is of interest to use high-resolution T2 neuroimages of brain damage to predict AQ. However, sparse regression models show marked evidence of heteroscedastic error even after transformations are applied. This violation of the homoscedasticity assumption can lead to biased and inconsistent standard errors of estimated coefficients and prediction intervals (PI) with improper length. Bayesian heteroscedastic linear regression models relax the homoscedastic error assumption but can enforce restrictive prior assumptions on parameters, and many are computationally infeasible in the high-dimensional setting. This paper proposes estimating high-dimensional heteroscedastic linear regression models using a heteroscedastic partitioned empirical Bayes Expectation Conditional Maximization (H-PROBE) algorithm. H-PROBE is a computationally efficient maximum *a posteriori* estimation approach that requires minimal prior assumptions and can incorporate covariates known or hypothesized to impact heterogeneity. We apply

this method by using high-dimensional neuroimages to predict and provide PIs for AQ that accurately quantify predictive uncertainty. Our analysis demonstrates that H-PROBE can provide narrower PI widths than standard methods without sacrificing coverage. Narrower PIs are clinically important for determining the risk of moderate to severe aphasia. Additionally, through extensive simulation studies, we exhibit that H-PROBE results in superior prediction, variable selection, and predictive inference compared to alternative methods.

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

1 Introduction

Much of the current literature on sparse linear regression methods for high-dimensional data assumes that the errors have a constant variance. However, in practice, this assumption is often violated. A clinical application where high-dimensional heteroscedastic data arises is in treatment decisions for neurological disorders based on patient imaging data. Johnson et al. (2019) present results from a study on language rehabilitation in patients who experienced a left-hemispheric stroke and suffer from aphasia – a language disorder impacting speech. The outcome of interest is the subjects' Aphasia Quotient (AQ), a score quantifying language impairment vital to understanding patients' treatment options (Risser & Spreen, 1985). However, collecting AQ is a cumbersome task, particularly for patients who have recently had a stroke (Odekar & Hallowell, 2005). Consequently, it is of interest to develop models that can *predict* subjects' unknown AQ based on images of their brains (Lee, Ko, Park, & Kim, 2021).

While several studies have proposed methods to predict aphasia severity (Lee et al., 2021; Teghipco, Newman-Norlund, Fridriksson, Rorden, & Bonilha, 2023; Yourganov, Smith, Fridriksson, & Rorden, 2015), none provide *prediction intervals* (PIs) for AQ predictions. Effective decision-making in healthcare relies crucially on combining predictive models with uncertainty analyses (Begoli, Bhattacharya, & Kusnezov, 2019; Zou et al., 2023). For example, according to the Western Aphasia Battery, the severity of aphasia can be classified as follows: AQ of less than 26 is very severe, 26–50 is severe, 51–75 is moderate, and above 75 is mild (Kertesz, 2007). If a patient's PI for AQ spans several categories of aphasia severity, then a clinician could take this predictive uncertainty into account and order additional diagnostic tests or consider other factors in the patient's medical history (Zou et al., 2023). Since the PIs can better guide clinicians in defining patient treatment courses, we develop a PI-based approach to quantify the predictive uncertainty of AQ predictions.

In our motivating application, data is cross-sectional and consists of brain images obtained through T1 structural Magnetic Resonance Imaging (MRI), giving the lesion status (i.e., damaged or not damaged from stroke) of more than 5×10^6 three-dimensional ($1\ mm^3$) brain voxels. As displayed in Figure 1A, patients with little brain damage commonly score near the top of the 0–100 range, while those with more

substantial brain damage generally have lower AQ scores. However, there is considerable heterogeneity in this trend. When using a high-dimensional homoscedastic linear regression approach (PROBE, McLain, Zgodic, & Bondell, 2025), Figure 1B suggests that the total number of damaged voxels (TBD) has a positive relationship with the residual variance. As a result, the scope of our research concerns applications with a known low-dimensional set of variables that potentially predict heterogeneity. A common remedy to heteroscedastic residuals is to use a transformed version of the outcome. However, Figures 1C and 1D demonstrate that the relationship between residual error variance (from the PROBE model) and TBD appears to persist with log and square-root transformations.

This finding solidifies the need for a heteroscedastic approach to predicting patient AQ scores and providing accurate PIs to help clinical decision-making. If ignored, heterogeneity can harm multiple areas of analyses, including biased and inconsistent standard errors of estimated coefficients, as well as PIs with improper length (R.J. Carroll & Ruppert, 1988). These drawbacks can be particularly impactful in high-dimensional settings where the number of predictors is much larger than the sample size, and heterogeneity can lead to overfitting. From a clinical perspective, quantifying the uncertainty of predictive estimates from a machine learning model is essential because it allows practitioners to assess the reliability of individual predictions (Banerji, Chakraborti, Harbron, & MacArthur, 2023; Begoli et al., 2019). Crucially, in settings with marked differences in uncertainty among subjects, modeling heterogeneity enables individual-level risk assessment and supports informed decision-making under uncertainty.

1.1 High-dimensional heteroscedastic regression

In the classical low-dimensional heteroscedastic linear regression setting where the number of predictors p is much smaller than the sample size n, ordinary least squares (OLS) with White standard errors (White, 1980) can be used. Alternatively, restricted maximum likelihood (REML) approaches with models on both the mean and the variance (Smyth, 2002) can also be fitted to the data. When $p \ll n$, the coefficient estimates and standard errors under the OLS and REML approaches are unbiased and consistent (Eicker, 1967; Huber, 1967; White, 1980). However, in high-dimensional scenarios with $p \gg n$, regardless of heteroscedasticity, the model parameters are not identifiable without additional structure, and OLS and REML estimates are inherently unstable and non-unique due to the rank-deficient design matrix. To facilitate meaningful estimation in these high-dimensional scenarios, practitioners usually assume sparsity in the regression coefficients and use penalized regression approaches for both homoscedastic and heteroscedastic cases.

In heteroscedastic scenarios, penalized linear regression techniques have been expanded to down-weight outliers or anomalous observations with large error variances (Alfons, Croux, & Gelper, 2013; Curto, Pinto, Morais, & Lourenco, 2011; Rousseeuw & Van Driessen, 2006; Ziel, 2016) or use error criteria that are less sensitive to outliers (Belloni, Chernozhukov, & Wang, 2014; H. Wang, Li, & Jiang, 2007). However, these methods may not be suited for scenarios where *known* factors are hypothesized to be related to heterogeneity in the data, as in our AQ application, where TBD is

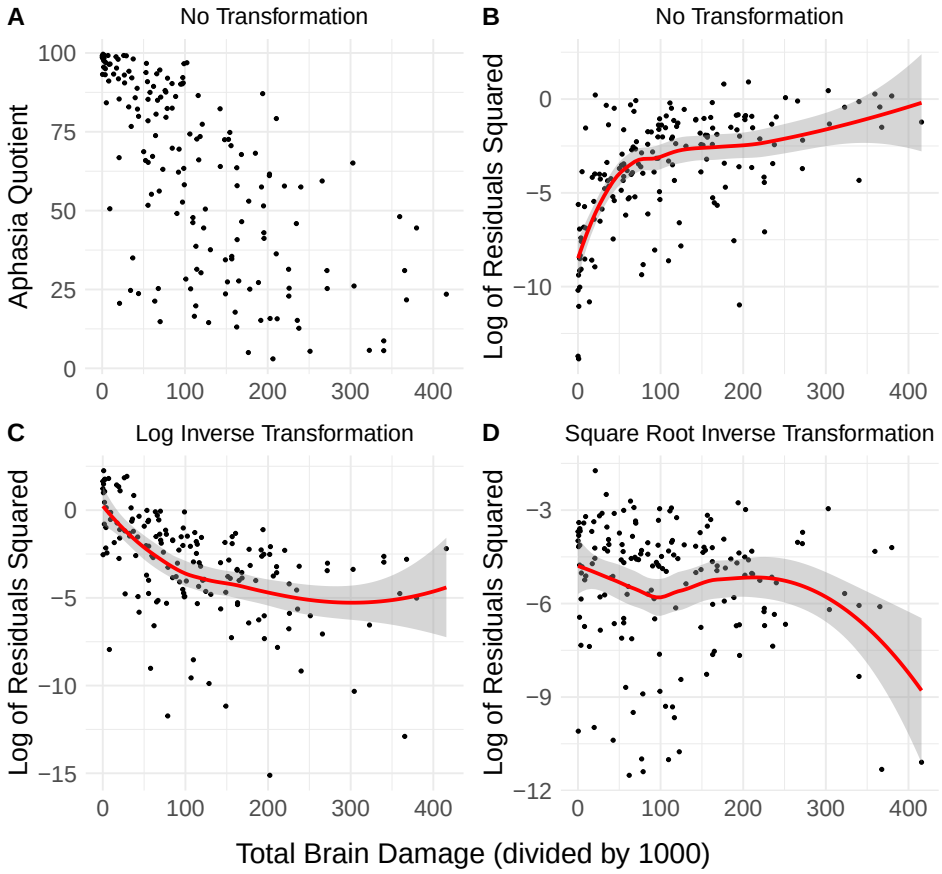

Fig. 1 This figure shows data from our real-world application. In Panel (A), the Aphasia Quotient (AQ) is plotted as a function of the total brain damage covariate (defined as the number of brain voxels with lesions). Panels (B), (C), and (D) show the log of squared residuals from homoscedastic high-dimensional linear regression performed via the PROBE method, using brain image data as predictors and AQ as the outcome. In Panel (B), PROBE modeled AQ without any transformation. In Panel (C), PROBE modeled a log-inverse transformation of AQ, $AQ_{log-inv} = \log AQ_{inv}$. In Panel (D), PROBE modeled a square-root-inverse transformation of AQ, $AQ_{sqrt-inv} = \sqrt{AQ_{inv}}$ where $AQ_{inv} = (100 - AQ)/100$. Red lines in Panels (B)–(D) represent a locally estimated scatterplot smoothing (LOESS) fit, along with its standard error in gray shading.

