JOINT LOW MUTUAL AND AVERAGE COHERENCE DICTIONARY LEARNING

Javad Parsa[1], Mostafa Sadeghi[1], Massoud Babaie-Zadeh[1], and Christian Jutten[2]

[1]Electrical Engineering Department, Sharif University of Technology, Tehran, IRAN
[2]GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP and Inst. Univ. de France.

ABSTRACT

Dictionary learning (DL) has found many applications in sparse approximation problems. Two important properties of a dictionary are maximum and average coherence (cross-correlation) between its atoms. Many algorithms have been presented to take into account the coherence between atoms during dictionary learning. Some of them mainly reduce the maximum (mutual) coherence whereas some other algorithms decrease the average coherence. In this paper, we propose a method to jointly reduce the maximum and average correlations between different atoms. This is done by making a balance between reducing the maximum and average coherences. Experimental results demonstrate that the proposed approach reduce the mutual and average coherence of the dictionary better than existing algorithms.

Index Terms— Compressed sensing, sparse coding, mutual coherence, average coherence, dictionary learning

1. INTRODUCTION

1.1. Dictionary learning

During the last decade, dictionary learning for sparse approximation has attracted a lot of attention due to its application in many different areas such as compressed sensing and image reconstruction [1, 2]. In the problem of dictionary learning, given a training dataset $\mathbf{Y} \triangleq [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_l]$, $\mathbf{y}_i \in \mathbb{R}^m$, a dictionary $\mathbf{D} \triangleq [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_n]$, $\mathbf{d}_i \in \mathbb{R}^m$, is learned in such a way that it provides sparse coefficients for $\mathbf{y}_i$'s. Each $\mathbf{d}_i$ is called an atom [1]. That is, the representations $\mathbf{X} \triangleq [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l]$, $\mathbf{x}_i \in \mathbb{R}^n$ are sufficiently sparse. This problem is usually formulated as follows [3]:

$$(\mathbf{D}^*, \mathbf{X}^*) = \underset{\mathbf{D}\in\mathcal{D}, \mathbf{X}\in\mathcal{X}}{\operatorname{argmin}} \ \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \qquad (1)$$

in which, $\mathcal{D}$ and $\mathcal{X}$ are defined as $\mathcal{D} = \{\mathbf{D} : \forall i, \|\mathbf{d}_i\|_2^2 \leq 1\}$ and $\mathcal{X} = \{\mathbf{X} : \forall i, \|\mathbf{x}_i\|_1 \leq \tau\}$, where, $\|.\|_1$ denotes ℓ_1 norm. To solve (1) , many dictionary learning algorithms have been introduced [2–6], which are mainly based on alternating minimization.

This work was supported in part by European project ERC-2012AdG-320684-CHESS, and in part by Iran National Science Foundation (contract No. 96000780).

1.2. Mutual and average coherence

Mutual coherence of a dictionary, denoted by $\mu(\mathbf{D})$, is defined as the maximum absolute value of cross-correlations between its atoms [7]. The average coherence of a dictionary, denoted by $\mu_{avg}(\mathbf{D})$, is defined as the average absolute value of the cross-correlations between atoms. For a dictionary with normalized columns, these two parameters are defined as:

$$\begin{aligned} \mu(\mathbf{D}) &= \max_{i\neq j} |\mathbf{D}^T\mathbf{D}|_{ij}, \\ \mu_{avg}(\mathbf{D}) &= \sqrt{\frac{\|\mathbf{D}^T\mathbf{D} - \mathbf{I}\|_F^2}{n(n-1)}}. \end{aligned} \qquad (2)$$

Mutual coherence plays an important role in sparse approximation. A signal with a sparse representation $\mathbf{x}$ with sparsity level s, i.e. with s nonzero coefficient s, can be recovered from $\mathbf{y} = \mathbf{D}\mathbf{x}$ through ℓ_1 minimization when [8]:

$$s \leq \frac{1}{2}(1 + \frac{1}{\mu}). \qquad (3)$$

According to (3), dictionaries with low mutual coherences are better for high sparsity levels in comparison to dictionaries with large mutual coherence. However, the mutual coherence is lower bounded [9], and it can be shown that:

$$\mathbf{D} \in \mathbb{R}^{m\times n} \rightarrow \sqrt{\frac{n-m}{m(n-1)}} \leq \mu(\mathbf{D}) \leq 1, \qquad (4)$$

where

$$\mu_{welch} = \sqrt{\frac{n-m}{m(n-1)}}.$$

Furthermore, dictionaries with low average coherences are favorable in compressed sensing applications [10]. During the recent years, many dictionary learning algorithms use the following cost function to learn low-coherence dictionaries [11–13]:

$$\min_{\mathbf{D}\in\mathcal{D}, \mathbf{X}\in\mathcal{X}} \ \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \frac{\lambda}{2}\|\mathbf{D}^T\mathbf{D} - \mathbf{I}\|_F^2. \qquad (5)$$

