

PARSIMONIOUS DICTIONARY LEARNING

Mehrdad Yaghoobi, Thomas Blumensath, Michael E. Davies

Institute for Digital Communications,
 Joint Research Institute for Signal and Image Processing,
 The University of Edinburgh, EH9 3JL, UK
 { m.yaghoobi-vaighan, thomas.blumensath, mike.davies }@ed.ac.uk

ABSTRACT

Sparse modeling of signals has recently received a lot of attention. Often, a linear under-determined generative model for the signals of interest is proposed and a sparsity constraint imposed on the representation. When the generative model is not given, choosing an appropriate generative model is important, so that the given class of signals has approximate sparse representations. In this paper we introduce a new scheme for dictionary learning and impose an additional constraint to reduce the dictionary size. Small dictionaries are desired for coding applications and more likely to “work” with sub-optimal algorithms such as Basis Pursuit. Another benefit of small dictionaries is their faster implementation, e.g. a reduced number of multiplication/addition in each matrix vector multiplication, which is the bottleneck in sparse approximation algorithms.

Index Terms— Sparse Approximation, Dictionary Learning, Majorization Method, Sparse Coding

1. INTRODUCTION

Let $\mathcal{Y} = \{y^{(i)} : 1 \leq i \leq L\}$ be a given set of training samples and $\mathcal{X} = \{x^{(i)} : 1 \leq i \leq L\}$ be the corresponding coefficient vectors. $\mathbf{Y}_{d \times L}$ and $\mathbf{X}_{N \times L}$ are the matrices generated by using the elements of \mathcal{Y} and \mathcal{X} as the column vectors, respectively. The dictionary learning problem can be formulated as follows. Given \mathbf{Y} , find a “dictionary” matrix \mathbf{D} and a coefficient matrix \mathbf{X} , such that the error $\epsilon = \mathbf{Y} - \mathbf{D}\mathbf{X}$ is small and \mathbf{X} is sparse. This is a challenging problem and researchers from different fields have introduced algorithms to solve it approximately [1–4]. Regardless of the sparsity measure, dictionary learning is a non-convex optimization problem and a locally optimum dictionary is often found [5]. Various additional constraints have been recently imposed on the dictionaries to constrain the dictionary search space. These constraints may come from *a priori* information about the dictionary [6, 7] or help to attain a fast implementation [8, 9].

One application of sparse approximation is sparse coding. In conventional sparse coding, indices of the selected columns of \mathbf{D} , called “atom”, and the associated coefficients are coded separately [10–12]. The coding cost of specifying the selected atoms is reduced by reducing the size of the dictionary. Therefore minimum size dictionaries are more desirable for a coding purpose. Also, when the size of the learnt dictionary reduces, matrix-vector multiplication can be done faster.

The application of parsimonious dictionary learning is not limited to coding. Dictionary size selection is also a challenging problem

in the sparse approximation of real signals. When the size of the dictionary is unknown, one can start with a oversized dictionary and find the minimum size learnt dictionary.

We here introduce a framework for parsimonious dictionary learning. The problem formulation is followed by a practical algorithm to find an approximate solution. We show that the proposed framework gives promising results in dictionary recovery. We then show that the learnt dictionary has advantages over the currently used dictionaries for sparse coding.

2. PARSIMONIOUS DICTIONARY LEARNING FORMULATION

Dictionary learning can be formulated as the minimization of a joint objective function based on \mathbf{D} and \mathbf{X} .

$$\min_{\mathbf{D}, \mathbf{X}} \phi(\mathbf{D}, \mathbf{X}) \text{ s.t. } \mathbf{D} \in \mathcal{D}; \quad (1)$$

$$\phi(\mathbf{D}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda J_{p,p}(\mathbf{X}),$$

where $\|\cdot\|_F$ is the Frobenius-norm, \mathbf{D} is a dictionary in an admissible set \mathcal{D} and $J_{p,p}$ is the penalty term over the diversity of the coefficients,

$$J_{p,q}(\mathbf{X}) = \sum_{i \in I} \left[\sum_{j \in J} |x_{ij}|^q \right]^{p/q}, \quad (2)$$