related to the residual variance. To address this gap, a second line of research focuses on modeling the variance of observations. We have located only five such proposals for the high-dimensional setting, four in the frequentist framework and one in the Bayesian framework. First, Daye, Chen, and Li (2012) proposed doubly regularized likelihood estimation with ℓ_1 penalty on parameters for both the mean and variance. Similarly, Chiou, Guo, and Ing (2020) as well as Peng, Chiou, Huang, and Ing (2025) use a greedy algorithm (Temlyakov, 2000) and backward elimination to select variables for the models on the mean and the variance. Third, L. Zhou and Zou (2021) leverage the conceptual framework of Daye et al. (2012) but use sample splitting to

select predictors via LASSO models on the mean and fit additional LASSO models to the residual variance. Finally, Pratola, Chipman, George, and McCulloch (2020) proposed heteroscedastic Bayesian additive regression trees (HBART), where, similarly to Daye et al. (2012) the same predictors are considered for the models on the mean and the variance.

While many of the above methods can sufficiently incorporate heteroscedasticity in the high-dimensional setting, very few have investigated creating PIs for future observations, especially in cross-sectional applications. In time series applications, PIs are commonly implemented in forecasting and planning, whereas in our high-dimensional cross-sectional application, PIs can vary from patient to patient and fewer approaches are available. As a result, estimating PIs is a key motivating factor for heteroscedastic regression models since the lengths of PIs may vary by patient factors available in the data. HBART can construct PIs for future observations that account for heteroscedasticity. However, HBART does not perform variable selection for the mean and, as illustrated in Section 3, is very computationally intensive to fit in the high-dimensional ($p \gg n$) setting. Conformal inference methods (Lei, G'Sell, Rinaldo, Tibshirani, & Wasserman, 2018; Tibshirani & Foygel, 2019; Vovk, Gammerman, & Shafer, 2005) can be used with penalized regression. However, their finite-sample coverage guarantees are marginal and may vary depending on certain predictor combinations. As demonstrated in our data analysis in Section 4, marginal properties are unsatisfactory for heterogeneous data and are particularly inefficient when researchers can well-hypothesize potential sources of heterogeneity.

Given these limitations, we propose a heteroscedastic high-dimensional linear regression model estimated with a Heteroscedastic PaRtitiOned empirical Bayes Expectation conditional maximization (H-PROBE) algorithm, for applications with a known low-dimensional set of residual variance predictor variables. We base H-PROBE on the previously established PROBE framework (McLain et al., 2025). PROBE is a computationally efficient maximum *a posteriori* (MAP) estimation approach based on a quasi Parameter-Expanded Expectation Conditional Maximization (PX-ECM) algorithm (Liu, Rubin, & Wu, 1998; Meng & Rubin, 1993). It requires minimal prior assumptions on the regression parameters through plug-in empirical Bayes estimates of hyperparameters in the E-step. The novelty of H-PROBE is that it expands PROBE by allowing for non-constant residual variance via incorporating covariates known or hypothesized to impact heterogeneity, assuming cross-sectional data and independence (no autocorrelation) between observations. Further, we propose methods to estimate prediction intervals for future observations.

We demonstrate the utility of incorporating heterogeneity when constructing PIs by analyzing simulated data and our study of AQ in patients with recent left-hemispheric stroke (Johnson et al., 2019). Our work makes both methodological and applied contributions. On the methodological side, to the best of our knowledge, H-PROBE is one of the first Bayesian variable selection approaches for high-dimensional data that includes both models on the mean and the variance. Further, we demonstrate that H-PROBE is much less computationally demanding than other Bayesian regression models fitted with Markov chain Monte Carlo (MCMC), such as HBART.

On the application side, we expand the literature on AQ prediction by outlining the first proposal that provides PIs and AQ predictions.

The remaining paper layout is as follows. We describe H-PROBE in Section 2 and numerical studies evaluating its performance in Section 3. Section 4 presents a comprehensive analysis of the AQ imaging study, and Section 5 concludes with a brief discussion. Supplementary Materials include more detailed technical content, additional simulations, and further data analysis results.

2 Methods

2.1 Model framework

In most linear regression models with outcome $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, predictors $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, and error term $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, it is assumed that $Var(\epsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$. As a result, the error term has the same variance for all observations (homoscedasticity). In a contrasting scenario, the error term may display a variance that differs from observation to observation. Then, the linear regression model for heteroscedastic data is written as

$$Y_i = \boldsymbol{X}_i\boldsymbol{\xi} + \epsilon_i, \tag{1}$$

where $\boldsymbol{\xi} \in \mathcal{R}^p$, $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma_i^2$. Let X_{ik} represent predictor k for observation i, with $n \times 1$ vector $\boldsymbol{X}_k = (X_{1k}, \ldots, X_{nk})$, $n \times p$ design matrix $\boldsymbol{X}$, and $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ as a diagonal $n \times n$ matrix. Assuming Gaussian errors and independence between observations, the distribution of the outcome is $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\xi}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix.

We leverage a Bayesian framework to accommodate a high-dimensional ($p \gg n$) setting and conduct sparse linear regression. Specifically, our model will allow for sparse and non-sparse predictors in the model on the mean, where the non-sparse predictors (e.g., an intercept) are denoted by $\boldsymbol{Z} \in \mathcal{R}^{n\times z}$. In practice, the model does not require non-sparse predictors $\boldsymbol{Z}$. The variance predictors are denoted by $\boldsymbol{V} \in \mathcal{R}^{n\times v}$. With this we rewrite the model in (1) as

$$\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\gamma} \circ \boldsymbol{\beta}) + \boldsymbol{Z}\boldsymbol{\varphi} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\gamma} \circ \boldsymbol{\beta}$ is a Hadamard product, $\boldsymbol{\gamma} \in \{0, 1\}^p$, and $\boldsymbol{\varphi} \in \mathcal{R}^z$. For brevity, let $\boldsymbol{\gamma\beta} \equiv \boldsymbol{\gamma} \circ \boldsymbol{\beta}$ for the remainder. 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 $\boldsymbol{\Sigma}$,

$$-\log\{diag(\boldsymbol{\Sigma})\} = \boldsymbol{V}\boldsymbol{\omega}, \tag{3}$$

where $\boldsymbol{\omega} \in \mathcal{R}^v$. The log transformation on the variances ensures positivity, can accommodate variances that vary over orders of magnitude, and has been long established in variance function modeling (R. Carroll, 1988; Cleveland, 1993). The complete Bayesian

framework includes the following prior information

$$\begin{aligned} p(\boldsymbol{\beta}) &= \prod_{k=1}^{p} f_\beta(\beta_k), \\ p(\boldsymbol{\gamma}|\pi) &= \pi^{p-|\gamma|}(1-\pi)^{|\gamma|}, \\ f_\beta(\beta_k) &\propto 1, \\ p(\boldsymbol{\varphi}) &\propto 1, \\ \pi &\sim \text{Uniform}(0,1), \\ \boldsymbol{\omega} &\sim \text{MLG}(\mathbf{0}, c^{1/2}\sigma_\omega^2 \boldsymbol{I}, c\mathbf{1}, c\mathbf{1}), \end{aligned}$$

where MLG denotes a Multivariate Log-Gamma distribution with $c, \sigma_\omega^2 > 0$ (Parker, Holan, & Wills, 2021), and $|\gamma| = \sum_k \gamma_k$. Throughout, we use $\sigma_\omega^{-1} = 10^{-5}$ and $c = 1000$ to yield a weakly informative prior on $\boldsymbol{\omega}$. The MLG distribution yields conjugate conditional posteriors in heteroscedastic linear models (Parker et al., 2021). MLG distributions are useful in contexts where both Gaussian posteriors and computational efficiency are desired. The MLG prior converges to a multivariate normal prior with mean $\mathbf{0}$ and variance $\sigma_\omega^2 \boldsymbol{I}$ as the value of c approaches infinity (Bradley, Holan, & Wikle, 2020; Parker et al., 2021).

Our proposed H-PROBE model differs from the PROBE method of McLain et al. (2025) since H-PROBE includes a model on the variance as shown in Equation (3). In the present work, we specifically leverage the MLG distribution and its conjugate properties to allow for heteroscedastic errors. In contrast, PROBE assumes homoscedastic errors and does not include variance predictors $\boldsymbol{V}$ or coefficients $\boldsymbol{\omega}$. The estimation procedure described in Section 2.2 also differs between H-PROBE and PROBE. Unlike PROBE, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are conditional on each other within each maximization step of H-PROBE. The estimated posterior covariance of parameters used to formulate PIs for H-PROBE also includes the effect of $\boldsymbol{\Sigma}$.

