Quantifying Uncertainty in Variable Selection with Arbitrary Matrices

Willem van den Boom*, David Dunson*, and Galen Reeves*†
*Department of Statistical Science, Duke University
†Department of Electrical and Computer Engineering, Duke University

***Abstract*—Probabilistically quantifying uncertainty in parameters, predictions and decisions is a crucial component of broad scientific and engineering applications. This is however difficult if the number of parameters far exceeds the sample size. Although there are currently many methods which have guarantees for problems characterized by large random matrices, there is often a gap between theory and practice when it comes to measures of statistical significance for matrices encountered in real-world applications. This paper proposes a scalable framework that utilizes state-of-the-art methods to provide approximations to the marginal posterior distributions. This framework is used to approximate marginal posterior inclusion probabilities for Bayesian variable selection.**

I. Introduction

This paper considers the problem of recovering an unknown p-dimensional parameter vector β from a set of n noisy linear measurements of the form

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{1}$$

where X is a fixed $n \times p$ matrix of features and ϵ is additive white Gaussian noise with variance σ^2. This problem arises throughout science and engineering and has been studied extensively within compressed sensing and statistics [1]. The problem of variable selection, also known as support recovery, is to determine which entries of β are nonzero [2]–[4]. This problem is particularly challenging in the high-dimensional setting where p is large and much greater than n, since evaluating all possible subsets of the support is computationally intractable.

A great deal of recent work has focused on providing statistical guarantees for high-dimensional inference. One line of work has studied approximate message passing (AMP) algorithms [5]–[7], which provide posterior approximations under a postulated prior. Another line of work has studied the statistical properties of optimization-based methods, such as lasso and other M-estimators, and used these to derive confidence regions [8]–[12]. A remarkable feature of all of these methods is that, for certain large random matrices, their performance can be characterized rigorously and precisely. Moreover, empirical studies also suggest that many of the theoretical guarantees also apply more broadly to certain types of structured matrix construction [13]. At this point, the key challenge is to understand the extent to which these ideas can be applied to the types of feature matrices that one frequently encounters in statistical applications, which often have high degrees of collinearity and non-uniformity.

The goal of this paper is to introduce a framework for marginal posterior approximation that harnesses the power of existing methods for high-dimensional inference while being less restrictive about the types of feature matrices that can be used. For each entry of the parameter vector, the quality of approximation is assessed in terms of a single number, which measures the influence of the noise and the other parameters. This characterization provides a clean comparison of existing methods, which allows one to select the method which is most appropriate for a given matrix. Moreover, it can be used to accurately predict functionals of the prior and approximated posterior distributions, such as the receiver operating characteristic (ROC).

A. The Bayesian model

For the purposes of this paper we assume throughout that the parameter vector β is drawn according to a known iid prior,

$$p(\beta) = \prod_{j=1}^{p} p(\beta_j). \tag{2}$$

Under this prior, the joint posterior is difficult to visualize and interpret, and one routinely bases inferences on summaries of marginal posterior distributions for univariate functionals of the parameters. The marginal posterior distribution of coefficient β_j is obtained by marginalizing out the other coefficients $\beta_{(-j)}$ and can be stated as

$$p(\beta_j|y) \propto \int \exp\big(-\tfrac{1}{2\sigma^2}\|y - X\beta\|^2\big) p(\beta) d\beta_{(-j)}.$$

This expression is challenging to compute in general since it requires evaluating a high-dimensional integral.

Section II introduces our framework for approximating marginal posterior distributions. This framework consists of two stages. First, a rotation is applied to the data to separate the parameter of interest from the other parameters. Then, the posterior distribution of a scalar auxiliary variable is approximated using a Gaussian distribution. The problem of computing the mean and variance of the approximation can be attacked using existing methods such as AMP, lasso and Bayesian compressed regression [14].

Section III shows how our framework can be used for the problem of Bayesian variable selection, where the goal is to compute the posterior inclusion probabilities $p(\beta_j \neq 0|y)$, $j = 1, 2, \ldots, p$. These inclusion probabilities provide summaries of the weight of evidence in favor of the respective hypotheses

$H_{1j} : \beta_j \neq 0$ representing that the jth feature plays a role in predicting the response.

