BAYESIAN L_1-NORM SPARSE LEARNING

Yuanqing Lin, Daniel D. Lee

GRASP Laboratory, Department of Electrical and Systems Engineering,
University of Pennsylvania, Philadelphia, PA 19104

ABSTRACT

We propose a Bayesian framework for learning the optimal regularization parameter in the L_1-norm penalized least-mean-square (LMS) problem, also known as LASSO [1] or basis pursuit [2]. The setting of the regularization parameter is critical for deriving a correct solution. In most existing methods, the scalar regularization parameter is often determined in a heuristic manner; in contrast, our approach infers the optimal regularization setting under a Bayesian framework. Furthermore, Bayesian inference enables an independent regularization scheme where each coefficient (or weight) is associated with an independent regularization parameter. Simulations illustrate the improvement using our method in discovering sparse structure from noisy data.

1. INTRODUCTION

Finding a sparse solution of a least-mean-square (LMS) problem is key to many applications in signal processing [1][2][3][4]. An effective approach for deriving sparse LMS solution is L_1-norm regularization [1][2][5], and the optimization problem is:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi w}\|^2 + \hat{\lambda}\sum_{i=1}^{M}|w_i|, \qquad (1)$$

where $\mathbf{y}$ is an $N \times 1$ data vector, $\mathbf{\Phi}$ is an $N \times M$ matrix, $\mathbf{w}^*$ is the $M \times 1$ weight vector that needs to be optimized, and $\hat{\lambda}$ is the regularization parameter that balances favoring the LMS fit versus the sparseness of the solution described by the L_1-norm.

Although the setting of the regularization parameter $\hat{\lambda}$ in Eq. 1 is critical for deriving a correct solution, it is often determined heuristically. For instance, for the special case where the columns of $\mathbf{\Phi}$ are orthogonal, S. S. Chen *et. al* [2] speculated that $\hat{\lambda} = \sigma\sqrt{2\log(M)}$ with σ being the noise level in amplitude; J.J. Fuchs [4] argued that $\hat{\lambda}$ should be proportional to the noise level and signal level; D. M. Malioutov *et. al* [6] considered solving a piece-wise linear problem with respect to $\hat{\lambda}$ to derive a complete set of possible solutions, from which one may select an appropriate solution according to some empirical criterion. However, the main drawback of these approaches originates from their lack of a generalized criterion for sparsity. In contrast, our approach models sparsity in a Bayesian framework and the optimal regularization parameter is inferred by maximizing the posterior distribution of the regularization parameter.

In our approach, we are able to deal with a more general form than Eq. 1, namely,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi w}\|^2 + \sum_{i=1}^{M}\hat{\lambda}_i|w_i|, \qquad (2)$$

where each element in $\mathbf{w}$ is associated with an independent regularization parameter. We will refer to Eq. 2 as independent regularization, and Eq. 1 as unform regularization. The extension in Eq. 2 was inspired by M. E. Tipping's work [7] which found that independent L_2-norm regularization was able to yield a much sparser solution than uniform L_2-norm regularization. Therefore, we can expect that Eq. 2 will yield stronger sparsity regularization than Eq. 1 if the regularization parameters are optimally inferred.

The remainder of this paper is organized as follows. In Section 2, Expectation-Maximization (EM) type update rules are derived from a Bayesian framework for iteratively estimating the optimal regularization parameters. In Section 3, we employ simulations to demonstrate the advantage of the Baysesian L_1-norm sparse learning. Finally, a brief discussion of these results is presented in Section 4.

2. BAYESIAN FRAMEWORK FOR L_1-NORM SPARSE LEARNING

In this section, the EM type update rules are derived for iteratively estimating the regularization parameter in Eq. 1. In the EM procedure, we introduce a variational method for overcoming the difficulties in inference with non-Gaussian probability distributions. To compute the mode of the distribution for its variational approximation, we present an algorithm for solving the L_1-norm penalized LMS problem via auxiliary function minimization. Then, we extend the EM procedure for computing the independent regularization parameters in Eq. 2.

