High-dimensional Bayesian modeling

Speakers: Joseph Antonelli and Antonio Linero

2021-06-24

1 Introduction to high-dimensional Bayesian inference

High-dimensional modeling has grown in popularity over the last couple of decades, for many different reasons. We need as working understanding of models in order to be up to speed with contemporary research and techniques. The Bayesian approach is particularly useful in high dimensions, since we can easily introduce non-linearity, introduce complex structures via hierarchical models, and handle missing data.

Frequentist estimators usually cannot provide confidence intervals for predictions or parameters, although some work is being done (Van de Geer et al., 2014; J. D. Lee et al., 2016).

The goal for today is to go from simple to complex models,

$$Y = \sum_{j=1}^{p} X_j \beta_j \xrightarrow{\text{GAM}} \sum_{j=1}^{p} f_j(X_j) \xrightarrow{\text{BART}} f(\boldsymbol{X}).$$

We are interested in predicting $\mathbb{E}\big[Y|\boldsymbol{X}\big] = f(\boldsymbol{X})$, with two simultaneous goals:

- Good prediction performance
- Identifying the q true important predictors in $\boldsymbol{X}$

Typically, $p > n$ and we assume that p grows with n (no further detail). Moreover we have P large enough that we need some sort of shrinkage or sparsity.

1.1 Spike and slab prior

The standard model is

$$Y = \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

although we can generalize to GLMs easily, $g^{-1}\big(\mathbb{E}\big[Y|\boldsymbol{X}\big]\big) = \sum_{j=1}^{p} X_j \beta_j$. Assuming β_0 is the true parameter,

$$q = \|\beta_0\|_0 = \sum_{j=1}^{p} \mathbb{1}(\beta_{0j} \neq 0).$$

We will discuss prior distributions for β such that they

1. Work when p is large
2. Identify nonzero elements of β_0
3. Are easy to implement computationally

We focus on prior distributions of the "spike-and-slab" form, i.e.

$$p(\beta_j|\gamma_j) \sim (1-\gamma_j)\delta_0 + \gamma_j \mathcal{N}(0, \sigma^2_\beta)$$

$$p(\gamma_j) = \tau^{\gamma_j}(1-\tau)^{1-\gamma_j}$$

Summing over γ_j, we have

$$p(\beta_j) = (1-\tau)\delta_0 + \tau\mathcal{N}(0, \sigma_\beta^2),$$

first introduced by Mitchell and Beauchamp (1988). This prior includes the belief that some covariates are not important while others are. We can also use a prior distribution of the form

$$p(\beta_j) = (1-\tau)\mathcal{N}(0, \sigma_0^2) + \tau\mathcal{N}(0, \sigma_1^2),$$

where $\sigma_0^2 < \sigma_1^2$ and is small so that it's sort of a spike near zero. This specification leads to easy update rules for the parameters, but requires good prior choice of σ_0^2 and σ_1^2.

Important: Standardize variables, otherwise you have performance issues and interpretation problems for the posterior inclusion probabilities.

Reasons to use spike and slab

1. We can look at the posterior probability $\mathbb{P}\big(\gamma_j = 1|\mathcal{D}\big)$ for variable importance.
2. Can look at the full posterior distribution $\mathbb{P}\big(\gamma_1, \ldots, \gamma_p|\mathcal{D}\big)$ to identify most likely models.
3. It still performs shrinkage of important coefficients, depending on δ_β^2.

The performance of the prior distribution depends on the hyperpriors, since

› τ : prior probability that a coefficient is zero, underlying sparsity.

› σ_β^2 : impacts the estimates, performs shrinkage and *has variable selection properties.*

1.2 Parameter updates

A traditional Gibbs sampler would update from full conditionals, which are hard to calculate for β_j and γ_j.
Unfortunately, if we condition on the current value of β_j, we have that

$$\mathbb{P}\big(\gamma_j|\beta_j, -\big) = \mathbb{1}(\beta_j \neq 0),$$

and γ_j is never going to change from the starting value. Therefore, we integrate out β_j when updating γ_j. Usually, we integrate parameters and condition on the data

$$p(\boldsymbol{\gamma}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_\gamma p(\mathcal{D}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}$$

and therefore sample in order

1. $p(\boldsymbol{\gamma}|\mathcal{D})$
2. $p(\boldsymbol{\beta}, \tau, \sigma_\beta^2|\boldsymbol{\gamma}, \mathcal{D})$ which is the same as a standard linear model.

