POTENTIALS AND PERILS OF PREPROCESSING

Alexander W. Blocker and Xiao-Li Meng

Maxime and Angela

April 9th, 2015

OVERVIEW OF THE PAPER

Addresses data preprocessing and its effect on later analysis in two main ways:

1. Points out potential problems with current preprocessing strategies
2. Describes a statistical framework for analyzing data preprocessing

WHAT IS THE PREPROCESSING PROBLEM?

Raw data is reduced based on assumptions about future analysis, which **constrains downstream data analysis**.

Yet, preprocessing is necessary:

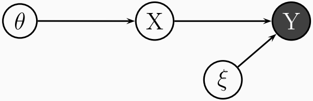
1. Even if original data was passed down, many users may not know how to process the data themselves.
2. The analyst performing the preprocessing may have detailed knowledge about the experimental situation that the data user does not have

THE MULTIPHASE SETUP

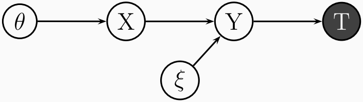
Two phases:

1. **Preprocessing** = Data generation, collection, preprocessing
2. **Downstream Analysis** = inference using output from phase 1

Preprocessor's model

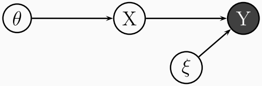


Downstream analyst's model

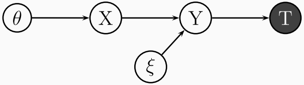


MISSING DATA

Preprocessor's model



Downstream analyst's model



1. **Missingness from data generation: X is missing** since preprocessor observes Y
2. **Missingness from inference process: Y and X are both missing** since data analyst observes T

Note that the analyst's T(Y) is a product of preprocessing, so it is a **design decision** of the preprocessor.

SQUARE KILOMETER ARRAY

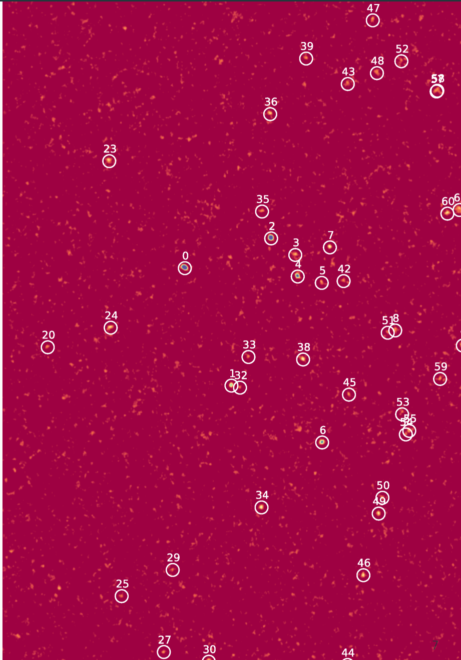
- Radio telescope to be built around 2018-2030
- Total collecting area = 1 km^2, between about 3000 dishes.
- Raw data: Each antenna produces **420Gb/s**

SQUARE KILOMETER ARRAY

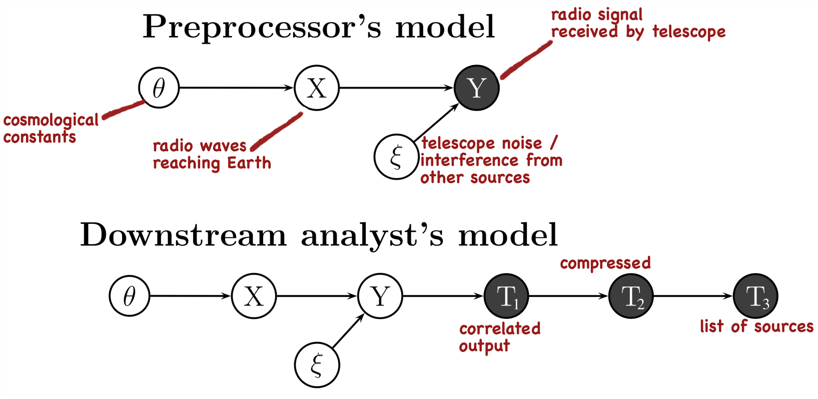
- First stage preprocessing: correlation of signals between dishes
- **Lossless** reduction to about 100 Tb/s ($\approx$ the internet)
- Standard lossless compression algorithms (SZIP, GZIP...) can further reduce the data by about a third (Rajeswaran & Winberg 2013).