In the probabilistic model for Eq.1, the data $\mathbf{y}$ are assumed to be coupled with additive I.I.D. zero-mean Gaussian noise,

namely,

$$P(\mathbf{y}|\mathbf{\Phi},\mathbf{w},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\|^2\right), \tag{3}$$

and the prior on the weights is a Laplacian distribution,

$$P(\mathbf{w}|\boldsymbol{\lambda}) = (\frac{\lambda}{2})^M \prod_{i=1}^{M} \exp\{-\lambda|w_i|\}. \tag{4}$$

The regularization parameter $\hat{\lambda}$ in Eq. 1 will be a function of σ^2 and λ. Then, the optimal regularization parameters (σ^2 and λ) are computed by maximizing the posterior distribution $P(\sigma^2,\lambda|\mathbf{\Phi},\mathbf{y})$. According to the Bayes' rule, if the hyper-prior distributions on σ^2 and λ are flat, maximizing the posterior is equivalent to maximizing the marginal likelihood:

$$\begin{aligned} P(\mathbf{y}|\lambda,\sigma^2,\mathbf{\Phi}) &= \int_{-\infty}^{+\infty} d\mathbf{w}\, P(\mathbf{y}|\mathbf{\Phi},\mathbf{w},\sigma^2)P(\mathbf{w}|\lambda) \quad (5)\\ &= \frac{\lambda^M}{2^M(2\pi\sigma^2)^{N/2}}\int_{-\infty}^{+\infty} d\mathbf{w}\, \exp[-F(\mathbf{w})] \end{aligned}$$

where

$$F(\mathbf{w}) = \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\|^2 + \lambda\sum_{i=1}^{M}|w_i|. \tag{6}$$

Unfortunately, this marginal likelihood can not be evaluated analytically, and thus it can not be maximized directly. Our strategy is to treat $\mathbf{w}$ as hidden variables, σ^2 and λ as parameters, and to optimize the marginal likelihood via Expectation-Maximization (EM) update rules:

$$\frac{1}{\lambda} \longleftarrow \frac{1}{M}\int_{-\infty}^{+\infty} d\mathbf{w}\, \sum_i |w_i| Q(\mathbf{w}) \tag{7}$$

$$\sigma^2 \longleftarrow \frac{1}{N}\int_{-\infty}^{+\infty} d\mathbf{w}\, \|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\|^2 Q(\mathbf{w}) \tag{8}$$

where the expectations are taken over the distribution $Q(\mathbf{w}) = \frac{1}{\mathcal{Z}_w}\exp[-F(\mathbf{w})]$ with normalization constant $\mathcal{Z}_w = \int_{-\infty}^{+\infty} d\mathbf{w}\, \exp[-F(\mathbf{w})]$. The EM procedure can be thought as iteratively re-estimating the optimal parameters (σ^2 and λ) from the current estimate of the weight statistics $Q(\mathbf{w})$. Because it is difficult to analytically compute the integrals in Eqs. 7 and 8, we seek to approximate the distribution $Q(\mathbf{w})$ around its mode, $\mathbf{w}^{MP}$.

2.1. Computing $\mathbf{w}^{MP}$

The mode, $\mathbf{w}^{MP}$, is defined as the $\mathbf{w}$ that maximizes the $Q(\mathbf{w})$, namely

$$\mathbf{w}^{MP} = \arg\min_w \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} + \lambda\sum_{i=1}^{M}|w_i|, \tag{9}$$

