BALSON: BAYESIAN LEAST SQUARES OPTIMIZATION WITH NONNEGATIVE L1-NORM CONSTRAINT

Jiyang Xie[1], *Zhanyu Ma*[1,*], *Guoqiang Zhang*[2], *Jing-Hao Xue*[3], *Jen-Tzung Chien*[4], *Zhiqing Lin*[1], *Jun Guo*[1]

[1]Pattern Recognition and Intelligent Systems Lab., Beijing University of Posts and Telecommunications, China
[2]School of Computing and Communications, University of Technology Sydney, Australia
[3]Department of Statistical Science, University College London, United Kingdom
[4]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

ABSTRACT

A Bayesian approach termed BAyesian Least Squares Optimization with Nonnegative L_1-norm constraint (BALSON) is proposed. The error distribution of data fitting is described by Gaussian likelihood. The parameter distribution is assumed to be a Dirichlet distribution. With the Bayes rule, searching for the optimal parameters is equivalent to finding the mode of the posterior distribution. In order to explicitly characterize the nonnegative L_1-norm constraint of the parameters, we further approximate the true posterior distribution by a Dirichlet distribution. We estimate the statistics of the approximating Dirichlet posterior distribution by sampling methods. Four sampling methods have been introduced. With the estimated posterior distributions, the original parameters can be effectively reconstructed in polynomial fitting problems, and the BALSON framework is found to perform better than conventional methods.

Index Terms— Bayesian learning, least squares optimization, L_1-norm constraint, Dirichlet distribution, sampling method

1. INTRODUCTION

In machine learning and statistics, optimization methods, including Newton's method [1], quasi-Newton method [1], sequence quadratic programming (SQP) method [2], gradient descent method [3], interior-point (IP) method [4], and Bayesian methods [5, 6, 7], are widely applied. The least squares optimization (LSO), which is one of the unconstrained optimization problems, includes the residual sum of squares (RSS) errors as the objective function. This optimization can be proved and solved by proven algorithms with low computational complexity [8, 9]. On this foundation, introduction of constraint conditions is beneficial to achieve numerical stability and increase predictive performance [9].

Sparsity is a common constraint to make the objective function depend on only a small number of model parameters. L_0- and L_1-norm regularizations are the commonly used constraints for sparsity. L_0-norm, denoted as $\|\cdot\|_0$, which can be defined as the number of non-zero elements in the parameter vector, performs the most precise sparsity of parameters, yet is difficult to implement in practice. L_1-norm, denoted as $\|\cdot\|_1$, which can be defined as the sum of the absolute values of the elements in a parameter vector, performs a strong sparsity constraint to the vector, and is convenient to be applied. With the constraint of L_1-norm regularization, the sparse representation [10], the nonlinear programming [11], and nonlinear time series prediction [12] are applied.

In addition to the aforementioned methods, solution under Bayesian framework is an alternative solution. With the probabilistic interpretation, the LSO problem (*i.e.*, the RSS objective function) is usually treated as Gaussian likelihood, and the constraint is considered as prior distribution. Combining the likelihood function with the prior distribution and with the Bayes theorem, finding the optimal solution to the constrained LSO problem is then equivalent to calculating the mode of the posterior distribution. This is a maximum *a posteriori* (MAP) solution to the constrained LSO problem. For example, with the L_1-norm constraint, the prior distribution is usually assumed to be a Laplacian [6, 13, 14]. Chien [5] proposed a Bayesian framework based on the Laplace prior of model parameters for sparse representation of sequential data. Finding the mode of the posterior distribution for Gaussian likelihood and Laplacian prior can solve the sparse optimization problem with numerical simulation.

There exists another type of regularization with nonnegative L_1-norm constraint, *i.e.*, the regularization term contains nonnegative elements only [9]. Nonnegative constraint plays an important role for solving the general nonnegative linear or nonlinear programming problems in physics (for example, fluid physics) [15] and engineering applications (for example, hyperspectral image processing, audio processing, web documents analysis, and bioinformatics data processing) [5, 10, 16]. In this case, Laplacian assumption cannot describe the constraint well as it has negative support.

