EURASIP Journal on Applied Signal Processing 2002:1, 38–45

Variable Selection by Perfect Sampling

Yufei Huang

Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-2350, USA
Email: yfhuang@ece.sunysb.edu

Petar M. Djurić

Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794-2350, USA
Email: djuric@ece.sunysb.edu

Received 31 July 2001 and in revised form 30 September 2001

Variable selection is very important in many fields, and for its resolution many procedures have been proposed and investigated. Among them are Bayesian methods that use Markov chain Monte-Carlo (MCMC) sampling algorithms. A problem with MCMC sampling, however, is that it cannot guarantee that the samples are exactly from the target distributions. This drawback is overcome by related methods known as perfect sampling algorithms. In this paper, we propose the use of two perfect sampling algorithms to perform variable selection within the Bayesian framework. They are the sandwiched *coupling from the past* (CFTP) algorithm and the Gibbs coupler. We focus our attention to scenarios where the model coefficients and noise variance are known. We indicate the condition under which the sandwiched CFTP can be applied. Most importantly, we design a detailed scheme to adapt the Gibbs coupler algorithm to variable selection. In addition, we discuss the possibilities of applying perfect sampling when the model coefficients and noise variance are unknown. Test results that show the performance of the algorithms are provided.

Keywords and phrases: variable selection, Markov chain Monte-Carlo, the Gibbs sampler, perfect sampling, coupling from the past, the Gibbs coupler.

1. INTRODUCTION

The problem of variable selection is of great importance in many science and engineering areas. In time-varying system identification, variable selection is used to select a set of basis sequences for representing the time varying coefficients [1]. In radar or sonar, it is applied to detection of number of reflections and for determination of relevant parameters of signal models [2]. In visual recognition [3, 4], variable selection plays an important role in determining the identity of the desired object from the data. In neural networks [5], variable selection is a necessary procedure in building neural as well as other classifiers.

Generally speaking, the basic goal of variable selection is to identify from a pool of available predictors the *best* subset with satisfactory predictive performance. For its resolution, many procedures have been proposed and investigated. A natural and direct approach is to exhaust all the possible subsets. However, this exhaustive search method become computationally burdensome if the available number of predictors is large. As a result, computationally efficient suboptimal approaches like genetic algorithms are used in practice [6].

In recent years, Bayesian signal processing methods that use Markov chain Monte-Carlo (MCMC) sampling algorithms [7, 8, 9, 10, 11] have drawn much attention. These methods usually provide better performance than deterministic approaches. MCMC sampling has been also applied to variable selection [12, 13], where it is referred to as *stochastic search variable selection*.

It is well known that with MCMC sampling, the generated samples are distributed according to a desired distribution only after the underlying Markov chain converges to equilibrium. However, in general, MCMC algorithms cannot detect the instants of convergence of the Markov chains. Therefore, MCMC algorithms may produce samples which follow desired distributions poorly, and as a result, subsequent estimations that use the generated samples may be quite inaccurate.

The drawback of MCMC is overcome by perfect sampling algorithms [14, 15, 16, 17]. Perfect sampling algorithms can attain i.i.d. samples exactly from desired distributions. The first perfect sampling algorithm is called *coupling from the past* (CFTP) [14]. This algorithm is only practical in ap-

plications that encounter distributions with small discrete state spaces. Efficient implementation of CFTP is only possible for large discrete state spaces if monotonic Markov chains can be constructed [14]. To increase the computational efficiency of CFTP, other algorithms have been proposed [15, 18]. Among them is an algorithm which we call the Gibbs coupler [19]. An appealing feature of the Gibbs coupler is that it can be applied efficiently to high-dimensional state spaces where monotonicity or anti-monotonicity do not exist.

In this paper, we propose to perform variable selection by two perfect sampling algorithms, the sandwiched CFTP and the Gibbs coupler. In particular, we focus our attention to scenarios where the model coefficients and the noise variance are known. We find the condition under which the sandwiched CFTP can be applied. We also develop a detailed scheme to accommodate the Gibbs coupler algorithm to variable selection and discuss the possibilities of applying perfect sampling when the model coefficients and the noise variance are also unknown.

The paper is organized as follows: the problem of variable selection is formulated in Section 2. A background on perfect sampling is provided in Section 3, where the sandwiched CFTP and the Gibbs coupler are reviewed. In Section 4, the implementations of the sandwiched CFTP and the Gibbs coupler are carefully studied. In Section 5, the possibilities of applying perfect sampling to more general settings of the problem are addressed. The performance of the methods is investigated through simulations in Section 6.

2. VARIABLE SELECTION

In many applications, observed data records can be represented by linear regression with Gaussian noise [13]

$$\mathbf{y} = \gamma_1\theta_1\mathbf{x}_1 + \gamma_2\theta_2\mathbf{x}_2 + \cdots + \gamma_q\theta_q\mathbf{x}_q + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{x}_i,\ i = 1,\ldots,q$ are q predictors, the θ's are the corresponding coefficients, the γ's are indicators which can take values of either 0 or 1, and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is a noise vector. All the vectors are of dimension $N \times 1$. Note that the indicators are utilized for selecting the right subset of the predictors as they indicate which predictors should be included in model (1). The primary task of variable selection is to choose the best subset of predictors from the available set, or equivalently, to determine the values of the indicators γ.

We focus on perfect sampling of the indicators γ, and so we concentrate our attention to the scenarios where the θ's and σ^2 are known to us. In a later section, we discuss ways of extending the proposed perfect sampling solution to situations where the θ's and σ^2 are also unknown.

