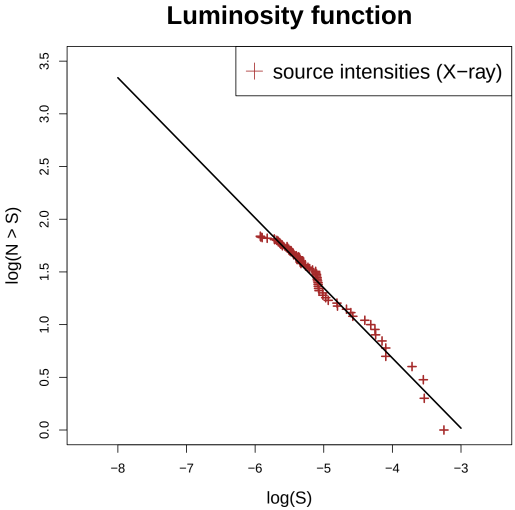
Estimation of Galaxy Luminosity Distributions from Incomplete X-ray and Optical Survey Data

Presenter: Max Autenrieth (Imperial College London)

Collaboration: David A. van Dyk (Imperial College London)
Vinay Kashyap, Aneta Siemiginowska (Harvard/CfA)
David Stenning (SFU), Andreas Zezas (University of Crete)

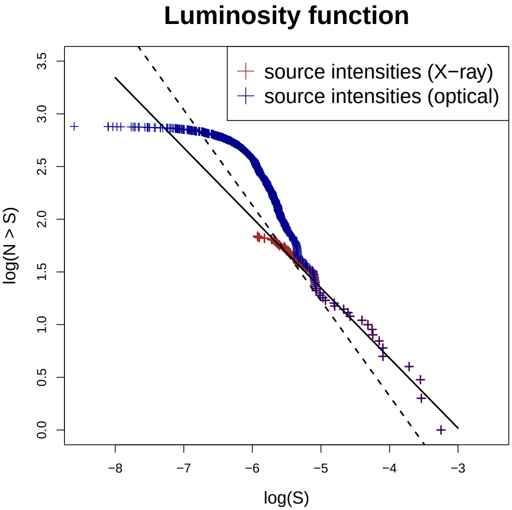
November 2, 2023

Motivation: log(N) – log(S)



Goal: Estimate X-ray source intensities and obtain luminosity function.

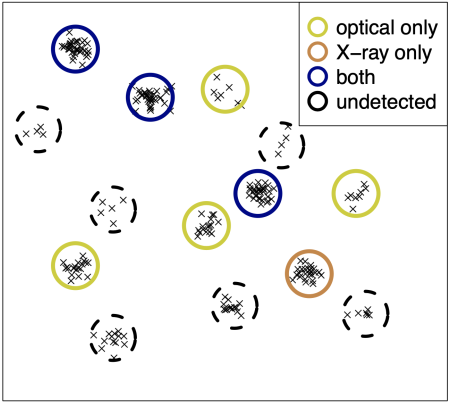
Motivation: log(N) – log(S)



Goal: Estimate X-ray source intensities and obtain luminosity function.

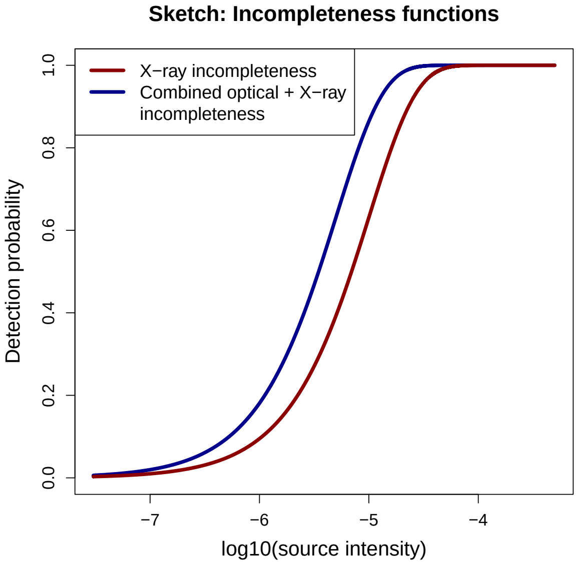
A New Strategy for a Key Challenge

Sketch: Source population (detection cases)

