Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

On Semi-Supervised Learning and Sparsity

Alexander Balinsky
Cardiff School of Mathematics
Cardiff University
Cardiff , United Kingdom
BalinskyA@cardiff.ac.uk

Helen Balinsky
Hewlett Packard Labs
Bristol, United Kingdom
Helen.Balinsky@hp.com

***Abstract*—In this article we establish a connection between semi-supervised learning and compressive sampling. We show that sparsity and compressibility of the learning function can be obtained from heavy-tailed distributions of filter responses or coefficients in spectral decompositions. In many cases the NP-hard problems of finding sparsest solutions can be replaced by l^1-problems from convex optimisation theory, which provide effective tools for semi-supervised learning. We present several conjectures and examples.**

***Index Terms*—Semi-supervised learning, compressive sampling, heavy-tailed distributions, sparsity.**

I. Introduction and Notations

There are striking similarities between goals in *semi-supervised learning* (SSL) and *compressive sampling* (CS). Semi-supervised learning refers to the problem of learning from labeled and unlabelled data [1]. In this case, the data set $X = (x_i)_{i\in[n]}$ can be divided into two parts: the points $X_l := (x_1, \ldots, x_l)$, for which values of the learning function F are provided, i.e. $y_i = F(x_i),\ i = 1, \ldots, l$, are given, and the points $X_u := (x_{l+1}, \ldots, x_n)$ for which values of F are not known. The main goal of SSL is to calculate the function F on X_u based on the "geometry" of the set X. The theory now known as "Compressed Sensing" or "Compressive Sampling" allows the faithful recovery of "compressible/sparse" signals from a very limited number of fixed measurements [2]. To relate SSL and CS we are going to establish connections between the "geometry" of the set X and the "compressibility/sparsity" of the learning function/signal F.

Graph-based methods in SSL give us a very natural way to represent the geometry of the training set X [1]. This geometry can be represented by a weighted graph $G = (X, E, w)$ where nodes X represent the training data and edges E with weights given by a positive function $w : E \to \mathbb{R}_+$ represent "similarities" between nodes. Here, the weight $w(e)$ of an edge e indicates the similarity of the incident nodes (and a missing edge corresponds to zero similarity). The weighted adjencency matrix (or weight matrix) W of the graph G is defined by

$$W_{ij} = \begin{cases} w(e) & \text{if} \quad e = (x_i, x_j) \in E, \\ 0 & \text{if} \quad (x_i, x_j) \notin E. \end{cases}$$

Without loss of generality we can assume that the graph G is a connected graph. If our training set X is a subset of some metric space, then the weight matrix W can be, for instance, the k-nearest matrix: $W_{ij} = 1$ iff x_i is among the k-nearest neighbors of x_j or vice versa (and is 0 otherwise). Another popular choice of weight matrix is the Gaussian kernel of width σ:

$$W_{ij} = e^{-\frac{d(x_i,x_j)^2}{2\sigma^2}}, \tag{1}$$

where $d(x_i, x_j)$ denotes the distance between x_i and x_j in the metric space.

The diagonal matrix D defined by $D_{ii} = \sum_j W_{ij}$ is called the degree matrix of G. Let us denote by P the matrix of transition probabilities $D^{-1}W$ of the graph G, i.e.

$$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}.$$

We have $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$ for all i.

For a real vector $z = (z_1, \ldots, z_N) \in \mathbb{R}^N$ we shall denote by $||z||_0$ the sparsity of z, i.e. the number of nonzero elements in $(z_1, \ldots, z_N)$. Also for $0 < p < \infty$, we set $||z||_p := \left(\sum_{j=1}^N |z_j|^p\right)^{1/p}$. $||\cdot||_p$ is a true norm on $\mathbb{R}^N$ only for $p \geq 1$. Very often we call $||\cdot||_p$ the l^p-norm even for $p < 1$. This is a standard abuse of terminology: $||\cdot||_0$ is not positive homogeneous and $||\cdot||_p$ does not satisfies the triangle inequality for $p < 1$ (i.e., the unit l^p-ball in $\mathbb{R}^N$ is not a convex set). Nevertheless, $||\cdot||_p^p$ satisfies the triangle inequality for $0 < p < 1$ and induces a metric on $\mathbb{R}^N$.