Comparing with (2), it is seen that these algorithms tend to reduce the average coherence, and by varying the value of λ, we can make a balance between minimizing the data representation error and average coherence of the dictionary. A

ISBN 978-90-827970-1-5

recent algorithm, called GSD, [14] targets reducing mutual coherence of dictionary by solving the following problem

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{H}\in\mathcal{H},\mathbf{X}\in\mathcal{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \frac{\lambda}{2}\|\mathbf{D}^T\mathbf{D}-\mathbf{H}\|_F^2, \quad (6)$$

where

$$\mathcal{H} \triangleq \left\{\mathbf{H}\in\mathbb{R}^{n\times n} : \mathbf{H}=\mathbf{H}^T, h_{ii}=1,\ \forall i \max_{i\neq j}|h_{ij}| \le \mu_0\right\}$$

$\mu_0 \ge \mu_{welch}$. The dictionary is updated using gradient descent. Then, $\mathbf{H}$ is updated using the following formula [14], in which k denotes the iteration number:

$$h_{ij}^k = \begin{cases} \tau & i\neq j, |\tau|\le\mu_0 \\ \mathrm{sgn}(\tau)\mu_0 & i\neq j, |\tau|\ge\mu_0 \\ 1 & i=j \end{cases} \quad (7)$$

in which, $\mathbf{G}^k=(\mathbf{D}^k)^T\mathbf{D}^k$ and $\tau=g_{ij}^k$.

As will be seen in the next section, the second term in (5) reduces also the correlation between "rows" of the dictionary whereas the second term in (6) decreases the mutual coherence. However, problem (5) reduces the average coherence better than problem (6), because its regularization term is exactly the average coherence. Inspired by this observation and considering the usefulness of reducing both average and mutual coherences in dictionary learning, we propose a problem to impose low average as well as low mutual coherences simultaneously. This is done by tuning two trade-off parameters which balance between these two terms. Our simulation results demonstrate the effectiveness of the proposed algorithm.

The rest of the paper is organized as follows. Section 2 presents the main idea and our algorithm and discussions related to it.Then, the algorithm is numerically studied in section 3

2. THE PROPOSED ALGORITHM

Although the papers that use (5) as their cost function hope that it indirectly reduces the mutual coherence of the dictionary, it is interesting to note that it also reduces the correlation between the "rows" of the dictionary. In fact, $\mathbf{I}_n$ denoting the $n\times n$ identity matrix, one can compute:

$$\begin{aligned}
&\|\mathbf{D}^T\mathbf{D}-\mathbf{I}_n\|_F^2 \\
&\quad= \mathrm{trace}((\mathbf{D}^T\mathbf{D}-\mathbf{I}_n)(\mathbf{D}^T\mathbf{D}-\mathbf{I}_n)) \\
&\quad= \mathrm{trace}(\mathbf{D}^T\mathbf{D}\mathbf{D}^T\mathbf{D}-2\mathbf{D}^T\mathbf{D}+\mathbf{I}_n) \\
&\quad= \mathrm{trace}(\mathbf{D}\mathbf{D}^T\mathbf{D}\mathbf{D}^T-2\mathbf{D}\mathbf{D}^T+\mathbf{I}_m)+n-m \\
&\quad= \mathrm{trace}((\mathbf{D}\mathbf{D}^T-\mathbf{I}_m)(\mathbf{D}\mathbf{D}^T-\mathbf{I}_m))+n-m \\
&\quad= \|\mathbf{D}\mathbf{D}^T-\mathbf{I}_m\|_F^2+n-m. \qquad (8)
\end{aligned}$$

So, we can write:

$$\|\mathbf{D}^T\mathbf{D}-\mathbf{I}_n\|_F^2 = \|\mathbf{D}\mathbf{D}^T-\mathbf{I}_m\|_F^2+n-m. \quad (9)$$