Problem: In certain settings this is tractable (e.g. linear models), but otherwise it requires knowledge of the marginal likelihood of the data.

Different Gibbs sampler

A different strategy iterates through the following parameter updates:

- $p(\beta_j, \gamma_j|-)$ jointly for each $j = 1, \ldots, p$
- $p(\tau|-)$
- $p(\sigma_\beta^2|-)$

The key is to sample together (β_j, γ_j), which is computationally easy and does not get stuck at a particular γ_j value. We can sample them from

- $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma_\beta^2, \mathcal{D})$
- $p(\beta_j|\gamma_j, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma_\beta^2, \mathcal{D})$

The β_j update is simply the full conditional, however $\gamma_j| -$ except β_j is not straightforward. However, we have a *probability trick* to do this.

Denote $\boldsymbol{\vartheta}$ = all parameters except γ_j and β_j.

With terrible notation, let

$$\begin{aligned}
\mathbb{P}\big(\gamma_j = 1|\boldsymbol{\vartheta}, \mathcal{D}\big) &= \frac{\mathbb{P}\big(\beta_j = 0, \gamma_j = 1|\boldsymbol{\vartheta}, \mathcal{D}\big)}{\mathbb{P}\big(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathcal{D}\big)} \\
&= \frac{\mathbb{P}\big(\boldsymbol{\vartheta}, \mathcal{D}|\overbrace{\beta_j = 0, \gamma_j = 1}^{\text{can remove } \gamma_j}\big)\mathbb{P}\big(\beta_j = 0, \gamma_j = 1\big)}{\mathbb{P}\big(\boldsymbol{\vartheta}, \mathcal{D}\big)\mathbb{P}\big(\beta_j|\gamma_j = 1, \boldsymbol{\vartheta}, \mathcal{D}\big)} \\
&= \frac{\mathbb{P}\big(\boldsymbol{\vartheta}, \mathcal{D}|\beta_j = 0\big)\mathbb{P}\big(\beta_j = 0, \gamma_j = 1\big)}{\mathbb{P}\big(\boldsymbol{\vartheta}, \mathcal{D}\big)\mathbb{P}\big(\beta_j|\gamma_j = 1, \boldsymbol{\vartheta}, \mathcal{D}\big)} \\
&\propto \frac{\mathbb{P}\big(\beta_j = 0, \gamma_j = 1\big)}{\mathbb{P}\big(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathcal{D}\big)} \\
&= \frac{\mathbb{P}\big(\beta_j = 0|\gamma_j = 1\big)\mathbb{P}\big(\gamma_j = 1\big)}{\mathbb{P}\big(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathcal{D}\big)} \\
&= \frac{\varphi(0|0, \sigma_\beta^2) \cdot \tau}{\varphi(0|m, v)}.
\end{aligned}$$

The update for γ_j is a bernoulli distribution, since

$$\mathbb{P}\big(\gamma_j = 0|\boldsymbol{\vartheta}, \mathcal{D}\big) \propto \frac{\mathbb{P}\big(\beta_j = 0|\gamma_j = 0\big)\mathbb{P}\big(\gamma_j = 0\big)}{\mathbb{P}\big(\beta_j = 0|\gamma_j = 0, \boldsymbol{\vartheta}, \mathcal{D}\big)}.$$

If the data does not support the inclusion of the coefficient, the denominator in $\mathbb{P}\big(\gamma_j = 1|\boldsymbol{\vartheta}, \mathcal{D}\big)$ is going to be big and the likelihood of γ_j being 1 is going to be low.

