Astronomical Data Compression: Algorithms & Architectures

Rob Seaman – National Optical Astronomy Observatory
William Pence – NASA / Goddard Space Flight Center
Richard White – Space Telescope Science Institute
Séverin Gaudet – National Research Council Canada

Agenda

• Overview – Rob
• Tile compression and CFITSIO – Bill
• Experiences with FITS compression in a large astronomical archive – Séverin
• Lossy compression – Rick
• Open discussion
• Door prize!

Thanks to Pete Marenfeld & Koji Mukai

5 October 2009
Overview

- FITS tile compression
- Rice algorithm
- CFITSIO / FPACK
- IRAF and community software
- The ubiquity of noise: optimal DN encoding
- The role of sparsity: compressive sensing
- An information theory example
Overview

- FITS tile compression
- Rice algorithm
- CFITSIO / FPACK
- IRAF and community software
- The ubiquity of noise: optimal DN encoding
- The role of sparsity: compressive sensing
- An information theory example
References

• Too many ADASS presentations to list

• See references within:
  
  “Lossless Astronomical Image Compression and the Effects of Noise”, Pence, Seaman & White, PASP v121 n878 2009, 
  http://arxiv.org/abs/0903.2140v1
FITS tile compression

• ADASS 1999 (Pence, White, Greenfield, Tody)
• FITS Convention v2.1, 2009
• Images mapped onto FITS binary tables
• Headers remain readable
• Tiling permits rapid RW access
• Supports multiple compression algorithms
• First & every copy can be compressed
Rice algorithm

• Fast (difference coding)
  – near optimum compression ratio
  – throughput is key, not just storage volume
• Numerical, not character-based like gzip
• Depends on pixel value so BITPIX = 32
  compresses to same size as BITPIX = 16
• fpack can be swapped in for gzip & funpack for gunzip
• Library support (eg, CFITSIO) allows jpeg-like access – compression built into the format
• More options means more parameters – setting appropriate defaults is key
IRAF and community software

- Tile compression can & should be supported by all software that *reads* FITS
- Instrument and pipeline software may benefit strongly from *writing* compressed FITS
- Transport & storage always benefit
- IRAF fitsutil package in beta testing
- Work on a new IRAF FITS kernel pending
- VO applications and services

5 October 2009
The ubiquity of noise

- Noise is incompressible
- Signals are correlated
  - physically
  - instrumentally
- Shannon entropy: \( H = - \Sigma p \log p \)
  - depends only on the probabilities of the states
  - measures “irreducible complexity” of the data
Optimal DN encoding

- CCD “square-rooting”
- Variance stabilization, more generally
  - many statistical methods assume homoscedasticity
  - generalized Anscombe transform
- Foundations of the empirical world view:
  - ergodicity (statistical homogeneity)
  - Markov processes (memoryless systems)

The role of sparsity

• For most astronomical data, compression ratio depends *only* on the background noise
  – Sparse signals are negligible (in whatever axes)
  – Noise is incompressible

\[
R = \frac{\text{BITPIX}}{(N_{\text{bits}} + K)}
\]

K is about 1.2 for Rice
Compressional ratio correlates closely with noise

Distinctive functional behavior

For three very different comp. algorithms

For flat-field and bias exposures as well as for science data

That is, for pictures of: the sky, a lamp in the dome
no exposure at all

Signal doesn’t matter!

天文データの圧縮
A better compression diagram

\[ R = \frac{\text{BITPIX}}{N_{\text{bits}} + K} \]

**Effective BITPIX:**

\[ \text{BIT}_{\text{EFF}} = \frac{\text{BITPIX}}{R} \]

\[ \text{BIT}_{\text{EFF}} = N_{\text{bits}} + K \]

Line with:
- **Slope** = 1
- **Intercept** = K

GZIP Rice Hcompress

**Lossy algorithms**

shift left to move down

**Limit for perfect algorithm**

**Margin for improvement**

Compressed bits per pixel

N_{\text{bits}} Noise
Compressive sensing

• Real world data are often sparse (*correlated*)
• Nyquist/Shannon sampling applies broadly
• But we can do even better if we sample against purpose-specific axes:
  