Hence, problem (5) is actually equivalent to the following problem

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{X}\in\mathcal{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \frac{\lambda}{2}\|\mathbf{D}\mathbf{D}^T-\mathbf{I}\|_F^2, \quad (10)$$

while it is intuitively less acceptable than (5), because it proposes to design the dictionary by minimizing the correlation of its rows. This problem does not exist in (6), because $\|\mathbf{D}^T\mathbf{D}-\mathbf{H}_n\|_F \neq \|\mathbf{D}\mathbf{D}^T-\mathbf{H}_m\|_F$. Actually, it is experimentally shown in [14] that (6) reduces the mutual coherence as well. Our experiments (will be seen in Fig.1 of section 3) with cost functions (5) and (6) shows also that increasing λ in (6) results in a reduction in the mutual coherence, while increasing λ in (5) cannot reduce the mutual coherence that much.

On the other hand, it is intuitively useful to reduce the average coherence, too, which is done in (5). Actually, in some papers like [14], the final average coherence is one of the performance measures of the dictionary learning algorithms. So, in this paper, we introduce an approach for reducing mutual and average coherences jointly. The new optimization problem is:

$$\begin{aligned}\min_{\mathbf{D}\in\mathcal{D},\mathbf{H}\in\mathcal{H},\mathbf{X}\in\mathcal{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\frac{\lambda_1}{2}\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_F^2+ \\ \frac{\lambda_2}{2}\|\mathbf{D}^T\mathbf{D}-\mathbf{H}\|_F^2,\end{aligned} \quad (11)$$

which is solved by alternating minimization. To solve (11) over $\mathbf{H}$, we use the same technique as in the GSD algorithm. To solve (11) over $\mathbf{D}$, we note that the update problem is equivalent to:

$$\min_{\mathbf{D}\in\mathcal{D}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\frac{\lambda_1+\lambda_2}{2}\|\mathbf{D}^T\mathbf{D}-\frac{\lambda_1\mathbf{I}+\lambda_2\mathbf{H}}{\lambda_1+\lambda_2}\|_F^2. \quad (12)$$

Note that, the gradients of (11) and (12) over $\mathbf{D}$ are the same. So, these two problems lead to the same steepest descent iterations for updating $\mathbf{D}$. Let us denote the above cost function by $F(\mathbf{D})$. The gradient of $F(\mathbf{D})$ is:

$$\begin{aligned}\nabla_{\mathbf{D}}F(\mathbf{D},\mathbf{X}) =& 2(\mathbf{D}\mathbf{X}-\mathbf{Y})\mathbf{X}^T+ \\ & 2(\lambda_1+\lambda_2)\mathbf{D}(\mathbf{D}^T\mathbf{D}-\frac{\lambda_1\mathbf{I}+\lambda_2\mathbf{H}}{\lambda_1+\lambda_2}).\end{aligned} \quad (13)$$

We then use steepest descent for solving (12):

$$\mathbf{D}^{k+1} = \mathbf{D}^k - \alpha\nabla_{\mathbf{D}}F(\mathbf{D}^k,\mathbf{X}^{k+1}). \quad (14)$$

The suitable value for the step-size α is chosen as $\alpha \in (0, 1/\mathcal{L}]$ where $\mathcal{L}$ is a Lipschitz constant of ∇F [15], that is,

$$\|\nabla_{\mathbf{D}}F(\mathbf{D}_2,\mathbf{X})-\nabla_{\mathbf{D}}F(\mathbf{D}_1,\mathbf{X})\|_F \le \mathcal{L}\|\mathbf{D}_2-\mathbf{D}_1\|_F. \quad (15)$$

By using the above definition, a Lipschitz constant of (13) can be obtained by writing:

$$\begin{aligned}
&\nabla_{\mathbf{D}}F(\mathbf{D}_2,\mathbf{X})-\nabla_{\mathbf{D}}F(\mathbf{D}_1,\mathbf{X}) = 2(\mathbf{D}_1-\mathbf{D}_2)\mathbf{X}\mathbf{X}^T+ \\
&2(\lambda_1+\lambda_2)(\mathbf{D}_1\mathbf{D}_1^T\mathbf{D}_1-\mathbf{D}_2\mathbf{D}_2^T\mathbf{D}_2)-2(\mathbf{D}_1-\mathbf{D}_2) \\
&(\lambda_1\mathbf{I}+\lambda_2\mathbf{H}) = 2(\mathbf{D}_1-\mathbf{D}_2)\mathbf{X}\mathbf{X}^T+2(\lambda_1+\lambda_2) \\
&((\mathbf{D}_1-\mathbf{D}_2)\mathbf{D}_1^T\mathbf{D}_1+\mathbf{D}_2(\mathbf{D}_1-\mathbf{D}_2)^T\mathbf{D}_1+\mathbf{D}_2\mathbf{D}_2^T \\
&(\mathbf{D}_1-\mathbf{D}_2))-2(\mathbf{D}_1-\mathbf{D}_2)(\lambda_1\mathbf{I}+\lambda_2\mathbf{H}).
\end{aligned} \quad (16)$$

ISBN 978-90-827970-1-5

Then, utilizing the inequalities $\|\mathbf{A}+\mathbf{B}\|_F \le \|\mathbf{A}\|_F + \|\mathbf{B}\|_F$ and $\|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ [16] we get

$$\|\nabla_{\mathbf{D}}F(\mathbf{D}_2,\mathbf{X}) - \nabla_{\mathbf{D}}F(\mathbf{D}_1,\mathbf{X})\|_F \le 2(\|\mathbf{X}\|_F^2 + (\lambda_1+\lambda_2) (\|\mathbf{D}_1\|_F^2 + \|\mathbf{D}_1\|_F\|\mathbf{D}_2\|_F + \|\mathbf{D}_2\|_F^2) + (\lambda_1\|\mathbf{I}\|_F + \lambda_2\|\mathbf{H}\|_F)) \|\mathbf{D}_1 - \mathbf{D}_2\|_F. \tag{17}$$

Assuming normalized dictionary atoms, we have:

$$\begin{cases} \|\mathbf{D}_1\|_F = \|\mathbf{D}_2\|_F = \sqrt{n}, \mathbf{D} \in \mathbb{R}^{m\times n} \\ \|\mathbf{I}\|_F = \sqrt{n} \end{cases} \tag{18}$$

Therefore, an upper bound of the Lipschitz constant is obtained as

$$\mathcal{L}' = 2\|\mathbf{X}\|_F^2 + 2(\lambda_1+\lambda_2)(3n) + 2(\lambda_1\sqrt{n} + \lambda_2\|\mathbf{H}\|_F). \tag{19}$$

$$\nabla_{\mathbf{D}}F(\mathbf{D}^k,\mathbf{X}^{k+1}) = 2(\mathbf{D}^k\mathbf{X}^{k+1} - \mathbf{Y})\mathbf{X}^{k+1^T} + 2 (\lambda_1+\lambda_2)\mathbf{D}^k(\mathbf{D}^{k^T}\mathbf{D}^k - \frac{\lambda_1\mathbf{I} + \lambda_2\mathbf{H}^k}{\lambda_1+\lambda_2}). \tag{20}$$

The final algorithm, which we call Reduced Average and Mutual Coherence (RAMC), is summarized in Algorithm 1. For the sparse approximation step, we use orthogonal matching pursuit (OMP) algorithm [17]. The existence of the constant γ is because the $\mathcal{L}'$ above is only an upper bound for the Lipschitz constant.

Algorithm 1 The proposed algorithm (RAMC)

initialization: Set initial dictionary $\mathbf{D}^1$
for $k = 1$ to MaxIteration **do**
  Sparse approximation: $\mathbf{X}^{k+1}$ =OMP($\mathbf{Y}$, $\mathbf{D}^k$, s)
  Dictionary update: $\mathbf{D}^{(k+1)}$=$\mathbf{D}^k$-$\frac{\gamma}{\mathcal{L}'}$ $\nabla_{\mathbf{D}}F(\mathbf{D}^k,\mathbf{X}^{k+1})$
  Normalize the columns of $\mathbf{D}^{k+1}$
  Update $\mathbf{H}^{k+1}$ using (7)
end for