From a Bayesian perspective, an optimum subset of the γ's can be defined as the one that maximizes the a posteriori probability of the γ's. When the Jeffreys' prior is adopted, the posterior distribution of the γ's can be shown as

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{q}\sum_{j=1}^{q}\theta_i\theta_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\gamma_i\gamma_j + \frac{1}{\sigma^2}\sum_{i=1}^{q}\theta_i\mathbf{x}_i^{\mathrm{T}}\mathbf{y}\gamma_i\right), \tag{2}$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_q]^{\mathrm{T}}$. Apparently, the maximization of (2) is a combinatorial optimization problem. The combinatorial optimization can always be solved by an exhaustive search method. The exhaustive search method computes the corresponding posterior probabilities of all possible combinations of γ's and selects the combination that has the largest posterior probability. However, when q is large, the computational requirement for the exhaustive search is prohibitive because we need to compute the probabilities of 2^q different models. As alternative solutions, MCMC sampling methods such as Gibbs sampling and reversible jump MCMC have been proposed [12, 13]. With an MCMC approach, the key is to draw random samples of $\boldsymbol{\gamma}$ from the posterior distribution (2). Once we accumulate the desired number of samples, the optimization of (2) can be approximated by

$$\hat{\boldsymbol{\gamma}}_{\mathrm{MAP}} \approx \arg\left\{\max_{\boldsymbol{\gamma}\in\{\boldsymbol{\gamma}^{(t)}\}_{t=1}^{M}} p(\boldsymbol{\gamma}|\mathbf{y})\right\}, \tag{3}$$

where $\boldsymbol{\gamma}^{(t)}$ represents the tth sample from the posterior distribution and M is the number of drawn samples. This Monte-Carlo optimization approach is also called *stochastic exploration* [20] since it explores the variable space of the unknowns using a random mechanism.

To generate random samples from an arbitrary desired distribution, MCMC sampling constructs a homogeneous Markov chain on the support of the unknowns, where the equilibrium distribution of the chain is the desired posterior distribution. However, a common drawback of all MCMC algorithms is that they lack theoretical diagnosis for detecting with certainty the convergence of the underlying Markov chain. This means that the samples obtained by them are very likely to only approximate the true posterior distribution. In addition, these samples are also correlated. Even though, theoretically, stochastic exploration does not require i.i.d. samples or even samples that follow the desired posterior distribution, the approximate and correlated samples could drastically reduce the efficiency of the approach. For instance, when the Markov chain is trapped in some local high density region, it tends to stay around there for a long time before it gets out.

The drawbacks of MCMC are ultimately overcome by perfect sampling algorithms. In the sequel, we present two perfect sampling algorithms and show how they can be used to draw perfect samples from the posterior distribution.

3. INTRODUCTION TO PERFECT SAMPLING ALGORITHMS

3.1. *Coupling from the past*

Perfect sampling algorithms are algorithms that can draw samples exactly from a desired distribution. Like MCMC,

they generate samples by running Markov chains, but they also possess the ability to determine the convergence time of the Markov chain. The first perfect sampling algorithm was proposed by Propp and Wilson [14], and is called *coupling from the past*. CFTP was initially designed for discrete variable spaces. However, efforts have been made to extend it to accommodate perfect sampling from continuous variable spaces [18, 21].

We now explain the idea behind the CFTP algorithm. Suppose that the desired discrete state space S is of size $M = |S|$ and that we initiate M Markov chains at every possible state of the state space S from some time $-T$ and run them to time 0. It is noted that all the Markov chains should have the desired distribution as their stationary distribution, and at any instant of time t in the past, the same random seed $R^{(t)}$ and updating function $\Phi(\cdot, R^{(t)})$ are applied. The updating function satisfies

$$\begin{aligned} x^{(t+1)} &= \Phi\left(x^{(t)}, R^{(t)}\right), \\ P\left(\Phi\left(x^{(t)}, R^{(t)}\right) = x^{(t+1)}\right) &= P\left(x^{(t+1)} \mid x^{(t)}\right), \end{aligned} \tag{4}$$

where $P(x^{(t+1)}|x^{(t)})$ is the probability of the Markov chain of going from state $x^{(t)}$ to state $x^{(t+1)}$. Suppose that there comes a time when all the chains reach the same state. From this time onward, due to the common random seed and updating function, all the chains follow the same path, and at time 0 they arrive at state $\bar{x}$. Now, if we restart a Markov chain from the infinite past but keep the same random seeds $R^{(t)}$ and update functions $\Phi(\cdot, R^{(t)})$ for the transitions in the interval $(-T, 0)$, this chain will also arrive at state $\bar{x}$ at time 0. The reason is the following: since the same random seeds $R^{(t)}$ and update functions $\Phi(\cdot, R^{(t)})$ are used for the transitions during $-T$ to 0, in this period, the chain initiated from the infinite past will follow one of the M paths that are determined before we restart the chain from the infinite past. Furthermore, since all the M chains have reached $\bar{x}$ at time 0, the path followed by the chain from the infinite past will also reach $\bar{x}$ at time 0. Apparently, since the chain has been propagated infinitely long from the past, the state $\bar{x}$ at time 0 is a steady state which follows the desired distribution exactly.