2.2 Estimation overview

We begin this Section with an overview of ECM and Parameter-Expanded (PX) Expectation-Maximization (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). For Bayesian methods, the EM results in MAP estimates of the parameters. For the model presented above, the standard M-step at iteration t consists of maximizing the expected complete-data log-posterior distribution

$E_{\boldsymbol{\gamma}}\left\{\log\ p(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\omega}|\boldsymbol{\mathcal{D}},\boldsymbol{\gamma})|\boldsymbol{\mathcal{D}},\boldsymbol{\beta}^{(t)},\boldsymbol{\varphi}^{(t)},\boldsymbol{\omega}^{(t)}\right\}$ where the expectation is over $\boldsymbol{\gamma}$,

$$\begin{aligned}\log\ p(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\omega}|\boldsymbol{\mathcal{D}},\boldsymbol{\gamma}) &= \log\left\{\det\left(\exp(-\boldsymbol{V}\boldsymbol{\omega})\right)^{-\frac{1}{2}}\right\}\\ &+\left\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\varphi})'\exp(-\boldsymbol{V}\boldsymbol{\omega})^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\varphi})\right\}\\ &+p\log(\pi)-|\boldsymbol{\gamma}|\log\left(\frac{\pi}{1-\pi}\right)\\ &+c\mathbf{1}'(c^{\frac{1}{2}}\sigma_\omega^2\boldsymbol{I})^{-1}\boldsymbol{\omega}-c\mathbf{1}'\exp\left\{(c^{\frac{1}{2}}\sigma_\omega^2\boldsymbol{I})^{-1}\boldsymbol{\omega}\right\}\\ &+\text{constants},\end{aligned} \tag{4}$$

and $(\boldsymbol{\beta}^{(t)},\boldsymbol{\varphi}^{(t)},\boldsymbol{\omega}^{(t)})$ denote the current MAP estimates. Note that this is a non-convex optimization over the high-dimensional vector $(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\omega})$.

In the ECM algorithm, the single M-step is replaced by multiple computationally simpler CM-steps (Meng & Rubin, 1993). Specifically, each CM-step maximizes the expectation of (4) over a subvector of $(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\omega})$ while holding the remaining values at their current MAP estimates. For example, to find the current MAP estimate of β_k, let $\boldsymbol{W}_k = \boldsymbol{X}_{\setminus k}(\boldsymbol{\gamma}_{\setminus k}\boldsymbol{\beta}_{\setminus k}) + \boldsymbol{Z}\boldsymbol{\varphi}$ where $\boldsymbol{A}_{\setminus k}$ indicates a matrix, or vector without column or element k. This gives $E(\boldsymbol{Y}|\boldsymbol{W}_k) = \boldsymbol{X}_k\beta_k + \boldsymbol{W}_k$, where $\boldsymbol{W}_k$ encompasses the impact of all predictors except $\boldsymbol{X}_k$.

The PX-EM is another extension of the EM, where the model is rewritten with auxiliary terms (i.e., expanded parameters) to help with stability and convergence (Liu et al., 1998). Here, since $\boldsymbol{W}_k$ is estimated, we use parameter expansion to include expanded parameter α_k that adjusts for the impact of $\boldsymbol{W}_k$ when updating β_k for all $k = 1,\dots,p$. This gives $E(\boldsymbol{Y}|\boldsymbol{W}_k) = \boldsymbol{X}_k\beta_k + \boldsymbol{W}_k\alpha_k$. Similarly, let $\boldsymbol{W}_{p+1} = \boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\beta}$ and α_{p+1} be the expanded parameter for $\boldsymbol{\varphi}$ such that $E(\boldsymbol{Y}|\boldsymbol{W}_{p+1}) = \boldsymbol{Z}\boldsymbol{\varphi} + \boldsymbol{W}_{p+1}\alpha_{p+1}$. The expanded parameters are jointly optimized with their corresponding original parameter sub-vector. The PX-EM contains a remapping step, which is critical to eliminate over-parameterization, maintain identifiability, and preserve the expected complete-data log-posterior evaluated at the remapped parameters is the same as at the expanded parameters. In general, Jaakkola and Qi (2006) demonstrate that Parameter-Expanded Variational Bayes (PX-VB) – which has similarities with the proposed PX-ECM since both use coordinate-wise updates – improves the convergence over standard VB by reducing the dependence between the coordinate-wise updates.

The complete M-step of our PX-ECM results in updated MAP values for $(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\omega})$ and the expanded parameters $\boldsymbol{\alpha}_\beta = (\alpha_1,\dots,\alpha_p)$ and α_{p+1}. Here, $\boldsymbol{\alpha}_\beta$ improves the estimates of $\boldsymbol{\beta}$ by reducing the dependence between the coordinate-wise updates. However, the actual values of $\boldsymbol{\alpha}_\beta$ play no functional role. Conversely, α_{p+1} is used in the remapping step, which remaps $\boldsymbol{\beta}$ from the expanded to original parameter space. Specifically, our use of the PX-ECM proceeds as follows. First, we update the MAP estimates of $\beta_k|(\gamma_k = 1)$ and α_k for $k = 1,\dots,p$. Second, we obtain the current MAP estimates of $\boldsymbol{\varphi}$ and α_{p+1}. Third, the $\boldsymbol{\beta}$ estimates are remapped from the expanded to

the original parameter space. Fourth, the MAP estimate of $\boldsymbol{\omega}$ is calculated. Finally, the E-step is performed.

In Section A of the Supplementary Materials, we derive the modes of the expected log-conditional posterior distributions of all quantities. The MAP estimate of $\beta_k|(\gamma_k = 1)$ and α_k at step t are given by

$$(\hat{\beta}_k^{(t)}, \hat{\alpha}_k^{(t)})' = \{(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)}\}^{-1}(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1})^{(t-1)}\boldsymbol{Y}, \tag{5}$$

where $\boldsymbol{C}_k = (\boldsymbol{X}_k\ \boldsymbol{W}_k)$ and $(\boldsymbol{C}_k\boldsymbol{\Sigma}^{-1})^{(t-1)}$ and $(\boldsymbol{C}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_k)^{(t-1)}$ are used to denote the expectations of these quantities given the MAP estimates of the other parameters, i.e., $(\boldsymbol{C}_k\boldsymbol{\Sigma}^{-1})^{(t-1)} \equiv E(\boldsymbol{C}_k\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}_{\setminus k}^{(t-1)}, \boldsymbol{\varphi}^{(t-1)}, \boldsymbol{\omega}^{(t-1)})$. These expectations are functions of the expectations of $\boldsymbol{W}_k$ and $\boldsymbol{W}_k^2$ (obtained in the E-step). The updates for $\boldsymbol{\varphi}$ and α_{p+1} are similar to (5) with $\boldsymbol{C}_{p+1} = (\boldsymbol{Z}\ \boldsymbol{W}_{p+1})$ in place of $\boldsymbol{C}_k$. The remapped β_k values are $\alpha_{p+1}^{(t)}\beta_k^{(t)}$ for all k. The MAP estimate of $\boldsymbol{\omega}$ does not have a closed form and is obtained via quasi-Newton optimization (Fletcher, 1987).

The aim of the E-step is to obtain updates for $E(\boldsymbol{W}_k)$ and $E(\boldsymbol{W}_k^2)$ for all k, where the expectations are over $\boldsymbol{\gamma}_{\setminus k}$. This requires estimates $p_k = E(\gamma_k)$ for all k, made via a novel application of the empirical Bayes estimator commonly used in the two-groups approach to multiple testing (Efron, 2008; Liang, Paulo, Molina, Clyde, & Berger, 2008). This procedure requires estimates of the posterior variances of ϕ and $\beta_k|\gamma_k = 1$ for all k, which are obtained by assuming their marginal posteriors are Gaussian (given in the Supplemental Materials). The p_k estimates update the moments $\boldsymbol{W}_k$ and $\boldsymbol{W}_k^2$ for all k. More discussion contrasting PROBE to other extensions of the EM algorithm is available in McLain et al. (2025). All technical details for performing the estimation procedures are provided in Section A of the Supplementary Materials. The H-PROBE method is implemented in the `probe` `R` package available at https://github.com/alexmclain/PROBE.

2.3 Estimates and model checks

The H-PROBE method converges when subsequent changes in the expected $\boldsymbol{W}_{p+1}$ values are small, as they capture changes in all regression parameters. Specifically, convergence at iteration t is quantified via $CC^{(t)} = \log(n)\max_i\left\{(W_{i,p+1}^{(t)} - W_{i,p+1}^{(t-1)})^2/Var(W_{i,p+1}|\boldsymbol{\beta}^{(t)}, \boldsymbol{p}^{(t)})\right\}$, where the ECM algorithm has converged when $CC^{(t)} < \chi^2_{1,0.1}$ and $\chi^2_{1,0.1}$ represents the 0.1th quantile of a χ^2 distribution with 1 degree of freedom. Here, $\log(n)$ controls for the impact of sample size on the maximum of χ^2 random variables (Embrechts, Klüppelberg, & Mikosch, 2013). We initiate the algorithm using $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ 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. For the elements of $\boldsymbol{\omega}^{(0)}$, we initialize the first element to $\log(s_Y^2)$, where s_Y^2 is the sample variance of $\boldsymbol{Y}$, and all remaining elements to 0. These initial values lead to estimates of $\beta_k^{(1)}$ that correspond to the coefficient of a simple linear regression for each $\boldsymbol{X}_k$ on $\boldsymbol{Y}$. Algorithm 1 in the Supplementary Materials shows H-PROBE steps in sequence.