In this paper, *we propose a Bayesian learning framework to solve this LSO problem with nonnegative L_1-norm con-*

*Corresponding author.

straint. In order to capture the distribution for the constraint term precisely, the Dirichlet distribution is applied. Bhattacharya et al. [17] introduced Dirichlet-Laplace priors for optimal shrinkage. Sato et al. [18] used the parametric mixture model with Dirichlet prior for knowledge discovery of multiple-topic document. Since combining the Gaussian likelihood for the model residual and the Dirichlet prior for the model parameters does not lead to a Dirichlet posterior, this paper approximates the posterior distribution with a Dirichlet distribution. Therefore, the optimal solution to LSO problem with nonnegative L_1-norm constraint can be solved by finding the mode of the approximating Dirichlet posterior distribution.

However, there is no analytically tractable solution to find the parameters of the aforementioned approximating Dirichlet posterior distribution. Sampling is an available method to analyze the statistical property of posterior distributions. Girolami et al. [19] used the importance sampling to calculate the corresponding moments with respect to the posterior distribution over the Dirichlet parameters. In addition to the importance sampling, other sampling methods including the rejection sampling [20] can also be applied. We propose an approach, called BAyesian Least Squares Optimization for Nonnegative L_1-norm constrain (BALSON), which utilizes sampling methods to approximate the required statistical properties (including mode) of posterior distributions.

2. BAYESIAN LEAST SQUARES OPTIMIZATION WITH NONNEGATIVE L1-NORM CONSTRAINT

2.1. Problem Formulation

A LSO problem with nonnegative L_1-norm constraint can be defined as

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \|y - f(\boldsymbol{x};\boldsymbol{\theta})\|_2^2, \\ \text{s.t.} \quad &\textstyle\sum_i \theta_i \le C, \theta_i \ge 0, i = 1, \cdots, K \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is the input of the model, $f(\cdot;\boldsymbol{\theta})$ is the model function with parameters $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]^{\mathrm{T}}$, y is the observed target value, and C is a constant. In this case, the optimization problem is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \|y - f(\boldsymbol{x};\boldsymbol{\theta})\|_2^2 + \lambda \mathbf{1}^{\mathrm{T}}\boldsymbol{\theta}, \\ \text{s.t.} \quad &\theta_i \ge 0, i = 1, \cdots, K \end{aligned} \tag{2}$$

where λ is a Lagrangian multiplier, and $\mathbf{1}$ is a column vector of ones. As we know, transformation $\ln t$ requires a nonnegative variable t which is suitable for our nonnegative constraint. Thus, we can introduce a log-barrier penalty with a group of positive hyperparameters $\mu_i, i = 1, \cdots, K$ to deal with the nonnegative constraint on the $\theta_i, i = 1, \cdots, K$, and the condition $\theta_i \ge 0$ can be replaced by $\sum_{i=1}^K \mu_i \ln \theta_i = M$, where M is a constant [9].

With nonnegative L_1-norm constraint, the problem is presented as

$$\min_{\boldsymbol{\theta}} \underbrace{\|y - f(\boldsymbol{x};\boldsymbol{\theta})\|_2^2}_{\mathbb{A}} + \underbrace{\lambda \mathbf{1}^{\mathrm{T}}\boldsymbol{\theta} - \sum_{i=1}^K \mu_i \ln \theta_i}_{\mathbb{B}}, \tag{3}$$

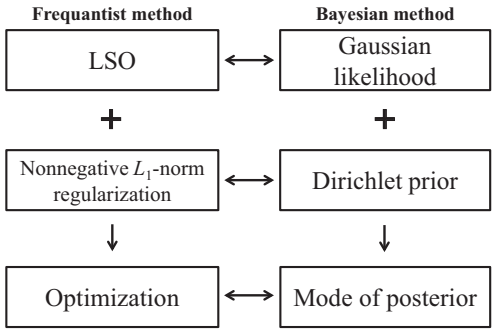


Fig. 1: The relation between the LSO problem with nonnegative L_1-norm constraint and the proposed Bayesian framework.