Typically we use $\tau \sim \text{Beta}(C, p)$ so that updates are conjugate, where C is a constant, and the expected sparsity parameter is

$$\mathbb{E}\big[\tau\big] = \frac{C}{C + p}.$$

σ_β^2 is trickier, since it both performs shrinkage and impact variable selection. Moreover, σ_β^2 shows both in the numerator and the denominator of

$$\frac{\varphi(0|0, \sigma_\beta^2) \cdot \tau}{\varphi(0|m, v)}.$$

We have that

- σ_β^2 too small: can't distinguish spike from slab
- σ_β^2 too big: posterior probability of inclusion goes down

Remark

Assigning a diffuse prior on the slab is not possible, since it leads to bad inferences when it is not appropriate. A discussion of prior variance for model selection is Liang et al. (2008).

We can place a conjugate prior on σ_β^2, such as

$$\sigma_\beta^2 \sim \text{Inv-Gamma}(a, b),$$

or allow a separate slab variance for each covariate (Mitra and Dunson, 2010)

$$\sigma_{\beta_j}^2 \sim \text{Inv-Gamma}(a, b),$$

which can reduce shrinkage for larger coefficients, since it is not a "one size fits all" approach. Lastly, τ and σ_β^2 can be estimated by empirical Bayes methods.

2 Nonlinear models

2.1 Usual approach

Nonlinear models are of the form

$$\mathbb{E}\big[Y|\boldsymbol{X}\big] = \beta_0 + \sum_{j=1}^{p} f_j(X_j),$$

which we want to intuitively model with a spike and slab prior over functions, where the function is flat if it is not present in the model.

We can make a parametric assumption about $f_j(\cdot)$ as a basis function (splines, wavelets, ...)

$$\begin{aligned} f_j(X_j) &= \sum_{k=1}^{K} b_k(X_j)\beta_{jk} \\ &= \tilde{X}_j \beta_j \end{aligned}$$

If $\beta_j = \mathbf{0}$, then $f(X_j) = 0$, therefore we use

$$p(\boldsymbol{\beta}_j|\gamma_j) \sim (1-\gamma_j)\delta_{\mathbf{0}} + \gamma_j \mathcal{N}_K(\mathbf{0}, \Sigma_\beta),$$

which is similar to the univariate case and has lots of connections to grouped – either all in or all out – variable selection approaches (Bai, Moran, et al., 2020; Bai, Rockova, et al., 2021). As a prior

distribution, we can use some natural choices in order not to have too many parameters, which work reasonably well in practice

$$\Sigma_\beta = \begin{cases} \sigma_\beta^2 (\boldsymbol{X}_j^\top \boldsymbol{X}_j)^{-1} \\ \sigma_\beta^2 I_k \end{cases}$$

The orthogonal matrix can be used for orthogonalized covariates.

2.2 Nonparametric approach

We can place a prior on the function $f_j(\cdot)$ via Gaussian processes, which have been shown to be very flexible and to work well empirically,

$$f_j \sim \mathcal{GP}\big(\mu_j(X_j), K_j(X_j, X_j^\top)\big).$$

Here, $\mu_j(\cdot)$ is a mean function, zero or linear, and $K_j(X_j, X_j^\top)$ is a kernel function that reflects the similarity/distance between X_j and $X_j^\top$.

Since GPs are a very cool way of specifying multivariate distributions, we have for each finite collection of points the following result,

$$\Big(f_j(X_{j1}), \dots, f_j(X_{jn})\Big) \sim \mathcal{N}\Big(\big(\mu_j(X_{j1}), \dots, \mu_j(X_{jn})\big)^\top, \Sigma_j\Big),$$

where $\Sigma_j = \big(K(X_{ji}, X_{jk})\big)_{i,k=1,\dots,n}$. We can embed this in a spike-and-slab framework as (Reich et al., 2009)

$$f_j(\boldsymbol{X}_j) \sim \mathcal{N}(\boldsymbol{0}, \sigma_j \Sigma_j)$$

$$\sigma_j \sim (1-\gamma_j)\delta_0 + \gamma_j G,$$

where G is a continuous distribution on the positive real line. Alternatively, we can model it as

$$f_j(\boldsymbol{X}_j) \sim (1-\gamma_j)\delta_{\boldsymbol{0}} + \gamma_j \mathcal{N}_n(\boldsymbol{0}, \sigma_j \Sigma_j).$$

Instead of including this variable as a linear model, these approaches include it as a Gaussian Process nonparametric regression function.

