BAYESIAN LEARNING FOR SPARSE SIGNAL RECONSTRUCTION

David P. Wipf and Bhaskar D. Rao

University of California, San Diego
Department of Electrical and Computer Engineering
La Jolla, CA 92093-0407 USA
e-mail: {dwipf, brao}@ece.ucsd.edu

ABSTRACT

Sparse Bayesian learning and specifically relevance vector machines have received much attention as a means of achieving parsimonious representations of signals in the context of regression and classification. We provide a simplified derivation of this paradigm from a Bayesian evidence perspective and apply it to the problem of basis selection from overcomplete dictionaries. Furthermore, we prove that the stable fixed points of the resulting algorithm are necessarily sparse, providing a solid theoretical justification for adapting the methodology to basis selection tasks. We then include simulation studies comparing sparse Bayesian learning with Basis Pursuit and the more recent FOCUSS class of basis selection algorithms, empirically demonstrating superior performance in terms of average sparsity and success rate of recovering generative bases.

1. INTRODUCTION

Sparse signal representations from overcomplete dictionaries find increasing relevance in a large number of application domains [1, 2]. Moreover, attaining such representations is tantamount to solving regularized linear inverse problems that have far-reaching significance. Consequently, deeper insight into these issues is of both theoretical and practical importance. The canonical form of this problem is given by,

$$t = \Phi w + \epsilon, \tag{1}$$

where $\Phi \in \Re^{N \times M}$ is a matrix whose columns represent a possibly overcomplete basis ($M \gg N$), w is the vector of weights to be learned, ϵ is noise, and t is a vector of targets. In this vein, we seek to find weight vectors whose entries are predominantly zero.

When Φ is selected such that $\Phi_{i,j} = K(x_i, x_j)$ for training vectors x_i, x_j and kernel function $K(\cdot, \cdot)$ satisfying Mercer's condition, we recover the standard support vector machine (SVM) model. While successful for classification and regression problems, however, SVMs are inadequate for finding sparse signal representations from possibly overcomplete bases. In fact, SVM discriminate functions are usually only quasi-sparse, with the number of support vectors growing steeply with the size of the training set [3]. This is partially due to the fact that all outliers typically become support vectors. Moreover, SVMs unfortunately require the estimation of a trade-off parameter.

More recently, relevance vector machines (RVM) have been fashioned from a Bayesian perspective to address these limitations to SVMs [3]. Although initially developed for regression and classification problems, RVMs, or more generally, the sparse Bayesian learning (SBL) framework, provide a viable candidate for finding sparse signal representations. In this paper, we prove that the SBL cost function leads to sparse solutions of underdetermined linear inverse problems, providing solid theoretical justification for adapting it to basis selection tasks. Furthermore, we empirically substantiate the algorithm by comparing it with Basis Pursuit [2] (which finds minimum ℓ_1-norm solutions via linear programming (LP)) and the FOCUSS class of algorithms [4, 5] (which find minimum p-norm-like ($\ell_{(p \leq 1)}$) solutions via gradient factorizations). However, first we will introduce a simplified derivation of the SBL algorithm as a model selection tool based on maximizing Bayesian evidence.

2. SPARSE BAYESIAN LEARNING

In contrast to the statistical learning theory that underlies SVMs, SBL arises from a probabilistic perspective. In [3], SBL is presented/derived as an approximation of the posterior distribution of all unknowns given the data. Herein, we derive the SBL cost function as an exact evaluation of the Bayesian evidence. First, we will describe the two levels of Bayesian inference that motivate SBL and subsequently, we will detail how appropriate sparsifying weight priors are estimated from the data.

This research was partially supported by the National Science Foundation Grant No. CCR-9902961 and DiMI grant #22-8376 sponsored by Nissan.

