SUPPLEMENTARY MATERIALS for "Parameter Expanded Variational Bayes for Well-Calibrated High-Dimensional Linear Regression with Spike-and-Slab Priors"

Peter Olejua[1], Enakshi Saha[1], Rahul Ghosal[1], Ray Bai[2], Alexander McLain[1*]

[1]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, 29208, USA.
[2]Department of Statistics, George Mason University, Fairfax, VA, 22030, USA.

*Corresponding author(s). E-mail(s): mclaina@mailbox.sc.edu;
Contributing authors: polejua@email.sc.edu; esaha@mailbox.sc.edu; rghosal@mailbox.sc.edu; rbai2@gmu.edu;

A *spexvb* updates derivation

A.1 Updates for $q(b_j, s_j)$

The updates for $q(b_j, s_j)$ have been given in Carbonetto and Stephens (2012) and Ray and Szabó (2022). We provide a full derivation here for completeness.

$$\begin{aligned}
\log q_j(\theta_j) &\stackrel{+}{=} E_{-j}\left[\log p(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{y})\right] \\
&= E_{\{b_k, s_k : k \neq j\}, \pi}\left[\log p(b_j, s_j, \boldsymbol{b}_{-j}, \boldsymbol{s}_{-j}, \pi, \boldsymbol{y})\right] \\
&\stackrel{+}{=} E_{-j}\left[\frac{n}{2}\log(\tau_\epsilon) - \frac{1}{2}\tau_\epsilon \sum_{i=1}^{n}\left[y_i - \boldsymbol{x}_i^\top(\boldsymbol{b}\odot\boldsymbol{s})\right]^2 + \frac{p}{2}\log(\tau_\epsilon\tau_b) - \frac{1}{2}\tau_\epsilon\tau_b\sum_{k=1}^{p} b_k^2\right. \\
&\qquad \left. + \log\left(\frac{\pi}{1-\pi}\right)\sum_{k=1}^{p} s_k + (a_\pi - 1)\log(\pi) + (p + b_\pi - 1)\log(1-\pi)\right]
\end{aligned}$$

$$
\begin{aligned}
&\stackrel{+}{=} -\frac{1}{2}\tau_\epsilon\left[\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i \sum_{k=1}^{p} x_{ik}s_k b_k + \sum_{i=1}^{n}\left(\sum_{k=1}^{p} x_{ik}s_k b_k\right)^2\right] \\
&\quad -\frac{1}{2}\tau_\epsilon\tau_b \sum_{k=1}^{p} E_{-j}(b_k^2) + E_{-j}\left[\log\left(\frac{\pi}{1-\pi}\right)\sum_{k=1}^{p} s_k\right] \\
&\stackrel{+}{=} -\frac{1}{2}\tau_\epsilon\left\{-2\sum_{i=1}^{n} y_i E_{-j}(x_{ij}s_j b_j)\right. \\
&\quad \left. + \sum_{i=1}^{n} E_{-j}\left[(x_{ij}s_j b_j)^2 + 2x_{ij}s_j b_j \sum_{k\neq j} x_{ik}s_k b_k + \left(\sum_{k\neq j} x_{ik}s_k b_k\right)^2\right]\right\} \\
&\quad -\frac{1}{2}\tau_\epsilon\tau_b b_j^2 + s_j E_{-j}[\mathrm{logit}(\pi)] \\
&\stackrel{+}{=} \tau_\epsilon\left[s_j b_j \sum_{i=1}^{n} y_i x_{ij} - \frac{1}{2}(s_j b_j)^2 \sum_{i=1}^{n} x_{ij}^2 - s_j b_j \sum_{i=1}^{n} x_{ij} \sum_{k\neq j} x_{ik} E_{-j}\left(s_k b_k\right)\right] \\
&\quad -\frac{1}{2}\tau_\epsilon\tau_b b_j^2 + s_j E_{-j}[\mathrm{logit}(\pi)]
\end{aligned}
$$