Let L be a linear subspace of $\mathbb{R}^N$ and $b \in \mathbb{R}^N$. The following l^p-optimisation problem plays a crucial role in many applications

$$\min_{z\in L}\{||b - z||_p\}, \tag{2}$$

i.e., we want to approximate the vector b by elements from L with minimum l^p-error. For $p \geq 1$ this is a convex optimisation problem and many algorithms have been developed to find a minimiser. The case $p = 1$ (also called *basis pursuit*) is a special case since the l^1-sphere in $\mathbb{R}^N$ is not a smooth manifold. For $p > 1$ the problem (2) has a unique minimiser. In applications the linear subspace L appears in two forms: (1) as the image of a linear transform $A : \mathbb{R}^M \to \mathbb{R}^N$ with $M < N$, or (2) as the kernel of a linear transform $A : \mathbb{R}^N \to \mathbb{R}^M$ with $M < N$. The first form appears, for example, in the context of channel codding, and the second form appears in the context of sampling theory. The case $p = 2$ is the standard least squares problem. The optimisation problem (2) with $p < 1$ and especially $p = 0$ has attracted a lot of interest recently in the context of compressive sampling and statistics with high dimensionality [2], [3]. The l_0-problem (2) is combinatorial

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

and generally NP-hard [4]. One of the main achievements of CS theory is the fact that, under some mild conditions, the convex l_1-problem (2) yields the same minimiser as the l_p-problem (2) with $0 \leq p < 1$.

In Sec.II we show how the optimisation problem (2) with the subspace L in the first form is related to SSL by the statistics of filter response. In Sec.III we relate the compressibility of the learning function with the optimisation problem (2) and L in the second form.

As an example of the application of our results, we consider the *colourisation problem* of natural images [5] as a problem of SSL. We are given a grey image and an expert marks several pixels in the image with colours. After that we would like to learn the colours of all other pixels. We are working in the colour space YUV with intensity Y (grey image) and chromacity channels U and V. The graph-based setting of this problem can be performed as follows. We consider the grey image as a surface in $\mathbb{R}^3$ (the graph of the function Y). This surface will be our data set X and X_l will be the coloured points. The learning function is U (and equivalently V). As the weight matrix for this data set we are using (1) with $d(x_i, x_j)$ a distance in $\mathbb{R}^3$. Such choice of the weight matrix appear in image processing under the name of *bilateral filters*.

II. Sparsity of Filter Response

For semi-supervised learning to work, certain assumptions will have to hold. One of the most popular of such assumptions is the *smoothness assumption of semi-supervise learning*: If two points x_i, x_j are close, then so should be the corresponding outputs $F(x_i), F(x_j)$. If we want to take into account the density of the input, then we can also say that if two points x_i, x_j in a high-density region are close, then so should be the corresponding outputs $F(x_i), F(x_j)$ (see [1] for details).

We can make the smoothness assumption more precise by saying that for the learning function F we should have "small" fluctuations, i.e.

$$\gamma(F)(x_i) := F(x_i) - \sum_{j=1}^{n} P_{ij} F(x_j) \approx 0,$$

where P_{ij} are transition probabilities. Using the language of image processing we shall call $\gamma(F)(x_i)$ a filter response at point x_i.

We can now say that one of the main tasks of SSL is to find a function F under the constraints $F(x_i) = y_i,\ i = 1, \ldots, l$, such that the filter responses $\gamma(F)(x_i)$ are "small". To make the problem precise we should define what "small" means.

If given values $y_1, \ldots, y_l$ are not all the same, then we can't make all $\gamma(F)(x_i)$ zero. This fact is a discrete analogue of the famous Liouville theorem which says that harmonic functions on compact connected Riemannian manifolds are all constants. In our case we can prove this as follows. Suppose that F is a function on the data set X such that all filter responses $\gamma(F)(x_i)$ are zero. Since X is a finite set, then F has a maximum value at some points x_{i_0}. From $F(x_{i_0}) = \sum_{j=1}^{n} P_{i_0 j} F(x_j)$ and $P_{i_0 j} \geq 0$, $\sum_j P_{i_0 j} = 1$ we can conclude that $F(x_j) = F(x_{i_0})$ for all neighbours x_j of x_{i_0} in the graph G. Now we can do the same for all neighbours x_j and so on. Finally, connectivity of the graph implies that F is a constant.

