Cosmology and its Data Flood Challenge

Jean-Luc Starck

http://jstarck.cosmostat.org

CEA, IRFU, AIM, Service d'Astrophysique, France,

Samual Farrens    Francois Lanusse    Sandrine Pires    Fred Ngolè
Cosmology and its Data Flood Challenge

• Introduction to the standard cosmological model.
  Three Sources of Uncertainty:
  ➡ Stochastics
  ➡ Systematics
  ➡ Approximations

• Cosmology in the Big Data era
  ➡ Big Data Today
  ➡ Big Data Tomorrow

• Euclid Space Mission
  ➡ Point Spread Function and Deconvolution
  ➡ Dark matter mass mapping.
The Standard Cosmological Model

GW
Lensing
Supernovae
Dark Matter 26.8%
Ordinary Matter 4.9%
Dark Energy 68.3%

ESA/Planck
CMB
Galaxy distribution
The Standard Cosmological Model

- Afterglow Light Pattern 375,000 yrs.
- Dark Ages
- Development of Galaxies, Planets, etc.
- Dark Energy Accelerated Expansion
- Inflation
- Quantum Fluctuations
- 1st Stars about 400 million yrs.
- Big Bang Expansion 13.77 billion years
Precision Cosmology

First Source of Uncertainty: Stochastics
- Noise
- Cosmic Variance

- New instruments with better sensitivity (hardware)
- **Collect more Data** => large survey (SDSS, WMAP, Planck, KIDS, DES, etc)
Precision Cosmology

First Source of Uncertainty: Stochastics
- Noise
- Cosmic Variance

- New instruments with better sensitivity (hardware)
- Collect more Data => large survey (SDSS, WMAP, Planck, KIDS, DES, etc)

===> Virtual Observatory (CDS Strasbourg)

- Data access, web services, interoperability, data model, etc
First Source of Uncertainty: Stochastics

- Better statistical tools (Bayesian modeling, sparsity, BSS, machine learning, etc): beyond the second order statistics

Astrophysics + Statistics/Applied math => Astrostatistics

• Two International organizations:
  ➡ International Astrostatistics Association (IAA)
  ➡ Commission on Astroinformatics and Astrostatistics within the International Astronomical Union (IAU)

• Two important U.S. national organizations:
  ➡ the Working Group in Astroinformatics and Astrostatistics within the American Astronomical Society (AAS),
  ➡ the Interest Group in Astrostatistics within the American Statistical Association (ASA).

• One project-level organization: the Informatics and Statistics Science Collaboration of the Large Synoptic Survey Telescope (LSST)

• Astrostatistics laboratories
  ➡ USA: Penn State University, Berkeley, CMU, Cornell
  ➡ Europe:
    • Imperial Center for Inference and Cosmology (ICIC) at Imperial College
    • CosmoStat laboratory, CEA-Saclay
Second Source of Uncertainty: Systematics

**BICEP2:** March 2014 - Primordial Gravitationnal Wave detection claimed by BICEP2

==> it happened to be a dust signature, dust from our own galaxy !!!
Second Source of Uncertainty: Systematics

**BICEP2:** March 2014 - Primordial Gravitationnal Wave detection claimed by BICEP2

==> it happened to be a dust signature, dust from our own galaxy !!!

Fig: As datasets grow, systematic errors swamp statistical errors and new disparities appear.

Credit: David Van Dick
Second Source of Uncertainty: Systematics

Second Source of Uncertainty: Systematics

The need of Numerical Simulations (physics + instrument)

- to test the pipeline and its ability to measure accurately the cosmological parameters.

- to build the covariance matrices that are required to fit the cosmological parameters.

Numerical simulations are a very important aspect of new big projects.
Third Source of Uncertainty: Approximation

- We now perfectly how to calculate some estimators and their covariance matrices, but the volume of data is so large that it is impossible to do it, even on HPC infrastructures.

Theoretical and algorithmic work is necessary to well control errors and biases introduced by the approximations.