Now, let $w_{ij} = \sum_{k\neq j} x_{ik} E_{-j}\left(s_k b_k\right)$ and $\boldsymbol{w}_j = (w_{1j}, w_{2j}, ..., w_{nj})$. Then

$$
\begin{aligned}
\log q_j(\theta_j) &\stackrel{+}{=} \tau_\epsilon\left[s_j b_j \boldsymbol{x}_j^\top \boldsymbol{y} - \frac{1}{2}(s_j b_j)^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j - s_j b_j \boldsymbol{x}_j^\top \boldsymbol{w}_j - \frac{1}{2}\tau_b b_j^2 + s_j E_{-j}[\mathrm{logit}(\pi)]\right] \\
&= \tau_\epsilon\left[b_j\left(s_j \boldsymbol{x}_j^\top \boldsymbol{y} - s_j \boldsymbol{x}_j^\top \boldsymbol{w}_j\right) - \frac{b_j^2}{2}\left(s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_b\right) + s_j E_{-j}[\mathrm{logit}(\pi)]\right] \\
&= -\frac{1}{2}\left\{b_j^2\left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon\tau_b\right) - 2\tau_\epsilon b_j\left[s_j \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{w}_j)\right] + s_j E_{-j}[\mathrm{logit}(\pi)]\right\} \\
&= -\frac{1}{2}\left[b_j^2\left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon\tau_b\right) - 2b_j\left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon\tau_b\right)\frac{s_j \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{w}_j)}{s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_b}\right. \\
&\quad + \left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon\tau_b\right)\left(\frac{s_j \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{w}_j)}{s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_b}\right)^2 \\
&\quad \left. - \left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon\tau_b\right)\left(\frac{s_j \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{w}_j)}{s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_b}\right)^2 + s_j E_{-j}[\mathrm{logit}(\pi)]\right] \\
&= -\frac{1}{2}\frac{\left[b_j - \mu_j(s_j)\right]^2}{\sigma_j^2(s_j)} + \frac{1}{2}\frac{\mu_j(s_j)^2}{\sigma_j^2(s_j)} + s_j E_{-j}[\mathrm{logit}(\pi)]
\end{aligned}
$$

where

$$\mu_j(s_j) = \frac{s_j \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{w}_j)}{s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_b}, \text{ and}$$
$$\sigma_j^2(s_j) = \left(\tau_\epsilon s_j^2 \boldsymbol{x}_j^\top \boldsymbol{x}_j + \tau_\epsilon \tau_b\right)^{-1}.$$

A.2 Updates for $\boldsymbol{\pi}$

The update for π would be

$$\begin{aligned}
\log q(\pi) &\stackrel{+}{=} E_{-j}\left[\log(p(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{y}))\right] \\
&= E\left[\log(p(\pi, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{y}))\right] \\
&\stackrel{+}{=} E\left[\frac{n}{2}\log(\tau_\epsilon) - \frac{1}{2}\tau_\epsilon ||\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{b} \odot \boldsymbol{s})||^2 + \frac{p}{2}\log(\tau_b) - \frac{1}{2}\tau_\epsilon \tau_b ||\boldsymbol{b}||^2 \right. \\
&\qquad \left. + \log\left(\frac{\pi}{1-\pi}\right)\sum_{k=1}^{p} s_k + (a_\pi - 1)\log(\pi) + (p + b_\pi - 1)\log(1-\pi)\right] \\
&\stackrel{+}{=} \log\left(\frac{\pi}{1-\pi}\right)\sum_{k=1}^{p} E(s_k) + (a_\pi - 1)\log(\pi) + (p + b_\pi - 1)\log(1-\pi) \\
&= (M + a_\pi - 1)\log(\pi) + (p + b_\pi - M - 1)\log(1-\pi) \\
&= \log\left(\pi^{(M + a_\pi - 1)}(1-\pi)^{(p + b_\pi - M - 1)}\right),
\end{aligned}$$

where $M = \sum_{k=1}^{p} E(s_k)$. Exponentiating both sides gives

$$q(\pi) \propto \pi^{a_\pi + M - 1} \times (1-\pi)^{p + b_\pi - M - 1}.$$

Therefore, we have $q(\pi) \sim \text{Beta}\,(a_\pi + M, p + b_\pi - M)$.