where $\{\mu_i\}_{i=1}^K$ are the Lagrangian multipliers. It is assumed that the model residual $e = y - f(\boldsymbol{x};\boldsymbol{\theta})$ follows the Gaussian distribution with zero mean and unit variance. Therefore, term $\mathbb{A}$ in (3) can be considered as the negative logarithm of a Gaussian likelihood with zero mean and unit variance up to a constant difference as

$$\mathbb{A} = -\ln \mathcal{N}(y - f(\boldsymbol{x}; C\boldsymbol{\omega}); 0, 1) + C_{\mathbb{A}}, \tag{4}$$

where $\boldsymbol{\omega} = \frac{\boldsymbol{\theta}}{C}$. Moreover, term $\mathbb{B}$ can be considered as the negative logarithm of a Dirichlet prior up to a constant difference as

$$\mathbb{B} = -\ln \mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha}) + C_{\mathbb{B}}. \tag{5}$$

A Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha})$ with parameter vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_K]^{\mathrm{T}}, \alpha_i > 0, i = 1, \cdots, K$ is defined as

$$\mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K \omega_i^{\alpha_i - 1}, \tag{6}$$

where order $K \ge 2$, $\sum_{i=1}^K \omega_i = 1$ and $\omega_i \ge 0, i = 1, \cdots, K$, and $B(\boldsymbol{\alpha})$ is the multivariate beta function as a normalization constant. It is worthy to note that this L_1-norm constraint is guaranteed by the definition of the Dirichlet, and $\mathbf{1}^{\mathrm{T}}\boldsymbol{\theta} = \sum_i \theta_i = C \cdot \sum_i \omega_i = C$ is a constant. Hence, the term $\lambda \mathbf{1}^{\mathrm{T}}\boldsymbol{\theta}$ in $\mathbb{B}$ can be neglected. Then, we can convert the original minimization problem in (3) to a maximization problem as

$$\max_{\boldsymbol{\omega}} \left[\ln \mathcal{N}(y - f(\boldsymbol{x}; C\boldsymbol{\omega}); 0, 1) + \ln \mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\alpha})\right]. \tag{7}$$

The maximization operation in (7) is equivalent to calculating the mode of the posterior distribution characterized by a Gaussian likelihood described in (4) and a Dirichlet prior distribution in (5). The relation between the LSO problem with nonnegative L_1-norm constraint and the proposed Bayesian framework is shown in Figure 1.

2.2. Implementation Procedure

The true posterior distribution characterized by a Gaussian likelihood and a Dirichlet prior distribution has a complex form which is not feasible in practice. In order to preserve the nonnegative L_1-norm constraint, we further *assume* that the posterior distribution follows a Dirichlet distribution. Although under such conditions there is no analytically tractable

solution to get the parameters, we can approximate the actual posterior distribution by matching the moments [21] in the approximating Dirichlet posterior $\mathrm{Dir}(\boldsymbol{\omega}; \boldsymbol{\alpha}^*)$. A numerical solution is proposed here to estimate the moments in the Dirichlet posterior by sampling methods. The first and second order moments of the Dirichlet posterior distribution are

$$\begin{aligned} \mathbb{E}[\omega_i] &= \frac{\alpha_i^*}{\alpha_0^*}, \\ \mathrm{Var}[\omega_i] &= \mathbb{E}[(\omega_i - \mathbb{E}[\omega_i])^2] = \frac{\alpha_i^*(\alpha_0^* - \alpha_i^*)}{(\alpha_0^*)^2(\alpha_0^*+1)}, \end{aligned} \tag{8}$$

where $\mathbb{E}[\omega_i]$ and $\mathrm{Var}[\omega_i]$ can also be estimated by mean value and variance of the i^{th}-dimensional samples respectively, $i = 1, \cdots, K$, and $\alpha_0^* = \sum_i \alpha_i^*$. Then, $\boldsymbol{\alpha}^*$ can be computed directly by solving the linear equations in (8). With the estimated parameters $\boldsymbol{\alpha}^*$, the optimal $\boldsymbol{\omega}^*$ to the LSO problem with nonnegative L_1-norm constraint can be obtained by computing the mode of the Dirichlet posterior distribution directly as

$$\omega_i^* = \begin{cases} \frac{\alpha_i^* - 1}{\sum_{j|\alpha_j^*>1} \alpha_j^* - K^*} & \alpha_i^* > 1 \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where K^* is the number of $\alpha_j^* > 1, j = 1, \cdots, K$. From (6), we can observe that, when α_i^* is smaller than 1, the marginal distribution on the i^{th} dimension is with a convex shape. Hence, the mode of this dimension is at 0 where the likelihood is infinity, and the other dimensions should be normalized as in (9). This ensures the sparse property of nonnegative L_1-norm constraint. Then, the optimal $\boldsymbol{\theta}^*$ in the original problem (1) can be computed by $\boldsymbol{\omega}^*$ as

$$\boldsymbol{\theta}^* = C\boldsymbol{\omega}^*. \tag{10}$$

The algorithm of BALSON is summarized in Algorithm 1. Four sampling methods have been applied in the algorithm: rejection sampling (RS), importance sampling (IS), rejection sampling importance resampling (RSIRS) and importance sampling importance resampling (ISIRS), which are described in Section 3.1.