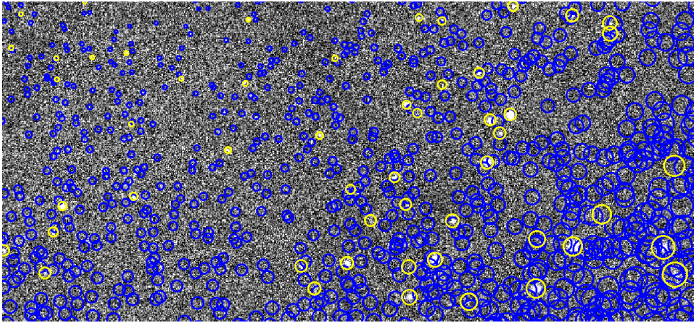


- **Challenge:** Incompleteness in X-ray data
- Low-intensity sources most prevalent, but least likely detected in X-ray
- **Strategy:** Leverage optical data, combining X-ray and optical surveys
- Account for missing sources via incompleteness functions

X-ray and Optical Incompleteness



Additional Complications:



Patch of Chandra Deep Field Survey

- Background contamination, exposure and off-axis corrections
- Overlapping optical sources (blue)
- X-ray source detection (yellow), and optical matching

Source Model

- In source region i, observed photon counts Y_i are sum of latent background $\mathcal{B}_i$ and source $\mathcal{S}_i$

$$Y_i = \mathcal{S}_i + \mathcal{B}_i. \tag{1}$$

- Arrival of photons at detector as Poisson process, with source model

$$\mathcal{S}_i \big| \lambda_i \overset{\text{indep}}{\sim} \text{Poisson}(e_i \lambda_i \mathcal{T}) \tag{2}$$

- Background count $\mathcal{B}_i$ with rate ξ in source region i modeled via

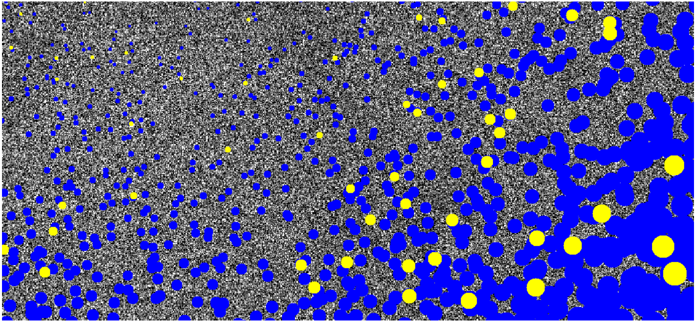
$$\mathcal{B}_i \big| \xi \overset{\text{indep}}{\sim} \text{Poisson}(a_i \xi \mathcal{T}). \tag{3}$$

Data	**Description**
a_i	area of the source region
Y_i	counts collected in source region (of area a_i)
e_i	telescope effective area [cm^2] at source location
A	area of the background region
X	collected background counts

Background Model

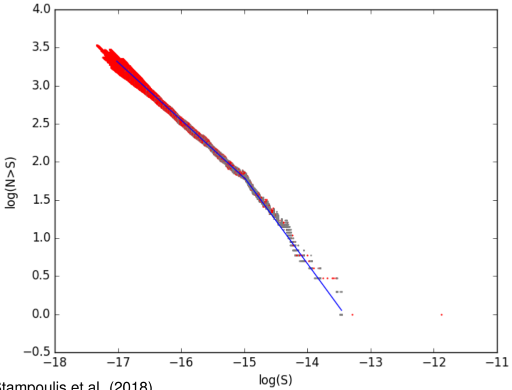
- Background rate ξ (count/s/pixel) **uniform** across source regions.
- Observed photon count in background modeled via

$$X \big| \xi \sim \text{Poisson}(A\xi\mathcal{T}), \tag{4}$$



Patch of Chandra Deep Field Survey

Estimation of the Luminosity Function



Stampoulis et al. (2018)

- Piece-wise linear $\log(N) - \log(S)$ relation assumed.
- Breakpoints τ_1, τ_2, slopes θ_1, θ_2

