VLBI Data & Errors

Lindy Blackburn

Event Horizon Telescope Collaboration
Center for Astrophysics | Harvard & Smithsonian
• Interferometry and Synthesis in Radio Astronomy
  Thomson, Moran, Swenson (TMS)

• Synthesis Imaging in Radio Astronomy II
  Ed: Taylor, Carilli, Perley

• KVN Lecture Notes
  Sasao, Fletcher

• Synthesis and Imaging Workshop 2018 Presentations
  NRAO
Van Cittert-Zernike theorem

- Relates spatial coherence of wavefront with brightness distribution of distant source

\[
\langle E_1 E_2^* \rangle = S_\nu
\]

point source

\[
h E_1 = E_2^* = S_\nu
\]
Van Cittert-Zernike theorem

- Relates spatial coherence of wavefront with brightness distribution of distant source

\[ \delta \phi = -2\pi u \cdot \sigma \]

\[ \langle E_1 E_2^* \rangle = S_\nu \]

point source

\[ \langle E_1 E_2^* \rangle = e^{-2\pi u \cdot \sigma} S_\nu \]

shifted point source
Van Cittert-Zernike theorem

- Relates spatial coherence of wavefront with brightness distribution of distant source

\[ \langle E_1 E_2^* \rangle = S_\nu \]

point source

\[ \langle E_1 E_2^* \rangle = e^{-2\pi u \cdot \sigma} S_\nu \]

shifted point source

extended source (integration over many point sources)

\[ \langle E_1 E_2^* \rangle = \iiint e^{-2\pi u \cdot \sigma} I_\nu(\sigma) \, d\Omega \]

\[ = \mathcal{V}(u) \]

“Visibility function” encodes 2D complex spatial frequency components of the sky brightness
Van Cittert-Zernike theorem

- Relates spatial coherence of wavefront with brightness distribution of distant source

Time-shift antenna to form baseline vector taken in the plane of propagation (Linearize about phase center)

\[ \langle E_1 E_2^* \rangle = S_\nu \]

point source

\[ \langle E_1 E_2^* \rangle = e^{-2\pi \mathbf{u} \cdot \mathbf{\sigma}} S_\nu \]

shifted point source

\[ \langle E_1 E_2^* \rangle = \int \int e^{-2\pi \mathbf{u} \cdot \mathbf{\sigma}} I_{\nu}(\mathbf{\sigma}) \, d\Omega \]

extended source (integration over many point sources)

“Visibility function” encodes 2D complex spatial frequency components of the sky brightness
Ingredients of a VLBI measurement

We just need to measure $E_1$ and $E_2$ at various locations in the plane of propagation, but..

1. Earth is round & moving
2. Irregular delays from troposphere/ionosphere
3. Different atmospheric and receiver noise
4. Various electronics and path delays
5. Independent and imperfect clocks at all stations
6. Post-digitization artifacts
7. Unexpected data issues

In data reduction, we are asked to “hide” as many of these effects as possible (without ruining the data)
VLBI data and calibration pathway

- **Raw signals** [PB]
- **Correlation** [TB]
- **Calibration** [MB]
- **Analysis** [kB]
There are two important bandpass effects $H(f)$ and $G(f)$, sometimes factored into a real (autocorr) and complex BP.

Non-linear effects (delay-rate, atmospheric phase) must be described using time-dependent factors.
Propagation of the astrophysical signal $E$ through measurement $v$ can be characterized by complex gain factors $g$

$$v(t, f) = g(t, f)E(t, f) \quad \langle v_1v_2^* \rangle = g_1g_2^* \langle E_1E_2^* \rangle$$

Signal and ensemble averages are parameterized in *time* and *frequency*, which requires that $g$ is varying (relatively) slowly