where $p \leq 1$. λ is a Lagrangian multiplier. In this paper we use $p = 1$ which makes the minimization over \mathbf{X} convex, if \mathbf{D} is fixed. Various admissible sets have been used for dictionary learning (e.g. see [5]). We use bounded column-norm and bounded Frobenius-norm sets as the admissible sets to make the dictionary update a convex problem for a fixed \mathbf{X} . The bounded column-norm admissible set is defined as follows,

$$\mathcal{D}_F = \{\mathbf{D}_{d \times N} : \|\mathbf{D}\|_F \leq c_F^{1/2}\}, \quad (3)$$

where c_F is a constant. The bounded Frobenius-norm admissible set is defined by,

$$\mathcal{D}_C = \{\mathbf{D}_{d \times N} : \|\mathbf{d}_i\|_2 \leq c_C^{1/2}\}, \quad (4)$$

where \mathbf{d}_i is the i^{th} column of the dictionary \mathbf{D} and c_C is a constant. To get a dictionary of minimum size, we now include an additional penalty on the dictionary size. The new joint optimization problem is as follows,

$$\min_{\mathbf{X}, \mathbf{D}} \phi_{\theta,0,\infty}(\mathbf{D}, \mathbf{X}) \text{ s.t. } \mathbf{D} \in \mathcal{D};$$

$$\phi_{\theta,0,\infty}(\mathbf{D}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda J_{1,1}(\mathbf{X}) + \theta \|\max_i \{\|\mathbf{D}\}_{i,j}\|_0.$$

where $\|\cdot\|_0$ is an operator that counts the number of non-zero elements, and is therefore related to the size of the dictionary, and $\{\mathbf{D}\}_{i,j}$ is the element (i,j) of \mathbf{D} . Because $\phi_{\theta,0,\infty}$ is non-convex and non-continuous, we replace the objective function with a relaxed version as follows,

$$\min_{\mathbf{X}, \mathbf{D}} \phi_{\theta,1,q}(\mathbf{D}, \mathbf{X}) \text{ s.t. } \mathbf{D} \in \mathcal{D};$$

$$\phi_{\theta,1,q}(\mathbf{D}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda J_{1,1}(\mathbf{X}) + \theta J_{1,q}(\mathbf{D}^T) \quad (5)$$

where $q \geq 1$. By selecting $q = 1$, the objective function penalizes any non-zero element of the dictionary. With some changes, this would be useful for sparse dictionary learning as introduced in [13]. When $q > 1$, the objective function penalizes the number of atoms more than the sparsity of the atoms which is our aim in this paper. The parameter θ is then the regularization parameter which controls the sparsity of the dictionary. By increasing θ , one can get a smaller dictionary.

This objective function can be minimized using alternating minimization. Although this method is guaranteed to reduce the objective in each step, the objective function is not convex and has various local minima. The proposed method optimizes \mathbf{X} and \mathbf{D} alternately keeping the other parameter is fixed. In this framework, the non-convex optimization problem is broken into two convex optimization problems, which can be solved using any convex optimization method. Here we use a majorization minimization method.

3. MAJORIZATION METHOD FOR SPARSE APPROXIMATION AND DICTIONARY UPDATE

We use the majorization minimization method [14] to minimize (5). In the majorization method, the objective function is replaced by a surrogate objective function which majorizes it and can be minimized easier. Here we are interested in the surrogate functions in which the parameters are decoupled, so that the surrogate function can be minimized element-wise.

A function ψ majorizes ϕ when it satisfies the following conditions,

$$\begin{aligned} \phi(\omega) &\leq \psi(\omega, \xi), \quad \forall \omega, \xi \in \Upsilon \\ \phi(\omega) &= \psi(\omega, \omega), \quad \forall \omega \in \Upsilon, \end{aligned} \quad (6)$$

where Υ is the parameter space. The surrogate function has an additional parameter ξ . We choose this parameter as the current value of ω and find the optimal update for ω .

$$\omega_{new} = \arg \min_{\omega \in \Upsilon} \psi(\omega, \xi). \quad (7)$$

We then update ξ with ω_{new} . The algorithm continues until we find an accumulation point. In practice the algorithm could be terminated when the distance between ω and ω_{new} is less than a threshold.