B Additional Computational Details

In the simulation studies and data analysis, how `spexvb` obtains initial values, treats hyperparameters, and sets the convergence criteria are inherited from Ray and Szabó (2022), so that comparisons best isolate the impact of parameter expansion. This section provides details on these modeling choices; see Ray and Szabó (2022) for more information.

To obtain the initial values for `spexvb`, cross-validated lasso and ridge models are fitted to the data. From the lasso model with the regularization parameter that minimizes the mean cross-validated error, denoted by $\lambda_{\min}$, we obtain the number of non-zero coefficients, denoted by $nz_{\lambda_{\min}}$, and $\hat{y}$, the predicted outcome. Similarly, from the lasso model with the largest regularization parameter that has cross-validated error within one standard error of the minimum, we obtain $nz_{\lambda_{1se}}$, the number of non-zero

coefficients. Next, we calculate $nz_{\min} = \min(nz_{\lambda_{1se}}, n-2)$, $c_\pi = e^{1/2} \cdot \max(nz_{\min}, 1)$, $d_\pi = p - c_\pi$, and $\tau_\epsilon^{-1} = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2/(n - nz_{\min} - 1)$ following Reid et al. (2016).

The initial variational mean $\boldsymbol{\mu}^{(0)}$ is obtained as the coefficients of the ridge model corresponding to $\lambda_{\min}$. Finally, the updating order is determined by sorting the absolute values of the entries in $\boldsymbol{\mu}^{(0)}$ in descending order and taking the arranged indices. For $\boldsymbol{\omega}^{(0)}$, the jth entry is 1 if the corresponding coefficient has a non-zero value in the lasso $\lambda_{\min}$ model, and $\max(nz_{\min}, 1)/p$ otherwise.

We initialize $\boldsymbol{W}$ as $\boldsymbol{W} \leftarrow \boldsymbol{X} \odot (\boldsymbol{\mu}_0 \odot \boldsymbol{\omega}_0)$. To efficiently update this quantity during the coordinate descent iterations, we can leverage the fact that only one element of $\boldsymbol{b}$ and $\boldsymbol{s}$ changes at a time. That is, we subtract the old contribution of the j-th coordinate: $\boldsymbol{W}_j \leftarrow \boldsymbol{W} - \boldsymbol{x}_j \mu_j \omega_j$; update μ_j and ω_j; and add back its contribution: $\boldsymbol{W} \leftarrow \boldsymbol{W}_j + \boldsymbol{x}_j \mu_j \omega_j$. For $\boldsymbol{W}^2$, which is only required for the update of α, we initialize it with $\boldsymbol{W}^2 \leftarrow (\boldsymbol{X} \odot \boldsymbol{X}) \cdot (\boldsymbol{\mu}_0 \odot \boldsymbol{\mu}_0 \odot \boldsymbol{\omega}_0 \odot (1 - \boldsymbol{\omega}_0)) + (\boldsymbol{W} \odot \boldsymbol{W})$ and update it before Step II.

The convergence criterion is based on the entropy of the posterior probability vector $\boldsymbol{\omega}$ and a predefined tolerance level. The entropy of the vector $\boldsymbol{\omega}$ is calculated via

$$H(\omega_j) = -\omega_j \log_2(\omega_j) + (1 - \omega_j) \log_2(1 - \omega_j)$$

Our stopping criteria is $\max_j \|H(\omega_{j,\text{new}}) - H(\omega_{j,\text{old}})\| \leq tol$ for some $tol > 0$.