Algorithm 1 BALSON

Require: $\boldsymbol{x}$: input data; $\boldsymbol{y}$: observed target data; K: dimension of model parameters

Ensure: $\boldsymbol{\theta}^*$: estimated model parameters

1: Initial values: $\boldsymbol{\alpha}$: parameter vector of Dirichlet prior; L: number of sampling points.
2: Sample from objective function in (7) using RS, IS, RSIRS or ISIRS and obtain L samples (and their importance weights in IS and IRS).
3: Compute the estimated parameters of Dirichlet posterior $\boldsymbol{\alpha}^*$ with the moments of L samples in (8).
4: Compute the mode $\boldsymbol{\omega}^*$ with the method in (9).
5: Compute the optimal $\boldsymbol{\theta}^*$ with the method in (10).

3. BAYESIAN INFERENCE AND INTERPRETATION

3.1. Sampling Solutions

3.1.1. Rejection Sampling

The rejection sampling (RS) method allows us to generate enough acceptable samples from relatively complex target distributions $p(z)$, and reject samples which do not satisfy the target distribution, subject to certain constraints [20]. A simpler proposal distributions $q(z)$, such as a Gaussian or Dirichlet distribution, and a constant k whose value is selected to satisfy $kq(z) \geq p(z)$ for all z are needed to draw samples easier. As an example, generating a one-dimensional sample z_0 needs three steps: (1) Generate a number z_0 from the distribution $q(z)$; (2) Generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$; (3) Accept z_0 if $u_0 \leq p(z_0)$, otherwise reject z_0. After having drawn L (enough) samples, the moments can be calculated.

3.1.2. Importance Sampling

The importance sampling (IS) method allows us to sample from $p(z)$ only to approximate the moments instead of drawing samples [20]. All samples drawn from $q(z)$ directly are accepted and weighted. The normalized importance weights $\tilde{\boldsymbol{r}} = \{\tilde{r}_1, \cdots, \tilde{r}_L\}$ and the corresponding samples $\boldsymbol{z} = \{z_1, \cdots, z_L\}$ are used to calculate the moments.

3.1.3. Rejection Sampling Importance Resampling

The rejection sampling importance resampling (RSIRS) is a combination of RS and IS methods. Firstly, we run RS one time and estimate parameter $\boldsymbol{\alpha}^{(\mathbf{0})}$ of the approximating Dirichlet posterior distribution. Then, the IS is implemented R rounds and the parameters $\{\boldsymbol{\alpha}^{(\mathbf{1})}, \cdots, \boldsymbol{\alpha}^{(\boldsymbol{R})}\}$ are estimated. In the i^{th} iteration of IS, the estimated parameter $\boldsymbol{\alpha}^{(\boldsymbol{i})}$ of Dirichlet posterior is approximated by combining Gaussian likelihood of data and Dirichlet prior with the parameter $\boldsymbol{\alpha}^{(\boldsymbol{i-1})}$.

3.1.4. Importance Sampling Importance Resampling

Similar to RSIRS, the importance sampling importance resampling (ISIRS) is a combination of IS methods which will work several iterations. The only difference with RSIRS is that the first step is IS method instead of RS method.

3.2. Relationship with Bayesian LASSO

Bayesian LASSO method is a popular Bayesian framework combining Gaussian likelihood and Laplace prior for sparse representation of model parameters [6]. Moreover, MAP problem in the Bayesian LASSO method can be converted to a negative logarithm form as a LSO problem with L_1-norm constraint as

$$\begin{aligned} &\min_{\tilde{\boldsymbol{\theta}}} \left\| y - f(\boldsymbol{x}; \tilde{\boldsymbol{\theta}}) \right\|_2^2, \\ \text{s.t.} \quad &\textstyle\sum_i \left| \tilde{\theta}_i \right| \leq C \end{aligned} \tag{11}$$

where $\tilde{\boldsymbol{\theta}} = \left[\tilde{\theta}_1, \cdots, \tilde{\theta}_{\tilde{K}}\right]^{\mathrm{T}}$ are model parameters.

