Supplementary material for variational Bayes for high-dimensional linear regression with sparse priors

Kolyan Ray* and Botond Szabó†

Imperial College London and Vrije Universiteit Amsterdam

Abstract

In the supplementary material, we provide additional simulation results (Section A), full oracle results and all proofs (Section B), additional methodological details, including proofs of the variational update equations (Section C), and examples of compatible design matrices (Section D).

A Additional numerical results

A.1 Ozone interaction data

We apply our method to the real world ozone interaction data investigated in [1]. The dataset contains $n = 203$ readings of maximal daily ozone measured in the Los Angeles basin and $p = 134$ variables modeling the pairwise interaction of 9 meteorological and 3 time variables. We firstly normalize the design matrix by centering and rescaling each column to have Euclidean norm equal to $\sqrt{n}$ and then add a column vector of ones to add an intercept to the model.[1] We apply the four methods investigated above (i.e. our method `sparsevb` [7], `varbvs`, `EMVS`, `SSLASSO`) with unknown noise variance ς^2, using the method settings described in Section 5.2. We also tried to apply the ebreg method, but due to the highly co-linear nature of the design matrix, the code gave errors when trying to compute the Cholesky decomposition.

As we do not know the underlying truth, we consider the 10-fold cross validation prediction error, i.e. we use nine folds to compute the posterior mean or MAP $\hat{\theta}$ and then use the 10th fold to compute the prediction error $\|Y - X\hat{\theta}\|_2$. We report the averaged out cross-validation errors in Table 3, together with the runtimes and number of selected covariates. Our method outperforms the other approaches in cross-validated prediction loss. Furthermore, while there is some overlap between the models selected by the various methods, the results are quite different, see Figure 2.

A.2 Comparing the VB algorithms

We compare our VB method with Laplace slabs (Algorithm 1) with different variations of the VB algorithm. First, we consider the other mean-field VB posterior $\widetilde{Q}$ derived from the

*Department of Mathematics, Imperial College London. E-mail: kolyan.ray@imperial.ac.uk

†Department of Mathematics, Vrije Universiteit Amsterdam. E-mail: b.t.szabo@vu.nl

Botond Szabó received funding from the Netherlands Organization for Scientific Research (NWO) under Project number: 639.031.654.

[1]Except for EMVS, since adding an intercept resulted in an error message.

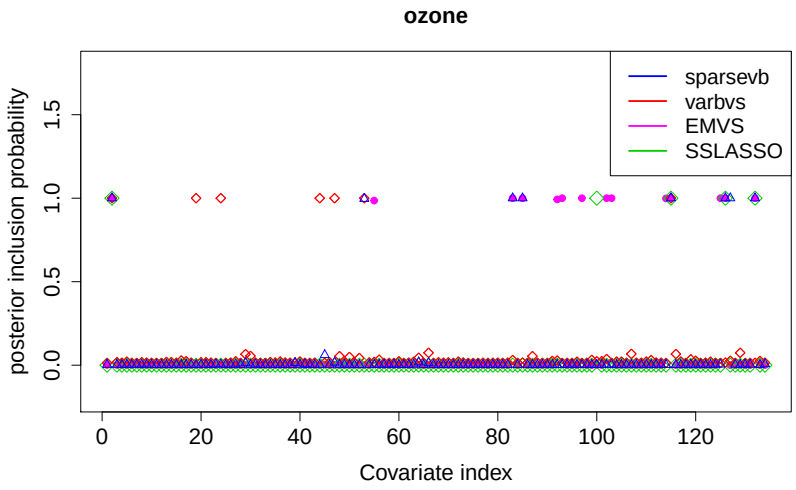

Figure 2: Marginal inclusion probabilities of the variables for the ozone interaction data using sparsevb (blue), EMVS (purple), SSLASSO (green) and varbvs (red).

Table 3: Cross-validated ℓ_2-estimation error of Bayesian model selection methods

data \ Method	sparsevb	varbvs	EMVS	SSLASSO
CV error	**16.43**	59.49	74.45	53.28
model size	9	7	14	5
runtime (sec)	1.49	1.14	**0.02**	0.10

variational class $\mathcal{Q}_{MF}$ (Algorithm 4 in Section C.2). Next, we consider the VB method with Gaussian prior slabs, which is the standard choice in the literature, see for instance [10, 4, 9], both with component-wise and batch-wise computational approaches, see Algorithms 2 and 3 in Section C.2. To compensate for the over-shrinkage of the posterior mean caused by the light tail of the Gaussian slabs, we also consider centered Gaussian prior slabs with standard deviation set to the (unknown) oracle $\rho = \|\theta_0\|_2$, as proposed by [6] for the sequence model (i.e. $X = I$ the identity matrix).

In all experiments, we placed the non-zero signal components $\theta_i = A$ at the beginning of the signal. In the first experiment, (i) we take the identity design matrix $X = I_n$ and set $n = p = 400$, $s = 40$, $A = 4\sqrt{\log n}$. In the other three experiments, we consider a Gaussian design matrix with entries $X_{ij} \overset{iid}{\sim} N(0, \tau^2)$ and vary the parameters n, p, s, τ and A. We take (ii) $(n, p, s, \tau) = (100, 200, 20, 1)$, $A \overset{iid}{\sim} U(0, 2\log n)$; (iii) $(n, p, s, \tau) = (200, 800, 40, 0.1)$, $A = 2\log n$; (iv) $(n, p, s, \tau) = (100, 400, 15, 0.5)$, $A \overset{iid}{\sim} U(-8, 8)$. In all experiments, we take $\varsigma = 1$ assumed to be known. The results over 200 runs are reported in Table 4 and we plot the outcome of a typical run in Figure 3.

Our Laplace VB method (sparsevb) with variational class $\mathcal{P}_{MF}$ typically outperforms the other VB algorithms. From the identity design case (i), it is clear that Gaussian prior slabs provide suboptimal recovery for θ_0 unless the prior slab variance is rescaled by the norm of θ_0. However, the rescaled Gaussian slabs perform much less well in the Gaussian design cases (ii)-(iv). The other mean-field variational class $\mathcal{Q}_{MF}$ performs similarly to our main method in the identity design case, but significantly worse in the more complicated Gaussian design cases. This is due to discrete nature of the variational parameter $\gamma \in \{0, 1\}$ in this family, which makes the optimization problem even more difficult, causing the method to frequently get stuck at a poor local minimum. We do not report run times as the `sparsevb` R-package is optimized for computation and therefore runs substantially faster than the other methods, which are more simply implemented.

A.3 The effect of the hyper-parameter λ

Theorem 1 states that for a wide range of hyper-parameter values $\lambda \in [\frac{\|X\|}{p}, \frac{C\|X\|\sqrt{\log p}}{s_0}]$, our VB algorithm has good asymptotic properties. However, the finite-sample performance depends on λ as we now investigate. We ran our algorithm for different choices of λ, ranging from 1/20 to 20, on simulated data similar to that in the preceding subsections.

We consider four different settings, each with Gaussian design with entries $X_{ij} \overset{iid}{\sim} N(0, \tau^2)$, non-zero signal components set to $\theta_i = A$ and noise variance $\varsigma^2 = 1$ assumed to be known. We take (i) $(n, p, s, \tau) = (200, 300, 15, 0.5)$, $A = 2\log n$; (ii) $(n, p, s, \tau) = (500, 1000, 50, 1)$, $A = 2\log n$; (iii) $(n, p, s, \tau) = (200, 500, 20, 0.2)$, $A \overset{iid}{\sim} U(-10, 10)$; and (iv) $(n, p, s, \tau) = (1000, 2000, 15, 2)$, $A \overset{iid}{\sim} U(-8, 8)$. In all cases, the non-zero signal components are located at the beginning of the signal. We ran each algorithm 200 times and report the results in Table 5. The choice of λ can indeed significantly influence the finite-sample behaviour of the algorithm (e.g. cases (ii) and (iii)), but not always ((i) and (iv)). There was not clear evidence to support a particular fixed choice of λ, since larger values sometimes performed better ((ii) and (iv)) and sometime worse ((i) and (iii)). This suggests using a data-driven choice of λ may be helpful in practice. As expected, larger choices for λ, which cause more shrinkage, result in smaller FDR and TPR. The runtime across hyper-parameter choices were broadly comparable.

Metric	Method\ Experiment	(i)	(ii)	(iii)	(iv)
$\ell_2 -$ error	Laplace $\mathcal{P}_{MF}$	8.80 ± 0.85	**0.60 ± 0.90**	**9.25 ± 9.73**	**1.08 ± 0.20**
	Laplace $\mathcal{Q}_{MF}$	8.80 ± 0.85	7.07 ± 1.48	39.98 ± 6.88	6.56 ± 1.97
	Gauss	31.06 ± 0.49	0.78 ± 1.14	43.58 ± 2.94	1.40 ± 0.29
	Gauss (batch-wise)	31.11 ± 0.48	16.38 ± 0.79	66.98 ± 0.00	18.03 ± 0.00
	Gauss ($\rho = \|\theta_0\|_2$)	**6.26 ± 0.72**	5.97 ± 6.16	58.12 ± 19.01	2.05 ± 3.59
FDR	Laplace $\mathcal{P}_{MF}$	**0.00 ± 0.00**	**0.00 ± 0.01**	**0.03 ± 0.11**	**0.00 ± 0.02**
	Laplace $\mathcal{Q}_{MF}$	**0.00 ± 0.00**	0.70 ± 0.07	0.45 ± 0.08	0.55 ± 0.14
	Gauss	**0.00 ± 0.00**	**0.00 ± 0.00**	0.50 ± 0.03	0.01 ± 0.03
	Gauss (batch-wise)	**0.00 ± 0.00**	0.87 ± 0.01	0.62 ± 0.03	0.82 ± 0.03
	Gauss ($\rho = \|\theta_0\|_2$)	**0.00 ± 0.00**	0.25 ± 0.36	0.57 ± 0.21	0.06 ± 0.21
TPR	Laplace $\mathcal{P}_{MF}$	**1.00 ± 0.00**	**0.89 ± 0.02**	**0.99 ± 0.06**	**0.81 ± 0.03**
	Laplace $\mathcal{Q}_{MF}$	**1.00 ± 0.00**	0.81 ± 0.06	0.88 ± 0.08	0.74 ± 0.07
	Gauss	**1.00 ± 0.00**	**0.89 ± 0.02**	0.94 ± 0.05	**0.81 ± 0.03**
	Gauss (batch-wise)	**1.00 ± 0.00**	0.88 ± 0.07	0.82 ± 0.07	0.68 ± 0.08
	Gauss ($\rho = \|\theta_0\|_2$)	**1.00 ± 0.00**	0.81 ± 0.10	0.58 ± 0.17	0.78 ± 0.10

Table 4: Linear regression with (i) identity design $X = I_n$, and $(ii) - (iv)$ Gaussian design $X_{ij} \overset{iid}{\sim} N(0, \tau^2)$. The non-zero coefficients are located in the beginning of the signal. The parameters (n, p, s, A) are set to (i) $(400, 400, 40, 4\sqrt{\log n})$; (ii) $(100, 200, 20, U(0, 2\log(n)))$; (iii) $(200, 800, 40, 2\log n)$; (iv) $(100, 400, 15, U(-8, 8))$. We set (ii) $\tau = 1$; (iii) $\tau = 0.1$; (iv) $\tau = 0.5$. We compare the means and standard deviations over 200 runs for our method and other variations of the VB algorithm.

