Chapter 14

Sparsity in Bayesian Signal Estimation

Ishan Wickramasingha, Michael Sobhy and Sherif S. Sherif

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70529

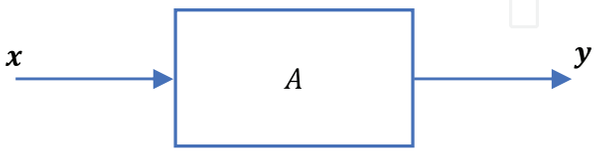
Abstract

In this chapter, we describe different methods to estimate an unknown signal from its linear measurements. We focus on the underdetermined case where the number of measurements is less than the dimension of the unknown signal. We introduce the concept of signal sparsity and describe how it could be used as prior information for either regularized least squares or Bayesian signal estimation. We discuss compressed sensing and sparse signal representation as examples where these sparse signal estimation methods could be applied.

Keywords: inverse problems, signal estimation, regularization, Bayesian methods, signal sparsity

1. Introduction

In engineering and science, a system typically refers to a physical process whose outputs are generated due to some inputs [1, 2]. Examples of systems include measuring instruments, imaging devices, mechanical and biomedical devices, chemical reactors and others. A system could be abstracted as a block diagram,



where x and y represent the inputs and outputs of the system, respectively. The block, A, formalizes the relation between these inputs and the outputs using mathematical equations [2, 3]. Depending on the nature of the system, the relation between its inputs and outputs

could be either linear or nonlinear. For a linear relation, the system is called a *linear system* and it would be represented by a set of linear equations [3, 4]

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{1}$$

In this chapter, we will restrict our attention to linear systems, as they could adequately represent many actual systems in a mathematically tractable way.

When dealing with systems, two typical types of problems arise, forward and inverse problems.

1.1. Forward problems

In a *forward problem*, one would be interested in obtaining the output of a system due to a particular input [5, 6]. For linear systems, this output is the result of a simple matrix-vector product, $\boldsymbol{Ax}$. Forward problems usually become more difficult as the number of equations increases or as uncertainties about the inputs, or the behavior of the system, are present [6].

1.2. Inverse problems

In an *inverse problem*, one would be interested in inferring the inputs to a system $\boldsymbol{x}$ that resulted in observed outputs, i.e., measured $\boldsymbol{y}$ [5, 6]. Another formulation of an inverse problem is to identify the behavior of the system, i.e., construct $\boldsymbol{A}$, from knowledge of different input and output values. This problem formulation is known as *system identification* [1, 7, 8]. In this chapter, we will only consider the input inference problem. The nature of the input $\boldsymbol{x}$ to be inferred further leads to two broad categories of this problem: *estimation*, and *classification*. In input estimation, the input could assume an infinite number of possible values [4, 9], while in input classification the input could assume only a finite number (usually small) of possible values [4, 9]. Accordingly, in input classification, one would like to only assign an input to a predetermined signal class. In this chapter, we will only focus on estimation problems, particularly on restoring an input signal $\boldsymbol{x}$ from noisy data $\boldsymbol{y}$ that is obtained using a linear measuring system represented by a matrix $\boldsymbol{A}$.

2. Signal restoration as example of an inverse problem

To solve the above signal restoration problem, we need to estimate input signal $\boldsymbol{x}$ through the inversion of matrix $\boldsymbol{A}$. This could be a hard problem because in many cases the inverse of $\boldsymbol{A}$ might not exist, or the measurement data, $\boldsymbol{y}$, might be corrupted by noise. The existence of the inverse of $\boldsymbol{A}$ depends on the number of acquired independent measurements relative to the dimension of the unknown signal [5, 10]. The conditions for the existence of a stable solution of any inverse problem, i.e., for an inverse problem to be well-posed, have been addressed by Hadamard as:

- *Existence*: for measured output $\boldsymbol{y}$ there exists at least one corresponding input $\boldsymbol{x}$.
- *Uniqueness*: for measured output $\boldsymbol{y}$ there exists only one corresponding input $\boldsymbol{x}$.
- *Continuity*: as the input $\boldsymbol{x}$ changes slightly, the output $\boldsymbol{y}$ changes slightly, i.e., the relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is continuous.

These conditions could be applied to linear systems as conditions on the matrix $\boldsymbol{A}$. Let the matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, such that $\mathbb{R}^{n \times m}$ denotes the set of matrices of dimension $n \times m$ with its elements being real values. The matrix equation, $\boldsymbol{y}_{n \times 1} = \boldsymbol{A}_{n \times m}\, \boldsymbol{x}_{m \times 1}$, is equivalent to n linear equations with m unknowns. The matrix $\boldsymbol{A}$ is a linear transformation that maps input signals from its domain $\mathfrak{D}(\boldsymbol{A}) = \mathbb{R}^m$ to its range $\mathcal{R}(\boldsymbol{A}) = \mathbb{R}^n$ [4, 5, 10]. For any measured output signal $\boldsymbol{y} \in \mathbb{R}^n$, we could identify three cases based on the values of n and m.

2.1. Underdetermined linear systems

In this case, $n < m$, i.e., the number of equations is less than the number of unknowns,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}. \tag{2}$$

If these equations are consistent, *Hadamard's Existence* condition will be satisfied. However, Hadamard's *Uniqueness* condition is not satisfied because the *Null Space*$(\boldsymbol{A}) \neq \{\mathbf{0}\}$, i.e., there exist $\boldsymbol{z} \neq \mathbf{0} \in$ *Null Space*$(\boldsymbol{A})$ such that,

$$\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{z}) = \boldsymbol{y}. \tag{3}$$

This linear system is called *under-determined* because its equations, i.e., system constraints, are not enough to uniquely determine $\boldsymbol{x}$ [4, 5]. Thus, the inverse of $\boldsymbol{A}$ does not exist.

2.2. Overdetermined linear systems

In this case, $m > n$, the number of equations is more than the number of unknowns,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}. \tag{4}$$

If these equations are consistent, Hadamard's *Existence* condition will not be satisfied. However, *Hadamard*'s *Uniqueness* condition will be satisfied, if $\boldsymbol{A}$ has full rank. In this case, *Null Space* $(\boldsymbol{A}) = \{\mathbf{0}\}$, i.e.,

$$\boldsymbol{A}(\boldsymbol{x} + \mathbf{0}) = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{5}$$

This linear system is called *over-determined*, because its equations, i.e., system constraints, are too many for $\boldsymbol{x}$ to exist [4, 5]. Also, the inverse of $\boldsymbol{A}$ does not exist.

2.3. Square linear systems

The case where $m = n$, the number of equations is equal to the number of unknowns,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}. \tag{6}$$

If A has full rank, its *Null Space*(A) = {**0**} and both Hadamard's *Existence* and *Uniqueness* conditions will be satisfied. In addition, if A has a small condition number, the relation between x, y will be continuous, and Hadamard's *Continuity* condition will be satisfied [4, 5, 10]. In this case, the inverse problem formulated by this system of linear equations is well-posed.

3. Methods for signal estimation

In this section, we will focus on the estimation of an input signal x from a noisy measurement y of the output of a linear system A.

