Preprocessing, Multiphase Inference, and Massive Data in Theory and Practice

Alexander W Blocker
Department of Statistics
Harvard University

MMDS 2012 — July 13, 2012

Joint work with Xiao-Li Meng
Outline

1. Perspective on preprocessing
2. Motivating examples
3. Framework
4. Theoretical cornerstones
5. Concluding remarks
Defining preprocessing

- **Formally**: Transformations and reductions of observed data for subsequent analyses
Defining preprocessing

- **Formally:** Transformations and reductions of observed data for subsequent analyses
- **Informally:** Everything that happens before statistical modeling and your favorite algorithms
Defining preprocessing

- **Formally**: Transformations and reductions of observed data for subsequent analyses
- **Informally**: Everything that happens before statistical modeling and your favorite algorithms
- Examples: aggregation, smoothing, calibration — all feature engineering
Defining preprocessing

- **Formally**: Transformations and reductions of observed data for subsequent analyses
- **Informally**: Everything that happens before statistical modeling and your favorite algorithms
- Examples: aggregation, smoothing, calibration — all feature engineering
- Widely considered an art, domain knowledge driven
Defining preprocessing

- **Formally**: Transformations and reductions of observed data for subsequent analyses
- **Informally**: Everything that happens before statistical modeling and your favorite algorithms
- Examples: aggregation, smoothing, calibration — all feature engineering
- Widely considered an art, domain knowledge driven
- Aim to enhance domain knowledge with formal, mathematical theory
Perils and promise

Destructive preprocessing

- Most non-trivial preprocessing is irreversible
- Assumptions matter — and the wrong ones cause a lot of damage
- Preprocessing decisions constrain all later analyses, no matter the scale
Perils and promise

Destructive preprocessing

- Most non-trivial preprocessing is irreversible
- Assumptions matter — and the wrong ones cause a lot of damage
- Preprocessing decisions constrain all later analyses, no matter the scale

Alleviating complexity

- Smaller data and less complex modeling required
- Separation of effort among analysts (e.g. pipelines and workflows)
Massive data

Scale is not always a savior

- First step: preprocess and extract features
- If information is discarded or distorted by preprocessing, scale will not always save you
- Conversely, simple and huge with good preprocessing often beats complex
Massive data

Scale is not always a savior

- First step: preprocess and extract features
- If information is discarded or distorted by preprocessing, scale will not always save you
- Conversely, simple and huge with good preprocessing often beats complex

Systematic errors at scale

- With massive data, systematic errors dominate statistical noise (e.g. Szalay)
- Observation/sensor models are vital (e.g. Ré)
- Help or harm depends on what is passed forwards
Theory vs. practice

Statistical theory
- Model generative process for observed data
- Evaluate procedures in their entirety
Theory vs. practice

Statistical theory
- Model generative process for observed data
- Evaluate procedures in their entirety

Statistical practice
- Delineate between pre- and post-modeling work
- Formal evaluation only after preprocessing
Theory vs. practice

Statistical theory
- Model generative process for observed data
- Evaluate procedures in their entirety

Statistical practice
- Delineate between pre- and post-modeling work
- Formal evaluation only after preprocessing

Closing the gap
- Want theoretical foundations for statistical practice
- Building this under banner of multiphase inference from theory of missing data
Multiphase, graphically

1. Align & Pre-process
2. Construct Templates
3. Estimate Occupancies
4. Screen Positions
5. Summarize Results
Multiphase, graphically

Align & Pre-process → Construct Templates → Estimate Occupancies → Screen Positions → Summarize Results
Multiphase, graphically

Align & Pre-process

Construct Templates

Estimate Occupancies

Screen Positions

Summarize Results

Phase

Phase

Phase

Phase

Phase
Preprocessing ubiquitous in massive-data astrophysics (Richards & Szalay)
Preprocessing ubiquitous in massive-data astrophysics (Richards & Szalay)