Examples:

• Two point correlation functions
• Covariance matrices
• Example of ongoing work at CMU: simulate N-body simulations using machine learning.
• Approximate Bayesian Modelling (ABC) (likelihood free approach to approximating posterior where likelihood function is not specified).
The Big Data Today

GAIA space telescope

- Map the milky way in 3D
- Stellar physics
- Dark matter
- Extrasolar planets
- 50 Gbyte/day; 1 Pbyte total data product
- 3D catalogue of ~1 billion astronomical objects
The Big Data Today

GAIA space telescope

- Map the milky way in 3D
- Stellar physics
- Dark matter
- Extrasolar planets
- 50 Gbyte/day; 1 Pbyte total data product
- 3D catalogue of ~1 billion astronomical objects
The Big Data Today - LOFAR: the LOw Frequency ARray

- Giant digital & multi-purpose radio telescope distributed across Europe
- Radio interferometer composed of \( \sim 48 \) phased arrays (stations)
- Working bands: LBA 30-80 MHz & HBA 120-240 MHz
- Improved angular (arcsec), temporal (µs), spectral (kHz) resolutions
- High sensitivity (\( \sim \)mJy) \( 1 \text{ Jy} = 10^{-26} \text{ W.m}^{-2}.\text{Hz}^{-1} \)
- 24 Pbyte/day raw data - 1Pbyte per year in archive

**Map of LOFAR stations:**
- Chilbolton
- Norderstedt
- Effelsberg
- Jülich
- Tautenburg
- Unterweilenbach
- Nançay
- Potsdam
- Onsala

**Image:** NL Station
Multi-element interferometer

\[ N \] antennas/telescopes

\[ \frac{N(N-1)}{2} \] independent baselines

1 projected baseline

= 1 sample in the Fourier « u,v » plane

\[ V(u, v) = \int \int T(l, m) e^{-i2\pi(lu+vm)} \, dl \, dm \]

http://arxiv.org/abs/1406.7242
Colorsacle: reconstructed 512x512 image of Cygnus A at 151 MHz (with resolution 2.8" and a pixel size of 1"). Contours levels are [1,2,3,4,5,6,9,13,17,21,25,30,35,37,40] Jy/Beam from a 327.5 MHz Cyg A VLA image (Project AK570) at 2.5" angular resolution and a pixel size of 0.5". **Recovered features in the CS image correspond to real structures observed at higher frequencies.**
The Big Data Tomorrow - SKA: Square Kilometer Array
The Big Data Tomorrow - SKA: Square Kilometer Array
SKA Science Data Pipeline

 LOW-FREQUENCY APERTURE ARRAY

Science Data Processors
200 – 300 pflop systems

Correlator

Ingest
Calibration
Imaging

~2 Tb/s
Ave = 3.8 Tb/s

6.8 Tb/s

DISH

1 PB/day
1 x 10^8 €/yr

SKA Phase 1 is a 10% scale system of the full SKA

Credit: Melanie Johnston-Hollitt, COSMO21 conf, Chania, May 2016
What are we keeping?

1.6 x 10^{14} voxels

50K pixels

65K channels

Credit: Melanie Johnston-Hollitt, COSMO21 conf, Chania, May 2016
- **Dark matter, dark energy**, cosmology (spatial distribution of galaxies, gravitational lensing, supernovae, quasars).
- **Time domain** (cosmic explosions, variable stars).
- **The Solar System** structure (asteroids)
- **The Milky Way** structure (stars).

