Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Data Modeling: Visual Psychology Approach and $L_{1/2}$ Regularization Theory

Zongben Xu*

Abstract

Data modeling provides data analysis with models and methodologies. Its fundamental tasks are to find structures, rules and tendencies from a data set. The data modeling problems can be treated as cognition problems. Therefore, simulating cognition mechanism and principles can provide new subtle paradigm and can solve some basic problems in data modeling.

In pattern recognition, human eyes possess a singular aptitude to group objects and find important structure in an efficient way. I propose to solve a clustering and classification problem through capturing the structure (from micro to macro) of a data set from a dynamic process observed in adequate scale spaces. Three types of scale spaces are introduced, respectively based on the neural coding, the blurring effect of lateral retinal interconnections, the hierarchical feature extraction mechanism dominated by receptive field functions and the feature integration principle characterized by Gestalt law in psychology.

The use of L_1 regularization has now been widespread for latent variable analysis (particularly for sparsity problems). I suggest an alternative of such commonly used methodology by developing a new, more powerful approach – $L_{1/2}$ regularization theory. Some related open questions are raised in the end of the talk.

Mathematics Subject Classification (2000). 6IH30, 68T10, 62-07, 94A12.

Keywords. Data modeling, sparse signal recovery, visual psychology approach, L_1 regularization, $L_{1/2}$ regularization.

*The research was supported by National 973 Program (2007CB311002) and NSF Projects (60975036,60905003) of China.

Department of Mathematics & Institute for Information and System Sciences, Xi'an Jiaotong University, XI'an, 710049, P.R. China. E-mail: zbxu@mail.xjtu.edu.cn.

1. Introduction

We are in the era of knowledge economy. One of the main features is the rapid growing of the information technology which has become the most lucrative segment of the world economy, with much of the growth occurring in the development, management, and application of prodigious streams of data for scientific, medical, engineering, and commercial purposes. Responding to the rapid advances in information technology, data analysis has been developed at breakneck pace in the last few decades. It has now been a very significant, or even main part of science and engineering, as predicted by John Tukey [1] forty years ago.

The main purpose of data analysis is to help people to understand the meaning and value of the data. Initiated from statistics, data analysis has, however, strong connections with many other disciplines such as computer science, information processing and pattern recognition. It is inarguably accepted as a part of information technology today.

Data Modeling provides data analysis with models and methodologies. In other words, data modeling yields the data analysis techniques that have solid mathematical basis. Different from traditional mathematical modeling that aims to formulate a phenomenon, a principle or a system, data modeling models a data set. This is perhaps a basic form of applications of mathematics nowadays.

The fundamental tasks of data modeling are to find patterns, structures, rules, relations or tendencies from a data set, which serves then to explain which measurement(s) or attribute(s) is relevant to the phenomenon of interest, or what kind of structures or rules existed in a collection of data. The aims are provision of computational models which makes it possible that data can be automatically perceived and understood for decision. The basic problems of data modeling include clustering, classification, regression and latent variable analysis [2].

Clustering is a problem of partitioning a data set into subgroups based on similarity among data. It seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar. Very basic to knowledge discovery, the clustering is capable of finding new concepts, new phenomenon or new patterns of data. *Classification* is a problem of seeking a general discriminative rule (normally, a function) to categorize the data by their attributes. The sought discriminative function is then used in discriminative analysis, and therefore, laid the basis of any pattern recognition application. *Regression* aims to determine a quantitative cause/result relationship between variables in data, where M variables in the data are quantitative response variables, and the other N variables are used to predict it. This quantitative relationship is generally modeled as a continuous function (say, a polynomial or a neural net), and mainly used for prediction/forecasting application. *Latent variable analysis* attempts to identify the intrinsic variables from the observation, fundamental to vi

alization, feature extraction and motion modeling. In such a problem, we are given

$$\mathbf{y} = A\mathbf{x},\ \mathbf{x} \in R^N,\ \mathbf{y} \in R^M$$

$\mathbf{y}$ is a observation, $\mathbf{x}$ is a unobserved latent variables, and A is a linear transformation converting one into the other. The hope is that a few underlying latent variables are responsible for essentially the structure we see in the observation, and by uncovering those variables, we can achieve important insights. We easily see that the latent variable analysis problem can be reexpressed as a *sparsity problem* [3], as will be explained latter in section 3 of this talk.

All the above problems can be tackled within the frameworks of statistics and information science. A great number of useful and effective tools and techniques, for instance, have been developed from those methodologies. The k-means, Graph-based Clustering, Fisher Discriminant Analysis, Support Vector Machine, Neural Networks, Fuzzy Systems, Boosting, PCA, Manifold Learning are just a few of the popularly used techniques. Nevertheless, all those techniques face challenges when applied to real data sets we are meeting today and in future.

The challenges come mainly from several striking features of real data sets: (i) *massiveness*, say, think of the huge volumes of data automatically generated by a satellite; (ii) *high dimensionality*, say, think of the DNA microarrays for patients, where genes are huge, but relatively few patients with a given genetic disease; (iii)) *inhomogeneity*, say, think of a muti-medium data set which contains images, texts, media, and video in the same time; and (iv) *uncertainty*, say, think of hyperspectral imagery, internet portals, and financial tick-by-tick data, in which noise and inaccuracy are inevitably involved in gathering or measurement. All these features may make the existing techniques either infeasible or ineffective.

To be further, for example, the massiveness of a data set may cause ineffectiveness for any algorithms related to inversion of a matrix, which takes $\bigcirc(N^3)$ operations and for large N (say in the millions) is prohibitively expensive. The high dimensionality may lead to infeasibility and ineffectiveness of most techniques based on traditional statistical methodology. This is because, in traditional statistical methodology, we assumed many observations and a few, well-chosen variables (namely, $M \gg N$, the *large sample problem*). The data set today is, however, towards more observations but even more so, to radically larger number of variables. We are seeing examples where the observations gathered on individual instances are curves, or spectrums, or images, or even movies, so that a single observation has dimensions in the thousands or billions, while there are only tens or hundreds of instances available for study (thus, $M \ll N$, the *small sample problem*). Such high-dimension/small sample problems cannot be solved effectively by the large sample algorithms.

The challenges will get more serious if we take it into account that our purpose of data modeling, hopefully, is to provide computational models for automatical understanding of data (such type of models are referred to as

Machine Cognition Models). In other words, a machine cognition model provides a technique that can perform an automatical data analysis without any other assistance. From this sense, most of existing techniques are still far from the end.

It is unlikely to have all the problems being solved simultaneously. For some special and separate cases, however, some significant progresses can be made. In this talk, I review some of these progresses.

As the terminology "Pattern Recognition" implies, pattern recognition (essentially, a classification problem) could be accomplished by repeating the human cognition rules (that is, Re-cognition is the way to solve the problem). Through viewing a data modeling problem as a cognition problem, clustering, classification and regression problems can be tackled by mimicking visual psychology. Such visual psychology approach brings many benefit, defines machine cognition models of the problems, and provides satisfactory solutions to several long-standing problems in data analysis. We summarize the related works in the next section.

The way how our visual system encodes observation naturally motivates the methodology for solving latent variable analysis problem. Such an approach could be considered in a more general framework, sparsity problems — to find sparse solution(s) of a representation or an underdetermined equation. A common practice for solution of sparsity problems is L_1 *regularization*, formalized by Tibshirani [4] and Chen, Donoho, and Saunders [5]. The use of L_1 regularization has become so widespread that it could arguably be considered the "modern least squares" [6]. However, for many applications, the solutions of the L_1 regularization are often less sparse than that expected. As an alternative, $L_{1/2}$ regularization then has been developed in recent years by my group. I introduce such new methodology in section 3.

In section 4 I propose problems open to be answered along the line of research topics talked here.

2. Visual Psychology Approach

We begin with an observation that for most of the data modeling problems in low dimensions (say, $N = 1, 2$), the solutions of problems can always be promptly captured with our eyes. Why it is so is due to the unrivaled cognition ability of human being! The approach I will introduce in this section just follows this modus of human being to solve a data modeling problem.

Thus, my basic point of view is: *A data modeling problem is a cognition problem.* Although this is supported only with the low dimensional problems, we can solve the problem through modeling it in the way of human beings in low dimensions, and then generalizing it to the high dimensions through formalization plus mathematical justification.

Let us first explain how a data set can be transformed into an object that can be observed by our eyes. Naturally, such an object should be somewhat

an image, and we call it the *Data Image*. The data image is a real one only in low dimensions, but imaginary in high dimensional cases. Given a data set $D = \{z_i = (x_i, y_i)\}_{i=1}^{M}$ with $x_i \in R^N, y_i \in R^1$, the data image of data set D can be defined with its empirical distribution respective to the problems we are tackling. For example, for clustering problem, the data image can be defined as

$$g_D(z) = \frac{1}{M} \sum_{i=1}^{M} \delta(z - z_i) \tag{1}$$

For classification problem, it is then defined by

$$g_D(x) = \frac{1}{M_+ + M_-} \left(\sum_{i=1}^{M_+} \delta(x - x_i^+) - \sum_{i=1}^{M_-} \delta(x - x_i^-) \right) \tag{2}$$

where the classification problem is assumed to be canonical, that is, a two-class problem, and the data set is correspondingly splitted into two parts:

$$D = \{(x_i^+, +1)\}_{i=1}^{M_+} \bigcup \{(x_i^-, -1)\}_{i=1}^{M_-}.$$

Data images are very special images without color and continuous texture information. A data image, however, contains various macro-information like cluster structure, separation structure, tendency, dependence, all of those interested us. According to physics, any macro-structure must consist of micro-structures. The macro-structure of a data set thus can be observed only when various micro-structures of the data have been perceived. What types of micro-structures have been captured then when we observe a data image? The psychology experiments conducted by Santos and Marqures [7] suggested the following ingredients:

- ***Density Feature*** It is the distribution difference feature of data, which can be measured with the number of data in a certain volume of data space; A data set with uniform distribution is normally accepted as no feature because no visual difference is perceived.
- ***Connectedness Feature*** It is the feature of a date set in which some data look like the samplings on a curve or a manifold. When they are observed from appropriately far away, those data appear as continuous curves or manifolds.
- ***Orientation Feature*** A datum together with its surrounding data defines a subregion of data space. If the subregion has a distinct principle direction, the datum is said to have local orientation; If the local orientation of some data are almost same, those data are said to have a structure direction. Whenever there exists structure direction in the set, the data set is said to have orientation feature.

Those structures are reexpressed in [8] with computational models. We remark that the micro-structures of a data set is by no means accountable, and it actually depends on the *visual attention* and what type of *observation purpose* is taking. For example, when a discrimination task is taking, the separation extent (margin) and boundary may be also perceived, besides the features mentioned above.