For two orthogonal feeds of an antenna, this can be written in matrix form,

$$
\begin{pmatrix}
  v_L \\
  v_R
\end{pmatrix}
= 
\begin{pmatrix}
  g_L & 0 \\
  0 & g_R
\end{pmatrix}
\begin{pmatrix}
  E_L \\
  E_R
\end{pmatrix}
= 
\begin{pmatrix}
  \langle v_1Lv_2^* \rangle & \langle v_1Lv_2^* \rangle \\
  \langle v_1Rv_2^* \rangle & \langle v_1Rv_2^* \rangle
\end{pmatrix}
= 
\begin{pmatrix}
  g_1L & 0 \\
  0 & g_1R
\end{pmatrix}
\begin{pmatrix}
  \langle E_1LE_2^* \rangle & \langle E_1LE_2^* \rangle \\
  \langle E_1RE_2^* \rangle & \langle E_1RE_2^* \rangle
\end{pmatrix}
\begin{pmatrix}
  g_2L & 0 \\
  0 & g_2R
\end{pmatrix}
$$

Tracking various physical propagation effects, as well as non-zero off-diagonal “$D$” terms (leakage across feeds, or change of polarization basis), leads to Jones matrix formalism used by the Measurement Equation

$$v = J_a J_b \cdots J_z E \quad \langle v_1v_2^\dagger \rangle = J_{1a} J_{1b} \cdots J_{1z} \langle E_1E_2^\dagger \rangle J_{2z}^\dagger \cdots J_{2a}^\dagger J_{2b}^\dagger \quad \text{(see Smirnov 2011)}$$

Why so many? Physical model generally allows for least complexity. Note that matrices do not necessarily commute!

This is a very useful structure! One still must adopt good models for all the Jones matrices.. also track noise..
Sky signal [PB]

\[ E_6 \]

\[ r_{ij} = \frac{\langle v_i v_i^* \rangle}{\eta Q \sqrt{\langle v_i v_i^* \rangle \langle v_j v_j^* \rangle}} \]
Correlation

\[ E_6 \rightarrow g_6 \ast E_6 \]

\[ \text{atmosphere} \]

\[ r_{ij} = \frac{\langle v_i v_j^* \rangle}{\eta Q \sqrt{\langle v_i v_i^* \rangle \langle v_j v_j^* \rangle}} \]

\[ r_{ij} < 10^{-4} \quad \text{is small} \]

\[ \rightarrow \text{average} > 10^8 \text{ samples for detection} \]
The correlation coefficient is normalized by the system noise in the separate receiving systems.

Relating this to physical units of correlated flux density requires a calibration of the noise power:

$$ r_{ij} = \frac{\langle v_i v_j^* \rangle}{\eta Q \sqrt{\langle v_i v_i^* \rangle \langle v_j v_j^* \rangle}} $$

$$ |V_{ij}| = \sqrt{\text{SEFD}_i \times \text{SEFD}_j} |r_{ij}|. $$

This is encapsulated into the system-equivalent flux density (SEFD) at each site, which is the (measured) noise power in units of flux density from an unpolarized astrophysical source (above the atmosphere).

The SEFD is calibrated separately from the data using first principles, known bright calibrators (planets), and noise sources of known temperature placed directly in front of receiving elements, and is taken “a priori”.

For a heterogeneous array such as the EHT, SEFD can range by orders of magnitude $\sim 10^2$ to $\sim 10^5$ Jy.

EHTC 2019 ApJL 875 (Paper III), Issaoun+ 2017-CE-02
Closure relationships

At mm-frequencies, phase transfer from nearby calibration targets is very difficult or impossible so we have essentially no a priori information about station phase

\[ r_{12} = \frac{\langle x_1 x_2^* \rangle}{\eta_Q \sqrt{\langle x_1^* x_1 \rangle \langle x_2 x_2^* \rangle}} = \frac{e^{i\theta_1} e^{-i\theta_2} \mathcal{V}_{12}}{\sqrt{\text{SEFD}_1 \times \text{SEFD}_2}} \]

However there are \(N(N-1)/2\) baseline measurements of phase, yet only \((N-1)\) unknown station phases, so the measurements do capture structural phase information about the source

