Compressed sensing

Robust recovery of sparse signals from limited measurements

ANZIAM 2008

6 February, 2008

Terence Tao (UCLA)

Linear measurement

A classic problem in linear algebra is to solve the equation

$$Ax = b$$

where

- $x \in \mathbf{R}^n$ or $x \in \mathbf{C}^n$ is an n-dimensional unknown vector;
- $b \in \mathbf{R}^m$ or $b \in \mathbf{C}^m$ is a vector of m linear measurements; and
- A is a known $m \times n$ matrix, which we will assume to be full rank.

In practice, we might also consider noisy models such as $b = Ax + z$ where z is a Gaussian noise vector, or $b = Ax + e$ where e is a sparse corruption vector.

For our intended applications, one should think of n and m as being moderately large, e.g. between 10^3 and 10^6.

When the number of measurements m is greater than or equal to the number of degrees of freedom n, the problem $Ax = b$ is overdetermined or determined and the problem is easily solved.

When the number of measurements m is less than n, the problem is underdetermined, and x lies on an $n - m$-dimensional subspace. If one assumes that x is likely to have as small an energy $\|x\|_{l^2}$ as possible, one can propose the least squares solution

$$x^{\#} := \operatorname{argmin}_{x:Ax=b} \|x\|_{l^2} = A^*(AA^*)^{-1}b$$

as the “best” guess for x.

However, in many situations the least squares solution is not satisfactory. For instance, consider the problem of reconstructing a one-dimensional discrete signal $f : \{1, \dots, n\} \to \mathbb{C}$ from a partial collection $\hat{f}(\xi_1), \dots, \hat{f}(\xi_m)$ of Fourier coefficients

$$\hat{f}(\xi) := \frac{1}{n} \sum_{x=1}^{n} e^{-2\pi i x \xi} f(x).$$

The least squares solution $f^\#$ to this problem is easily seen to be the partial Fourier series

$$f^\# := \sum_{j=1}^{m} \hat{f}(\xi_j) e^{2\pi i x \xi_j}$$

which, when m is small, is often very different from the original signal f, especially if f is "spiky" (consider for instance a delta function signal).

It is thus of interest to obtain a good estimator for underdetermined problems such as $Ax = b$ in the case in which x is expected to be "spiky" - that is, concentrated in only a few of its coordinates. A model case occurs when x is known to be S-sparse for some $1 \leq S \leq n$, which means that at most S of the coefficients of x can be non-zero.

A typical example of when this assumption is reasonable is in imaging. An image may consist of $\sim 10^6$ pixels and thus require a vector of $n \sim 10^6$ to fully represent. But, if expressed in a suitable wavelet basis, and the image does not contain much noise or texture, only a small fraction (e.g. 10^4) of the wavelet coefficients should be significant. (This is the basis behind several image compression algorithms, e.g. JPEG2000.)

Intuitively, an S-sparse vector x has only S degrees of freedom, and so one should now be able to reconstruct x using only S or so measurements. This is the philosophy of compressed sensing (or compressive sensing, or compressive sampling): the number of measurements needed to accurately capture an object should be comparable to its compressed size, not its uncompressed size.

Compressed sensing is advantageous whenever

- signals are sparse in a known basis;
- measurements (or computation at the sensor end) are expensive; but
- computations at the receiver end are cheap.

Such situations can arise in

- Imaging (e.g. the "single-pixel camera")
- Sensor networks
- MRI
- Astronomy
- ...

Is compressed sensing even possible?

It is easy to see that the answer is yes:

> **Lemma.** Suppose that the $m \times n$ measurement matrix A is such that every set of $2S$ columns of A are linearly independent. Then an S-sparse vector $x \in \mathbf{C}^n$ can be reconstructed uniquely from $Ax \in \mathbf{C}^m$.

In principle, this shows that one can sense S-sparse vectors accurately with as few as $m = 2S$ measurements.

Proof. Suppose unique reconstruction failed; then there would exist two S-sparse vectors $x, x' \in \mathbf{C}^n$ such that $Ax = Ax'$. But then $A(x - x') = 0$. Since $x - x'$ is $2S$-sparse, this means that $2S$ of the columns are linearly dependent, contradiction. $\square$

The above argument shows that x is in fact the sparsest solution to $Ax = b$, i.e.

$$x = \text{argmin}_{x:Ax=b} \|x\|_0$$

where $\|x\|_0 = \sum_{i=1}^{n} |x_i|^0 = \#\{1 \leq i \leq n : x_i \neq 0\}$ is the sparsity of x. This should be compared with the least squares solution

$$x^{\#} = \text{argmin}_{x:Ax=b} \|x\|_2.$$

Unfortunately, the problem of finding the sparsest solution

$$\text{argmin}_{x:Ax=b} \|x\|_0$$

to a linear system $Ax = b$ is computationally expensive in general (in fact this problem contains the *subset-sum problem* as a special case, and is thus NP-complete!). Brute-force methods, such as looping over all $\binom{n}{S}$ possible collections of S columns of A, and solving a separate linear algebra problem for each such collection, are clearly impractical.

Basis pursuit

It turns out that for "generic" choices of matrix A, one can resolve this difficulty by replacing the non-convex norm $\|\|_0$ by its convex relaxation $\|\|_1$, thus solving the basis pursuit problem

$$x^* := \text{argmin}_{x:Ax=b} \|x\|_1.$$

By using standard linear programming tools, this problem is computationally feasible for $n, m \sim 10^6$.

Basis pursuit was introduced empirically in the sciences (e.g. in seismology by Claerbout-Muir and others) in the 1970s, and then studied mathematically in the 1990s by by Chen, Donoho, Huo, Logan, Saunders, and others. Near-optimal performance guarantees emerged in the 2000s by Candés-Romberg-Tao, Donoho, and others.

