SPARSE CODING WITH A GLOBAL CONNECTIVITY CONSTRAINT

R.M. Thomas[1], *S. Yatawatta*[2], *and C. Keysers*[1]

[1]Netherlands Institute for Neuroscience, Amsterdam, The Netherlands
[2]The Netherlands Institute for Radio Astronomy, Dwingeloo, The Netherlands
Email: r.thomas@nin.knaw.nl, yatawatta@astron.nl, c.keysers@nin.knaw.nl

ABSTRACT

Basis pursuit via sparse coding techniques have generally enforced sparseness by using L1-type norms on the coefficients of the bases. When applied to natural scenes these algorithms famously retrieve the Gabor-like basis functions of the primary visual cortex (V1) of the mammalian brain. In this paper, inspired further by the architecture of the brain, we propose a technique that not only retrieves the Gabor basis but does so respecting global power-law type connectivity patterns. Such global constraints are beneficial from a biological perspective in terms of efficient wiring, robustness etc. We draw on the similarity between sparse coding and neural networks to formulate the problem and impose such global connectivity patterns.

Index Terms— sparse coding; scale-free networks; biologically inspired

1. INTRODUCTION

Since Olshausen and Field [1, 2] first published their papers on the emergence of Gabor type wavelets via the statistics of sparse coding, there has been much excitement in using sparse coding to explain the tuning curves of neuronal ensembles, especially in the primary sensory cortices [3, 4]. Although these efforts have been successful to a large extent, they have not attempted to explain how this sparse coding might be implemented in the brain as a network in "hardware". While an exact description of the neural correlates requires substantial research, in this article we propose to use a network model following the formalism of an autoencoder construction [5], used extensively in the "deep learning" community. We translate sparse coding in its language and more importantly develop regularisations grounded on neurobiology.

There is evidence from neuroscience [6, 7, 8] that at various levels of hierarchy in the mammalian brain the network architecture follows a power-law in its degree distribution, specially that of a scale-free network [9]. Although there have been attempts to explain this architecture [7], the underlying reasons of its emergence in the brain is debatable. Scale-free like small world networks do have properties that are also biologically advantageous, like reduced wiring costs and its robustness to random attacks on nodes.

Natural images are imbued with statistical dependencies that often require n-point correlations to describe, with $n > 2$ typically [10]. This high order dependencies arise because of the prevalence of oriented edges, curves and other fractal-like forms in a natural scene. An efficient code, under the "principle of redundancy reduction" [11], should extract these dependencies. Given an image I, the primary objective of the coding algorithm is to find a set of basis functions ϕ_i, that minimizes the least squares cost $\| I - \sum_{i=1}^{n} a_i \phi_i \|^2$, where a_i are the coefficients of the representation. The network architecture of this is given in Fig. 1.

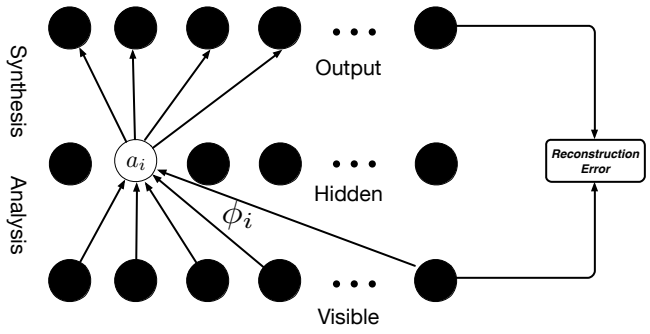


Fig. 1. The network model: Each pixel of the input image I is a node on the visible layer at the bottom. The visible to the hidden layer consists of the "analysis" step and the incoming signals are weighted by a vector ϕ_i which is the basis vector. The activation of the hidden node a_i and ϕ_i are then composed as $\sum_i a_i \phi_i$ to form the synthesis stage from hidden to the output layer. This output is then compared to the input image to generate the reconstruction error which is used to learn and adapt the activation and the basis.

In addition to minimizing the least squares error, the cost function is often augmented by the constraint that coefficients a_i are "sparse". The reasons to impose this constraint include: (i) the resulting basis are qualitatively similar to the simple cell receptive fields, (ii) the intuition that natural images can

be described succinctly with fewer appropriate structure primitives like lines, edges and curves, and (iii) the post-facto phenomenon that this sparseness minimizes wiring costs [12] and facilitates efficient metabolism [13].