There are different ways to derive a surrogate function. Jensen's inequality and Taylor series have often been used for this purpose [14]. When \mathbf{D} or \mathbf{X} are fixed, the surrogate function for the quadratic part of (5) can be found [15] by adding $\pi_{\mathbf{X}}(\mathbf{X}, \mathbf{X}^{[n-1]}) := c_X \|\mathbf{X} - \mathbf{X}^{[n-1]}\|_F^2 - \|\mathbf{D}\mathbf{X} - \mathbf{D}\mathbf{X}^{[n-1]}\|_F^2$ or $\pi_{\mathbf{D}}(\mathbf{D}, \mathbf{D}^{[n-1]}) := c_D \|\mathbf{D} - \mathbf{D}^{[n-1]}\|_F^2 - \|\mathbf{D}\mathbf{X} - \mathbf{D}^{[n-1]}\mathbf{X}\|_F^2$ respectively, where $c_X > \|\mathbf{D}^T \mathbf{D}\|$ and $c_D > \|\mathbf{X}^T \mathbf{X}\|$ are two constants and $\|\cdot\|$ is defined as the spectral norm. $\mathbf{X}^{[n-1]}$ and $\mathbf{D}^{[n-1]}$ are the old values of \mathbf{X} and \mathbf{D} respectively which are the auxiliary parameter ξ in the surrogate objective. In the next two subsections, we show how this method can be used for optimizing (5) in an alternating minimization scheme.

3.1. Matrix Valued Sparse Approximation

In this subsection we briefly show how the majorization method is used for matrix valued sparse approximation. We add $\pi_{\mathbf{X}}$ to (5) and minimize the surrogate objective based on \mathbf{X} , followed by updating $\mathbf{X}^{[n-1]}$ with the new value of \mathbf{X} . Let $\mathbf{A} := \frac{1}{c_X}(\mathbf{D}^T \mathbf{Y} + (c_X \mathbf{I} - \mathbf{D}^T \mathbf{D})\mathbf{X}^{[n-1]})$. It can be shown that (7) can be solved, for the proposed surrogate objective, by shrinking elements in \mathbf{A} , as follows:

$$\{\mathbf{X}^{[n]}\}_{i,j} = \begin{cases} a_{i,j} - \lambda/2 \text{ sign}(a_{i,j}) & \lambda/2 < |a_{i,j}| \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

The convergence of this algorithm is studied in [16] for vector valued coefficients. This proof can also be extended to matrix valued problems.

3.2. Dictionary Update

The objective function is convex when \mathbf{X} is fixed. For fixed \mathbf{X} , to minimize over \mathbf{D} , the joint sparsity penalty is decoupled by adding $\pi_{\mathbf{D}}$ to the objective function,

$$\psi_{\theta,1,q}(\mathbf{D}, \mathbf{D}^{[n-1]}) = \phi_{\theta,1,q}(\mathbf{D}, \mathbf{X}) + \pi_{\mathbf{D}}(\mathbf{D}, \mathbf{D}^{[n-1]}). \quad (9)$$

By separating the terms depending on \mathbf{D} , the surrogate cost can be written as,

$$\psi_{\theta,1,q}(\mathbf{D}, \mathbf{D}^{[n-1]}) \propto c_{str}\{\mathbf{D}\mathbf{D}^T - 2\mathbf{B}\mathbf{D}^T\} + J_{1,q}(\mathbf{D}^T) \quad (10)$$

where $\mathbf{B} := \frac{1}{c_D}(\mathbf{Y}\mathbf{X}^T + \mathbf{D}^{[n-1]}(c_D \mathbf{I} - \mathbf{X}\mathbf{X}^T))$. The dictionary constraint is introduced into the objective function using Lagrangian multipliers. Let \mathbf{d}_j and \mathbf{b}_j be the j^{th} columns of \mathbf{D} and \mathbf{B} respectively. The objective function, using the bounded column-norm (4), can be written as,