Several researches in the area of SSL used the quadratic criterion (see [1] for example) as a measure of smallness of the filter responses. More precisely they proposed to find F by minimisation of the quadratic error:

$$\min \sum_{i=1}^{n} (\gamma(F)(x_i))^2$$

under constraints $F(x_i) = y_i,\ i = 1, \ldots, l$. As it happens very often with the least squares method, for the resulting minimising function F too many of the $\gamma(F)(x_i)$ are not zero. When there are outliers in the data, the quadratic criterion often has poor performance.

From the probabilistic point of view the quadratic criterion means that we consider $\gamma(F)(x_i)$ as Gaussian noise. Similar to image processing, such an assumption results in oversmoothing of the learning function and blurring boundaries between clusters.

Let us perform a simple Bayesian analysis of the SSL problem: We are given the graph G and the restriction F_0 of the learning function F on the subset X_l, $F|_{X_l} = F_0$. For any event A let us denote by $P_G(A)$ the conditional probability $P(A|G)$. Then we wish to maximise $P_G(F|F_0)$. Applying Bayes' formula results in maximising

$$P_G(F_0|F) \cdot P_G(F),$$

or equivalently to find

$$\arg\max_F P_G(F) \tag{3}$$

under condition $F|_{X_l} = F_0$.

Remark. If we want to keep F_0 exactly, then $P_G(F_0|F)$ is 1 if $F|_{X_l} = F_0$ and zero otherwise. If we assume that we can have some errors in F_0 then $P_G(F_0|F)$ can be modelled as a Gaussian distribution for $F|_{X_l} - F_0$.

We model $P_G(F)$ by marginal probability of filter responses. Let us remind that *sparse distributions* are distributions that generate values for the component of a vector v close to zero most of the time, but occasionally far from zero. Sparse distributions are defined as being more likely than Gaussian of the same mean and variance to generate values near zero and also more likely to generate values far from zero. These occasional high values can convey substantial information. Distributions with this character are also called *heavy-tailed*. In the real world many filter responses are heavy-tailed and are very far from Gaussian [6], [7], [8], [9], [10]. Sparseness has been defined in a variety of different ways. Sparseness of a distribution is sometimes linked to a high value of a measure called *kurtosis*. Kurtosis of the distribution $p(v)$ is defined as

$$k = \frac{\int dv\ p(v)(v - \overline{v})^4}{(\int dv\ p(v)(v - \overline{v})^2)^2} - 3$$

with $\overline{v} = \int dv\ p(v)v$, and it takes the value zero for a Gaussian distribution. Positive values of k are taken to imply sparse distributions, which are also called super-Gaussian or leptocurtotic. Distributions with $k < 0$ are called sub-Gaussian or platykurtotic. This is a slightly different definition of sparseness from being heavy-tailed. Several researches also suggest that the kurtosis should be avoided as a sparseness measure and recommend tanh-functions for measuring noisy sparseness.

We propose the following enhancement of the smoothness assumption of semi-supervised learning.
Sparseness assumption of semi-supervised learning:
Distributions of filter responses are heavy-tailed. The vector $(\gamma(F)(x_1), \dots, \gamma(F)(x_n))$ *should be sparse.*

This sparseness assumption leads to the following optimisation problem:

$$\begin{array}{l} \min \|(\gamma(F)(x_1), \dots, \gamma(F)(x_n))\|_0 \\ \text{with } \; F(x_i) = y_i, \text{for } i = 1, \dots, l. \end{array} \tag{4}$$

Very often the combinatorial problem (4) can be replaced by the simpler problem

$$\begin{array}{l} \min \|(\gamma(F)(x_1), \dots, \gamma(F)(x_n))\|_p, \;\; 0 < p \leq 1, \\ \text{with } \; F(x_i) = y_i, \text{for } i = 1, \dots, l. \end{array} \tag{5}$$

Let us show that the optimisation problem (5) is in fact the optimisation problem (2) with the linear subspace L in the second form. Denote by A the $n \times n$-matrix $E_n - P$, where E_n is the identity $n \times n$-matrix and P is the matrix of transition probabilities. Denote b the n-dimensional vector $A \cdot (y_1, \dots, y_l, 0, \dots, 0)^t$ and the linear subspace L to be the image of the $(n - l)$-dimensional space of vectors with first l coordinates zero, i.e. $(0, \dots, 0, z_1, \dots, z_{n-l})$, under linear transformation A. Then it is easy to see that (5) becomes (2) with such b and L.

