Spike-and-Slab Generalized Additive Models and Scalable Algorithms for High-Dimensional Data

Supporting Infromation

Boyi Guo, Byron C. Jaeger, AKM Fazlur Rahman, D. Leann Long, Nengjun Yi

Supplementary Information 1: Marginal Distribution of γ_j^*

Given that $\gamma_j^*|\gamma_j, \theta_j \sim Bin(1, \gamma_j\theta_j)$ where $\gamma_j|\theta_j \sim Bin(1, \theta_j)$, we can derive the the marginal distribution of γ_j^* with the following manipulation.

$$
\begin{aligned}
&\Pr(\gamma_j^* = 1|\theta_j) = \Pr(\gamma_j^* = 1, \gamma_j = 1|\theta_j) + \Pr(\gamma_j^* = 1, \gamma_j = 0|\theta_j) \\
=&\Pr(\gamma_j^* = 1, \gamma_j = 1|\theta_j) + 0 \qquad \text{[hierarchical structure between } \gamma^* \text{ and } \gamma.] \\
=&\Pr(\gamma_j^* = 1|\gamma_j = 1, \theta_j)\Pr(\gamma_j = 1|\theta_j) \\
=&\theta_j^2 \\
&\Pr(\gamma_j^* = 0|\theta_j) = \Pr(\gamma_j^* = 0, \gamma_j = 1|\theta_j) + \Pr(\gamma_j^* = 0, \gamma_j = 0|\theta_j) \\
=&\Pr(\gamma_j^* = 0, \gamma_j = 1|\theta_j) + \Pr(\gamma_j^* = 0, \gamma_j = 0|\theta_j) \\
=&\Pr(\gamma_j^* = 0|\gamma_j = 1, \theta_j)\Pr(\gamma_j = 1|\theta_j) + \Pr(\gamma_j^* = 0|\gamma_j = 0, \theta_j)\Pr(\gamma_j = 0|\theta_j) \\
=&(1 - \theta_j)\theta_j + 1(1 - \theta_j) = 1 - \theta_j^2
\end{aligned}
$$

Supplementary Information 2: EM-IWLS Algorithm for Fitting Bayesian Hierarchical Additive Models

Similar to the EM-CD algorithm, the EM-IWLS algorithm is an iterative EM-based algorithm where the iterative weighted least squares algorithm is used to find the estimate of $\boldsymbol{\beta}, \phi$ that maximizes $E(Q_1)$. The iterative weighted least squares algorithm was originally proposed to fit the classical generalized linear models, and generalized to fit some Bayesian hierarchical models.[@Gelman2013] Yi and Ma [@Yi2012] formulated Student's t-distribution and double exponential distribution as hierarchical normal distributions such that generalized linear models with shrinkage priors can be easily fitted using IWLS in combination with EM algorithm. In this work, we adapt the EM-IWLS paradigm to fit BHAM with spike-and-slab spline prior .

A double exponential prior, $\beta|S \sim DE(0, S)$ can be formulated as a hierarchical normal prior with unknown variance τ^2 integrated out:

$$\begin{aligned}\beta|\tau^2 &\sim N(0, \tau^2)\\ \tau^2|S &\sim Gamma(1, 1/(2S^2)),\end{aligned}$$

For the mixture double exponential priors, we can define the scale parameter $S = (1-\gamma)s_0 + \gamma s_1$ following Equation (**??**). The change in the prior formulation in turn leads to the change in the log posterior density function, as Q_1 needs to account for the hyperprior of τ^2:

$$Q_1(\boldsymbol{\beta}, \phi) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{p}\left[\log f(\beta_j|\tau_j^2) + \log f(\tau_j^2|S_j) + \sum_{k=1}^{K_j}\{\log f(\beta_{jk}^*|{\tau^*}_{jk}^2) + \log f({\tau^*}_{jk}^2|S_j^*)\}\right]. \quad (1)$$

Since $\boldsymbol{\tau}^2$ are not of our primary interest, we treat them as the "missing" data in addition to the latent indicators $\boldsymbol{\gamma}$, and hence construct the expectation $E_{\boldsymbol{\gamma},\boldsymbol{\tau}^2|\Theta^{(t-1)}}(Q_1)$ in the E-step. To note, unlike the same latent indicator γ_j^* which is shared by the coefficients of the nonlinear terms β_{jk}^* for $k = 1, \ldots, K_j$, τ_{jk}^2 is coefficient specific for β_{jk}^*. $E({S_j}^{-1}|\beta_j, s_0, s_1), E({S^*}_j^{-1}|\boldsymbol{\beta}_j^*, s_0, s_1), E(\tau_j^2|S_j, \beta_j)$ and $E({\tau^*}_{jk}^2|S_j^*, \beta_{jk}^*)$ needs to be calculated to formulate $E(Q_1)$. As neither $E({S_j}^{-1}|\beta_j, s_0, s_1)$ nor $E({S^*}_j^{-1}|\boldsymbol{\beta}_j^*, s_0, s_1)$ depends on τ^2s, they can be derived using Equation (**??**). On the other hand, τ^2, following gamma distributions, is a conjugate prior for the normal variance, and the conditional posterior density of τ^{-2} is an inverse Gaussian distribution. $E(\tau_j^{-2})$ and $E({\tau^*}_{jk}^{-2})$ are calculated using the closed form equation