$$\begin{aligned} \psi_{\theta,1,q}(\mathbf{D}, \mathbf{D}^{[n-1]}) &\propto \sum_j (\text{tr}\{\tau_j^2 \mathbf{d}_j \mathbf{d}_j^T - 2\mathbf{b}_j \mathbf{d}_j^T\} + \frac{\theta}{c_D} \|\mathbf{d}_j\|_q) \\ &= \sum_j (\tau_j^2 \mathbf{d}_j^T \mathbf{d}_j - 2\mathbf{d}_j^T \mathbf{b}_j + \frac{\theta}{c_D} \|\mathbf{d}_j\|_q) \\ &\propto \sum_j ((\tau_j \mathbf{d}_j - \mathbf{b}_j / \tau_j)^2 + \frac{\theta}{c_D \tau_j} \|\tau_j \mathbf{d}_j\|_q) \\ &= \sum_j \psi_q^{\frac{\theta}{c_D \tau_j}}(\tau_j \mathbf{d}_j, \mathbf{b}_j / \tau_j) \end{aligned} \quad (11)$$

where $\psi_q^\alpha(\mathbf{v}, \mathbf{w}) = (\mathbf{w} - \mathbf{v})^2 + \alpha \|\mathbf{v}\|_q$, $\tau_j = (1 + \gamma_j / c_D)^{1/2}$ and γ_j are the Lagrangian multipliers. To minimize (11), we can minimize the first term by minimizing ψ_q^α for each \mathbf{d}_j independently. With the help of two lemmas presented in [17], we can find the optimum of ψ_q^α based on \mathbf{d}_j for $q = 1, 2$ and ∞ . The minimum of $\psi_q^\alpha(\mathbf{v}, \mathbf{w})$ based on \mathbf{v} [17, Lemma 4.1] is,

$$\min_{\mathbf{v}} \psi_q^\alpha(\mathbf{v}, \mathbf{w}) = \mathbf{w} - \mathcal{P}_\alpha^{q'}(\mathbf{w}) \quad (12)$$

where $\mathcal{P}_\alpha^{q'}$ is the orthogonal projection onto the dual norm ball with radius \mathbf{w} and the dual norm is defined as $\|\cdot\|_{q'}$ with $1/q' + 1/q = 1$. This minimization problem can be solved analytically for some

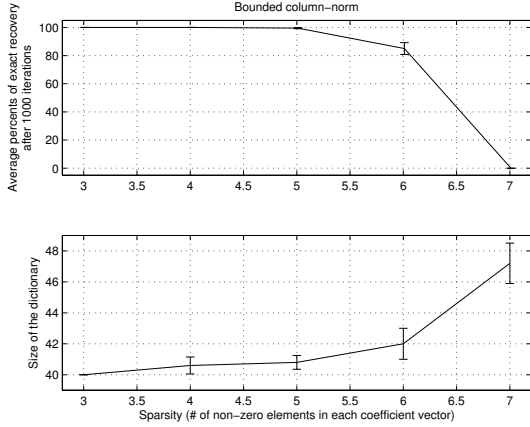


Fig. 1. Exact recovery with the constrained column-norm.

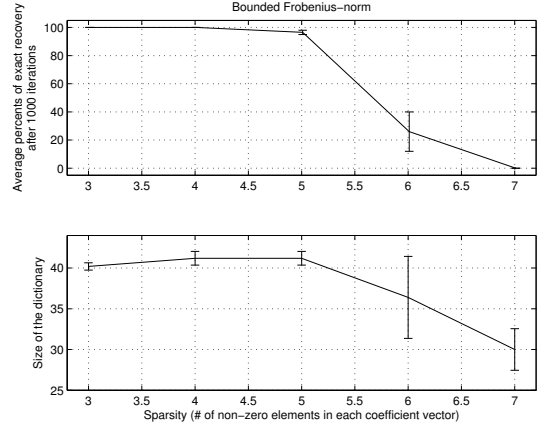


Fig. 2. Exact recovery with the bounded Frobenius column-norm.

q [17, Lemma 4.2]. In this paper we derive the dictionary update formula for $q = 2$.

$$\mathbf{b}_j^* = \arg \min_{\mathbf{d}_j} \psi_{2, \tau_j}(\tau_j \mathbf{d}_j, \mathbf{b}_j / \tau_j) = \begin{cases} \frac{1}{\tau_j^2} \left(1 - \frac{\theta}{2c_D \|\mathbf{b}_j\|_2}\right) \mathbf{b}_j & \frac{\theta}{2c_D} < \|\mathbf{b}_j\|_2 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

When all γ_j are non-negative, for any inadmissible \mathbf{b}_j^* with $\tau_j = 1$ ($\gamma_j = 0$), one can decrease $\|\mathbf{d}_j^*\|_2$ to $c_c^{1/2}$ by increasing τ_j to satisfy the K.K.T conditions. The dictionary update is therefore done by calculating \mathbf{B} followed first by (13) ($\tau_j = 1$) and secondly by orthogonal projection onto the convex set (4).