25 TByte per night. After 10 years, half of the sky will be imaged ~100 PB of data (10 years). Credit: Željko Ivezić, SCMA6 conf, Pittsburg, June 2016
Understand the origin of the Universe’s accelerating expansion:

- probe the properties and nature of dark energy, dark matter, gravity and distinguish their effects decisively

- by tracking their observational signatures on the
  
  - geometry of the universe: Weak Lensing + Galaxy Clustering

  - cosmic history of structure formation: WL, z-space distortion, clusters of galaxies

- Controling systematic residuals to an unprecedented level of accuracy, that cannot be reached by any other competing missions/telescopes

Gains in space:

Stable data: homogeneous data set over the whole sky

- Systematics are small, understood and controlled

- Homogeneity: Selection function perfectly controlled

~150 PB of data.
Weak Lensing
Image Forming Process

- Point source
- Circular aperture
- Observer screen

- "Unlensed" galaxy
- Lensing
- Optical PSF
- Photo-sensors integration
- Electronic noise
- Galaxy image
• Spatial variability
• Temporal variability: jitter, temperature of instrument
• Wavelength dependency
Weakly Lensed Galaxies

Galaxy

PSF

Galaxy*PSF

+ Background + Noise
Image Forming Process: Stars and Point Spread Function

• Spatial variability
• Temporal variability: jitter, temperature of instrument
• Wavelength dependency
Space Variant PSF

Hubble Ultra Deep Field 2014

NASA and ESA

STScI-PRC14-27a
Space Variant PSF
Euclid PSF Modeling

Observation model

\[ y_k = M_k x_k + n_k, \]

- **Joint estimations** of super-resolved PSFs at stars positions
  - Positivity constraint
  - Low rank constraint: Constraint the PSFs to be a linear combination of the *eigenvectors* PSFs
  - Smoothness constraint on each PSF
  - Proximity constraint: the closer are the stars, the more the coefficients of the linear combination are similar.
Monochromatic PSFs joint super-resolution

\[
X = [x_1, \ldots, x_p] = \Phi X \text{ columns sparse} = X^T = VW \text{ with } W \text{ columns sparse}
\]

Pixel domain features dictionary

Spatial frequencies dictionary
**Matrix Factorization**

\[
\mathbf{S} = [s_1, \ldots, s_p]
\]

\[
\text{PSF}^{(k)} = x_k = \sum_{i=1}^{r} a_{i,k} s_i
\]

\( s_i \) are ”eigen PSF”


\[
\min_{\alpha, \mathbf{S}} \frac{1}{2} \| \mathbf{Y} - \mathcal{F}(\mathbf{S}\alpha \mathbf{V}^T) \|_F^2 + \sum_{i=1}^{r} \| \mathbf{w}_i \odot \Phi \mathbf{s}^{(c)}_i \|_1
\]

s.t. \( \| \alpha[l, :] \|_0 \leq \eta_l, \ l = 1, \ldots, rp \) and \( \mathbf{S}\alpha \mathbf{U}^T \geq 0 \)
**Numerical Experiments**

**Data: 500 Euclid-like PSFs (Zemax), field observed with different SNRs**

Theese PSFs account for mirrors polishing imperfections, manufacturing and alignments errors and thermal stability of the telescope.

**Quality assessment : shape parameters**

\[
\gamma(X) = [e_1(X), e_2(X)]^T
\]

\[
e_1(X) = \frac{\mu_{2,0}(X) - \mu_{0,2}(X)}{\mu_{2,0}(X) + \mu_{0,2}(X)}
\]

\[
e_2(X) = \frac{2\mu_{1,1}(X)}{\mu_{2,0}(X) + \mu_{0,2}(X)}.
\]

\[
\mathbf{E}_\gamma = \sum_{i=1}^{p} \frac{\|\gamma(X_i) - \gamma(\hat{X}_i)\|_2}{p}
\]

\[
\text{Disp}_\gamma = \|\text{ME}_\gamma\|_*
\]

\[
\text{ME}_\gamma = [\gamma(X_1) - \gamma(\hat{X}_1), \ldots, \gamma(X_p) - \gamma(\hat{X}_p)]
\]
Numerical Experiments

With undersampling (upsampling factor of 2)

Center

Local

Corner

Obs  Ref  PSFEX  RCA
Numerical Experiments

With undersampling (upsampling factor of 2)

$\log_{10}(E_\gamma)$

Linear SNR
Optimal transport:

$$X_5 \approx \arg\min_X \sum_{i=1}^{4} w_i(P_5) W_2(X_i, X)^2$$
Astronomical Image Deconvolution