Metric	Method	(i)	(ii)	(iii)	(iv)
$\ell_2 -$ error	$\lambda = 1/20$	**0.56 ± 0.11**	2.92 ± 7.84	2.49 ± 0.50	0.09 ± 0.02
	$\lambda = 1/4$	0.57 ± 0.12	0.97 ± 3.65	**2.34 ± 0.50**	0.08 ± 0.02
	$\lambda = 1$	0.57 ± 0.11	**0.34 ± 0.04**	**2.38 ± 0.48**	0.08 ± 0.02
	$\lambda = 4$	0.67 ± 0.12	0.35 ± 0.04	3.56 ± 0.51	**0.07 ± 0.02**
	$\lambda = 20$	1.85 ± 0.22	0.47 ± 0.05	11.94 ± 1.01	**0.07 ± 0.02**
FDR	$\lambda = 1/20$	**0.00 ± 0.00**	0.09 ± 0.27	**0.00 ± 0.00**	**0.00 ± 0.00**
	$\lambda = 1/4$	**0.00 ± 0.01**	0.03 ± 0.15	**0.00 ± 0.00**	**0.00 ± 0.00**
	$\lambda = 1$	**0.00 ± 0.01**	**0.00 ± 0.00**	**0.00 ± 0.01**	**0.00 ± 0.00**
	$\lambda = 4$	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.01**	**0.00 ± 0.01**
	$\lambda = 20$	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.02**	**0.00 ± 0.00**
TPR	$\lambda = 1/20$	**1.00 ± 0.00**	**1.00 ± 0.00**	0.91 ± 0.04	0.95 ± 0.03
	$\lambda = 1/4$	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.92 ± 0.04**	0.96 ± 0.03
	$\lambda = 1$	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.92 ± 0.04**	0.97 ± 0.03
	$\lambda = 4$	**1.00 ± 0.00**	**1.00 ± 0.00**	0.90 ± 0.04	**0.98 ± 0.03**
	$\lambda = 20$	**1.00 ± 0.00**	**1.00 ± 0.00**	0.59 ± 0.07	**0.98 ± 0.03**
runtime (sec)	$\lambda = 1/20$	**0.34 ± 0.29**	3.23 ± 0.96	**0.51 ± 0.10**	**4.23 ± 0.46**
	$\lambda = 1/4$	**0.27 ± 0.07**	3.40 ± 0.98	0.52 ± 0.13	**4.23 ± 0.50**
	$\lambda = 1$	**0.45 ± 0.41**	2.98 ± 0.72	**0.49 ± 0.09**	**4.23 ± 0.46**
	$\lambda = 4$	**0.41 ± 0.47**	2.46 ± 0.59	0.51 ± 0.12	4.25 ± 0.50
	$\lambda = 20$	**0.32 ± 0.24**	**2.17 ± 0.46**	0.50 ± 0.07	4.29 ± 0.65

Table 5: Performance of sparsevb for different hyper-parameter values λ. We take Gaussian design $X_{ij} \overset{iid}{\sim} N(0, \tau^2)$, place the non-zero signal coefficients $\theta_{0,i} = A$ at the beginning of the signal, and set the parameters (n, p, s, τ, A) equal to (i) $(200, 300, 15, 0.5, 2\log n)$; (ii) $(500, 1000, 50, 1, 2\log n)$; (iii) $(200, 500, 20, 0.2, U(-10, 10))$; (iv) $(1000, 2000, 15, 2, U(-8, 8))$.

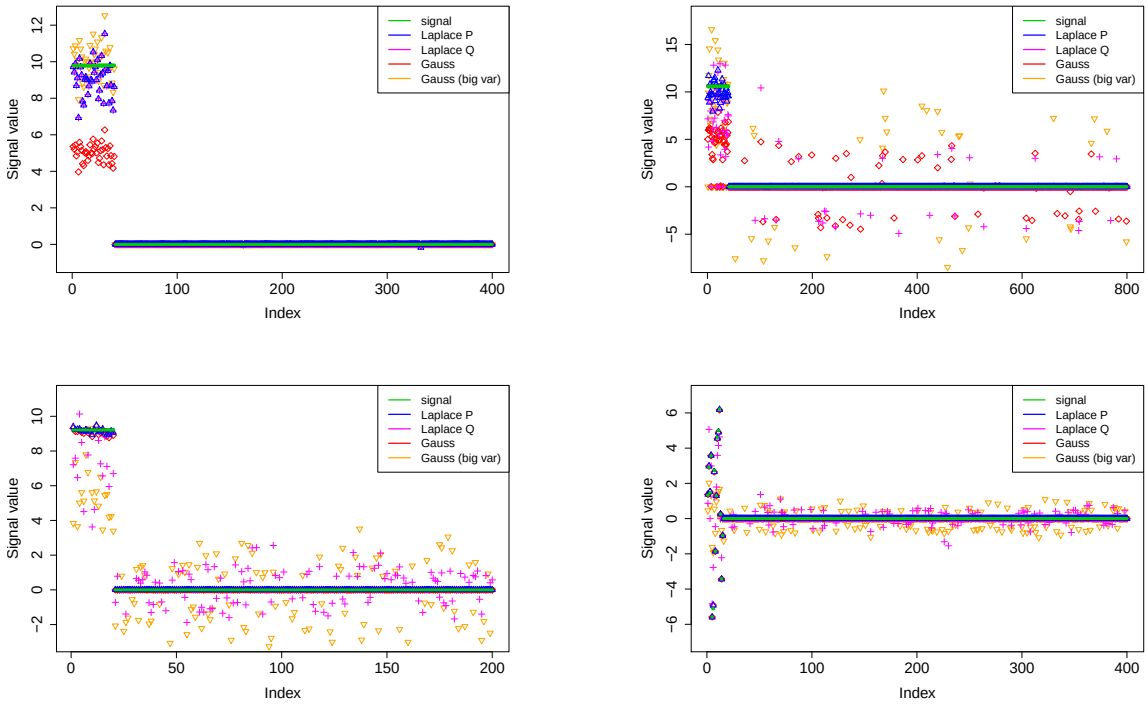

Figure 3: Linear regression with (i) identity design $X = I_n$ and (ii)-(iv) Gaussian design $X_{ij} \overset{iid}{\sim} N(0, \tau^2)$. We plot the underlying signal with non-zero components $\theta_i = A$ (green) and the posterior means of $\widetilde{\Pi}$ (blue), $\widetilde{Q}$ (purple), VB with Gaussian slabs (red) and VB with rescaled Gaussian slabs (orange). From left to right and top to bottom, we set the parameters (n, p, s, A): (i) $(400, 400, 40, 4\sqrt{\log n})$; (ii) $(100, 200, 20, 2\log n)$; (iii) $(200, 800, 40, 2\log n)$; (iv) $(100, 400, 15, U(-8, 8))$. We set (ii) $\tau = 1$; (iii) $\tau = 0.1$; (iv) $\tau = 0.5$.

Metric	Method	(i) $N(0,1)$	(ii) Lap$(0,1)$	(iii) $U(-2,2)$	(iv) Student t_3
ℓ_2-error	sparsevb	0.18 ± 0.05	**0.24 ± 0.04**	**0.21 ± 0.03**	0.30 ± 0.06
	varbvs	**0.17 ± 0.03**	**0.24 ± 0.04**	**0.21 ± 0.03**	0.30 ± 0.06
	EMVS	0.59 ± 0.03	1.03 ± 0.14	0.89 ± 0.16	1.13 ± 0.43
	SSLASSO	5.99 ± 0.98	4.07 ± 1.02	4.88 ± 0.62	4.87 ± 0.78
	ebreg	0.26 ± 0.05	0.26 ± 0.07	0.23 ± 0.05	**0.23 ± 0.05**
FDR	sparsevb	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	varbvs	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.01**
	EMVS	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.01**
	SSLASSO	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	ebreg	0.01 ± 0.02	0.01 ± 0.05	0.01 ± 0.05	0.01 ± 0.03
TPR	sparsevb	**1.00 ± 0.01**	**1.00 ± 0.00**	**0.95 ± 0.00**	**0.90 ± 0.01**
	varbvs	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.95 ± 0.01**	**0.90 ± 0.01**
	EMVS	0.95 ± 0.02	0.92 ± 0.02	0.89 ± 0.02	0.81 ± 0.04
	SSLASSO	0.67 ± 0.04	0.72 ± 0.05	0.64 ± 0.02	0.64 ± 0.04
	ebreg	**1.00 ± 0.01**	**1.00 ± 0.00**	**0.95 ± 0.01**	**0.90 ± 0.01**
runtime (sec)	sparsevb	0.52 ± 0.28	0.51 ± 0.18	0.44 ± 0.07	0.61 ± 0.35
	varbvs	0.57 ± 0.22	0.61 ± 0.31	0.45 ± 0.08	0.47 ± 0.10
	EMVS	2.17 ± 0.82	2.41 ± 1.06	1.64 ± 0.20	1.74 ± 0.23
	SSLASSO	**0.31 ± 0.15**	0.22 ± 0.08	**0.24 ± 0.05**	**0.28 ± 0.07**
	ebreg	29.37 ± 7.10	24.89 ± 3.78	27.72 ± 4.51	28.19 ±4.51

Table 6: Noise misspecification: we compare the robustness of Bayesian model selection methods under misspecified noise. We take Gaussian design $X_{ij} \overset{iid}{\sim} N(0, 2^2)$, set the model parameters $n = 200$, $p = 400$, $s = 20$, and take non-zero coefficients $\theta_i \overset{iid}{\sim} U(-10, 10)$ located in the beginning of the signal. We ran each experiment 200 times and report the means and standard deviations.

A.4 Noise misspecification

We investigate the robustness of the Bayesian model selection methods to misspecification of the noise distribution in practice. Note that our theoretical results are also robust to some misspecification, see Remark B.1 in Section B below. We consider Gaussian design $X_{ij} \overset{iid}{\sim} N(0, 2^2)$, set the model parameters $n = 200$, $p = 400$, $s = 20$, and take non-zero signal coefficients $\theta_i \overset{iid}{\sim} U(-10, 10)$ located in the beginning of θ. We compare the correctly-specified Gaussian noise case (i) $Z_i \overset{iid}{\sim} N(0, 1)$ in model (1) with the misspecified noise cases: (ii) Laplace noise $Z_i \overset{iid}{\sim} \text{Lap}(0, 1)$; (iii) uniform noise $Z_i \overset{iid}{\sim} U(-2, 2)$; (iv) Student noise with 3 degrees of freedom $Z_i \overset{iid}{\sim} t_3$. We apply the same parametrizations of the methods as in Section 5.2. We ran each experiments 200 times and collect the results in Table 6. Our method (sparsevb) gave similar results to varbvs, ebreg and EMVS, while the SSLASSO performed slightly worse. The noise distribution does not seem to have a major effect on the results, hence these algorithms seem robust to noise misspecification. It is worthwhile to further investigate this phenomenon both empirically and analytically.

A.5 Bayesian variable selection methods under correlated inputs

We lastly consider the common situation of correlated input variables. We take each row $X_{i\cdot} \overset{iid}{\sim} N_p(0, \Sigma)$ with $\Sigma_{jk} = \rho$ for $j \neq k$ and $\Sigma_{jj} = 1$, giving standard normal predictors with non-zero correlation ρ. We take (i) $(n, p, s, \varsigma) = (100, 400, 10, 0.2)$, correlation $\rho = 0.3$ and

Metric	Method	(i)	(ii)	(iii)	(iv)
$\ell_2 -$ error	sparsevb	**0.12 ± 0.06**	0.89 ± 1.40	**1.97 ± 0.37**	**4.85 ± 1.29**
	varbvs	0.13 ± 0.06	**0.30 ± 0.10**	2.10 ± 0.43	27.18 ± 23.59
	EMVS	4.80 ± 0.21	5.29 ± 0.26	4.04 ± 0.30	7.04 ± 0.98
	SSLASSO	1.62 ± 0.35	0.97 ± 0.36	56.70 ± 7.78	79.17 ± 4.95
	ebreg	0.34 ± 0.06	0.56 ± 0.14	5.41 ± 0.67	6.41 ± 1.21
FDR	sparsevb	**0.00 ± 0.00**	0.18 ± 0.34	**0.00 ± 0.01**	**0.00 ± 0.00**
	varbvs	**0.00 ± 0.00**	**0.00 ± 0.00**	0.01 ± 0.02	0.31 ± 0.26
	EMVS	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.01**	0.14 ± 0.08
	SSLASSO	**0.00 ± 0.00**	**0.00 ± 0.00**	0.18 ± 0.16	0.41 ± 0.19
	ebreg	**0.00 ± 0.00**	**0.00 ± 0.00**	0.43 ± 0.05	0.28 ± 0.08
TPR	sparsevb	**0.96 ± 0.05**	0.95 ± 0.10	**1.00 ± 0.00**	**1.00 ± 0.00**
	varbvs	0.95 ± 0.05	**1.00 ± 0.00**	**1.00 ± 0.00**	0.69 ± 0.32
	EMVS	0.01 ± 0.03	0.02 ± 0.04	**1.00 ± 0.00**	**1.00 ± 0.00**
	SSLASSO	0.48 ± 0.04	0.81 ± 0.08	0.34 ± 0.10	0.18 ± 0.05
	ebreg	0.90 ± 0.01	0.96 ± 0.05	**1.00 ± 0.00**	**1.00 ± 0.00**
runtime (sec)	sparsevb	0.64 ± 0.26	0.51 ± 0.20	1.84 ± 2.00	2.61 ± 0.54
	varbvs	0.84 ± 0.24	1.53 ± 0.60	21.16 ± 22.46	77.39 ± 14.84
	EMVS	**0.23 ± 0.23**	**0.23 ± 0.14**	1.20 ± 1.27	1.21 ± 0.08
	SSLASSO	0.82 ± 0.26	0.40 ± 0.12	**0.20 ± 0.11**	**0.20 ± 0.03**
	ebreg	13.90 ± 1.62	15.06 ± 2.96	74.14 ± 6.36	65.98 ± 4.41