A crucial requirement of real-world systems, biological or otherwise, is their ability to perform unhindered under attack or during the malfunctioning of one or many of their nodes/connections and power-law type small-world networks have been shown to possess these properties [14]. Also, empirical evidences suggest that functional organization of the brain can be well approximated using scale-free model [6] and it has been claimed [13] that sparse power-law networks do implement optimal energy consumption.

Relation to prior work: In this paper, we modify the regularization term to impose sparseness with a structure, i.e., nodes follow a degree distribution d that has a power-law $\propto d^{-\alpha}$ where α is positive. Sparseness [1, 2, 15, 16] and scale-freeness [17, 6, 18, 7] have already been enforced in a significant amount of existing work, but almost always as individual constraints. We enforce both these constraints together in this work, and that is the novelty of this paper.

The paper is organized as follows: §2 introduces the formalism of the sparse coding problem and the algorithm used in this work. In §3, we give results of our proposed algorithm applied onto a set of images of natural scenes. We also give comparisons with existing sparse coding algorithms that do not enforce scale-freeness. Finally, we draw our conclusions in §4.

Notation: All matrices are represented as bold uppercase letters. The Frobenious norm is given by $\|.\|$ while the L1 norm is given by $\|.\|_1$.

2. MATHEMATICAL PRELIMINARIES

Consider a set of images, that are similar in resolution and dynamic range. Let each image to have L pixels, and consider having B such images. We represent the full set of images as a matrix $\mathbf{X}$ ($\in \mathbb{R}^{L\times B}$). Consider a set of M basis functions that are used to reconstruct the images in $\mathbf{X}$. Ideally, we have

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{A}$ ($\in \mathbb{R}^{L\times M}$) gives the basis, where each column represents one basis function. The coefficients of the decomposition is given by $\mathbf{S}$ ($\in \mathbb{R}^{M\times B}$), where each column of $\mathbf{S}$ represents the coefficients required to reconstruct one image.

The objective of this work is to find $\mathbf{S}$ for each given $\mathbf{X}$, such that $\mathbf{S}$ is sparse. Moreover, we enforce the scale-freeness into columns of $\mathbf{A}$. Given a Gaussian noise model, the cost function that needs to be minimized (or the log-likelihood that needs to be maximized) is

$$f(\mathbf{A},\mathbf{S}) = \|\mathbf{X}-\mathbf{AS}\|^2 = \text{trace}\left((\mathbf{X}-\mathbf{AS})^T(\mathbf{X}-\mathbf{AS})\right). \tag{2}$$

Note that in (2), $f(\mathbf{A},\mathbf{S})$ is dependent on the value of $\mathbf{X}$ and $\mathbf{X}$ depends on the input images. We impose additional constraints on $\mathbf{S}$ and $\mathbf{A}$ to get the required properties.

1. *Sparsity*: We enforce that the coefficients of $\mathbf{S}$ are sparse, in other words, we use a regularization term $\lambda_s\|\text{vec}(\mathbf{S})\|_1$ where λ_s is a scalar. With this, we estimate $\mathbf{S}$ as

$$\mathbf{S} = \arg\min_{\mathbf{S}} \left(f(\mathbf{A},\mathbf{S}) + \lambda_s\|\text{vec}(\mathbf{S})\|_1\right) \tag{3}$$

2. *Scale Freeness*: We enforce scale freeness onto the columns of $\mathbf{A}$, where we denote the i-th column as $\mathbf{A}(:,i)$. Following [18], although ideally the scale freeness should be imposed by using L0 norm as a regularization term, we use (for the log-likelihood) $\sum_i \log(\|\mathbf{A}(:,i)\|_1+\epsilon)$ as a surrogate and we have

$$\mathbf{A} = \arg\min_{\mathbf{A}} \left(f(\mathbf{A},\mathbf{S}) + \lambda_a \sum_i \log(\|\mathbf{A}(:,i)\|_1+\epsilon)\right) \tag{4}$$