C Additional Simulation Results

Here, we include additional plots of the simulation results for other $n = 1000$ that were not included in the main article. See Figures 1 - 4.

D Alternative expanded model derivations

In the previous model derivations, we used $b_j \sim N\{0, (\tau_\epsilon \tau_b)^{-1}\}$ as the prior for the *expanded parameters*. This results in the prior for the *original parameters* being a function of α, and the need to recalculate $\tau_b^{(t+1)}$ at each iteration.

Alternatively, we could have set $\tilde{b}_j \sim N\{0, (\tau_\epsilon \tau_b)^{-1}\}$ as the prior for the *original parameters*, then the prior for the *expanded parameters* is $b_j \sim N\{0, (\alpha^2 \tau_\epsilon \tau_b)^{-1}\}$. Setting all other priors the same yields the expanded model likelihood

$$\begin{aligned}
\log p(\boldsymbol{\theta}, \alpha, \boldsymbol{X}, \boldsymbol{y}) \stackrel{+}{=} & \frac{n}{2} \log(\tau_\epsilon) \\
& - \frac{1}{2} \tau_\epsilon ||\boldsymbol{y} - \alpha \boldsymbol{X}(\boldsymbol{b} \odot \boldsymbol{s}))||^2 \\
& - \frac{p}{2} \log(\tau_\epsilon \tau_b \alpha^2) - \frac{\tau_\epsilon \tau_b \alpha^2}{2} ||\boldsymbol{b}||^2 \\
& - \frac{1}{2} \log(\tau_\epsilon \tau_a) - \frac{\tau_\epsilon \tau_a}{2} (\alpha - \mu_a)^2 \\
& + (M_1 + a_\pi - 1) \log \pi + (M_0 + b_\pi - 1) \log(1 - \pi).
\end{aligned}$$

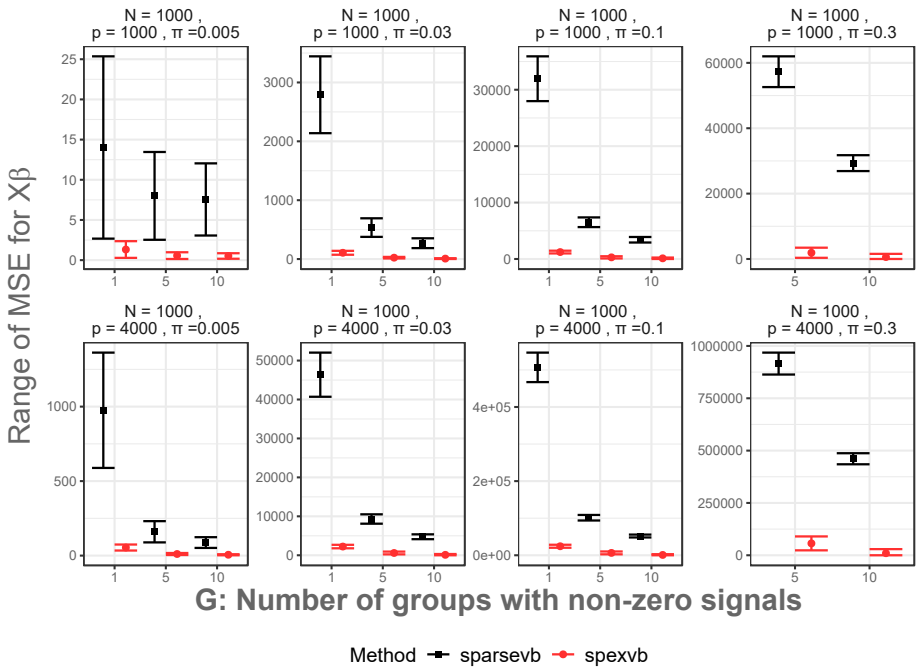

Fig. 1 The range (maximum-minimum) of test set MSE comparing the true $X\beta$ against the estimated $X\hat{\beta}$ over all $\tau_b^{1/2} \in (1/20, 1/4, 1, 4, 20)$. In the boxplots, the centers represent the mean range of the MSE and the whiskers ± 1 standard deviation.