B. Relation to previous work

A number of recent papers have focused on statistical guarantees corresponding to specific recovery methods. For instance, [8] shows how confidence intervals can be obtained for various M-estimators while [9]–[12] do the same for test statistics derived from additional processing of the lasso solution. The theory for these confidence intervals depends on the consistency of the estimator used and asymptotic convergence rates of certain normal approximations in the limit as n and p scale to infiinty. The behavior for large iid subgaussian random feature matrices is determined precisely in [11].

There has also been a great deal of interest in AMP based algorithms, which are based on Gaussian and quadratic approximations to loopy belief propagation. The statistical behavior of AMP can be characterized theoretically for large iid subgaussian random feature matrices [15], and empirical results suggest that the theory holds more generally for certain types of nonrandom matrices [13]. One of the challenges with AMP, however, is that for arbitrary matrices convergence of the AMP iterations may require dampening [16] or serial updates [17]. Recent work has shown that stable points of the AMP iterations correspond to stationary points of an approximation to the Bethe free energy [16] and developed optimization methods which attempt to minimize the approximate Bethe free energy directly [18]. While this leads to methods with guaranteed convergence, the statistical behavior of the solution is not fully understood for general matrices.

Modeling the data in a Bayesian fashion provides an alternative. It forms a natural framework to evaluate statistical evidence via the posterior for general feature matrices. The posterior can however be computationally intractable, especially in the high-dimensional case. So even though many Bayesian variable selection methods exist [19], they typically rely on Monte Carlo sampling for inference [2], [20], [21], which does not scale well with the number of candidate predictors. This has motivated a rich literature on better samplers [22], [23] but since the dimensionality of the posterior grows exponentially in p, these are still not scalable to high-dimensional problems.

Our approximation framework focuses on computing the one-dimensional posterior distribution of a single entry of the parameter vector. The other parameters are viewed as "nuisance" parameters and their combined influence is summarized in terms of a signal auxiliary variable, which is approximated as Gaussian. The key contribution of this paper is to describe a simple two-stage procedure where one first estimates the mean and variance of the auxiliary variable, and then combines these estimates with the prior to produce the final posterior approximation. This two-stage procedure has the property that estimates for the mean and variance in the first-stage are statistically independent of the parameter of interest.

In comparison to many of the classical Bayesian approximation methods, our framework can handle posteriors which are multimodal and posteriors which are discrete-continuous mixtures. This is not possible using methods based on direct normal-type approximations or Laplace's method [24]–[26].

II. General Approximation Framework

This section describes our general framework for approximating posterior marginal distributions. The key assumption we make is that the entries of β are independent, and hence the prior distribution can be decomposed as in (2).

A. Decomposition of posterior marginal

The first step is to introduce *rotated* data, which focuses the influence of a parameter of interest into a single response. We will assume henceforth that we are interested in the posterior marginal distribution of the jth entry of β.

Let $q_1 = x_j/\|x_j\|$ be a unit vector in the direction of the jth column of X and let Q_2 be an $n \times (n-1)$ matrix chosen so that $Q = [q_1|Q_2]$ is orthonormal, i.e., $QQ^T = I_n$. The rotated data are defined as

$$\tilde{y} = Q_2^T y, \qquad z = q_1^T y.$$

To characterize the distribution on $(\tilde{y}, z)$, we define

$$s = \|x_j\|^2, \qquad z_0 = q_1^T X_{(-j)}\beta_{(-j)}, \qquad \tilde{X} = Q_2^T X_{(-j)},$$

where $\beta_{(-j)}$ denotes the $p-1$ vector with the jth entry removed and $X_{(-j)}$ the $n \times (p-1)$ matrix with the jth column removed. Following the rotational invariance of the Gaussian distribution, the distribution of the rotated data is expressed as

$$\tilde{y}|\beta \sim \mathcal{N}\Big(\tilde{X}\beta_{(-j)}, \sigma^2 I_{n-1}\Big), \tag{3}$$

$$z|\beta \sim \mathcal{N}(\sqrt{s}\beta_j + z_0, \sigma^2). \tag{4}$$

The important property of this decomposition is that $\tilde{y}$ and β_j are independent.