The above idea is implemented by an iterative scheme. At the very beginning, M copies of the Markov chain are started at every possible state from some time $-T$ to time 0. Then the output of the chains at time 0 is observed to check if they have coalesced into a singleton. If they have, then the coalesced output is recorded and taken as a perfect sample from the desired distribution. If they have not, the above procedure is repeated from another time further back in the past with the old random seeds from the corresponding Markov chain transitions being reused. Note that the reuse of the random seeds assures that the size of the state space at time t is not increasing from iteration to iteration which is critical for coalescence. The repeated procedure terminates until coalescence occurs at time 0. The pseudocode of the CFTP algorithm can be described as follows:

```
CFTP (T)
    t ⟵ −T
    B_t ⟵ S
    while t < 0
        t ⟵ t + 1
        for all x ∈ B_{t−1}
            B_t ⟵ Φ(x, U_t)

    if |B_t| = 1 then
        return (B_0)
    else
        CFTP (T′ > T)
```

It has been proved that CFTP draws a perfect sample in finite time on finite discrete state spaces with probability one [14]. However, the use of CFTP often becomes prohibitive due to the heavy computation in tracing all the chains in large spaces. A practical and an efficient simulation can be accomplished when the designed Markov chain is *monotonic* [14]. A monotonic Markov chain has an updating function that preserves a partial order $\prec$ on its state space, that is, $\Phi(x, R) \prec \Phi(y, R)$ for all R whenever $x \prec y$. According to the partial order, a maximal and a minimal state can be determined on the state space. If we implement CFTP with monotonic Markov chains, at any instant of time during the propagation of the chain towards 0, the monotonicity will cause all the chains starting from different states to be *sandwiched* between the paths started from the two extreme states. Obviously, when these two extreme paths coalesce at time 0, all the other paths, too, coalesce into the same state. Thus, efficient CFTP can be implemented by a sandwiched algorithm where only chains from the two extreme states are traced and examined for coalescence at time 0.

3.2. The Gibbs coupler

In Section 3.1, we have indicated that efficient CFTP algorithms are only possible for monotonic Markov chains. Unfortunately, in many cases, a monotonic Markov chain either is difficult to construct or even does not exist. In [22, 23], a component-updated perfect sampling scheme is proposed. A prominent feature of the algorithm is that neither monotonicity nor anti-monotonicity property is required. In [19], we independently proposed a similar scheme, which we refer to as the Gibbs coupler because it accommodates Gibbs sampling in the implementation of CFTP in order to achieve perfect samples from high-dimensional sampling spaces. The basic framework of the Gibbs coupler still follows that of CFTP, which means that the coupling algorithm first starts from some time $-T$, and its coalescence is checked at time 0. If coalescence happens, the algorithm terminates and a perfect sample is recorded. Otherwise, the coupling algorithm is repeated again from some time $-T'$ with $T' > T$. This framework actually guarantees that unbiased samples from the stationary distribution are obtained. However, there is an important difference with the CFTP algorithm discussed above in that instead of tracing all the possible paths generated by the Gibbs sampler, the main coupling algorithm of

the Gibbs coupler only records the possible values that the CFTP algorithm may generate.

In detail, at any time t, when updating the ith component, a new *support* of a component is generated from the full conditional distribution of the component, where the conditioning is on the latest updated supports of the remaining components. Similarly as in the Gibbs sampler, the updates are carried out in a componentwise fashion. Therefore, the advantage of the Gibbs sampler on high-dimensional spaces is preserved by the Gibbs coupler. The overall algorithm starts from some time T in the past by setting the initial support of each component to be the component's space and terminates when coalescence happens at time $t = 0$. *Coalescence occurs when the support of each component becomes a singleton.*

Now, let $S^{(t)} = \{S_1^{(t)}, S_2^{(t)}, \ldots, S_N^{(t)}\}$ denote the support of $\mathbf{x}$ at time t where $S_i^{(t)}$ represents the support of the component x_i at time t. In cases where the variable space is binary, as in our problem of variable selection, $S_i^{(t)} \in \{0, 1\}$, $i = 1, 2, \ldots, N$, the basic layout of the Gibbs coupler can be summarized as follows:

```
Gibbs coupler (T):
  t ⟵ −T
  while t < 0
    t ⟵ t + 1
    i ⟵ 0
    while i ≤ N
      update S_i^(t) using p(x_i | x_1, ..., x_{i−1}^(t), x_{i+1}^(t), ..., x_N^(t))
      for all x_j ∈ S_j^(t) with j = 1, 2, ..., N and j ≠ i

  if size of S_i^(0) for all i is equal to 1 then
    return (S^(0))
  else
    Gibbs coupler (T′ with T′ > T)
```

It is proved in [22] that the algorithm terminates almost surely and produces an unbiased sample from a desired distribution if there exists an $n < \infty$ such that the probability of coalescence of the whole space in n steps is greater than 0. Thus, the crucial point of the algorithm is the update procedure of locating the possible values in $S_i^{(t)}$. The detailed scheme, which may vary and is problem dependent, determines not only the coalescence but also the efficiency of the algorithm.