Two solar observatories: SDO and ATST
Massive solar data

- Preprocessing ubiquitous in massive-data astrophysics (Richards & Szalay)
- Two solar observatories: SDO and ATST
- Terabytes of data per day

Image credit: NASA/SDO
Preprocessing ubiquitous in massive-data astrophysics (Richards & Szalay)

Two solar observatories: SDO and ATST

Terabytes of data per day

Raw data inaccessible (SDO) or completely unavailable (ATST)
Preprocessing ubiquitous in massive-data astrophysics (Richards & Szalay)

Two solar observatories: SDO and ATST

Terabytes of data per day

Raw data inaccessible (SDO) or completely unavailable (ATST)

Disputes in solar science community; cannot correct errors later on
Indirect observations

- Dust clouds are centers of star formation

Image credit: NASA/JPL
Indirect observations

- Dust clouds are centers of star formation
- Want to understand dynamics
Indirect observations

- Dust clouds are centers of star formation
- Want to understand dynamics
- Key relationship between temperature and dust density

Image credit: NASA/JPL
Indirect observations

- Dust clouds are centers of star formation
- Want to understand dynamics
- Key relationship between temperature and dust density
- Must estimate these properties; cannot directly observe

Image credit: NASA/JPL
Indirect observations

- Dust clouds are centers of star formation
- Want to understand dynamics
- Key relationship between temperature and dust density
- Must estimate these properties; cannot directly observe
- Incorrect preprocessing leads to backwards estimates of relationship (Kelly et al. 2012)

Image credit: NASA/JPL
Modern technologies measure thousands of genes or proteins at a time.
High-throughput biology

- Modern technologies measure thousands of genes or proteins at a time
- Sequencing and microarrays are most popular

Image credit: PNL
High-throughput biology

- Modern technologies measure thousands of genes or proteins at a time
- Sequencing and microarrays are most popular
- Measure brightness of points on array; infer gene expression

Image credit: PNL
High-throughput biology

- Modern technologies measure thousands of genes or proteins at a time
- Sequencing and microarrays are most popular
- Measure brightness of points on array; infer gene expression
- Sequencing brings its own, complex error processes

Image credit: PNL
Role of preprocessing — Microarrays

Standard methods based on heavily processed data
- Raw signals adjusted for background contamination
- Subsequent calibration for variation between arrays
- Then, statistical analysis of preprocessed results
Biology

Role of preprocessing — Microarrays

Standard methods based on heavily processed data

- Raw signals adjusted for background contamination
- Subsequent calibration for variation between arrays
- Then, statistical analysis of preprocessed results

Why not a joint model?

- Computational scale
- Complexity of measurement process
- Separation of knowledge and effort is needed
Pitfalls and improvements

Missing pieces

- Measures of uncertainty not retained
- Irreversible calibration
- Processed results often insufficient for follow-up
- E.g.: Observe $Y = S + B$, correct for B, pass only point estimate of log $S$. Problems?
Pitfalls and improvements

Missing pieces
- Measures of uncertainty not retained
- Irreversible calibration
- Processed results often insufficient for follow-up
- E.g.: Observe $Y = S + B$, correct for $B$, pass only point estimate of $\log S$. Problems?

Remedies
- Wring rich summaries from observation models
- Retain summaries of uncertainty (e.g. Rick Steven’s “graphs with probabilities” to replace alignments)
- Integrate richer information into downstream analyses
A model for two phases, in two phases

Building a framework to analyze and develop preprocessing techniques

Notation

- $Y \in \mathbb{R}^N$ are observed data
- $X \in \mathbb{R}^J$ are scientific variables of interest
- $\theta$ are parameters governing scientific process
- $T$ is the output from preprocessing
Model

A model for two phases, phase two

Data-generating process — Preprocessor’s model

\[ p(Y, X|\theta) = p_Y(Y | X) \cdot p_X(X | \theta) \]

- Conditional independence structure
- Separation of knowledge
- \(X\) are missing data
A model for two phases, phase two