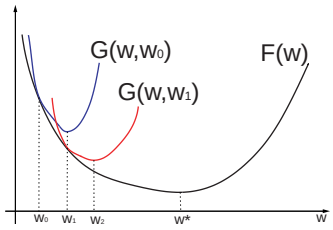

Fig. 1. The iterative procedure of minimizing $F(\mathbf{w})$ via auxiliary functions $G(\mathbf{w},\tilde{\mathbf{w}})$, with $\tilde{\mathbf{w}} = \mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, ...$

where $\mathbf{A} = \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}$, and $\mathbf{b} = \sigma^{-2}\mathbf{\Phi}^T\mathbf{y}$. Note that Eq. 9 is equivalent to Eq. 1 with $\hat{\lambda} = \sigma^2\lambda$. We will also represent the objective function in Eq. 9 with $F(\mathbf{w})$. This minimization problem can be solved with several different optimization techniques such as the simplex method and interior point methods. However, we introduce here a method that solves this optimization problem by constructing auxiliary functions. Because of the concavity of a square-root function, $|w_i| = (w_i^2)^{1/2}$ is upper bounded as $|w_i| \le |\tilde{w}_i| + \frac{1}{2|\tilde{w}_i|}(w_i^2 - \tilde{w}_i^2)$ for any $\tilde{w}_i$, and equality holds only when $w_i = \tilde{w}_i$. As a result, we construct the auxiliary function:

$$G(\mathbf{w},\tilde{\mathbf{w}}) = \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + \mathbf{b}^T\mathbf{w} + \sum_i \frac{\lambda_i}{2|\tilde{w}_i|}w_i^2 + \sum_i \frac{\lambda_i}{2}|\tilde{w}_i|, \tag{10}$$

which satisfies the two conditions: 1) $G(\tilde{\mathbf{w}},\tilde{\mathbf{w}}) = F(\tilde{\mathbf{w}})$, and 2) $G(\mathbf{w},\tilde{\mathbf{w}}) \ge F(\mathbf{w})$ where the equality holds only when $\mathbf{w} = \tilde{\mathbf{w}}$. Then, the iterative update rule, $\tilde{\mathbf{w}} \longleftarrow \arg\min_{\mathbf{w}} G(\mathbf{w},\tilde{\mathbf{w}})$, will converge to a local minimum of $F(\mathbf{w})$ [8], which is also the global minimum since $F(\mathbf{w})$ in Eq. 9 is convex. An example for illustrating the iterative scheme is shown in Fig. 1. At each iterative step, since the auxiliary function is a quadratic function, its optimal solution can be computed analytically:

$$\tilde{\mathbf{w}}^* = (\mathbf{A}+\mathbf{\Lambda})^{-1}\mathbf{b} \tag{11}$$

where $\mathbf{\Lambda} = \mathrm{diag}([\lambda_1/|\tilde{w}_1|, \lambda_2/|\tilde{w}_2|, ..., \lambda_M/|\tilde{w}_M|])$. Because the columns in $\mathbf{\Phi}$ associated with zero solutions during the iterations can be pruned, the matrix inversion in Eq. 11 is performed on a gradually reduced matrix. Generally, the resulting algorithm for solving the optimization problem in Eq. 11 is easy to implement, has excellent convergence property, and is computationally efficient when the optimal solution is sparse.

2.2. Approximating $Q(\mathbf{w})$

After $\mathbf{w}^{MP}$ is computed, one may approximate $Q(\mathbf{w})$ as a δ-function at $\mathbf{w}^{MP}$. Unfortunately, this simple treatment may cause divergence of the updates when σ^2 and λ are not initialized properly. Here we adopt a similar approximation scheme recently developed for nonnegative deconvolution [9].

The solution of $\mathbf{w}^{MP}$ naturally partitions itself into two distinct groups: non-zero elements (indexed by J) and zero elements (indexed by I). As a result, we choose to approximate the joint distribution $Q(\mathbf{w})$ as a factorized distribution, namely, $Q(\mathbf{w}) \approx Q_J(\mathbf{w}_J)Q_I(\mathbf{w}_I)$.

Since $\mathbf{w}_J \neq 0$, and the first order derivative $(\nabla F(\mathbf{w})|_{\mathbf{w}^{MP}})_J = 0$, $Q_J(\mathbf{w}_J)$ is approximated as a Gaussian distribution with mean $\mathbf{w}_J^{MP}$ and variance $\mathbf{A}_{JJ}^{-1}$ with $\mathbf{A}_{JJ}$ being the sub-matrix of $\mathbf{A}$.