The crucial problems are: How those structure features have been organized into a macro-structure, and how the macro-structures have in turn been captured by our human eyes? This is the key to any attempt of solving the data modeling problems in the same or similar ways as our human beings do. The complete solutions are clearly in brain science, cognition science, and perhaps whole sciences, are in future but still unknown today. However, in recent years, physiological discoveries and researches in computer-aided neuroanatomy, neurobiology, and psychology have advanced several quite accurate computational models of primary visual system, each modeling some parts of the human visual system at a particular level of details. By simulating those known facts and discoveries, it is possible to form data modeling techniques more or less like the human eyes. Taking clustering problem as an example, I introduce those progresses below.

2.1. Scale Space Based Approach. One of our common visual experiences is that how clearly we observe an object depends on the distance of our observation. This is the principle of blurring effect of lateral retinal interconnections in primary visual system. The scale space theory, which models the blurring effect by applying Gaussian filtering to a digital image, was introduced by Witkin [9] in 1983. Suppose $P(x)$ is the intensity distribution of one object in nature and $P(x,\sigma)$ is the intensity distribution of the projected image of the object on the retina, where σ is a scale, understood either as the distance between the object and eyes or as the curvature of crystalline lens. Then $P(x,\sigma)$ can be mathematically described by

$$\begin{cases} \frac{\partial P(x,\sigma,)}{\partial \sigma} = \Delta_x P(x,\sigma) \\ P(x,0) = P(x) \end{cases}, \tag{3}$$

the solution of which is explicitly given by

$$P(x,\sigma) = P(x) * g(x,\sigma) = \int g(x-y)P(y)dy \tag{4}$$

where '$*$' denotes the convolution operation and $g(x,\sigma)$ is the Gaussian function

$$g(x,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\|x\|^2/2\sigma^2}. \tag{5}$$

In this way, $P(x)$ has been embedded into a continuous family $P(x,\sigma)$ of gradually smoother versions of $P(x)$. The original image corresponds to the scale

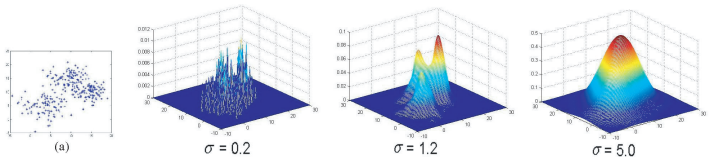

Figure 1. How evolves the data set (a) in scale space.

$\sigma = 0$, and, as the scale σ increases, $P(x, \sigma)$ gives a more and more blurring while simplified representation of $P(x)$ without creating spurious structure. Due to this, $P(x, \sigma)$ is referred to as *a multi-scale representation* of the image $P(x)$, and $\{P(x, \sigma)\}_{\sigma \geq 0}$ is a scale space. For any σ, $P(x, \sigma)$ is called *a scale space image*.

Interestingly, it can be shown that the above representation is unique if the retina property is assumed to be isotropic and spatially invariant. Without those assumptions, nevertheless, several other complicated PDE models, say, Anisotropic Diffusion Models, can be built. These models can not be directly applied to the approach introduced here.

Now, applying the scale space theory to the data image (1), we have the following multi-scale representation of data set D

$$P(x, \sigma) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} \tag{6}$$

which coincides with the Parzen distribution estimations of D with Gaussian window function. Figure 1 illustrates how a data set evolves in the scale space, i.e., what a multi-scale representation of a data set looks like.

As demonstrated in Figure 1, the data set appears as a data image with each datum being a light point attached with a uniform luminous flux. As we blur this image, each datum first becomes a light blob. Throughout the blurring process, smaller blobs merge into larger ones until the whole image contains only one light blob at a low enough level of resolution. In the process, small blobs always merge into large ones and new ones are never created. If we equate each blob with a cluster, the above blurring process seems providing a natural hierarchical clustering with resolution being the height of a dendrogram.

This is the point of our approach. That is, our idea is to capture the structure (from micro to macro) of a data set from the dynamic process observed in the scale space. This is a natural way to structure-finding, as inspired by the function of a lens in the visual system and our everyday visual experience.

However, to formalize this idea into a standard procedure of data clustering, three questions must be answered. *(i)* What means a cluster and how it can be formalized? *(ii)* How the continuous scale σ can be discretized so as not to affect our observation (say, not cause the loss of important structures)? and *(iii)* Does the blobs (clusters) evolve in an somewhat regular way? We answer those questions one by one below.

First, each blob can be defined as a cluster. So, for each fixed scale σ, we define a cluster (a light blob) as being the region in data set (corresponding to scale $\sigma = 0$) that satisfies

$$C_{y_\sigma} = \left\{ x_0 \in R^N : \lim_{t\to\infty} x(t, x_0) = y_\sigma \right\},$$

where $x(t, x_0)$ is the solution of gradient flow

$$\begin{cases} \frac{dx}{dt} = \nabla_x P(x, \sigma) \\ x(0) = x_0 \end{cases} \tag{7}$$

Here y_σ is a maxima of scale space image $P(x, \sigma)$, and referred as the blob center or cluster center of C_{y_σ}. Thus, at each scale σ, all blobs in $P(x, \sigma)$ produce a partition of data set D with each point belonging to a unique blob (cluster) except the boundary point. Each blob has its own survival range of scale, and larger blobs are made up of smaller blobs through the evolution. In consequence, a higher scale partition of D can be deduced from its lower scale partition, as long as the evolution of clusters is regular, leading to the third question in turn.

Second, we discretize the continuous scale σ according to the way of our human being. In psychophysics, Weber's law says that the minimal size of the difference ΔI in stimulus intensity which can be sensed is related to the magnitude of standard stimulus intensity I by $\Delta I = kI$, where k is a constant called Weber fraction. Coren [10] experimentally showed that $k = 0.029$ in one-dimensional observation. Consequently, we suggest the following discretization scheme for our observation:

$$\sigma_i - \sigma_{i-1} = k\sigma_{i-1}$$

where k is any constant not larger than Weber fraction. According to psychology, such a discretization scheme provides us a guarantee with which we cannot sense the difference between any two scale space images $P(x, \sigma_i)$ and $P(x, \sigma_{i-1})$.

The third question is essentially concerned with whether the cluster number, $\pi(\sigma)$, can be monotonically decreasing in the scale space. Define the cluster center curve $\Gamma = \{y_\sigma : \sigma \geq 0\}$. The following Theorem 2.1 justifies that Γ exactly consists of N simple curves, like Figure 2. So the monotonically deceasing of $\pi(\sigma)$ follows.

Theorem 2.1 ([11]). *For almost all data sets, we have: 1) zero is a regular value of $\nabla_x P(x, \sigma)$; 2) as $\sigma \to 0$, the clustering obtained for $P(x, \sigma)$ with $\sigma > 0$ induces a clustering at $\sigma = 0$ in which each datum is a cluster and the corresponding partition is a Voronoi tessellation, i.e., each point in the scale space belongs to its nearest-neighbor datum, and 3) as σ increases from $\sigma = 0$, there are N maximal curves in the scale space with each of them starting from a datum of the data set.*

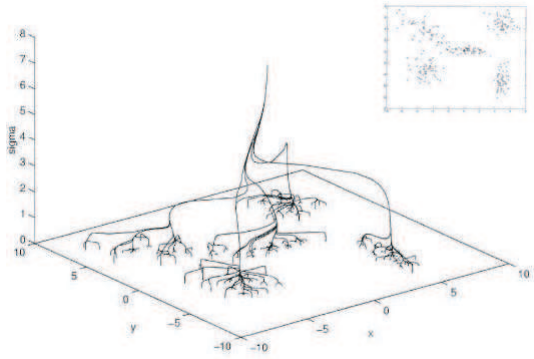

Figure 2. The cluster center curves defined by maxima of scale space data images.

Theorem 2.1 not only shows the simplicity of the cluster center curves that contains no forking, but also implies that for sufficiently small scale, the cluster center curves consist exactly of N branches with each datum being a cluster center. This shows that the deduced approach is independent of initialization. In addition, "zero is a regular value of $\nabla_x P(x, \sigma)$" implies the local uniqueness of stationary state of system (7), thus underlies the convergence of the gradient flow.

Based on the expositions above, a complete procedure, called Clustering by Scale Space Filtering (CSSF), for data clustering is developed. See [11] for the details.

The clustering approach made here has many exclusive advantages: Some readily observed advantages, for example, are: *(i)* The patterns of clustering are highly consistent with the perception of human eyes; *(ii)* The algorithms thus derived are computationally stable and insensitive to initialization; *(iii)* They are totally free from solving difficult global optimization problems; *(iv)* It allows cluster in a partition to be obtained at different scales, and more subtle clustering, such as the classification of land covers, can be obtained; and (v) The algorithms work equally well in small and large data sets with low and high dimensions.

The most promising advantage of the approach, however, is the provision of a cognitive answer to the long-standing problem of *cluster validity*. Cluster validity is a vexing but very important problem in cluster analysis because each clustering algorithm always finds clusters even if the data set is entirely random. While many cluster algorithms can be applied to a given problem, there is in general no guarantee that any two algorithms will produce consistent answers (so, it is why clustering has been regarded as a problem with a part art form and part scientific undertaking [2]).

What is a meaningful (real) cluster? The basis of human visual experience that the real cluster should be perceivable over a wide range of scales leads us to adopt the notion of *"lifetime"* of a cluster as its validity criterion: A cluster

with longer lifetime is preferred to a cluster with shorter lifetime; The cluster with longest lifetime in the scale space is the most meaningful or real cluster of a data set. We define the lifetime of a cluster and the lifetime of a clustering respectively as follows:

Definition 1. Lifetime of a cluster is the range of logarithmic scales over which the cluster survives, i.e., the logarithmic difference between the point when the cluster is formed and the point when the cluster is absorbed into or merged with other clusters.

Definition 2. Let $\pi(\sigma)$ be the number of clusters in a clustering achieved at a given scale σ. Suppose C_σ is a clustering obtained at σ with $\pi(\sigma) = m$. The σ-lifetime of C_σ is defined as the supremum of the logarithmic difference between two scales within which $\pi(\sigma) = m$.

The reasons why logarithmic scale is used was proven in [11] based on the experimental tests reported in [12], which experimentally justified that $\pi(\sigma)$ decays with scale σ according to $\pi(\sigma) = ce^{-\beta\sigma}$, where c is a constant and β is an unknown parameter.

See Figure 2, by Definitions 1 and 2, the data set D thus contains 5 real clusters, and the partitions of multi-scale representation of D at $\sigma = 1.5 \sim 2.5$ result in the most valid clustering, precisely consistent with the perception of the human eyes.