Remarkably, x^* recovers the sparsest solution exactly for many choices of matrix A, if we make the number of measurements of m slightly larger than the sparsity S! For instance:

> Theorem (Candés-Romberg-T. 2004). Let $\xi_1, \ldots, \xi_m \in \{1, \ldots, n\}$ be chosen randomly. Then with high probability, every S-sparse signal $f : \{1, \ldots, n\} \to \mathbf{C}$ can be recovered from $\hat{f}(\xi_1), \ldots, \hat{f}(\xi_m)$ via basis pursuit, so long as $m \geq CS \log n$ for some absolute constant C.

Numerical experiments suggest that most S-sparse signals are in fact recovered exactly once $m \geq 4S$ or so.

Theorem (Donoho 2004, Candés-T. 2004). Suppose that the entries of A are iid Gaussians (or Bernoulli signs ± 1). Then any given S-sparse signal x can be recovered from Ax by basis pursuit with high probability as long as $m \geq CS \log \frac{n}{S}$.

Theorem (Candés-T. 2005). Suppose that A obeys the restricted isometry property (RIP): every collection of $4S$ columns are almost orthogonal, in that the top $4S$ singular values range between 0.9 and 1.1. Then any given S-sparse signal x can be recovered from Ax by basis pursuit.

This particular result is elementary, using nothing more sophisticated than the triangle inequality and the Cauchy-Schwarz inequality.

The RIP has been shown to hold for many random matrix ensembles, as long as m is logarithmically larger than S. (Donoho, Candés-Tao, Rudelson-Vershynin, Mendelson-Pajor-Tomczak-Jaegermann, ...)

More generally, it seems that compressed sensing works whenever the measurement matrix is sufficiently "incoherent", in that the measurement basis is radically different from the basis in which the signal is sparse. (If the measurement basis aligned too closely with the sparsity basis, it is possible that sparse signals might not be detected at all by taking just a few measurements.)

Noise

Now suppose we are given noisy measurements $b = Ax + z$, where z is a noise vector of size $\|z\|_{l^2} \leq \varepsilon$. To recover x approximately from b, one can now solve the modified basis pursuit problem

$$x^* := \text{argmin}_{x:\|Ax-b\|_2 \leq \varepsilon} \|x\|_1.$$

If z is adversarially chosen, one cannot expect the accuracy $\|x - x^*\|_{l^2}$ to be much better than ε. Remarkably, though, this bound is attained up to constants!

Theorem (Candés-T. 2005). Suppose that A obeys the RIP. Then for any S-sparse x and any $b := Ax + z$ with $\|z\|_{l^2} < \varepsilon$, the basis pursuit solution x^* obeys $\|x^* - x\|_{l^2} \leq C\varepsilon$.

One can also do slightly better when z is non-adversarial (e.g. Gaussian white noise), leading to a near-optimal statistical selector for x from $Ax + z$.

Tails

In many applications, the hypothesis that x is S-sparse is unrealistic. A more reasonable assumption is that the coefficient magnitudes of x are distributed by some sort of power law, or more generally that x is compressible. In this case, basis pursuit gives a solution x^* comparable in accuracy to the hard-thresholded vector x_S, defined as the vector consisting of the S largest coefficients of x (with all other coefficients zero).

More precisely:

Theorem (Candés 2006). Suppose that A obeys the RIP. Then for any x, the minimiser $x^* := \mathrm{argmin}_{x':Ax'=Ax} \|x'\|_{l^1}$ obtained by basis pursuit obeys $\|x^* - x\|_{l^1} \leq 4\|x_S - x\|_{l^1}$.

In the case of power law decay (e.g. the k^{th} largest coefficient of x decays like k^{-c} for some $c > 1$) one can also obtain further l^2 error bounds (see Candés-Tao 2004).

Linear coding

We have talked about compressed sensing, in which the signal is sparse, the noise is non-sparse, and one has fewer measurements m than degrees of freedom n. But there is also an interesting dual situation in which the signal is non-sparse, one takes more measurements m than degrees of freedom n, but now the error is sparse. This situation arises in linear coding, in which one converts data $x \in \mathbb{R}^n$ to a longer string $Ax \in \mathbb{R}^m$ to transmit in m packets over a network. Assuming that some small proportion of these packets get corrupted, the receiver obtains a vector $Ax + e$ for some unknown sparse error e.

It turns out that the problem of recovering x from $y := Ax + e$ is a compressed sensing problem in disguise. Let B be an annihilator to A, thus $BA = 0$. Then from $y = Ax + e$ we have $By = Be$. One can then use compressed sensing methods to recover the sparse error e from $By = Be$, and then reconstruct x from $y = Ax + e$ and e by linear algebra.

(In digital applications, one would use a binary field $\mathbb{F}_2$ instead of $\mathbb{R}$. In such cases, compressed sensing methods do not work well; but other methods are available in this case, such as turbo codes.)

Other algorithms

There are several other algorithms for compressed sensing nowadays, besides basis pursuit. For instance, there is matching pursuit, in which individual basis elements are located which "match" (or "correlate" with the measurements, and then are projected out (e.g. via orthogonal projection), and the process repeated. There are also several hybrid strategies combining basis and matching pursuit.

Generally speaking, matching pursuit is faster, but has fewer proven guarantees regarding robustness with respect to noise or tails.

In situations in which the measurement matrix can be prescribed to one's specifications, extremely fast, robust, and low-memory algorithms are available (Gilbert, Strauss, Tropp, Muthukrishnan, Vershynin, ...).

There is much active research in optimising these algorithms and extending to other contexts (e.g. streaming data, matrix-valued or nonlinear measurements, etc.).