$$E(\tau_j^{-2}|S_j, \beta_j) = {S_j}^{-1}/|\beta_j| \qquad E({\tau^*}_{jk}^{-2}|S_j^*, \beta_{jk}^*) = {S_j^*}^{-1}/|\beta_{jk}^*|,$$

where S_j and S_j^* are replaced by the expectation and βs are replaced with $\beta^{(t-1)}$. With simplification (up to constant additive terms), we have

$$E(Q_1) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) - \sum_{j=1}^{p}\left[2E({\tau_j}^{-2}){\beta_j}^2 + \sum_{k=1}^{K_j} 2E({\tau_{jk}^*}^{-2}){\beta_{jk}^*}^2\right]. \quad (2)$$

$2E(\tau^{-2})\beta^2$ can be seen as the kernel of a normal density with mean 0 and variance $E(\tau^2)$, and we can formulate the coefficients $\boldsymbol{\beta}$ as a multivariate normal distribution with means $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma}_{\tau^2}$, where $\boldsymbol{\Sigma}_{\tau^2}$ is a diagonal matrix with $E(\tau^2)$s on the diagonal,

$$\boldsymbol{\beta} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\tau^2}).$$

Meanwhile, following the classical IWLS, we can approximate the generalized model likelihood at each iteration with a weighted normal likelihood:

$$f(\mathbf{y}|\boldsymbol{\beta}, \phi) \approx \text{MVN}(\mathbf{z}|\boldsymbol{X}\boldsymbol{\beta}, \phi\boldsymbol{\Sigma})$$

where the 'normal response' z_i and 'weight' w_i are called the pseudo-response and pseudo-weight respectively. The pseudo-response and the pseudo-weight are calculated by

$$z_i = \hat{\eta}_i - \frac{L^{'}(y_i|\hat{\eta}_i)}{L^{''}(y_i|\hat{\eta}_i)} \quad w_i = -L^{''}(y_i|\hat{\eta}_i),$$

where $\hat{\eta}_i = (\boldsymbol{X}\hat{\boldsymbol{\beta}})_i$, $L^{'}(y_i|\hat{\eta}_i, \hat{\phi})$ and $L^{''}(y_i|\hat{\eta}_i, \hat{\phi})$ are the first and second derivative of the log density, $\log f(\mathbf{y}_i|\boldsymbol{\beta}, \phi)$ with respect to η_i.

With $\boldsymbol{z} \sim \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \phi\boldsymbol{\Sigma})$ and $\boldsymbol{\beta} \sim \text{MVN}(0, \phi\boldsymbol{\Sigma}_{\tau^2})$, we can augment the two multivariate normal distributions and update the estimates for $\boldsymbol{\beta}$ and ϕ via least squares in each iteration of the EM algorithm. We create the augmented response, augmented data, and augmented variance-covariance matrix following

$$\boldsymbol{z}_* = \begin{bmatrix} \boldsymbol{z} \\ \mathbf{0} \end{bmatrix} \qquad \boldsymbol{X}_* = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{I} \end{bmatrix} \qquad \boldsymbol{\Sigma}_* = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\tau^2}/\phi \end{bmatrix},$$

such that

$$\boldsymbol{z}_* \sim \text{MVN}(\boldsymbol{X}_*\boldsymbol{\beta}, \phi\Sigma_*).$$

Using the least squares estimators to update $\boldsymbol{\beta}$ and ϕ, we have

$$\boldsymbol{\beta}^{(t)} = (\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_* \qquad \phi^{(t)} = \frac{1}{n}(\boldsymbol{z}_* - X_*\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_* - X_*\boldsymbol{\beta}^{(t)}).$$

To note, the variance-covariance matrix of the coefficient estimates variance-covariance matrix can be derived in the EM-IWLS algorithm and in turn can be used for statistical inferences,

$$\text{Var}(\boldsymbol{\beta}^{(t)}) = (\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_*)^{-1}\phi^{(t)}.$$

Totally, the proposed EM-IWLS algorithm is summarized as follows:

1) Choose a starting value $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. For example, we can initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and $\boldsymbol{\theta}^{(0)} = \mathbf{0}.5$
2) Iterate over the E-step and M-step until convergence