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

Parameter	Estimate	Equation	Parameter definition
$\boldsymbol{\beta}$	$\tilde{\boldsymbol{\beta}}$	(2), SM	Vector of sparse regression coefficients for predictors in $\boldsymbol{X}$, conditional on $\boldsymbol{\gamma} = 1$, in the model on the conditional mean.
$\boldsymbol{S}^2$	$\tilde{\boldsymbol{S}}^2$	SM	Vector of posterior variances of $\boldsymbol{\beta}\|(\boldsymbol{\gamma} = 1)$.
$\boldsymbol{\gamma}$	$\tilde{\boldsymbol{p}}$	(2), SM	Vector of inclusion indicators for sparse coefficients $\boldsymbol{\beta}$ associated with predictors in $\boldsymbol{X}$.
$\boldsymbol{\varphi}$	$\tilde{\boldsymbol{\varphi}}$	(2)	Vector of non-sparse regression coefficients in the model on the conditional mean.
$\boldsymbol{W}_{p+1}$	$\tilde{\boldsymbol{W}}_{p+1}$	SM	Overall (non-partitioned) latent parameter $\boldsymbol{W}_{p+1} = \boldsymbol{X}(\boldsymbol{\gamma\beta})$.
α_{p+1}	$\tilde{\alpha}_{p+1}$	SM	Overall (non-partitioned) coefficient adjusting for the impact of overall (non-partitioned) $\boldsymbol{W}_{p+1}$.
$\boldsymbol{\omega}$	$\tilde{\boldsymbol{\omega}}$	(3)	Vector of non-sparse regression coefficients in the model on the variance.
$\boldsymbol{\Sigma}$	$\tilde{\boldsymbol{\Sigma}}$	(3)	Diagonal variance matrix $\boldsymbol{\Sigma} = \exp\{-(\boldsymbol{V\omega})\}$.

Upon convergence, H-PROBE provides the MAP estimate of $(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\omega})$ and empirical Bayes estimates of p_k for all k. Table 1 provides a summary of the critical parameters $\tilde{\boldsymbol{W}}_{p+1}$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{\varphi}}$, $\tilde{\alpha}_{p+1}$, $\tilde{\boldsymbol{\omega}}$, $\tilde{\boldsymbol{\Sigma}}$, and $\tilde{\boldsymbol{S}}^2$, their estimates, as well as their definitions. In Numerical Studies (Section 3) and Data Analysis (Section 4), we use $\tilde{\alpha}_{p+1}(\tilde{\boldsymbol{p}}\tilde{\boldsymbol{\beta}})$ – a combination of $E(\boldsymbol{\gamma\beta})$ and the MAP estimate of α_{p+1} – to estimate the impact of sparse predictors $\boldsymbol{\gamma\beta}$. While the properties of non-sparse predictor coefficients $\tilde{\boldsymbol{\varphi}}$ are not the focus of this research, we do wish to account for the uncertainty they contribute to the PIs constructed below. We use $\tilde{\Psi}$ to designate the estimated posterior covariance of $(\tilde{\boldsymbol{\varphi}}, \tilde{\alpha}_{p+1})$, which we use in formulating PIs in Section 2.4.

We provide some guiding principles on how to examine the known low-dimensional set of variance predictors for inclusion in the model. The first step is to fit a high-dimensional homoscedastic regression model (via LASSO or PROBE) and extract the residuals. Second, using the extracted residuals and the known low-dimensional set of variance predictors, we apply the Breusch-Pagan or White tests (Breusch & Pagan, 1979; White, 1980). Here, we apply the White test as it allows non-linear relationships. Third, we visually inspect plots of the log-squared residuals by the variables identified in step two to determine appropriate parametric forms. Additionally, for situations where heterogeneity may depend on a grouping variable, Levene's, Bartlett's, and Brown-Forsythe tests can be used (Seber & Lee, 2003). Section A.5 of the Supplementary Materials provides additional information and examples for model checks. In situations where variance predictors are not necessary, the PROBE algorithm can be used as an alternative to H-PROBE.

2.4 Prediction intervals

This research aims to develop point estimates and PIs for a future observation not included in the training set with predictor data $\boldsymbol{X}_{new}$, $\boldsymbol{Z}_{new}$, and $\boldsymbol{V}_{new}$. MAP estimation does not provide posterior distributions of model parameters, just their mode, limiting predictive inference capabilities. As a result, we assume that the estimates of the posterior variance of $\tilde{\boldsymbol{\varphi}}$ and $\tilde{\beta}_k|\gamma_k = 1$ for all k can be used to capture the posterior variability of these parameters. To predict for a new observation, we use the MAP estimate of the predicted value $\tilde{Y}_{new} = \boldsymbol{Z}_{new}\tilde{\boldsymbol{\varphi}} + \tilde{W}_{p+1,new}\tilde{\alpha}_{p+1}$, where $\tilde{W}_{p+1,new} = \boldsymbol{X}_{new}(\tilde{\boldsymbol{p}}\tilde{\boldsymbol{\beta}})$. Further, the variance of $\tilde{W}_{p+1,new}$ is estimated with $\tilde{\mathcal{V}}_{new} = \boldsymbol{X}^2_{new}\left\{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{S}}^2 + \tilde{\boldsymbol{\beta}}^2\tilde{\boldsymbol{p}}(1-\tilde{\boldsymbol{p}})\right\}$. To estimate the variance of $\tilde{Y}_{new}$ while acknowledging the uncertainty in $\tilde{W}_{p+1,new}$, we use

$$Var(\tilde{Y}_{new}) = \tilde{\boldsymbol{C}}'_{new}\tilde{\Psi}\tilde{\boldsymbol{C}}_{new} + \tilde{\mathcal{V}}_{new}\left\{Var(\tilde{\alpha}_{p+1}) + \tilde{\alpha}^2_{p+1}\right\},$$

where $\tilde{\boldsymbol{C}}_{new} = (\boldsymbol{Z}_{new}, \tilde{W}_{p+1,new})'$, which is motivated by the measurement error literature (Buonaccorsi, 1995). Prediction intervals (PIs) can be formed using the appropriate critical values with $Var(\tilde{Y}_{new}) + \tilde{\sigma}^2_{new}$ where $\tilde{\sigma}^2_{new} = \exp\{-(\boldsymbol{V}'_{new}\tilde{\boldsymbol{\omega}})\}$ denotes the MAP estimate of the variance for a *new* subject. In the following sections, we evaluate the empirical coverage probabilities of PIs using this approach for test data via simulation studies and a clinical application.

3 Numerical Studies

We perform numerical studies to evaluate the performance of H-PROBE. We generate the outcome using $Y_i = \boldsymbol{X}'_i(\boldsymbol{\gamma}\boldsymbol{\beta}) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2_i)$, $\sigma^2_i = \exp(-\boldsymbol{V}'_i\boldsymbol{\omega})$, $\boldsymbol{\beta} \sim U(0, 2\eta_\beta)$, and $\omega_j = \bar{\omega}$ for all j. We set $\bar{\omega}$ such that the expected signal-to-noise ratio (SNR) is $SNR = 1$ or 2, where $SNR = E\{Var(\boldsymbol{X}'_i\boldsymbol{\gamma}\boldsymbol{\beta})/\exp(-\boldsymbol{V}'_i\boldsymbol{\omega})\}$. We generated correlated continuous predictors $\boldsymbol{X}_i \sim MVN(\boldsymbol{0}+a_i, \Sigma)$ where $a_i \sim N\left(0, \frac{3}{4}\right)$ and Σ is a squared exponential covariance function. Specifically, all predictors are superimposed on a $\sqrt{p} \times \sqrt{p}$ grid, where $\boldsymbol{d}_k = (d_{1k}, d_{2k})$ denotes coordinates of X_k. The (k, k') element of the covariance matrix is $\exp\{-||(\boldsymbol{d}_k - \boldsymbol{d}_{k'})/\Sigma_{cor}||^2_2\}$ where $||\cdot||_2$ denotes the ℓ_2-norm. Σ_{cor} quantifies the overall strength of the dependence between predictors.

Our correlation-inducing concept is motivated by Gaussian random fields (Schlather, Malinowski, Menck, Oesting, & Strokorb, 2015) and mimics the real-world AQ application, where brain voxels closer together are more correlated than voxels further away from each other. Similarly, the SNR settings mirror the AQ application and, in some scenarios, make estimation more challenging (i.e., those with lower SNR). We also tested correlated binary predictors generated by applying the indicator that the continuous predictors are less than zero. $\boldsymbol{\gamma}$ was generated similarly to the binary predictor variables, such that $\sum\gamma_k = p\pi$ in each iteration. Finally, we generated $\boldsymbol{V}$ to include an intercept along with an equal number of standard normal and Bernoulli(0.5) predictor variables.

Simulation settings were varied by the number of predictors in $\boldsymbol{X}$, $p = (20^2, 75^2)$, the number of predictors in $\boldsymbol{V}$ including an intercept, $v = (3, 7)$, the proportion

of non-zero $\boldsymbol{\beta}$ coefficients, $\pi = (0.01, 0.05)$, the signal-to-noise ratio, $SNR = (1, 2)$, the average effect size of $\boldsymbol{\beta}$, $\eta_\beta = (0.3, 0.8)$, and the dependence between predictors, $\Sigma_{cor} = (10, 20, 30)$. All simulations presented herein had $n = 400$ observations and were repeated 400 iterations. For brevity, we focus below on the results for $\eta_\beta = 0.8$, $\Sigma_{cor} = 20$, and binary $\boldsymbol{X}$ predictors. Results for $\eta_\beta = 0.3$ and continuous $\boldsymbol{X}$ were nearly identical and are omitted while results for $n = 200$ and $\Sigma_{cor} = (10, 30)$ were very similar and are presented in Sections B.1-B.3 of the Supplementary Materials. All simulations were performed on an Intel Xeon 8358 Platinum processor with 2.6GHz CPU and 128 GB memory.