The linear system shown in **Figure 1**, could be modeled as,

$$y = Ax + v. \tag{7}$$

where v is additive Gaussian noise. As a consequence of the *Central Limit Theorem*, this assumption of Gaussian distributed noise is valid for many output measurement setups.

Statistical Estimation Theory allows one to obtain an estimate $\hat{x}$ of a signal x that is input to a known system A from measurement y (see **Figure 2**) [11, 12]. However, this estimate $\hat{x}$ is not unique, as it depends on the choice of the used estimator from the different ones available. In addition to measurement y, if other information about the input signal is available, it could be

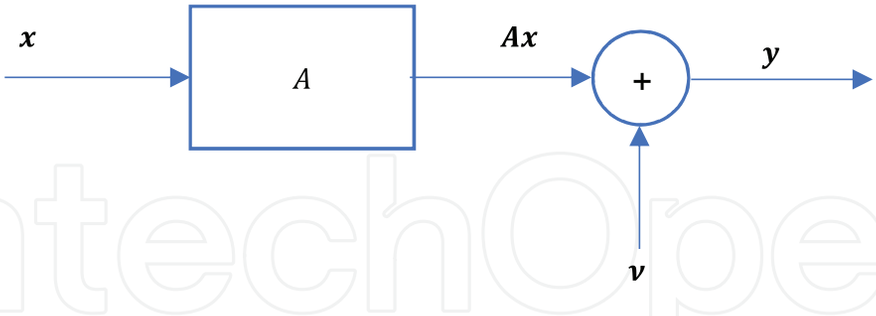


Figure 1. Linear system with noisy output measurement.

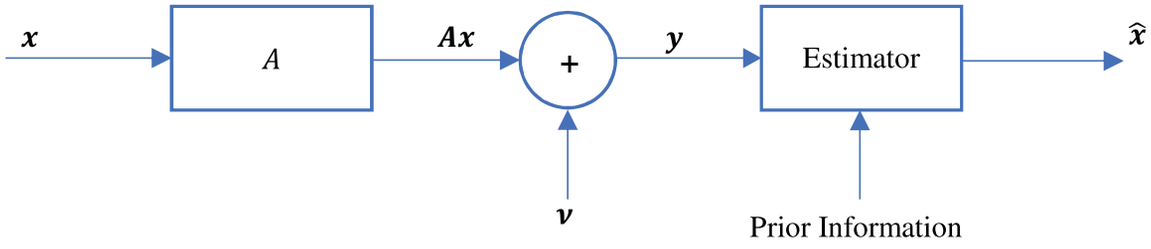


Figure 2. Signal estimation using prior information.

used as prior information to constrain the estimator to produce a better estimate of $\boldsymbol{x}$. Signal estimation for overdetermined systems could be achieved without any prior information about the input signal. However, for underdetermined systems, prior information is necessary to ensure a unique estimate.

3.1. Least squares estimation

If there is no information available about the statistics of the measured data,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}, \tag{8}$$

least squares estimation could be used. The least squares estimate is obtained by minimizing the square of the L_2 norm of the error between the measurement and the linear model, $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$. It is given by

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2. \tag{9}$$

The L_2 norm is a special case of the p-norm of a vector, where $p = 2$, that is defined as $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}}$. In Eq. (9), the unknown $\boldsymbol{x}$ is considered deterministic, so its statistics are not required. The noise $\boldsymbol{v}$ in this formulation is implicitly assumed to be white noise with variance σ^2 [13, 14]. Least squares estimation is typically used to estimate input signals $\boldsymbol{x}$ in overdetermined problems. Since $\widehat{\boldsymbol{x}}$ is unique in this case, no prior information, additional constraints, for $\boldsymbol{x}$ is necessary.

3.2. Weighted least squares estimation

If the noise $\boldsymbol{v}$ in Eq. (8) is not necessarily white and its second order statistics, i.e., mean and covariance matrix, are known, then weighted least squares estimation could be used to further improve the least squares estimate. In this estimation method, measurement errors are not weighted equally, but a weighting matrix $\boldsymbol{C}$ explicitly specifies such the weights. The weighted least squares estimate is given by

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{C}^{-1/2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})\|_2^2. \tag{10}$$

We note that the least squares problem, Eq. (9), is a special case of the weighted least squares problem, Eq. (10), when $\boldsymbol{C} = \sigma^2\boldsymbol{I}$.

3.3. Regularized least squares estimation

In underdetermined problems, the introduction of additional constraints on $\boldsymbol{x}$, also known as *regularization*, could ensure the uniqueness of the obtained solution. Standard least squares estimation could be extended, through regularization, to solve underdetermined estimation problems. The regularized least squares estimate is given by

$$\arg\min_{x} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{L}\boldsymbol{x}\|_2, \tag{11}$$

where $\boldsymbol{L}$ is a matrix specifying the additional constraints and λ is a *regularization parameter* whose value determines the relative weights of the two terms in the objective function. If the combined matrix $\begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{L} \end{bmatrix}$ has full rank, the regularized least squares estimate $\widehat{\boldsymbol{x}}$ is unique [4]. In this optimization problem, the unknown $\boldsymbol{x}$ is once again considered deterministic, so its statistics are not required. It is worthwhile noting that while *regularization* is necessary to solve underdetermined inverse problems, it could also be used to improve numerical properties, e.g., condition number, of either linear overdetermined or linear square inverse problems.

3.4. Maximum likelihood estimation

If the probability distribution function (pdf) of the measurement $\boldsymbol{y}$, parameterized by an unknown deterministic input signal $\boldsymbol{x}$, is available, then the *maximum likelihood* estimate of $\boldsymbol{x}$ is given by,

$$\widehat{\boldsymbol{x}} = \arg\max_{x} f(\boldsymbol{y}|\boldsymbol{x}). \tag{12}$$

This maximum likelihood estimate $\widehat{\boldsymbol{x}}$ is obtained by assuming that measurement $\boldsymbol{y}$ is the most likely measurement to occur given the input signal $\boldsymbol{x}$. This corresponds to choosing the value of $\boldsymbol{x}$ for which the probability of the observed measurement $\boldsymbol{y}$ is maximized. In maximum likelihood estimation, the negative log of the likelihood function, $f(\boldsymbol{y}|\boldsymbol{x})$, is typically used to transform Eq. (12) into a simpler minimization problem. When, $f(\boldsymbol{y}|\boldsymbol{x})$ is a Gaussian distribution, $\boldsymbol{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{C})$, minimizing the negative log of the likelihood function is equivalent to solving the weighted least squares estimation problem.