4. VARIABLE SELECTION BY PERFECT SAMPLING ALGORITHMS

4.1. Variable selection by the sandwiched CFTP

The purpose of the perfect sampling algorithms is to take samples from the posterior distribution (2). Once the desired number of samples is drawn, the MAP estimator of γ is the sample which produces the largest posterior probability. In implementing the CFTP algorithm, the Gibbs sampler is a natural choice for constructing the Markov chain. The conditional distributions required for the Gibbs sampler have the form

$$p(\gamma_i = 1 | \boldsymbol{\gamma}_{-i}, \mathbf{y}) = \left(1 + \exp\left(\sum_{j=1, j\neq i}^{q} \frac{1}{\sigma^2}\theta_i\theta_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\gamma_j - \frac{1}{\sigma^2}\theta_i\mathbf{x}_i^{\mathrm{T}}\mathbf{y}\right)\right)^{-1} \tag{5}$$

for $i = 1, 2, \ldots, q$, where $\boldsymbol{\gamma}_{-i}$ represents a vector of all the γ's except γ_i. In order to apply the sandwiched CFTP here, we need to verify first whether the update function of the Gibbs sampler preserves the monotonic property. The update function of the Gibbs sampler above can be shown as

$$\gamma_i = \begin{cases} 1 & \text{if } R_i \leq p(\gamma_i = 1 | \boldsymbol{\gamma}_{-i}, \mathbf{y}), \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where R_i is a random number from $U(0, 1)$. It follows that the monotonic property only exists if $\theta_i\theta_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$ for all i and j are all equal and nonpositive. One such instance of monotonicity is that all the predictors are orthogonal, that is, $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j = 0$ for all $i \neq j$. This case, however, is not of any interest because variable selection then is trivial. Also, the condition provided above is very stringent in that not many practical predictors satisfy the monotonic condition. Therefore the use of the sandwiched CFTP on variable selection seems really limited.

In fact, the sandwiched algorithm can always be used, regardless of the existence of monotonicity on variable spaces. This is so for as long as among all the CFTP chains there exist two chains which sandwich all the other chains in between at all times. By sandwiching, we mean that two chains always have the largest and the smallest probability of generating a 1 by the Gibbs sampling transition. In the variable selection problem, we notice that two chains initiated from all 1 state and all 0 state indeed always have the largest and the smallest probability to generating a 1 for as long as $\theta_i\theta_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$ for all i and j are nonpositive. The observation indicates that the sandwiched CFTP can be actually implemented in a larger variety of situations.

4.2. Variable selection by the Gibbs coupler

It was pointed out in Section 4.1 that the sandwiched CFTP algorithm can only be applied in special cases. A better perfect sampling solution to a general setting of the problem is provided by the Gibbs coupler. Recall that when applying the Gibbs coupler, the key issue is to determine the possible values that the Gibbs sampler might produce in the corresponding support. In [22, 23], detailed Gibbs coupler algorithms are constructed for problems modeled by Markov random fields where neighboring properties can be used to facilitate the computation. However here, no such neighboring properties exist. So the computation spent to determine the content of the support would be overwhelmingly expensive. Apparently, these algorithms are not suitable for solving the variable selection problem. Instead, to apply the Gibbs coupler, a special algorithm is needed for efficient determination of the support content.

Efficient determination of the support can be achieved by introducing *sandwich distributions* at every update. In particular, at any instant of time t, the *sandwich distributions* are defined by

$$\begin{aligned} &L_i^{(t)}(\gamma_i = 1) \\ &\quad \le p(\gamma_i = 1 | \boldsymbol{\gamma}_{-i}, \mathbf{y}) \le U_i^{(t)}(\gamma_i = 1), \quad i = 1, 2, \ldots, q \end{aligned} \tag{7}$$

with $\boldsymbol{\gamma}_{-i} \in S_{-i}^{(t)}$ where

$$S_{-i}^{(t)} = \{S_1^{(t)}, S_2^{(t)}, \ldots, S_{i-1}^{(t)}, S_{i+1}^{(t)}, \ldots, S_q^{(t)}\} \tag{8}$$

denotes the collection of supports of $\boldsymbol{\gamma}_{-i}$ at time t with the individual component supports at time t being $\{0\}$, $\{1\}$, or $\{0, 1\}$. We notice that definition (7) indicates that the two sandwich distributions bound all the probabilities of $\gamma_i = 1$ for every Markov chain in between. Therefore, if $L_i^{(t)}(\gamma_i = 1) > R_i^{(t)}$ (or $U_i^{(t)}(\gamma_i = 1) < R_i^{(t)}$), where $R_i^{(t)}$ is a uniform random seed for the ith update at time t, then samples from all the Markov chains will take values 1 (0) with certainty. Therefore, a sample of value 1 (0) is included as a unique value in the support. On the other hand, if the random seed is in between $L_i^{(t)}(\gamma_i = 1)$ and $U_i^{(t)}(\gamma_i = 1)$, that is, $L_i^{(t)}(\gamma_i = 1) < R_i^{(t)} < U_i^{(t)}(\gamma_i = 1)$, then the values that all the Markov chains can take will be uncertain to us. In this case, we leave the support $S_i^{(t)}$ as $\{0, 1\}$. According to these observations, the update of the support $S_i^{(t)}$ can be formulated as

$$S_i^{(t)} = \Phi\left(S_{-i}^{(t)}, R_i^{(t)}\right) = \begin{cases} \{1\}, & \text{if } R_i^{(t)} \le L_i^{(t)}(\gamma_i = 1), \\ \{0\}, & \text{if } R_i^{(t)} \ge U_i^{(t)}(\gamma_i = 1), \\ \{0, 1\}, & \text{otherwise.} \end{cases} \tag{9}$$

From (9), it can be seen directly that the probability for coalescence at time t is

$$p_t = \prod_{i=1}^{N} \left(L_i^{(t)}(x_i = 1) + \left(1 - U_i^{(t)}(x_i = 1)\right)\right). \tag{10}$$