With the lifetime criterion for cluster validity, we can also answer some questions like whether or not there is a valid structure in a data set. The answer for example is: If $\pi(\sigma)$ takes a constant over a wide range of the scale, a valid structure exists, otherwise, no structure in the data. We can also apply the lifetime criterion to do outlier check. The deduced criterion, say, is that if C_i contains a small number of data and survives a long time, then C_i is an outlier, otherwise, it is a normal cluster.

The scale space based approach thus can provide us an automatic validity check and result in the final most valid clustering. It is also robust to noise in the data.

The scale space approach has provided a unified framework for scale-related clustering techniques derived recently from many other fields such as estimation theory, recurrent signal processing, statistical mechanics, and artificial neural networks. The approach has been extensively applied nowadays as a useful clustering analysis tool in science and engineering. Examples, e.g., see the series of works conducted in Laurence's lab on protein structure identification [13].

2.2. Receptive Field Function Based Approach. This is also a scale space approach, but, different from the last subsection where a continuous scale space is used. I introduce a discrete scale space approach in this subsection.

The continuous scale space approach provides a promising paradigm for clustering. However the high expense is obvious: The scale needs to be discretized and generation of partition at each fixed scale requires an iteration,

too. As a result, two theoretically infinite processes have to be executed in order that a clustering analysis task is accomplished. Moreover, the CSSF can be essentially understood as the Gaussian kernel density based clustering. It works perfectly for the data sets with Gaussian distribution, but not necessarily good (actually very bad sometimes) for non-Gaussian data sets. We hope to generalize the approach to cope with any complex data set, while within a discrete scale space framework.

Some more intrinsic visual mechanism and principles are thus needed. I summarize those preliminary knowledge ([14] [15] [16]) on Visual Information Processing (VIP) and Receptive Field Mechanism first in the following.

2.2.1. VIP and Receptive Field Mechanism. Visual system is a highly complex biological system, which is mainly composed of the retina, primary visual cortex and extra-striate visual cortex. As justified in physiology and anatomy, visual information is transmitted through a certain pathway layer by layer in visual system. Visual information are firstly captured by photoreceptor cells, and then received by ganglion cells. After this retina level, visual information will be transmitted through optic nerves to cross the lateral geniculate and finally reach the primary visual cortex. At the retina and primary visual cortex level, the main function of information processing is *Feature Extraction*. Then the visual information is transmitted into advanced visual cortex for *Feature Integration* or Concept Recognition.

VIP with large connected neurons is very complex, however, it can be easily described and simulated with electrophysiology. Many tests show that each neuron of a certain level corresponds to a spatial region of front layer, where neurons transform visual information to the neuron, and the region is called *Receptive Field* of the neuron (RF) [17] [18]. Each neuron has a certain response pattern (prototype) on the corresponding RF which is called *Receptive Field Function* (RFF). Physiological and biological tests reveal that the shapes of the RF are spatially variant in visual cortex. The RFs of ganglion cells are mainly concentric circle, while the RFs of neurons in visual cortex are more complex.

Given a stimulus $I(x)$, the response of a neuron in primary visual system can be measured by

$$\begin{aligned} f(x;\Theta) &= I(x) * R(x;\Theta) \\ &= \int I(y-x)R(y;\Theta)dy \end{aligned} \tag{8}$$

where $R(x;\Theta)$ is RFF of the neuron, and Θ is a set of parameters. In Eq.(8), $f(x;\Theta)$ is the response of the neuron with stimulus $I(x)$, which is the filtering response and called as a *feature* of $I(x)$.

Different features of a visual input can be extracted by different neurons at different layer. In terms of Eq.(8), this can be equivalently made by different RFFs. Some of the well recognized RFFs in visual system are Gaussian function [11], Gaussian derivative function [19], Gabor function [20], DoG (different

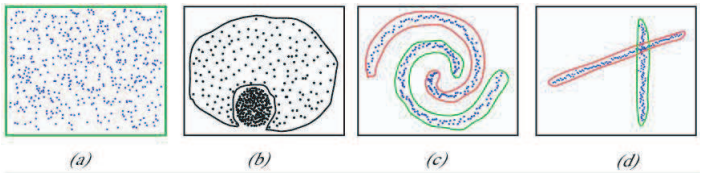

Figure 3. Structure features of data image: (a) No structure (uniform distributed); (b) Density feature; (c) Connectedness feature; (d) Orientation feature.

of Gaussian) function and 3-Gaussian functions. With these different RFFs, various features of visual input can be extracted. These extracted features can then be integrated to form a more complicated feature until a concept is identified.

There are various investigations into feature integration mechanism. However no general solution is rosolved up to now. Gestalt principle in psychology, nevertheless, summarizes some very fundamental perception rules of human being, which provides us useful guidance on how features can be organized. Gestalt principle summarizes the perception laws of how the objects (features) are perceived as a whole [21]. It says that human being tend to order our experience in a manner that is regular, orderly, symmetric, and simple. These are formalized respectively as the Law of Proximity, the Law of Continuity, the Law of Similarity, the Law of Closure and the Law of Symmetry. According to these laws, the objects with spatial or temporal proximity, with similar properties such as density, color, shape and texture, with connectedness and orientation features, with symmetric structure, are prone to be perceived as a whole. Our human being tends to group objects to an entity or a closure even it is actually not.

In this view, we can regard the VIP as a procedure of the hierarchical feature extraction dominated by RFFs and the feature integration characterized with Gestalt laws.

2.2.2. Receptive Field Function when Data Image Is Perceived. As the first step towards formalization of a more generic approach for scale space clustering, according to the VIP mechanism, we must first answer what type of RFFs should be adopted in the feature extraction process.

When a data set is observed, the receptive fields of neurons are adaptively formed. In other words, the RFFs are adaptive to the structure features of data image, particularly those of *Density Feature, Connectedness Feature* and *Orientation Feature,* as shown in Figure 3. Let χ be the data space. In [8], the following RFF was then suggested:

$$R(x;y,\Theta)=\min_{x\in\Gamma(y),y\in\Gamma(x)}\left\{\hat{R}(x;y,\Theta),\hat{R}(y;x,\Theta)\right\} \tag{9}$$

where $x, y \in \chi$ is any element,

$$\hat{R}(x; y, \Theta) = \exp\left(-\frac{1}{2} V(x, y; k) A(y; m) V^T(x, y; k)\right) \tag{10}$$

and $\Theta = \{m, k : m, k \; are \; integers\}$ is a parameter set that is used to confine the neighborhoods of a data set on which the data features are extracted.

In (10), $V(x, y; k)$ is a vector, called *manifold vector*, designed to model the connectedness features of the data image, defined by

$$V(x, x_j; k) = \begin{cases} \frac{x_j - x}{||x - x_j||} d_g(x, x_j; k) & x \neq x_j \\ 0 & x = x_j \end{cases}$$

where $d_g(x, y; k)$ is the geodesic metric between x and y, k is a neighborhood size parameter in computation of geodesic distance. It is clear that with such a definition, the manifold vector $V(x, y; k)$ is a vector from x to y with its norm being geodesic metric between x and y. The matrix $A(y; m)$ in (10), called *anisotropy matrix,* is designed to describe the orientation feature of the data set. Assume $\Gamma(x)$ is a chosen m-neighborhood of x, and $A(y; m)$ is then defined as $A(x; m) = B^{-1}(x; m)$ with $B(x; m)$ being the covariance matrix

$$B(x; m) = \frac{\sum_{x_i \in \Gamma(x)} (x - x_i)(x - x_i)^T}{|\Gamma(x)|}$$

where $|\Gamma(x)|$ denotes the number of data contained in $\Gamma(x)$.

It is immediate to see from (9) that the RFF so defined is a symmetric function. The symmetrization procedure in (9) was invented to characterize the density feature of the data set.

As suggested in real visual system, the RFF defined here is spatially localized, anisotropic and orientation selective. When $A(x; m) = I$ and $V(x, x_j; k) = x_j - x$, the RFF defined by (9)–(10) coincides with exactly the Gaussian function (5) used in CSSF.

2.2.3. Discrete Scale Space. With the RFF specified as in (9)–(10), according to VIP mechanism, a set of features of data image can then be extracted by formula (8). In effect, viewed as a data image, each datum of the data is a light point, which projects into χ at a certain location on retina. Suppose that each light point corresponds to a neuron (a photoreceptor cell) on retina photoreceptor level, and, for any $x \in \chi$, it most activates the neuron x' at the t-th layer of VIP. Then the receptive field, $\Gamma(x')$, of x' is a region of pattern space (or photoreceptor cell) which contains x, and RFF of x' is a function $R(x; x', \Theta)$ such that

(i) The nontrivial domain of R coincides with $\Gamma(x')$, i.e.,

$$\Gamma(x') = \{x \in \chi : R(x, x'; \Theta) \neq 0\}$$

(ii) The response of x' is given by

$$f(x;\Sigma) = X * R(x', x; \Theta) = \sum_{x_k \in \Gamma(x)} R(x_k - x; 0, \Theta) x_k \tag{11}$$

Let $X(t)$ be the feature of data set D extracted by VIP at t-th layer, and $X(0)$ simply corresponds to D. Then, $X(t)$ can be expressed as

$$X(t+1) = U(D)X(t),\ X(0) = D \tag{12}$$

with

$$U(D) = [u_{ij}]_{1 \leqslant i,j \leqslant N} = \left[\frac{R(x_i; x_j, k)}{\sum_{s=1}^{N} R(x_i; x_s, k)} \right]_{1 \leqslant i,j \leqslant N} \tag{13}$$

Here $X(t)$ and $X(0)$ are understood as $M \times N$ matrices.

The representation (12) defines a discrete scale space $\{X(t) = U(D)^t D : t \geq 0\}$. We call it the *discrete scale space* of data set D deduced from its feature. Correspondingly, it defines a multi-scale representation of data set D based on its features.

2.2.4. A Visual Clustering Framework (VClust). As in the continuous scale space case, a generic clustering procedure, called VClust in [8], can now be defined as follows:

$$\begin{cases} X(t+1) = U(D)X(t),\ X(0) = D; \quad t = 1, 2, ... \\ P_t(X) = G_1(\{X(t)\}_{t=0,1,\cdots,t}). \\ P(t) = G_2(\{P_t(X)\}) \end{cases}$$

where operator G_1 is the operation to get partition (clustering) of D at scale t, and G_2 is the operation to read the final most valid clustering of D. Both G_1 and G_2 can be defined completely similar to the case in CSSF.

It can be justified that VClust maintains all the promising properties of CSSF, while dismissing the two crucial drawbacks of CSSF: the high complexity and infeasibility to non-Gaussian data sets. Table 1 provides a direct support for this assertion. It further demonstrates the feasibility, effectiveness and robustness of VClust, as compared with some other competitive clustering techniques.