Consider now our example of the colourisation problem. To simplify the model, we assume that $\gamma(F)(x_i), \; i = 1, \dots, n$, are *i.i.d.* random variables, i.e.

$$P_G(F) \propto \prod_{i=1}^{n} p(\gamma(F)(x_i)).$$

In ([11]) it was shown that for natural colour images the distribution $p(\gamma(F)(x_i))$ can be modelled by a Generalised Gaussian Distribution (GGD)

$$\frac{1}{Z} e^{-|\gamma(F)(x_i)/s|^{\alpha}},$$

where Z is a normalising constant, s the scale parameter and α the shape parameter. The GGD gives a Gaussian or Laplacian distribution when $\alpha = 2$ or 1, respectively. When $\alpha < 1$ we have a heavy-tailed distribution. Figure 1 shows a typical example (with the vertical axis on a log scale) of the histogram of pixel intensity that is observed after filtering a natural colour image. Fitted to the data is a GGD. We also overlay on the figure the classical parabola shaped Gaussian distribution which clearly shows the difference in the tails between the two. Such differences highlight the importance of choosing the correct distribution as a prior knowledge when using Bayesian analysis.

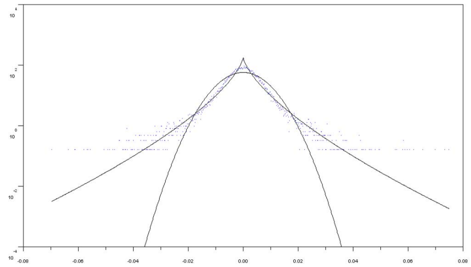

Fig. 1.

Taking $\log(P_G(F))$, the problem (3) leads to an equivalent minimisation objective ($p = \alpha < 1$) (5).

Figure 2 show an example of a colourised image based on sparse prior. In ([12]) it was shown that l^1-optimisation in many cases out-perform the case $\alpha = 2$.

Similar to the cases of natural images and modelling of human behaviour, we would like to finish this section with the following
Conjecture: *Many learning functions from real life have heavy-tailed distributions of filter responses.*

III. Compressibility of the learning function and sparse recovery

In this section we look at sparsity of the learning function from the point of view of compressibility. It is very natural to assume that learning functions should be very special, and that they depend only on a small number of parameters. There are two reasons for this: 1) the learning function ideally should learn from its values on the labelled set X_l, i.e. depend on a maximum of l parameters; 2) any supervisor uses a small number of rules and heuristics. This means that the class of learning functions should be highly compressible.

The smoothness assumption of semi-supervised learning suggests that the learning function should be well approximated by a few "slowly variable" functions. But we can't completely forget about higher frequencies since we would like also to be able to have some jumps between clusters. This suggest that we should be able to represent the learning function by a small number of special functions on the graph G. Such classes of special functions can be, for example, eigenfunctions of the matrix of transition probabilities P or graph Laplacians. We can also have more special functions than the size of the data set X by restricting to the graph G eigenfunctions of a larger graph with more unlabelled points.