Following the approach in [22], we extend the real vector $\tilde{\boldsymbol{\theta}}$ into a nonnegative real vector of K ($K = 2\tilde{K}$) dimensions $\boldsymbol{\theta} \triangleq \left[(\tilde{\boldsymbol{\theta}}^+)^{\mathrm{T}}, (\tilde{\boldsymbol{\theta}}^-)^{\mathrm{T}}\right]^{\mathrm{T}}$ as

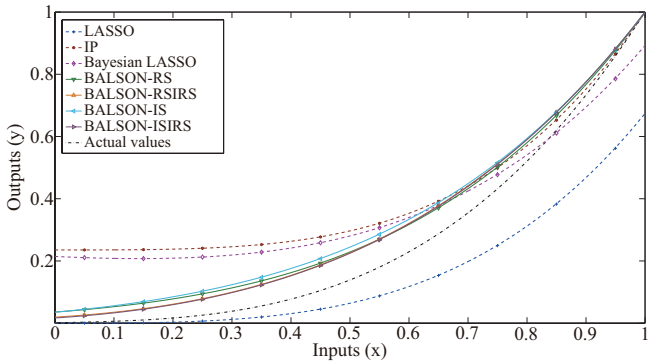


Fig. 2: Actual and fitting curves using different methods.

$$\theta_i = \tilde{\theta}_i^+ = \begin{cases} \tilde{\theta}_i & \tilde{\theta}_i > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

$$\theta_{i+\tilde{K}} = \tilde{\theta}_i^- = \begin{cases} -\tilde{\theta}_i & \tilde{\theta}_i < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

where $\tilde{\theta}_i^+$ and $\tilde{\theta}_i^-$ are the elements of $\tilde{\boldsymbol{\theta}}^+$ and $\tilde{\boldsymbol{\theta}}^-$, respectively, and $\tilde{\theta}_i^+ \geq 0$, $\tilde{\theta}_i^- \geq 0$, $i = 1, \cdots, \tilde{K}$, which means that $\boldsymbol{\theta}$ is a nonnegative vector. Therefore, $\tilde{\theta}_i$ has the relationship with $\tilde{\theta}_i^+$ and $\tilde{\theta}_i^-$ as

$$\tilde{\theta}_i \triangleq \tilde{\theta}_i^+ - \tilde{\theta}_i^-. \tag{14}$$

It is noticed that $|\tilde{\theta}_i| = \tilde{\theta}_i^+ + \tilde{\theta}_i^-$.Thus, the L_1-norm constraint $\sum_i |\tilde{\theta}_i| \leq C$ can be represented as $\sum_i (\tilde{\theta}_i^+ + \tilde{\theta}_i^-) = \sum_i \theta_i \leq C$, which is a nonnegative L_1-norm constraint.

According to the aforementioned approach, the Bayesian LASSO method can also be solved by our BALSON framework. Moreover, we can apply BALSON framework to the LSO problem with both nonnegative and common L_1-norm constraints, but the Bayesian LASSO method can only solve the LSO problem with common L_1-norm constraint, which means that the proposed BALSON framework is more general than the Bayesian LASSO method.

4. EXPERIMENTAL RESULT AND DISCUSSION

4.1. Polynomial Fitting Problem

Bayesian polynomial curve fitting is an important problem in signal processing for its excellent performance on standard denoising and speech segmentation problems [23, 24]. We apply the proposed BALSON framework to solve the polynomial fitting problem for illustrative purposes. The polynomial fitting problem aims to fit the K-dimensional polynomial parameters $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]^{\mathrm{T}}$, which can be expressed as

$$f(x; \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}} \Phi(x), \tag{15}$$

where x is the input scalar variable and the output is a scalar as well, and $\Phi(x) = \left[1, x, \cdots, x^{K-1}\right]^{\mathrm{T}}$ is a polynomial kernel as the input vector variable in (7). It is worth to note that, although we take scalar input and output as an example, the proposed method can also be applied to vector input and vector output.