where ϵ and λ_a are scalar constants. However, solving (4) is computationally not feasible and as in [18], we use the majorization-minimization algorithm [19], where we find an upper bound to the regularization term as

$$\begin{aligned}&\log(\|\mathbf{A}(:,i)\|_1+\epsilon) \\ &\quad\le \log(\|\mathbf{A}^n(:,i)\|_1+\epsilon) + \frac{\|\mathbf{A}(:,i)\|_1+\epsilon}{\|\mathbf{A}^n(:,i)\|_1+\epsilon} - 1 \\ &\quad= c_1 + \left(\frac{1}{\|\mathbf{A}^n(:,i)\|_1+\epsilon}\right)\|\mathbf{A}(:,i)\|_1\end{aligned} \tag{5}$$

where $\mathbf{A}^n(:,i)$ is constant (value at n-th iteration of the optimization). The regularization weight (for the i-th column) is $\frac{1}{\|\mathbf{A}^n(:,i)\|_1+\epsilon}$ while c_1 is constant. Hence, we minimize the majorized version of (4) to find $\mathbf{A}$.

We see that both (3) and (4) can be re-cast as L1 regularized least squares minimization problems. Note that in (4), the regularization weight also changes with iterations. There is a wide variety of algorithms to solve this problem and we use *L1General* [20] software to perform this minimization. The only extra information needed is the gradient of the least squares cost function

$$\frac{\partial f}{\partial \mathbf{S}} = -2\mathbf{A}^T(\mathbf{X}-\mathbf{AS}),\ \ \frac{\partial f}{\partial \mathbf{A}} = -2(\mathbf{X}-\mathbf{AS})\mathbf{S}^T. \tag{6}$$

With the basic principles described above, our proposed algorithm (which is essentially similar to [2]) for sparse coding can be described as follows:

1. Initial value for $\mathbf{A}$ is chosen randomly. The learning rate $\eta \in (0,1]$ is given.

2. Randomly sample the set of images to build $\mathbf{X}$. Using (3), estimate $\mathbf{S}$, say $\widehat{\mathbf{S}}$.
3. Using (4), find the estimate for $\mathbf{A}$, say $\widehat{\mathbf{A}}$.
4. Update $\mathbf{A} \leftarrow \mathbf{A} + \eta(\widehat{\mathbf{A}} - \mathbf{A})$
5. Normalize $\mathbf{A}$ to preserve $\|\mathbf{X}\| \approx \|\mathbf{A}\widehat{\mathbf{S}}\|$.
6. If maximum number of iterations has reached, stop, else go to step 2.

3. RESULTS ON NATURAL IMAGES

We used a publicly available set of images [21] to test our algorithm. From this database we selected approximately $B = 1000$ images of size 2048×2048 pixels. All images were pre-processed according to [2]: We first pre-whiten the images by an exponential function $f \exp(-(f/f_o)^4)$ where f_o is a fraction of size of the image ($f_o = 0.4 \times 2048$), and we rescale the filtered images such that the average image variance is a number in $[0, 1]$. Square patches of size 16×16 ($L = 256$) were extracted from the natural scenes to train the network, which had $M = 256$ hidden nodes.

Fig. 2 shows the comparison between basis functions trained with a Cauchy prior as in [1] in (a) and using the prior in §2 in (b). Both the schema do provide us with a complete basis set. Although oriented edge-like features are present in both cases, it appears qualitatively that the scale-free networks have fewer redundant basis functions. In what is to follow we see that scale-free network performs better in different metrics considered here.