Standard deconvolution framework:

\[ y = Hx + n \]
Standard deconvolution framework:

\[ \mathbf{y} = \mathbf{Hx} + \mathbf{n} \]

Standard deconvolution framework:

\[
\arg \min_{\mathbf{X}} \frac{1}{2} \| \mathbf{Y} - \mathbf{HX} \|_2^2 + \| \Phi^t \mathbf{X} \|_p \quad \text{s.t.} \quad \mathbf{X} \geq 0
\]
Astronomical Image Deconvolution

Standard deconvolution framework:

\[ y = Hx + n \]

Standard deconvolution framework:

\[
\arg\min_{X} \frac{1}{2} \| Y - HX \|_2^2 + \| \Phi^T X \|_p \quad \text{s.t.} \quad X \geq 0
\]

H is huge !!!
Big Astronomical Image Deconvolution

Object Oriented Deconvolution

For each galaxy, we use the PSF related to its center pixel:

\[
Y = \mathcal{H}(X) + N
\]

\[
\arg\min_X \frac{1}{2} \| Y - \mathcal{H}(X) \|^2 + \lambda \| \Phi^t X \|_p \quad \text{s.t.} \quad X \geq 0
\]
Big Astronomical Image Deconvolution
Galaxy images have similar properties.
Big Astronomical Image Deconvolution

\[
\arg\min_X \frac{1}{2} \| \mathbf{Y} - \mathcal{H}(\mathbf{X}) \|_2^2 + \lambda \| \mathbf{X} \|_* \quad \text{s.t.} \quad \mathbf{X} \geq 0
\]

\[
\| \mathbf{X} \|_* = \sum_i \sigma_i
\]

Galaxy images have similar properties.
Algorithm: Choose the proximal parameters $\tau > 0$, $\varsigma > 0$, the positive relaxation parameter, $\xi$, and the initial estimate $(X_0, Y_0)$. Then iterate, for every $k \geq 0$.

\begin{align*}
1 : \quad & \tilde{X}_{k+1} = \text{prox}_{\tau G}(X_k - \tau \nabla F(X_k) - \tau \mathcal{L}^*(Y_k)) \\
2 : \quad & \tilde{Y}_{k+1} = Y_k + \varsigma \mathcal{L}(2\tilde{X}_{k+1} - X_k) - \varsigma \text{prox}_{K/\varsigma} \left( \frac{Y_k}{\varsigma} + \mathcal{L}(2\tilde{X}_{k+1} - X_k) \right) \\
3 : \quad & (X_{k+1}, Y_{k+1}) := \xi (\tilde{X}_{k+1}, \tilde{Y}_{k+1}) + (1 - \xi) (Y_k, Y_k)
\end{align*}
The Simulated Data

- 10,000 space-based galaxy images derived from COSMOS data.
- Each image is a $41\times41$ pixel postage stamp around the centre of the galaxy.
- Images are free from PSF effects.
The Simulated Data

- 600 spatially varying Euclid-like PSFs

- Each galaxy image is convolved with a random PSF.
- Different levels of Gaussian noise is added.

\[
\sigma = 0.01 \quad \sigma = 0.02 \quad \sigma = 0.05 \quad \sigma = 0.07 \quad \sigma = 0.10 \quad \sigma = 0.15
\]
Results

\[ X \{ \]

\[ Y \{ \]

\[ \hat{X} \{ \]

Clean Image

Data

Sparse Recovery

Low Rank

jeudi 20 octobre 16
Results

\[ P_{err} = \frac{\| X - \hat{X} \|_2}{\| X \|_2} \]

Pixel Error for 100 Galaxy Images


Code available at https://github.com/sfarrens/psf
Results

\[ P_{\text{err}} = \frac{\| X - \hat{X} \|_2}{\| X \|_2} \]