The data sets in Figure 4 are all with complicated structures (particularly, non-Gaussian). The algorithms used for comparison are all well developed, representatives of respective approaches. Besides CSSF, the Chameleon [22] is derived from the graph-based approach, the spectral-Ng [23] and spectral-shi [24] are spectrum-based, the DBSCan [25] and the Gaussian Blurring Mean Shift (GBMS) are density-based. The latest LEGClust algorithm [26] based on the information entropy is also tested. In comparsion, NMI, the normalized mutual information, was taken as the criterion for measuring the performance of each algorithm.

Table 1. Performance comparison of different clustering algorithms when applied to data sets in Figure 4, measured with NMI.

Methods \Data sets	(a)	(b)	(c)	(d)	(e)	(f)
VClust	1.0	1.0	1.0	1.0	1.0	1.0
CSSF	0.4357	0.7682	0.4732	0.3269	0.2718	0.4966
Chameleon	1.0	0.9379	0.6824	1.0	0.5991	0.6425
Spectral-Ng	1.0	0.8326	1.0	0.4157	0.4921	0.7103
Spectral-Shi	0.8726	0.9721	1.0	0.7892	0.5283	0.6947
LEGClust	1.0	0.9846	0.4919	1.0	0.3721	1.0
DBScan	0.4115	0.7286	0.4351	0.3924	0.2362	0.4529

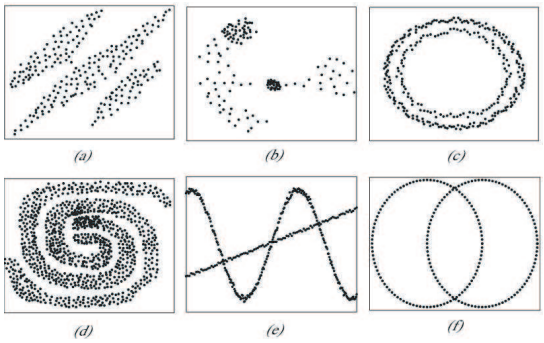

Figure 4. Some data sets with complicated structures used for comparison of different clustering techniques.

2.3. Neural Coding Based Approach. The scale space approach for clustering has been extended to classification problems. A similar idea was also used to do model selection for Gaussian Support Vector Machine, and in particular, a very useful data-driven formulae for Gaussian width parameter σ was discovered [27] (cf. Figure 5). Nevertheless, a much more significant extension of the scale space approach is the development of a new methodology: A neural coding based approach for data modeling.

In our brain, a neuron receives information from other neurons and processes/ responses through integrating information from other neurons, then sends the integrated information to others. We can generally classify the neurons into two types: the *stimulative neurons* (understood as the photoreceptor cells in visual system), which stimulate other neurons, and the *active neurons,* which receive information from stimulative neurons and produce response. Let $X = \{X_i\}_{i=1}^{M}$ be stimulative neurons and $Y = \{Y_j\}_{j=1}^{N}$ the active neurons, where X_i is a canonical stimulus, and Y_j is the receptive field function of neuron j that characterizes its response property. Let $e_j(X_i)$ denote the activation extent of active neuron j when the stimulative neuron i is stimulated, and let

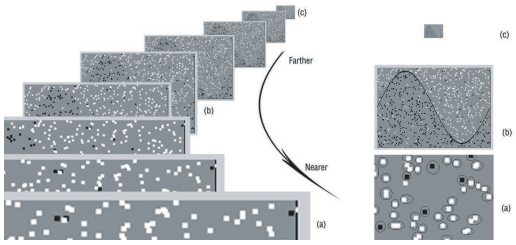

Figure 5. When a data set is observed from different distances, different structures are perceived.

$S(X_i, Y_j)$ denote the matching degree, or say, the similarity between the stimulus X_i and RFF Y_j. Then there holds a very fundamental coding principle: For any stimulative input, we response always maximally. That is to say, the neural coding in brain system is always such that for every input X, it maximizes the following response function

$$E(Y) = \left\{ \sum_{i,j} e_j(X_i) S(X_i, Y_j) \right\} \tag{14}$$

In preliminary visual system, neural coding is basically linear. Thus, let $f(X) = (f_1(X_1), f_2(X_2), ..., f_M(X_M))^\intercal$ be a stimulation mode, and $R(Y_j, X; \Theta)$ be the RFF of neuron j. Then we have [16] [28]

$$S(X_i, Y_j) = |f_j(X; \Theta)|$$

and

$$e_j(X_i) = \begin{cases} \frac{f_j(X;\Theta)}{|f_j(X;\Theta)|}; & if\ X_i \in \Gamma(Y_j) \\ 0; & otherwise \end{cases} \tag{15}$$

where $\Gamma(Y_j)$ is the receptive field of neuron j and $f_j(X; \Theta)$ is the response of Y_j given by

$$f_j(X; \Theta) = f(X) * R(Y_j, X; \Theta) = \sum_{x_k \in \Gamma(Y_j)} R(Y_j - X_k; 0, \Theta) f_k(X_k).$$

In this case, the response function (14) becomes

$$E_\Theta(Y) = \sum_{i,j} e_j(X_i) S(X_i, X_j) = \sum_j f(X) * R(Y_j, X; \Theta).$$

If one takes the parameter Θ be σ, then $\{E_\sigma(Y) : \sigma \geq 0\}$ gives the continuous scale space, and maximization of the response function directly leads to CSSF.

We naturally consider the nonlinear coding case. Different from linear case, nonlinear neural coding theory [29] [30] views the relationship between stimulative neurons and active neurons nonlinear. The theory says that the response

of a neuron is accomplished in two stages. In the first stage, as linear case, it integrates all stimuli from input cells, according to linear coding

$$U_{ij}^{(1)} = f(X) * R(Y_j, X; \Theta) \tag{16}$$

and in the second stage, it goes to two successive independent nonlinear procedures: within-pathway-nonlinearity and the divisive gain control nonlinearity,

$$e_j(X_i) = \frac{\left[U_{ij}^{(1)}\right]^p}{\left[C_2^p + \sum\limits_k U_{ik}^{(1)}\right]^p} \times \frac{\left[U_{ij}^{(1)}\right]^r}{\left[C_1^r + U_{ij}^{(1)}\right]^r} \tag{17}$$

where C_1 and C_2 are semi-saturation constants; r, p are the normalization parameters, controlling the degree of increasing response to the most sensitive stimulus, and decreasing the effect of insensitive stimulus.

With a neural coding scheme, a data modeling problem can be tackled in the subsequent way: Let $X = \{X_i\}_{i=1}^N$ be the data set, and $Y = \{Y_j\}_{j=1}^M$ be the solution we would like to find. We model the data modeling problem as an optimization problem

$$\max_Y \left\{ E(Y) = \sum_{i,j} e_j(X_i) S(X_i, Y_j) \right\} \tag{18}$$

through defining an appropriate similarity measure $S(X_i, Y_j)$, where $e_j(X_i)$ is any specified neural coding.

Examples are as follows:

Let $X = \{X_i\}_{i=1}^N$ be a data set with M clusters. Y_j is centroid of j-th cluster; d_{kj} is distance between X_k and the centroid Y_j of the j-th cluster; $g(\frac{1}{d_{kj}})$ is similarity between X_k and the centroid Y_j of the j-th cluster, and $g(\cdot)$ is any an increasing function. Then, (18) degenerates to CSSF when $e_j(X_i)$ is taken as the linear neural coding.

The Improved Probabilistic C-Means [31] provides an example with the nonlinear coding, where $S(X_i, Y_j) = 1/d_{kj}$. The technique improves substantially on Fuzzy C-means, noise clustering, and possible C-means. A comparison between PCM and its neural coding based counterpart is shown in Figure 6.

I suggest a methodology for solving a generic regression problem in section 4.

3. $L_{1/2}$ Regularization Theory

Latent variable analysis aims to identity the intrinsic variables from observation, while *Neural Coding* in neurobiology is concerned with how sensory and other information is represented in the brain by neurons. The aims of these two seemingly irrelevant subjects coincide with each other. So, borrowing the

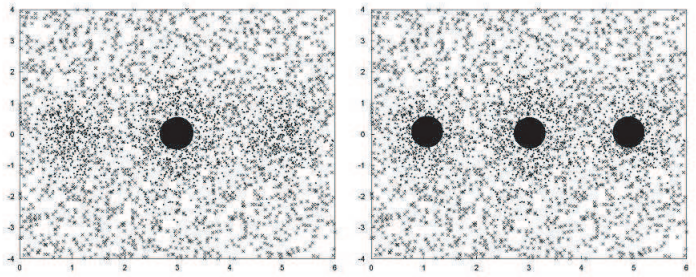

Figure 6. Comparison of clustering results with PCM and its neural coding based revision, where · is data point, • is cluster center and × denotes noisy data point.

methodology from neural coding can shed light on the way we solve a latent variable analysis problem.

The most striking feature of neural coding is its sparsity, which means that only a relatively small set of neurons in brain have strong response when a stimulus is received. Substantial biological evidence for such property occurs at earlier stages of processing across a variety of organisms, for example, auditory system of rats, visual system of primates and layer 6 in the motor cortex of rabbits [32]. Olshausen and Field [33] developed a mathematical model of sparse coding of natural image in visual system. Validated by neurobiological experiments, the receptive fields of simple cells in mammalian primary visual cortex are characterized as being spatially localized, oriented and bandpass. They demonstrated that such receptive fields emerge in their model when only the two global objectives are placed on a linear coding of natural images. In this case, the information of natural image is preserved, and the representation is sparse. Their model reads as

$$\min \left\{ ||I - Bx||_2^2 + \lambda p(x) \right\} \tag{19}$$

where I denotes the grey scale value of an image patch, B denotes the basis matrix consisted of the simple-cell receptive fields that are learned from samples, x is the sparse representation of natural image, and $p(x)$ is the sparse-promoting function which could be chosen as $-e^{-x^2}$, $log(1+x^2)$ or $|x|_1$. The research conducted by Olshausen and Field is important. It shows not only that the neural coding in primary visual processing (mainly with simple cells) does be sparse and can be linear, but also that the visual sparse coding can be simulated and found via a mathematical model. Such study has been generalized to complex cells in [34]. We observe that the model (19) is nothing else but a regularization scheme for solution of general sparsity problems.

Mathematically, a sparsity problem can be described as a problem of finding sparse solution(s) of an representation or a underdetermined equation. Besides the neural coding problem introduced above, variable selection, graphical modeling, error correction, matrix completion and compressed sensing (particularly,

signal recovery and image reconstruction) are all the typical examples. All these problems can be described as the following:

> Given a $M \times N$ matrix A and a procedure of generating observation y such that $y = Ax$, we are asked to recover x from observation y such that x is of the sparsest structure (that is, x has the fewest nonzero components).