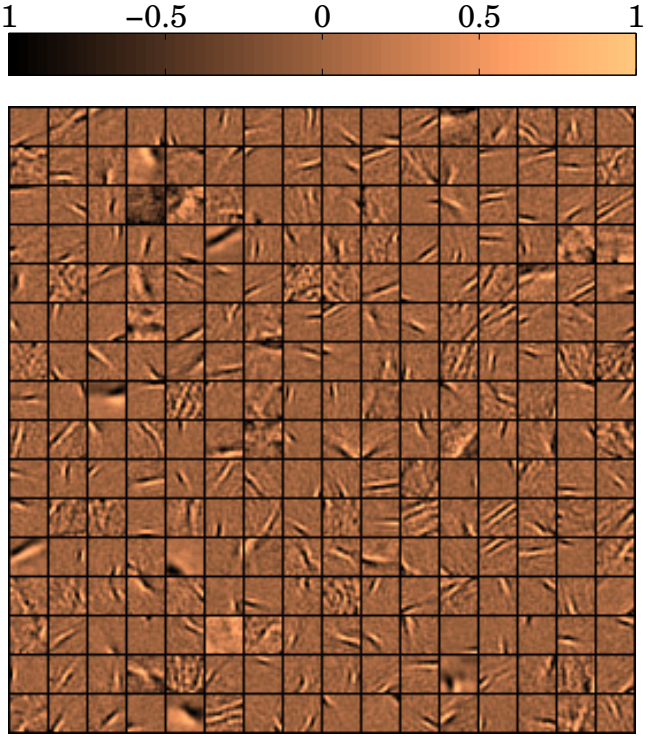
3.1. Degree distribution

The degree distribution of the nodes are compared in Fig. 3. Because this is a weighted network as in Fig. 1, we estimate the degree distribution following [22]; the total strength of the connectivity at node i is calculated as $s_i = \sum |\mathbf{A}(:, i)|$ and the degree k_i is such that $s_i \approx k_i \langle \mathbf{A}(:, :) \rangle$ where $\langle \mathbf{A}(:, :) \rangle$ is the average of absolute value of the connection weights in the network.

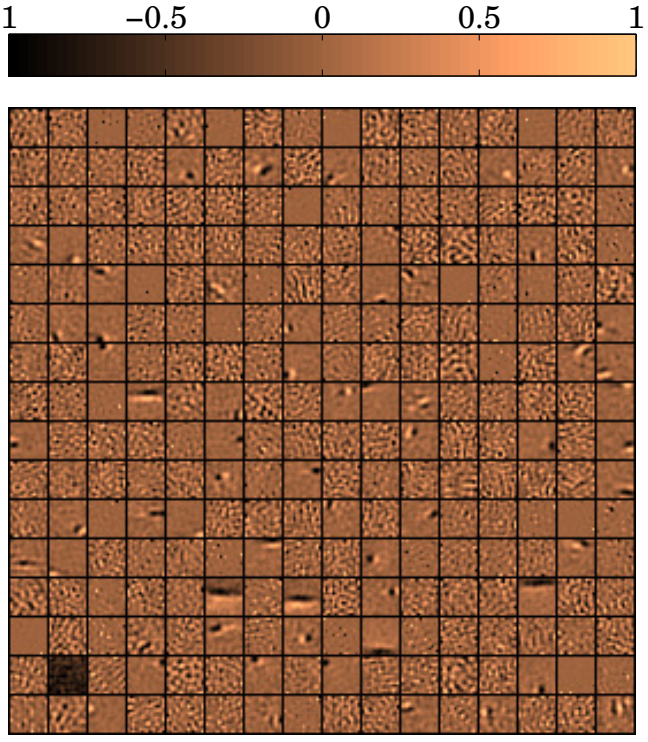
As is evident from Fig. 3, our algorithm has achieved a degree distribution that is power-law for about two orders of magnitude, whereas the sparse code as in [1] has its degree concentrated in a small window. There is an over abundance of degrees at the high k end and we presume that this will improve if we include a much larger training dataset.

3.2. Reconstruction error

In order to estimate the reconstruction error given the basis functions, we randomly selected areas from the B images and reconstructed them. The residual error is shown in Fig. 4. The vertical lines show the mean of the corresponding error. The standard deviations (dotted) of the scale-free network is



(a)



(b)

Fig. 2. (a) Basis functions learnt using [1]. (b) Using the scale free connectivity constraint as in §2