- $(\lambda_1, \ldots, \lambda_n)$ independent with double-Pareto population

$$f(\lambda_i|\theta_1, \theta_2, \tau_1, \tau_2) = \left(\frac{\theta_1}{\tau_1}\right)\left(\frac{\lambda_i}{\tau_1}\right)^{-(\theta_1+1)} 1_{\{\tau_1 \le \lambda_i \le \tau_2\}} + \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}\left(\frac{\theta_2}{\tau_2}\right)\left(\frac{\lambda_i}{\tau_2}\right)^{-(\theta_2+1)} 1_{\{\tau_2 \le \lambda_i \le \infty\}}$$

Incompleteness Correction

- For each source in population, introduce indicator Z_i, with

$$Z_i = \begin{cases} 0 & \text{if source } i \text{ not observed} \quad (\text{w/p } 1 - g(\lambda_i) - h(\lambda_i|\phi)) \\ 1 & i \text{ observed in optical only} \quad (\text{w/p } h(\lambda_i|\phi)) \\ 2 & i \text{ observed in both} \quad (\text{w/p } g(\lambda_i), \text{ with } g \text{ known}) \end{cases} \tag{5}$$

- Joint distributions of Z and λ via

$$p(\lambda, Z = 2|\boldsymbol{\theta}) = f(\lambda;\boldsymbol{\theta})g(\lambda) \tag{6}$$

$$p(\lambda, Z = 1|\boldsymbol{\theta}, \phi) = f(\lambda;\boldsymbol{\theta})h(\lambda|\phi) \tag{7}$$

- Binning based on λ. Let $k = 1, \ldots K$ be bins on λ, with bin size Δ_{λ_k}

$X_1^{(k)}$ = # of sources in bin k with $Z = 1$, with $X_1^{(k)} \sim Pois(f(\lambda_k;\boldsymbol{\theta})\Delta_{\lambda_k} h(\lambda_k|\phi)\psi)$

$X_2^{(k)}$ = # of sources in bin k with $Z = 2$, with $X_2^{(k)} \sim Pois(f(\lambda_k;\boldsymbol{\theta})\Delta_{\lambda_k} g(\lambda_k)\psi)$

Computational Details:

Incompleteness function:

- Incompleteness function dependend on nuisance parameters: L (off-axis angle), e (effective area), ξ (background rate)

$$p(\lambda, L, \xi, e, Z = 2|\boldsymbol{\theta}) = p(\lambda, L, \xi, e|\boldsymbol{\theta})P(Z = 2|\lambda, L, \xi, e) \tag{8}$$

- Assuming independence of λ and L, ξ, e, this yields for λ

$$p(\lambda, Z = 2|\boldsymbol{\theta}) = f(\lambda|\boldsymbol{\theta}) \int p(L)p(\xi)p(e)P(Z = 2|\lambda, L, \xi)dLd\xi de$$

Metropolis-within-Gibbs sampler:

- Updating source/background model and incompleteness via

$$1.) \quad (\xi, \boldsymbol{\lambda})^{(t)} \sim p(\xi, \boldsymbol{\lambda}|\boldsymbol{\theta}^{(t-1)}, \psi^{(t-1)}, \phi^{(t-1)}, \boldsymbol{X}, \boldsymbol{Y}, N_{\text{obs}}) \tag{9}$$

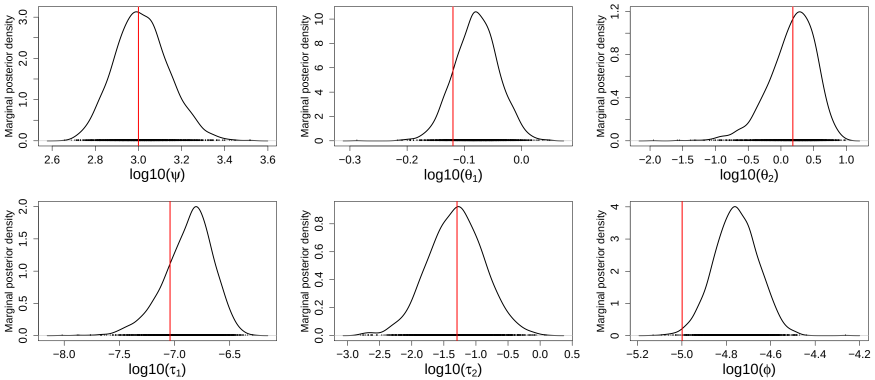
$$2.) \quad (\boldsymbol{\theta}, \psi, \phi)^{(t)} \sim p(\boldsymbol{\theta}, \psi, \phi|\lambda^{(t)}, Z_{\text{obs}}), \tag{10}$$