2.1. Levels of Bayesian Inference

Per the discussion in [6], a statistical model $\mathcal{H}$ of data $\mathcal{D}$ is characterized by three components: (i) a functional form parameterized by some weights w, (ii) a prior distribution over these weights, $p(w|\mathcal{H})$, and (iii) the distribution of the data given the model and fixed weights, i.e., the likelihood $p(\mathcal{D}|w,\mathcal{H})$. As we will see later, the power of modern Bayesian inference lies in its ability to optimally select (ii) and (iii).

The first level of inference assumes a given model $\mathcal{H}$ is fixed and deduces the parameters w by maximizing the posterior weight density $p(w|\mathcal{D},\mathcal{H}) \propto p(\mathcal{D}|w,\mathcal{H})p(w|\mathcal{H})$. The normalizing term $p(\mathcal{D}|\mathcal{H})$ (referred to as the evidence for $\mathcal{H}$) is not needed since it is independent of w. But how do we know which model $\mathcal{H}$ is most appropriate?

The second level of Bayesian inference involves comparing models/hypotheses, e.g., $\mathcal{H}_1$ and $\mathcal{H}_2$, with respect to the data. This is accomplished by evaluating the evidence $P(\mathcal{D}|\mathcal{H}_i)$, which can be computed by integrating over the weights:

$$P(\mathcal{D}|\mathcal{H}_i) = \int P(\mathcal{D}|w,\mathcal{H}_i)P(w|\mathcal{H}_i)dw. \qquad (2)$$

Models with excessive weights will typically have sharply peaked likelihoods $P(\mathcal{D}|w,\mathcal{H}_i)$ relative to the size of the weight space. Furthermore, if the prior $P(w|\mathcal{H}_i)$ is more or less uniformly distributed, then large weight spaces will be characterized by small prior densities in any given region. Thus, when we integrate (2) we obtain a small value for the evidence.

In contrast, if we have significantly fewer weights, then the likelihood becomes less sensitive to the exact weight values and $P(\mathcal{D}|w,\mathcal{H}_i)$ tends to spread more about its peak value. Moreover, the smaller weight space will also have a higher density producing a larger evidence upon integration.

Thus, the evidence favors models that are parsimonious and less sensitive to finely tuned weights.

2.2. Model Selection

Like SVMs, the functional form of our statistical model $\mathcal{H}$ remains fixed as in (1). Thus, most of our remaining effort in model selection must focus on finding an appropriate weight prior $p(w|\mathcal{H})$ and the likelihood $p(\mathcal{D}|w,\mathcal{H})$. We should note that the modern Bayesian methodology does not attempt to select the 'right' priors or the FOCUSS approach of selecting a fixed sparsity inducing prior. Rather, many different priors can be compared corresponding to different hypothesis about underlying truth. These hypothesis can be empirically compared by evaluating the evidence for each model [6].

So how is this accomplished? Following the reasoning in [3], we first handle the likelihood by assuming a gaussian noise model with unknown variance. Then we assume targets t are distributed as $p(t|w,\mathcal{H}) \propto \mathcal{N}(t|\Phi w, \sigma^2 I)$ where σ^2 is unknown . We must now select an appropriate form for the weight prior that reflects a preference for less complex functions, e.g.,

$$p(w|\mathcal{H}) = p(w|\gamma) = \prod_{i=1}^{M} \mathcal{N}(w_i|0,\gamma_i), \qquad (3)$$

where γ is a vector of M hyperparameters controlling the prior variance of each weight. Also, we may specify a hyperprior $p(\gamma_i)$ on each γ_i if we so desire. We are now in a position to formulate the evidence for each candidate model distinguished by γ and σ^2,

$$\begin{aligned} p(\mathcal{D}|\mathcal{H}) &= p(t|\gamma,\sigma^2) \\ &= \int p(t|w,\sigma^2)p(w|\gamma)dw \\ &= (2\pi)^{-N/2}\,|\Sigma_t|^{-1/2} \exp\left[-\frac{1}{2}t^T\Sigma_t^{-1}t\right] \end{aligned} \qquad (4)$$

where $\Sigma_t \triangleq \sigma^2 I + \Phi\Gamma\Phi^T$ and we have introduced the notation $\Gamma \triangleq \text{diag}\{\gamma\}$. The greater the evidence $p(t|\gamma,\sigma^2)$, the more plausible γ and σ^2, which collectively demarcate $\mathcal{H}$.