We compare H-PROBE to PROBE and LASSO for all settings. We also compare H-PROBE to heteroscedastic BART (HBART), a Bayesian linear model with a horseshoe prior (with 6000 MCMC iterations, 1000 used as the burn-in, Carvalho, Polson, & Scott, 2010) and an empirical Bayes approach for prediction in sparse high-dimensional linear regression (EBREG, Martin & Tang, 2020). Due to the high computational cost, we only ran 100 repetitions of these Bayesian competitors for settings for $p = 20^2$. For LASSO, we used the `glmnet R` package to implement ten-fold cross-validation (CV) to select parameters requiring tuning. For HBART, we used the `rbart R` package with models on the mean and variance as well as default parameters. We specified the Bayesian model with a horseshoe prior using the `horseshoe R` package. For the EBREG approach, we used the `ebreg R` package. We considered comparisons with Daye et al. (2012), but the computation requirements were prohibitive, with many settings running for over one hour per iteration.

The LASSO, PROBE, EBREG, and horseshoe approaches do not model heteroscedasticity, while H-PROBE and HBART do. We compared the performance of the methods with Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD, in Section B.1 of the Supplementary Materials) of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$, where data $\boldsymbol{X}$ consists of new observations not used during estimation (test set). Figures 2 and B.2 (in the Supplementary Materials) show that H-PROBE had the lowest RMSE and MAD for nearly all simulation settings. The horsehoe approach had lower MAD than H-PROBE when $\pi = 1\%$. H-PROBE led to marked efficiency gains when the proportion of signals, the number of predictors on the mean, or the effect size of $\boldsymbol{\beta}$ were higher. EBREG, horsehoe, and HBART methods had computation times that were at least eight-fold that of H-PROBE, PROBE, or LASSO in the smallest p settings ($p = 20^2$). In the settings with $SNR = 1$ and $\pi = 1\%$, computation time grew to at least 18-fold the computational time of H-PROBE, and up to 173-fold that of H-PROBE.

We obtained empirical coverage probabilities (ECPs) of 95% prediction intervals (PIs), given as the proportion of PIs that contained $Y_{i,test}$. We compared the PI ECPs of H-PROBE to PROBE and a Conformal Inference approach based on the LASSO (*split* variant, `conformalInference R` package), which estimates PIs for an existing model (Tibshirani & Foygel, 2019). Figure 3A shows that the average ECPs for H-PROBE and Conformal Inference PIs consistently remain centered at 0.95. In contrast, the average ECPs for PROBE can exceed 0.95 with interquartile ranges above the 0.95 level, particularly when $p = 75^2$. The average PI lengths for Conformal Inference are larger than those for H-PROBE, particularly when $p = 75^2$ (Figure 3B). Since both have ECPs at the nominal level, this is an indication that H-PROBE is using

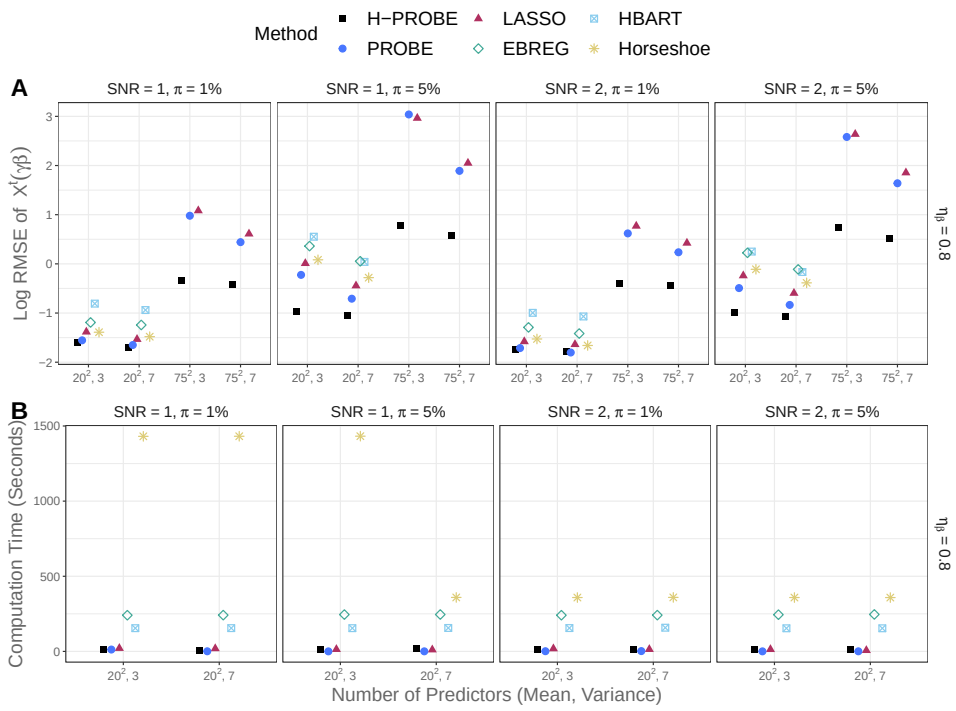

Fig. 2 This figure shows the model performance and computation time results from the numerical studies. Panel (A) shows the log Root Mean Squared Errors (RMSE) of $\boldsymbol{X}'(\boldsymbol{\gamma\beta})$, where $\boldsymbol{X}$ consists of new observations not used during estimation (test set) for six methods. Panel (B) shows the computation time in seconds for all 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). Figures showing results from additional settings are provided in Section B.1 of the Supplementary Materials.

the heterogeneity information for efficient PI computations. While the focus of our research is on the construction of PIs, we examined the ECPs of credible intervals for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ in Section B.1 of the Supplementary Materials. We found that the ECPs of $\boldsymbol{\beta}$ were accurate and close to the nominal level, whereas the ECPs of $\boldsymbol{\omega}$ were below the nominal level with a wide range. Additional results concerning the performance of PIs are also reported in Section B.2 of the Supplementary Material.

To compare the variable selection abilities of the methods, we calculated True Positive Rate (TPR) $TPR = \sum_{k;\gamma_k=1} \hat{\gamma}_k/|\boldsymbol{\gamma}|$ and the False Discovery Rate (FDR) $FDR = \sum_{k;\gamma_k=0} \hat{\gamma}_k/|\hat{\boldsymbol{\gamma}}|$ where $\hat{\gamma}_k = 1$ if variable k was 'selected' for the given method. For H-PROBE and PROBE $\hat{\gamma}_k = I(\tilde{p}_k > 0.5)$, while a predictor was selected for LASSO if the estimated coefficient was non-zero. H-PROBE performed well in variable selection. Figure 4A shows that H-PROBE correctly selects the highest proportion of the true signals in all settings. Further, Figure 4B shows that H-PROBE has a lower FDR than LASSO in all settings. Comparing the FDR between PROBE and H-PROBE, we see they are similar for $p = 20^2$ and lower for PROBE when $p = 75^2$

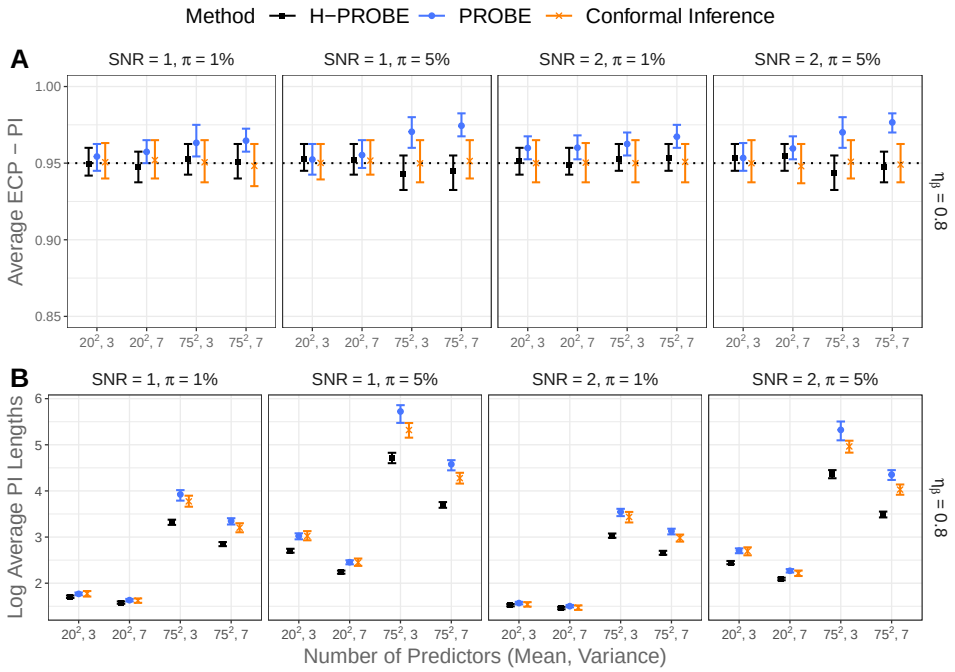

Fig. 3 This figure shows results related to Prediction Intervals (PIs) and Empirical Coverage Probabilities (ECP) from the numerical studies. In Panel (A), ECPs are defined as the proportion of PIs that contained the value $Y_{i,test}$ from the test set. In Panel (B), PI lengths are the difference between the upper and lower PI bounds for $Y_{i,test}$. The methods compared in this figure are H-PROBE (black squares), PROBE (blue circles), and Conformal Inference (orange crosses), for selected simulation settings. Vertical lines represent the first and third quartiles of the distributions of ECPs and PI lengths. Figures showing PI-related results from additional settings are provided in Section B.2 of the Supplementary Materials.