When we are looking for a bounded Frobenius-norm dictionary, the dictionary update could be derived using a similar approach, using orthogonal projection onto (3) instead of (4).

4. SIMULATION

We evaluate the proposed method with synthetic and real data. Using synthetic data with random dictionaries helps us to examine the ability of the proposed methods to recover dictionaries exactly (to within an acceptable squared error). To evaluate the performance on real data, we chose audio signals. We then used the learnt dictionary for audio coding and show improvements in Rate-Distortion performance compared to coding with classical dictionaries.

4.1. Synthetic Data

A 20×40 matrix \mathbf{D} was generated by normalizing a matrix with i.i.d. uniform random entries. The number of non-zero elements in each of the coefficient vectors was selected between 3 and 7. The locations of the non-zero coefficients were selected uniformly at random. We generated 1280 training samples where the absolute values of the non-zero coefficients were selected uniformly between 0.2 and 1. We debiased all the sparse approximations by orthogonally projecting onto the space spanned by atoms with non-zero coefficients.

We assume that the desired dictionary size is unknown but bounded. The simulations were started with four times overcomplete dictionaries (two times larger than the desired dictionary size). The

dictionary updates were based on the joint sparsity objective function (5) (with $\theta = 0.05$, $p = 1$ and $q = 2$). The average percentage of exact atom recovery, i.e. absolute inner product of the learnt atom with one of the atoms in the original dictionary is more than 0.99, for 5 trials are shown in Fig. 1 and 2. We plotted the percentage of the exact recovery of the original atoms, regardless of the learnt dictionary size. In the lower plot, we show the size of dictionary after 1000 iterations. With this θ we identified the size correctly but for less sparse signals (higher k) we got less accurate results.

4.2. Parsimonious Dictionary Learning for Sparse Audio Coding

In this section we demonstrate the performance of the proposed dictionary learning method on audio signals. An audio sample of more than 8 hours was recorded from BBC radio 3, which plays mostly classical music. We used the proposed method with the bounded Frobenius-norm constraint to learn a dictionary based on a training set of 8192 blocks, each 1024 samples long.

In this experiment, instead of fully optimizing over one parameter (\mathbf{X} or \mathbf{D}) before switching to the other one, we update each parameter for a small number of iterations and then switch to the other one. This type of alternate optimization was found to be faster in practice.

We chose a 2 times overcomplete sinusoid dictionary (frequency oversampled DCT) as the initialization point and ran the simulations with different lambda values for 5000 iterations of alternative optimization of (11). The number of appearances of each atom, which are sorted based on their ℓ_2 norms, are shown in Fig. 3. To design an efficient encoder we only used atoms that were used frequently in the representations. Therefore we were able to further shrink the dictionary size. In this test we chose a threshold of 40 appearances (out of 8192) as the selection criteria. This dictionary was used to find the sparse approximations of 4096 different random blocks, each of 1024 samples, from the same data set. We then encoded the location (significant bit map) and magnitude of the non-zero coefficients separately. In this paper we used a uniform scalar quantizer with a double zero bin size to code the magnitude. We estimated the entropy of the coefficients to approximate the required coding cost. To encode the significant bit map, we assumed an i.i.d. distribution for the location of the non-zero atoms. The same coding strategy was used to code sparse approximations with a two times frequency over-