Therefore, by recalling the condition of coalescence of the Gibbs coupler, we see that as long as $L_i^{(t)}(x_i = 1) \neq 0$ and $U_i^{(t)}(x_i = 1) \neq 1$, the proposed scheme can terminate almost surely and produce an unbiased sample from the posterior distribution. Furthermore, notice that the choice of sandwich distributions will affect the rate of coalescense. In order to get faster coalescence, we prefer a larger p_t, where p_t depends on the choice of $L_i^{(t)}(x_i)$ and $U_i^{(t)}(x_i)$, $i = 1, 2, \ldots, N$. In particular, among all distributions that satisfy (7), we prefer to choose the pair of distributions that entail the largest probability of coalescence p_t. Now, if the distributions are chosen according to

$$\begin{aligned} L_i^{(t)}(\gamma_i = 1) &= \min_{\boldsymbol{\gamma}_{-i}^{(t)} \in S_{-i}^{(t)}} \left\{p(\gamma_i = 1 | \boldsymbol{\gamma}_{-i}^{(t)}, \mathbf{y})\right\}, \\ U_i^{(t)}(\gamma_i = 1) &= \max_{\boldsymbol{\gamma}_{-i}^{(t)} \in S_{-i}^{(t)}} \left\{p(\gamma_i = 1 | \boldsymbol{\gamma}_{-i}^{(t)}, \mathbf{y})\right\}, \end{aligned} \tag{11}$$

they achieve the largest probability of coalescense, and hence their use leads to fastest coalescense.

In the implementation of the above scheme to variable selection, we need to determine the sandwich distributions from the full conditional distributions (5) for every instant t. Surprisingly, we find that the sandwich distributions (11) can easily be obtained only by checking the sign of $\theta_i \theta_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$. To be specific, the sandwiched distributions defined by (11) can be written as

$$\begin{aligned} &L_i^{(t)}(\gamma_i = 1) \\ &\quad = \left(1 + \exp\left(-\frac{1}{\sigma^2}\left(-\theta_i \mathbf{x}_i^{\mathrm{T}} \mathbf{y} + \sum_{k \in \mathbf{I}_{i1}^{(t-1)}} |\, \theta_i \theta_k \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k \,| \right.\right.\right. \\ &\qquad \left.\left.\left. + \sum_{k \in \mathbf{I}_{i3}^{(t-1)}} \theta_i \theta_k \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k \gamma_k^{(t-1)}\right)\right)\right)^{-1}, \end{aligned} \tag{12}$$

$$\begin{aligned} &U_i^{(t)}(\gamma_i = 1) \\ &\quad = \left(1 + \exp\left(-\frac{1}{\sigma^2}\left(-\theta_i \mathbf{x}_i^{\mathrm{T}} \mathbf{y} - \sum_{k \in \mathbf{I}_{i2}^{(t-1)}} |\, \theta_i \theta_k \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k \,| \right.\right.\right. \\ &\qquad \left.\left.\left. - \sum_{k \in \mathbf{I}_{i3}^{(t-1)}} \theta_i \theta_k \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k \gamma_k^{(t-1)}\right)\right)\right)^{-1}, \end{aligned} \tag{13}$$

where $\mathbf{I}_{i1}^{(t-1)}$, $\mathbf{I}_{i2}^{(t-1)}$, and $\mathbf{I}_{i3}^{(t-1)}$ are subsets of $\{1, 2, \ldots, i-1, i+1, \ldots, q\}$. The union $\mathbf{I}_{i1}^{(t-1)} \cup \mathbf{I}_{i2}^{(t-1)}$ contains the indices of the elements in $\{\gamma_k^{(t-1)}\}_{k=1, k\neq i}^{q}$ that have not coalesced at time $t-1$. However, the elements of $\mathbf{I}_{i1}^{(t-1)}$ have associated terms $\theta_i \theta_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j > 0$, whereas the elements of $\mathbf{I}_{i2}^{(t-1)}$ have associated terms $\theta_i \theta_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \le 0$. Additionally, $\mathbf{I}_{i3}^{(t-1)}$ contains the indices of the remaining elements in $\{\gamma_i^{(t-1)}\}_{k=1, k\neq i}^{q}$ (the ones that have coalesced at time $t-1$). Thus, at any time t, the algorithm updates the support of the ith component according to the function (9). The coalesced state at time 0 is then a perfect sample from (2).

5. DISCUSSION

In the previous sections, we have shown how we address variable selection with perfect sampling when the model coefficients and the noise variance are all known. In this section, we discuss the possibilities of applying perfect sampling when these parameters are not known.

When the θ's and the σ^2 are all unknown continuous and unbounded variables, a major difficulty is the drawing of perfect samples. Currently, most of the algorithms for perfect sampling are designed for discrete variable spaces. A theoretical realization of the perfect sampling, especially the Gibbs coupler on unbounded continuous state spaces remains problematic. One approximate solution could be based on quan-

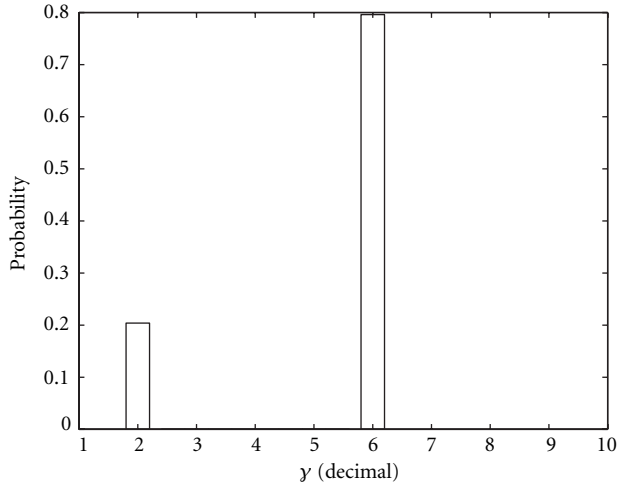


Figure 1: A histogram of 1000 samples of $\boldsymbol{\gamma}$ obtained by the Gibbs coupler in experiment 1. The true decimal value of $\boldsymbol{\gamma}$ is 6.