3.5. Bayesian estimation

If the conditional pdf of the measurement $\boldsymbol{y}$, given an unknown random input signal $\boldsymbol{x}$, is known, in addition to the marginal pdf of $\boldsymbol{x}$, representing prior information about $\boldsymbol{x}$, is given, then a Bayesian estimation method would be possible. The first step to obtain one of the many possible Bayesian estimates of $\boldsymbol{x}$ is to use Bayes rule to obtain the *a posteriori* pdf,

$$f(\boldsymbol{x}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}{\int f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}. \tag{13}$$

Once this *a posteriori* pdf is known, different Bayesian estimates $\widehat{\boldsymbol{x}}$ could be obtained. For example, the *minimum mean square error* estimate is given by,

$$\widehat{\boldsymbol{x}}_{MMSE} = E_x[f(\boldsymbol{x}|\boldsymbol{y})] = E_x\left[\frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}{\int f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}\right], \tag{14}$$

while the *maximum a priori* (MAP) estimate is given by,

http://dx.doi.org/10.5772/intechopen.70529

$$\widehat{\boldsymbol{x}}_{MAP} = \arg\ \max_{\boldsymbol{x}} f(\boldsymbol{x}|\boldsymbol{y}) = \arg\ \max_{\boldsymbol{x}} f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}). \tag{15}$$

We note that the maximum likelihood estimate, Eq. (12), is a special case of the MAP estimate, when $f(\boldsymbol{x})$ is a uniform pdf over the entire domain of $\boldsymbol{x}$. The use of prior information is essential to solve underdetermined inverse problems, but it also improves the numerical properties, e.g., condition number, of either linear overdetermined or linear square inverse problems.

3.5.1. Bayesian least squares estimation

In least squares estimation, the vector $\boldsymbol{x}$ is assumed to be an unknown deterministic variable. However, in Bayesian least squares estimation, it is considered a vector of scalar random variables that satisfies statistical properties given by an *a priori* probability distribution function [5]. In addition, in least squares estimation, the L_2 norm of the measurement error is minimized, while in Bayesian least squares estimation, it is the estimation error, $\boldsymbol{e} = \widehat{\boldsymbol{x}} - \boldsymbol{x}$, not measurement error, that is used [5]. Since $\boldsymbol{x}$ is assumed to be a random vector, the estimation error $\boldsymbol{e}$ will also be a random vector. Therefore, the Bayesian least squares estimate could be obtained by minimizing the condtional mean of the square of the estimation error, given measurement, $\boldsymbol{y}$,

$$\widehat{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}} E\left[(\widehat{\boldsymbol{x}} - \boldsymbol{x})^T (\widehat{\boldsymbol{x}} - \boldsymbol{x}) | \boldsymbol{y}\right]. \tag{16}$$

When $\boldsymbol{x}$ has a Gaussian distribution and $\boldsymbol{A}$ represents a linear system, then measurement $\boldsymbol{y}$ will also have a Gaussian distribution. In this case, the Bayesian least squares estimate given by Eq. (16) could be reinterpreted as a regularized least squares estimate given by,

$$\widehat{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\mu - \boldsymbol{x}\|, \tag{17}$$

where μ is the mean of the *a priori* distribution of $\boldsymbol{x}$ [5]. Therefore, a least squares Bayesian estimate is analogous to a regularized least squares estimate, where *a priori* information about $\boldsymbol{x}$ is expressed as additional constraints on $\boldsymbol{x}$ in the form of a *regularization* term.

3.5.2. Advantages of Bayesian estimation over other estimation methods

Bayesian estimation techniques could be used, given that a reliable *a priori* distribution is known, to obtain an accurate estimate of a signal $\boldsymbol{x}$, even if the number available measurements is smaller than the dimension of the signal to estimated. In this underdetermined case, Bayesian estimation could accurately estimate a signal while un-regularized least squares estimation or maximum likelihood estimation could not. The use of prior information in Bayesian estimation could also improves the numerical properties, e.g., condition number, of either linear overdetermined or linear square inverse problems. This could be understood by keeping in mind the mathematical equivalence between obtaining one scalar measurement related to $\boldsymbol{x}$, and specifying one constraint that $\boldsymbol{x}$ has to satisfy. Therefore, as the number of available measurements significantly increases, both Bayesian and maximum likelihood estimates would converge to the same estimate.

Bayesian estimation also could be easily adapted to estimate dynamic signals that change over time. This is achieved by sequentially using past estimates of a signal, e.g., x_{t-1}, as prior information to estimate its current value x_t. More generally, Bayesian estimation could be easily adapted for *data fusion*, i.e., combination of multiple partial measurements to estimate a complete signal in remote sensing, stereo vision and tomographic imaging, e.g., Positron emission tomography (PET), Magnetic resonance imaging (MRI), computed tomography (CT) and optical coherence tomography (OCT). Bayesian methods could also easily fuse all available prior information to provide an estimate based on measurements, in addition to all known information about a signal.

Bayesian estimation techniques could be extended in straight forward ways to estimate output signals of nonlinear systems or signals that have complicated probability distributions. In these cases, numerical Bayesian estimates are typically obtained using Monte Carlo methods.

3.5.3. Sparsity as prior information for underdetermined Bayesian signal estimation

Sparse signal representation means the representation of a signal in a domain where most of its coefficients are zero. Depending on the nature of the signal, one could find an appropriate domain where it would be sparse. This notion could be useful in signal estimation because assuming that the unknown signal x is sparse could be used as prior information to obtain an accurate estimate of it, even if only a small number of measurements are available. The rest of this chapter will focus on using signal sparsity as prior information for underdetermined Bayesian signal estimation.

4. Sparse signal representation

As shown in **Figure 3**, a sinusoid is a dense signal in the time domain. However, it could be represented by a single value, i.e., it has a sparse representation, in the frequency domain.

We note that any signal could have a sparse representation in a suitable domain [15]. A sparse signal representation means a representation of the signal in a domain where most of its coefficients are zero. Sparse signal representations have many advantages including:

1. A sparse signal representation requires less memory for its storage. Therefore, it is a fundamental concept for signal compression.

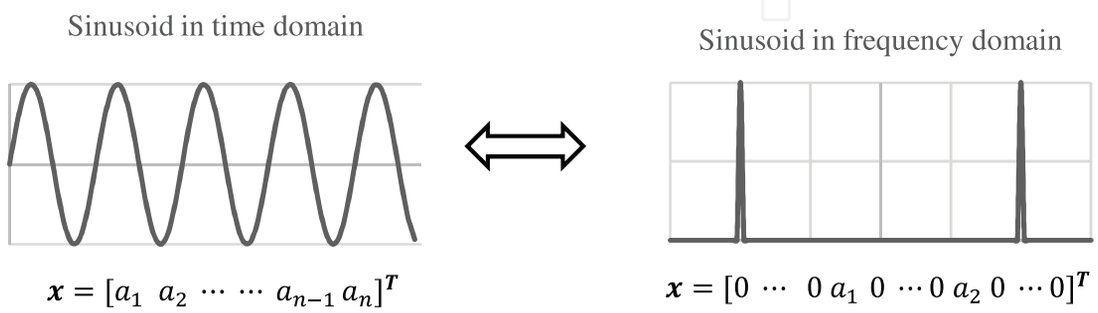


Figure 3. A sinusoid in time and frequency domains.

2. A sparse signal representation could lead to simpler signal processing algorithms. For example, signal denoising could be achieved by simple thresholding operations in a domain where the signal is known to be sparse.

3. Sparse signal representations have fewer coefficients than dense signal representations. Therefore, the computational cost for sparse representations would be lower than for dense representations.