Table 7: Linear regression with correlated Gaussian design $X_{i\cdot} \overset{iid}{\sim} N_p(0, \Sigma)$, with correlation $\Sigma_{jk} = \rho$ for $j \neq k$ and $\Sigma_{jj} = 1$. The noise variance ς^2 is unknown and the non-zero signal coefficients equal $\theta_i = A$. We take the parameters $(n, p, s, A, \rho, \varsigma)$ equal to (i) $(100, 400, 10, \overset{iid}{\sim} U(-3, 3), 0.3, 0.2)$ (non-zero coefficients at the beginning); (ii) $(100, 400, 10, \overset{iid}{\sim} U(-3, 3), 0.7, 0.2)$ (at the beginning); (iii) $(200, 800, 20, 2\log n, 0.3, 5)$ (at the end); (iv) $(200, 800, 20, 2\log n, 0.7, 5)$ (at the end). We compare the means and standard deviations over 100 runs.

non-zero coefficients $\theta_i \overset{iid}{\sim} U(-3, 3)$ at the beginning of the signal; (ii) the same setting as in (i), but with higher correlation $\rho = 0.7$; (iii) $(n, p, s, \varsigma) = (200, 800, 20, 5)$, correlation $\rho = 0.3$ and non-zero coefficients $\theta_i = 2\log n$ at the end of the signal; (iv) the same setting as in (iii), but with higher correlation $\rho = 0.7$. We apply the same parametrizations of the methods as in Section 5.2. The results are summarized in Table 7.

One might expect that mean-field VB methods should not perform so well under correlated inputs due to their factorizable structure. This was not the case in our simulations, where the VB methods perform competitively with the other methods, often providing the best results (except perhaps in (iv), where varbvs sometimes sometimes gave large ℓ_2 error). The correlated design also does not seem to substantially influence the run time.

While our simulations are certainly not extensive, they suggest that mean-field VB can perhaps still be effective in certain correlated input settings and understanding the exact effect of correlation on VB seems to be a subtle question. It is currently not well understood how VB, or indeed even the true posterior, behaves in general correlated design settings. This important and practically very relevant setting requires further investigation, both theoretically and empirically.

B Proofs

B.1 Full oracle results

The proofs of the full oracle results in Theorems B.1 and B.2 below rely on Theorem 5, which allows one to exploit exponential probability bounds for the posterior to control the corresponding probability under the variational approximation. To prove our results, it therefore suffices to show that on a suitable event, one can (a) control the KL divergence between the variational approximation and the true posterior and (b) establish the appropriate posterior tail inequality (14). Part (a) is dealt with in Section B.2 and (b) in Section B.3 below. Define the events

$$\mathcal{T}_0 = \{\|X^T(Y - X\theta_0)\|_\infty \leq 2\|X\|\sqrt{\log p}\} \tag{B.1}$$

and

$$\mathcal{T}_1 = \mathcal{T}_1(\Gamma, \varepsilon, \kappa) = \mathcal{T}_0 \cap \Big\{\Pi\big(\theta : |S_\theta| > \Gamma \big| Y\big) \leq 1/4\Big\} \cap \Big\{\Pi\big(\theta : \|\theta - \theta_0\|_2 > \varepsilon | Y\big) \leq e^{-\kappa}\Big\}, \tag{B.2}$$

for $\Gamma, \varepsilon, \kappa > 0$. The middle event in $\mathcal{T}_1$ says that the posterior puts most of its mass on models of dimension at most Γ; the number $1/4$ is unimportant and any number less than $1/2$ suffices. The third event says the posterior places all but exponentially small probability on an ℓ_2-ball of radius ε about the truth and is used for a localization argument when bounding the KL divergence. The proof uses an iterative structure, using successive posterior localizations to eventually bound the KL divergence in Section B.2. This idea is a useful technique from Bayesian nonparametrics, see e.g. [11].

For parameters $\theta_0, \theta_* \in \mathbb{R}^p$, set $S_* = S_{\theta_*}$ and $s_* = |S_*|$ and define

$$\Delta_* = (1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}})s_* \log p + \|X(\theta_0 - \theta_*)\|_2^2. \tag{B.3}$$

This quantity appears in the posterior exponential probabilities, which take the form $e^{-c\Delta_*}$. We require the following parameter choices for the event $\mathcal{T}_1$ in (B.2):

$$\begin{aligned}
\Gamma = \Gamma_{\theta_0,\theta_*} &= s_* + \frac{12}{A_4}\left(1 + \frac{16}{\phi(S_*)^2}\frac{\lambda}{\bar{\lambda}}\right)s_* + \frac{12\|X(\theta_0 - \theta_*)\|_2^2}{A_4 \log p} = s_* + \frac{12\Delta_*}{A_4 \log p}, \\
\varepsilon = \varepsilon_{\theta_0,\theta_*} &= \frac{ML_0^{1/2}}{\|X\|\widetilde{\psi}_{L_0+2}(S_0)^2}\left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right], \\
\kappa = \kappa_{\theta_0,\theta_*} &= (\Gamma_{\theta_0,\theta_*} + 1)\log p, \\
L_0 &= \max(3 + 12/A_4, 2 + A_4/2)
\end{aligned} \tag{B.4}$$

for some $M > 0$ large enough depending only on A_1, A_3, A_4.

Lemma B.1. *(i) The event $\mathcal{T}_0$ defined in* (B.1) *satisfies*

$$\inf_{\theta_0 \in \mathbb{R}^p} P_{\theta_0}(\mathcal{T}_0) \geq 1 - 2/p.$$

(ii) Suppose the prior satisfies (4) *and* (5)*. For $\theta_0 \in \mathbb{R}^p \backslash \{0\}$, let $\theta_* \in \mathbb{R}^p$ be any vector satisfying $1 \leq s_* = |S_{\theta_*}| \leq |S_{\theta_0}| = s_0$,*

$$\frac{s_*}{\phi(S_*)^2} \leq \frac{s_0}{\phi(S_0)^2} \quad \text{and} \quad \|X(\theta_0 - \theta_*)\|_2^2 \leq (s_0 - s_*)\log p.$$

Then the event $\mathcal{T}_1$ given in (B.2) *with parameters $\Gamma, \varepsilon, \kappa$ chosen according to* (B.4) *satisfies*

$$P_{\theta_0}(\mathcal{T}_1) \to 1$$

uniformly over all θ_0 and θ_* as above.

Proof. (i) Under P_{θ_0}, $X^T(Y - X\theta_0) = X^T Z \sim N_p(0, X^T X)$. Since $(X^T Z)_i \sim N(0, (X^T X)_{ii})$ and $(X^T X)_{ii} \le \|X\|^2$ for all $1 \le i \le p$, a union bound and the standard Gaussian tail inequality give

$$P_{\theta_0}(\mathcal{T}_0^c) = P(\|X^T Z\|_\infty \ge 2\|X\|\sqrt{\log p}) \le \sum_{i=1}^{p} P(|N(0,1)| \ge 2\sqrt{\log p}) \le p\frac{2}{\sqrt{2\pi}} e^{-2\log p}.$$

(ii) Applying Markov's inequality and Lemma B.5 below with $M = 3$ gives

$$\begin{aligned}
&P_{\theta_0}\left(\left\{\Pi\big(\theta : |S_\theta| > \Gamma_{\theta_0,\theta_*}|Y\big) > 1/4\right\} \cap \mathcal{T}_0\right) \\
&\quad \le 4E_{\theta_0}\Pi\big(\theta : |S_\theta| > \Gamma_{\theta_0,\theta_*}|Y\big)1_{\mathcal{T}_0} \\
&\quad \le C(A_2, A_4)\exp\left(-\left(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}}\right)s_* \log p\right) \\
&\quad \le C(A_2, A_4)e^{-s_* \log p} \le C(A_2, A_4)e^{-\log p}.
\end{aligned}$$

Since the right-hand side does not depend on θ_0 or θ_*, the probability tends to zero uniformly as required.

Under the assumptions on θ_*,

$$\begin{aligned}
(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}})s_* \log p + \|X(\theta_0 - \theta_*)\|_2^2 &\le s_* \log p + \tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\bar{\lambda}} s_0 \log p + (s_0 - s_*)\log p \\
&= (1 + \tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\bar{\lambda}})s_0 \log p.
\end{aligned} \tag{B.5}$$

Therefore, applying Lemma B.6 with $L \ge 1$ yields

$$\begin{aligned}
&E_{\theta_0}\Pi\left(\theta : \|\theta - \theta_0\|_2 > \frac{ML^{1/2}}{\|X\|\overline{\psi}_{L+2}(S_0)^2}\left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right]\Big|Y\right)1_{\mathcal{T}_0}, \\
&\le C\exp\left(-\left[L \wedge \tfrac{4(L+2)}{A_4}\right]\left[(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}})s_* \log p + \|X(\theta_0 - \theta_*)\|_2^2\right]\right).
\end{aligned}$$

Using Markov's inequality and the last display with $L = L_0 = \max(3 + 12/A_4, 2 + A_4/2)$,

$$\begin{aligned}
&P_{\theta_0}\left(\left\{\Pi\big(\theta : \|\theta - \theta_0\|_2 > \varepsilon|Y\big) > e^{-\kappa}\right\} \cap \mathcal{T}_0\right) \\
&\le e^{\kappa}E_{\theta_0}\Pi\big(\theta : \|\theta - \theta_0\|_2 > \varepsilon|Y\big)1_{\mathcal{T}_0} \\
&\le C\exp\left(-\left[L \wedge \tfrac{4(L+2)}{A_4} - \tfrac{12}{A_4}\right]\left[(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}})s_* \log p + \|X(\theta_0 - \theta_*)\|_2^2\right] + (s_* + 1)\log p\right) \\
&\le Ce^{-s_* \log p} \le Ce^{-\log p}.
\end{aligned}$$

Since the right-hand side again does not depend on θ_0 or θ_*, the probability tends to zero uniformly as required. □

Theorem B.1 (Full oracle recovery)**.** *Suppose the model selection prior* (3) *satisfies* (4) *and* (5)*. For $\theta_0 \in \mathbb{R}^p \backslash \{0\}$, let $\theta_* \in \mathbb{R}^p$ be any vector satisfying $1 \le s_* = |S_{\theta_*}| \le |S_{\theta_0}| = s_0$ and*

$\|X(\theta_0-\theta_*)\|_2^2 \le (s_0-s_*)\log p$. *Then the variational Bayes posterior* $\widetilde{\Pi}$ *satisfies, uniformly over all* θ_0 *and* θ_* *as above,*

$$E_{\theta_0}\widetilde{\Pi}\left(\theta : \|X(\theta-\theta_0)\|_2 \ge \frac{M\rho_n^{1/2}}{\overline{\psi}_{\rho_n}(S_0)}\left[\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0-\theta_*)\|_2\right]\right)$$
$$\lesssim \frac{1}{\rho_n}\left\{1 + \frac{\log(1/\widetilde{\phi}(\Gamma))}{\log p} + \frac{\lambda s_0}{\|X\|\widetilde{\psi}_{L_0+2}(S_0)^2\phi(S_0)\widetilde{\phi}(\Gamma)^2\sqrt{\log p}}\right\} + o(1)$$

for any $\rho_n > 2$, *where* Γ, L_0 *are given in* (B.4). *Moreover, both*

$$E_{\theta_0}\widetilde{\Pi}\left(\theta : \|\theta-\theta_0\|_1 > \|\theta_0-\theta_*\|_1 + \frac{M\rho_n}{\overline{\psi}_{\rho_n}(S_0)^2}\left[\frac{s_*\sqrt{\log p}}{\|X\|\phi(S_*)^2} + \frac{\|X(\theta_0-\theta_*)\|_2^2}{\|X\|\sqrt{\log p}}\right]\right),$$

$$E_{\theta_0}\widetilde{\Pi}\left(\theta : \|\theta-\theta_0\|_2 > \frac{M\rho_n^{1/2}}{\|X\|\widetilde{\psi}_{\rho_n}(S_0)^2}\left[\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0-\theta_*)\|_2\right]\right),$$

satisfy the same inequality. Furthermore, the exact same inequalities hold for the variational Bayes posteriors $\widetilde{Q}$ *and* $\hat{Q}$.