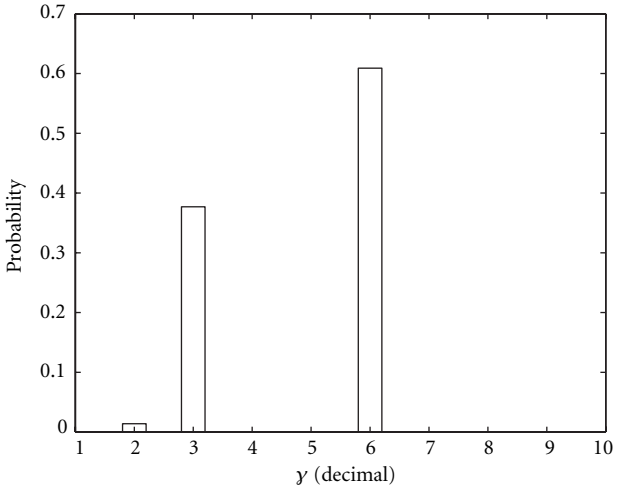


Figure 2: A histogram of 1000 samples of $\boldsymbol{\gamma}$ obtained by the Gibbs coupler in experiment 2. The true decimal value of $\boldsymbol{\gamma}$ is 6.

tizing the spaces with desired accuracy. The so-defined discrete spaces can further be expressed on binary spaces. Note that when the θ's and the σ^2 are all binary, the Gibbs coupler algorithm discussed in the paper can readily be applied where the sandwich distributions can be calculated in the same way as was discussed in Section 4. Consequently, if we perform quantization as is commonly done in digital signal processing and communications, perfect sampling is possible. The complexity of the perfect sampling implementation would depend on the number of quantization levels.

Often it is also desirable to obtain samples of the θ's and σ^2 directly from their continuous spaces. In theory, the generation of perfect samples by a Gibbs coupler algorithm is possible if both the θ's and σ^2 are bounded. One can use the rejection coupler [18] for the component coupling requirement of the Gibbs coupler. Then we might argue that after seeing the data, the state spaces can be bounded. Thus, perfect sampling of the unbounded θ's and σ^2 can always be tackled by bounding the variable spaces of the unknowns. In practice, however, when the dimension of the predictors, q, is large or the bounds are very loose, the coalescing time of the Gibbs coupler could be very long.

6. SIMULATION RESULTS

We tested the Gibbs coupler in several experiments. In the first experiment, $q = 5$. The predictor vectors $\mathbf{x}_i$ were independent and identically distributed according to a multivariate Gaussian density, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and their corresponding coefficients were chosen as $\theta_1 = \theta_2 = 0.8$, $\theta_3 = \theta_4 = 0.7$, and $\theta_5 = 0.6$. Fifty data records were generated by

$$\mathbf{y} = \mathbf{x}_3\theta_3 + \mathbf{x}_4\theta_4 + \boldsymbol{\epsilon}, \tag{14}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Apparently, the true model included only two variables with $\boldsymbol{\gamma}^{\mathrm{T}} = [0\ 0\ 1\ 1\ 0]$. A histogram of 1000 samples obtained by the Gibbs coupler is plotted in Figure 1,

Table 1: Number of trials with wrong solutions out of 100 trials by the Gibbs coupler and the least squares methods in the third experiment.

	Exp. 1	Exp. 2
Gibbs coupler	2	13
LS	7	44

where for convenience, the value of $\boldsymbol{\gamma}$ is represented by its corresponding decimal integer. (For example, $\boldsymbol{\gamma}^{\mathrm{T}} = [0\ 1\ 1\ 1\ 1]$ is represented by 15 and $\boldsymbol{\gamma}^{\mathrm{T}} = [1\ 1\ 1\ 1\ 1]$ by 31.) From the histogram, it is clear that the highest probability occurs at 6, or the MAP estimator is $\boldsymbol{\gamma}^{\mathrm{T}} = [0\ 0\ 1\ 1\ 0]$, which is the true setting of $\boldsymbol{\gamma}$.

In the second experiment, the only change was made in the definition of $\mathbf{x}_3$. There $\mathbf{x}_3 = \mathbf{x}_5 + 0.15\mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which introduced correlation between the third and fourth predictors (with a correlation coefficient equal to 0.989). A histogram of 1000 samples obtained by the Gibbs coupler is plotted in Figure 2. The largest value of the histogram occurs at $\boldsymbol{\gamma}$ (decimal) equal to 6, that is, $\boldsymbol{\gamma}^{\mathrm{T}} = [0\ 0\ 1\ 1\ 0]$, which is again the true value of $\boldsymbol{\gamma}$.