Next, we express the posterior marginal distribution of β_j in terms of the rotated data. Using the fact that there is a one-to-one mapping from y to $(\tilde{y}, z)$ and Bayes' rule yields

$$p(\beta_j|y) = p(\beta_j|\tilde{y}, z) = \frac{p(\beta_j, z|\tilde{y})}{p(z|\tilde{y})}.$$

Furthermore, the numerator can be expressed as

$$\begin{aligned} p(\beta_j, z|\tilde{y}) &= \iint p(\beta_j, z|z_0, \tilde{y}) p(z_0|\tilde{y}) dz_0 \\ &= \iint p(\beta_j, z|z_0) p(z_0|\tilde{y}) dz_0, \end{aligned}$$

where the last step follows from the fact that the pair (β_j, z) is conditionally independent of $\tilde{y}$ given z_0. The first probability inside the integral corresponds to a one-dimensional regression problem and is given by

$$p(\beta_j, z|z_0) = p(z|\beta_j, z_0) p(\beta_j).$$

By (4), the first probability on the right-hand side is a Gaussian. The second probability is simply the prior distribution of β_j. Putting everything together yields

$$p(\beta_j|y) \propto \iint \exp\big(-\tfrac{1}{2\sigma^2}(z - \sqrt{s}\beta_j - z_0)^2\big) p(\beta_j) \times p(z_0|\tilde{y}) dz_0. \tag{5}$$

B. The approximation step

The main challenge in using the formulation of the previous section to efficiently approximate the marginal posterior of β_j is that computation of the distribution function $p(z_0|\tilde{y})$ is intractable. The key step in our approach is approximating this distribution using a Gaussian distribution.

We approximate the distribution of $z_0|\tilde{y}$ using a Gaussian distribution of the form

$$p(z_0|\tilde{y}) \approx \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(z_0-\mu)^2}{2\tau^2}\right), \tag{6}$$

where the mean μ and variance τ^2 are functions of $\tilde{y}$. These can be estimated by applying existing methods to the linear regression problem defined by the rotated data $(\tilde{y}, \tilde{X})$. Plugging the approximation in (6) back into (5) yields the approximation on $p(\beta_j|y)$.

As a heuristic justification for the Gaussian approximation of $p(z_0|\tilde{y})$, consider a setting in which the entries of the $(p-1)\times 1$ vector $q_1^T X_{(-j)}$ are of roughly the same order. Then, the a priori distribution of z_0 is approximately Gaussian by the central limit theorem for sums of independent variables. Provided that the entries of $\beta_{(-j)}$ given $\tilde{y}$ are weakly correlated, it can then be argued that the posterior distribution of z_0 is also approximately Gaussian. Using ideas from [15], this line of reasoning can be made rigorous for iid random feature matrices with iid subguassian entries. It is important to note that approximate Gaussianity of the predictive distribution does not hold in the setting where a small number of other columns of X are highly collinear with x_j.

III. Bayesian Variable Selection

This section applies the approximation framework described in Section II to the problem of variable selection. A common model for sparse vectors is to consider a mixture distribution of the form

$$\beta_j \overset{\text{iid}}{\sim} (1-\lambda)\delta_0 + \lambda\mathcal{N}(0, \sigma^2\psi), \quad j = 1, \ldots, p, \tag{7}$$

where δ_0 is a point mass at zero, $\mathcal{N}(0, \sigma^2\psi)$ is a Gaussian distribution, and $\lambda \in (0, 1)$ is the prior inclusion probability. This is known alternatively as the spike-and-slab prior or the Bernoulli-Gaussian prior. Note that the scaling of the variance of β by σ^2 means the signal-to-noise ratio is controlled by ψ and does not depend on the error variance.

The posterior distribution of β can be expressed as a Gaussian mixture model. However, the number of mixtures grows exponentially in p and thus direct computation is infeasible for large p.

A. The approximation and its quality

The approximation described in Section II requires finding the parameters μ and τ^2 in the approximation of $p(z_0|\tilde{y})$ from (6). Following the details outlined in [27], this can be accomplished by running AMP on the rotated data. Alternatively, following ideas given in [8], it is possible to obtain approximations based on the lasso estimate.