Pixel Error for 1000 Galaxy Images


Code available at  https://github.com/sfarrens/psf
Results

\[ P_{err} = \frac{\|X - \hat{X}\|_2}{\|X\|_2} \]

Pixel Error for 10000 Galaxy Images


Code available at https://github.com/sfarrens/psf
Few undesampled images of a given galaxy

Many PSF at other positions

PSF superresolution + Interpolation + Shape Measurement

Shear Catalog & Map
\[ \hat{\kappa} = P_1 \hat{\gamma}_1 + P_2 \hat{\gamma}_2 \]

\[ P_1(k) = \frac{k_1^2 - k_2^2}{k^2} \]

\[ P_2(k) = \frac{2k_1k_2}{k^2} \]
\[ \hat{\kappa} = P_1 \hat{\gamma}_1 + P_2 \hat{\gamma}_2 \]

\[ P_1(k) = \frac{k_1^2 - k_2^2}{k^2} \]

\[ P_2(k) = \frac{2k_1k_2}{k^2} \]

**Missing data** (mask and limited number densities):

**Shape noise:**
Handling Missing Data (no noise): Binning+Smoothing

Galaxy catalogue with 30 gal/arcmin²

Input

Galaxy catalogue with 30 gal/arcmin²

Kaiser-Squires with 0.05' bins + 0.1' smoothing

Kaiser-Squires with 1' bins

Kaiser-Squires with 0.05' bins

jeudi 20 octobre 16
Mass mapping as an inverse problem

Binned data: \( \gamma = F^* PF \kappa \)

Unbinned data: \( \gamma = T^* PF \kappa \)

\( T = \) Non Equispaced Discrete Fourier Transform (NDFT)

\[
\min_{\kappa} \frac{1}{2} \| \gamma - P \kappa \|_2^2 + C(\kappa)
\]

with \( P = T^* PF \)

\[
C(\kappa) = \sum_\theta \| (\Phi^t \kappa)_\theta \|_p = \sum_j \| \alpha_j \|_p
\]
Example with 93 % of missing data

10’ x 10’, z=0.3 cluster,
Lens plane at redshift zs = 1.2

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Example with 93 % of missing data

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Example with 93% of missing data

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Example with 93 % of missing data

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Example with 93% of missing data

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Example with 93% of missing data

Galaxy distribution: 93% of missing pixels, corresponding to 30 galaxies per square arcminute
Missing Data + Noise

10’ x 10’, z=0.3 cluster, \( n_g = 30/\text{arcmin}^2 \)

Kaiser-Squires + 1.0’ smoothing

GLIMPSE 2D

Kaiser-Squires + 0.5’ smoothing

jeudi 20 octobre 16
Computational Astrophysics

SKA/LSST/Euclid is **BIG DATA**, but also: **rich** and very **complex** data, which require sophisticated statistical methods, astrophysical models and a huge amount of additional simulated data $\Rightarrow$ **Big computational challenges**

**Computational Astrophysics** = Astrophysics + Statistics/Applied math + Computer Science (machine learning, HPC, simulations, etc)
Computational Astrophysics

SKA/LSST/Euclid is **BIG DATA**, but also: rich and very complex data, which require sophisticated statistical methods, astrophysical models and a huge amount of additional simulated data ==> **Big computational challenges**

**Computational Astrophysics** = Astrophysics + Statistics/Applied math + Computer Science (machine learning, HPC, simulations, etc)

**USA:** The Center for Computational Astrophysics, created in 2016, led by David Spergel, Manhattan, NY, USA

60 open positions
Conclusions

- Upcoming surveys (SKA, LSST and Euclid) will provide fantastic new data set.
- To the first two kinds of uncertainties, stochastics and systematics, a third one has now to be considered: Approximation.
- Mathematical challenges: higher order statistics, combine probes, etc.
- HPC is a concern for processing of the final products, and not anymore only for pre-processing or simulations.
- Astrostatistics teams need to extended to include people from computer science (machine learning, HPC, etc)
- The future astronomers may not ask for observing time to collect data, but rather for computing time to process the data that are already collected.