   E-step: calculate $E(\gamma_j)$, $E(\gamma_j^*)$ and $E(\tau_j^{-2})$, $E(\tau^{*-2}_{jk})$ with the estimates $\Theta^{(t-1)}$ from the previous iteration

   M-step:

   a) Based on the current value of β, calculate the pseudo-data $z_i^{(t)}$ and the pseudo-weights $w_i^{(t)}$
   b) Update $\boldsymbol{\beta}^{(t)}$ by runing the augmented weighted least squared
   c) If ϕ is present, update ϕ

Similar to EM-CD, we assess convergence by the criterion, $|d^{(t)} - d^{(t-1)}|/(0.1 + |d^{(t)}|) < \epsilon$, where ϵ is a small value (say 10^{-5}).

Other Priors

With the re-parameterization step of the basis function matrix $\boldsymbol{X}$, it is possible to generalized the SSL prior to other priors, for example normal priors for ridge-type regularization and mixture normal prior for spike-and-slab regularization. These priors would work better in low and medium dimensional settings where the sparse assumption is not necessary. Here we elaborate the mixture normal prior as a demonstration of applying continuous spike-and-slab prior in BHAM.

A spike-and-slab mixture normal spline prior can be expressed as

$$\beta_j|\gamma_j, s_0, s_1 \sim N(0, (1-\gamma_j)s_0 + \gamma_j s_1)$$
$$\beta^*_{jk}|\gamma^*_j, s_0, s_1 \overset{\text{iid}}{\sim} N(0, (1-\gamma^*_j)s_0 + \gamma^*_j s_1), k = 1, \ldots, K_j.$$

Similar to the spike-and-slab spline prior in Equation (**??**), $0 < s_0 < s_1$ are tuning parameters and can be optimized via cross-validation. One of the critics received by the spike-and-slab mixture normal prior is that the tails of a normal distribution diminishes to zero too fast, which causes problems when estimating the large effects. Distributions with heavier tails can be used as an alternative, for example mixture Student's t distribution with small degree of freedom.

Supplementary Information 3: Predictive Performance of Linear Simulations

P	mgcv	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	0.38 (0.01)	0.39 (0.01)	0.31 (0.08)	0.29 (0.11)	0.38 (0.01)	0.35 (0.01)	0.39 (0.01)
10	0.36 (0.02)	0.38 (0.01)	0.35 (0.03)	0.34 (0.04)	0.39 (0.01)	0.33 (0.02)	0.39 (0.01)
50	0.09 (0.09)	0.37 (0.01)	0.30 (0.06)	0.30 (0.36)	0.38 (0.01)	0.32 (0.03)	0.37 (0.01)
100	-	0.37 (0.01)	0.28 (0.07)	0.34 (0.04)	0.38 (0.01)	0.29 (0.07)	0.35 (0.01)
200	-	0.36 (0.01)	0.26 (0.08)	0.31 (0.06)	0.38 (0.03)	0.28 (0.06)	0.33 (0.02)

Table 1: The average and standard deviation of the out-of-sample R^2 measure for Gaussian outcomes over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM), linear LASSO model (LASSO), component selection and smoothing operator (COSSO), adaptive COSSO, mgcv, sparse Bayesian generalized additive model (SB-GAM), and spikeSlabGAM mdoel. mgcv doesn't provide estimation whe number of parameters exceeds sample size i.e. p = 100, 200.

P	mgcv	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	0.79 (0.01)	0.79 (0.01)	0.76 (0.04)	0.75 (0.04)	0.78 (0.01)	0.76 (0.01)	0.79 (0.01)
10	0.77 (0.01)	0.79 (0.01)	0.78 (0.01)	0.78 (0.01)	0.78 (0.01)	0.75 (0.01)	0.79 (0.01)
50	0.62 (0.01)	0.78 (0.01)	0.75 (0.03)	0.73 (0.04)	0.74 (0.07)	0.75 (0.02)	0.77 (0.01)
100	-	0.78 (0.01)	0.73 (0.04)	0.69 (0.05)	0.73 (0.07)	0.74 (0.02)	0.76 (0.02)
200	-	0.78 (0.01)	0.71 (0.05)	0.67 (0.05)	0.73 (0.06)	0.73 (0.03)	0.72 (0.03)

Table 2: The average and standard deviation of the out-of-sample area under the curve measures for binomial outcomes over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM), linear LASSO model (LASSO), component selection and smoothing operator (COSSO), adaptive COSSO, mgcv, sparse Bayesian generalized additive model (SB-GAM), and spikeSlabGAM mdoel. mgcv doesn't provide estimation whe number of parameters exceeds sample size i.e. p = 100, 200.