The problem then can be modeled as the following L_0 optimization problem

$$\min \|x\|_0 \ \textit{subject to}\ y = Ax \tag{20}$$

where (and henceforth) $\|x\|_0$, formally called L_0 norm, is the number of nonzero components of x. Obviously, when $M \ll N$ (namely, the high dimension/small sample case), the sparsity problems are seriously ill-posed and may have multiple solutions. A common practice is then to apply regularization technique for the solution(s). Thus, the sparsity problems can be frequently transformed into the following so called L_0 regularization problem

$$\min_{x \in R^N} \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_0 \right\} \tag{21}$$

where $x = (x_1, \cdots, x_N)^T \in R^N$ and $\lambda > 0$ is a regularization parameter.

The L_0 regularization can be understood as a penalized least squares with penalty $\|x\|_0$, in which parameter λ functions as balancing the two objective terms. The complexity of the model is proportional with the number of variables, and solving the model generally is intractable, particularly when N is large (It is a NP-hard problem, see, e.g., [35]). In order to overcome such difficulty, many researchers have suggested to relax L_0 regularization and instead, to consider the following L_1 regularization

$$\min_{x \in R^N} \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_1 \right\} \tag{22}$$

where $\|x\|_1$ is the L_1 norm of R^N.

The use of the L_1 norm as a sparsity-promoting function appeared early in 1970's. Taylor, Banks and McCoy [36] proposed the use of L_1 norm to deconvolve seismic traces by improving on earlier ideas of Claerbout and Muir [37]. This idea was latter refined to better handle observation noise [38], and the sparsity-promoting nature of L_1 regularization was empirically confirmed. Rigorous uses of (22) began to appear in the late-1980's, with Donoho and Stark [39] and Donoho and Logan [40] quantifying the ability to recover sparse reflectivity functions. The application areas of L_1 regularization began to broaden in the mid-1990's, as the LASSO algorithm [4] was proposed as a method in statistics for sparse model selection, Basis Pursuit [5] was proposed in computational harmonic analysis for extracting a sparse signal representation from highly overcomplete dictionaries, and a technique known as total variation minimization was proposed in image processing [41, 42].

The L_1 regularization has now become so widespread that it could arguably be considered the "modern least squares" [6]. This is promoted not only by the sparsity-promoting nature of L_1 norm and the existence of very fast algorithms for solution of the problem, but also by the fact that there are conditions guaranteeing a formal equivalence between the combinatorial problem (21) and its relaxation (22)[43].

The L_1 regularization is, however, still far from satisfication. For many applications, the solutions of the L_1 regularization are less sparse than those of the L_0 regularization. It can not handle the collinearity problem, and may yield inconsistent selections [44] when applied to variable selection; It often introduces extra bias in estimation [45], and can not recover a signal or image with the least measurements when applied to compressed sensing. Thus, a mandatory and crucial question arises: Can the sparsity problems be solved by some other means? As shown below, I suggest the use of following alternative: the $L_{1/2}$ regularization

$$\min_{x\in R^N}\left\{\|y-Ax\|_2^2+\lambda\|x\|_{1/2}^{1/2}\right\}. \tag{23}$$

3.1. Why $L_{1/2}$ Regularization? We may seek other sparsity-promoting functions $p(x)$ to replace $\|x\|_1$ in (22). The generality of polynomial functions then naturally leads us to try $p(x)=\|x\|_q^q$ with $q\geq 0$. The geometry of Banach space implies, as suggested also by the classical least squares, $q>1$ may not lead to the sparsity-promoting property of functions $p(x)$. So $q\in(0,1]$ are only candidates. In consequence, the L_q regularizations have been suggested [46], that is, instead of L_1 regularization (22), using

$$\min_{x\in R^N}\left\{\|y-Ax\|_2^2+\lambda\|x\|_q^q\right\} \tag{24}$$

where $\|x\|_q$ is the L_q quasi-norm of R^N, defined by$\|x\|_q=\left(\sum_{i=1}^N|x_i|^q\right)^{1/q}$.

The problem is which q is the best? By using the phase diagram tool introduced by Donoho and his collaborators [47, 48], Wang, Guo and Xu [49] provided an affirmative answer to the question. Through applying the L_q regularizations to the typical sparsity problems of variable selection, error correction and compressed sensing with the reweighted L_1 technique suggested in [46], they experimentally showed that the L_q regularizations can assuredly generate more sparse solutions than L_1 regularization does for any $q\in(0,1)$, and, while so, the index $1/2$ somehow plays a representative role: Whenever $q\in[1/2,1)$, the smaller q, the sparser the solutions yielded by L_q regularizations, and, whenever $q\in(0,1/2]$, the performances of L_q regularizations have no significant difference (cf. Figures 7 and 8). From this study, the special importance of $L_{1/2}$ regularization is highlighted.

Figure 7 shows how sparsity (k/M, k is the number of nonzeros in x, and M is number of rows in A) and indeterminacy (M/N) affect the success of L_q regularizations. The contours indicate the success rates for each combination

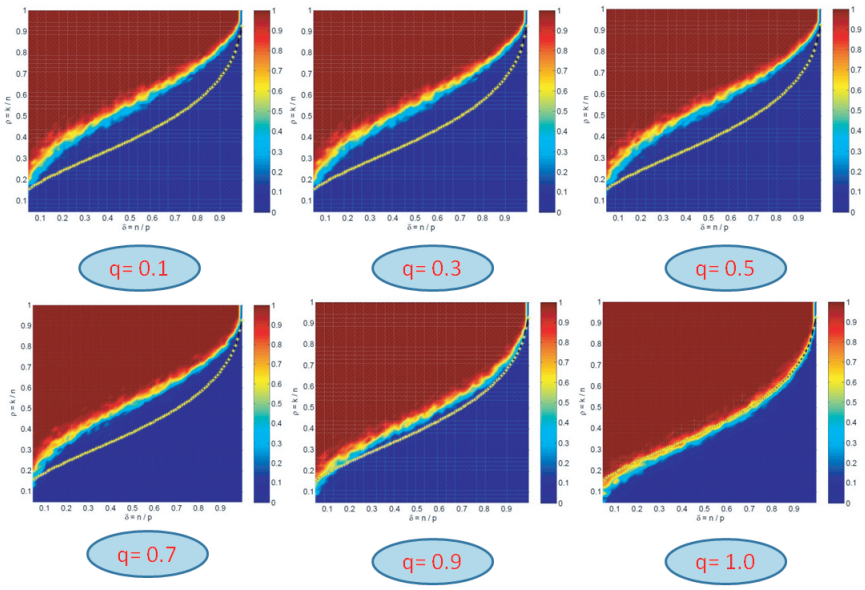

Figure 7. Phase diagrams of L_q $(q = 0.1, 0.3, 0.5, 0.7, 1.0)$ when applied to a sparsity problem (signal recovery).

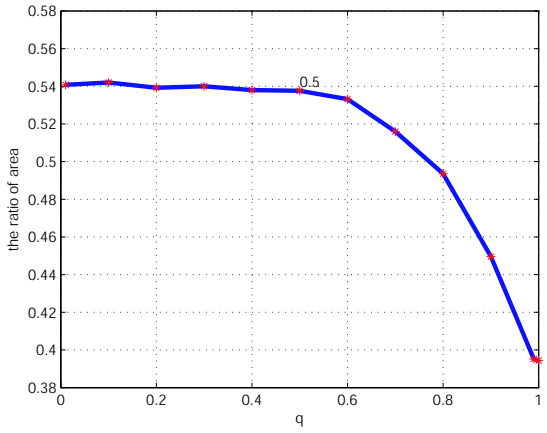

Figure 8. The interpolated success percentage curve of L_q regularizations, when applied to signal recovery.

of $\{k, M, N\}$, where red means the 0% success, blue means 100% success, the belt area means others. In the figure, the commonly occurred yellow curves are *Theoretical* L_1/L_0 *Equivalence Threshold Curve* found by Donoho [47, 48], which consists of the values at which equivalence of the solutions to the L_1 and

L_0 regularizations breaks down. The curve delineates a phase transition from the lower region where the equivalence holds, to the upper region, where the equivalence does not hold. Along the x-axis the level of underdeterminedness decreases, and along the y-axis the level of sparsity of the underlying model increases. The belt area in each case roughly defines a curve, which can be referred to as *A L_q/L_0 Equivalence Threshold Curve.* Then, Figure 7 exhibits that the L_q/L_0 equivalence threshold curves are always above of the theoretical L_1/L_0 equivalence threshold curve, showing a preferable sparsity-promoting nature of L_q regularizations.

Figure 8 shows the interplolated success percentage curve of L_q regularizations. Here the success percentage for a regularization is defined as the ratio of the blue region in the whole region of the phase plane. It is very clearly demonstrated that the $L_{1/2}$ regularization is nearly best, and therefore, can be taken as a representative of L_q regularizations with all q in $(0, 1]$.

Another reason why $\mathrm{L}_{1/2}$ is selected is due to its privilege of permitting fast solution, as that for L_1 regularization.

3.2. How $\mathbf{L_{1/2}}$ Fast Solved? The increasing popularity of L_1 regularization comes mainly from the fact that the problem is convex and can be very fast solved. The $L_{1/2}$ regularization, however, is a nonconvex, non-smooth and non-Lipschitz optimization problem. There is no directly available fast algorithm for the solution. Fortunately, I and my PhD students recently found such a fast algorithm for $L_{1/2}$ regularization problem [50].

The found fast algorithm is an iterative method, called the iterative *half* thresholding algorithm or simply *half* algorithm, which reads as

$$x_{n+1} = H_{\lambda_n \mu_n, \frac{1}{2}}(x_n + \mu_n A^T(y - Ax_n)) \tag{25}$$

Here $H_{\lambda\mu,\frac{1}{2}}$ is a diagonally nonlinear, thresholding operator specified as in Theorem 3.1, μ_n are parameters to control convergence and λ_n are adaptive regularization parameters. The derivation of the algorithm is based on a fundamental property of $L_{1/2}$ regularization problem, the thresholding representation property, as defined and proved in [50].