Simulation Study – Setup:

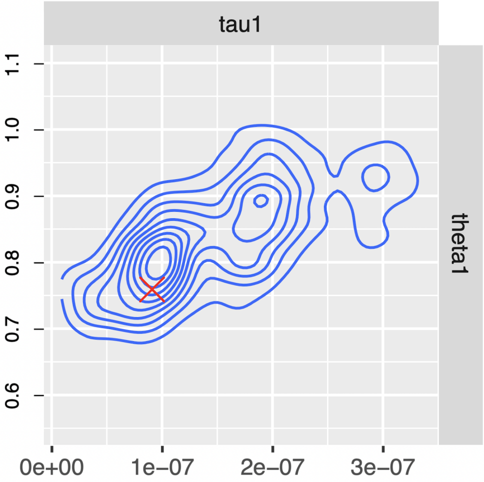
- Simulate realistic X-ray sources, mimicking data from the Chandra Deep Field South (CDFS)
- Assuming optical and X-ray survey incompleteness (resembling previous studies (Stampoulis 2018))
- Expected population size $\psi = 1000$, with only $N_{\text{obs}} = 124$ observed (in either both, or only optical survey)

Simulation Study – Results:

Marginal posterior distributions (parameters of interest):



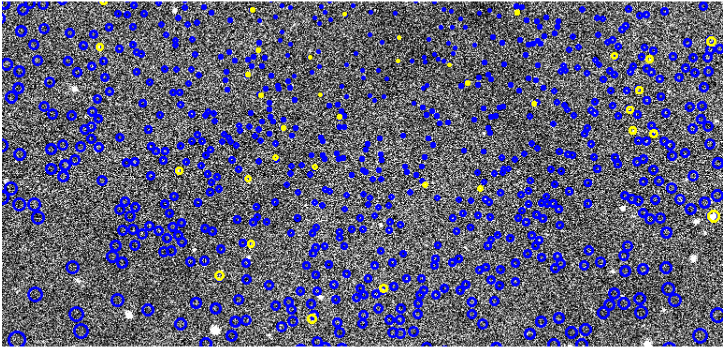
Simulation Study – Results:



- Correlation between τ_1 (population minimum) and θ_1 (first power-law slope)
- Higher τ_1 leads to steeper $\log(N) - \log(S)$
- Lower τ_1 leads to strong overestimation of population size, and flat $\log(N) - \log(S)$

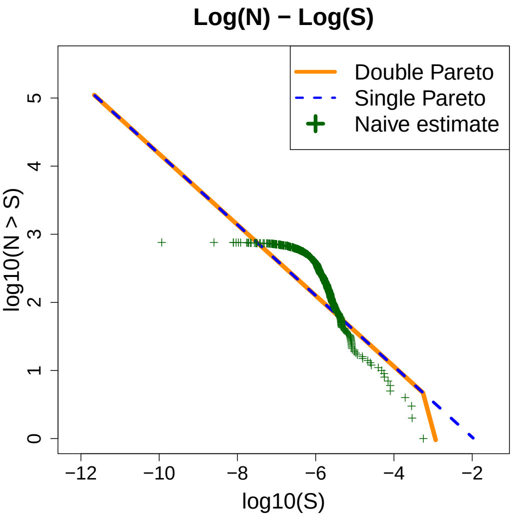
Data Analysis: Chandra Deep Field South (CDFS)

- 1524 galaxies observed (without overlap) in subregion of CDFS
- 70 galaxies observed in X-ray and optical, 1454 in optical only
- Source counts extracted at optical locations