Supplementary Information 4: Variable Selection Performance of Simulations

P	Metric	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	Precision	1.00	1.00	1.00	1.00	1.00	1.00
10	Precision	0.58	0.71	0.69	0.93	0.88	0.89
50	Precision	0.38	0.60	0.59	0.77	0.80	0.52
100	Precision	0.35	0.62	0.65	0.82	0.77	0.42
200	Precision	0.29	0.65	0.57	0.21	0.74	0.36
4	Recall	0.54	0.61	0.54	0.38	0.97	0.55
10	Recall	0.46	0.50	0.54	0.38	0.97	0.54
50	Recall	0.34	0.30	0.30	0.32	0.90	0.55
100	Recall	0.27	0.30	0.25	0.27	0.92	0.54
200	Recall	0.27	0.25	0.29	0.56	0.92	0.53
10	MCC	0.18	0.35	0.36	0.47	0.86	0.55
50	MCC	0.25	0.34	0.36	0.45	0.83	0.47
100	MCC	0.24	0.37	0.37	0.45	0.82	0.43
200	MCC	0.23	0.36	0.36	0.23	0.81	0.40

Table 3: The variable selection performance of binomial simulations, measured by positive predictive value (precision), true positive rate (recall), and Matthews correlation coefficient (MCC), for the high-dimensional methods averaged over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM), linear LASSO model (LASSO), component selection and smoothing operator (COSSO), adaptive COSSO, sparse Bayesian generalized additive model (SB-GAM), and spikeSlabGAM mdoel. MCC is ill-defined when $p = 4$ simulation (no true negative), and hence omitted for all methods.

P	Metric	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	Precision	1.00	1.00	1.00	1.00	1.00	1.00
10	Precision	0.59	0.97	0.97	0.90	0.56	0.99
50	Precision	0.43	0.74	0.84	0.89	0.39	0.99
100	Precision	0.34	0.53	0.72	0.91	0.29	0.99
200	Precision	0.27	0.42	0.52	0.92	0.33	0.99
4	Recall	1.00	0.87	0.84	1.00	1.00	1.00
10	Recall	1.00	0.98	0.96	1.00	1.00	1.00
50	Recall	1.00	0.85	0.99	1.00	1.00	1.00
100	Recall	1.00	0.84	0.97	1.00	1.00	1.00
200	Recall	1.00	0.78	0.88	1.00	0.98	1.00
10	MCC	0.59	0.96	0.94	0.91	0.54	0.99
50	MCC	0.59	0.76	0.90	0.92	0.57	1.00
100	MCC	0.54	0.61	0.81	0.94	0.48	1.00
200	MCC	0.49	0.52	0.64	0.95	0.52	0.99

Table 4: The variable selection performance of linear Gaussian simulations, measured by positive predictive value (precision), true positive rate (recall), and Matthews correlation coefficient (MCC), for the high-dimensional methods averaged over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM), linear LASSO model (LASSO), component selection and smoothing operator (COSSO), adaptive COSSO, sparse Bayesian generalized additive model (SB-GAM), and spikeSlabGAM mdoel. MCC is ill-defined when $p = 4$ simulation (no true negative), and hence omitted for all methods.

P	Metric	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	Precision	1.00	1.00	1.00	1.00	1.00	1.00
10	Precision	0.61	0.97	0.98	0.55	0.74	0.91
50	Precision	0.35	0.59	0.68	0.26	0.59	0.61
100	Precision	0.28	0.47	0.51	0.30	0.46	0.57
200	Precision	0.26	0.44	0.42	0.38	0.41	0.38
4	Recall	1.00	0.85	0.78	1.00	1.00	1.00
10	Recall	1.00	0.96	0.97	0.99	1.00	1.00
50	Recall	1.00	0.88	0.74	0.92	1.00	1.00
100	Recall	1.00	0.78	0.56	0.89	1.00	0.99
200	Recall	1.00	0.67	0.48	0.89	0.99	0.98
10	MCC	0.58	0.95	0.96	0.49	0.72	0.91
50	MCC	0.52	0.66	0.65	0.37	0.73	0.74
100	MCC	0.49	0.55	0.47	0.43	0.65	0.72
200	MCC	0.49	0.49	0.39	0.47	0.62	0.58

Table 5: The variable selection performance of linear binomial simulations, measured by positive predictive value (precision), true positive rate (recall), and Matthews correlation coefficient (MCC), for the high-dimensional methods averaged over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM), linear LASSO model (LASSO), component selection and smoothing operator (COSSO), adaptive COSSO, sparse Bayesian generalized additive model (SB-GAM), and spikeSlabGAM mdoel. MCC is ill-defined when $p = 4$ simulation (no true negative), and hence omitted for all methods.