To assess the quality of our approximations, we would ideally like to compare the approximated and true posterior marginal inclusion probabilities. Unfortunately, we are unable to make such a comparison since direct computation of the posterior marginal inclusion probabilities is computationally intractable for problem sizes of interest.

As an alternative, we consider the empirical receiver operating characteristic (ROC) curves corresponding to different methods of posterior approximation. These ROC curves allow us to compare the performance of different methods on different types of matrices. Moreover, if the curve of one method is uniformly higher than that of another, then the first method dominates the second, in the sense that it provides a strictly better posterior approximation.

B. Simulation

We used numerical simulations to obtain empirical ROC curves for posterior approximations made using our two-stage framework using two different methods (AMP and lasso) and two different types of matrices. In all cases, $(n, p) = (100, 200)$, the error variance was $\sigma^2 = 1$, and the parameter vector β was drawn according to a spike-and-slab prior.

For the first simulation, the entries of X were drawn iid zero-mean Gaussian with variance $1/n$ and the variance and prior inclusion probability of β were set to $(\psi, \lambda) = (50, 0.2)$. For the second simulation, the columns of X were set equal to $x_1 = \xi_1$ and $x_j = 0.7x_{j-1} + \sqrt{1-0.7^2}\xi_j$, $j = 2, \ldots, p$ where $\xi_j \overset{\text{iid}}{\sim} \mathcal{N}(0, I_n)$ such that the correlation between columns i and j was given by $0.7^{|i-j|}$. The variance and prior inclusion probability of β were set to $(\psi, \lambda) = (2.5, 0.01)$. Only the first predictor β_1 was used for construction of the ROC. The results are illustrated in Figure 1.

The results of the first simulation show that AMP outperforms lasso for the uncorrelated matrix. This is consistent with the asymptotic theory for large iid matrices which predicts that AMP will provide more accurate estimates of the mean and variance of the posterior distribution of the auxiliary variable z_0. Interestingly, though, the second simulation shows that for a feature matrix with highly correlated columns, lasso can provide a better posterior approximation. This is consistent with the observation that lasso is less sensitive to the properties of the matrix.

Acknowledgements

This material is based upon work supported in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

References

[1] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. Springer, 2011.