(where PROBE is markedly conservative). Additional variable selection results are reported in Section B.3 of the Supplementary Materials.

Sections B.4-B.6 of the Supplementary Materials include additional simulation results concerning the sensitivity of H-PROBE to initial values for the algorithm and misspecification of the variance model. We also considered $p \ll n$ scenarios where we compared H-PROBE with traditional approaches for low-dimensional heteroscedastic modeling. Briefly, we found that H-PROBE is robust to initial values, and a misspecified variance model has some impact on model accuracy, however, average ECPs were largely unaffected. Finally, when $p \ll n$, traditional approaches sacrificed model accuracy. REML (Smyth, 2002) resulted in conservative PIs while the other methods had ECPs markedly below the nominal level.

4 Data Analysis

We return to our motivating example to illustrate the use and distinctive features of the H-PROBE method. While several researchers have predicted aphasia severity or

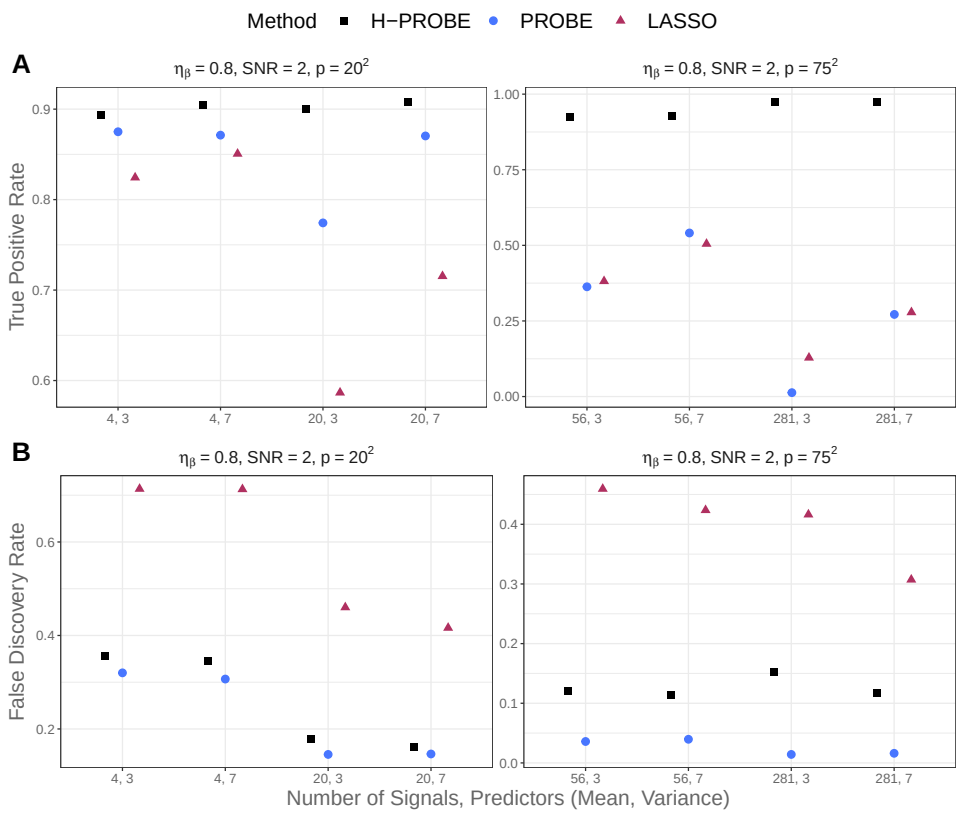

Fig. 4 This figure shows the True Positive Rate (TPR) in Panel (A) and False Discovery Rate (FDR) in Panel (B) for the H-PROBE (black squares), PROBE (blue circles), or LASSO (maroon triangles) methods plotted against the number of true signals ($|\gamma| = p\pi$) and the number of predictors on the variance. Figures showing TPR and FDR results from additional settings are provided in Section B.3 of the Supplementary Materials.

type using brain images (Lee et al., 2021; Teghipco et al., 2023; Yourganov et al., 2015), none of these works addressed the fundamental issue of *uncertainty quantification*. Point estimate predictions are often insufficient in clinical settings. Uncertainty quantification is crucial in precision medicine because it allows providers to assess the reliability of predictions and formulate optimal treatment plans (Banerji et al., 2023; Begoli et al., 2019; Zou et al., 2023). Our application aims to use patients' imaging and brain damage data to predict *and* quantify uncertainty of AQ score, which in turn guides post-stroke aphasia treatment decisions.

The data include $n = 167$ patients who have recently experienced a left-hemispheric stroke and are candidates for language rehabilitation therapy (Johnson et al., 2019; Yourganov et al., 2015). All individuals were scanned using a 3T MRI scanner, and an expert identified the lesion boundaries by hand via a high-resolution T2 scan. The lesions were then coregistered to the individual's T1 scan and warped to have a common size and shape through an enantiomorphic normalization clinical toolbox

(Nachev, Coulthard, Jäger, Kennard, & Husain, 2008; Rorden, Bonilha, Fridriksson, Bender, & Karnath, 2012). The resulting data contain over 5×10^5 binary features capturing the presence or absence of lesions in each 1 mm^3 brain voxel. Before comparing the performance of heteroscedastic and homoscedastic approaches in this clinical scenario, we streamlined the analysis by performing a marginal screening procedure based on X. Wang and Leng (2016) and retained 3×10^4 candidate imaging predictors. X. Wang and Leng (2016)'s screening procedure relies on a projection from a high-dimensional to a low-dimensional space, retaining the predictors whose components have the most impact on AQ in the projection. It is those predictors that are 'screened into' the next steps of the analysis. Total Brain Damage (TBD) was retained, i.e., the total number of voxels with lesions in the brain (out of 5×10^5).

Figure 1A displays the relationship between AQ and TBD. Most AQ values are concentrated near the top of the range for lower brain damage and diffuse as brain damage increases. The residuals (squared and log-transformed) from homoscedastic PROBE by TBD are shown in Figure 1B. There is a strong non-linear relationship between the error variance and TBD. We use the White test to confirm that the variance of the (squared and log-transformed) PROBE residuals is dependent on TBD as outlined in Equation (6), with a p-value less than 0.001. Transformations of AQ also show evidence of heterogeneity (see Figure C.13 and transformation details in Section C of the Supplementary Materials). As a result, we analyze AQ on the original scale and account for heteroscedasticity in this application. We use the H-PROBE approach, where the model on the variance includes TBD as well as its square-root transformation as predictors

$$\boldsymbol{\Sigma} = diag(\exp\left\{-\left(\omega_1 \mathbf{1} + \omega_2 \times \mathbf{TBD} \ + \omega_3 \times \sqrt{\mathbf{TBD}}\right)\right\}. \tag{6}$$

Our analyses include Conformal Inference (LASSO model, *split* variant, Tibshirani & Foygel, 2019), PROBE, and EBREG, a Bayesian method based on empirical priors for prediction in sparse high-dimensional linear regression (Martin & Tang, 2020). EBREG provides a simple algorithm based on a local search improvement rule that correctly identifies the support of the regression coefficients and subsequently provides PIs. We also include two traditional methods for heteroscedastic linear regression in the low-dimensional setting: OLS with White standard errors (OLSW, White, 1980), as well as heteroscedastic regression with models on the mean and the variance using restricted maximum likelihood (REML, Smyth, 2002).

For all methods, we model the AQ ($\boldsymbol{Y}$) as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For the variable selection and penalization approaches, $\boldsymbol{X}$ contains all 3×10^4 candidate imaging predictors. As discussed in Section 1.1, the traditional non-sparse methods (OLSW and REML) cannot be used when $p \gg n$ and it is necessary to further reduce the dimension of $\boldsymbol{X}$ so that $p \ll n$. To do so, we applied an additional principal component analysis to the 3×10^4 imaging predictors and selected the first 16 principal components to form a new $\boldsymbol{X}$ for OLSW and REML. These 16 principal components accounted for 74% of the total variance of the imaging predictors and aligned with p in the numerical studies in Section B.6 of the Supplementary Materials. As a result of this, the comparison between the high- and low-dimensional approaches is not direct. Nevertheless, this

Table 2 Performance metrics for H-PROBE and three high-dimensional homoscedastic comparison methods, PROBE, Conformal Inference, and EBREG, for the Aphasia Quotient (QA) analysis. Two heteroscedastic but low-dimensional methods are also evaluated, OLSW and REML, and are indicated by * in the Table.

Method \Metric	MAD	MSPE	Average PI Length	ECP
H-PROBE	8.635	220.671	60.749	0.952
PROBE	11.929	380.673	71.558	0.922
Conformal Inference	13.393	509.903	100.192	0.946
EBREG	17.677	614.314	91.582	0.940
OLSW*	12.300	381.735	68.104	0.916
REML*	63.227	5342.808	311.772	1.000

allows us to construct PIs based on OLWS and REML. We used `R` packages `probe`, `conformalInference`, `ebreg`, and `statmod` with 5-fold CV and default parameters (McLain & Zgodic, 2021; Tang & Martin, 2021; Tibshirani & Foygel, 2019). Due to the prohibitive computational demands of fitting horseshoe and HBART for moderate p, we omitted them from our analysis.