For $Q_I(\mathbf{w}_I) = Q(\mathbf{w})|_{\mathbf{w}_J=\mathbf{w}_J^{MP}}$, because $\mathbf{w}_I = 0$ and the first order derivative $(\nabla F(\mathbf{w})|_{\mathbf{w}^{MP}})_I \neq 0$, we approximate it with a factorized asymmetric Laplacian distribution, namely

$$\hat{Q}_I(\mathbf{w}_I) = \prod_{i\in I} \hat{Q}_i(w_i), \tag{12}$$

with

$$\hat{Q}_i(w_i) = \begin{cases} \frac{\mu_i^-}{2} e^{\mu_i^- w_i} & \text{when} \quad w_i < 0 \\ \frac{\mu_i^+}{2} e^{-\mu_i^+ w_i} & \text{when} \quad w_i \geq 0, \end{cases} \tag{13}$$

where the variational parameters $\boldsymbol{\mu}^+ \geq 0$ and $\boldsymbol{\mu}^- \geq 0$ is defined by minimizing the KL-divergence between Q_I and $\hat{Q}_I$, yielding the optimization problem:

$$\min_{\boldsymbol{\mu}\geq 0} \hat{\mathbf{b}}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^T\hat{\mathbf{A}}\boldsymbol{\mu} - \sum_i \ln \mu_i, \tag{14}$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}^+; \boldsymbol{\mu}^-]$, $\hat{\mathbf{b}} = [(\mathbf{A}\mathbf{w}^{MP} + \mathbf{b} + \lambda\mathbf{e})_I; (-\mathbf{A}\mathbf{w}^{MP} - \mathbf{b} + \lambda\mathbf{e})_I]$ with $\mathbf{e} = [1, 1, ..., 1]^T$, and $\hat{\mathbf{A}} = \begin{bmatrix} \hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12} \\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{22} \end{bmatrix}$, where $\hat{\mathbf{A}}_{11} = \hat{\mathbf{A}}_{22} = \frac{1}{2}\mathbf{A}_{II} + \frac{3}{2}\text{diag}(\mathbf{A}_{II})$, and $\hat{\mathbf{A}}_{21} = \hat{\mathbf{A}}_{12} = \frac{1}{2}\mathbf{A}_{II} - \frac{1}{2}\text{diag}(\mathbf{A}_{II})$ with $\mathbf{A}_{II}$ being the sub-matrix of $\mathbf{A}$. Since this minimization problem can not be solved analytically, we employ the same auxiliary function developed in [9], resulting multiplicative update rules for iteratively estimating $\boldsymbol{\mu}$ with guaranteed convergence:

$$\mu_i \longleftarrow \mu_i \frac{-\hat{b}_i + \sqrt{\hat{b}_i^2 + 4(\hat{\mathbf{A}}^+\boldsymbol{\mu})_i[(\hat{\mathbf{A}}^-\boldsymbol{\mu})_i + \frac{1}{\mu_i}]}}{2(\hat{\mathbf{A}}^+\boldsymbol{\mu})_i}. \tag{15}$$

After the variational parameters $\boldsymbol{\mu}$ are derived, the mean $\bar{\mathbf{w}}$, the absolute mean $\overline{|w_i|}, i = 1, 2, ..., M$, and the covariance $\mathbf{C}$ of $\mathbf{w}$ under the approximated distribution can be computed:

$$\bar{w}_i = \begin{cases} w_i^{ML} & \text{if} \quad i \in J \\ (\mu_i^+ - \mu_i^-)/2 & \text{if} \quad i \in I \end{cases}, \tag{16}$$

$$\overline{|w_i|} = \begin{cases} |w_i^{ML}| & \text{if} \quad i \in J \\ (\mu_i^+ + \mu_i^-)/2 & \text{if} \quad i \in I \end{cases}, \tag{17}$$

$$C_{ij} = \begin{cases} (\mathbf{A_{JJ}}^{-1})_{ij} & \text{if } i, j \in J \\ \delta_{ij}[\frac{(\mu_i^+ + \mu_i^-)^2}{4} + \frac{(\mu_i^+)^2 + (\mu_i^-)^2}{2}] & \text{otherwise.} \end{cases}$$