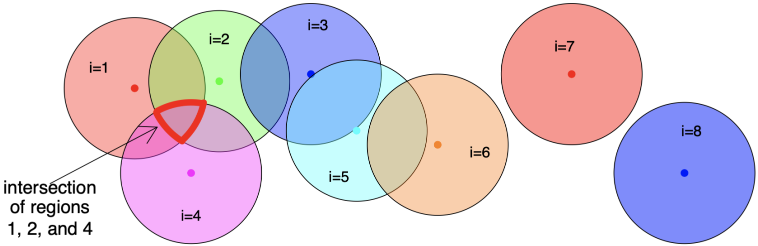
Patch of Chandra Deep Field Survey

Data Analysis: Chandra Deep Field South



- Expected population size $\psi = 110,704$ (mean posterior).

Ongoing Work: Overlapping Sources



- Photon counts $Y_{\mathcal{I}(s)}$ modelled in each of 14 segments s of overlap.
- For highlighted segment: $\mathcal{I}(s) = \{1, 2, 4\}$ with counts $Y_{\mathcal{I}(s)}$.
- $Y_{\mathcal{I}(s)}$ consists of mixture of photons from sources in s and background

$$Y_{\mathcal{I}(s)} = \sum_{i \in \mathcal{I}(s)} \mathcal{S}_{s,i} + \mathcal{B}_{\mathcal{I}(s)}. \tag{11}$$

Concluding Remarks

- Estimation of X-ray luminosity distributions is challenging!
  - Strong incompleteness in the high density region of the population
- Optical surveys can be leveraged, complementing sources in the incomplete, low-intensity end
  - Principled incompleteness corrections via hierarchical Bayes model

- Further challenges/extensions:
  - Incorporating overlapping sources
  - Improving survey matching
  - Uncertainty on source locations
  - Zero-inflated population model

References I

Stampoulis, V. (2018). *Bayesian estimation of luminosity distributions and model based classification of astrophysical sources*. PhD thesis, Imperial College London.

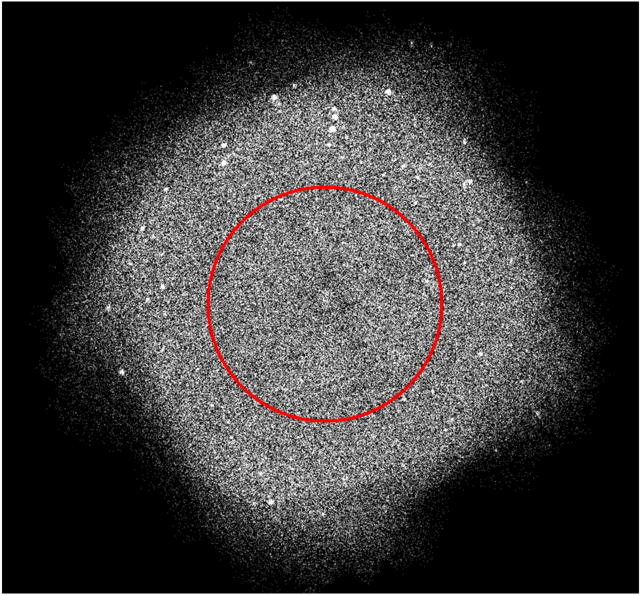
Udaltsova, I. S. (2014). *The Universe at Your Fingertips: Bayesian Modeling and Computation in Problems of Observational Cosmology*. University of California, Davis.

Wang, L., Kashyap, V. L., van Dyk, D. A., and Zezas, A. (2022). Bayesian methods for modeling source intensities. *Paper Draft*.

Wright, N. J., Drake, J. J., Guarcello, M. G., Kashyap, V. L., and Zezas, A. (2015). Simulating the sensitivity to stellar point sources of chandra x-ray observations. *arXiv preprint arXiv:1511.03943*.

Thank you very much for your time!