Proof. Suppose first that $s_*/\phi(S_*)^2 \le s_0/\phi(S_0)^2$. Let $\mathcal{T}_1$ denote the event in (B.2) with parameters (B.4), which by Lemma B.1(ii) satisfies $P_{\theta_0}(\mathcal{T}_1) \to 1$ uniformly over all θ_0, θ_* in the theorem hypothesis. Set

$$\Theta_n = \left\{\theta : \|X(\theta-\theta_0)\|_2 \ge \frac{M\rho_n^{1/2}}{\overline{\psi}_{\rho_n}(S_0)}\left[\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0-\theta_*)\|_2\right]\right\}$$

and note $E_{\theta_0}\widetilde{\Pi}(\Theta_n) \le E_{\theta_0}\widetilde{\Pi}(\Theta_n)1_{\mathcal{T}_1} + o(1)$. We now apply Theorem 5 with this choice of Θ_n on the event $\mathcal{T}_1$. For Δ_* defined in (B.3), it holds that $\Delta_* \le (1 + \frac{16}{\phi(S_0)^2}\frac{\lambda}{\bar{\lambda}})s_0\log p$ by (B.5). Using Lemma B.6 below with $L+2 = \rho_n$ thus gives

$$E_{\theta_0}\Pi(\Theta_n|Y)1_{\mathcal{T}_0} \le Ce^{-c\rho_n\Delta_*},$$

for p large enough depending on A_1, A_3, A_4, and where $C, c > 0$ also depend only on the prior parameters. Since $\mathcal{T}_1 \subset \mathcal{T}_0$ by (B.2), condition (14) is satisfied on $\mathcal{T}_1$ with $\delta_n = c\rho_n\Delta_*$. Applying Theorem 5 gives

$$E_{\theta_0}\widetilde{\Pi}(\Theta_n)1_{\mathcal{T}_1} \le \tfrac{2}{c\rho_n\Delta_*}\mathrm{KL}(\widetilde{\Pi}\|\Pi(\cdot|Y))1_{\mathcal{T}_1} + o(1).$$

Note that the parameters (B.4) satisfy $\Gamma\log p \lesssim \Delta_*$ and $\varepsilon \lesssim \frac{\sqrt{s_0\log p}}{\|X\|\widetilde{\psi}_{L_0+2}(S_0)^2\phi(S_0)}$. Using this and Lemma B.4 below,

$$\tfrac{2}{c\rho_n\Delta_*}\mathrm{KL}(\widetilde{\Pi}\|\Pi(\cdot|Y))1_{\mathcal{T}_1} \lesssim \frac{1}{\rho_n}\left\{1 + \frac{\log(1/\widetilde{\phi}(\Gamma))}{\log p} + \frac{\lambda s_0}{\|X\|\widetilde{\psi}_{L_0+2}(S_0)^2\phi(S_0)\widetilde{\phi}(\Gamma)^2\sqrt{\log p}}\right\} + o(1)$$

as required.

If $s_*/\phi(S_*)^2 > s_0/\phi(S_0)^2$, then $\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0-\theta_*)\|_2 > \frac{\sqrt{s_0\log p}}{\phi(S_0)}$. The desired inequality then immediately follows from the stronger inequality with $\theta_* = \theta_0$ just established above.

The results for ℓ_1 and ℓ_2 loss follow exactly as above by using the respective inequalities for the ℓ_1 and ℓ_2 oracle contraction rates in Lemma B.6 to establish (14).

Similarly, the results for the variational Bayes posteriors $\hat{Q}$ and $\widetilde{Q}$ based on the mean-field variational families (9) and (10) follow identically upon using Lemmas B.2 and B.3 instead of Lemma B.4 to control the Kullback-Leibler divergence. □

Theorem B.2 (Full oracle dimension)**.** *Suppose the model selection prior* (3) *satisfies* (4) *and* (5)*. For $\theta_0 \in \mathbb{R}^p \backslash \{0\}$, let $\theta_* \in \mathbb{R}^p$ be any vector satisfying $1 \leq s_* = |S_{\theta_*}| \leq |S_{\theta_0}| = s_0$ and $\|X(\theta_0 - \theta_*)\|_2^2 \leq (s_0 - s_*) \log p$. Then the variational Bayes posterior $\widetilde{\Pi}$ satisfies, uniformly over all θ_0 and θ_* as above,*

$$E_{\theta_0} \widetilde{\Pi} \left(\theta : |S_\theta| \geq |S_*| + \tfrac{4(\rho_n+2)}{A_4} \left[\left(1 + \tfrac{16}{\phi(S_*)^2} \tfrac{\lambda}{\bar{\lambda}} \right) |S_*| + \tfrac{\|X(\theta_0 - \theta_*)\|_2^2}{\log p} \right] \right)$$
$$\lesssim \frac{1}{\rho_n} \left\{ 1 + \frac{\log(1/\widetilde{\phi}(\Gamma))}{\log p} + \frac{\lambda s_0}{\|X\| \widetilde{\psi}_{L_0+2}(S_0)^2 \phi(S_0) \widetilde{\phi}(\Gamma)^2 \sqrt{\log p}} \right\} + o(1)$$

for any $\rho_n > 0$, where Γ, L_0 are given in (B.4)*. Furthermore, the exact same inequality holds for the variational Bayes posteriors $\widetilde{Q}$ and $\hat{Q}$.*

Proof. The proof follows similarly to that of Theorem B.1 by applying Theorem 5 with

$$\Theta_n = \left\{ \theta : |S_\theta| \geq |S_*| + \tfrac{4(\rho_n+2)}{A_4} \left[\left(1 + \tfrac{16}{\phi(S_*)^2} \tfrac{\lambda}{\bar{\lambda}} \right) |S_*| + \tfrac{\|X(\theta_0 - \theta_*)\|_2^2}{\log p} \right] \right\},$$

again taking the event $A = \mathcal{T}_1$ and using Lemma B.5 with $M = \rho_n + 2$ instead of Lemma B.6 to verify (14). □

Remark B.1 (Misspecification of the error distribution)**.** *The Gaussian error distribution is assumed in model* (1) *for concreteness and can be relaxed. For recovery and dimension control (Theorems 1 and 2), inspection of the contraction rate proofs in [5] and the KL bounds in Section B.2 show that it suffices that there exists a constant $C > 0$ such that*

$$P_{\theta_0}(\|X^T(Y - X\theta_0)\|_\infty > C\|X\|\sqrt{\log p}) \to 0,$$

which holds for much more general noise distributions. This condition is commonly imposed when studying the LASSO, see e.g. [2]. For the full oracle bounds, we further need that Lemma 3 of [5], which concerns a change of measure, holds. This indeed holds under a wider range of noise distributions, see Remark 1 of [5]. The results for VB in this paper are thus robust under noise misspecification as for the true posterior [5], see also Section A.4 for an empirical study of noise misspecification for our method.

B.2 Kullback-Leibler divergences between variational classes and the posterior

We now show that on the event $\mathcal{T}_1$ in (B.2), we can bound the (minimized) Kullback-Leibler divergences between the posterior and the approximating variational classes. In particular, we need oracle-type bounds on the KL divergence to obtain our oracle results. This is the major technical difficulty in establishing our result. We first consider the family $\mathcal{Q}$ of distributions (9), which consists of products of non-diagonal multivariate normal distributions with Dirac delta distributions for a single fixed support set S.

For a given model $S \subseteq \{1, \ldots, p\}$, let X_S denote the $n \times |S|$-submatrix of the full regression matrix X, where we keep only the columns $X_{\cdot i}$, $i \in S$. Let $\hat{\theta}_S = (X_S^T X_S)^{-1} X_S^T Y$ be the least squares estimator in the restricted model $Y = X_S \theta_S + Z$. If the restricted model were correctly specified, then $\hat{\theta}_S$ would have distribution $N_S(\theta_{0,S}, (X_S^T X_S)^{-1})$ under P_{θ_0}. We approximate the posterior with a $N_S(\hat{\theta}_S, (X_S^T X_S)^{-1}) \otimes \delta_{S^c}$ distribution, where S is a suitable approximating set to which the posterior assigns sufficient probability.

Lemma B.2. *If $4e^{1+\Gamma \log p - \kappa} \leq 1$, then the variational posterior $\hat{Q}$ arising from the family (9) satisfies*

$$\mathrm{KL}(\hat{Q} \| \Pi(\cdot|Y)) 1_{\mathcal{T}_1} \leq \Gamma \log p + \frac{\lambda \Gamma}{\widetilde{\phi}(\Gamma)^2} \left(2 s_0^{1/2} \varepsilon + \frac{3\sqrt{\log p}}{\|X\|} \right) + \log(4e).$$

Proof. We construct our posterior approximation on the event $\mathcal{T}_1$ in (B.2). The posterior takes the form

$$\Pi(\cdot|Y) = \sum_{S \subseteq \{1,\ldots,p\}} \hat{q}_S \Pi_S(\cdot|Y) \otimes \delta_{S^c}, \tag{B.6}$$

where the weights $\hat{q} = (\hat{q}_S : S \subseteq \{1, ..., p\})$ lie in the 2^p-dimensional simplex and $\Pi_S(\cdot|Y)$ is the posterior for $\theta_S \in \mathbb{R}^{|S|}$ in the restricted model $Y = X_S \theta_S + Z$. Since

$$\Pi(\theta : \|\theta_{0,S_\theta^c}\|_2 > \varepsilon | Y) \leq \Pi(\theta : \|\theta - \theta_0\|_2 > \varepsilon | Y),$$

it follows that on $\mathcal{T}_1$,

$$\sum_{\substack{S : |S| \leq \Gamma \\ \|\theta_{0,S^c}\|_2 \leq \varepsilon}} \hat{q}_S \geq 1 - \frac{1}{4} - e^{-\kappa} \geq \frac{3}{4} - \frac{1}{4e} e^{-\Gamma \log p} \geq \frac{1}{2}$$

for all p since $\Gamma > 0$. Note further that

$$\Big| \big\{ S \subseteq \{1, ..., p\} : |S| \leq \Gamma \big\} \Big| = \sum_{s=0}^{\lfloor \Gamma \rfloor} \binom{p}{s} \leq \sum_{s=0}^{\lfloor \Gamma \rfloor} \frac{p^s}{s!} \leq e p^\Gamma.$$

Together, the last two displays show that on $\mathcal{T}_1$ and for all p, there exists a set $\tilde{S}$ satisfying

$$|\tilde{S}| \leq \Gamma, \quad \|\theta_{0,\tilde{S}^c}\|_2 \leq \varepsilon, \quad \hat{q}_{\tilde{S}} \geq (2e)^{-1} p^{-\Gamma}. \tag{B.7}$$

Since an $N_S(\mu_S, \Sigma_S) \otimes \delta_{S^c}$ distribution is only absolutely continuous with respect to the $\hat{q}_S \Pi_S(\cdot|Y) \otimes \delta_{S^c}$ term of the posterior (B.6),

$$\begin{aligned} \inf_{Q \in \mathcal{Q}} \mathrm{KL}(Q || \Pi(\cdot|Y)) &= \inf_{S, \mu_S, \Sigma_S} E_{\theta \sim N_S(\mu_S, \Sigma_S) \otimes \delta_{S^c}} \log \frac{dN_S(\mu_S, \Sigma_S) \otimes \delta_{S^c}}{\hat{q}_S d\Pi_S(\cdot|Y) \otimes \delta_{S^c}} \\ &\leq \log \frac{1}{\hat{q}_{\tilde{S}}} + \inf_{\mu_{\tilde{S}}, \Sigma_{\tilde{S}}} \mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}}) \| \Pi_{\tilde{S}}(\cdot|Y) \big), \end{aligned} \tag{B.8}$$

where the last Kullback-Leibler divergence is over $|\tilde{S}|$-dimensional distributions. On $\mathcal{T}_1$, $\log(1/\hat{q}_{\tilde{S}}) \leq \log(2ep^\Gamma) = \log(2e) + \Gamma \log p$. It thus remains to bound the second term in (B.8).