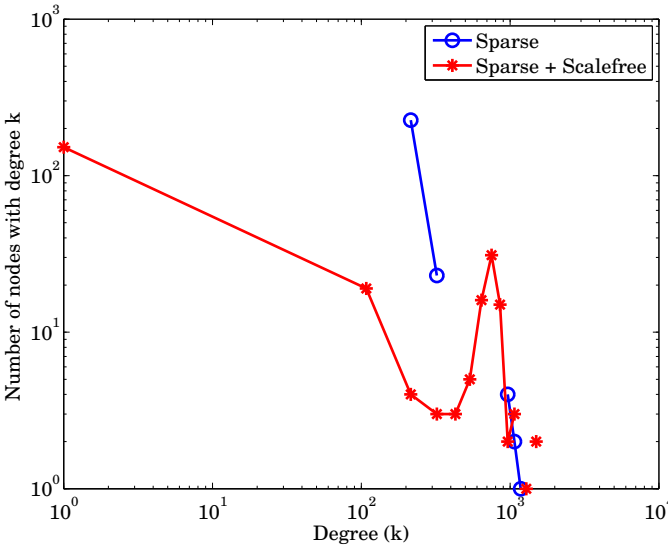


Fig. 3. Comparison of the degree distribution between [1] (blue) and scale-free connectivity (red). The scale-free network indeed spans a large range of degrees according to a power-law.

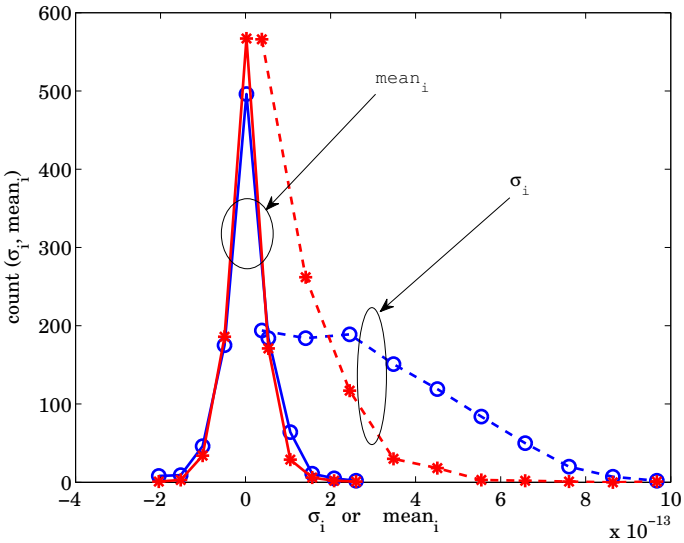


Fig. 4. Comparison of the reconstruction error standard deviation (dotted) and mean (solid) between [1] (blue) and scale-free connectivity (red). Note: the standard deviation has been equally scaled down to fit in the figure.

indeed lower than that of sparse-only implementation. This is not something that was explicitly modelled but is a welcome outcome.

3.3. Activation distribution

Finally, in Fig. 5, the distribution of the average activation is shown. The activation peaks at zero for both models because we expect the sparse code to be efficient and most of the time (any given reconstruction) only a few basis functions are required and the rest is inactive.

We also see here that scale-free (red) implementation is more skewed towards zero suggesting that fewer basis functions were required for the reconstruction.

4. CONCLUSIONS

We have implemented a neural network scheme that does sparse coding with the additional constraint that the degree distribution follows a power-law. The motivation for this comes from biology where we see scale-free behaviour in neural circuits at various levels. It turns out that this scale-free nature also provides desirable error and activation properties. In the future this framework will be used to simulate the robustness of the network to random attacks and also how re-training a network like this after an attack can alter its functionality – plausible scenarios in a real-world network.

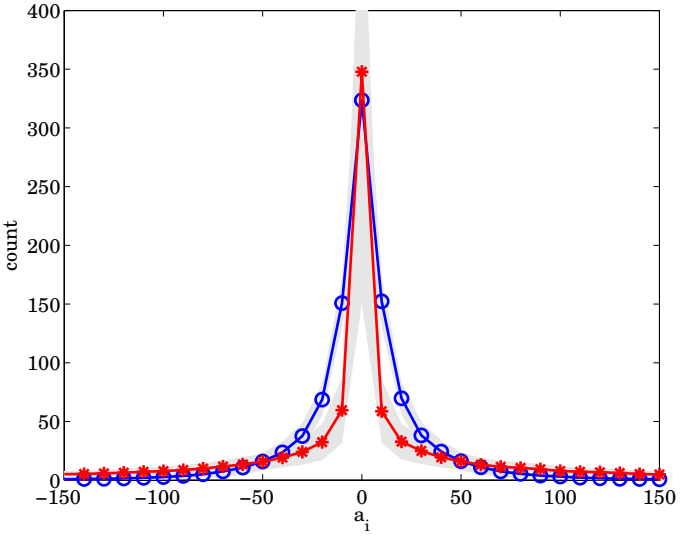


Fig. 5. Distribution of the activation of nodes for [1] (blue) and scale-free connectivity (red). Both are highly peaked at zero as expected but the scale-free network is more sharply peaked indicating that the basis might be "more optimal" for image representation.