The derivation of Step 1 is the same as the main text. For Step 2, having α in the prior for $\boldsymbol{b}$ means that portion of the equation cannot be ignored, and we seek to maximize

$$
\begin{aligned}
E\left\{\log p(\boldsymbol{\theta}, \alpha, \boldsymbol{X}, \boldsymbol{y})\right\} &\stackrel{+}{=} -\frac{1}{2}\tau_\epsilon E\left\{||\boldsymbol{y} - \alpha\boldsymbol{X}(\boldsymbol{b}\odot\boldsymbol{s}))||^2 + \tau_a(\alpha-\mu_a)^2 + \tau_b\alpha^2||\boldsymbol{b}||^2\right\} \\
&\quad -\frac{p}{2}\log(\tau_\epsilon\tau_b\alpha^2) \\
&\stackrel{+}{=} -\frac{1}{2}\tau_\epsilon\Big(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\alpha\boldsymbol{X}E(\boldsymbol{b}\odot\boldsymbol{s}) \\
&\quad +\alpha^2 E\left[\{\boldsymbol{X}(\boldsymbol{b}\odot\boldsymbol{s})\}^\top\{\boldsymbol{X}(\boldsymbol{b}\odot\boldsymbol{s})\}\right] \\
&\quad +\tau_a\alpha^2 - 2\tau_a\alpha\mu_a + \mu_a^2 + \tau_b\alpha^2 E||\boldsymbol{b}||^2\Big) \\
&\quad -p\log(\alpha) - \frac{p}{2}\log(\tau_\epsilon\tau_b) \\
&\stackrel{+}{=} \frac{1}{2}\tau_\epsilon\left\{2\alpha\boldsymbol{y}^\top(\boldsymbol{W}+\tau_a\mu_a) - \alpha^2(\boldsymbol{W}^2 + \tau_a + \tau_b E||\boldsymbol{b}||^2)\right\} - p\log(\alpha) \\
&\stackrel{+}{=} \frac{1}{2}\tau_\epsilon\left\{-\alpha^2 A + 2\alpha B - 2\log(\alpha)C\right\}.
\end{aligned}
$$