Theorem 3.1 ([50]). *The $L_{1/2}$ regularization permits a thresholding representation, i.e., there is a thresholding function h such that any of its solution, x, can be represented as*

$$x = H(Bx) \tag{26}$$

where H is a thresholding operator deduced from h and B is a linear operator from R^N to R^N. More specifically, one can take in (26) that for any fixed λ, $\mu > 0$,

$$B(x) = B_\mu(x) = x + \mu A^T(y - Ax) \tag{27}$$

$$H(x) = H_{\lambda\mu,1/2} = (h_{\lambda\mu/2}(x_1), h_{\lambda\mu/2}(x_2), ..., h_{\lambda\mu/2}(x_N))^\intercal \tag{28}$$

where $h_{\lambda\mu/2}(t)$ *is defined by*

$$h_{\lambda\mu/2}(t) = \begin{cases} \frac{2}{3}t\left(1+\cos\left(\frac{2\pi}{3}-\frac{2\varphi_\lambda}{3}\right)\right), & |t| > \frac{3}{4}(\lambda\mu)^{\frac{2}{3}} \\ 0, & otherwise \end{cases}. \tag{29}$$

with

$$\varphi_{\lambda\mu} = \arccos\left(\frac{\lambda\mu}{8}\left(\frac{|t|}{3}\right)^{-\frac{3}{2}}\right). \tag{30}$$

With the thresholding representation (27)-(30), the iterative algorithm (25) then can be seen as the successive approximation for common fixed point of operators H and B. The diagonal nonlinearity of the thresholding operator $H_{\lambda\mu,1/2}$ makes it possible to implement the iteration (25) component-wisely. The high efficiency and fastness of the $half$ algorithm thus follows. The thresholding representation (27)-(30) also has other meaningful consequences, say, it can be applied to justify the finiteness of local minimizers of $L_{1/2}$ regularization problem. This is an unusual, very useful property of a nonconvex problem, which distinguishes the $L_{1/2}$ regularization strikingly from other optimization problems.

Theorem 3.1 can also be used to derive an alternative theorem on solutions of $L_{1/2}$ regularization problem. From the theorem, some almost optimal parameter setting strategies can then be suggested. For example, the following parameter-setting strategy in (25) has been recommended in [50]:

$$\mu_n = \frac{(1-\varepsilon)}{\|A\|^2}, \lambda_n = \frac{4}{3}\|A\|^2 \left|[B_{\mu_n}(x_n)]_k\right|^{3/2}$$

where ε is any small fixed positive constant.

The half algorithm has been applied to a wide rang of applications associated with signal recovery, image reconstruction, variable selection and matrix completion in [50]. The applications consistently support that the algorithm is a fast solver of $L_{1/2}$ regularization, comparable with and corresponding to the well known iterative $soft$ thresholding algorithm for L_1 regularization.

It is interesting to ask a question here: *Is there other index* q *in* $(0,1)$, *except* $1/2$, *which permits a thresholding representation for* L_q *regularization?* In [50], an observation was made to guess that only with $q = 1, 2/3, 1/2$, L_q regularization admits a (27)-(28) like representation. A general answer is still unknown.

3.3. What Theory Says? The following theorem justify the convergence of the iterative half thresholding algorithm.

Theorem 3.2 ([51]). *Assume* $\mu_n \in (0, \|A\|^{-2})$ *and* λ_n *is monotonically decreasing to a fixed* $\lambda \geq 0$. *Then the* $half$ *thresholding algorithm converges to a local minimum of* $L_{1/2}$ *regularization problem (23). Furthermore, if any* k

columns of A (denoted by A_k) are linear independent, and μ_n, $\lambda_n$satisfies

$$\mu_n < 1/s_{\min}(A_k^T A_k); \lambda_n = \frac{4}{3}\|A\|^2 |[B_{\mu_0}(x_n)]_k|^{3/2}$$

where $s_{\min}(A_k^T A_k)$ is the smallest eigenvalue of matrix $A_k^T A_k$, then the algorithm converges to a k-sparsity solution of the $L_{1/2}$ regularization.

For the proof of Theorem 3.2, we refer to [51]. The proof depends upon a very careful analysis on the thresholding operator H defined as in (28). In the considered case, H is deduced intrinsically from the resolvent of gradient of $\|.\|_{1/2}^{1/2}$. Unlike the L_1 regularization case, where $\|x\|$ is a convex function, so that $\partial(\|x\|)$ is maximal monotone and the resolvent operator $(I+\partial(\|x\|))^{-1}$ is nonexpansive. In the $L_{1/2}$ regularization case, however, $\|x\|_{1/2}^{1/2}$ is non-convex and non-Lipschitz, so that the resolvent operator $(I+\lambda\partial(\|\cdot\|_{1/2}^{1/2}))^{-1}$ is only restrainedly defined and is not nonexpansive.

When applying a nonconvex sparsity-promoting function as a penalty, a problem we commonly worry about is the local minimum problem: The algorithm might only converge to a local minimum. Sometimes, this becomes the reason why a nonconvex regularization scheme would not be adopted in practice. However, due to the finiteness of local minima of $L_{1/2}$ regularization problem, Theorem 3.2 provides a promise that it can find the global optimal solution provided we run the algorithm many times with uniformly distributed random initial values.

With application to latent variable analysis or compressed sensing, the independence condition in Theorem 3.2 can be very intuitively explained. In the later case, for example, we have $A = \Psi\Phi$ with Ψ being $M \times N$ sampling matrix and Φ a basis matrix. Theorem 3.2 then says that a k-sparsity signal x can be recovered from M measurements with $M \ll N$ only if the sampling Ψ is such that every k columns of $\Psi\Phi$ are independent. This is obviously reasonable, and in fact constitutes the basis how the sampling should be taken.

The next theorem shows the condition when $L_{1/2}/L_0$ equivalence occurs. Recall that a matrix is said to possess *Restricted Isometry Property* (RIP) if

$$(1-\delta_k)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta_k)\|x\|_2^2 \text{ whenever } \|x\|_0 \le k$$

The restricted isometry constant $\delta_k(A)$ is the smallest constant for which the RIP holds for all k-sparsity vector x.

Theorem 3.3 ([52]). *Any k-sparsity vector x can be exactly recovered via $L_{1/2}$ regularization if $\delta_{2k}(A) < 1/2$.*

Note that Candès and Tao showed the L_1/L_0 equivalence when $\delta_{3k}(A)+\delta_{4k}(A) < 2$ [53], and later Candès relaxed to $\delta_{2k}(A) < \sqrt{2}-1 \approx 0.414$ [54], Foucart and Lai [55] verified the L_q/L_0 equivalence under the condition

$\delta_{2k}(A) < 2(3-\sqrt{2})/7 \approx 0.4531$. Theorem 3.3 provided a looser $L_{1/2}/L_0$ condition $\delta_{2k}(A) < 0.5$.

It is interesting to compare the convergence condition in Theorem 3.2 with the $L_{1/2}/L_0$ equivalence condition $\delta_{2k}(A) < 1/2$ in Theorem 3.3. In effect, the condition "any k columns of A (denoted by A_k) are linear independent" in Theorem 3.2 can be reformulated as $\delta_k(A) < 1$, which is much looser than $\delta_{2k}(A) < 1/2$. This leads to a natural question: Whether Theorem 3.3 is still true when the condition $\delta_{2k}(A) < 1/2$ is relaxed to $\delta_k(A) < 1$. I guess this is the case. However the real answer is open.

Theorems 3.4 and 3.5 below summarize two important statistical properties of $L_{1/2}$ regularization. Consider the linear model

$$y = X^{\intercal}\beta + \varepsilon, E\varepsilon = 0, Cov(\varepsilon) = \sigma^2 I \tag{31}$$

where $y = (y_1, y_2, ..., y_M)^{\intercal}$ is an $M \times 1$ response vector, $X = (X_1, X_2, ..., X_M)$ $(X_i \in R^N)$ and $\beta = (\beta_1, \beta_2, ..., \beta_N)^{\intercal}$ is unknown target vector, ε is a random error and σ is a constant. For any $1 \le k \le N$, let β_k denote the k-sparsity vector of β, that is, the vector whose k components coincide with those of β whenever the corresponding components β_i are among the k largest ones in magnitude, and other $N-k$ components are zeros. Note that when $L_{1/2}$ regularization is applied to problem (31), its solution is given by

$$\widehat{\beta} = \arg\min_{\beta \in R^N} \left\{ \sum_{i=1}^{M} (\beta^{\intercal} X_i - y_i)^2 + \lambda \|x\|_{1/2}^{1/2} \right\}. \tag{32}$$

Theorem 3.4 ([56]). *Let β^* be any solution of (31) and $\widehat{\beta}$ any solution of (32). Then for any $a > 0$ and under some mild conditions, for any $\delta \in (0,1)$ with probability larger than $1-\delta$, there holds the following estimation*

$$\left\| \widehat{\beta} - \beta^* \right\|_2 \le \bigcirc(\lambda\sqrt{k} + \|\beta^* - \beta_k^*\|_2 + \|\beta^* - \beta_k^*\| / \sqrt{l}) \tag{33}$$

where l is any constant satisfying $k \le l \le (N-k)/2$, t is constant satisfying $0 < t \le C(k,l)$, $\lambda \ge \frac{8(2-t)}{t} \max\{\sqrt{C_0}, 1\} \left(a\sigma \sqrt{\frac{2}{M} \ln \frac{2N}{\delta}} \right)$, and β_k^* is the k-sparsity vector of β^*.

The estimation (33), which measures how well the solution yielded by $L_{1/2}$ regularization approximates the target solution, can be shown to be optimal in the sense of achieving an ideal bound. It reveals that even though the number of samples is much smaller than that of the dimension of parameters, the solutions of $L_{1/2}$ regularization can achieve a loss within logarithmic factor of the ideal mean squared error one would achieve with an oracle. This shows that $L_{1/2}$ regularization is good at tacking the high-dimension/small sample problems.

One of direct applications of model (31) is variable selection. Fan [57] has ever suggested a standard of measuring how well an algorithm performs varaible

selection via the model (31). That is the so called oracle property: An ideal varaible selection algorithm should automatically set the irrelevant variables to zero. The following Theorem 3.5 shows that $L_{1/2}$ regularization has the oracle property. Without loss of generality, we assume that the target vector $\beta^* = (\beta_1^{*\intercal}, \beta_2^{*\intercal})^\intercal$ with β_1^* having no zero component and $\beta_2^* = 0$.

Theorem 3.5 ([56]). *If $\lambda = \circ(M^{1/4})$, then the $L_{1/2}$ regularization possesses the following properties:*

(i) Consistency in variable selection: $lim_{M\to\infty} P(\widehat{\beta}_2 = 0\} = 1$;

(ii) Asymptotic normality: $\sqrt{M}((\widehat{\beta}_1 - \beta_1^*) \to_d N(0, \sigma^2 C)$.

Theorem 3.5 shows that $L_{1/2}$ regularization is an idea variable selection method.