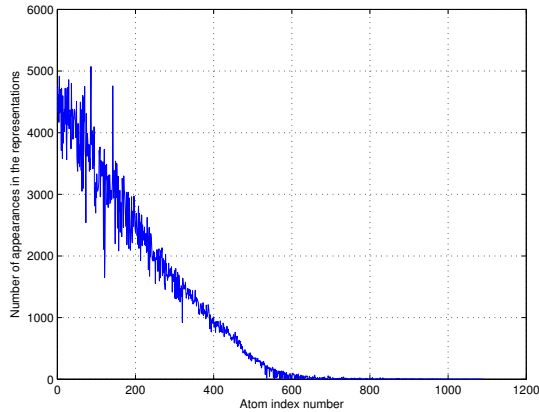


Fig. 3. Number of appearances in the representations of the training blocks (of size 8192).

complete DCT (the initial dictionary used for learning) followed by shrinking based on the number of appearances. For reference we calculated the rate-distortion of the DCT coefficient encoding of the same data, using the same method of significant bitmap and non-zero coefficients coding. The performance is compared in Fig. 4. In the sparse coding methods, the convex hulls of the rate-distortion performances calculated with different dictionaries, each optimized and shrunk for different bit-rates, are shown in this figure. Using the learnt dictionaries for sparse approximation is superior to using the DCT or overcomplete DCT for the range of bit-rates shown.

5. CONCLUSIONS

We introduced a formulation for parsimonious dictionary learning. We have shown how we can solve the dictionary learning problem approximately, by imposing a penalty on the size of the dictionary, using a majorization method. A small set of simulations showed that the algorithm often recovers a dictionary with the correct size. We then used the learnt dictionary for sparse coding. We showed the advantages over standard overcomplete and orthogonal dictionaries, specially at low bit-rate. Although the results are promising, more investigations are needed to find a method to determine the parameter θ .

6. REFERENCES

- [1] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.
- [2] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations," *Neural Comp*, vol. 12, no. 2, pp. 337–365, 2000.
- [3] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comp.*, vol. 15, pp. 349–396, 2003.
- [4] M. Aharon, E. Elad, and A.M. Bruckstein, "K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

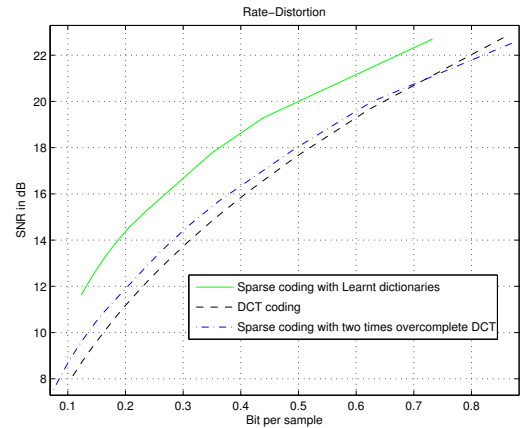


Fig. 4. Estimated Rate-Distortion for the audio coding.

- [5] M. Yaghoobi, T. Blumensath, and M. Davies, "Regularized dictionary learning for sparse approximation," in *EUSIPCO*, 2008.
- [6] T. Blumensath and M.E. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 50–57, 2006.
- [7] P. Jost, S. Lesage, P. Vandergheynst, and R. Gribonval, "MOTIF: An efficient algorithm for learning translation invariant dictionaries," in *ICASSP*, 2006.
- [8] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," in *ICASSP*, 2005.
- [9] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Modeling and Simulation*, vol. 7, no. 1, pp. 214–241, 2008.
- [10] P. Frossard, P. Vandergheynst, R. Figueras i Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *IEEE Trans. on Signal Processing*, vol. 52, no. 2, pp. 525–535, 2004.
- [11] M.E. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, vol. 86, no. 3, pp. 457–470, 2006.
- [12] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1999.
- [13] R. Rubinstein, M. Zibulevsky, and M. Elad, "Sparsity, take 2: Accelerating sparse-coding techniques using sparse dictionaries," submitted.
- [14] K. Lange, *Optimization*, Springer-Verlag, 2004.
- [15] M. Yaghoobi, T. Blumensath, and M. Davies, "Dictionary learning for sparse approximations with the majorization method," submitted.
- [16] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, pp. 1413–1541, 2004.
- [17] M. Fornasier and H. Rauhut, "Recovery algorithms for vector valued data with joint sparsity constraints," to appear in *SIAM Journal of Numerical Analysis*, 2007.