Table 1: MSE and sparsity using different methods.

Method	Frequentist method		Bayesian method
	LASSO	IP	Bayesian LASSO
MSE	0.0209	0.0257	0.0370
Sparsity	0.6660	0.5353	0.3833

Method	BALSON (Bayesian method)			
	RS	RSIRS	IS	ISIRS
MSE	0.0143	**0.0116**	0.0159	0.0158
Sparsity	0.6751	0.8124	0.6785	**0.8675**

Three methods have been implemented as reference methods. These methods can be categorized into two classes. One class of method is the frequentist method, which contains the least absolute shrinkage and selection operation (LASSO) [25] and the interior-point (IP) [4] algorithm. Another class of method belongs to Bayesian methods, from which we select the Bayesian LASSO [6]. The proposed BALSON (with four different implementations named as BALSON-RS, BALSON-IS, BALSON-RSIRS, and BALSON-ISIRS, respectively) has been compared with the aforementioned methods, and the performance has been evaluated in terms of mean squared error (MSE) and sparsity. Here, MSE is defined as

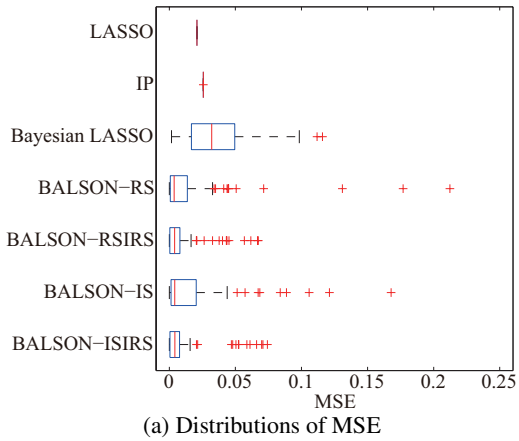
$$\text{MSE} = \frac{1}{N_{te}} (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}} (\boldsymbol{y} - \hat{\boldsymbol{y}}), \tag{16}$$

where $\boldsymbol{y}$ is the actual value vector, $\hat{\boldsymbol{y}}$ is the polynomial fitting value vector, and N_{te} is the number of test points. Moreover, the sparsity [26] denotes the degree of sparseness of the estimated polynomial parameters $\hat{\boldsymbol{\theta}}$, which is defined as

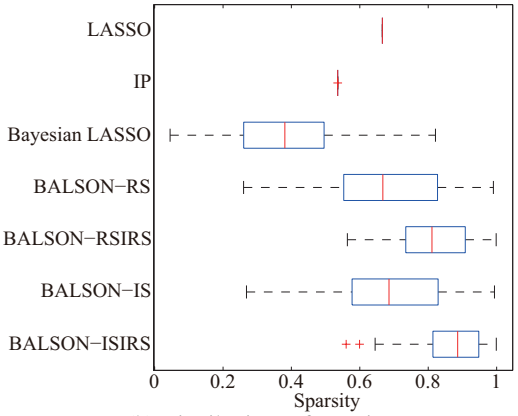
$$\text{Sparsity} = \frac{\sqrt{K} - \frac{\|\hat{\boldsymbol{\theta}}\|_1}{\|\hat{\boldsymbol{\theta}}\|_2}}{\sqrt{K} - 1}, \tag{17}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote L_1- and L_2-norm, respectively.

Data of the polynomial fitting problem are generated for both training and test. In training step, a group of training points is generated by a 5-dimensional polynomial parameter vector $\boldsymbol{\theta} = [0.0013, 0.0380, 0.0102, 0.9082, 0.0423]^{\mathrm{T}}$ (sparsity$= 0.9200$) and x obtained from 0 to 1 with fixed intervals according to the polynomial curve fitting model in (15), and added noise which follows the Gaussian distribution with zero mean and unit variance. Then, in the test step, we pick test points on the estimated and actual curves at the same input values which are obtained from 0 to 1 with fixed intervals. These points are used to measure the differences between estimated and actual curves by MSE defined in (16). We set the number of training points $N_{tr} = 100$ and number of test points $N_{te} = 1000$, and $C = 1$. All the seven methods (*i.e.*, LASSO, IP, Bayesian LASSO, BALSON-RS, BALSON-IS, BALSON-RSIRS, BALSON-ISIRS) have been conducted 100 times with the same data to obtain the distribution of MSE and sparsity. Figure 2 shows the comparisons among the estimated curves of different methods and the actual curve. The solid and dashed curves indicate estimated