The spectral theory on graphs plays an important role in many areas of data analysis: diffusion maps, dimension reductions, feature extractions etc. For example, in [13] it was shown that the first few eigenvectors and eigenvalues of the matrix P of transition probabilities of the graph G contain useful geometric information and that the diffusion

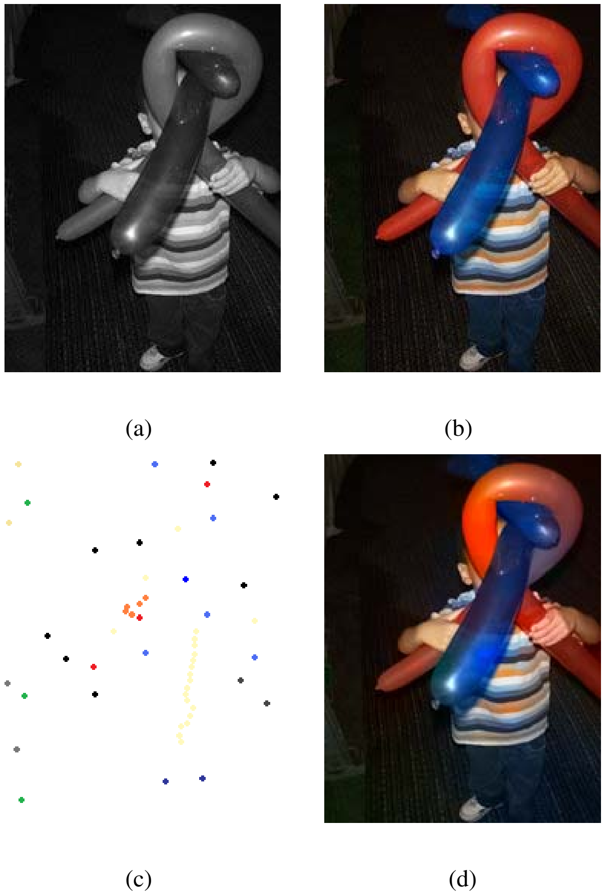

Fig. 2. Colourisation example. (a) An example gray image; (b) the original colour image for reference; (c) a set of coloured pixels arbitrary places; (d) shows colourisation using l^1 optimisation.

map converts the diffusion distance into Euclidean distance. In [14] multiresolution analysis for data was also developed.

There are different ways of defining the graph Laplacian, the two most prominent of which are the normalised graph Laplasian, $\mathcal{L}$, and unnormalised graph Laplacian, L:

$$\begin{aligned} &\mathcal{L} := E_n - D^{-1/2} W D^{-1/2}, \\ &L := D - W. \end{aligned}$$

Now, if we have some set of special function $\{f_1, \ldots, f_N\}$ $(N \geq n)$ on the graph G, we would like to define the learning function F as a function with $F(x_i) = y_i, \quad i = 1, \ldots, l$, and with the smallest number of nonzero coefficients $a_1, \ldots, a_N$ in the decompositions $F = \sum_{i=1}^N a_i f_i$, i.e.

$$\begin{aligned} &\min ||(a_1, \ldots, a_N)||_0 \\ &\text{with } \ F(x_i) = y_i, \text{for } i = 1, \ldots, l, \\ &\text{and } F = \textstyle\sum_{i=1}^N a_i f_i. \end{aligned} \tag{6}$$

We can justify the optimisation problem (6) by Bayesian analysis, similar to the Sec.II. We model $P_G(F)$ by marginal probability distribution of the coefficients $a_1, \ldots, a_N$. If $\{f_1, \ldots, f_N\}$ is an orthonormal basis, then $a_i =< F, f_i >$, the scalar product of F and f_i. To simplify the model, we assume that a_i are *i.i.d.* random variables, i.e. $P_G(F) \propto \prod_{i=1}^n p(a_i)$. Similar to wavelet coefficients in signal processing, the distributions $p(a_i)$ in many examples are heavy-tailed. For example, in the colourisation problem and f_i being eigenvectors of the normalised Laplacian $\mathcal{L}$, the distributions $p(a_i)$ can be also modelled by a GGD. So, similar to the Sec.II, we can justify using the problem (6) as an optimisation problem for finding F.

We introduce now
Compressibility assumption of semi-supervised learning:
Distributions of the coefficients $a_1, \ldots, a_N$ are heavy-tailed. The vector $(a_1, \ldots, a_N)$ should be sparse.