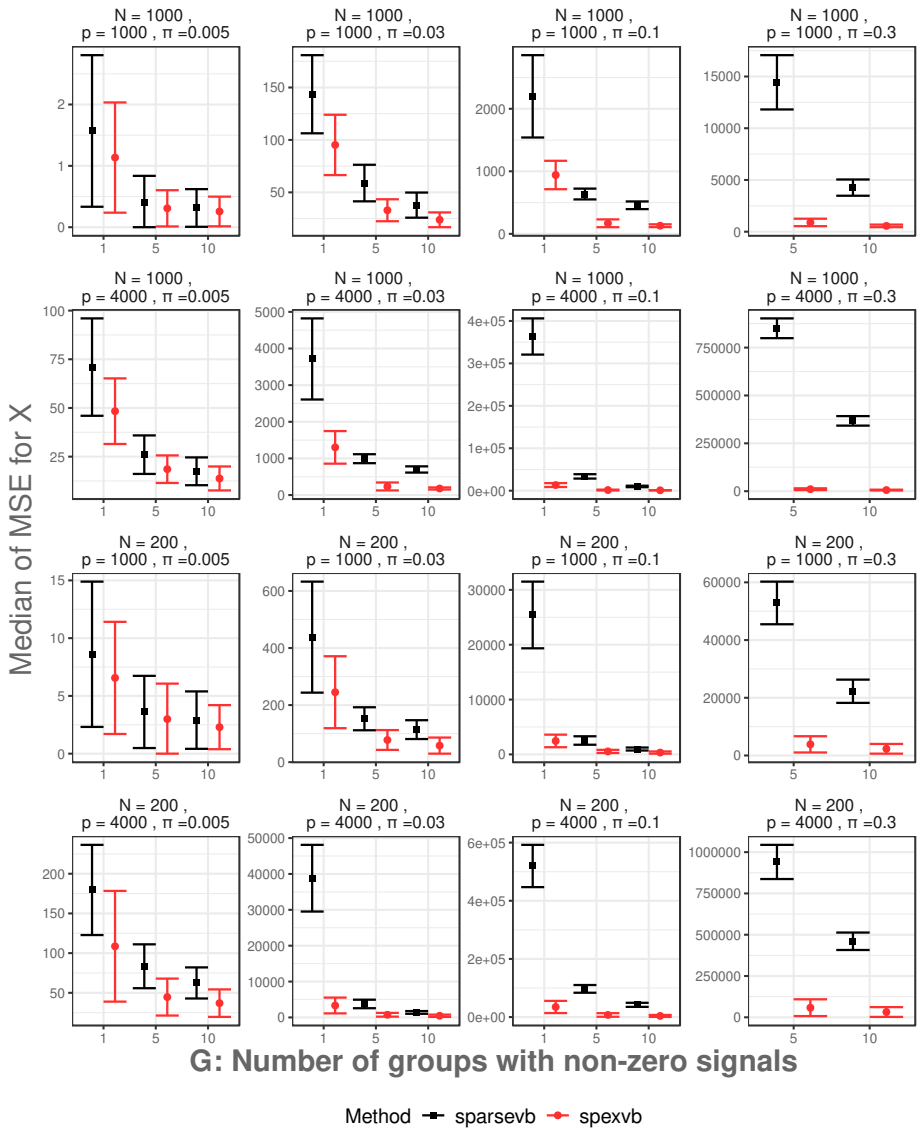

Fig. 2 Test set median MSE comparing the true $X\beta$ against the estimated $X\hat{\beta}$ over all $\tau_b^{1/2} \in (1/20, 1/4, 1, 4, 20)$. In the boxplots, the centers represent the mean value and the whiskers the ± 1 standard deviation.

where $A = \boldsymbol{W}^2 + \tau_a + \tau_b E||\boldsymbol{b}||^2$, $B = \boldsymbol{y}^\top(\boldsymbol{W} + \tau_a \mu_a)$, and $C = p/\tau_\epsilon$. Taking the derivative with respect to α and setting it to zero gives

