BPOL Data Analysis

R. Stompor (APC), C. Baccigalupi, A. Balbi,
J. Bartlett, E. Hivon, A. Jaffe, P. Natoli

The three-fold challenge

- **Due to:**

  - The huge data volume - as determined by the noise characteristics of the available detectors.

  - The minute signal amplitude - as predicted by the current models.

  - The signature goals of the mission (large angular scales) - as relevant for the satellite mission.

CMB data analysis process

- Data management.
- Data analysis methods and algorithms:
  1. Time domain processing.
  2. Map-making.
  3. Component separation/foreground cleaning.
  4. Power spectrum estimation.
- Computer resources.

BPOL data set (1)

- In time domain:

- Defined by the required sensitivity and the characteristic noise of the detectors.

- For example, from the SAMPAN proposal:
  - 5uK.arcmin - the target sensitivity;
  - 140 uK rt(sec) - the detector sensitivity;
  - 10msec - the sampling rate.

⇒ 1,000 - 10,000 detector-years of total integration (uniform sky coverage);

⇒ $n_t \sim 10^{13}$-10^{14} total time samples (10-10^2x Planck);

⇒ $n_{det} \sim 10^3$ detectors (5,000 for SAMPAN) for each frequency channel.

⇒ $O(10^1$-$10^2)$ PetaBytes of the storage for raw and intermediate data products.

BPOL data set (2)

- In pixel domain:
- What beam size/pixel size to choose ?
  - 30' : $N_{pix} = 10^5$
  - 8' : $N_{pix} = 10^6$
  - 2' : $N_{pix} = 10^8$

  of pixels per Stokes parameter per frequency channel.
- Beam vs pixel ?
  - beam over-sampling factor:
    - x 3 ? (like for T);
    - x 10 ?!
- Not a Planck-like data set (but also not like a typical ground-based experiment.)

BPOL data set (3)

- Huge;
- Complex (multiple detectors, frequency channels, etc).

Data management challenge:

•Storage (object diversity);
•Retrieval (speed);
•Keeping track of (reliability).

DA: time domain processing

- **Resources consuming:**
  - Typically time domain operations (FFTs, noise estimations, filtering, deglitching, etc) scale like $O(n_t)$ with a prefactor up to hundreds and more;
  - Parallel efficient I/O …
- **Automated:**
  - to cope with the sheer volume of the data;
- **Needs specialized tools for the TOD model testing and validating:**
  - no comfort of manually manipulating each detector data.
- **Model building and verification.**

DA: map-making

- **Well-understood and researched but not a perfect tool:**
  - beam asymmetries (main lobe and side lobes);
  - pixelization effects (including pixelization schemes);
  - degeneracies and near-degeneracies.

- **Resource consuming:**
  - $O(n_{iter} n_t \ln \lambda)$; $n_{iter} \sim 10^2$, $\ln \lambda \sim 10$ - if no cross detector correlations;
  - $O(n_{det}^2 n_{iter} n_t \ln \lambda)$; $n_{iter} \sim 10^2$, $\ln \lambda \sim 10$ - if all detectors are correlated.
  - $O(n_{det}^{(1-2)} n_{stokes}^2 n_{pix}^3)$; if the map covariance needs to be explicitly estimated.

- **Simplifying the pixel domain noise structure:**
  - Operation optimization: e.g., scanning strategy (well-crosslinked, pixel revists on multiple timescales).
  - Instrument design (extra modulations).

- **Low (super beam) resolution map-making.**

DA: component separation/removal

- Needed to access as much of the sky as possible.

- Separate the sky components => to create foreground template;
- Clean the sky to get foreground-free sky map and its uncertainty.

- "Blind":
  - Ability to take advantage of basic properties such as CMB isotropy and/or non-Gaussianity;
  - CPU/FLOPs cheap;
  - Limited by the assumptions;
  - Difficult to estimate the errors due to separation/cleaning in any other way than extensive Monte Carlos => well-suited for pseudo-cl methods and small/intermediate angular scales;

- Parametric:
  - Need model consistency verification ("goodness-of-fit", "model selection") procedure;
  - Allow for the (approximate) estimations of the component errors.

- Beam sizes at different frequencies ?!

DA: power spectra

- Optimal (maximum likelihood) methods (pixel or time domain based);
  - Direct: e.g., quadratic estimators (Newton-Raphson, QML);
  - Sampling algorithms, e.g., Gibbs sampling;
  - Hybrid (e.g., QML + MC).
- Foreground/systematic effects template marginalization.
- Require precise reliable uncertainty estimates (explicit and/or implicit);
- Beam issues (asymmetries, side lobes, channel dependence…);
- Simulations/random sampling: pixel or time domain based … => noise modeling in pixel domain,

Computer resources

- A dedicated 1 PetaFlop computational platform (scaling-up from Planck).
- $O(10^2)$ PetaByte storage: short, intermediate and archiving.
- The resources are significant but not overwhelming => c.f., LOFAR, SKA, LSST which are to be operational on the similar or shorter timescales.
- The good news:
  - Moore's law is likely to continue over the next decade or so.
  - Computer libraries (linear algebra, FFT, wavelets, etc.) keep on improving.

Conclusions

- The data analysis of the anticipated BPOL data set will be a challenging but not hopeless task.

- Need for R&D:
  - new algorithms and methods;
  - automatization (humans a potential bottleneck);
  - keeping up with the available computer hardware and software.

- Not alone … (many common problems with the forthcoming ground-based and balloon-borne experiments) but likely one of the most challenging …