3. SIMULATION RESULTS

In this section, we compare our proposed algorithm, RAMC, with MOD, SGK [18], and GSD algorithms in synthetic dictionary recovery [4]. The performance measures are root mean square error (RMSE) defined as $\varepsilon_k = \frac{\|\mathbf{Y}-\mathbf{D}^k\mathbf{X}^k\|_F}{\sqrt{ml}}$ [12], percentage of atom recovery, mutual coherence and average coherence (2). Assuming that $\mathbf{D}_t$ is the true dictionary and $\mathbf{D}$ is the recovered dictionary, we say that the ith atom of dictionary $\mathbf{D}$ is successfully recovered if:

$$\min_j(1- \mid \mathbf{D}(:,i)^T\mathbf{D}_t(:,j) \mid) < 0.01. \tag{21}$$

For OMP, we used the available MATLAB code at `http://www.cs.technion.ac.il/~ronrubin/software.html`. We generated a Gaussian random matrix $\mathbf{D}_t \in \mathbb{R}^{30\times 50}$ with zero mean and unit variance. Then 1000 training data $\{\mathbf{y}_i\}_{i=1}^{1000}$ were generated by random linear combinations of dictionary atoms.To better compare our results with GSD, we defined $\lambda = \lambda_1+\lambda_2$, $\beta_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}$ and $\beta_2 = \frac{\lambda_2}{\lambda_1+\lambda_2}$. According to the size of the dictionary, the Welch bound (4) is obtained as $\mu_{welch} = 0.1166$ and we chose $\mu_0 = \mu_{welch}$. By varying the values of β_1 and β_2, we can make a trade-off between the representation error, mutual coherence and average coherence of the dictionary. We performed 100 iterations between the sparse coding and dictionary updating. The dictionary was initialized by randomly choosing different signals from the training set followed by a normalization. We repeated all simulations 500 times and the averaged results are reported here. The value of γ was empirically chosen equal to 40 in all simulation.

Figure 1 is to verify our primary motivation concerning that (5) is suitable for decreasing the average coherence while (6) is suitable for decreasing the mutual coherence.

Figures 2 to 9 are the results of simulation of our algorithm and its comparisons with other mentioned algorithms. As observed from these figures, when β_1 increases, the average coherence reduces, whereas if β_2 increases, the mutual coherence decreases. This was expected, as β_1 and β_2 determine the contribution of the identity matrix (responsible for average coherence) and matrix $\mathbf{H}$ (responsible for mutual coherence) in the coherence reduction term; see (12). Note that, the higher λ, the higher RMSE. Based on the simulations, our algorithm reduces mutual and average coherence better than the GSD algorithm, while both of them have nearly the same error and percentage of recovery. The key point of this simulation, we should set β_1 and β_2 so that the learned dictionary has a low coherence and a satisfactory percentage of recovery.

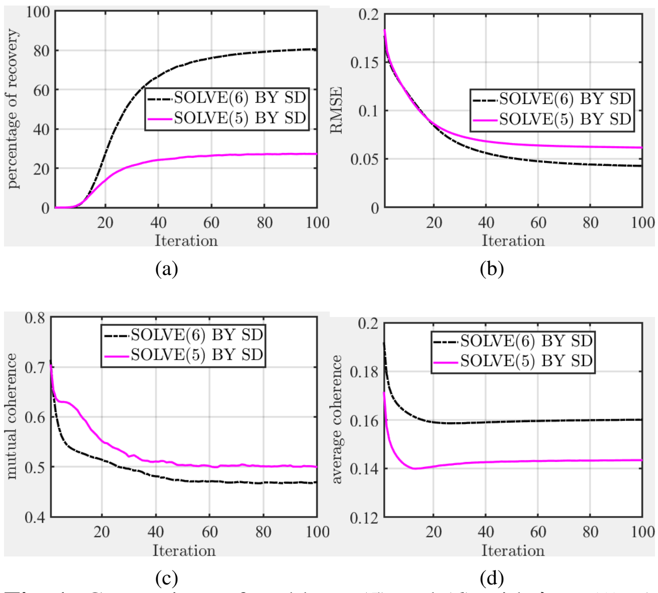

Fig. 1: Comparison of problems (5) and (6) with $\lambda = 15$. a) percentage of recovery; b) RMSE; c) mutual coherence; d) average coherence.

4. CONCLUSION

In this paper, we proposed a new algorithm for learning dictionaries with both low average and low mutual coherence. In contrast to previous approaches, our algorithm reduces mutual and average coherence of the dictionary jointly. This is

ISBN 978-90-827970-1-5

done by adding two terms responsible for reducing the coherences. Our simulations showed the superiority of our algorithms compared to a set of well-known algorithms.