5. REFERENCES

[1] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun 1996.

[2] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, Dec 1997.

[3] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, Feb 2006.

[4] Andrew M. Saxe, Maneesh Bhand, Ritvik Mudur, Bipin Suresh, and Andrew Y. Ng, "Unsupervised learning models of primary cortical receptive fields and receptive field plasticity," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, Eds., pp. 1971–1979. 2011.

[5] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009, Also published as a book. Now Publishers, 2009.

[6] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, "Scale-free brain functional networks," *Phys. Rev. Lett.*, vol. 94, no. 1, pp. 018102, Jan 2005.

[7] J. Piersa, F. Piekniewski, and T. Schreiber, "Theoretical model for mesoscopic-level scale-free self-organization of functional brain networks," *IEEE Trans Neural Netw*, vol. 21, no. 11, pp. 1747–1758, Nov 2010.

[8] C. A. S. Batista, A. M. Batista, J. A. C. de Pontes, R. L. Viana, and S. R. Lopes, "Chaotic phase synchronization in scale-free networks of bursting neurons," *Phys. Rev. E*, vol. 76, pp. 016218, Jul 2007.

[9] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun 1998.

[10] D.J. Field, "Scale-invariance and Self-similar 'Wavelet' Transforms: an Analysis of Natural Scenes and Mammalian Visual Systems," in *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications*, M. Farge, J. Hunt, and Vassilicos, Eds., pp. 151 – 193. Oxford University Press, Oxford, 1993.

[11] H.B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, W.A. Rosenblith, Ed., pp. 217 – 234. MIT Press, Cambridge, MA, 1961.

[12] P. Foldiak, "Sparse coding in the primate cortex," in *The handbook of brain theory and neural networks*, M.A. Arbib, Ed., pp. 895 – 989. MIT Press, Cambridge, MA, 1995.

[13] R. Baddeley, "Visual perception. An efficient code in V1?," *Nature*, vol. 381, no. 6583, pp. 560–561, Jun 1996.

[14] Hidefumi Sawai, "A small-world network immune from random failures and resilient to targeted attacks," *Procedia Computer Science*, vol. 18, pp. 976 – 985, 2013.

[15] Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun, "Sparse feature learning for deep belief networks," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., pp. 1185–1192. MIT Press, Cambridge, MA, 2008.

[16] Shenghua Gao, I.W.-H. Tsang, and Liang-Tien Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 92–104, Jan 2013.

[17] A. Defazio and T. S. Caetano, "A Convex Formulation for Learning Scale-Free Networks via Submodular Relaxation," in *Advances in Neural Information Processing Systems*, 2012, pp. 1259–1267.

[18] Q. Liu and A. T. Ihler, "Learning scale free networks by reweighted L1 regularization," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 40–48.

[19] T. T. Wu and K. Lange, "The MM alternative to EM," *Statistical Science*, vol. 25, no. 4, pp. 492–505, 2010.

[20] Mark Schmidt, Glenn Fung, and Rómer Rosales, "Fast optimization methods for L1 regularization: A comparative study and two new approaches," in *Proceedings of the 18th European Conference on Machine Learning*, Berlin, Heidelberg, 2007, ECML '07, pp. 286–297, Springer-Verlag.

[21] W.S. Geisler and J.S. Perry, "Statistics for optimal point prediction in natural images ," *Journal of Vision*, vol. 11, no. 12, 2011.

[22] Marc Barthálemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani, "Characterization and modeling of weighted networks," *Physica A: Statistical Mechanics and its Applications*, vol. 346, pp. 34 – 43, 2005.