An Invitation to Compressive Sensing

I-Liang Chern

August 24, 2016

Outline of applications [1]

- Sampling Theory
- Sparse Approximation
- Error Correction
- Statistics and Machine Learning
- Low-Rank Matrix Recovery and Matrix Completion
- ...
- See more from Compressive Sensing Resources

[1] This part of the note is mainly copied from: [1] Simon Foucart Holger Rauhut, A Mathematical Introduction to Compressive Sensing, Springer 2013.

1. Sampling Theory

- Sampling theory is to reconstruct a continuous-time signal from a discrete set of samples.
- Examples include image processing, sensor technology in general, and analog-to-digital conversion appearing in audio entertainment systems or mobile communication devices [1].
- Sampling theory is based on the Shannon sampling Theorem:

Theorem (Shannon)

If f is banded limited, that is, Supp$\hat{f} \subset [-B, B]$, then f can be represented by $f(t) = \sum_{k \in \mathbb{Z}} f\left(k\frac{\pi}{B}\right)$ sinc$(Bt - k\pi)$.

Remark. Sample density (or Nyquist rate): $\frac{|\text{supp } \hat{f}|}{2\pi} = B/\pi$.

Proof.

WLOG, we assume $B = \pi$. Thus, $\hat{f}$ is a function on $[-\pi, \pi]$ which can be expanded as

$$\hat{f}(\xi) = \sum_{k \in \mathbb{Z}} c_k e^{-ik\xi}.$$

where

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\xi} \hat{f}(\xi) \, d\xi = \frac{1}{\sqrt{2\pi}} f(k).$$

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{it\xi} \hat{f}(\xi) \, d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{it\xi} \sum_{k \in \mathbb{Z}} c_k e^{-ik\xi} \, d\xi \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k \in \mathbb{Z}} c_k \int_{-\pi}^{\pi} e^{i(t-k)\xi} \, d\xi = \sum_{k \in \mathbb{Z}} f(k) \frac{\sin \pi(t-k)}{\pi(t-k)} \end{aligned}$$

- For periodic function $f(t)$ with period 1 and band width M, it can be represented as

$$f(t) = \sum_{k=-M}^{M} \hat{f}_k e^{2\pi ikt}$$

- $\hat{f}_k, k = -M, ..., M$ discrete, periodic, can be represented as (discrete Fourier)

$$\hat{f}_k = \sum_{j=0}^{2M} c_j e^{2\pi ikj/(2M+1)},$$

$$c_j = \sum_{k=-M}^{M} \hat{f}_k e^{-2\pi kj/(2M+1)} = f\left(\frac{j}{2M+1}\right).$$

- Thus, f can also be represented by

$$f(t) = \frac{1}{2M+1} \sum_{j=0}^{2M} f\left(\frac{j}{2M+1}\right) D_M\left(t - \frac{j}{2M+1}\right).$$

where $D_M(t) = \sum_{k=-M}^{M} e^{2\pi ikt} = \frac{\sin(2M+1)\pi t}{\sin(\pi t)}$.

- In practice, $M >> 1$, but $\hat{f}$ may only have sparse s modes, i.e. $\|\hat{f}\|_0 = s$
- Question: can we take fewer samples in time-domain to reconstruct f exactly?

CS formulation for compressed sampling

Answer:

- Choose $t_1, ..., t_m$ independently and randomly, uniformly distributed on $[0, 1]$.
- We sample f at t_ℓ, $\ell = 1, ..., m$:

$$y_\ell = f(t_\ell) = \sum_k e^{2\pi i k t_\ell} \hat{f}_k = \sum_k A_{\ell,k} x_k, \ell = 1, ..., m$$

$$A_{\ell,k} = e^{2\pi i k t_\ell}, \ell = 1, ..., m,\ \ k = -M, ..., M,$$

- The CS problem is

$$\min \|x\|_0 \text{ subject to } Ax = y.$$

- CS: One can reconstruct f with high probability from its m samples $f(t_1), ..., f(t_m)$ provided that $m \geq Cs \ln(N)$.

- E Candes, J Romberg, T Tao, Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information, 2004

2. Sparse Approximation for images

- In image processing, we would like to represent an image y in terms of some elements $\{a_1, ..., a_N\}$

$$y = \sum_{j=1}^{N} x_j a_j = [a_1, ..., a_N]x.$$

- The elements a_j are called atoms, the set $\{a_1, ..., a_N\}$ called dictionary. Usually, they are redundant.
- Examples are wavelets + Fourier basis or + Gabor basis.
- We usually assume that the image is sparse in this dictionary. That is, we want to find

$$\min \|x\|_0 \text{ subject to } Ax = y.$$

CS application to image processing

- Compression: using the sparse representation

$$\min \|x\|_0 \quad \text{subject to } y = Ax,$$

we only store those s significant coefficients $x_i's$.