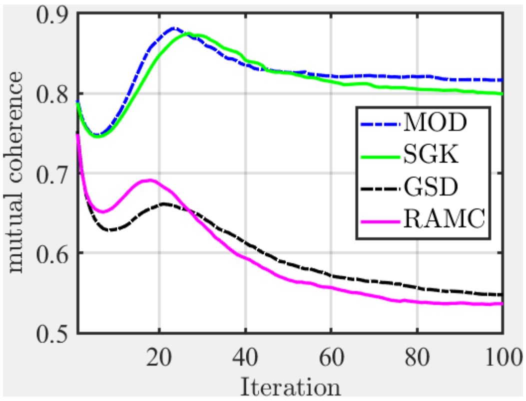

Fig. 2: Evaluation of mutual coherence with assumptions: SNR $=30$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 8$, for both algorithms RAMC and GSD.

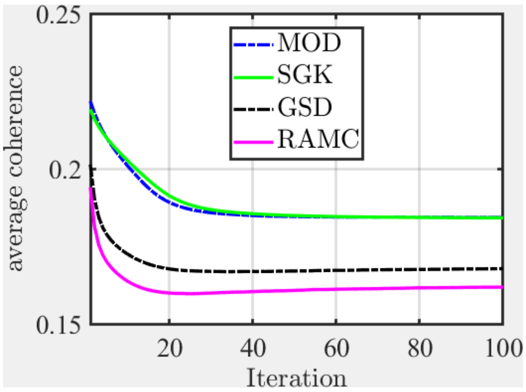

Fig. 3: Evaluation of average coherence with assumptions: SNR $=30$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 8$, for both algorithms RAMC and GSD.

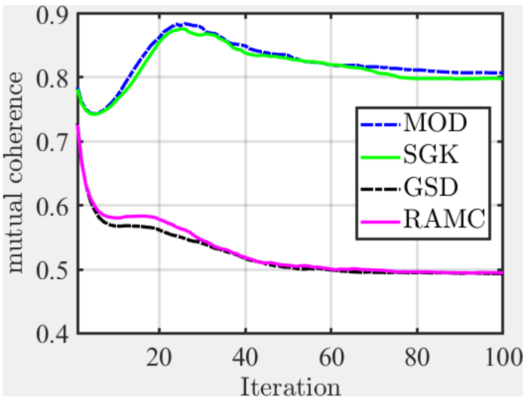

Fig. 6: Evaluation of mutual coherence with assumptions: SNR $=30$, $\beta_1 = 0.15$, $\beta_2 = 0.85$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 12$, for both algorithms RAMC and GSD.

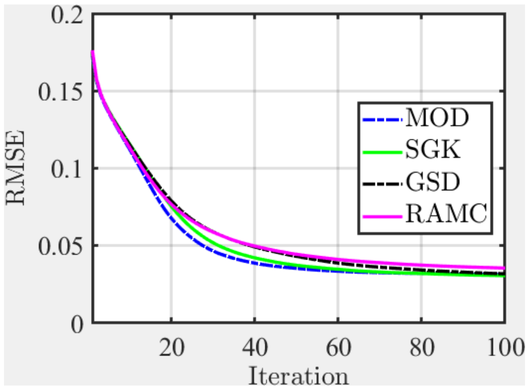

Fig. 4: Evaluation of RMSE with assumptions: SNR $=30$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 8$, for both algorithms RAMC and GSD.

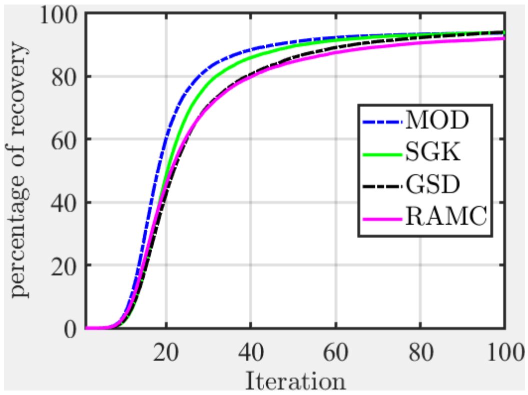

Fig. 5: Evaluation of recovery with assumptions: SNR $=30$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 8$, for both algorithms RAMC and GSD