  [http://www dsp ece rice edu/cs](http://www.dsp.ece.rice.edu/cs)
  
  [http://nuit-blanche.blogspot.com](http://nuit-blanche.blogspot.com)

• Herschel proof of concept, Starck, et al.
• CS is about the sampling theorem
• Optimal encoding is about quantization
An information theory example

Find the fake coin in 3 measurements

Putting the same number of coins on each side of the scale constitutes a measurement.
To Win, use the Ankh to verify the heavy coin or the Feather to verify the light coin.

Find the fake coin in 3 measurements


1A 2B 3C 4D 5E 6F 7G 8H 9I 10J 11K 12L

http://www.mapsofconsciousness.com/12coins

If you like this game send feedback email: Joe

天文データの圧縮
Compression = optimal representation

A. 11 coins all the same
   + 1 coin, identical except for weight

B. Scale to weigh groups of coins

C. In only 3 steps, must identify:
   the coin that is different and
   whether it is light or heavy

“The 12-balls Problem as an Illustration of the Application of Information Theory”
How to solve a problem

• First, define the problem
  – second, entertain solutions
  – third, iterate (don’t give up)

• More basic yet, what is the goal?
  – to solve the problem?
  – or to understand how to solve it?

• Stating a problem constrains its solutions
What do we know?

• One bit discriminates two equally likely alternatives
  To select between $N$ equal choices, $N_{\text{bits}} = \log_2 N$

• For 12-coin problem, $N_{\text{bits}} = \log_2 (12) + 1 = \log_2 24$
  (must also distinguish light vs. heavy)

• Information provided in each measurement is $\log_2 3$
  (3 positions for scale: left, right, balanced)

• For three weighings, $W_{\text{bits}} = \log_2 3^3 = \log_2 27$
  Meets necessary condition that $W_{\text{bits}} \geq N_{\text{bits}}$
Necessary, but not sufficient

• A strategy is also necessary such that
  \[ W_{\text{bits}} \geq N_{\text{bits}} \ (\text{remaining}) \]
  is satisfied at each step to the solution

• \( N_{\text{bits}} \) is the same thing as the entropy \( H \)

\[
H = - \sum p \log p \quad \text{where} \quad p = \frac{1}{N} \\
= - \sum \left(\frac{1}{N}\right) \log \left(\frac{1}{N}\right) = \left(\sum \left(\frac{1}{N}\right)\right) \log N = \log N \\
H = \log_2 N \quad \text{(in bits)}
\]
What else do we know?

• Physical priors!
  – only one coin is fake
  – astronomical data occupy sparse phase space

• FITS arrays = images (physical priors)
  – of astrophysical sources
  – taken through physical optics
  – recorded by physical electronics
  – digitization is restricted by information theory
  – possessing a distinctive noise model

5 October 2009
Putting the same number of coins on each side of the scale constitutes a measurement.

To Win, use the Ankh to verify the heavy coin or the Feather to verify the light coin.
2 of 3 measurements used.

Putting the same number of coins on each side of the scale constitutes a measurement.

To Win, use the Ankh to verify the heavy coin or the Feather to verify the light coin.
Putting the same number of coins on each side of the scale constitutes a measurement.
To Win, use the Ankh to verify the heavy coin or the Feather to verify the light coin.
You Win! 12 Coins in 3 of 3 measurements!
2094 seconds.

Putting the same number of coins on each side of the scale constitutes a measurement.
To Win, use the Ankh to verify the heavy coin or the Feather to verify the light coin.
Observations about observations

- The sequence of three measurements can occur in any order
- The systematization of the solution occurs during its definition, not at run time
Try it yourself

http://heasarc.gsfc.nasa.gov/fitsio/fpack

http://www.mapsofconsciousness.com/12coins