SQUARE KILOMETER ARRAY

- That’s still more raw data than is reasonable to store on hard drives
- it will be necessary to **further reduce** the data in **real time**



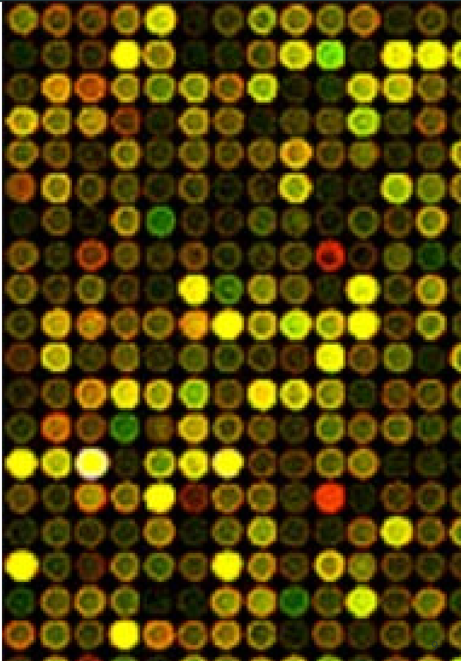
SQUARE KILOMETER ARRAY


Preprocessor's model
radio signal received by telescope
θ
X
Y
ξ
cosmological constants
radio waves reaching Earth
telescope noise / interference from other sources
Downstream analyst's model
θ
X
Y
ξ
T1
T2
T3
compressed
correlated output
list of sources

- 3 possible levels of preprocessing

HIGH-THROUGHPUT BIOLOGY: EXPRESSION MICROARRAYS

- **Raw data** are probe intensity measurements
- **Want to study** log fold change in gene expression between different conditions



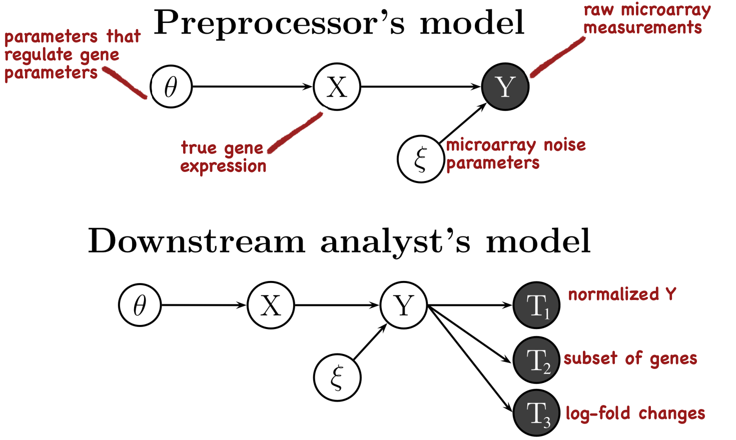
high-throughput biology: expression microarrays

Many levels of preprocessing:

1. **background correction** to reduce noise- many different algorithms, mostly passing down point estimates
2. **normalization across different microarrays** to reduce systematic error
3. screening for data corruption, etc.