4.1. Signal representation using a dictionary

A *dictionary* $\mathcal{D}$ is a collection of vectors $\{\phi_n\}_{n\epsilon\Gamma}$, indexed by a parameter $n\ \epsilon\ \Gamma$ equal to the dimension of a signal *f*, where we could represent *f* as a linear combination [16],

$$f = \sum_{n\epsilon\Gamma} c_n \phi_n. \tag{18}$$

If the vectors $\{\phi_n\}_{n\epsilon\Gamma}$ are linearly independent, then such dictionary is called a *basis*. Representing a signal as a linear combination of sinusoids, i.e., using a *Fourier* dictionary, is very common. *Wavelet* dictionaries and *Chirplet* dictionaries are also common dictionaries for signal representation. Dictionaries could be combined together to obtain a larger dictionary, where $n\ \epsilon\Gamma$ is larger than the dimension the signal *f*, that is called an *overcomplete* dictionary or a frame.

4.1.1. Signal representation using a basis

A set of vectors form a basis for $\mathbb{R}^n$ if they span $\mathbb{R}^n$ and are linearly independent. A basis in a vector space *V* is a set *X* of linearly independent vectors such that every vector in *V* is a linear combination of elements in *X*. A vector space *V* is finite dimensional if it has a finite number of basis vectors [17].

Depending on the properties of $\{\phi_n\}_{n\epsilon\Gamma}$, bases could be classified into different types, e.g., orthogonal basis, orthonormal basis, biorthogonal basis, global basis and local basis. For an orthogonal basis, its basis vectors in the vector space *V* are mutually orthogonal,

$$\langle \phi_m, \phi_n \rangle = 0 \ \text{ for } \ m \neq n. \tag{19}$$

For an orthonormal basis, its basis vectors in the vector space *V* are mutually orthogonal and have unit length,

$$\langle \phi_m, \phi_n \rangle = \delta(m-n), \tag{20}$$

where $\delta(m-n)$ is the Kronecker delta function. For a biorthogonal basis, its basis vectors are not orthogonal to each other, but they are orthogonal to vectors in another basis, $\left\{\widetilde{\phi}_n\right\}_{n\epsilon\Gamma}$, such that

$$\left\langle \phi_m, \widetilde{\phi}_n \right\rangle = \delta(m-n). \tag{21}$$

In addition, depending on the domain (support) on which these basis vectors are defined, we could also classify a basis as either global or local. Sinusoidal basis vectors used for the discrete Fourier transform are defined on the entire domain (support) of *f*, so they are considered a *global* basis. Many wavelet basis vectors used for the discrete wavelet transform are defined on only part of the domain (support) of *f*, so they are considered a *local* basis.

4.1.2. Signal representation using a frame

A frame is a set of vectors $\{\phi_n\}_{n\epsilon\Gamma}$ that spans $\mathbb{R}^n$ and could be used to represent a signal *f* from the inner products $\{\langle f, \phi_n\rangle\}_{n\epsilon\Gamma}$. A frame allows the representation of a signal as a set of frame coefficients, and its reconstruction from these coefficients in a numerically stable way

$$f = \sum_{n\epsilon\Gamma} \langle f, \phi_n \rangle \phi_n. \tag{22}$$

Frame theory analyzes the completeness, stability, and redundancy of linear discrete signal representations [18]. A frame is not necessarily a basis, but it shares many properties with bases. The most important distinction between a frame and a basis is that the vectors that comprise a basis are linearly independent, while those comprising frame could be linearly dependent. Frames are also called *overcomplete* dictionaries. The redundancy in the representation of a signal using frames could be used to obtain sparse signal representations.

4.2. Sparse signal representation as a regularized least squares estimation problem

If designed to concentrate the energy of a signal in a small number of dimensions, an orthogonal basis would be the minimum-size dictionary that could yield a sparse representation of this signal [15]. However, finding an orthogonal basis that yields a highly sparse representation for a given signal is usually difficult or impractical. To allow more flexibility, the orthogonality constraint is usually dropped, and *overcomplete* dictionaries (frames) are usually used. This idea is well explained in the following quote by Stephane Mallat:

> *"In natural languages, a richer dictionary helps to build shorter and more precise sentences. Similarly, dictionaries of vectors that are larger than bases are needed to build sparse representations of complex signals. Sparse representations in redundant dictionaries can improve pattern recognition, compression, and noise reduction but also the resolution of new inverse problems. This includes super resolution, source separation, and compressed sensing" [15].*

Thus representing a signal using a particular *overcomplete* dictionary has the following goals [16]

- Sparsity—this representation should be more sparse than other representations.
- Super resolution—the resolution of the signal when represented using this dictionary should be higher than when represented in any other dictionary.
- Speed—this representation should be computed in $O(n)$ or $O(n\ log(n))$ time.

A simple way to obtain an *overcomplete* dictionary $\boldsymbol{A}$ is to use a union of basis $\boldsymbol{A_i}$ that would result in the following representation of a signal $\boldsymbol{y}$,

$$(\boldsymbol{y}) = \underbrace{([\boldsymbol{A}_1][\boldsymbol{A}_2][\boldsymbol{A}_3][\boldsymbol{A}_4][\boldsymbol{A}_5])}_{\boldsymbol{A}}(\boldsymbol{x}) \Rightarrow \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{23}$$

where $\boldsymbol{A}$ is a $n \times m$ matrix representing the dictionary and $\boldsymbol{x}$ are the coefficients representing $\boldsymbol{y}$ in the domain defined by $\boldsymbol{A}$. Since $\boldsymbol{A}$ represents an *overcomplete* dictionary, the number of its rows will be less than the number of its columns. Eq. (23) is a formulation of the signal representation problem as an underdetermined inverse problem.

To obtain a sparse solution for Eq. (23) one needs to find an $m \times 1$ coefficient vector $\widehat{\boldsymbol{x}}$, such that,

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_0, \tag{24}$$

where $\|\boldsymbol{x}\|_0$ is the cardinality of vector $\boldsymbol{x}$, i.e., its number of nonzero elements, and $\lambda > 0$ is a regularization parameter that quantifies the tradeoff between the signal representation error,$\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, and its sparsity level, $\|\boldsymbol{x}\|_0$ [19]. The cardinality of vector $\boldsymbol{x}$ is sometimes referred to as the L_0 norm of $\boldsymbol{x}$, even though $\|\boldsymbol{x}\|_0$ is actually a *pseudo norm* that does not satisfy the requirements of a norm in $\mathbb{R}^m$. This sparse signal representation problem, Eq. (24), has a form similar to the regularized least squares estimation problem, Eq. (11), that would be *underdetermined* in the case of an *overcomplete* dictionary. Because of the correspondence between regularized least squares estimation and Bayesian estimation, the problem of finding a sparse representation of a signal could be formulated as a Bayesian estimation problem.

5. Compressed sensing

Compressed sensing involves the estimation of a signal using a number of measurements that are significantly less than its dimension [20]. By assuming that the unknown signal is sparse in the domain where the measurements were acquired, one could use this sparsity constraint as prior information to obtain an accurate estimate of the signal from relatively few measurements.

Compressed sensing is closely related to *signal compression* that is routinely used for efficient storage or transmission of signals. Compressed sensing was inspired by this question: instead of the typical signal acquisition followed by signal compression, is there a way to acquire (sense) the compressed signal in the first place? If possible, it would significantly reduce the number of measurements and the computation cost [20]. In addition, this possibility would allow acquisition of signals that require extremely high, hence impractical, sampling rates [21]. As an affirmative answer to this question, compressed sensing was developed to combine signal compression with signal acquisition [20]. This is achieved by designing the measurement setup to acquire signals in the domain where the unknown signal is assumed to be sparse.