(a) Distributions of MSE



(b) Distributions of sparsity

Fig. 3: Boxplots for distributions of MSE and sparsity using different methods.

values of the proposed methods and the referred methods, respectively, and the dashdot curve indicates the actual values. It can be observed that the proposed methods yield the curves that are closer to the actual one than the referred methods.

To quantitatively evaluate the performance, the mean values of MSE and sparsity are shown in Table 1. The smallest MSE and the highest sparsity in Table 1 are highlighted in bold. Moreover, the distributions of the MSE and the sparsity are shown in Figure 3(a) and 3(b), respectively. The proposed methods have smaller mean and median values of MSE than the referred methods, and the BALSON-RSIRS has the smallest mean value of MSE. Meanwhile, both of the BALSON-RSIRS and BALSON-ISIRS methods have higher mean and median values of sparsity which are larger than 0.8. Generally speaking, the mean and median values of sparsity of the proposed methods are higher than the referred methods.

Moreover, paired *t*-test on MSE and sparsity are conducted by setting the significance level as 0.05, respectively. The corresponding *p-values* are shown in Table 2 and 3. Most of the *p-values* of MSE computed between the proposed and the referred methods are smaller than 0.05, except for the *p-values* computed between BALSON-IS and LASSO, BALSON-ISIRS and LASSO, and BALSON-ISIRS and IP. In addition, most of the *p-values* of sparsity are smaller than 0.05, except for the *p-values* computed between BALSON-RS and LASSO, IS and LASSO. Therefore, only

Table 2: *P-values* of MSE.

	LASSO	IP	Bayesian LASSO
BALSON-RS	4.00E-02	4.86E-04	1.92E-06
BALSON-RSIRS	2.43E-03	7.27E-06	1.52E-08
BALSON-IS	7.55E-02	6.13E-04	4.34E-07
BALSON-ISIRS	3.74E-01	8.47E-02	3.93E-04

Table 3: *P-values* of sparsity.

	LASSO	IP	Bayesian LASSO
BALSON-RS	6.18E-01	4.44E-14	1.85E-19
BALSON-RSIRS	4.12E-24	4.33E-48	4.35E-38
BALSON-IS	5.15E-01	1.66E-13	4.75E-18
BALSON-ISIRS	5.73E-37	3.83E-58	1.61E-44

BALSON-RSIRS method has statistically significant performance improvements from all the referred methods on both MSE and sparsity. It is worth to note that we have conducted several experiments with different parameter settings (*i.e.*, different $\boldsymbol{\theta}$) and similar performances can be observed. Due to the limitation of space, we report only one example here.

4.2. Discussion

According to the results of the polynomial fitting experiments, the proposed methods have smaller MSE and higher sparsity. To specify, BALSON-RSIRS has the lowest MSE, and both BALSON-RSIRS and BALSON-ISIRS have higher sparsity than other methods. However, only BALSON-RSIRS method has statistically significant performance improvement from all the referred methods on both MSE and sparsity according to the results of the *t*-test. Therefore, by considering both MSE and sparsity together, we suggest to apply BALSON-RSIRS method to perform simulation for the proposed Bayesian framework.

The computational complexity of BALSON-RS and BALSON-IS are only related to number of samples L, number of training data N_{tr}, and dimension of polynomial parameters K when the rate of rejection is acceptable, which can be shown as $O(LK(K+N_{tr}))$. Thus, the computational complexity of BALSON-RSIRS and BALSON-ISIRS are $O(LK(K+N_{tr})R)$, where R is the number of rounds in importance resampling.