Let E_{μ_S, Σ_S} denote the expectation under the law $\theta_S \sim N_S(\mu_S, \Sigma_S)$. Setting

$$\mu_{\tilde{S}} = (X_{\tilde{S}}^T X_{\tilde{S}})^{-1} X_{\tilde{S}}^T Y \quad \text{and} \quad \Sigma_{\tilde{S}} = (X_{\tilde{S}}^T X_{\tilde{S}})^{-1}, \tag{B.9}$$

one can check that the resulting normal distribution has density function proportional to $e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2}$, $\theta_{\tilde{S}} \in \mathbb{R}^{|\tilde{S}|}$. Therefore,

$$\begin{aligned}\mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}}) \| \Pi_{\tilde{S}}(\cdot|Y)\big) &= E_{\mu_{\tilde{S}},\Sigma_{\tilde{S}}} \log \frac{D_\Pi e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{0,\tilde{S}}\|_1}}{D_N e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1}} \\ &= E_{\mu_{\tilde{S}},\Sigma_{\tilde{S}}} \lambda(\|\theta_{\tilde{S}}\|_1 - \|\theta_{0,\tilde{S}}\|_1) + \log(D_\Pi/D_N),\end{aligned} \tag{B.10}$$

with $D_\Pi = \int_{\mathbb{R}^{|\tilde{S}|}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1} d\theta_{\tilde{S}}$ and $D_N = \int_{\mathbb{R}^{|\tilde{S}|}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{0,\tilde{S}}\|_1} d\theta_{\tilde{S}}$ the normalizing constants.

We firstly upper bound $\log(D_\Pi/D_N)$. Define

$$B_{\tilde{S}} = \{\theta_{\tilde{S}} \in \mathbb{R}^{|\tilde{S}|} : \|\theta_{\tilde{S}} - \theta_{0,\tilde{S}}\|_2 \le 2\varepsilon\}.$$

Let $\bar{\theta}_{\tilde{S}}$ denote the extension of a vector $\theta_{\tilde{S}} \in \mathbb{R}^{|\tilde{S}|}$ to $\mathbb{R}^p$ with $\bar{\theta}_{\tilde{S},j} = \theta_{\tilde{S},j}$ for $j \in \tilde{S}$ and $\bar{\theta}_{\tilde{S},j} = 0$ for $j \notin \tilde{S}$. On $\mathcal{T}_1$, using (B.6) and (B.7),

$$\begin{aligned}\Pi_{\tilde{S}}(B_{\tilde{S}}^c|Y) &\le \frac{\hat{q}_{\tilde{S}}}{\hat{q}_{\tilde{S}}} \Pi_{\tilde{S}}(\theta_{\tilde{S}} \in \mathbb{R}^{|\tilde{S}|} : \|\bar{\theta}_{\tilde{S}} - \theta_0\|_2 > 2\varepsilon - \|\theta_{0,\tilde{S}^c}\|_2|Y) \\ &\le \hat{q}_{\tilde{S}}^{-1} \Pi(\theta \in \mathbb{R}^p : \|\theta - \theta\|_2 > \varepsilon|Y) \\ &\le 2ep^\Gamma e^{-\kappa} = 2e^{1+\Gamma\log p - \kappa} \le 1/2,\end{aligned}$$

where the last inequality holds by assumption. Using Bayes formula, this yields

$$\Pi_{\tilde{S}}(B_{\tilde{S}}|Y) 1_{\mathcal{T}_1} = \frac{\int_{B_{\tilde{S}}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1} d\theta_{\tilde{S}}}{\int_{\mathbb{R}^{|\tilde{S}|}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1} d\theta_{\tilde{S}}} 1_{\mathcal{T}_1} \ge \frac{1}{2} 1_{\mathcal{T}_1}$$

almost surely. In particular, $D_\Pi \le 2\int_{B_{\tilde{S}}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1} d\theta_{\tilde{S}}$ on $\mathcal{T}_1$. Therefore on $\mathcal{T}_1$,

$$\begin{aligned}\log \frac{D_\Pi}{D_N} &\le \log \frac{2\int_{B_{\tilde{S}}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{\tilde{S}}\|_1} d\theta_{\tilde{S}}}{\int_{B_{\tilde{S}}} e^{-\frac{1}{2}\|Y-X_{\tilde{S}}\theta_{\tilde{S}}\|_2^2 - \lambda\|\theta_{0,\tilde{S}}\|_1} d\theta_{\tilde{S}}} \\ &\le \sup_{\theta_{\tilde{S}} \in B_{\tilde{S}}} \log e^{\lambda\|\theta_{0,\tilde{S}}\|_1 - \lambda\|\theta_{\tilde{S}}\|_1} + \log 2 \\ &\le \sup_{\theta_{\tilde{S}} \in B_{\tilde{S}}} \lambda\|\theta_{\tilde{S}} - \theta_{0,\tilde{S}}\|_1 + \log 2 \\ &\le \sup_{\theta_{\tilde{S}} \in B_{\tilde{S}}} \lambda|\tilde{S}|^{1/2}\|\theta_{\tilde{S}} - \theta_{0,\tilde{S}}\|_2 + \log 2 \\ &\le 2\lambda\Gamma^{1/2}\varepsilon + \log 2,\end{aligned}$$

where in the fourth inequality we have applied Cauchy-Schwarz.

We now turn to the first term in (B.10). On $\mathcal{T}_1$, using the triangle inequality and Cauchy-Schwarz,

$$\begin{aligned}\lambda E_{\mu_{\tilde{S}},\Sigma_{\tilde{S}}}(\|\theta_{\tilde{S}}\|_1 - \|\theta_{0,\tilde{S}}\|_1) &\le \lambda\|\mu_{\tilde{S}} - \theta_{0,\tilde{S}}\|_1 + \lambda E_{0,\Sigma_{\tilde{S}}}\|\theta_{\tilde{S}}\|_1 \\ &\le \lambda\Gamma^{1/2}\big(\|\mu_{\tilde{S}} - \theta_{0,\tilde{S}}\|_2 + \mathrm{Tr}(\Sigma_{\tilde{S}})^{1/2}\big)\end{aligned} \tag{B.11}$$

since $E_{0,\Sigma_{\tilde{S}}}\|\theta_{\tilde{S}}\|_2^2 = \mathrm{Tr}(\Sigma_{\tilde{S}})$. Let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ denote the smallest and largest eigenvalues, respectively, of a symmetric, positive definite matrix A. Using the variational characterization of maximal/minimal eigenvalues ([8], p. 234), for any $S \subseteq \{1, \ldots, p\}$,

$$\Lambda_{\min}(X_S^T X_S) = \min_{v \in \mathbb{R}^{|S|}: v \neq 0} \frac{v^T X_S^T X_S v}{\|v\|_2^2} = \min_{u \in \mathbb{R}^p: u \neq 0, u_{S^c}=0} \frac{\|Xu\|_2^2}{\|u\|_2^2} \geq \|X\|^2 \widetilde{\phi}(|S|)^2. \tag{B.12}$$

Therefore,

$$\mathrm{Tr}(\Sigma_{\tilde{S}}) \leq \Gamma \Lambda_{\max}((X_{\tilde{S}}^T X_{\tilde{S}})^{-1}) \leq \frac{\Gamma}{\Lambda_{\min}(X_{\tilde{S}}^T X_{\tilde{S}})} \leq \frac{\Gamma}{\|X\|^2 \widetilde{\phi}(\Gamma)^2}.$$

Under P_{θ_0}, using (1) and (B.9), the bias term can be decomposed as

$$\|\mu_{\tilde{S}} - \theta_{0,\tilde{S}}\|_2 \leq \|(X_{\tilde{S}}^T X_{\tilde{S}})^{-1} X_{\tilde{S}}^T X_{\tilde{S}^c} \theta_{0,\tilde{S}^c}\|_2 + \|(X_{\tilde{S}}^T X_{\tilde{S}})^{-1} X_{\tilde{S}}^T Z\|_2 = I + II.$$

For I, note first that the ℓ_2-operator norm of $(X_{\tilde{S}}^T X_{\tilde{S}})^{-1}$ is bounded by $1/(\|X\|^2 \widetilde{\phi}(|\tilde{S}|)^2)$ by (B.12). On $\mathcal{T}_1$, using Cauchy-Schwarz,

$$\begin{aligned}
\|X_{\tilde{S}}^T X_{\tilde{S}^c} \theta_{0,\tilde{S}^c}\|_2^2 &= \sum_{i \in \tilde{S}} \left(\sum_{k=1}^n \sum_{j \in \tilde{S}^c} X_{ki} X_{kj} \theta_{0,j} \right)^2 \\
&= \sum_{i \in \tilde{S}} \left(\sum_{j \in \tilde{S}^c} \langle X_{\cdot i}, X_{\cdot j} \rangle \theta_{0,j} \right)^2 \\
&\leq \|X\|^4 \sum_{i \in \tilde{S}} \left(\sum_{j \in \tilde{S}^c \cap S_0} |\theta_{0,j}| \right)^2 \\
&\leq \|X\|^4 |\tilde{S}| s_0 \|\theta_{0,\tilde{S}^c}\|_2^2.
\end{aligned}$$

Together with (B.7), this gives

$$I \leq \frac{\|X\|^2 |\tilde{S}|^{1/2} s_0^{1/2} \|\theta_{0,\tilde{S}^c}\|_2}{\|X\|^2 \widetilde{\phi}(|\tilde{S}|)^2} \leq \frac{\Gamma^{1/2} s_0^{1/2} \varepsilon}{\widetilde{\phi}(|\tilde{S}|)^2}.$$

Using the same bound on the ℓ_2-operator norm and (1), on the event $\mathcal{T}_1 \subset \mathcal{T}_0$ it holds that

$$II \leq \frac{\|X_{\tilde{S}}^T Z\|_2}{\|X\|^2 \widetilde{\phi}(|\tilde{S}|)^2} = \frac{1}{\|X\|^2 \widetilde{\phi}(|\tilde{S}|)^2} \left(\sum_{i \in \tilde{S}} \left(X^T (Y - X\theta_0) \right)_i^2 \right)^{1/2} \leq \frac{2|\tilde{S}|^{1/2} \sqrt{\log p}}{\|X\| \widetilde{\phi}(|\tilde{S}|)^2}.$$

Combining all of the above bounds and using that $|\tilde{S}| \leq \Gamma$, on the event $\mathcal{T}_1$,

$$\lambda E_{\mu_{\tilde{S}}, \Sigma_{\tilde{S}}}(\|\theta_{\tilde{S}}\|_1 - \|\theta_{0,\tilde{S}}\|_1) \leq \frac{\lambda \Gamma}{\widetilde{\phi}(\Gamma)^2} \left(s_0^{1/2} \varepsilon + \frac{2\sqrt{\log p}}{\|X\|} + \frac{\widetilde{\phi}(|\tilde{S}|)}{\|X\|} \right).$$

Together with (B.10), the bound $\log(D_\Pi / D_N) \leq 2\lambda \Gamma^{1/2} \varepsilon + \log 2$ derived above and that $\widetilde{\phi}(|\tilde{S}|) \leq \widetilde{\phi}(1) \leq 1$, this yields

$$\mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}}) \| \Pi_{\tilde{S}}(\cdot|Y)\big) 1_{\mathcal{T}_1} \leq \frac{\lambda \Gamma}{\widetilde{\phi}(\Gamma)^2} \left(2 s_0^{1/2} \varepsilon + \frac{3\sqrt{\log p}}{\|X\|} \right) + \log 2.$$

Combining this with (B.8) and that $\log(1/\hat{q}_{\tilde{S}}) \leq \log(2e) + \Gamma \log p$ completes the proof. $\square$

We next consider the mean-field subclass $\mathcal{Q}_{MF}$ of $\mathcal{Q}$ given by (10). This again selects a single fixed support S but further requires the fitted normal distribution to have diagonal covariance matrix. We consider a diagonal version of $N_S(\hat{\theta}_S, (X_S^T X_S)^{-1}) \otimes \delta_{S^c}$ considered in Lemma B.2.