In compressed sensing, we consider the estimation of an input signal $\boldsymbol{x} \in \mathbb{R}^n$ from m linear measurements, where $m \ll n$. As discussed above, this problem could be written as an underdetermined linear system,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{25}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ represent the measurements and measurement (sensing) matrix, respectively.

Assuming that the unknown signal $\boldsymbol{x}$ is s-sparse, i.e., $\boldsymbol{x} \in \sum_s$ has only s nonzero elements, in the domain specified by the measurement (sensing) matrix $\boldsymbol{A}$, and assuming that $\boldsymbol{A}$ satisfies the restricted isometry property (RIP) of order $2s$, i.e., there exists a constant $\delta_{2s} \in (0, 1)$ such that,

$$(1 - \delta_{2s})||\boldsymbol{z}||_2^2 \leq ||\boldsymbol{A}\boldsymbol{z}||_2^2 \leq (1 + \delta_{2s})||\boldsymbol{z}||_2^2, \tag{26}$$

for all $\boldsymbol{z} \in \sum_{\boldsymbol{2s}}$, then $\boldsymbol{x}$ could be reconstructed from $m \geq s$ measurements by different optimization algorithms [20]. When the measurements $\boldsymbol{y}$ are noiseless, $\boldsymbol{x}$ could be exactly estimated from,

$$\min_{\boldsymbol{x}} ||\boldsymbol{x}||_0 \ \text{ subject to } \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{27}$$

However, when the measurements $\boldsymbol{y}$ are contaminated by noise,$\boldsymbol{x}$ could be obtained as the regularized least squares estimate,

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||_2^2 + \lambda\, ||\boldsymbol{x}||_0. \tag{28}$$

This minimization problem could also be mathematically reformulated and solved as a Bayesian estimation problem.

6. Obtaining sparse solutions for signal representation and signal estimation problems

From Sections 4 and 5 we note that the problem of obtaining a sparse signal representation, Eq. (24) and the problem of sparse signal estimation in compressed sensing, Eq. (28), both have the same mathematical form [11, 22],

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_0 . \tag{29}$$

In this section, we describe different approaches to solving this minimization problem. From Eq. (29), we note that the first term of its RHS, $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, represents either signal reconstruction error (sparse signal representation problem) or measurement fitting error (sparse signal estimation in compressed sensing problem), while the second term of its RHS,$\|\boldsymbol{x}\|_0$, represents the cardinality (number of nonzero coefficients) of the unknown signal. The regularization parameter λ specifies the tradeoff between these two terms in the objective function. The selection of an appropriate value of λ to balance the reconstruction, or fitting error, and signal sparsity is very important. Regularization theory and Bayesian approaches could provide ways to determine optimal values of λ [23–26].

Convex optimization problems is a class of optimization problems that are significantly easier to solve compared to nonconvex problems [34]. Another advantage of convex optimization problems is that any local solution, e.g., a local minimum, is guaranteed to be a global solution. We note that obtaining an exact solution for the minimization problem in Eq. (29) is difficult because it is nonconvex. Therefore, one could either seek an approximate solution to this nonconvex problem or approximate this problem by a convex optimization whose exact solution could be obtained easily.

Considering the general regularized least squares estimation problem,

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_p, \tag{30}$$

we note that it is a nonconvex optimization problem for $0 \leq p < 1$ and a convex optimization problem for $p \geq 1$. One alternative to approximate Eq. (29) by a convex optimization problem, one could relax the strict condition of minimizing the cardinality of the signal, $\|\boldsymbol{x}\|_0$, by replacing by it by the sparsity-promoting condition of minimizing the L_1 norm of the signal, $\|\boldsymbol{x}\|_1$. Another alternative to approximate Eq. (29) by another nonconvex optimization problem that is easier to solve than the original problem using a Bayesian formulation, is to replace $\|\boldsymbol{x}\|_0$ by $\|\boldsymbol{x}\|_p$, $0 < p < 1$. The minimization of Eq. (30) using $\|\boldsymbol{x}\|_p$, $0 < p < 1$ would result in a higher degree of signal sparsity compared to when $\|\boldsymbol{x}\|_1$ is used. This could be understood visually by examining **Figure 4**, that shows the shapes of two-dimensional unit balls using (pseudo)norms with different values of p.

We explain further details in the following subsections.

6.1. Obtaining a sparse signal solution using L_0 minimization

The sparsest solution of the regularized least squares estimation problem, Eq. (29) would be obtained when $p = 0$ in $\|\boldsymbol{x}\|_p$. As shown in **Figure 5**, the solution of the regularized least squares problem, $\widehat{\boldsymbol{x}}$, is given by the intersection of the circles, possibly ellipses, representing the

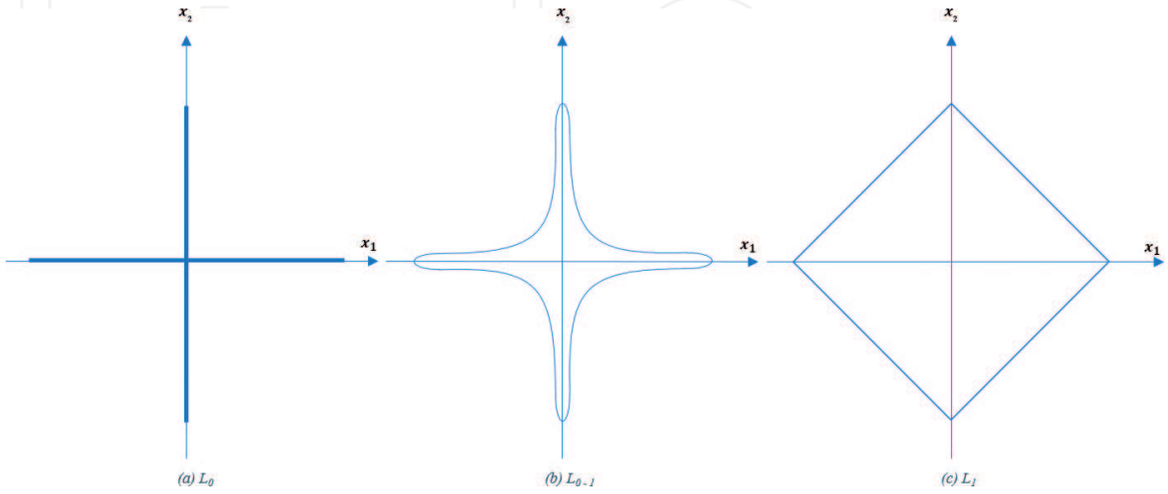


Figure 4. Two-dimensional unit ball using different (pseudo)norms. (a) L_0, (b) L_{0-1}, and (c) L_1.