3.4. How Useful? The $L_{1/2}$ regularization has been applied to solve various sparsity problems, and among them compressed sensing is a very typical example. The compressed sensing (CS) has been one of the hottest research topics in recent years. Different from the traditional Shannon/Nyquist theory, CS is a novel sampling paradigm that goes against the common wisdom in data acquisition. Given a sparse signal in a high dimensional space, one wishes to reconstruct the signal accurately from a number of linear measurements much less than its actual dimensionality. The problem can be modeled as the sparsity problem (20) with

$$A = \Psi\Phi \tag{34}$$

where Ψ is a $M \times N$ sampling matrix, Φ is a $N \times N$ basis matrix and A is called a sensing matrix. A very fundamental requirement here is $M \ll N$. Given fewer measurements $y = Ax = \Psi\Phi x$ on a signal, we then are asked to reconstruct the signal x from y.

Let us take MRI as a concrete example. In MRI, the scanner takes slices from two dimensional Fourier domain of an image [58]. In order to reduce scan time and the exposure of patients to electromagnetic radiation, it is desirable to take fewer measurements. In this case, we hope to exploit the sparsity of the image in the Fourier or wavelet domain for reconstructing the image from fewer measurements. In application, the measurements are normally accomplished via sampling the image in its Fourier specture domain. According to [59], when sampling in this way on L rays in the domain and taken a Gaussian sampling on each ray, the resultant sensing matrix is Gaussian random, satisfying the so called RIP condition ([60]) so that the image can be exactly reconstructed.

We experimented with the standard Shepp-Logan phantom, a 256×256 MRI image shown as in Figure 9(a). The *half* thresholding algorithm (*Half*) in [50] and the Reweighted L_1 method (*RL1*) in [61] for $L_{1/2}$ regularization were applied in comparison with L_1 regularization. In implementation of L_1 regularization, the well known L_1 magic algorithm (*L1magic*) [62] and the soft thresholding algorithm (*soft*) [63] were applied, while the hard thresholding

algorithm (*hard*) [64] was adopted to perform L_0 regularization. We ran the simulations by varying the measurements from $L = 70$ to 40. The simulations reveal that before $L = 60$, all the algorithms succeeded in exactly recovery. Nevertheless, when L reduced to under 55, the L_1 regularization algorithms failed, but $L_{1/2}$ algorithm still succeeded, as listed in Table 2. It is seen that when sampling are taken on 52 rays, the $half$ and *hard* algorithms both can recover the image, with $half$ having the highest precision. When we reduce the sampling rays to $L = 40$, the algorithms L_1 magic, $RL1$, $soft$ and *hard* are all perform very poor, while the $half$ algorithm reconstructed the image with a very high precision, which distinguishes the $half$ from other competitive algorithms very obviously. See Figure 9 and Table 2 for the reconstructed images.

Table 2. The image reconstruction results

L	Method	MSE	Time	L	Method	MSE	Time
	L1magic	fail	∞		L1magic	9.3458 (fail)	1008.72
	RL1	fail	∞		RL1	4.6881 (fail)	3650.24
40	soft	8.2469	882.0637	**52**	soft	0.9812 (fail)	1795.5
	hard	**15.3978**	1038.1		hard	7.98e-6	105.0087
	half	**5.30e-7**	2738.8		half	3.15e-7	181.6311

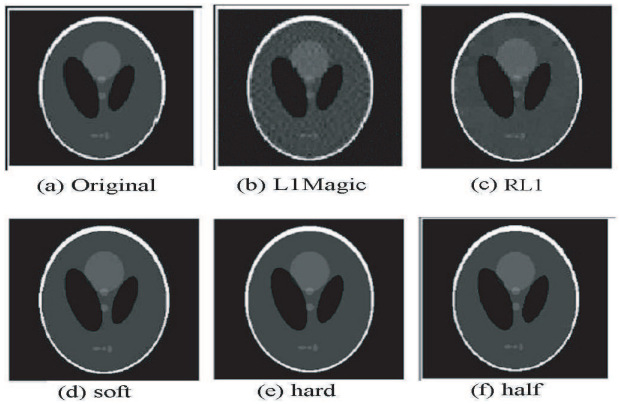

Figure 9. The reconstructed images by different regularization algorithms when $L = 52$.

This application demonstrates the outperformance of $L_{1/2}$ regularization over L_1 regularization. Such outperformance of $L_{1/2}$ regularization is also consistently supported by other experiments and applications.

Before ending this section, I would like to make an observation on overall features of L_p regularizations when p takes over entire real axis. The L_1 regularization is well known, that has the sparsity-promoting property and leads to a convex problem easy to be solved; When $p > 1$, the L_p regularizations

have not maintained the sparsity-promoting property any more, but possess a stronger convex property (uniformly convex property) and the problems get more and more easily solved; While when $p < 1$, the L_p regularizations have a stronger and stronger sparsity-promoting property, but have not maintained the convex property any more, and the problems get more and more difficult to be solved. This demonstrates a threshold or center position of $p = 1$ over which the sparsity-promoting property, the convex property and the easiness of solution all break down. In this sense, we can see that L_1 regularization just is the scheme with the weakest sparsity-promoting property and the weakest convex property (so, the weakest scheme), but more positively, it provides the best convex approximation to L_0 regularization and the best sparsity-promoting approximation to L_2 regularization. It is well known that all p with $1 \leq p \leq \infty$ constitute a complete system within which $p = 2$ plays a very special role. I therefore guess that $p = 1/2$ might somehow plays also a special role in another system $\{p : 0 \leq p \leq 1\}$. The study on $\mathrm{L}_{1/2}$ regularization is providing a direct support to this view.

4. Concluding Remarks

Data modeling is emerging as a cross-disciplinary, fast developing discipline. New ideas and new methodologies have been called for. In this talk I have introduced two new methodologies which seems meaningful and potentially important. Along the line of research in this talk, however, there are many problems open. As final remarks, I list some of those problems for further study.

Problem 1. Towards $L_{1/2}$ regularization theory

I first summarize the open questions I have raised in exposition of the last section. Firstly, *Does any other L_q regularizations rather than $q = 1/2$ permit a thresholding representation?* Following the idea in [50], it is not difficult to say "yes" for $q = 2/3$, but how about for other q in $(0, 1)$? The answer for this question is meaningful to development of other more effective sparsity-promoting algorithms. Secondly, we have shown the superiority and representative of $L_{1/2}$ regularization among L_q regularizations with $q \in (0, 1)$ based on a phase diagram study. This is certainly an experiment based approach. So, *Does the representative role of $L_{1/2}$ regularization can be justified in a somewhat theoretical way?* An tightly relevant question arises from an observation of phase diagrams in Figure 7. The belt area in each diagram roughly defines an empirical L_q/L_0 equivalence threshold curve, which fundamentally characterizes the sparsity-promoting capability of each corresponding regularization scheme. *Does there exist theoretical L_q/L_0 equivalence threshold curves for any L_q regularization? Are those L_q/L_0 equivalence threshold curves in Figure 3 the theoretical ones?* Thirdly, we have proved the convergence of the $L_{1/2}$ regularization algorithm (half thresholding algorithm) under the condition $\delta_k(A) < 1$, while justified the

$L_{1/2}/L_0$ equivalence under the much tighter condition $\delta_{2k}(A) < 1/2$. A natural question is: *Whether $\delta_k(A) < 1$ is also a sufficient condition for $L_{1/2}/L_0$ equivalence?*

Problem 2. Towards geometry of $L_{1/2}$ space

Let $\Gamma = \{p : 1 \leq p \leq \infty\}$. It is well known that with any $p \in \Gamma$, L_p space (understood either as function spaces or as sequence spaces) is a Banach space, and, within the duality framework $\frac{1}{p} + \frac{1}{q} = 1$, L_2 is self-dual and can be characterized with Parallelogram Law or equivalently Binomial Formula

$$\|x + y\|_2^2 = \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2, \forall x, y \in L_2$$

It is such characteristic identity law that makes many mathematical tools available, say, Fourier analysis and wavelet analysis. The Hilbert characteristic identity law was extended by Xu and Roach [65] into Banach space setting, which states that a Banach space X is uniformaly convex if and only if there is a positive function σ_p such that

$$\|x + y\|^p \geq \|x\|^p + p\langle J_p x, y\rangle + \sigma_p(x, y)\|y\|^p, \forall x, y \in X \tag{35}$$

and it is uniformaly smooth if and only if there is a positive function δ_p such that

$$\|x + y\|^p \leq \|x\|^p + p\langle J_p x, y\rangle + \delta_p(x, y)\|y\|^p, \forall x, y \in X \tag{36}$$

where J_p is the duality mapping with the guage t^p/p, σ_p is uniquely determined by the convexity modulus of X and δ_p uniquely determined by the smoonthness modulus of X. These Banach characteristic inequality laws admit two sets of explicit homogenous forms in L_p spaces with $1 < p < \infty$, since in this case, the spaces are both uniformal convex and uniformaly smoonth. A space with two or one of the two inequalities of the form (35) and (36) is very fundamental. In the case, many quantitative analysis and mathematical deductions then can be made in the space.

Let $\Sigma = \{p : 0 \leq p \leq 1\}$. It is then known that for any $p \in \Sigma$, L_p is not a Banach space, but is a quasi-normed space. Promoted by studying L_q regularization, I would like to know the geometry of quasi-normed spaces L_p with $p \in \Sigma$. More particularly, due to the speciality of $L_{1/2}$ regularization, I want to ask: *Does there exist a some kind of duality framework (say, $p+q=1$) such that within the framework $L_{1/2}$ space is self-dual?* Also, for studying $L_{1/2}$ regularization purpose, I would like to know: *Does there hold some kinds of characteristic laws like (35) and (36)?* If so, the convergence of $L_{1/2}$ algorithm and $L_{1/2}/L_0$ equivalence can be done in a straightforward way.

Problem 3. Towards a neural coding based machine learning theory

The neural coding based data modeling suggests also a new paradigm for solution of generic learning problem. In effect, assume X is a feature space,

Y is a response space and $Z = X \times Y$ is the data space. For a given set of training examples $D = \{z_i = (x_i, y_i)\}_{i=1}^M$ which are drawn $i.i.d$ from an unknown distribution P on Z, and a preset family of functions $\digamma = \{f : X \to Y\}$, a learning problem is asked to seek a function f^* in $\digamma$ such that the expected risk $E(f)$ is minimized, that is,

$$f^* = \arg\min_{f \in \digamma} E(f) = \int l(f, z) dP$$

The ERM principle suggests to use the empirical error $E_{emp}(f)$ to replace $E(f)$ and find f^* through

$$f^* = \arg\min_{f \in \digamma} \left\{ \frac{1}{M} \sum_{i=1}^{M} l(z_i, f) \right\} \tag{37}$$

while regularization principle is to solve the problem through

$$f^* = \arg\min_{f \in \digamma} \left\{ \frac{1}{M} \sum_{i=1}^{M} l(z_i, f)\} + \lambda \|f\|_p^p \right\}$$

where $l(., f.)$ is a loss measure when f is taken as a solution, and $p \geq 0$ is a parameter.