Data-generating process — Preprocessor’s model

\[ p(Y, X | \theta) = p_Y(Y | X) \cdot p_X(X | \theta) \]

- Conditional independence structure
- Separation of knowledge
- \( X \) are missing data

Downstream analyst’s model

\[ p(T, Y, X | \theta) = p_T(T | Y) \cdot p_Y(Y | X) \cdot p_X(X | \theta) \]

- Additional layer of processing
- \( X \) and \( Y \) are missing data
Revisiting examples

Indirect measurements in astrophysics

- $Y$ are measurements from telescope, $X$ are true features of dust cloud
- $p_Y$ characterizes telescope’s responses
- $p_X$ and $\theta$ characterize structure of dust cloud
Revisiting examples

Indirect measurements in astrophysics
- Y are measurements from telescope, X are true features of dust cloud
- $p_Y$ characterizes telescope’s responses
- $p_X$ and $\theta$ characterize structure of dust cloud

Microarray measurement errors
- Y are measurements from array, X are true gene expressions
- $p_Y$ characterizes measurement error
- $p_X$ and $\theta$ characterize biological mechanisms
Defining multiphase procedures

Basic setup

- First phase provides $T = (\hat{X}, S)$
- Second phase has estimator $\hat{\theta}$ for each such $T$
- Different practical constraints induce different outputs
### Basic setup

- First phase provides $T = (\hat{X}, S)$
- Second phase has estimator $\hat{\theta}$ for each such $T$
- Different practical constraints induce different outputs

### Adaptation

- Second phase adapts; e.g., $T_0 = \hat{X}$ leads to simple mean, $T_1 = (\hat{X}, S)$ leads to weighted mean
- Extends to more than two constraints; e.g., aggregating data at multiple resolutions
- Need to regulate adaptation; for example, should do better with higher-resolution data
Regulating procedures

- Need sensible adaptation for theory and practice
Regulating procedures

- Need sensible adaptation for theory and practice
- Improve performance with more information
Regulating procedures

- Need sensible adaptation for theory and practice
- Improve performance with more information
  - Risk-monotonicity: deterministically more information ⇒ weakly lower risk
  - Self-efficiency (Meng 1994): No improvement from using less information
Regulating procedures

- Need sensible adaptation for theory and practice
- Improve performance with more information
  - Risk-monotonicity: deterministically more information ⇒ weakly lower risk
  - Self-efficiency (Meng 1994): No improvement from using less information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information
Regulating procedures

- Need sensible adaptation for theory and practice
- Improve performance with more information
  - Risk-monotonicity: deterministically more information \( \Rightarrow \) weakly lower risk
  - Self-efficiency (Meng 1994): No improvement from using less information
- Actually a strong constraint on mutual knowledge — no mis\underline{use} of additional information
- How to generate such procedures? Bayes rules from model \( P(Y|\theta) \), MLEs for such models (asymptotically)
Regulating procedures

- Need sensible adaptation for theory and practice
- Improve performance with more information
  - Risk-monotonicity: deterministically more information $\Rightarrow$ weakly lower risk
  - Self-efficiency (Meng 1994): No improvement from using less information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information
- How to generate such procedures? Bayes rules from model $P(Y|\theta)$, MLEs for such models (asymptotically)
- Models with principled estimation as generators of procedures
Role of constraints

Constraints are key

- Without tight constraints, dead ends and trivial results
- For example, optimal method simply computes optimal estimator with $Y$ then passes it on
- Pragmatic constraints can yield deep theory (e.g. MI)
Constraints

Role of constraints

Constraints are key
- Without tight constraints, dead ends and trivial results
- For example, optimal method simply computes optimal estimator with $Y$ then passes it on
- Pragmatic constraints can yield deep theory (e.g. MI)