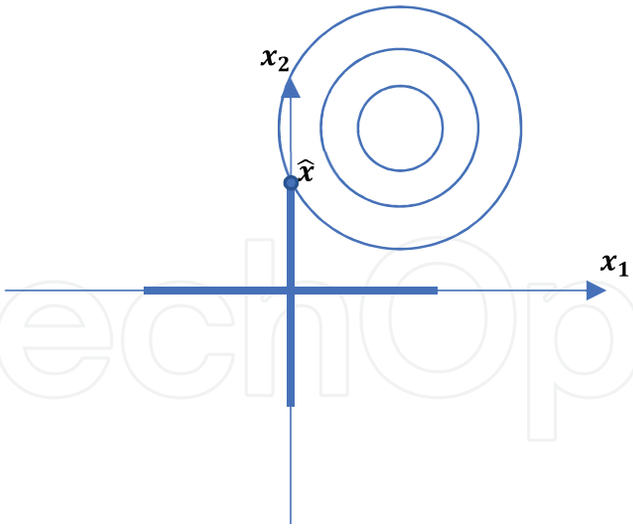


Figure 5. Regularized least squares using L_0.

solution of the unconstrained least squares estimation problem and the unit ball using L_0 representing the constraint of minimizing L_0. In this case of minimizing L_0, the unconstrained least squares solution will always intersect the unit ball at an axis, this yielding the most possible sparse solution. However, as mentioned earlier, this L_0 minimization problem is difficult to solve because it is nonconvex. Approximate solutions for this problem could be obtained using greedy optimization algorithms, e.g., Matching Pursuits [27] and Least Angle Regression (LARS) [28].

6.2. Obtaining a sparse signal solution using L_1 minimization

On relaxing the nonconvex regularized least squares using L_0 minimization problem, by setting $p = 1$, we obtain the convex L_1 minimization problem. As shown in **Figure 4(c)**, the unit ball using the L_1 norm covers a larger area than the unit ball using the L_0 pseudo norm, shown in **Figure 4(a)**. Therefore, as shown in **Figure 6**, the solution for the regularized least squares problem using the L_1 minimization would be sparse, but it should not be expected to be as sparse as the L_0 minimization problem.

This L_1 minimization problem could be solved easily using various algorithms, e.g., Basis Pursuits [16], Method of frames (MOF) [29], Lasso [30, 31], and Best Basis Selection [32, 33]. A Bayesian formulation of this L_1 minimization problem is also possible by assuming that the *a priori* probability distribution of $\boldsymbol{x}$ is Laplacian, $\boldsymbol{x} \sim e^{-|x|}$.

6.3. Obtaining a sparse signal solution using L_{0-1} minimization

As discussed above, solving the regularized least squares problem with L_0 minimization should yield the sparsest signal solution. However, only approximate solutions are available for this difficult nonconvex problem. Alternatively, solving the regularized least squares problem with L_1 minimization should yield an exact sparse solution that would be less sparse than in the L_0 case, but it is considerably easier to obtain.

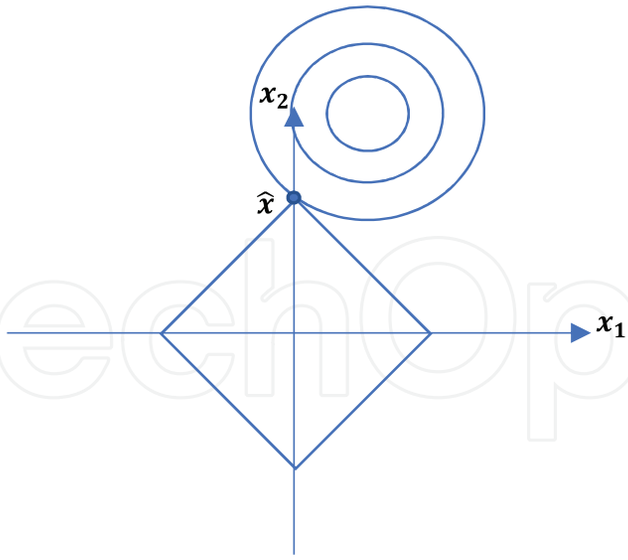


Figure 6. Regularized least squares using L_1.

The regularized least squares problem could also be formulated as an $L_{0\ -\ 1}$ minimization problem. As $\|\boldsymbol{x}\|_p$, $0 < p < 1$,that we abbreviate as $L_{0\ -\ 1}$, is not an actual norm, this optimization problem would be nonconvex [34]. The advantage of using $L_{0\ -\ 1}$ minimization is that, as shown in **Figure 4(b)**, compared to unit ball using the L_1 norm, the unit ball using the $L_{0\ -\ 1}$ pseudo norm has a narrower area that is concentrated around the axes. Therefore, as shown in **Figure 7**, the $L_{0\ -\ 1}$ minimization problem should yield a sparser solution compared to the L_1 minimization problem.

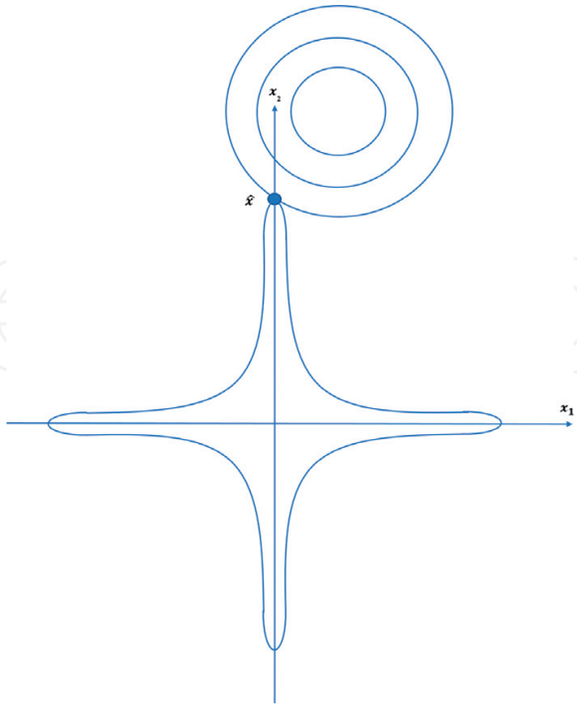


Figure 7. Regularized least squares using L_{0-1}.

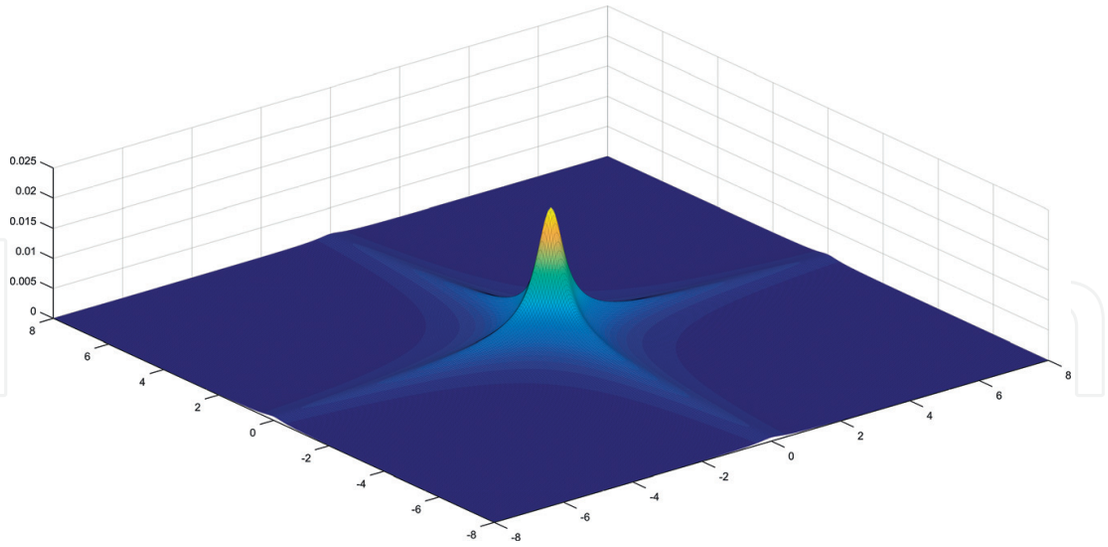


Figure 8. Product of two *student-t* probability distributions.