This information is captured by the “closure phases”

\[ \delta_1, \delta_2, \delta_3 \]

\(1\) insensitive to relative phase of each antenna:
\(N-1\) degrees of freedom removed from baselines
The correlation coefficients measured by the interferometer relies on finite averages to estimate expectation value

$$\langle v_1 v_2^* \rangle = g_1 g_2^* \langle E_1 E_2^* \rangle$$

What are the limits from the source?

$$\langle E_1 E_2^* \rangle = \int \int e^{-2\pi \mathbf{u} \cdot \sigma} I_\nu(\sigma) \, d\Omega$$

$$= \mathcal{V}(\mathbf{u})$$

Interferometer sweeps through ~FOV/beam measurements in 24h

For EHT sources of ~few^2 independent pixels, coherence length ~hours

A ~few pixels across a spatial dimension means >10% fractional bandwidth can be averaged without affecting independent measurements

Compact EHT sources implies intrinsic smoothness/stability in time and frequency for the model visibility

EHTC 2019 ApJL 875 (Paper II)
What about variability in gain parameters?

\[ \langle v_1 v_2^* \rangle = g_1 g_2^* \langle E_1 E_2^* \rangle \]

First-order phase systematics

\[ \Delta \phi = \frac{\partial \phi}{\partial \nu} \Delta \nu + \frac{\partial \phi}{\partial t} \Delta t \]

Delay  Delay-rate (rate)

\[ \tau = \frac{1}{2\pi} \frac{\partial \phi}{\partial \nu} \]
\[ \dot{\tau} = \frac{1}{2\pi \nu} \frac{\partial \phi}{\partial t} \]

Large delays and rates taken out at Correlator at high time-frequency resolution using a priori Earth model (calc)

We only worry about residual clock errors

0.5s × 0.5 MHz dump time, bandwidth

→ rates within ~2 ps/s (1.3 mm/s)
→ delays within ~1 μs
Fringe fitting involves self-calibration of residual clock errors to extract and average correlation coefficient.

At high frequencies, there are linear and non-linear residuals in phase vs frequency and phase vs time.

\[ \Delta \phi_{12}(t, f, pp) = \phi_0 \text{ (a priori phase corrections)} \]
First correction is generally an instrumental phase bandpass because it is stable across the experiment and can be solved on an ensemble of bright calibrators

$$\Delta \phi_{12}(t, f, pp) = \phi_0 + \phi_{2-1}(f)$$
After removing non-linear phase vs frequency, we can extract a clean linear fit to delay for this scan.

\[
\Delta \phi_{12}(t, f, \text{pp}) = \phi_0 + \phi_{2.1}(f) + 2\pi(f - f_{\text{ref}})\tau_{\text{pp}}
\]
As well as delay-rate, although this is poorly defined in the presence of rapid atmospheric fluctuations

$$\Delta \phi_{12}(t, f, \text{pp}) = \phi_0 + \phi_{2.1}(f) + 2\pi(f-f_{\text{ref}})\tau_{\text{pp}} + 2\pi f(t-t_{\text{ref}})\tau'_{\text{pp}}$$
And finally we can estimate and correct for atmospheric phase, here referencing to the first antenna

\[ \Delta \phi_{12}(t, f, pp) = \phi_0 + \phi_{2-1}(f) + 2\pi(f-f_{ref})\tau_{pp} + 2\pi(f(t-t_{ref})\tau^{'}_{pp} + \phi_{2-1}(t) \]

now we can average over the entire scan and bandwidth
Phase calibration pipeline

For mm-VLBI such as EHT, custom pipelines are required due to uniqueness of data and systematics. Purpose of steps is to fit as simple a model as possible, using as much S/N as available, and maintain closure (station-based gains).
Some things that can go wrong

- Too many free parameters for available S/N
- Introduce calibration noise
  Overfit data: bias amplitude upward, bias phase toward model
  Underutilize array constraints and gain priors
- Averaging over visibilities when gain is not stable
  Introduce non-closing errors (averaged product of station gains may not factor)
- Leaving in bad data / Ignoring systematics
- Wrong calibration solutions
  Systematic errors drive solution under the assumption of Gaussian thermal noise only
Thermal (statistical) error due to contribution from independent system noise at each site. For a normalized correlation coefficient and white noise, this follows from the central limit theorem,

\[ \sigma_{r,ij}^2 \frac{1}{2 \Delta t \Delta \nu} \]

Thermal noise is Gaussian and independent in real, imaginary components, and thus scales very simply under vector average and scaling by any visibility amplitude factors. Still, it is always good to check!