Lemma B.3. *If $4e^{1+\Gamma \log p - \kappa} \leq 1$, then the variational posterior $\widetilde{Q}$ arising from the family (10) satisfies*

$$\mathrm{KL}(\widetilde{Q} \| \Pi(\cdot|Y)) 1_{\mathcal{T}_1} \leq \Gamma \log \frac{p}{\widetilde{\phi}(\Gamma)} + \frac{\lambda \Gamma}{\widetilde{\phi}(\Gamma)^2} \left(2 s_0^{1/2} \varepsilon + \frac{3\sqrt{\log p}}{\|X\|} \right) + \log(4e).$$

Proof. We showed in the proof of Lemma B.2 that on the event $\mathcal{T}_1$ given in (B.2), there exists a set $\tilde{S}$ satisfying (B.7). Arguing as in (B.8),

$$\inf_{Q \in \mathcal{Q}_{MF}} \mathrm{KL}(Q || \Pi(\cdot|Y)) \leq \log \frac{1}{\hat{q}_{\tilde{S}}} + \inf_{\mu_{\tilde{S}}, D_{\tilde{S}}} \mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}}) \| \Pi_{\tilde{S}}(\cdot|Y)\big),$$

where the last Kullback-Leibler divergence is over the $|\tilde{S}|$-dimensional distributions and $D_{\tilde{S}}$ ranges over diagonal positive definite matrices. On $\mathcal{T}_1$ and for all p, we have $\log(1/\hat{q}_{\tilde{S}}) \leq \log(2ep^{\Gamma}) = \log(2e) + \Gamma \log p$ by (B.7).

The latter Kullback-Leibler divergence equals

$$\mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}}) \| \Pi_{\tilde{S}}(\cdot|Y)\big) = E_{\mu_{\tilde{S}}, D_{\tilde{S}}} \left[\log \frac{dN_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}})}{dN_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}})} + \log \frac{dN_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}})}{d\Pi_{\tilde{S}}(\cdot|Y)} \right] \tag{B.13}$$

for any covariance matrix $\Sigma_{\tilde{S}}$. For the first term in (B.13), the formula for the Kullback-Leibler divergence between two multivariate Gaussians gives

$$\mathrm{KL}\big(N_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}}) \| N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}})\big) = \tfrac{1}{2}\big(\log(|\Sigma_{\tilde{S}}|/|D_{\tilde{S}}|) - |\tilde{S}| + \mathrm{Tr}(\Sigma_{\tilde{S}}^{-1} D_{\tilde{S}})\big),$$

where $|A|$ denotes the determinant of a square matrix A. Set now $\mu_{\tilde{S}} = (X_{\tilde{S}}^T X_{\tilde{S}})^{-1} X_{\tilde{S}}^T Y$, $\Sigma_{\tilde{S}} = (X_{\tilde{S}}^T X_{\tilde{S}})^{-1}$ as in (B.9) and define the diagonal matrix $D_{\tilde{S}}$ via $(D_{\tilde{S}})_{ii} = 1/(\Sigma_{\tilde{S}}^{-1})_{ii} = 1/(X_{\tilde{S}}^T X_{\tilde{S}})_{ii}$. This gives $\mathrm{Tr}(\Sigma_{\tilde{S}}^{-1} D_{\tilde{S}}) = |\tilde{S}|$, so that it remains to control $\frac{1}{2}\log(|\Sigma_{\tilde{S}}|/|D_{\tilde{S}}|) = \frac{1}{2}\log(|\Sigma_{\tilde{S}}||D_{\tilde{S}}^{-1}|)$. For our choice of $D_{\tilde{S}}$,

$$|D_{\tilde{S}}^{-1}| = \prod_{j=1}^{|\tilde{S}|} (\Sigma_{\tilde{S}}^{-1})_{jj} = \prod_{j=1}^{|\tilde{S}|} (X_{\tilde{S}}^T X_{\tilde{S}})_{jj} \leq \|X\|^{2|\tilde{S}|},$$

while for $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ the smallest and largest eigenvalues, respectively, of a matrix A and using (B.12),

$$|\Sigma_{\tilde{S}}| \leq \Lambda_{\max}((X_{\tilde{S}}^T X_{\tilde{S}})^{-1})^{|\tilde{S}|} = (1/\Lambda_{\min}(X_{\tilde{S}}^T X_{\tilde{S}}))^{|\tilde{S}|} \leq 1/(\|X\| \widetilde{\phi}(|\tilde{S}|))^{2|\tilde{S}|}.$$

This yields that $\mathrm{KL}(N_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}}) \| N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}})) \leq |\tilde{S}| \log(1/\widetilde{\phi}(|\tilde{S}|)) \leq \Gamma \log(1/\widetilde{\phi}(\Gamma))$.

Note that the second term in (B.13) is identical to the expression (B.10), except that the expectation is taken under $\theta_{\tilde{S}} \sim N_{\tilde{S}}(\mu_{\tilde{S}}, D_{\tilde{S}})$ instead of $\theta_{\tilde{S}} \sim N_{\tilde{S}}(\mu_{\tilde{S}}, \Sigma_{\tilde{S}})$. One may therefore use the exact same arguments as in Lemma B.2 with the only difference occurring

in the second term in (B.11), where one instead has $\lambda E_{0,D_{\tilde{S}}}\|\theta_{\tilde{S}}\|_1 \leq \lambda|\tilde{S}|^{1/2}(E_{0,D_{\tilde{S}}}\|\theta_{\tilde{S}}\|_2^2)^{1/2} = \lambda|\tilde{S}|^{1/2}\mathrm{Tr}(D_{\tilde{S}})^{1/2}$. For e_i the i^{th} unit vector in $\mathbb{R}^p$,

$$\mathrm{Tr}(D_{\tilde{S}}) = \sum_{i=1}^{|\tilde{S}|} \frac{1}{(X_{\tilde{S}}^T X_{\tilde{S}})_{ii}} = \sum_{i\in\tilde{S}} \frac{1}{\|Xe_i\|_2^2} \leq \sum_{i\in\tilde{S}} \frac{1}{\|X\|^2\|e_i\|_2^2\widetilde{\phi}(1)^2} = \frac{|\tilde{S}|}{\|X\|^2\widetilde{\phi}(1)^2},$$

so that $\lambda|\tilde{S}|^{1/2}\mathrm{Tr}(D_{\tilde{S}})^{1/2} \leq \lambda\Gamma/(\|X\|\widetilde{\phi}(1))$. Combining the bounds as in Lemma B.2 then gives the result. □

Lemma B.4. *If $4e^{1+\Gamma\log p-\kappa} \leq 1$, then the variational posterior $\widetilde{\Pi}$ arising from the family (7) of spike-and-slab distributions satisfies*

$$\mathrm{KL}(\widetilde{\Pi}\|\Pi(\cdot|Y))1_{\mathcal{T}_1} \leq \Gamma\log\frac{p}{\widetilde{\phi}(\Gamma)} + \frac{\lambda\Gamma}{\widetilde{\phi}(\Gamma)^2}\left(2s_0^{1/2}\varepsilon + \frac{3\sqrt{\log p}}{\|X\|}\right) + \log(4e).$$

Proof. Since $\mathcal{Q}_{MF} \subset \mathcal{P}_{MF}$, we have $\mathrm{KL}(\widetilde{\Pi}\|\Pi(\cdot|Y)) \leq \mathrm{KL}(\widetilde{Q}\|\Pi(\cdot|Y))$. The result then follows from Lemma B.3. □

B.3 Oracle contraction rates for the original posterior distribution

Oracle type contraction rates for the original posterior were established in Castillo et al. [5]. However, their results are not stated with exponential bounds as needed in (14), so we must reformulate them in order to apply our Theorem 5. The required exponential bounds in fact follow from their proofs; we recall here the required results and, since [5] is a rather technical article, we provide a brief explanation why the exponential bounds hold.

Lemma B.5 (Theorem 10 of [5])**.** *Suppose the prior satisfies (4) and (5). Then for p large enough depending on A_2, A_4, any $M > 0$ and any $\theta_0, \theta_* \in \mathbb{R}^p$,*

$$\begin{aligned} E_{\theta_0}\Pi\left(\theta : |S_\theta| \geq |S_*| + \frac{4M}{A_4}\left(1 + \frac{16}{\phi(S_*)^2}\frac{\lambda}{\bar{\lambda}}\right)|S_*| + \frac{4M\|X(\theta_0-\theta_*)\|_2^2}{A_4\log p}\Big|Y\right)1_{\mathcal{T}_0} \\ \leq C(A_2,A_4)\exp\left(-(M-2)\left(1+\tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}}\right)|S_*|\log p - (M-1)\|X(\theta_0-\theta_*)\|_2^2\right), \end{aligned}$$

where $S_* = S_{\theta_*}$ and $\mathcal{T}_0$ is the event in (B.1).

Proof. Following the proof of Theorem 10 of [5], one obtains using (6.3) and the second display on p. 2008 of [5] that for $\bar{\lambda} = 2\|X\|\sqrt{\log p}$, any θ_* and any measurable set $B \subseteq \mathbb{R}^p$,

$$\sup_{\theta_0\in\mathbb{R}^p} E_{\theta_0}\Pi(B|Y)1_{\mathcal{T}_0} \leq e^{\|X(\theta_0-\theta_*)\|_2^2}\left(\frac{ep^{2s_*}}{\pi_p(s_*)}e^{\frac{8\lambda\bar{\lambda}s_*}{\|X\|^2\phi(S_*)^2}}\int_B e^{-(\lambda/4)\|\theta-\theta_*\|_1+\lambda\|\theta\|_1}d\Pi(\theta)\right)^{1/2}.$$

Setting now $B = \{\theta : |S_\theta| > R\}$ for $R \geq s_*$, the third display on p. 2008 of [5] shows that

$$\begin{aligned} \int_B e^{-(\lambda/4)\|\theta-\theta_*\|_1+\lambda\|\theta\|_1}d\Pi(\theta) &\leq \pi_p(s_*)4^{s_*}\left(\frac{4A_2}{p^{A_4}}\right)^{R+1-s_*}\sum_{j=0}^{\infty}\left(\frac{4A_2}{p^{A_4}}\right)^j. \\ &\leq C(A_2,A_4)\pi_p(s_*)4^{s_*}\left(\frac{4A_2}{p^{A_4}}\right)^{R+1-s_*} \end{aligned}$$

for p large enough that $4A_2/p^{A_4} < 1$. Substituting this into the second last display and using that $\bar{\lambda}^2 = 4\|X\|^2 \log p$,

$$\sup_{\theta_0 \in \mathbb{R}^p} E_{\theta_0} \Pi(B|Y) 1_{\mathcal{T}_0} \leq C(A_2, A_4) e^{\|X(\theta_0 - \theta_*)\|_2^2} (2p)^{s_*} e^{\frac{16\lambda s_* \log p}{\bar{\lambda}\phi(S_*)^2}} \left(\frac{4A_2}{p^{A_4}}\right)^{(R+1-s_*)/2}.$$

Choosing $R = (2\delta+1)s_* - 1 + 2\eta$, the right-hand side equals

$$C(A_2, A_4) \exp\big\{ \|X(\theta_0 - \theta_*)\|_2^2 + (\log 2 + \delta \log(4A_2))\, s_* + \Big(1 + \tfrac{16\lambda}{\bar{\lambda}\phi(S_*)^2} - \delta A_4\Big) s_* \log p \\ + \eta(\log(4A_2) - A_4 \log p) \big\}.$$

Further picking $\delta = 2M(1 + 16\lambda/(\bar{\lambda}\phi(S_*)^2))/A_4$ and $\eta = 2M\|X(\theta_0 - \theta_*)\|_2^2/(A_4 \log p)$, the right-hand side is bounded by

$$C(A_2, A_4) \exp\{-(M-2)(1 + \tfrac{16\lambda}{\bar{\lambda}\phi(S_*)^2}) s_* \log p - (M-1)\|X(\theta_0 - \theta_*)\|_2^2\}$$

for p large enough depending on A_2, A_4, as required. □

The following result is a modified version of the oracle inequality in Theorem 3 of [5] with $S_* = S_0$. Since it is stated somewhat differently in [5], we sketch why this is true.