Another advantage of using $L_{0\ -\ 1}$ minimization is that this nonconvex optimization problem could be easily formulated as a Bayesian estimation problem that could be solved using Markov Chain Monte Carlo (MCMC) methods. As shown in **Figure 8**, the product of *student-t* probability distributions has a shape similar to the unit ball using the $L_{0\ -\ 1}$ pseudo norm, so *student-t* distributions could be used as *a priori* distributions to approximate the $L_{0\ -\ 1}$ pseudo norm.

6.4. Bayesian method to obtain a sparse signal solution using $L_{0\ -\ 1}$ minimization

As mentioned in Section 3.5, the first step to obtaining one of the many possible Bayesian estimates of $\boldsymbol{x}$ is to use Bayes rule to obtain the *a posteriori* pdf,

$$f(\boldsymbol{x}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}{\int f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}\,. \tag{31}$$

Using this *a posteriori* distribution, one could obtain a sparse signal solution using $L_{0\ -\ 1}$ minimization, as the *maximum a posteriori* (MAP) estimate given by Eq. (15). Compared to other Bayesian estimates, the MAP estimate could be easier to obtain because the calculation of the normalizing constant, $\int f(\boldsymbol{y}\,|\,\boldsymbol{x})f(\boldsymbol{x})$,would not be needed. The maximization of the product of conditional probability distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ and the *a priori* distribution of $\boldsymbol{x}$ is equivalent to the minimizing of the sum of their negative logarithms,

$$\widehat{\boldsymbol{x}}_{MAP} = \arg\min_{\boldsymbol{x}} \left[-\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{x})\right]. \tag{32}$$

In the case of white Gaussian measurement noise, $p(\boldsymbol{y}\,|\,\boldsymbol{x}) \sim N_x(\boldsymbol{Ax}, \sigma^2\boldsymbol{I})$ where $-\log p(\boldsymbol{y}|\boldsymbol{x}) \propto \|\boldsymbol{y} - \boldsymbol{Ax}\|_2^2$, which the first term of the RHS of Eq. (30). As discussed in the previous section, the *a priori* probability $p(\boldsymbol{x})$ corresponding to $L_{0\ -\ 1}$ minimization could be represented as a product of univariate *student-t* probability distribution functions [14],

$$p(\boldsymbol{x}) = \prod_{i=1}^{M} stud_{\boldsymbol{x}_i}[0, 1, \vartheta] = \prod_{i=1}^{M} \frac{\Gamma\left(\frac{\vartheta+1}{2}\right)}{\sqrt{\vartheta\pi}\Gamma\left(\frac{\vartheta}{2}\right)} \left(1 + \frac{\boldsymbol{x}_i^2}{\vartheta}\right)^{-\frac{(\vartheta+1)}{2}}, \tag{33}$$

where Γ is the Gamma function, and ϑ is the number of degrees of freedom of the *student-t* distribution. Since this *a priori* distribution function is not an exponential function, we would use Eq. (15) instead of Eq. (32) to obtain the MAP estimate.

Because the prior is not a Gaussian distribution, there is no simple closed form expression for the posterior, $p(\boldsymbol{x}|\boldsymbol{y})$ with a *student-t a priori* probability distribution. However, we could express each *student-t* distribution as an infinite weighted sum of Gaussian distributions, where the hidden variables h_i determine their variances [14].

$$p(\boldsymbol{x}) = \prod_{i=1}^{M} \int N_{\boldsymbol{x}_i}(0, 1/h_i)\, Gam_{h_i}[\vartheta/2, \vartheta/2] dh_i = \int N_{\boldsymbol{x}}\left(\boldsymbol{0}, \boldsymbol{H}^{-1}\right) \prod_{i=1}^{M} Gam_{h_i}[\vartheta/2, \vartheta/2] d\boldsymbol{H}, \tag{34}$$

where the matrix $\boldsymbol{H}$ contains the hidden variables $\{h_i\}_{i=1}^{M}$ on its diagonal and has zeros elsewhere, and $Gam_{h_i}[\vartheta/2, \vartheta/2]$ is the gamma probability distribution function with parameters $(\vartheta/2, \vartheta/2)$. Using this approximation, the *a posteriori* pdf could be written as

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) &= N_{\boldsymbol{x}}\left(\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}\right) \int N_{\boldsymbol{x}}\left(\boldsymbol{0}, \boldsymbol{H}^{-1}\right) \prod_{i=1}^{M} Gam_{h_i}\left[\frac{\vartheta}{2}, \frac{\vartheta}{2}\right] d\boldsymbol{H} \\ &= \int N_{\boldsymbol{x}}\left(\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}\right) N_{\boldsymbol{x}}\left(\boldsymbol{0}, \boldsymbol{H}^{-1}\right) \prod_{i=1}^{M} Gam_{h_i}\left[\frac{\vartheta}{2}, \frac{\vartheta}{2}\right] d\boldsymbol{H}. \end{aligned} \tag{35}$$

The product of two Gaussian distributions is also a Gaussian distribution [35],

$$N_{\boldsymbol{x}}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) N_{\boldsymbol{x}}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right) = k.N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{36}$$

where the mean and covariance $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the new Gaussian distribution in Eq. (36) is given by,

$$\boldsymbol{\mu} = \left({\boldsymbol{\Sigma}_1}^{-1} + {\boldsymbol{\Sigma}_2}^{-1}\right)^{-1}\left({\boldsymbol{\Sigma}_1}^{-1}\boldsymbol{\mu}_1 + {\boldsymbol{\Sigma}_2}^{-1}\boldsymbol{\mu}_2\right) \text{and}\, \boldsymbol{\Sigma} = \left({\boldsymbol{\Sigma}_1}^{-1} + {\boldsymbol{\Sigma}_2}^{-1}\right)^{-1}, \tag{37}$$

and k is a constant. Therefore, we could simplify the product of two the Gaussian distributions given in Eq. (35) as,

$$N_{\boldsymbol{x}}\left(\boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}\right).N_{\boldsymbol{x}}\left(\boldsymbol{0}, \boldsymbol{H}^{-1}\right) = k.N_{\boldsymbol{x}}\left(\left(\sigma^{-2}I + \boldsymbol{H}\right)^{-1}\left(\sigma^{-2}\boldsymbol{A}\boldsymbol{x}\right), \left(\sigma^{-2}I + \boldsymbol{H}\right)^{-1}\right). \tag{38}$$