5. CONCLUSIONS

To solve the least squares optimization problem with non-negative L_1-norm constraint, a novel Bayesian optimization method, BALSON, is proposed. The error distribution of data fitting is described by Gaussian likelihood while the Dirichlet prior and the approximating Dirichlet posterior were applied to satisfy the conjugate match requirement. As no analytically tractable solution exists, we estimate the properties of the Dirichlet posterior of the parameters by implementing sampling methods. With the estimated posterior distributions, the original parameters can be effectively reconstructed. In order to evaluate the performance of the proposed methods, the BALSON framework has been applied in the polynomial fitting problems. Compared with several referred methods, it achieved the best performance. In addition to the polynomial fitting problems, the proposed methods can be extended

to other parameter estimation problems in many applications, such as hyperspectral image processing, audio signal processing, web documents analysis, and bioinformatics data processing. Future work will take into account optimizing by using variational Bayes methods [27, 28, 29] for approximating the posterior distribution under the Kullback-Leibler (KL) divergence constraint. In addition, the relation between BALSON and compressive sensing will be explored.

6. REFERENCES

[1] J. Nocedal and S. J. Wright, *Numerical Optimization, Second Edition*, Springer-Verlag, 2006.

[2] P. E. Gill, W. Murray, and M. A. Saunders, "SNOPT: An SQP algorithm for large-scale constrained optimization," *SIAM Review*, vol. 477, pp. 99–131, 2005.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[4] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming*, vol. 89, pp. 149–185, 2000.

[5] J. Chien, "Laplace group sensing for acoustic models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 909–922, 2015.

[6] T. Park and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, pp. 681–686, 2008.

[7] R. B. Ohara and M. J. Sillanpaa, "A review of Bayesian variable selection methods: what, how and which," *Bayesian Analysis*, vol. 4, no. 1, pp. 85–117, 2009.

[8] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[9] M. Schmidt, "Least squares optimization with L1-norm regularization," *CS542B Project Report*, vol. 504, pp. 195–221, 2005.

[10] H. Cheng, Z. Liu, L. Yang, and X. Chen, "Sparse representation and learning in visual recognition: Theory and applications," *Signal Processing*, vol. 93, pp. 1408–1425, 2013.

[11] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*, MIT Press, 2011.

[12] O. Kocadagli and B. Asikgil, "Nonlinear time series forecasting with Bayesian neural networks," *Expert Systems with Applications*, vol. 41, pp. 6596–6610, 2014.

[13] A. Mohammad-Djafari, "Bayesian approach with prior models which enforce sparsity in signal and image processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 52–70, 2012.

[14] P. M. Williams, "Bayesian regularisation and pruning using a Laplace prior," *Neural Computation*, vol. 7, pp. 117–143, 1995.

[15] H. Nagarajan and K. B. Nakshatrala, "Enforcing the nonnegativity constraint and maximum principles for diffusion with decay on general computational grids," *International Journal for Numerical Methods in Fluids*, vol. 67, pp. 820–847, 2011.

[16] G. Saito, H. W. Corley, J. M. Rosenberger, T. Sung, and A. Noroziroshan, "Constraint optimal selection techniques (COSTs) for nonnegative linear programming problems," *Applied Mathematics and Computation*, vol. 251, pp. 586–598, 2015.

[17] A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson, "Dirichlet-Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, vol. 110, pp. 1479–1490, 2015.

[18] I. Sato and H. Nakagawa, "Knowledge discovery of multiple-topic document using parametric mixture model with Dirichlet prior," in *Proceedings of Knowledge Discovery and Data Mining*, 2007, pp. 590–598.

[19] M. Girolami and S. Rogers, "Hierarchic Bayesian models for kernel learning," in *Proceedings of International Conference on Machine Learning*, 2005, vol. 119, pp. 241–248.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media LLC., 2006.

[21] R. Franke, T. S. Jang, and S. Sacht, "Moment matching versus Bayesian estimation: Backward-looking behaviour in a New-Keynesian baseline model," *The North American Journal of Economics and Finance*, vol. 31, pp. 126–154, 2015.

[22] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, vol. 58, pp. 267–288, 1996.

[23] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Transactions on Signal Processing*, vol. 50, pp. 747–758, 2002.

[24] P. Fearnhead, "Exact Bayesian curve fitting and signal segmentation," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2160–2166, 2005.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of The Royal Statistical Society Series B-statistical Methodology*, vol. 73, pp. 273–282, 2011.

[26] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, pp. 4723–4741, 2009.

[27] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 876–889, 2015.

[28] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2160–2173, 2011.

[29] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognition*, vol. 47, no. 9, pp. 3143–3157, 2014.