To evaluate and compare the methods, we used 5-fold CV to calculate Mean Squared Predictive Error (MSPE), Median Absolute Deviation (MAD), and empirical coverage probability (ECP) of 95% PIs where coverage implies that the PI for a subject in the *test* fold included their actual observation. Figures in Section C of the Supplementary Materials provide additional results.

Table 2 shows that H-PROBE had the lowest MAD and MSPE, followed by PROBE, Conformal Inference, and EBREG. H-PROBE also had the shortest average PI length with ECP close to the nominal 0.95 level. This pattern is consistent with H-PROBE providing accurate predictions for new observations and down-weighting observations with high estimated variance. For the traditional low-dimensional heteroscedastic methods, OLSW had similar model performance metrics to those of sparse approaches, while REML had poor overall performance. Figures C.16 and C.17 in the Supplementary Materials provide per-voxel statistical brain maps that compare the performance of H-PROBE to PROBE and LASSO. There is a large overlap between the voxels with positive $\boldsymbol{\beta}$ estimates between the H-PROBE to PROBE. However, H-PROBE provided a more sparse model. This may be due to PROBE's misspecified homoscedastic model leading to more false discoveries. There was little overlap between the voxels selected by H-PROBE and LASSO.

Figure 5 shows PI lengths for each patient and each method by fold. The Conformal Split and EBREG methods had similar PI lengths, while REML (omitted from Figure 5) had very wide PI lengths compared to all methods. Specifically, the average REML

PI lengths by CV fold were 317, 310, 311, 311, and 310, respectively. As anticipated, H-PROBE displayed the widest range of PI lengths, reflected by the differing estimated $\tilde{\sigma}_i^2$ by subject (a Figure of $\tilde{\sigma}_i^2$ is given in Figure C.14 of the Supplementary Materials). While H-PROBE had the lowest average PI length across observations, it had the largest ECP, so the narrower PIs did not sacrifice coverage. We show predictions and PIs for two subjects from the study data in Figure 6. The panel for Subject 4 shows moderately wide PIs for H-PROBE and wider PIs for other methods, spanning multiple AQ severity levels. The AQ predictions for Subject 23 are mild using all methods but one, and only H-PROBE provided a tight PI that does not go below a the mild aphasia severity threshold.

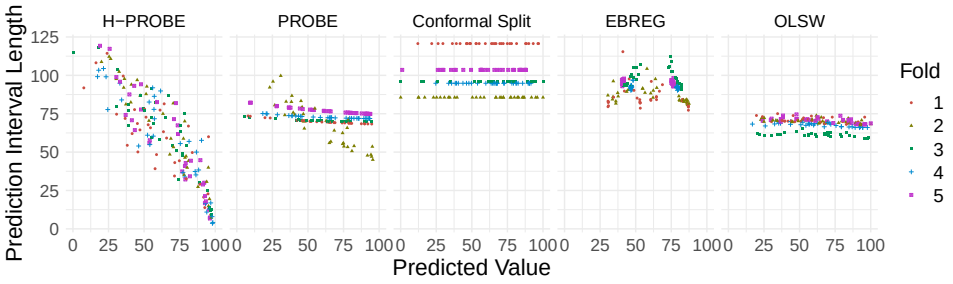

Fig. 5 This figure presents Prediction Interval (PI) lengths as a function of each patient's predicted Aphasia Quotient (AQ). H-PROBE is compared to three high-dimensional but homoscedastic methods, PROBE, Conformal Inference, and EBREG, as well as two low-dimensional but heteroscedastic methods, OLSW and REML (REML is omitted from the figure). The color legend represents the cross-validation fold in which predictions were obtained. A version of this figure which includes REML is shown in Figure C.15 of the Supplementary Materials.

Shorter PIs impact the judgment of clinical outcomes like aphasia severity category (AQ less than 26 is very severe, 26–50 is severe, 51–75 is moderate, and above 75 is mild) (Kertesz, 2007). For example, H-PROBE was the only method with PIs that spanned one aphasia severity category, where 29 subjects were predicted to have mild aphasia with the lower limit of their PIs above 75. None of the other methods provided PIs that spanned only one category. EBREG, REML, Conformal Split, PROBE, OLWS, REML, and H-PROBE had 100%, 100%, 96%, 77%, 77%, and 63% of their PIs cover at least three severity categories (out of four), respectively. Large uncertainty hinders the utility of Conformal Split, EBREG, and PROBE in clinical scenarios as it deprives clinicians of an accurate understanding of patients' potential outcomes. Accurate predictions are always critical as they give the most likely patient outcome. Nevertheless, correctly estimated uncertainty allows clinicians to gauge the risk of worse (or better) outcomes. In risk-sensitive healthcare settings, understanding the risk is crucial for making treatment decisions that balance patient safety and treatment efficacy (Banerji et al., 2023). Furthermore, methods that yield correctly estimated uncertainty can help avoid unnecessary additional diagnostic procedures, thereby preserving patient care and reducing medical costs. For example, patients with PIs that are limited to one AQ severity category would not require further testing; a

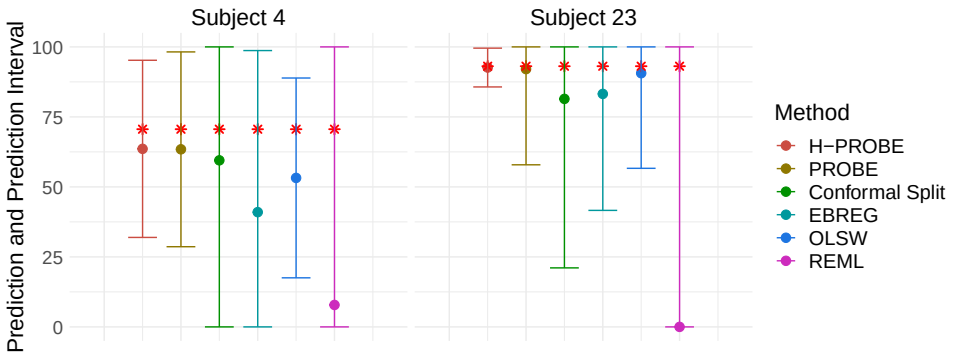

Fig. 6 In this figure, predictions and Prediction Intervals (PIs) are compared between six methods indicated by the color legend. H-PROBE is compared to three high-dimensional but homoscedastic methods, PROBE, Conformal Inference, and EBREG, as well as two low-dimensional but heteroscedastic methods, OLSW and REML. Red stars indicate the subject's true Aphasia Quotient (AQ) score, circles indicate the predicted AQ, and the vertical bars represent PI length. AQ ranges that indicate aphasia severity are: AQ less than 26 is very severe, 26–50 is severe, 51–75 is moderate, and above 75 is mild.

benefit to the patient, as manual testing for AQ is a difficult task for patients who have recently had a stroke, and reduces medical expenditures.

5 Discussion

In this paper, we developed a novel approach to conduct high-dimensional linear regression for heteroscedastic data. H-PROBE uses a Bayesian framework with parameter expansion and minimally informative priors on the parameters. H-PROBE is a computationally effective solution to sparse linear regression in heteroscedastic settings that combines an empirical Bayes estimator with the PX-ECM algorithm. Simulation studies illustrated that accounting for heterogeneity in variance errors via H-PROBE generally resulted in more accurate estimation and prediction, as shown by lower MSEs and MADs for model predictions, compared to PROBE, LASSO, HBART, EBREG, and the horseshoe. Empirical coverage probabilities of prediction intervals were consistently at the nominal 95% level, with smaller PI lengths than other methods. H-PROBE is one of the few Bayesian approaches to address these issues in high-dimensional settings by including models on both the mean and the variance. Compared to HBART, H-PROBE is more scalable and can perform variable selection, enhancing H-PROBE's interpretability in clinical settings.

On the application side, we contributed to the literature on the prediction of aphasia severity by constructing PIs for AQ scores in stroke patients. Our analyses reinforced that appropriately accounting for non-constant error variances can improve predictive ability and PI lengths while maintaining coverage. Accurately assessing the predictive uncertainty is essential to the acceptance and effectiveness of models in risk-sensitive healthcare settings (Banerji et al., 2023; Begoli et al., 2019). In such cases, the level of uncertainty helps physicians gauge the risk associated with a predictive

estimate (Zou et al., 2023). For example, a predicted AQ with a narrow 95% PI may lead to a specific treatment plan for post-stroke aphasia, whereas a wide PI may indicate that additional testing is necessary. H-PROBE provided narrow PIs where other methods could not (in patients with mild aphasia, AQ > 75), which allows clinicians to define treatment courses that weigh a relevant range of potential treatment outcomes. PIs and PI lengths that capture patient-level heteroscedasticity enable personalized treatment decisions based on each patient's stroke-induced total brain damage and MRI imaging. In this application, homoscedastic methods such as Conformal Inference, PROBE, and EBREG are likely to yield less accurate predicted AQ values and, consequently, suboptimal treatment decisions compared to H-PROBE.