2.3. Sparse Bayesian Learning Algorithm

With SBL, the first level of Bayesian inference is trivial by design; given the gaussian weight priors from (3), the posterior density of the weights is gaussian, i.e.,

$$p(w|t,\gamma,\sigma^2) \propto \mathcal{N}(\mu,\Sigma), \qquad (5)$$

with $\mu = \sigma^{-2}\Sigma\Phi^T t$ and $\Sigma = (\sigma^{-2}\Phi^T\Phi + \Gamma^{-1})^{-1}$. Thus, the onus remains in estimating γ and σ^2. To accomplish this, we employ the EM algorithm to maximize $P(\gamma,\sigma^2|t) \propto P(t|\gamma,\sigma^2)p(\gamma)p(\sigma^2)$ = evidence $\times$ hyperprior. This produces the update rule

$$\gamma_i^{\text{new}} = \Sigma_{i,i} + \mu_i^2, \qquad (6)$$

which is iterated until convergence. Likewise, an update rule for σ^2 can be simply derived [3]. Also, when Φ is formed from kernel functions, we obtain RVMs.

3. GLOBAL CONVERGENCE TO SPARSE SOLUTIONS

Sparse solutions are formally equivalent to the basic solutions in LP, i.e., solutions with at most $N \ll M$ nonzero entries. In this section, we prove that all stable fixed points of the SBL algorithm outlined above are sparse solutions. To accomplish this we express our cost function $\mathcal{L}$ as the log of $P(\gamma,\sigma^2|t)$ giving,

$$\mathcal{L} = -\frac{1}{2}\log|\Sigma_t| - \frac{1}{2}t^T\Sigma_t^{-1}t \tag{7}$$

where both terms come from the evidence and we have assumed a uniform hyperprior on γ. Before we can proceed further, we must introduce three intermediate results:

Lemma 1: $-\log|\Sigma_t|$ is convex with respect to Γ (or equivalently γ).

Proof: By definition, Σ_t is an affine transformation of Γ. Furthermore, in the space of psd matrices, $-\log|x|$ is a convex function of x (see e.g., [7]). Thus, it remains to show that a convex function of an affine transformation is convex (assuming the transformation preserves positive semidefiniteness).

Let $f : \Re^{N\cdot N} \rightarrow \Re$ and $g : \Re^M \rightarrow \Re^{N\cdot N}$ be functions such that $h \triangleq f \circ g : \Re^M \rightarrow \Re$, with f convex. By application of the chain rule for vectors, we can compute H^h, the Hessian of h, as a function of H^f and $H^{g_n}, n \in \{1,\ldots,N\cdot N\}$, which represent the Hessians of f and g respectively:

$$H^h = \sum_{m=1}^{N\cdot N}\sum_{n=1}^{N\cdot N}\left[\frac{\partial g_m}{\partial x}\frac{\partial g_n}{\partial x}^T H^f_{m,n} + H^{g_n}\text{diag}\left(\frac{\partial f}{\partial g_n}\right)\right]. \tag{8}$$

Now let $g(x) = Ax + b$. Then $H^{g_n} = 0\ \forall n$ and $\frac{\partial g_m}{\partial x} \triangleq w_m$ is constant $\forall x$. We may then rewrite H^h as,

$$H^h = \sum_{m=1}^{N\cdot N}\sum_{n=1}^{N\cdot N} w_n w_m^T H^f_{m,n}. \tag{9}$$