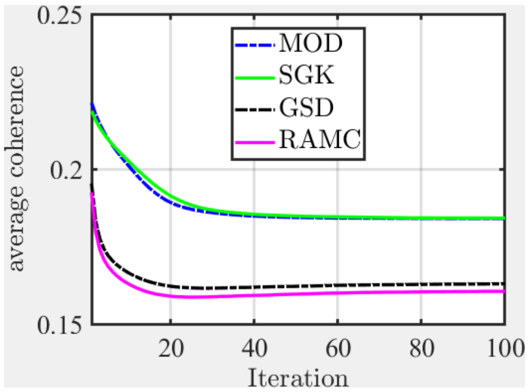

Fig. 7: Evaluation of average coherence with assumptions: SNR $=30$, $\beta_1 = 0.15$, $\beta_2 = 0.85$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 12$, for both algorithms RAMC and GSD.

ISBN 978-90-827970-1-5 © EURASIP 2018

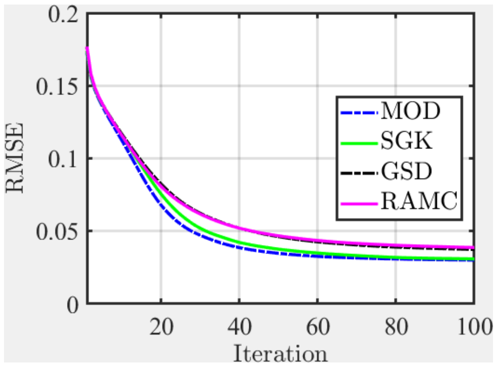

Fig. 8: Evaluation of RMSE with assumptions: SNR $= 30$, $\beta_1 = 0.15$, $\beta_2 = 0.85$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 12$, for both algorithms RAMC and GSD.

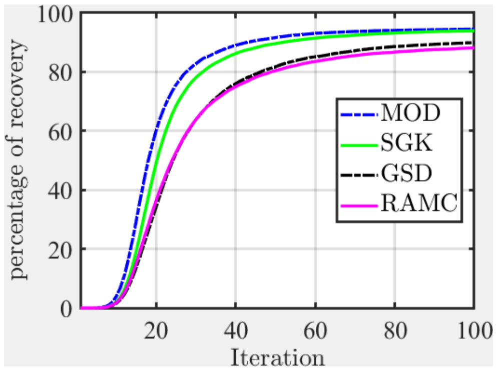

Fig. 9: Evaluation of recovery with assumptions: SNR $= 30$, $\beta_1 = 0.15$, $\beta_2 = 0.85$, sparsity level=5, $\mu_0 = \mu_{welch}$, $\lambda = 12$, for both algorithms RAMC and GSD.

5. REFERENCES

[1] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[2] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[3] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Dictionary learning for sparse representation: A novel approach," *IEEE Signal Proc. Letters*, vol. 20, no. 12, pp. 1195–1198, 2013.

[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[5] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proceedings of IEEE ICASSP*, 1999, vol. 5.

[6] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Learning over-complete dictionaries based on atom-by-atom updating," *IEEE Trans. on Signal Proc.*, vol. 62, no. 4, pp. 883–891, 2014.

[7] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[8] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Aca. Sci*, vol. 100, no. 5, pp. 2197–2202, 2003.

[9] T. Strohmer and Heath R. W., "Grassmannian frames with applications to coding and communication," *Applied and Computational Harmonic Analysis*, vol. 14, no. 3, pp. 257–275, 2003.

[10] W. Chen, M. R. D. Rodrigues, and J. I. Wassell, "Projection design for statistical compressive sensing: A tight frame based approach," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2016–2029, 2013.

[11] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Learning dictionaries with bounded self-coherence," *IEEE Signal Proc. Letters*, vol. 19, no. 12, pp. 861–864, 2012.

[12] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "A new algorithm for learning overcomplete dictionaries," in *Proceedings of 21th European Signal Processing Conference (EUSIPCO 2013)*, 2013.

[13] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Regularized low-coherence overcomplete dictionary learning for sparse signal decomposition," in *Proceedings of 24th European Signal Processing Conference (EUSIPCO 2016)*, 2016.

[14] G. li, Z. Zhu, H. Bai, and A. Yu, "A new framework for designing incoherent sparsifying dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[16] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[17] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *In Proc. Asilomar Conf. Signal Syst. Comput.*, 1993.

[18] S.K Sahoo and A Makur, "Dictionary training for sparse representation as generalization of k-means clustering," *IEEE Signal Proc. Letters*, 2013.