$$\alpha^{(t+1)} = \frac{B + \sqrt{B^2 - 4AC}}{2A} \tag{1}$$

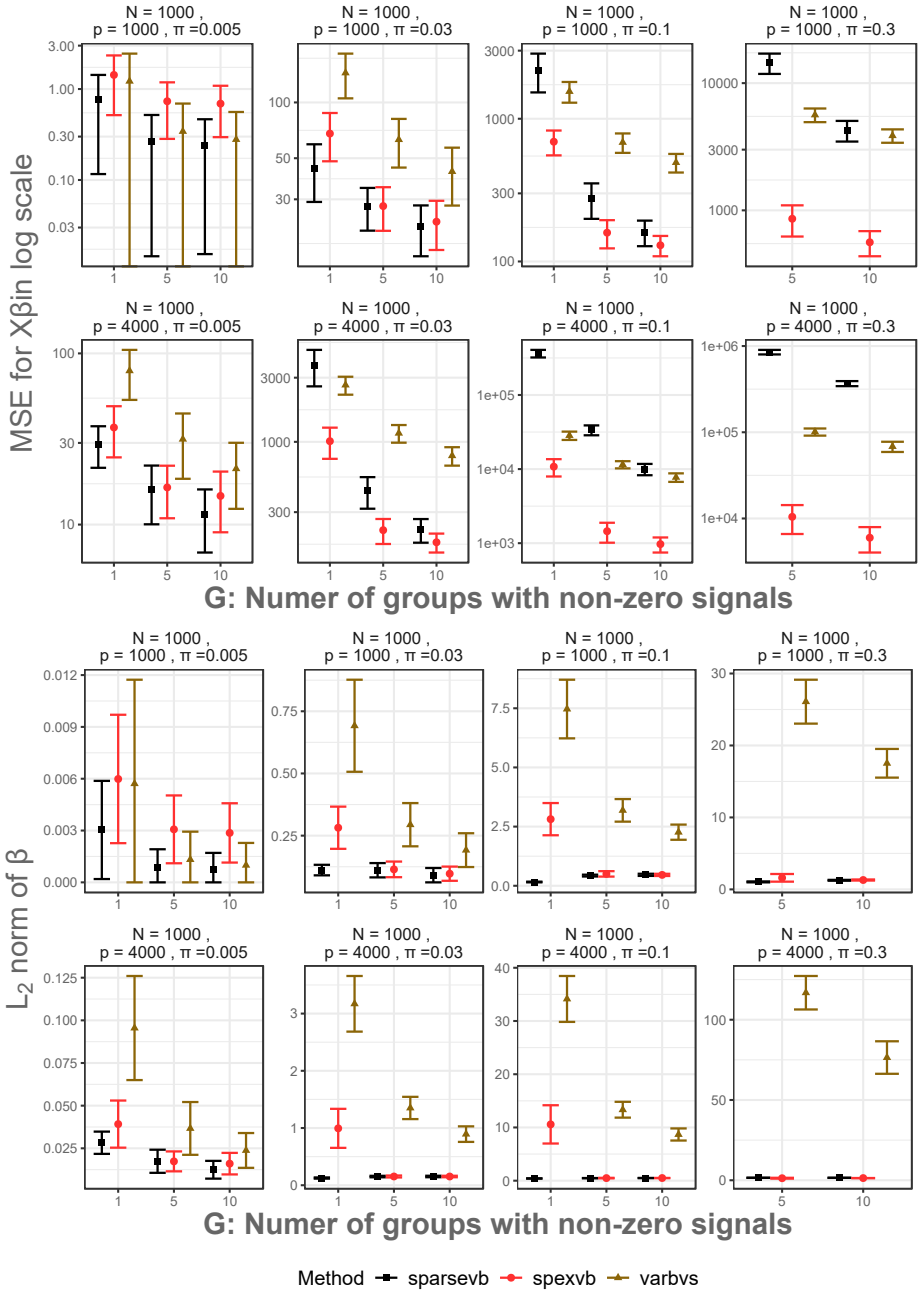

Fig. 3 (top) Test set MSE comparing the true $X\beta$ against the estimated $X\hat{\beta}$, and (bottom) L_2 norm of the $\beta - \hat{\beta}$. In the boxplots, the centers represent the mean value and the whiskers the ± 1 standard deviation.