“Closure-phase” differencing, e.g. Ortiz+ 2016

Amplitude scatter, e.g. Wielgus+ 2019-CE-02
Thermal errors: non-Gaussianity

Thermal error is Gaussian in complex visibility, not necessarily in amplitude & phase

common estimators of phase error will give large reduced chi-square at low S/N

Figure 12.

Phase error vs expectation

KL Divergence vs Normal

Figure 2.
Systematic errors: closing vs non-closing

Closing errors (manageable)

Errors in gain calibration: \[ V_{ij} = g_i g_j r_{ij} + n_{ij} \]

Non-closing errors (try to minimize)

Non-thermal baseline errors: \[ V_{ij} = g_i g_j r_{ij} + n_{ij} + e_{ij} \]

Difficult to estimate, possibly reflected in trivial closure phases and amplitudes polarization leakage, band-pass non-overlap, coherence issue, etc

Commonly modeled as additional Gaussian RV:

\[ \sigma^2 = \sigma_{\text{th}}^2 + s^2 \]

\( s \sim 1-2\% \)

but be careful! most likely not independent across data points (do not average..)

EHTC 2019 ApJL 875 (Paper IV)

EHTC 2019 ApJL 875 (Paper III), Wielgus+ 2019

Possibly reflected in high/low band comparison, pipeline comparison, etc

If uncertain, best left for self-calibration (do not “inflate” data errors)
The noise properties of the correlation coefficients from the correlator are very simple:

**Gaussian noise** in real and imaginary components, independent across all data products

This is ideal for model fitting, calculating likelihoods, goodness-of-fit, etc.. messing with the data just makes it worse

**Simple example — Gain error:** \( V_{ij} = [g_i g_j^*] r_{ij} + n_{ij} \)

If Gaussian, can be captured by covariance matrix (e.g. for log amplitude)

\[
\Sigma_a = \begin{pmatrix}
\sigma_{12}^2 + \sigma_{g,1}^2 + \sigma_{g,2}^2 & \sigma_{g,1}^2 & \sigma_{g,2}^2 \\
\sigma_{g,1}^2 & \sigma_{23}^2 + \sigma_{g,1}^2 + \sigma_{g,3}^2 & \sigma_{g,3}^2 \\
\sigma_{g,2}^2 & \sigma_{g,3}^2 & \sigma_{23}^2 + \sigma_{g,2}^2 + \sigma_{g,3}^2
\end{pmatrix}
\]

**More complicated example — Closure phase**

\[
\Sigma_\psi = \begin{pmatrix}
\sigma_{12}^2 + \sigma_{23}^2 + \sigma_{13}^2 & \sigma_{12}^2 & -\sigma_{13}^2 \\
\sigma_{12}^2 & \sigma_{12}^2 + \sigma_{24}^2 + \sigma_{14}^2 & \sigma_{14}^2 \\
-\sigma_{13}^2 & \sigma_{14}^2 & \sigma_{13}^2 + \sigma_{34}^2 + \sigma_{14}^2
\end{pmatrix}
\]

covariance over 3 closure phases

Ignoring covariant errors often leads to confidence intervals which are too small!

It can be fun and instructive to use covariant errors, but make sure there is a **very good reason** before moving away from forward modeling into simple data products..
After a lot of work, we want to make sure we have good data, and not bad data.¹

¹https://www.bu.edu/blazars/songs/baddata.html
http://bit.ly/HandlingDataEval