From these statistics, the integrals in Eqs. 7 and 8 can be evaluated analytically, and the update rules for estimating λ and σ^2 becomes:

$$\lambda \longleftarrow \frac{M}{\sum_{i=1}^M \overline{|w_i|}} \tag{18}$$

$$\sigma^2 \longleftarrow \frac{1}{N}[(\mathbf{y} - \boldsymbol{\Phi}\bar{\mathbf{w}})^T(\mathbf{y} - \boldsymbol{\Phi}\bar{\mathbf{w}}) + \text{Tr}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{C})] \tag{19}$$

2.3. Extension to independent L_1-norm regularization

We can adapt the Bayesian framework to infer the optimal regularization parameters in Eq. 2 by assuming a Laplacian distribution with independent decay parameters, namely,

$$P(\mathbf{w}|\boldsymbol{\lambda}) = \prod_{i=1}^M \frac{\lambda_i}{2} \exp\{-\lambda_i|w_i|\}. \tag{20}$$

Then, the Bayesian formulation would be similar as the uniform regularization case except that Eqs. 6, 7 and 18 respectively become

$$F(\mathbf{w}) = \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w})^T(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}) + \sum_{i=1}^M \lambda_i|w_i|, \tag{21}$$

$$\frac{1}{\lambda_i} \longleftarrow \int_{-\infty}^{+\infty} d\mathbf{w}\, |w_i| Q(\mathbf{w}), \tag{22}$$

$$\lambda_i \longleftarrow \frac{1}{\overline{|w_i|}}. \tag{23}$$

3. SIMULATION

In this section, we employ simulation to demonstrate that the update rules derived in Section 2 converge to the correct noise level (σ^2) and the Bayesian L_1-norm sparsity learning is better at discovering sparse solutions. In particular, with inferred optimal independent regularization parameters, the optimization problem in Eq. 2 is able to accurately resolve the correct sparseness of the solution even in very noisy data.

We use deconvolution as an example for simulation, and demonstrate that sparsity regularization can be utilized to achieve high temporal resolution (or sub-sample resolution) in FIR filter identification. A speech segment (1024 samples, sampling frequency was 16,000Hz) was employed as the source signal $\mathbf{s}$. The simulated sparse FIR filter $\mathbf{w}$ has nonzero amplitudes of -0.5, 0.35, 1, 0.6, and -0.4 at -9.75T_s, -6T_s, 1T_s, 2.5T_s, and 7.75T_s (T_s denotes the sample interval), respectively, and has zero amplitude elsewhere. Then the observation $\mathbf{y}$ is the convolution of the source and the simulated filter corrupted by I.I.D. sampled zero-mean Gaussian noise. The task of the deconvolution is to discover the filter $\mathbf{w}$ given the source $\mathbf{s}$ and the observation $\mathbf{y}$.

In deconvolution, the columns of the designed matrix $\boldsymbol{\Phi}$ are the delayed patterns of the source with delays from $-10T_s$

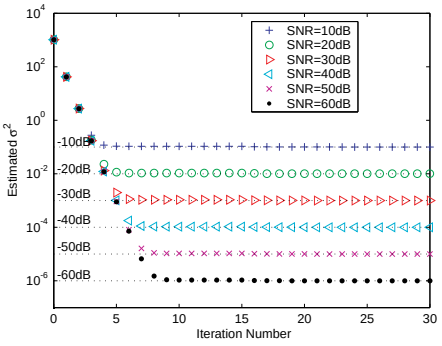

Fig. 2. Convergence of σ^2 estimation in Bayesian L_1-norm sparse learning. The signal was normalized so that it had unit power.