HIGH-THROUGHPUT BIOLOGY: EXPRESSION MICROARRAYS

Let’s say our goal is to assess if there is a significant difference in expression levels of 10 different genes between people who have blue eyes and people who have brown eyes



Preprocessor’s model
parameters that regulate gene parameters
raw microarray measurements
θ
X
Y
true gene expression
ξ
microarray noise parameters
Downstream analyst’s model
θ
X
Y
ξ
T1
normalized Y
T2
subset of genes
T3
log-fold changes

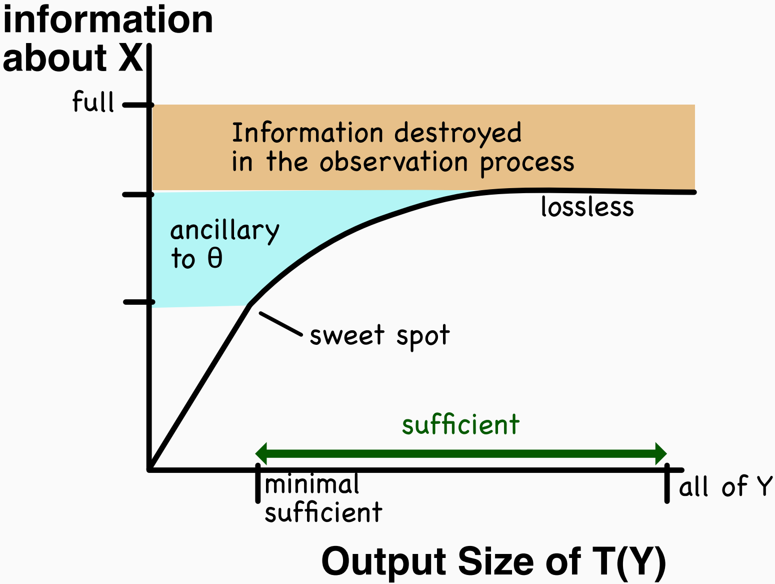
WHY ARE PREPROCESSING AND ANALYSIS SEPARATE?

1. **Data warehousing limitations**- cannot store the amount of data produced by radio telescopes
2. **Division of labor**-
   - inefficient for each biologist to reproduce a whole analysis pipeline
   - many companies that provide microarrays include algorithms and methods for data preprocessing immediately from raw output
3. **Inability to provide whole datasets, only preprocessed versions**- data anonymization concerns, for example

WHAT SHOULD WE RETAIN?

- **Optimal:** lossless compression to **minimal sufficient** statistic for a given research model
  - Sufficiency: $\Pr(\theta \mid Y, T(Y)) = \Pr(\theta \mid T(Y))$
  - This is only practical if the analyst's model $X \mid \theta$ is specified and known before preprocessing.
- **More realistic:** Hope downstream analyst's scientific model is related (congenial) to preprocessor's observation model

WHAT SHOULD WE RETAIN?



REGRESSION WITH NUISANCE PARAMETERS: AN EXAMPLE

Data: Each processor i observes $y_{ij} \sim N(\beta_0 + \beta_{1i} x_j, \sigma^2)$ for $j = 1 \ldots m$

Downstream analyst wants to estimate β_0, with β_{1i} and σ^2 as nuisance parameters so that ξ = $(\sigma^2, \beta_{11} \ldots \beta_{1n})$

Preprocessor reduces data to: $T_i = \frac{1}{m} \sum \left(y_{ij} - \hat{\beta_{1i}} x_j \right)$ where $\hat{\beta_{1i}}$ is the OLS estimator of β_{1i}.

Distribution of T_i still depends on σ^2, but is free of β_{1i}

WHAT SHOULD WE RETAIN?

- **Problem:** The “sweet spot” depends on the scientific model $X \mid \theta$. Different analysts will have different models $X \mid \eta$ with a different optimal $\tilde{T}(Y)$
    - and even those might not be known in advance.
- **Goal:** Find a statistic that retains enough information to keep most people happy.

DECISION THEORY TO THE RESCUE.