We focused our research on the situation where known heterogeneity markers exist in the data. However, in practice, the variables' impact on heterogeneity may be unknown. A valuable extension of H-PROBE is to the situation where model selection on mean and variance parameters is required (Chiou et al., 2020; K. Zhou, Li, & Zhou, 2021). If the heterogeneity markers are appropriately included in the variance model, observations with high residual variability will be accordingly down-weighted. A limitation of H-PROBE is that it assumes a linear model, which may be insufficiently flexible. In contrast, HBART models both the mean and the variance nonparametrically (Pratola et al., 2020). As a trade-off, however, HBART is much more computationally demanding than H-PROBE. One way to achieve greater flexibility for H-PROBE while retaining its computational advantages over HBART is to use a semiparametric additive model $Y_i = \sum_{k=1}^{p} f(X_{ik}) + \epsilon_i$. In this case, we could approximate each $f(X_{ik})$ as a linear combination of basis functions and use H-PROBE to regularize basis coefficients to zero, similar to Bai, Moran, Antonelli, and Boland (2022) and Guo, Jaeger, Rahman, Long, and Yi (2022). This is a useful extension for future work. Another relevant extension of H-PROBE is to settings where the observations are dependent or autocorrelated. In this case, it would be necessary to estimate a covariance matrix $\mathbf{\Sigma}$ with nonzero off-diagonal entries.

Abbreviations. AQ, aphasia quotient; PI, prediction interval; TBD, total brain damage; MRI, magnetic resonance imaging; MAP, maximum a posteriori; PX, parameter expanded; EM, expectation-maximimization; ECM, expectation conditional maximization; SNR, signal-to-noise; CV, cross-validation; ECP, empirical coverage probability; RMSE, root mean squared error; MSE, mean squared error; MAD, median absolute deviation; TPR, true positive rate; FDR, false discovery rate; MSPE, mean squared predictive error; LAD, Least Absolute Deviation; LTS, least trimmed squares.

Supplementary information. *Supplementary Materials*: The Supplementary Materials contain an expanded Methods section as well as additional results from simulations and real data analyses. (.pdf).

Funding. The authors acknowledge that this research has been supported in part by a National Institute on Deafness and Other Communication Disorders grant (R01 DC009571).

Compliance with Ethical Standards. The authors have no competing interests to declare that are relevant to the content of this article.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. Authors are unable to make data available due to the ongoing clinical trials through which these data are collected. The clinical trials and the research data collected are governed by an Institutional Review Board and are not shared at this time.

Materials availability. Not applicable.

Author contribution. All authors have accepted responsibility for the entire content of this manuscript and approved its submission. *Anja Zgodic*: Conceptualization, Methodology, Software, Writing. *Ray Bai*: Conceptualization, Methodology, Writing, Supervision. *Jiajia Zhang*: Writing - Review & Editing. *Yuan Wang*: Writing - Review & Editing. *Christopher Rorden*: Data Curation, Writing - Review & Editing. *Alexander McLain*: Conceptualization, Methodology, Writing, Software, Supervision.

References

Alfons, A., Croux, C., Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, *7*(1), 226–248,

Bai, R., Moran, G.E., Antonelli, J.L., Boland, M.R. (2022). Spike-and-slab group lassos for grouped regressions and sparse generalized additive models. *Journal of the American Statistical Association*, *117*(537), 184–197,

Banerji, C., Chakraborti, T., Harbron, C., MacArthur, B. (2023). Clinical ai tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, *29*, 2996–2998,

Begoli, E., Bhattacharya, T., Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*, 20–23,

Belloni, A., Chernozhukov, V., Wang, L. (2014). Pivotal Estimation via Square-Root LASSO in Nonparametric Regression. *The Annals of Statistics*, *42*(2), 757–788,

Bradley, J., Holan, S., Wikle, C. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, *115*(532), 2037–2052,

Breusch, T., & Pagan, A. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294,

Buonaccorsi, J.P. (1995). Prediction in the presence of measurement error: General discussion and an example predicting defoliation. *Biometrics*, 1562–1569,

Carroll, R. (1988). *Transformation and weighting in regression*. New York: Chapman and Hall.

Carroll, R.J., & Ruppert, D. (1988). *Transformation and weighting in regression* (Vol. 30). CRC Press.

Carvalho, C.M., Polson, N.G., Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465-480,

Chiou, H.T., Guo, M., Ing, C.K. (2020). Variable selection for high-dimensional regression models with time series and heteroscedastic errors. *Journal of Econometrics*, *216*(1), 118–136, https://doi.org/10.1016/j.jeconom.2020.01.009 Retrieved from https://doi.org/10.1016/j.jeconom.2020.01.009

Cleveland, W. (1993). *Visualizing data*. Hobart Press.

Curto, J., Pinto, J., Morais, A., Lourenco, I. (2011). The heteroskedasticity- consistent covariance estimator in accounting. *Review of Quantitative Finance and Accounting*, *37*(4), 427–449,

Daye, Z.J., Chen, J., Li, H. (2012). High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis. *Biometrics*, *68*(1), 316–326, https://doi.org/10.1111/j.1541-0420.2011.01652.x

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).

Embrechts, P., Klüppelberg, C., Mikosch, T. (2013). *Modelling extremal events: for insurance and finance* (Vol. 33). Springer Science & Business Media.

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

Guo, B., Jaeger, B.C., Rahman, A.K.M.F., Long, D.L., Yi, N. (2022). Spike-and-slab least absolute shrinkage and selection operator generalized additive models and scalable algorithms for high-dimensional data analysis. *Statistics in Medicine*, *41*(20), 3899-3914,

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).

Jaakkola, T., & Qi, Y. (2006). Parameter expanded variational bayesian methods. *Advances in Neural Information Processing Systems*, *19*, ,

Johnson, L., Basilakos, A., Yourganov, G., Cai, B., Bonilha, L., Rorden, C., Fridriksson, J. (2019). Progression of aphasia severity in the chronic stages of stroke. *American Journal of Speech-Language Pathology*, *28*(2), 639–649,

Kertesz, A. (2007). *Western aphasia battery–revised (wab-r).* London, UK: Pearson.

Lee, J., Ko, M., Park, S., Kim, G. (2021). Prediction of aphasia severity in patients with stroke using diffusion tensor imaging. *Brain Sciences*, *11*(3), 304,

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111,

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

Martin, R., & Tang, Y. (2020). Empirical priors for prediction in sparse high-dimensional linear regression. *Journal of Machine Learning Research*, *21*, 1–30,

McLain, A.C., & Zgodic, A. (2021, 9). *Fitting high-dimensional linear regression models with probe.* https://github.com/alexmclain/PROBE. Retrieved from https://github.com/alexmclain/PROBE

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

Nachev, P., Coulthard, E., Jäger, H.R., Kennard, C., Husain, M. (2008). Enantiomorphic normalization of focally lesioned brains. *Neuroimage*, *39*(3), 1215–1226,

Odekar, A., & Hallowell, B. (2005). Comparison of alternatives to multidimensional scoring in the assessment of language comprehension in aphasia. *American Journal of Speech-Language Pathology*, *14*(4), 337-345, https://doi.org/10.1044/1058-0360(2005/032)

Parker, P., Holan, S., Wills, S. (2021). A general bayesian model for heteroskedastic data with fully conjugate full-conditional distributions. *Journal of Statistical Computation and Simulation*, *91*(15), 3207–3227,

Peng, P., Chiou, H., Huang, H., Ing, C. (2025). Variable selection for high-dimensional heteroscedastic regression and its applications. *Journal of Computational and Graphical Statistics*, 1–11,

Pratola, M.T., Chipman, H.A., George, E.I., McCulloch, R.E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, *29*(2), 405–417,

Risser, A.H., & Spreen, O. (1985). The western aphasia battery. *Journal of clinical and experimental neuropsychology*, *7*(4), 463–470,

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., Karnath, H.-O. (2012). Age-specific CT and MRI templates for spatial normalization. *Neuroimage*, *61*(4), 957–965,

Rousseeuw, P., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, *12*(1), 29–45,

Schlather, M., Malinowski, A., Menck, P.J., Oesting, M., Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, *63*(8), 1–25, Retrieved from https://www.jstatsoft.org/v63/i08/

Seber, G.A., & Lee, A.J. (2003). *Linear regression analysis* (Vol. 330). John Wiley & Sons.

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

Tang, Y., & Martin, R. (2021). ebreg: Implementation of the empirical bayes method [Computer software manual]. Vienna, Austria. Retrieved from https://CRAN.R-project.org/package=ebreg (R package version 0.1.3)

Teghipco, A., Newman-Norlund, R., Fridriksson, J., Rorden, C., Bonilha, L. (2023). Distinct brain morphometry patterns revealed by deep learning improve prediction of aphasia severity. *Research Square*, ,

Temlyakov, V. (2000). Weak greedy algorithms. *Advances in Computational Mathematics*, *12*, 213-–227,

Tibshirani, R., & Foygel, R. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, ,

Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wang, H., Li, G., Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business and Economic Statistics*, *25*(3), 347–355, https://doi.org/10.1198/073500106000000251

Wang, X., & Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(3), 589–611, Retrieved 2022-10-08, from http://www.jstor.org/stable/24775353

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

Yourganov, G., Smith, K., Fridriksson, J., Rorden, C. (2015). Predicting aphasia type from brain damage measured with structural mri. *Cortex*, *73*, 203–2015,

Zhou, K., Li, K.-C., Zhou, Q. (2021). Honest confidence sets for high-dimensional regression by projection and shrinkage. *Journal of the American Statistical Association*, 1–20,

Zhou, L., & Zou, H. (2021). Cross-fitted residual regression for high-dimensional heteroscedasticity pursuit. *Journal of the American Statistical Association*, *0*(0), 1–10, https://doi.org/10.1080/01621459.2021.1970570

Ziel, F. (2016). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to AR-ARCH type processes. *Computational Statistics and Data Analysis*, *100*, 773–793, https://doi.org/10.1016/j.csda.2015.11.016 Retrieved from http://dx.doi.org/10.1016/j.csda.2015.11.016 arXiv:1502.06557

Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H. (2023). A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, *1*(1), 100003,