Dual-use datasets
- If we could target $T(Y)$ to a single analysis, it’s easy
- Practically, want preprocessed data for multiple uses
- Want $T(Y)$ both for inference on $X$ and as input for further analyses
- Interpretability and modelability, not just efficiency
Two constraints for multiphase

**Not too much more work**

- Require complete-data estimator (using $X$) to be a version of multiphase estimator (e.g. nested models)
- For example, weighted least-squares regression when $X$ would call for unweighted regression
Two constraints for multiphase

Not too much more work

- Require complete-data estimator (using $X$) to be a version of multiphase estimator (e.g. nested models)
- For example, weighted least-squares regression when $X$ would call for unweighted regression

Spreading the load

- Require that first-phase procedures are distributable across researchers
- Build preprocessing on factored models for $X$
Factored models and sufficiency

- Even simple tasks in multiphase are quite complex
Factored models and sufficiency

- Even simple tasks in multiphase are quite complex
- Suppose we want to distribute preprocessing across multiple researchers, each with their own experiments
Factored models and sufficiency

- Even simple tasks in multiphase are quite complex
- Suppose we want to distribute preprocessing across multiple researchers, each with their own experiments
- Assume particular model $p_X(X|\theta)$
Factored models and sufficiency

- Even simple tasks in multiphase are quite complex
- Suppose we want to distribute preprocessing across multiple researchers, each with their own experiments
- Assume particular model $p_X(X|\theta)$
- Basic question: how can they determine what’s needed to maintain sufficiency for $\theta$?
Factored models and sufficiency

- Even simple tasks in multiphase are quite complex
- Suppose we want to distribute preprocessing across multiple researchers, each with their own experiments
- Assume particular model $p_X(X|\theta)$
- Basic question: how can they determine what’s needed to maintain sufficiency for $\theta$?
- Conversely, for which models $p_X$ do we preserve sufficiency with given preprocessing?
Factored models

Mathematical structure

Assume observation model factors by researcher \((i)\)

\[
p_Y(Y|X) = \prod_i p(Y_i|X_i)
\]
Factored models

Mathematical structure

- Assume observation model factors by researcher \((i)\)
  \[
  p_Y(Y|X) = \prod_i p(Y_i|X_i)
  \]
- Introduce factored working model for \(X\) using \(\eta = \{\eta_i\}\)
  \[
  p_W(X|\eta) = \prod_i p_W(X_i|\eta_i)
  \]
Factored models

### Mathematical structure

- Assume observation model factors by researcher ($i$)

$$p_Y(Y|X) = \prod_i p(Y_i|X_i)$$

- Introduce factored working model for $X$ using $\eta = \{\eta_i\}$

$$p_W(X|\eta) = \prod_i p_W(X_i|\eta_i)$$

- Each researcher models only their own data using

$$p_W(Y_i|\eta_i) = \int p(Y_i|X_i)p_W(X_i|\eta_i) \, dX_i$$
Factored models

Mathematical structure

- Assume observation model factors by researcher \((i)\)
  \[
  p_Y(Y|X) = \prod_i p(Y_i|X_i)
  \]
- Introduce factored working model for \(X\) using \(\eta = \{\eta_i\}\)
  \[
  p_W(X|\eta) = \prod_i p_W(X_i|\eta_i)
  \]
- Each researcher models only their own data using
  \[
  p_W(Y_i|\eta_i) = \int p(Y_i|X_i)p_W(X_i|\eta_i) dX_i
  \]
- Each researcher preprocesses observations \(Y_i\) into \(T_i\)
Factored models

Mathematical structure

- Assume observation model factors by researcher \((i)\)
  \[
  \rho_Y(Y|X) = \prod_i \rho(Y_i|X_i)
  \]
- Introduce factored working model for \(X\) using \(\eta = \{\eta_i\}\)
  \[
  \rho_W(X|\eta) = \prod_i \rho_W(X_i|\eta_i)
  \]
- Each researcher models only their own data using
  \[
  \rho_W(Y_i|\eta_i) = \int \rho(Y_i|X_i)\rho_W(X_i|\eta_i) \, dX_i
  \]
- Each researcher preprocesses observations \(Y_i\) into \(T_i\)
- Constraining \(T_i\) to be sufficient for \(\eta_i\)
Factored models