- Denoising: Suppose $y = Ax + e$, e is the noise with $\|e\| \leq \eta$, and Ax is the sparse presentation. We can reconstruct image x (noiseless) by

$$\min \|x\|_0 \quad \text{subject to } \|Ax - y\| \leq \eta.$$

3. Data separation

- Suppose $y = y_1 + y_2$ with

$$y_1 = Ax_1,\ y_2 = Bx_2,$$

where $A = [a_1, ..., a_{N_1}]$ $B = [b_1, ..., b_{N_2}]$ be the dictionaries with different nature. For instance, sinunoidal and spikes.

- Assumption: x_1 and x_2 are sparse.
- Sparse reconstruction: we can reconstruct x_i from

$$\min_{x_1, x_2} \|x_1\|_0 + \|x_2\|_0 \text{ subject to } y = Ax_1 + Bx_2.$$

4. Error correction

- Suppose we need to transmit a vector $z \in \mathbb{R}^n$ and we want to design a method to correct transmittion errors.
- Method:
  1. Add redundancy: we convert $z \to v = Bz \in \mathbb{R}^N$, $N = n + m$.
  2. Transmit v but it may have error x. We receive $w = v + x$.
     Assumption: the error x is sparse.
  3. We choose $A \in \mathbb{R}^{m \times N}$ such that $AB = 0$
  4. We measure $y = Aw = A(v + x) = ABz + Ax = Ax$.
  5. We obtain x by solving

$$\min \|x\|_0 \quad \text{subject to } y = Ax.$$

  6. We recover z by solving overdetermined system $Bz = v = w - x$. (E.g. solving $B^T Bz = B^T v$)
  7. Usually, we choose A random Fourier modes, B a complement of A. In this case, $AB = 0$.

5. Statistics

- The goal of statistical regression is to predict an outcome based on certain input data. It is common to choose the linear model

$$y = Ax + e$$

  Here, A is the predictor matrix, collected from input, y output data, and e the noise. x is the parameter to be estimated.
- Example: A_{jk} is the clinical data of the jth patient such as blood pressure, weight, DNA data, etc. whereas y_j is the probability that the jth patient suffers a certain disease, x is the parameters to fit.

- In many cases, however, only a small number of parameters contribute towards the effect to be predicted. This leads to sparsity of x.
- In statistical terms, determining a sparse parameter vector x corresponds to selecting the relevant explanatory variables, i.e., the support of x. One also speaks of model selection.

- In statistics, one usually considers the so-called LASSO (least absolute shrinkage and selection operator)

$$\min \|Az - y\|_2 \text{ subject to } \|z\|_1 \leq \tau.$$

- Or the Dantzig selector

$$\min \|z\|_1 \text{ subject to } \|A^*(Az - y)\|_\infty \leq \lambda.$$

- Or Basis pursuit (also called LASSO)

$$\min \lambda\|z\|_1 + \|Az - y\|_2^2.$$

6. Low-Rank Matrix Recovery and Matrix Completion

- Recover a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ from incomplete information.
- Sparsity is replaced by the assumption that X has low rank.
- Singular value decomposition of X:

$$\mathbf{X} = \sum_{\ell=1}^{n} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^*, \; n = \min\{n_1, n_2\}.$$

$$\text{Rank } \mathbf{X} = \|\sigma(\mathbf{X})\|_0.$$

Netflex problem: a matrix completion problem

- Given an incomplete (rating) matrix whose row is the videos indices and whose column is the customers. The entry X_{jk} is the rating of the jth video by the kth customer.
- Often, if two customers both like some subset of products, then they will also both like or dislike other subsets of products (the types of customers are essentially limited). It is therefore assumed the rating matrix has low rank.
- The given data $\mathbf{Y}$ is incomplete, not every customer rates every video. The support of $\mathbf{Y}$ is only on Λ. Thus the problem is

$$\min \text{Rank } (\mathbf{X}) \text{ subject to } \mathbf{X}_{ij} = \mathbf{Y}_{ij} \text{ for } (i,j) \in \Lambda.$$

- It can also be approximated by

$$\min \|\mathbf{X}\|_* \text{ subject to } \mathbf{X}_{ij} = \mathbf{Y}_{ij} \text{ for } (i,j) \in \Lambda.$$

- Where, the nuclear norm is defined by

$$\|\mathbf{X}\|_* := \|\sigma(\mathbf{X})\|_1.$$