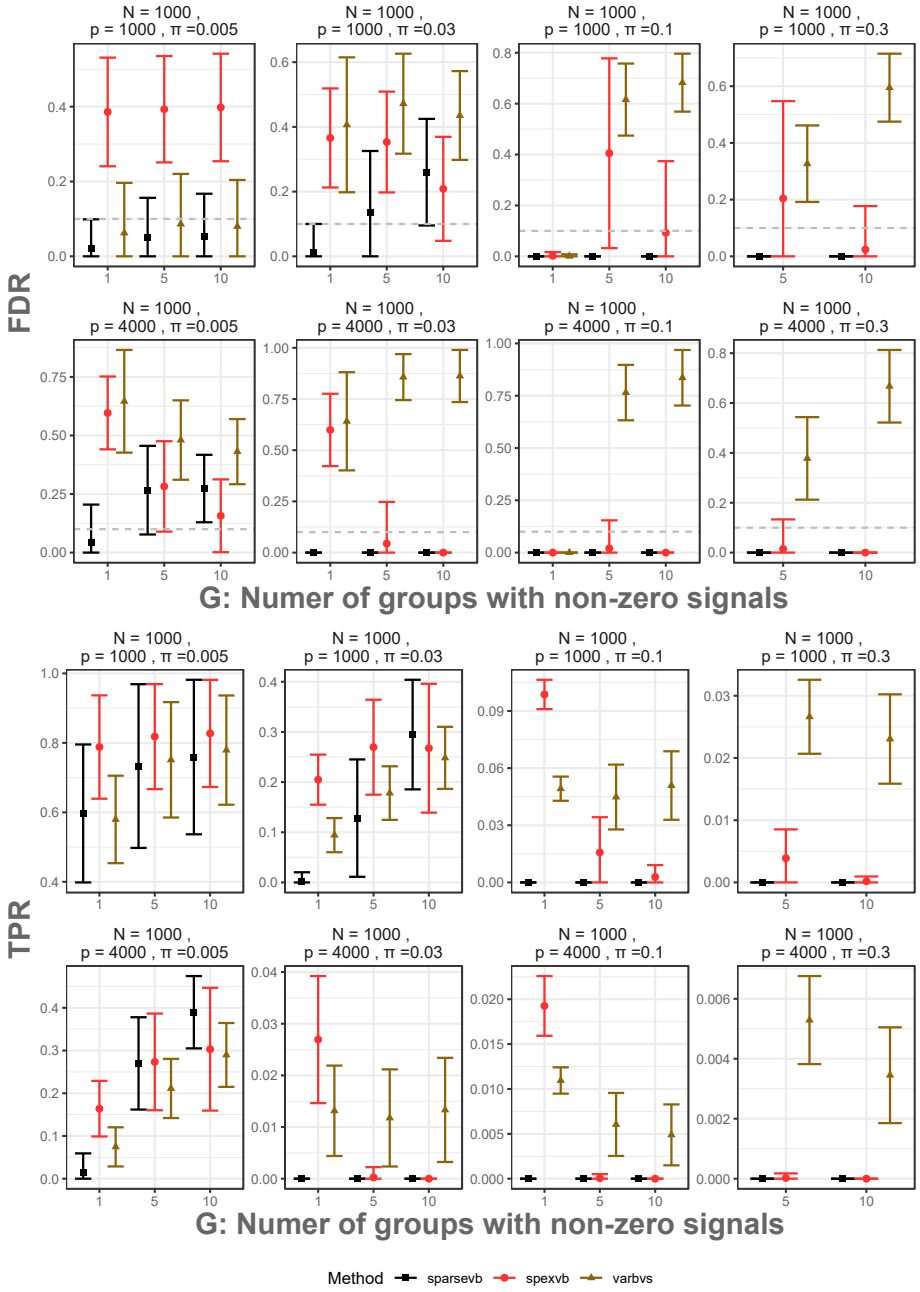

Fig. 4 (top) False discovery rate. (bottom) True Positive Rate. In the boxplots, the centers represent the mean value and the whiskers the ± 1 standard deviation.

as the maximum value. This equation has a similar form to the optimal expanded parameter for automatic relevance determination (ARD) given in Section 3.2 of Jaakkola and Qi (2006).

Step 3a would proceed similarly to the main text. Step 3b would not require any alteration of τ_b. However, a similar alteration of μ_a would need to be made to ensure that the expanded and original models have equal KL divergence.

References

Carbonetto, P., Stephens, M.: Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Analysis **7**(1), 73–108 (2012) https://doi.org/10.1214/12-BA703 . Accessed 2022-05-10

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

Ray, K., Szabó, B.: Variational bayes for high-dimensional linear regression with sparse priors. Journal of the American Statistical Association **117**(539), 1270–1281 (2022) https://doi.org/10.1080/01621459.2020.1847121

Reid, S., Tibshirani, R., Friedman, J.: A study of error variance estimation in lasso regression. Statistica Sinica, 35–67 (2016)