We can use the same trick to see that

$$\mathbb{P}\big(\gamma_j = 1 | \mathcal{D}, \boldsymbol{\vartheta}\big) \propto \tau \frac{\varphi(\boldsymbol{0}|\boldsymbol{0}, \sigma_j \Sigma_j)}{\varphi(\boldsymbol{0}|M, V)},$$

where $\varphi(\cdot)$ is now the n-dimensional multivariate normal density. The problem now is to calculate M and V, which requires the inversion of a $n \times n$ matrix and is extremely costly from a computational point of view. Some approximations include Gramacy and H. K. H. Lee (2009), S. Banerjee et al. (2008) and A. Banerjee et al. (2013).

Take-home

› If computation time is a concern, use the basis function approach

$$f_j(X_j) = \tilde{X}_j \boldsymbol{\beta}_j.$$

› Otherwise, use Gaussian processes as in Qamar and Tokdar (2014), such that

$$\mathbb{E}\big[Y|\boldsymbol{X}\big] = f_1(\boldsymbol{X}) + \ldots + f_k(\boldsymbol{X}),$$

where each f_j is a separate GP which includes only a subset of the covariates.

3 Bayesian Additive Regression Trees (BART)

BART can perform nonparametric regression and classification, variable selection with some extensions, and exploration of interactions between variables.

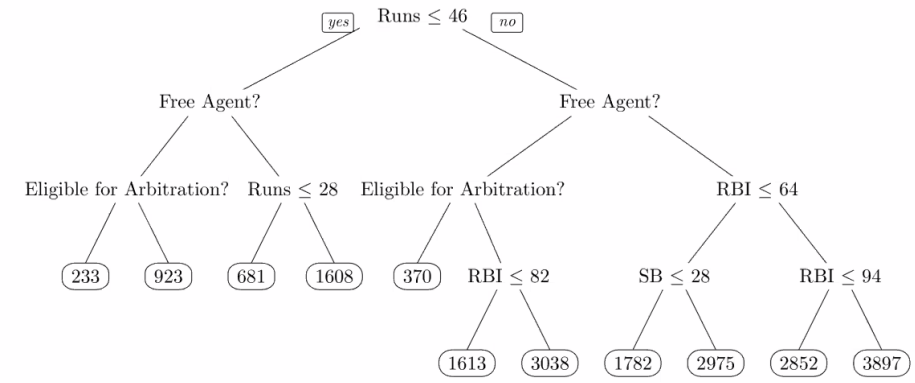

Figure 1: A simple decision tree, which implies variable selection through variables used to build rules.

Variable selection: If I don't split using a particular variable, it means that it is automatically selected as non-relevant.

BART (Chipman et al., 2010; Linero and Yang, 2018; Linero, Sinha, et al., 2020; Rockova and van der Pas, 2019) is a method that combines a bunch of bad decision trees in order to obtain a better predictor. Boosting is another method that falls in this category, and the gold standard is gradient boosted decision trees (`xgboost`).

We assume our function to be a sum of regression trees

$$r(x) = \mathcal{T}_1 + \mathcal{T}_2 + \ldots + \mathcal{T}_T$$

where on each tree $\mathcal{T}_t$ and each collection of leafs $\mathcal{M}_t$ we place a prior distribution. There are problems in placing a prior on T, since rjMCMC are quite hard to apply in this case. Tradition is to pick a standard value of $T = 50$ or $T = 200$, and it can be shown that if $T \to \infty$ there is a result that shows that this becomes an approximation to a Gaussian process if the tree leaves have prior variance $\mathcal{O}(T^{-1})$.

Formally,

$$r(x) = \sum_{t=1}^{T} g(x|\mathcal{T}_t, \mathcal{M}_t),$$

where $g(\cdot)$ is the associated step function to the tree. You can see this as an adaptive basis function expansion where the basis is made by step functions (LOL carina questa, mi piace).

Adding trees together induces smoothness, since different splits yield different values and there is a correlation between the values of the cells Figure 2.