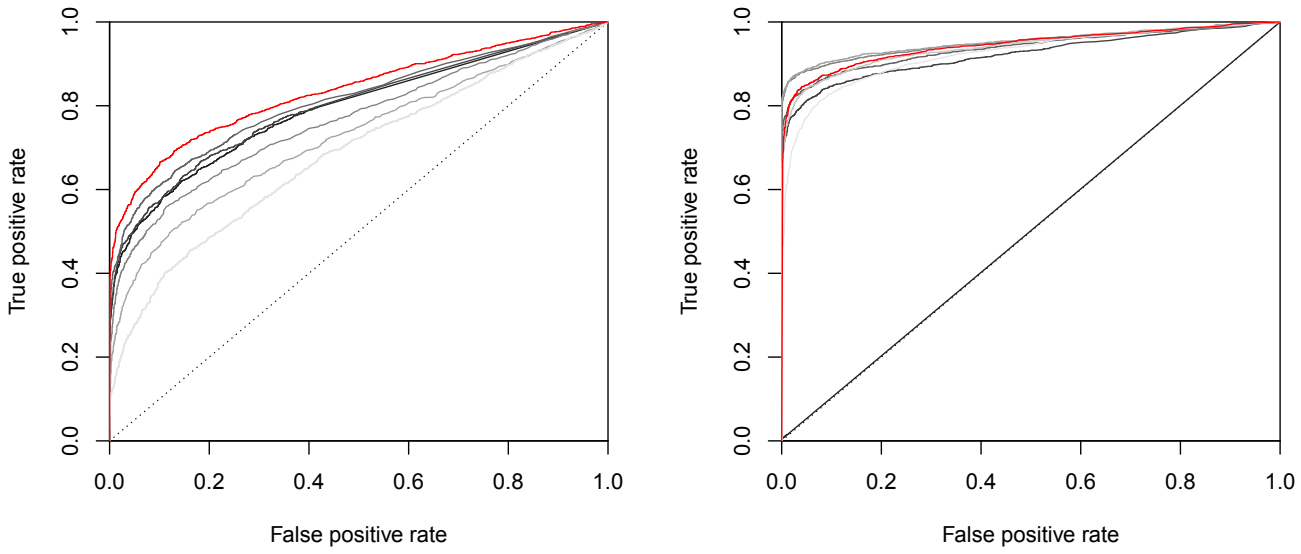

Fig. 1. Empirical ROC curves for variable selection resulting from 4000 Monte Carlo trials with $(n, p) = (100, 200)$. The results correspond to our two-stage approximation framework using AMP (red) and lasso with regularization parameter ranging from 0.01 (black) to 100 (light gray). The left panel corresponds to an iid Gaussian matrix and the right panel corresponds to a matrix drawn from a random ensemble with correlated columns.

[2] R. B. O'Hara and M. J. Sillanpää, "A review of Bayesian variable selection methods: What, how and which," *Bayesian Anal.*, vol. 4, no. 1, pp. 85–117, 2009.

[3] G. Reeves and M. Gastpar, "The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3065–3092, 2012.

[4] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3451–3465, 2013.

[5] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[6] S. Rangan, A. Fletcher, V. Goyal, and P. Schniter, "Hybrid generalized approximate message passing with applications to structured sparsity," in *IEEE International Symposium on Information Theory Proceedings*, pp. 1236–1240, 2012.

[7] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," *IEEE Transactions on Information Theory*, vol. 60, pp. 2969 – 2985, March 2014.

[8] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional statistical models," in *Advances in Neural Information Processing Systems 26*, pp. 1187–1195, 2013.

[9] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, no. 3, pp. 1166–1202, 2014.

[10] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 76, no. 1, pp. 217–242, 2014.

[11] A. Javanmard and A. Montanari, "Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6522–6554, 2014.

[12] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2869–2909, 2014.

[13] H. Monajemi, S. Jafarpour, M. Gavish, and D. L. Donoho, "Deterministic matrices matching the compressed sensing phase transitions of Gaussian random matrices," *Proceedings of the National Academy of Sciences*, vol. 110, no. 4, pp. 1181–1186, 2012.

[14] R. Guhaniyogi and D. B. Dunson, "Bayesian compressed regression," *Journal of the American Statistical Association*, 2015. Advance online publication.

[15] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and iterative algorithms," in *IEEE International Symposium on Information Theory*, (Boston, MA), July 2012.

[16] S. Rangan, P. Schniter, E. Riegler, A. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," in *IEEE International Symposium on Information Theory*, 2013.

[17] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Swept approximate message passing for sparse estimation," in *Proceedings of The 32nd International Conference on Machine Learning*, pp. 1123–1132, 2015.

[18] S. Rangan, A. K. Fletcher, P. Schniter, and U. Kamilov, "Inference for generalized linear models via alternating directions and Bethe free energy minimization," 2015. arXiv:1501.01797.

[19] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[20] E. I. George and R. E. McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[21] E. I. George and R. E. McCulloch, "Approaches for Bayesian variable selection," *Statistica Sinica*, vol. 7, pp. 339–374, 1997.

[22] D. J. Nott and R. Kohn, "Adaptive sampling for Bayesian variable selection," *Biometrika*, vol. 92, no. 4, pp. 747–763, 2005.

[23] M. A. Clyde, J. Ghosh, and M. L. Littman, "Bayesian adaptive sampling for variable selection and model averaging," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 80–101, 2011.

[24] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 82–86, 1986.

[25] R. Kass, L. Tierney, and J. Kadane, "The validity of posterior expansions based on Laplace's method," in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (J. S. Geisser, S. P. Hodges, and A. Zellner, eds.), pp. 473–488, North-Holland: Elsevier Science Publishers B.V., 1990.

[26] Y. Miyata, "Approximate marginal posterior distributions using asymptotic modes," *Communications in Statistics - Theory and Methods*, vol. 39, no. 7, pp. 1129–1140, 2010.

[27] W. van den Boom, G. Reeves, and D. B. Dunson, "Scalable approximations of marginal posteriors in variable selection," 2015. arXiv:1506.06629.