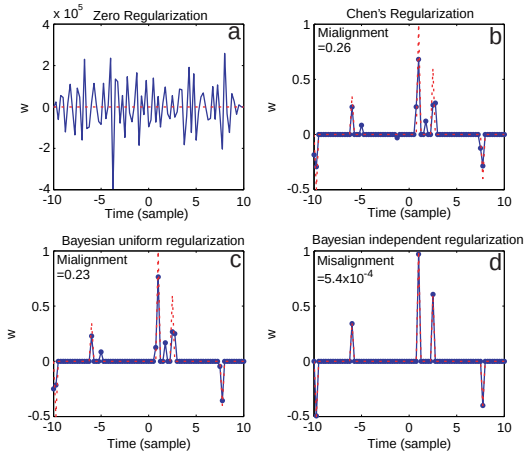

Fig. 3. Deconvolution result by different L_1-norm regularization schemes. a) no regularization; b) the regularization proposed by S. S. Chen *et. al* [2]($\hat{\lambda} = 0.94$); c)Bayesian uniform regularization, ($\lambda = 28$ and $\sigma^2 = 0.1$, thus $\hat{\lambda} = 2.8$); d) Bayesian independent regularization. The dot lines in the figures indicate the ground truth of the filter, while the solid lines with dots are the estimates.

to $+10T_s$ incremented by $0.25T_s$. Due to the fact that the adjacent columns in $\mathbf{\Phi}$ are very similar to each other and the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ is ill-conditioned, sparsity regularization is crucial for deriving a correct solution.

Figure 2 illustrates the convergence of the σ^2 estimation under different noise levels (from -60dB to -10dB) using the update rules of Bayesian L_1-norm sparse learning derived in Section 2. In the simulation, uniform regularization was employed in the first 15 iterations, and then independent regularization was utilized in the next 15 iterations to further refine the solution. From Figure 2, we observe that the σ^2 estimate often converges to the true value even with bad initialization.

The resulting filter estimate when SNR=10dB is shown in Figure 3 (d). Compared to the estimate of the first 15 iterations with uniform regularization (shown in Figure 3 (c)), the result of an additional 15 iterations with independent regularization exhibits the same sparseness as the true solution and has very small misalignment (defined as $\|\hat{\mathbf{w}} - \mathbf{w}_0\|^2/\|\mathbf{w}_0\|^2$ with $\hat{\mathbf{w}}$ being the estimate and $\mathbf{w}_0$ being the ground truth). By contrast, other approaches that empirically determine the regularization parameter often yield sub-optimal solutions, as shown in Figure 3 (b). Because the simulated deconvolution is ill-conditioned without sparsity regularization, the estimate in Figure 3 (a) with no regularization fluctuates widely, containing little information about the true filter.

4. CONCLUSION

We have developed a Bayesian framework for inferring the optimal regularization parameters for L_1-norm regularized LMS problem. We have demonstrated that, by extending a uniform regularization to independent regularization, our Bayesian L_1-norm sparse learning algorithm is able to precisely resolve the sparse solution even in very noisy conditions.

Our work provides an unified probabilistic framework for L_1-norm sparse learning. It can be easily adapted to other variants of L_1-norm regularized problems(such as nonnegative LMS [9]), and can be used to elucidate the role of sparsity in these algorithms. Future work will concentrate on extending our Bayesian L_1-norm sparse learning algorithm to other applications in signal processing.

5. REFERENCES

[1] Robert Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Socity, Series B, (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[2] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] Donald L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 508–518, 2000.

[4] J. J. Fuchs, "Multipath time-delay detection and estimation," *IEEE Transactions on Signal Processing*, vol. 47, pp. 237–243, 1999.

[5] B.A. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for nature images," *Nature*, vol. 381, pp. 607–609, 1996.

[6] Dmitry M. Malioutov, Mujdat Cetin, and Alan S. Willsky, "Homotopy continuation for sparse signal representation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[7] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[8] W. J. Zangwill, *Nonlinear Programming: a unified approach*, Prentice-Hall, 1969.

[9] Yuanqing Lin and Daniel D. Lee, "Bayesian Regularization And Nonnegative Deconvolution (BRAND) for room impulse response estimation," *IEEE Trans. Signal Processing*, accepted for publication.