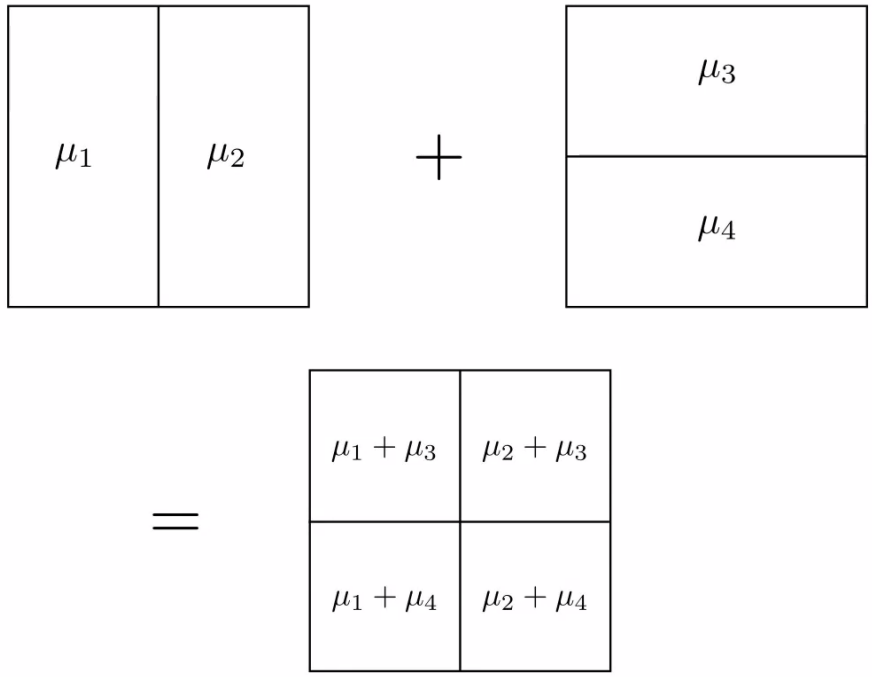

Figure 2: Smoothness and correlation induced by adding together decision trees.

To sample from the prior you start with a node and grow the tree by deciding whether to stop or keep splitting the tree.

3.1 Semiparametric regression

$$Y_i = r(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

also nonparametric probit regression and Poisson loglinear models

$$Y_i \sim \text{Ber}\big(\Phi\left\{r(X_i)\right\}\big)$$

$$Y_i \sim \text{Pois}\big(\exp\left\{r(X_i)\right\}\big).$$

Many other possibilities are used in practice, since there are collections of default hyperparameters which seem to work extremely well for some reason (see slide 'Magic Defaults').

Algorithm fitting is carried out via *Bayesian backfitting*, which can be proven to converge to the actual posterior distribution.

Algorithm 1 Bayesian backfitting

1:
2: **for** $t = 1, \ldots, T$ **do**
3: Compute residual $R_i = Y_i - \sum_{k \neq t} g(X_i | \mathcal{T}_k, \mathcal{M}_k)$
4: Propose $\mathcal{T}'$ from proposal distribution $Q(\mathcal{T}' | \mathcal{T})$
5: Compute marginal likelihood of $\mathcal{T}_t$ and $\mathcal{T}'$ as

$$\Lambda(\mathcal{T}) \prod_{\ell \in \mathcal{T}} \int \pi(\mu) \prod_{i \in \ell} \mathcal{N}(R_i | \mu, \sigma^2) \, d\mu$$

6: Set $\mathcal{T}_t \leftarrow \mathcal{T}'$ with probability

$$p = \frac{\Lambda(\mathcal{T}') \pi_{\mathcal{T}}(\mathcal{T}') Q(\mathcal{T}_t | \mathcal{T}')}{\Lambda(\mathcal{T}_t) \pi_{\mathcal{T}}(\mathcal{T}_t) Q(\mathcal{T}' | \mathcal{T}_t)}$$

7: Sample $\mathcal{M}_t$ from its full conditional
8: **end for**

Proposal distributions for sampling a new tree can be specified with different mechanisms:

› *Birth*: take a leaf node and split it into two leaves.

› *Death*: collapse two leaves into their parent.

› *Prior*: sample a new tree from the prior.