Next, in the third experiment, we compared the performance of the Gibbs coupler solution with the "least squares" solution. If $\mathbf{H} = [\mathbf{x}_1\theta_1\ \mathbf{x}_2\theta_2\ \cdots\ \mathbf{x}_5\theta_5]$, the ith component of the least squares solution $\hat{\boldsymbol{\gamma}}_{\mathrm{LS}}$ is defined by

$$\hat{\gamma}_i = \begin{cases} 1, & z_i \geq 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where z_i is the ith component of $\mathbf{z}$ which is computed by

$$\mathbf{z} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}. \tag{16}$$

We repeated the first and second experiments 100 times, respectively. In each trial, the Gibbs coupler and the least

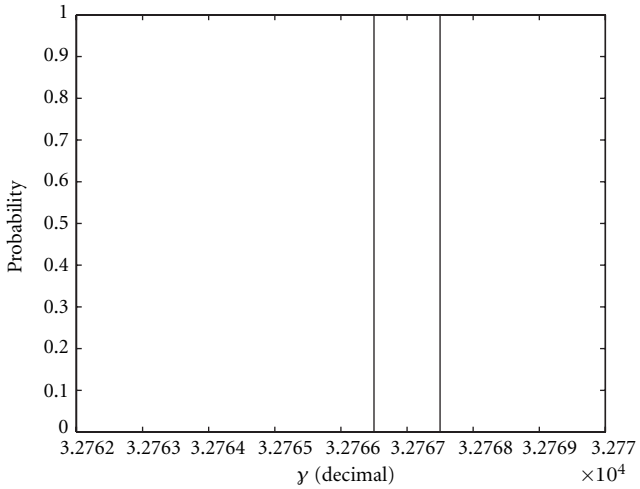


FIGURE 3: A histogram of 1000 samples of $\boldsymbol{\gamma}$ obtained by the Gibbs coupler in experiment 4. The true decimal value of $\boldsymbol{\gamma}$ is 32767.

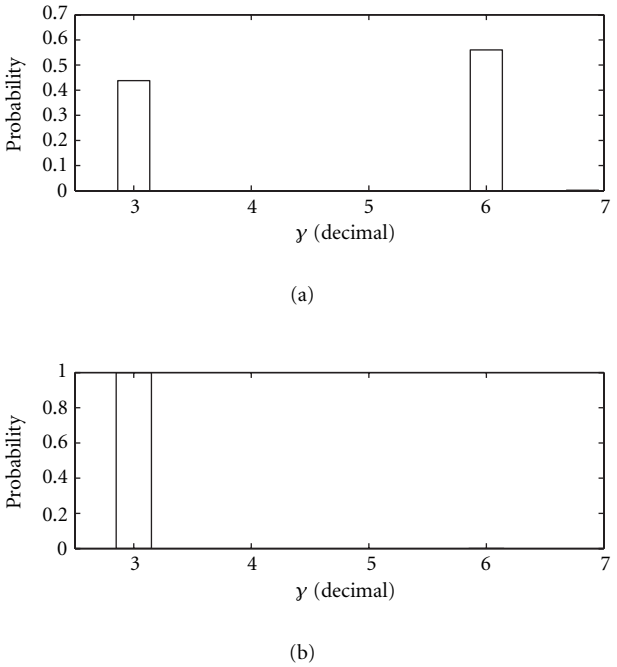


FIGURE 4: (a) A histogram of 500 samples of $\boldsymbol{\gamma}$ obtained by the Gibbs coupler in experiment 5. (b) A histogram of 500 samples of $\boldsymbol{\gamma}$ obtained by the Gibbs sampler in experiment 5. The true decimal value of $\boldsymbol{\gamma}$ is 6.

squares solutions were calculated. The number of trials with incorrect solutions by each algorithm was recorded. The results are displayed in Table 1. Notice that in both experiments, the Gibbs coupler performed better than the least squares method. The improved performance of the Gibbs coupler is especially emphasized in the second experiment, where it was more difficult to choose the right predictors due to their correlation.

In the fourth experiment, a scenario with $q = 20$ predictors was tested. The predictors were defined by $\mathbf{x}_i = \mathbf{u}_i + \mathbf{w}$, where the $\mathbf{u}_i$'s were independent and identically distributed according to a multivariate Gaussian density $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The vector $\mathbf{w}$ was independent from the $\mathbf{u}_i$'s and came from the same distribution. As a result, the pairwise correlation between the predictors was 0.5. In addition, the model coefficients were chosen as follows: $\theta_1 = \theta_2 = \cdots = \theta_5 = 1.5$, $\theta_6 = \theta_7 = \cdots = \theta_{10} = 1$, $\theta_{11} = \theta_{12} = \cdots = \theta_{15} = 2$, and $\theta_{16} = \theta_{17} = \cdots = \theta_{20} = 3$. A data vector of size 120 was generated according to the model $\mathbf{y} = \sum_{i=6}^{20} \theta_i \mathbf{x}_i + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma = 2$. Therefore,

$$\boldsymbol{\gamma}^{\mathrm{T}} = [0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1] \qquad (17)$$

in this experiment. 1000 samples were drawn by the Gibbs coupler. A histogram of the samples is displayed in Figure 3, and it shows that all the samples of $\boldsymbol{\gamma}$ took on the value 32767, which corresponds to the true binary value of $\boldsymbol{\gamma}^{\mathrm{T}}$, [0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1].

Finally, in the last experiment, we performed a comparison between the Gibbs coupler and the Gibbs sampler. The objective of the experiment was to stress the advantages of using perfect sampling over MCMC. We adopted the settings of the second experiment. Note that in that experiment there is a high correlation between the third and the fifth predictor. We ran the Gibbs coupler and collected 500 samples. The histogram of the drawn samples is plotted in Figure 4a. We clearly see two high peaks at 3 and 6. The binary representation of 3 and 6 is [0 0 0 1 1] and [0 0 1 1 0]. Note that the decimal value 6 is the right value for $\boldsymbol{\gamma}$. However, 3 is the value when the fifth predictor is mistaken by the third predictor. In this experiment, the Gibbs coupler was able to identify the true set of predictors. Next, we ran the Gibbs sampler on the same data. We chose the initial value of the Gibbs sampler to be 3. Again, 500 samples were gathered. The corresponding histogram is depicted in Figure 4b. The figure clearly shows that during the 500 runs, the Gibbs sampler is trapped at 3. Therefore the Gibbs sampler method cannot identify the right set of the predictors with 500 samples. To improve the performance, more samples are required. Apparently, the Gibbs coupler algorithm is more efficient in the sense that it requires less samples to achieve the same performance.