Lemma B.6 (Theorem 3 of [5])**.** *Suppose the prior satisfies* (4) *and* (5)*. Then there exists a constant $M > 0$ such that for p large enough, both depending only on A_1, A_3, A_4, any $L \geq 1$, and uniformly over all $\theta_0, \theta_* \in \mathbb{R}^p$ with $|S_{\theta_*}| \leq |S_{\theta_0}|$,*

$$E_{\theta_0} \Pi\left(\theta : \|X(\theta - \theta_0)\|_2 > \frac{ML^{1/2}}{\overline{\psi}_{L+2}(S_0)} \left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right] \Big| Y\right) 1_{\mathcal{T}_0} \\ \leq C \exp\left(-\left[L \wedge \tfrac{4(L+2)}{A_4}\right] \left[(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}}) s_* \log p + \|X(\theta_0 - \theta_*)\|_2^2\right]\right) \\ + C \exp(-L(1 + \tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\bar{\lambda}}) s_0 \log p),$$

where $s_0 = |S_{\theta_0}|$, $s_* = |S_{\theta_*}|$ and $C = C(A_2, A_4)$. Moreover, both

$$E_{\theta_0} \Pi\left(\theta : \|\theta - \theta_0\|_1 > \|\theta_0 - \theta_*\|_1 + \frac{ML}{\overline{\psi}_{L+2}(S_0)^2} \left[\frac{s_* \sqrt{\log p}}{\|X\|\phi(S_*)^2} + \frac{\|X(\theta_0 - \theta_*)\|_2^2}{\|X\|\sqrt{\log p}}\right] \Big| Y\right) 1_{\mathcal{T}_0},$$

$$E_{\theta_0} \Pi\left(\theta : \|\theta - \theta_0\|_2 > \frac{ML^{1/2}}{\|X\|\widetilde{\psi}_{L+2}(S_0)^2} \left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right] \Big| Y\right) 1_{\mathcal{T}_0},$$

satisfy the same inequality.

Proof. Unless otherwise stated, we use here the notation from [5]. As on p. 2008 of [5], define the event $E = \{\theta : |S_\theta| \leq D_* \wedge D_0\}$ for

$$D_* = D_*(L) = s_* + \frac{4(L+2)}{A_4} \left(1 + \frac{16}{\phi(S_*)^2}\frac{\lambda}{\bar{\lambda}}\right) s_* + \frac{4(L+2)\|X(\theta_0 - \theta_*)\|_2^2}{A_4 \log p}, \tag{B.14}$$

where $\bar\lambda = 2\|X\|\sqrt{\log p}$ and D_0 is the same expression with θ_* replaced by θ_0. Note that we take different constants than in (6.7) of [5] to obtain the required exponential tail bound. Lemma B.5 yields, with $M = L + 2$ and since $s_* \le s_0$,

$$\begin{aligned} E_{\theta_0}\Pi(E^c|Y)1_{\mathcal{T}_0} &= E_{\theta_0}\Pi(\theta : |S_\theta| > D_* \wedge D_0|Y)1_{\mathcal{T}_0} \\ &\le C(A_2, A_4)\exp(-L(1 + \tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\bar\lambda})s_0 \log p) \\ &\quad + C(A_2, A_4)\exp(-L(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar\lambda})s_* \log p - L\|X(\theta_0 - \theta_*)\|_2^2) \end{aligned} \tag{B.15}$$

for every $\theta_0 \in \mathbb{R}^p$, so we can intersect the desired set with E in what follows.

From definition (12), we have $\overline{\psi}_{L+2}(S_0) = \overline{\phi}(D_0 + s_0)$. Continuing through the proof, the third last display on p. 2009 of [5] (note that up to this point, the definitions of D_* and D_0 only affect the definition of the compatibility type constants) gives

$$\begin{aligned} &\Pi(\theta \in E : \|X(\theta - \theta_0)\|_2 > 4\|X(\theta_* - \theta_0)\|_2 + R|Y)1_{\mathcal{T}_0} \\ &\le \frac{e}{\pi_p(0)A_1^{s_*}} p^{(2+A_3)s_*} e^{\frac{32\bar\lambda^2(D_* + s_*)}{\|X\|^2\overline{\psi}_{L+2}(S_0)^2}} e^{-\frac{R^2}{8}} \sum_{s=0}^{p} \pi_p(s)2^s, \end{aligned}$$

where again $\bar\lambda = 2\|X\|\sqrt{\log p}$. By condition (4), $\sum_{s=0}^p \pi_p(s)2^s \le \pi_p(0)\sum_{s=0}^p (2A_2 p^{-A_4})^s \le \pi_p(0)C(A_2, A_4)$ for p large enough. Using this and taking $R^2 = \overline{M}^2(D_* + s_*)\log p/\overline{\psi}_{L+2}(S_0)^2$, the last display is bounded by

$$\begin{aligned} &C(A_2, A_4)\exp\left\{-s_*\log A_1 + (2 + A_3)s_*\log p + \frac{128(D_* + s_*)\log p}{\overline{\psi}_{L+2}(S_0)^2} - \frac{1}{8}R^2\right\} \\ &\le C(A_2, A_4)\exp\left\{-\left[\frac{\overline{M}^2}{8} - 130 - A_3 - \frac{|\log A_1|}{\log p}\right]\frac{(D_* + s_*)\log p}{\overline{\psi}_{L+2}(S_0)^2}\right\}, \end{aligned}$$

where we have also used $\overline{\psi}_{L+2}(S_0) \le \overline{\phi}(1) \le 1$ for any S_0. Using the definition (B.14) of D_*, that $\lambda/\bar\lambda \le 2$ and the inequality $\sqrt{x + y} \le \sqrt{x} + \sqrt{y}$ for any $x, y \ge 0$,

$$(D_* + s_*)^{1/2} \le Cs_*^{1/2}L^{1/2}/\phi(S_*) + CL^{1/2}\|X(\theta_0 - \theta_*)\|_2/\sqrt{\log p}$$

for a constant $C > 0$ depending only on A_4, yielding

$$R \le \frac{C\overline{M}L^{1/2}}{\overline{\psi}_{L+2}(S_0)}\left(\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right).$$

Combining this with the third last display gives

$$\begin{aligned} &\Pi\left(\theta \in E : \|X(\theta - \theta_0)\|_2 > \frac{ML^{1/2}}{\overline{\psi}_{L+2}(S_0)}\left[\frac{\sqrt{s_*\log p}}{\phi(S_*)} + \|X(\theta_0 - \theta_*)\|_2\right]\Big|Y\right)1_{\mathcal{T}_0} \\ &\le C(A_2, A_4)\exp(-(D_* + s_*)\log p/\overline{\psi}_{L+2}(S_0)^2) \end{aligned}$$

for some $M > 0$ large enough depending only on A_1, A_3, A_4. Using $\overline{\psi}_{L+2}(S_0) \le 1$ and the definition (B.14), the probability in the last display is smaller than that in (B.15) if $4(L + 2)/A_4 \ge L$. Considering these two cases separately establishes the required inequality for the prediction error $\|X(\theta - \theta_0)\|_2$.

For ℓ_1-loss, the result follows from that for prediction error and the first display on p. 2010 of [5].

For ℓ_2-loss, note that $\|X(\theta-\theta_0)\|_2 \geq \widetilde{\phi}(|S_{\theta-\theta_0}|)\|X\|\|\theta-\theta_0\|_2 \geq \widetilde{\psi}_{L+2}(S_0)\|X\|\|\theta-\theta_0\|_2$ for any $\theta \in E$. The result then follows from that for prediction error and that $\overline{\psi}_{L+2}(S_0) \geq \widetilde{\psi}_{L+2}(S_0)$ by Lemma D.1. □

C Additional methodological details

C.1 Proofs for the variational algorithm

We provide here the derivations of the formulas used in the CAVI update equations of our variational algorithm in Section 4.

Proof of (16): We compute the Kullback-Leibler divergence between $P_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}$ and the posterior $\Pi(\cdot|Y)$, conditional on $z_i = 1$, as a function of μ_i and σ_i. Since the variational probability distribution of θ_i conditional on $z_i = 1$ (i.e. $P_{\mu_i,\sigma_i|z_i=1}$) is singular to the Dirac measure δ_0, in the Radon-Nikodym derivative $dP_{\mu_i,\sigma_i|z_i=1}/d\Pi_i$, where Π_i is the prior for θ_i, it suffices to consider only the continuous part of the prior measure in the denominator. Write $\Pi(\theta|Y) = D_\Pi^{-1}e^{-\|Y-X\theta\|_2^2/2}d\Pi(\theta)$ with D_Π the normalizing constant. Using all of these and the prior product structure, $\mathrm{KL}(P_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}|z_i=1}\|\Pi(\cdot|Y))$ equals, as a function of μ_i and σ_i,

$$
\begin{aligned}
&E_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}|z_i=1}\left[\tfrac{1}{2}\|Y-X\theta\|_2^2 + \log D_\Pi + \log\frac{dP_{\boldsymbol{\mu}_{-i},\boldsymbol{\sigma}_{-i},\boldsymbol{\gamma}_{-i}}\otimes N(\mu_i,\sigma_i^2)}{d\Pi_{-i}\otimes \overline{w}\mathrm{Lap}(\lambda)}\right] \\
&= E_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}|z_i=1}\left[\tfrac{1}{2}(Y-X\theta)^T(Y-X\theta) + \log\frac{dP_{\boldsymbol{\mu}_{-i},\boldsymbol{\sigma}_{-i},\boldsymbol{\gamma}_{-i}}}{d\Pi_{-i}}(\theta_{-i}) - \log\sigma_i - \frac{(\theta_i-\mu_i)^2}{2\sigma_i^2} + \lambda|\theta_i|\right] + C,
\end{aligned}
$$

where $C > 0$ is independent of μ_i, σ_i and $\overline{w}_i = a_0/(a_0+b_0)$ is the prior mean for w_i. Recall that the expected value of the folded normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ is $\sigma\sqrt{2/\pi}e^{-\mu^2/(2\sigma^2)} + \mu(1-2\Phi(-\mu/\sigma))$. Using this and explicitly evaluating the expectation of the first term, the last display equals

$$
\begin{aligned}
&\mu_i\sum_{k\neq i}(X^TX)_{ik}\gamma_k\mu_k + \frac{1}{2}(X^TX)_{ii}(\sigma_i^2+\mu_i^2) - (Y^TX)_i\mu_i + \lambda\sigma_i\sqrt{2/\pi}e^{-\mu_i^2/(2\sigma_i^2)} \\
&\quad + \lambda\mu_i(1-2\Phi(-\mu_i/\sigma_i)) - \log\sigma_i + C',
\end{aligned}
$$

where $C' > 0$ is again independent of μ_i, σ_i. Minimizing the last display with respect to either μ_i or σ_i (but not jointly) gives the same minimizers as minimizing f_i and g_i in (16).

Proof of (17): Similarly to the derivation of (16) above, the KL divergence between $P_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}$ and $\Pi(\cdot|Y)$ as a function of γ_i equals

$$
E_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}\left[\tfrac{1}{2}\|Y-X\theta\|_2^2 + \log\frac{dP_{\boldsymbol{\mu}_{-i},\boldsymbol{\sigma}_{-i},\boldsymbol{\gamma}_{-i}}}{d\Pi_{-i}}(\theta_{-i}) + \log\frac{d(\gamma_i N(\mu_i,\sigma_i^2)+(1-\gamma_i)\delta_0)}{d(\overline{w}_i\mathrm{Lap}(\lambda)+(1-\overline{w}_i)\delta_0)}(\theta_i)\right] + C,
$$

where $C > 0$ is independent of γ_i and $\overline{w}_i = a_0/(a_0+b_0)$. Since on an event of $P_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}$-

probability one, $\theta_i = 0$ if and only if $z_i = 0$, the last display equals

$$
\begin{aligned}
&E_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}\left[\tfrac{1}{2}\|Y - X\theta\|_2^2 + 1_{\{z_i=1\}}\log\frac{\gamma_i dN(\mu_i,\sigma_i^2)}{\overline{w}_i d\mathrm{Lap}(\lambda)}(\theta_i) + 1_{\{z_i=0\}}\log\frac{1-\gamma_i}{1-\overline{w}_i}\right] + C \\
&= E_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}}\left[\tfrac{1}{2}\|Y - X\theta\|_2^2 + 1_{\{z_i=1\}}\left(\log\frac{\sqrt{2}}{\sqrt{\pi}\sigma_i\lambda} - \frac{(\theta_i-\mu_i)^2}{2\sigma_i^2} + \lambda|\theta_i|\right)\right] \\
&\qquad + \gamma_i\log\frac{\gamma_i}{\overline{w}_i} + (1-\gamma_i)\log\frac{1-\gamma_i}{1-\overline{w}_i} + C \\
&= \gamma_i\Bigg\{\mu_i\sum_{k\neq i}(X^TX)_{ki}\gamma_k\mu_k + \tfrac{1}{2}(X^TX)_{ii}(\sigma_i^2+\mu_i^2) - (Y^TX)_i\mu_i + \log\frac{\sqrt{2}}{\sqrt{\pi}\sigma_i\lambda} - \frac{1}{2} \\
&\qquad + \lambda\sigma_i\sqrt{2/\pi}e^{-\mu_i^2/(2\sigma_i^2)} + \lambda\mu_i(1-2\Phi(-\mu_i/\sigma_i)) + \log\frac{\gamma_i}{1-\gamma_i} + \log\frac{b_0}{a_0}\Bigg\} + \log(1-\gamma_i) + C \\
&=: h_i(\gamma_i|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\gamma}_{-i}) \qquad \text{(C.1)}
\end{aligned}
$$

where $C > 0$ may change from line to line and is independent of γ_i. Setting the derivative with respect to γ_i of this last expression equal to zero and rearranging gives (17).