3.2 Measuring variable importance

There are different possibilities for measuring importance:

a) A variable is relevant if it is included in *at least one branch* of the ensemble.

b) A variable is relevant if it is included in *many branches* of the ensemble.

> **Def. (Variable importance)**
> The *importance* of a variable j is $\mathbb{E}\big[m_j / B | \mathcal{D}\big]$, the average proportion of all branches which split on variable j.

Other ways of variable importance metrics are Sobol' indices (Horiguchi et al., 2020).

We can optimize the prior splitting proportion s_j for the j-th coordinate by an empirical Bayes procedure. Using an EM algorithm, we can start from a prior value P and iteratively update s_j using

$$s_j \leftarrow \frac{\mathbb{E}[m_j | \mathcal{D}]}{\mathbb{E}[B | \mathcal{D}]}.$$

Otherwise, we can use a $s \sim \text{Dirichlet}(\eta, \ldots, \eta)$ prior and obtain a MAP estimator by starting from $s = (P^{-1}, \ldots, P^{-1})$ and iterating

$$s_j \leftarrow \frac{\max\{\mathbb{E}[m_j + \eta - 1 | \mathcal{D}, s], 0\}}{\sum_k \max_k \{\mathbb{E}[m_k + \eta - 1 | \mathcal{D}, s], 0\}}.$$

3.3 Number of predictors

If B is the number of branches, the number of active variables Q under the default BART prior setting $s_j = P^{-1}$ is

$$\mathbb{E}_{\Pi}\big[Q|B\big] = B + O(P^{-1}),$$

which is not very informative. Using $s \sim \text{Unif}(\mathbb{S}_{P-1})$ is also not useful, since it leads to the same answer

$$\mathbb{E}_{\Pi}\big[Q|B\big] = B + O(P^{-1}).$$

We can use a sparsity-inducing prior using a $s \sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P)$ and it can be shown that

$$Q - 1 \approx \text{Pois}(\vartheta),$$

where

$$\vartheta = \alpha \sum_{i=1}^{B-1} \frac{1}{\alpha + i}.$$

This last result holds some reminiscence to the number of prior expected number of clusters under a Dirichlet Process prior.

References

Bai, R., Moran, G. E., et al. (2020). "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models". In: *Journal of the American Statistical Association* 0.0, pp. 1–14.

Bai, R., Rockova, V., et al. (2021). *Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO*. arXiv: 2010.06451 [stat].

Banerjee, A. et al. (2013). "Efficient Gaussian process regression for large datasets". In: *Biometrika* 100.1, pp. 75–89.

Banerjee, S. et al. (2008). "Gaussian predictive process models for large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, pp. 825–848.

Chipman, H. A. et al. (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Gramacy, R. B. and Lee, H. K. H. (2009). *Bayesian Treed Gaussian Process Models with an Application to Computer Modeling*. arXiv: 0710.4536 [stat].

Horiguchi, A. et al. (2020). *Assessing Variable Activity for Bayesian Regression Trees*. arXiv: 2005.13622 [stat].

Lee, J. D. et al. (2016). "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3.

Liang, F. et al. (2008). "Mixtures of g Priors for Bayesian Variable Selection". In: *Journal of the American Statistical Association* 103.481, pp. 410–423.

Linero, A. R., Sinha, D., et al. (2020). "Semiparametric mixed-scale models using shared Bayesian forests". In: *Biometrics* 76.1, pp. 131–144.

Linero, A. R. and Yang, Y. (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.5, pp. 1087–1110.

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404, pp. 1023–1032.

Mitra, R. and Dunson, D. (2010). "Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors". In: *The International Journal of Biostatistics* 6.1.

Qamar, S. and Tokdar, S. T. (2014). *Additive Gaussian Process Regression*. arXiv: 1411.7009 [stat].

Reich, B. et al. (2009). "Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes". In: *Technometrics* 51, pp. 110–120.

Rockova, V. and van der Pas, S. (2019). *Posterior Concentration for Bayesian Regression Trees and Forests*. arXiv: 1708.08734 [math, stat].

Van de Geer, S. et al. (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3, pp. 1166–1202.