Mixture condition

- When will working model preserve sufficiency for $\theta$?
- Formally, when will sufficiency for $\eta$ imply sufficiency for $\theta$?
Factored models

Mixture condition

- When will working model preserve sufficiency for $\theta$?
- Formally, when will sufficiency for $\eta$ imply sufficiency for $\theta$?
- Not enough for working model to be (marginally) correct for each $Y_i$
Factored models

Mixture condition

- When will working model preserve sufficiency for $\theta$?
- Formally, when will sufficiency for $\eta$ imply sufficiency for $\theta$?
- Not enough for working model to be (marginally) correct for each $Y_i$
- Sufficient condition: mixture

$$p_X(X|\theta) = \int_H \prod_i p_W(X_i|\eta_i) \, dP(\eta|\theta)$$
Factored models

Implications and extensions

**Applied guidance**

- Not enough to reduce data based on a correctly-specified model
- Must look to models that include yours hierarchically
- However, can obtain results without sufficiency for $X$
Implications and extensions

Applied guidance
- Not enough to reduce data based on a correctly-specified model
- Must look to models that include yours hierarchically
- However, can obtain results without sufficiency for $X$

Theoretical loose-ends
- Mixture condition not necessary
- Counterexamples to necessity based on unparameterized dependence
Classical bounds

Classical results

Doing the best with what you get

- For fixed preprocessing, what bounds performance?
- In large samples, fraction of missing information \( F = \mathcal{I}_Y^{-1} \mathcal{I}_Y | T \) determines lower bound on variance
- Formally, relative excess variance converges to \( F \):
  \[
  \frac{\text{Var}(\hat{\theta}(T))^{-1} \text{Var}(\hat{\theta}(T) - \hat{\theta}(Y))}{\mathcal{F}} \rightarrow F
  \]
Doing the best with what you get

- For fixed preprocessing, what bounds performance?
- In large samples, fraction of missing information \((F = I_Y^{-1}I_{Y|T})\) determines lower bound on variance
- Formally, relative excess variance converges to \(F\):
  \[
  \text{Var}(\hat{\theta}(T))^{-1} \text{Var}(\hat{\theta}(T) - \hat{\theta}(Y)) \to F
  \]

Giving all that you can

- What is good preprocessing with \(\hat{\theta}(T)\) fixed?
- All* admissible \(T(Y)\) are (generalized) Bayes rules
- Extension of standard complete-class results
- Further bounds from multiple imputation (MI) theory
Recap

Goal

- Building foundation for multiphase inference
- Descended from theory of missing data
- Motivated by real problems and practical constraints
Recap

**Goal**
- Building foundation for multiphase inference
- Descended from theory of missing data
- Motivated by real problems and practical constraints

**A formal framework for multiphase theory**
- Defined model and multiphase procedures
- Constraints crucial for theoretical development
Recap

Goal
- Building foundation for multiphase inference
- Descended from theory of missing data
- Motivated by real problems and practical constraints

A formal framework for multiphase theory
- Defined model and multiphase procedures
- Constraints crucial for theoretical development

Theoretical cornerstones
- Condition for distributed preprocessing
- Performance bounds for multiphase settings
Coming attractions

Theory

- Evaluation of preprocessing methods for design and analysis
- Constrained optimality results for broadly-applicable multiphase strategies
Coming attractions

Theory
- Evaluation of preprocessing methods for design and analysis
- Constrained optimality results for broadly-applicable multiphase strategies

Applications
- Improved multiphase methods for biological and astronomical problems
- Multiphase-based computational strategies for massive data
Acknowledgments

- Xiao-Li Meng
- Edo Airoldi
- Art Dempster
- Stephen Blyth
- Harvard Statistics Dept.