From Eqs. (35) and (38) we could write $p(\boldsymbol{x}|\boldsymbol{y})$ as,

$$p(\boldsymbol{x}|\boldsymbol{y}) = k \int N_{\boldsymbol{x}}\left(\left(\sigma^{-2}I + \boldsymbol{H}\right)^{-1}\left(\sigma^{-2}\boldsymbol{A}\boldsymbol{x}\right), \left(\sigma^{-2}I + \boldsymbol{H}\right)^{-1}\right) \prod_{i=1}^{M} Gam_{h_i}\left[\frac{\vartheta}{2}, \frac{\vartheta}{2}\right] d\boldsymbol{H}. \tag{39}$$

We still could not compute the integral in Eq. (39) in closed form. However, we could maximize the RHS of Eq. (39) over the hidden variables $\boldsymbol{H}$ to obtain an approximation for the *a posteriori* probability distribution function

$$p(\boldsymbol{x}|\boldsymbol{y}) \approx \arg\max_{\boldsymbol{H}} \left[N_{\boldsymbol{x}} \left(\left(\sigma^{-2} I + \boldsymbol{H}\right)^{-1} \left(\sigma^{-2} \boldsymbol{A}\boldsymbol{x}\right), \left(\sigma^{-2} I + \boldsymbol{H}\right)^{-1} \right) \prod_{i=1}^{M} Gam_{h_i} \left[\frac{\vartheta}{2}, \frac{\vartheta}{2} \right] \right]. \quad (40)$$

Eq. (40) would be a good approximation of $p(\boldsymbol{x}|\boldsymbol{y})$, if the actual distribution over the hidden variables is concentrated tightly around its mode [14]. When h_i has a large value, its corresponding ith component of the *a priori* probability distribution function $p(\boldsymbol{x})$ would have a small variance, $\frac{1}{h_i}$, so that this ith component of $p(\boldsymbol{x})$ could be set to zero. Therefore, this ith dimension of the prior $p(\boldsymbol{x})$ would not contribute to the solution of Eq. (30), thus increasing its sparsity.

Since both Gaussian and gamma pdfs in Eq. (40) are members of the exponential family of probability distributions, we could obtain $\widehat{\boldsymbol{x}}_{MAP}$ by maximizing the sum of their logarithms. Section 3.5 in [11] and Section 8.6 in [14] describe an iterative optimization method to obtain $\widehat{\boldsymbol{x}}_{MAP}$ from the approximate *a posteriori* probability distribution function given by Eq. (40).

7. Conclusion

In this chapter, we described different methods to estimate an unknown signal from its linear measurements. We focused on the underdetermined case where the number of measurements is less than the dimension of the unknown signal. We introduced the concept of signal sparsity and described how it could be used as prior information for either regularized least squares or Bayesian signal estimation. We discussed compressed sensing and sparse signal representation as examples where these sparse signal estimation methods could be applied.

Author details

Ishan Wickramasingha[1], Michael Sobhy[2] and Sherif S. Sherif[1]*

*Address all correspondence to: sherif.sherif@umanitoba.ca

1 Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada

2 Biomedical Engineering Graduate Program, University of Manitoba, Winnipeg, Canada

References

[1] Keesman KJ. System Identification: An Introduction. London: Springer Science & Business Media; 2011

[2] Von Bertalanffy L. General system theory. New York. 1968;**41973**(1968):40

[3] Chen C-T. Linear System Theory and Design. New York, NY: Oxford University Press, Inc.; 1999

[4] Moon TK, Stirling WC. Mathematical Methods and Algorithms for Signal Processing. Upper Saddle River, NJ: Prentice Hall; 2000

[5] Fieguth P. Statistical Image Processing and Multidimensional Modeling. New York, NY: Springer Science+Business Media, LLC; 2011

[6] Tarantola A. Inverse problem theory and methods for model parameter estimation. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2005. p. 1-37

[7] Ljung L. Perspectives on system identification. Annual Reviews in Control. 2010 Apr 30;**34**(1):1-2

[8] Wellstead PE. Non-parametric methods of system identification. Automatica. 1981 Jan 1;**17**(1):55-69

[9] Shanmugan KS, Breipohl AM. Random Signals: Detection, Estimation, and Data Analysis. New York, NY: Wiley; 1997

[10] Saad Y. Iterative Methods for Sparse Linear Systems. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003

[11] Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer; 2006

[12] Mendel JM. Lessons in Estimation Theory for Signal Processing, Communications, and Control. Englewood Cliffs, N.J.: Prentice-Hall; 1995

[13] Sorenson HW. Least-squares estimation: From Gauss to Kalman. IEEE Spectrum. 1970 Jul;**7**(7):63-68

[14] Prince SJ. Computer Vision: Models, Learning, and Inference. Cambridge: Cambridge University Press; 2012

[15] Mallat S. A Wavelet Tour of Signal Processing: The Sparse Way. Amsterdam: Academic Press; 2009

[16] Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. SIAM Review. 2001;**43**(1):129-159

[17] Paul R. Halmos. Finite-Dimensional Vector Spaces. Mineola, UNITED STATES: Dover Publications; 2017

[18] Mallat S. A Wavelet Tour of Signal Processing. San Diego: Academic Press; 1999

[19] Shannon CE. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 2001;**5**(1):3-55

[20] Eldar YC, Kutyniok G, editors. Compressed Sensing: Theory and Applications. Cambridge: Cambridge University Press; 2012

[21] Asif MS. Dynamic compressive sensing: Sparse recovery algorithms for streaming signals and video [Doctoral dissertation]. Georgia Institute of Technology

[22] Huang K, Aviyente S. Sparse representation for signal classification. In: NIPS. Vol. 19; 2006. pp. 609-616

[23] Poggio T, Torre V, Koch C. Computational vision and regularization theory. Nature. 1985 Sep 26;**317**(6035):314-319

[24] Tikhonov AN, Arsenin VI. Solutions of Ill-posed Problems. Washington, DC: Winston; 1977 Jan

[25] Wahba G, Wendelberger J. Some new mathematical methods for variational objective analysis using splines and cross validation. Monthly Weather Review. 1980 Aug;**108**(8): 1122-1143

[26] Lin Y, Lee DD. Bayesian L1-Norm Sparse Learning. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Toulouse, France: vol. 5; 2006. p. V–V.

[27] Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing. 1993 Dec;**41**(12):3397-3415

[28] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004 Apr;**32**(2):407-499

[29] Daubechies I. Time-frequency localization operators: A geometric phase space approach. IEEE Transactions on Information Theory. 1988 Jul;**34**(4):605-612

[30] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996 Jan 1;**58**(1):267–288

[31] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006 Feb 1;**68**(1): 49-67

[32] Coifman RR, Wickerhauser MV. Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory. 1992 Mar;**38**(2):713-718

[33] Rao BD, Kreutz-Delgado K. An affine scaling methodology for best basis selection. IEEE Transactions on Signal Processing. 1999 Jan;**47**(1):187-200

[34] Boyd S, Vandenberghe L. Convex Optimization. New York: Cambridge University Press; 2004.

[35] Bromiley P. Products and convolutions of Gaussian probability density functions. Tina-Vision Memo. 2003;**3**(4):1–13.