7. CONCLUSION

We have proposed the use of two perfect sampling algorithms for variable selection. They are the sandwiched CFTP algorithm and the Gibbs coupler. An efficient implementation of the CFTP algorithm requires construction of monotonic Markov chains, whereas the Gibbs coupler is efficient without the requirement of monotonicity. To implement the Gibbs coupler algorithm, we designed a detailed efficient coupling scheme. Simulation results showed good performance of the perfect sampling algorithms. In addition, we also discussed the possibilities and difficulties to employ the perfect sampling algorithms in more general settings of the problem.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments and suggestions.

This work was supported by the National Science Foundation under Award No. CCR-9903120.

REFERENCES

[1] M. K. Tsatsanis and G. B. Giannakis, "Time-varying system identification and model validation using wavelets," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3512–3523, 1993.

[2] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, pp. 387–392, 1985.

[3] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, Indiana, 1990.

[4] K. Funahashi, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

[5] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[6] A. Hoover, G. Jean-Baptiste, X. Jiang, et al., "An experimental comparison of range image segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, 1996.

[7] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[8] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.

[9] S. P. Brooks, "Markov chain Monte Carlo method and its application," *The Statistician*, vol. 47, no. 1, pp. 69–100, 1998.

[10] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *J. Amer. Statist. Assoc.*, vol. 85, pp. 398–409, 1990.

[11] M. A. Tanner, *Tools for Statistical Inference*, Springer-Verlag, New York, 1996.

[12] P. M. Djurić, "Variable selection by a reversible jump MCMC approach," in *Proc. of European Signal Processing Conference*, vol. 4, pp. 2013–2016, Rhodes, Greece, 1998.

[13] E. I. George and R. McCulloch, "Variable selection via Gibbs sampling," *J. Amer. Statist. Assoc.*, vol. 88, pp. 881–889, 1993.

[14] J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures Algorithms*, vol. 9, pp. 223–252, 1996.

[15] J. Møller, "Perfect simulation of conditionally specified models," *J. Roy. Statist. Soc. Ser. B*, vol. 61, no. 1, pp. 251–264, 1999.

[16] M. Fismen, "Exact simulation using Markov chains," Tech. Rep., Department of Mathematics, Norwegian University of Science and Technology, 1997.

[17] E. Thönnes, "A primer on perfect simulation," *http://dimacs.rutgers.edu/˜dbwilson/exact*, 2000.

[18] D. J. Murdoch and P. J. Green, "Exact sampling from a continuous state space," *Scand. J. Statist.*, vol. 25, pp. 483–502, 1998.

[19] Y. Huang and P. M. Djurić, "Multiuser detection of synchronous Code-Division-Multiple-Access signals by the Gibbs coupler," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, UT, 2001.

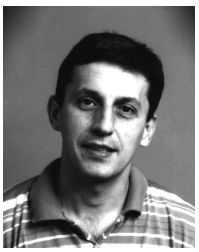
[20] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 1999.

[21] P. J. Green and D. J. Murdoch, "Exact sampling for Bayesian inference: towards general purpose algorithms," in *Bayesian Statistics*, J. M. Bernardo, J. O. Gerger, A. P. Dawid, and A. F. M. Smith, Eds., vol. 6, pp. 301–321, Oxford University Press, 1999.

[22] O. Häggström and K. Nelander, "On exact simulation of Markov random fields using coupling from the past," *Scand. J. Statist.*, vol. 26, no. 3, pp. 395–411, 1999.

[23] M. Huber, "Exact sampling and approximate counting techniques," in *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing*, pp. 31–40, 1998.

Yufei Huang was born in Shanghai, China, in 1973. He received the B.S. degree in applied electronics from Northwestern Polytechnic University, Xi'an, China in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York at Stony Brook, Stony Brook, NY in 1997 and 2001, respectively. He is now working as a post-doctoral researcher in the Department of Electrical and Computer Engineering, State University of New York at Stony Brook. His current research interests are in the theory of Monte-Carlo methods and their applications to array processing and multi-user communications.



Petar M. Djurić, Professor in the Department of Electrical and Computer Engineering at the State University of New York at Stony Brook, works in the area of statistical signal processing. He got his B.S. and M.S. degrees in Electrical Engineering from the University of Belgrade, Yugoslavia, in 1981 and 1986, respectively, and the Ph.D. degree in Electrical Engineering from the University of Rhode Island, US, in 1990. His primary interests are in the theory of modeling, detection, estimation, and time series analysis and its application to a wide variety of disciplines, including telecommunications, bio-medicine, and power engineering. Prof. Djurić has served on numerous Technical Committees for the IEEE and SPIE, and has been invited to lecture at universities in the US and overseas. He has been Associate Editor of the IEEE Transactions on Signal Processing, and currently he is the Vice Chair of the IEEE Signal Processing Society Committee on Signal Processing—Theory and Methods. He is also a Treasurer of the IEEE Signal Processing Conference Board, and a Member of the American Statistical Association and the International Society for Bayesian Analysis.