We know the learning function F only on the set X_l. So, we can write the known part of the decomposition $F = \sum_{i=1}^N a_i f_i$ as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} = \sum_{i=1}^N a_i \begin{pmatrix} f_i(x_1) \\ \vdots \\ f_i(x_l) \end{pmatrix} \tag{7}$$

with $l << N$. The problem (6) now become an l^0-problem of the type (2). More precisely, if $\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_N)^t$ is any fixed solution of (7), then the general solution of (7) is of the form $\tilde{a} - z$, where $z \in L = Ker A$ with

$$A = \begin{pmatrix} f_1(x_1), \ldots, f_N(x_1) \\ \vdots \qquad\qquad\qquad \vdots \\ f_1(x_l), \ldots, f_N(x_l) \end{pmatrix},$$

and we need to minimise $||\tilde{a} - z||_0$ over $z \in L$.

Now the strategy of finding the learning function is as follow:

1) Replace the l^0-problem by the l^1-problem, i.e.

$$\begin{aligned} &\min ||(a_1, \ldots, a_N)||_1 \\ &\text{with } \begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} = \sum_{i=1}^N a_i \begin{pmatrix} f_i(x_1) \\ \vdots \\ f_i(x_l) \end{pmatrix}. \end{aligned}$$

2) After finding such l^1-minimiser $a_1^0, \ldots, a_N^0$, the learning function F is

$$F = \sum_{i=1}^N a_i^0 f_i.$$

We would like to finish this section with the following
Conjecture: Many learning functions from real life have heavy-tailed distributions of the coefficients $a_1, \ldots, a_N$.

IV. Conclusions and Future Directions

Semi-supervised learning and compressive sampling both have a very similar goal: to recover a signal from a small number of measurements. Graph-based methods in semi-supervised learning allow us to construct two sets of filters. The first set of filters measure local fluctuations from being harmonic, and the second set of filters measure coefficients in spectral decompositions. If such filter responses have heavy-tailed (sparse) distributions then the problem of finding learning functions becomes an optimisation problem of finding sparse representations. Sparse distributions are defined as being more likely than Gaussian of the same mean and variance to generate values near zero and also more likely to generate values far from zero. These occasional high values can convey substantial information. In many cases we can replace these sparsity problems by tractable l^1-optimisation problems from the theory of compressive sampling. As an example, we have looked into the colourisation problem.

We have proposed several conjectures and assumptions in semi-supervised learning. We are planning to check these sparsity and compressibility conjectures for different known classes of learning function in supervised learning. It is also very interesting to apply these sparsity/compressibility assumptions to other problems of semi-supervised learning. We should also to use iteratively re-weighted least squares minimisation methods directly to the problem 2 with $p < 1$ as in [15].

References

[1] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[2] E. Candès, "Compressive sampling," in *Int. Congress of Mathematics, 3*, Madrid, Spain, 2006, pp. 1433–1452.

[3] J. Fan and R. Li, "Statistical challenges with high dimensionality: feature selection in knowledge discovery," in *Int. Congress of Mathematics, 3*, Madrid, Spain, 2006, pp. 596–622.

[4] B.K. Natarajan, "Sparse approximate solution to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[5] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Transaction on Graphics*, vol. 23, no. 3, pp. 689–694, 2004.

[6] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, Fort Collins*, 1999, pp. 541–547.

[7] S. I. Resnick, *Heavy-Tail Phenomena. Probabilistic and Statistical Modelling*, Springer Series in Operations Research and Financial Engineering. Springer, 2007.

[8] J. Beirlant, Y. Goegebeur, and J. Teugels, *Statistics of Extremes: Theory and Applications*, John Wiley & Sons Ltd, 2004.

[9] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*, Springer Series in Synergetics. Springer-Verlag, 2000.

[10] A. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, pp. 207–211, 2005.

[11] A. Balinsky and N. Mohammad, "Non-linear filter response distributions of natural colour images," *To appear in the Proc. of the 2009 Computational Color Imaging Workshop, Saint Etienne, Springer-Verlag Lecture Notes*.

[12] A. Balinsky and N. Mohammad, "Colorization of natural images via l^1 optimization," *Submitted*, 2009.

[13] R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7426–7431, May 2005.

[14] R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker, "Geometric diffusions as a tool for harmonics analysis and structure definition of data: Multiscale methods," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7432–7437, May 2005.

[15] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk, "Iteratively re-weighted least squares minimization for sparse recovery," *Preprint*, pp. 1–35, 2008.