It is then easy to show that H^h is psd since, for any $z \in \Re^M$, we have

$$\begin{aligned} z^T H^h z &= z^T\left(\sum_{m=1}^{N\cdot N}\sum_{n=1}^{N\cdot N} w_n w_m^T H^f_{m,n}\right) z \\ &= \sum_{m=1}^{N\cdot N}\sum_{n=1}^{N\cdot N}(z^T w_n)(w_m^T z)H^f_{m,n} \\ &= \sum_{m=1}^{N\cdot N}\sum_{n=1}^{N\cdot N} a_n a_m H^f_{m,n} \\ &= a^T H^f a \geq 0, \end{aligned} \tag{10}$$

where $a_n \triangleq w_n^T z \in \Re$, $a = [a_1,\ldots,a_{NN}]^T$, and the inequality follows since f is convex. Therefore H^h is psd everywhere and consequently h is convex.

Lemma 2: The term $t^T\Sigma_t^{-1}t$ is constant over all γ satisfying the N linear constraints $b = A\gamma$ if $b = \Phi\mu^*$ for some μ^* as defined in (5) and $A = \Phi\text{diag}\{\Phi^T(t-b)\}$.

Proof: By the matrix inversion lemma,

$$t^T\Sigma_t^{-1}t = \sigma^{-2}t^T(t - \Phi\mu). \tag{11}$$

Therefore, the constraint $\Phi\mu = b$ for some b clearly holds $t^T\Sigma_t^{-1}t$ constant. Moreover, it can be shown that over this constraint surface, γ and μ are linearly related by

$$\mu = \Gamma\Phi^T(t-b) = \text{diag}\{\Phi^T(t-b)\}\gamma, \tag{12}$$

completing the proof.

Lemma 3: Every local maximum of $\mathcal{L}$ is a sparse.

Proof: The proof is by contradiction. Assume for the moment that $\gamma = \gamma^*$ is a non-sparse local maximum of $\mathcal{L}$. Now consider the optimization problem,

$$\begin{aligned} \max: &\quad -\log|\Sigma_t| \\ \text{subject to:} &\quad A\gamma = b,\ \ \gamma \geq 0, \end{aligned} \tag{13}$$

where A and b are as defined before. From Lemma 2, the above constraints hold $t^T\Sigma_t^{-1}t$ constant defined on a closed, bounded convex polytope (i.e., we are maximizing (7) while holding the second term constant). Also, Lemma 1 dictates that the objective function is convex. If γ^* is truly a local maximum of $\mathcal{L}$, then a maxima of (13) should be $\gamma = \gamma^*$ as well. However, from [8] Theorem 6.5.3, all maxima of (13) are achieved at extreme points and additionally, Theorem 2.5 establishes the equivalence between extreme points and basic feasible solutions, i.e., solutions with at most N nonzero values. Thus, if γ^* is not a basic feasible solution, it cannot be a maximum of (13) and therefore cannot be a local maximum of $\mathcal{L}$. Consequently, all local maxima must be sparse.

Theorem: The SBL algorithm is globally convergent and the stable fixed points are sparse solutions.

Proof: The EM algorithm is provably convergent to fixed points. Furthermore, from Lemma 3, all stable fixed points are sparse, completing the proof.

4. RESULTS

To quantify the performance of SBL relative to other methods, we completed a simulation study of each method on synthetic data. For simplicity and ease of comparison, noiseless tests were performed first. This facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of the trade-off parameter (which balances sparsity and quality of fit) in the case of FOCUSS and Basis Pursuit [2, 9]. To accommodate the low noise case with SBL, we approximated the low noise limit by setting σ^2 to a small fixed value (10^{-8}). Experiment I below details these results. Experiment II involves a low-noise sparse filtering application. In all cases, FOCUSS weights were initialized using the ℓ_2-norm solution while SBL was initialized with small random γ values. Basis Pursuit was performed using the default Matlab $Linprog$ command.

4.1. Experiment I

Consistent with [4], we generated a random $N \times M$ Φ matrix whose entries where each drawn from a standardized gaussian distribution. The columns were then normalized to unit ℓ_2-norm. Sparse weight vectors w were randomly generated with R nonzero entries. The vector of target values is then computed as $t = \Phi w$. Each algorithm is then presented with t and Φ and attempts to learn w.