- Further reduction is often necessary, but **comes at a cost**
- formalised through **decision theory**: loss function $\mathsf{L}\left(\hat{\theta}, \theta_{\text{true}}\right)$
- risk = expected loss

$$\mathsf{R}\left(\hat{\theta}, \theta_{\text{true}}\right) = \mathbb{E}\left\{\mathsf{L}\left(\hat{\theta}, \theta_{\text{true}}\right)\right\}$$

- regret = how bad is this estimator?

$$\mathsf{R}\left(\hat{\theta}, \theta_{\text{true}}\right) - \mathsf{R}\left(\hat{\theta}^{*}, \theta_{\text{true}}\right)$$

- Idea: risk with respect to better estimator. In our case, using the **full data** instead of the **reduced data**.

$$\mathsf{R}\left(\hat{\theta}\left(\mathsf{T}\right), \hat{\theta}\left(\mathsf{Y}\right)\right) = \mathbb{E}\left\{\mathsf{L}\left(\hat{\theta}\left(\mathsf{T}\right), \hat{\theta}\left(\mathsf{Y}\right)\right)\right\}$$

DECISION THEORY TO THE RESCUE.

$$\mathsf{R}\left(\hat{\theta}(\mathsf{T}), \hat{\theta}(\mathsf{Y})\right) = \mathbb{E}\left\{\mathsf{L}\left(\hat{\theta}(\mathsf{T}), \hat{\theta}(\mathsf{Y})\right)\right\}$$

- e.g. mean-squared error

$$\mathsf{R}\left(\hat{\theta}(\mathsf{T}), \hat{\theta}(\mathsf{Y})\right) = \mathbb{E}\left\{\left(\hat{\theta}(\mathsf{T}) - \hat{\theta}(\mathsf{Y})\right)^2\right\}$$

- asymptotically $\mathsf{R}\left(\hat{\theta}(\mathsf{T}), \hat{\theta}(\mathsf{Y})\right) \rightarrow \mathrm{regret}$ ("asymptotic decorrelation")
- for multiple analysts with different models, the costs can be averaged:

$$\mathrm{regret}(\mathsf{T}) = \frac{1}{\mathsf{N}_{\mathsf{analysts}}}\left[\mathsf{R}\left(\hat{\theta}(\mathsf{T}), \hat{\theta}(\mathsf{Y})\right) + \mathsf{R}\left(\hat{\xi}(\mathsf{T}), \hat{\xi}(\mathsf{Y})\right) + \ldots\right]$$

PROJECTING INTO THE FUTURE

Likelihood as a minimal sufficient statistic- computationally efficient approximations of the likelihood function could be foundation for passing information between phases of downstream analysis.

Two things to consider:

1. Nuisance parameters
2. Downstream analysts may be constrained by the likelihood approximation chosen. For example, analysts may want to estimate the data and estimate the parameters, and going from likelihood approximation to estimates of the data, X, could require a lot of effort and computation.

CONNECTIONS TO MULTIPLE IMPUTATION

1. Relates to prior presentation on congeniality in MI
2. Both papers want to bound and measure the amount of degradation in inference when information is imperfectly combined
3. Congeniality paper concludes that nuisance parameters can be a stumbling block for MI, perhaps understanding role of preprocessing in addressing nuisance parameters could be helpful

CONCLUSION

Despite previous discussion and historical concerns about data preprocessing, not much work has been done to address issues related with preprocessing.

This paper has proposed a conceptual and mathematical framework to understand the issues that arise in preprocessing, and provides some suggestions for how researchers can begin to think about and investigate preprocessing effects.

REFERENCES

1. Babu, M. Madan. ”An Introduction to Microarray Data Analysis.” Microarray Data Analysis (2004).
2. Blocker, Alexander W., and Xiao-Li Meng. ”The potential and perils of preprocessing: Building new foundations.” Bernoulli 19.4 (2013): 1176-1211.
3. Rajeswaran, Karthik, and Simon Winberg. ”Lossless Compression of SKA Data Sets.” Communications and Network 2013 (2013).