Chandra Deep Field Survey South – Merged Event File



Additional Data available:

Data	Description
a_i	area of the source region
Y_i	counts collected in source region (of area a_i)
d_i	area of the background region (around source i)
X_i	background counts collected in source region (of area d_i)
e_i	telescope effective area [cm^2]
r_i	proportion of photons expected to fall in source region $\equiv 1$
bg-sur-bri	background counts /pixel (for incompleteness correction)
off-axis	(off-axis angle - needed for the incompleteness correction)
sign	(source S/N ratio)

- Data available from the Chandra Deep Field Catalogue.
- $n = 358$ X-ray sources detected in data set.
- Observation time $\mathcal{T}$ in seconds ($\mathcal{T} = 1960631$).
- The count-rate to flux conversion for the reference point is 1.06E-11 erg/s/cm^2/cnt/s

Likelihood Functions of Source Model

Background case 1:

- Likelihood function for $(\xi, \boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$,

$$L(\xi, \boldsymbol{\lambda} \big| \mathbf{D}) = \exp(-A\mathcal{T}\xi) \frac{(A\mathcal{T}\xi)^X}{X!} \prod_{i=1}^{n} \exp[-(a_i\xi + r_i e_i \lambda_i)\mathcal{T}] \frac{[(a_i\xi + r_i e_i \lambda_i)\mathcal{T}]^{Y_i}}{Y_i!}$$

Background case 2:

- Likelihood function for $(\boldsymbol{\xi}, \boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$

$$L(\boldsymbol{\xi}, \boldsymbol{\lambda} \big| \mathbf{D}) = \prod_{i=1}^{n} \exp(-a_i\mathcal{T}\xi_i) \frac{(a_i\mathcal{T}\xi_i)^{X_i}}{X_i!} \prod_{i=1}^{n} \exp[-(a_i\xi + r_i e_i \lambda_i)\mathcal{T}] \frac{[(a_i\xi + r_i e_i \lambda_i)\mathcal{T}]^{Y_i}}{Y_i!}$$

Likelihood and Posterior

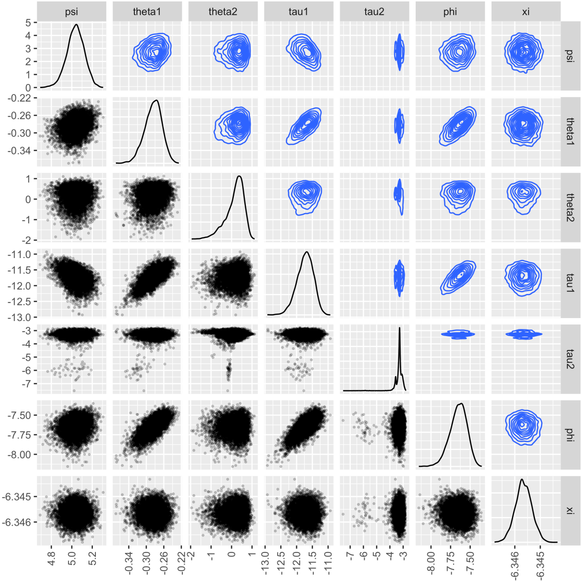
- Assume that $X_1^{(k)}, X_2^{(k)} \perp\!\!\!\perp X_1^{(l)}, X_2^{(l)} | \boldsymbol{\theta}, \phi, \psi$, for all $k \neq l$, with $k, l = 1, \ldots, K$.
- Log-likelihood (Assuming λ (observed sources) known):

$$\begin{aligned}
\ell(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) = &- \psi \sum_k f(\lambda^{(k)}; \boldsymbol{\theta}) \Delta_{\lambda_k} [h(\lambda^{(k)}|\phi) + g(\lambda^{(k)})] \\
&+ \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(f(\lambda^{(k)}; \boldsymbol{\theta})) + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(f(\lambda^{(k)}; \boldsymbol{\theta})) \\
&+ \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(h(\lambda^{(k)}|\phi)) + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(g(\lambda^{(k)})) \\
&+ N_{obs} \log(\psi)
\end{aligned}$$

- Full posterior distribution:

$$p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi | \boldsymbol{Y}, \boldsymbol{X}, Z_{\text{obs}}) \approx p(\boldsymbol{Y}, \boldsymbol{X} | \xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, Z_{\text{obs}}) p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K | \boldsymbol{\theta}, \phi, \psi) p(\boldsymbol{\theta}) p(\phi) p(\psi) p(\xi).$$

CDFS: Pairwise marginal posterior distributions



Literature

Stampoulis (2018);
Udaltsova (2014);
Wang et al. (2022);
Wright et al. (2015)