Fig. 1(a) below depict results from these tests with $R = 7$. We set $N = 20$ and M is varied from 30 to 100 allowing the overcompleteness ratio, i.e, M/N, to vary from 1.5 to 5.0.

The SBL algorithm clearly outperforms the others both in terms of success rate, i.e, the percentage of time the correct (and only correct) bases are selected, and in terms of average sparsity (not shown). Also, the higher success rates acquire added significance in the noiseless case because, given the conditions in [5], these are provably the minimum sparsity solutions.

One possible explanation for the performance discrepancy could be the different initialization strategies used with each method. We can dismiss this possibility however when we consider that the Basis Pursuit method achieves a global maximum of the ℓ_1-norm and is thus impervious to initial conditions. Thus, the superiority of SBL cannot be in its ability to consistently reach global maxima, but rather in the larger correlation between maxima in its cost function and maximally sparse solutions.

4.2. Experiment II

To explore a slightly more realistic scenario, we applied each of the above algorithms (without any accommodations for noise, e.g., the SBL algorithm was not allowed to adapt σ^2) to a low-noise filtering problem, namely, the recovery of sparse FIR filter weights. For this example, the Φ matrix was composed of delayed versions of an input white noise sequence distributed as $\mathcal{N}(0, 1)$. Random weights were generated as before and the targets were computed as $t = \Phi w + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.01 \cdot I)$. The results shown in Fig. 1(b) clearly show the superior robustness of SBL.

5. CONCLUSIONS

In this paper, we have motivated the SBL cost function as a vehicle for finding models with maximal Bayesian evidence. We have also proven that the local maxima of this cost function are necessarily sparse. Furthermore, our simulation studies clearly indicate uniformly superior performance over popular methods while retaining the desired theoretical optimality qualities the of FOCUSS class of algorithms. As such, we have demonstrated that SBL is a viable candidate for sparse signal reconstructions.

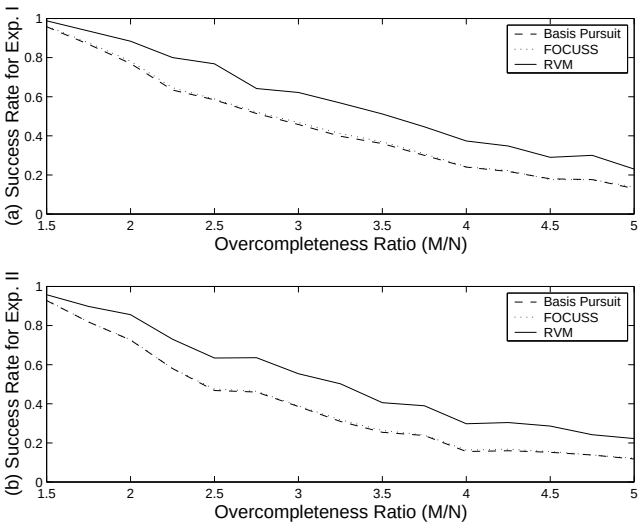

Fig. 1. Results with true sparsity = 7; (a) Experiment I (noiseless), (b) Experiment II (noisy)

6. REFERENCES

[1] B. D. Rao, "Signal processing with the sparseness constraint," *Proc. ICASSP*, vol. 3, pp. 1861–1864, May 1998.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[3] Michael E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning*, vol. 1, pp. 211–244, 2001.

[4] Bhaskar D. Rao and Kenneth Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, January 1999.

[5] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Trans. on Signal Processing*, vol. 45, no. 3, pp. 600–616, March 1997.

[6] David J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[8] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison–Wesley, Reading, Mass., 1984.

[9] B.D. Rao, *et. al.*, "Subset selection in noise based on diversity measure minimization," *To Appear in IEEE Trans. Signal Processing*, 2003.