The above learning principles are tightly connected with the neural coding methodology introduced in section 2.3. Actually, for any $f \in \digamma$, if we let $z = (f(x), x)$ be a candidate solution, then the loss $l(z_i, f)$ measures the dissimilarity between z_i and z, so $1/l(z_i, f)$ describes the similarity. Consequently, (37) can be recast as $f^* = \arg\max_f \sum_i S(z_i, z)$ with $S(z_i, z) = 1/l(z_i, f)$.

Based on the neural coding methodology, we thus propose to solve the learning problem by the revised ERM principle

$$f^* = \arg\min_{f \in \digamma} \left\{ \frac{1}{M} \sum_{i=1}^{M} w(z_i) l(z_i, f) \right\}$$

and the reviseed regularization principle

$$f^* = \arg\min_{f \in \digamma} \left\{ \frac{1}{M} \sum_{i=1}^{M} w(z_i) l(z_i, f)\} + \lambda \|f\|_p^p \right\} \tag{38}$$

where $w(z_i)$ is any fixed neural coding or something like. This then provides a more reasonable learning paradigm. The problems are: *Can we develop a similar statistical learning theory for such neural coding based paradigm? Can we develop a corresponding $L_{1/2}$ or L_1 theory for (38)?*

References

[1] J.W. Tukey, The future of data analysis, Ann. Math. Statist. 33(1962) 1–67.

[2] D.L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, American Math. Society Lecture—Match Challenges of the 21st Century, 2000.

[3] D.J. Bartholomew and M. Knott, Latent varianble methods and factor analysis, (2nded,). London: Arnold, 1999.

[4] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal. Statist. Soc B., 58(1996) 267–288.

[5] S.S. Chen, D.L. Donoho and M. A. Saundera, Automic decomposition by basis pursuit, SIAM Journal of Scientific Computing, 20(1998) 33–61.

[6] E. Candés, M. Wakin and S. Boyd, Enhancing sparsity by reweighted L_1 minimization. J.Fourier A, 14(2008) 877–905.

[7] J.M. Santos and J. Marqures, Human clustering on bi-dimensional data: an assessment, Techical Report 1, INEB-Instituto de Engenharia Biomedica, 2005.

[8] Z.B. Xu, C.Z. Li and J. Sun, Visual clustering: an approach inspired from visual psychology, Submitted to IEEE Trans. Pattern Analysis and Machine Intelligence, 2010.

[9] A.P. Witkin, Scale space filtering, Proc. Int'l Joint Conf. Artificial Intelligence, (1983) 1,019–1,022.

[10] S. Coren, L.M. Ward, and J.T. Enns, Sensation and Perception, Harcourt Brace College Publishers, 1994.

[11] Y. Leung, J.S. Zhang and Z.B. Xu, Clustering by scale-space filtering, IEEE Trans. Pattern Analysis and Machine Intelligence, 22(2000) 1396–1410.

[12] S.J. Roberts, Parametric and nonparametric unsupervised clustering analysis, Pattern Recognition, 30(1997) 261–272.

[13] L. Laurence, L. Dury and D. P. Vercauteren, Structural identification of local maxima in low-resolution promolecular electron density distributions, J.Phys. Chem. A, 107(2003), 9857–9886.

[14] S. Grossberg and E. Mingolla, Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading, Psychological Review, 92(1985) 173–211.

[15] S. E. Palmer, Model theories of gestalt perception, In: Humphreys G W, ed. Understanding Vision. CA: Blackwell, 1992.

[16] S. W. Kuffler, Discharge pattern and functional organizatin of the mammalian retina, Journal of Neurophysiology, 16(1953) 37–68.

[17] H. B. Barlow, Summation and inhibition in the frog's retina, Journal of Neurophysiology, 119(1953) 69–88.

[18] D. H. Huber and T. N. Wiesel, Receptive fields of cells in striate cortex of very young, visually inexperienced kittens, Journal of Neurophysiology, 26(1971) 994–1002.

[19] R. A. Young, R. M. Lesperance and W. W. Meyer, The Gaussian derivative model for spatial-temporal vision: I. Cortical model, Spatial Vision, 14(2001) 261–319.

[20] D. J. Field and D. J. Tolhurst, The structure and symmetry of simple-cell receptive-field profiles in the cat's visual cortex, Proceedings of the Royal Society of London. Series B, Biological Sciences, 228(1986) 379–400.

[21] S.E. Palmer, Model theories of Gestalt perception, in: Humphreys G.W., ed. Understanding Vision, CA:Blackwell, 1992.

[22] G. Karypis, E.-H.S. Han and V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling, Computer, 32(1999) 68–75.

[23] A. Y. Ng, M. I. Jordan and Y. Weiss, On spectral clustering analysis and algorithm, Advances in Neural Information Processing Systems, 14(2001) 849–856.

[24] J.B. Shi and J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Analysis and Machine Intelligence, 22(2000) 888–905.

[25] M. T. Ester, H. P. Kriegel, J. Scander and X. W. Wu, DBScan: a density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, (1996) 226–231.

[26] J. M. Santos, J. Marques and L. A. Alexandre, LEGClust – a clustering algorithm based on layered entropic subgraphs, IEEE Trans. Patten Analysis and Machine Intelligence, 30(2008) 62–75.

[27] Z.B. Xu, M.W. Dai and D.Y. Meng. A fast heuristic strategy for model selection of support vector machines, IEEE Trans. Systems, Man and Cybernetics, Part B. 39(2009) 1292–1307.

[28] D.H. Hubel and T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex. J. Physiol. London, (1968) 215–243.

[29] K. Naka and W. Rushton, S-potentials from luminosity units in the retina of fish, J. Physiology, 185(1996) 587–599.

[30] Olzak & Thomas, L.A. Olzak and J.P. Thomas, Neural recoding in human pattern vision: model and mechanisms, Vision Research, 39(1999) 231–256.

[31] J.S. Zhang and Y.W. Leung, Improved possibilistic C-Means clustering algorithm, IEEE Trans. Fuzzy System, 12(2004) 209–217.

[32] W.E. Vinje and J.L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision, Science, 287(2000) 1273–1276.

[33] B.A. Olshausen and D.J. Field, Emergence of simple cell receptive field properties by learning a sparse code for natural images, Nature, 381(1996) 607–609.

[34] Y. Karklin and M.S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. Nature, 457(2009) 83–86.

[35] B.K. Natarajan, Sparse approximate solutions to linear systems, SIAM. J. Comput. 24(1995) 227–234.

[36] H.L. Taylor, S.C. Banks and J.F. McCoy, Deconvolution with the l_1 norm, Geophysics, 44(1979) 39–52.

[37] J.F. Claerbout and F. Muir, Robust modeling with erratic data, Geophysics, 38(1973) 826–844.

[38] F. Santosa and W. W. Symes, Linear inversion of band-limited reflection seismograms, SIAM J. Sci. Stat. Comput., 7(1986) 1307–1330.

[39] D. L. Donoho and P. B. Stark, Uncertainty principles and signal recovery, SIAM J. Appl. Math., 49(1992) 906–931.

[40] D. L. Donoho and B. F. Logan, Signal recovery and the large sieve, SIAM J. Appl. Math., 52(1992) 577–591.

[41] L. I. Rudin, S. Osher and E. Fatemi, Nonlinear total variation based noise removal algorithms, Phys. D, 60(1992) 259–268.

[42] P. Blomgren and T. F. Chan, Color TV: total variation methods for restoration of vector-valued images, IEEE Trans. Image Processing, 7(1998) 304–309.

[43] D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. Information Theory, 47(2001) 2845–2862.

[44] P. Zhao and B. Yu, On model selection consistency of Lasso, Journal of Machine Learning Research, 7(2006) 2541–2563.

[45] N. Meinshausen and B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, Annals of Statistics, 720(2009) 246–270.

[46] Z.B. Xu, H. Zhang, Y. Wang and X.Y. Chang, $L_{1/2}$ regularization, Science in China Series F-Information Sciences, 40(2010) 1–11.

[47] D.L. Donoho. Neighborly polytopes and the sparse solution of underdetermined systems of linear equations. Technical Report, Statistics Department, Stanford University, 2005.

[48] D.L. Donoho. High-dimensional centrosymmetric polytopes with neighborliness proportional to dimension. Discrete and Computational Geometry, 35(2006) 617–652.

[49] Y. Wang, H.L. Guo, Z.B. Xu and H. Zhang, The representative of $L_{1/2}$ regularization among L_q $(0 < q < 1)$ regularizations: an experimental study based on a phase diagram, submitted.

[50] Z.B. Xu, F.M. Xu, X.Y. Chang and H. Zhang, $L_{1/2}$ regularization: an iterative half thresholding algorithm, submitted.

[51] Z.B.Xu, J.J.Wang and Z.S.Zhang, Convergence of iterative half thresholding algorithm for $L_{1/2}$ regularization, submitted.

[52] J.J. Wang, Z.B. Xu, Y. Wang and H. Zhang, Sparse signal recovery based on L_q $(0 < q \leq 1)$regularization, to appear in Science in China.

[53] E.J. Candés, J. Romberg and T. Tao, Signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math., 59(2005) 1207–1223.

[54] E.J. Candés, The restricted isometry property and its implications for compressed sensing, Comptas Rendus de ÍAcadèmie des Scidence, Serie I, 346(2008) 589–592.

[55] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via Lq-minimization for $0 < q \leq 1$, Applied and Computational Harmonic Analysis, 26(2009) 395–407.

[56] H.Zhang, Z.B. Xu, X.Y.Chang and Y. Liang, Varaible selection and sparse reconstruction via $L_{1/2}$ regularization, submitted.

[57] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96(2001) 1348–1360.

[58] Z.P. Liang and P.C. Lauterbur, Principles of magnetic resonance imaging: a signal processing perspective, Wiley Blackwell, 1999.

[59] E. J. Candés, J. Romberg and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory, 52(2004) 489–509.

[60] E.J. Candés and T. Tao. Decoding by linear programming, IEEE Trans. Information Theory, 51(2005) 4023–4215.

[61] E.J. Candés and J. Romberg, L1-maGIC: recovery of sparse signals via convex programming, Technical Report, Caltech, 2005.

[62] B. Efron, T. Haistie, I. Johnstone and R.Tibshirani. Least angle regression. Ann Statist, 32(2004) 407–499.

[63] I. M. Defrise and C.D. Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communication on Pure and Applied Mathematics, 11(2004) 1413–1457.

[64] T. Blumensath and M.E. Davies, Iterative hard thresholding for compressed sensing, to appear in Applied and Computational Harmonic Analysis, 2009.

[65] Z.B. Xu and G.F. Roach, Characteristic inequalities of unformaly convex and uniformaly smooth Banach spaces, J.Math.Anal. Appl. 157(1991) 189–210.