C.2 Algorithms for Gaussian slabs

We collect here for completeness the variational algorithms for the spike-and-slab prior with Gaussian slabs with which we have compared our method. First we give the component-wise update of the parameters as in [10], see Algorithm 2 below.

Algorithm 2 Component-wise variational Bayes for Gaussian prior slabs

Initialize: $(\Delta_H, \boldsymbol{\sigma}, \boldsymbol{\gamma})$, $\boldsymbol{\mu} := \hat{\mu}^{(0)}$ (for a preliminary estimator $\hat{\mu}^{(0)}$), $\boldsymbol{a} := order(|\boldsymbol{\mu}|)$
while $\Delta_H \geq \varepsilon$ **do**
  $\boldsymbol{\gamma}_{old} := \boldsymbol{\gamma}$
  for $i = 1$ to p **do**
    $\mu_i := \sigma_i^2\big((Y^TX)_i - \sum_{j\neq i}(X^TX)_{j,i}\mu_j\sigma_j\big)$
    $\sigma_i := 1/\sqrt{(X^TX)_{ii}+1}$
    $\gamma_i = \mathrm{logit}^{-1}\big(\log(a_0/b_0) + \log\sigma_i + \mu_i^2/(2\sigma_i^2)\big)$
  $\Delta_H := \max_i\{|H(\gamma_i) - H(\gamma_{old,i})|\}$

In [9] the authors argue that coordinate-wise parameter updates can accumulate error from each step leading to a suboptimal optimization procedure. To resolve this, they propose simultaneously updating the entire parameter vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ without using a CAVI type of algorithm. A version of their proposed algorithm is given in Algorithm 3, where $diag(v)$, $v \in \mathbb{R}^p$, creates a diagonal square matrix in $\mathbb{R}^{p\times p}$ with diagonal elements v (see also Algorithm 1 of [12] with $\alpha = 1$, $\sigma = 1$ and $\nu_1 = 1$ for a related implementation). As in the other cases, we have taken the ridge regression estimator $(X^TX + I)^{-1}X^TY$ as our initialization for μ.

Lastly, we provide the VB algorithm for the $\mathcal{Q}_{MF}$ mean-field variational class using Laplace slabs in the prior.

Algorithm 3 Batch-wise variational Bayes for Gaussian prior slabs

Initialize: $(\Delta_H, \boldsymbol{\sigma}, \boldsymbol{\gamma})$, $\boldsymbol{\mu} := \hat{\mu}^{(0)}$ (for a preliminary estimator $\hat{\mu}^{(0)}$), $\boldsymbol{a} := order(|\boldsymbol{\mu}|)$
while $\Delta_H \geq \varepsilon$ **do**
  $\boldsymbol{\gamma}_{old} := \boldsymbol{\gamma}$
  $\Gamma := diag(\gamma)$
  $\mu := (X^T X + \Gamma)^{-1} X^T Y$
  for $i = 1$ to p **do**
    $\sigma_i := 1/\sqrt{(X^T X)_{ii} + \gamma_i}$
    $\gamma_i := \text{logit}^{-1}\big(\text{logit}(1/p) + \log \sigma_i + \mu_i^2/(2\sigma_i^2)\big)$
  $\Delta_H := \max_i\{|H(\gamma_i) - H(\gamma_{old,i})|\}$

Algorithm 4 Variational Bayes for Laplace prior slabs and variational class $\mathcal{Q}_{MF}$

Initialize: $(\Delta_H, \boldsymbol{\sigma}, \boldsymbol{\gamma})$, $\boldsymbol{\mu} := \hat{\mu}^{(0)}$ (for a preliminary estimator $\hat{\mu}^{(0)}$), $\boldsymbol{a} := order(|\boldsymbol{\mu}|)$
while $\Delta_H \geq \varepsilon$ **do**
  $\boldsymbol{\gamma}_{old} := \boldsymbol{\gamma}$
  for $j = 1$ to p **do**
    $i := a_j$
    $\mu_i := \text{argmax}_{\mu_i} f_i(\mu_i | \boldsymbol{\mu}_{-i}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, z_i = 1)$ // see equation (16)
    $\sigma_i := \text{argmax}_{\sigma_i} g_i(\sigma_i |, \boldsymbol{\mu}, \boldsymbol{\sigma}_{-i}, \boldsymbol{\gamma}, z_i = 1)$ // see equation (16)
    $\gamma_i = \text{argmax}_{\gamma_i \in \{0,1\}} h_i(\gamma_i | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}_{-i})$ // see equation (C.1)
  $\Delta_H := \max_i\{|H(\gamma_i) - H(\gamma_{old,i})|\}$

D Examples of compatible design matrices

In addition to the compatibility type constants defined in Section 2.3, we also consider a stronger invertibility condition involving the 'mutual coherence' of the design matrix, which is the maximal correlation between the different predictors in X.

Definition D.1 (Mutual coherence)**.** *The mutual coherence number is*

$$\text{mc}(X) = \max_{1 \leq i \neq j \leq p} \frac{|\langle X_{\cdot i}, X_{\cdot j} \rangle|}{\|X_{\cdot i}\|_2 \|X_{\cdot j}\|_2}. \tag{D.2}$$

While we do not actually use the mutual coherence in our results, it provides an easy way to understand the compatibility constants in Definitions 1-3 in several well-studied design matrix examples below. The following result relates these notions.

Lemma D.1 (Lemma 1 of [5])**.** $\phi(S)^2 \geq \overline{\phi}(1)^2 - 15|S| mc(X)$, $\overline{\phi}(s)^2 \geq \widetilde{\phi}(s)^2 \geq \overline{\phi}(1)^2 - smc(X)$.

By evaluating the infimum in Definition 2 at the unit vectors, one obtains $\widetilde{\phi}(1) = \overline{\phi}(1) = \min_i \|X_{\cdot i}\|_2 / \|X\| = \min_{i \neq j} \|X_{\cdot i}\|_2 / \|X_{\cdot j}\|_2$, which is bounded away from zero if the columns of X have comparable Euclidean norms. In this case, Lemma D.1 implies that the compatibility numbers and sparse singular values are bounded away from zero for models of size $O(1/\text{mc}(X))$. The mutual coherence condition is thus the strongest of these notions. These conditions are illustrated via the following well-studied examples.

1. (Sequence model). We observe a vector $Y = (Y_1, \ldots, Y_n)$ of independent random variables with $Y_i \sim N(\theta_i, 1)$. This corresponds to model (1) with $n = p$ and $X = I_p$ the identity matrix, so that $\|X\| = \|X_{\cdot i}\|_2 = 1$ for all i, the compatibility numbers are 1 and $\mathrm{mc}(X) = 0$. In this setting, all results below are valid for all sparsity levels.

2. (Sequence model, multiple observations). We observe n independent $N(\theta_i, \sigma_n^2)$ random variables with $\sigma_n \to 0$. Defining Y_i as σ_n^{-1} times the original observations, this falls within the framework of model (1) with $X = \sigma_n^{-1} I_p$, so that $\|X\| = \|X_{\cdot i}\|_2 = \sigma_n^{-1}$ for all i, the compatibility numbers are 1 and $\mathrm{mc}(X) = 0$, similar to Example 1.

3. (Regression with orthogonal design). If X is an orthogonal design matrix such that $\langle X_{\cdot i}, X_{\cdot j} \rangle = 0$ for $i \neq j$, the regression problem can be transformed into a sequence model.

4. (Response model). Suppose the entries of the original regression matrix are i.i.d. random variables W_{ij}. We may then normalize the entries of the design matrix by defining $X_{ij} = W_{ij}/\|W_{\cdot j}\|_2$, so that the column lengths satisfy $\|X\| = \|X_{\cdot i}\|_2 = 1$ for all i. If $|W_{ij}| \leq C$ for a constant $C > 0$ and $\log p = o(n)$, or $E e^{t_0 |W_{ij}|^\alpha} < \infty$ for some $\alpha, t_0 > 0$ and $\log p = o(n^{\alpha/(4+\alpha)})$, then Theorems 1 and 2 of [3] show that $\sqrt{n/\log p}\,\mathrm{mc}(W) \xrightarrow{P} 2$ as $n \to \infty$. Since $\mathrm{mc}(W) = \mathrm{mc}(X)$, this shows that for any $\varepsilon > 0$, $P(\mathrm{mc}(X) > (2+\varepsilon)\sqrt{(\log p)/n}) \to 0$. Thus with probability approaching one, the compatibility numbers are bounded away from zero for sparsity levels $s_n = o(\sqrt{n/\log p})$.

   A classic example is $W_{ij} \overset{iid}{\sim} N(0,1)$. In this case, the above bound on the mutual coherence holds as long as $\log p = o(n^{1/3})$.

5. By rescaling the columns of X, one can set the $p \times p$ matrix $C := X^T X/n$ to take value one for all diagonal entries. Then $\|X\| = \|X_{\cdot i}\|_2 = \sqrt{n}$ for all i and the elements C_{ij}, $i \neq j$, are the correlations between columns. For some $m \in \mathbb{N}$, if $C_{ij} = r$ for a constant $0 < r < (1 + cm)^{-1}$ and all $i \neq j$ or $|C_{ij}| \leq c/(2m-1)$ for every $i \neq j$, then [13] show that models up to dimension m satisfy the 'strong irrepresentability condition' and are hence estimable. In particular, $\mathrm{mc}(X) = \max_{i \neq j} C_{ij} = O(1/m)$ and hence the compatibility numbers are bounded away from zero for sparsity levels $s_n = o(m)$.

References

[1] Breiman, L., and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc. 80*, 391 (1985), 580–619.

[2] Bühlmann, P., and van de Geer, S. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[3] Cai, T. T., and Jiang, T. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist. 39*, 3 (2011), 1496–1525.

[4] Carbonetto, P., and Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal. 7*, 1 (2012), 73–107.

[5] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Statist. 43*, 5 (2015), 1986–2018.

[6] Castillo, I., and van der Vaart, A. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist. 40*, 4 (2012), 2069–2101.

[7] Clara, G., Szabo, B., and Ray, K. *sparsevb: spike and slab variational Bayes for linear and logistic regression*, 2020. R package version 1.0.

[8] Horn, R. A., and Johnson, C. R. *Matrix analysis*, second ed. Cambridge University Press, Cambridge, 2013.

[9] Huang, X., Wang, J., and Liang, F. A variational algorithm for Bayesian variable selection. *ArXiv e-prints* (Feb. 2016), arXiv:1602.07640.

[10] Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC bioinformatics 11*, 1 (2010), 58.

[11] Nickl, R., and Ray, K. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist. 48*, 3 (2020), 1383–1408.

[12] Yang, Y., Pati, D., and Bhattacharya, A. α-variational inference with statistical guarantees. *Ann. Statist. 48*, 2 (2020), 886–905.

[13] Zhao, P., and Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res. 7* (2006), 2541–2563.