The Potential and Perils of Preprocessing: A Multiphase Investigation

Alexander W Blocker

Department of Statistics
Harvard University

May 11, 2012

Outline

1. Perspective on preprocessing
2. Motivating examples
   - Astrophysics
   - Microarrays
3. Framework
   - Model
   - Procedures
   - Multiple imputation and constraints
4. Theoretical cornerstones
   - Factored models
   - Fraction of missing information
   - Complete-class
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

Defining preprocessing

- **Formally:** Transformations and reductions of observed data for subsequent analyses
- **Informally:** Everything that happens before statistical modeling
- Examples:
  - Aggregation
  - Smoothing
  - Calibration

Perils and promise

Destructive preprocessing

- Most non-trivial preprocessing is irreversible
- Not assumption-free
- Preprocessing decisions constrain all later analyses

Perils and promise

Destructive preprocessing

- Most non-trivial preprocessing is irreversible
- Not assumption-free
- Preprocessing decisions constrain all later analyses

Alleviating complexity

- Less complex modeling required
- Smaller data
- Separation of effort among analysts

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**

Massive data

- Preprocessing ubiquitous in massive-data astrophysics

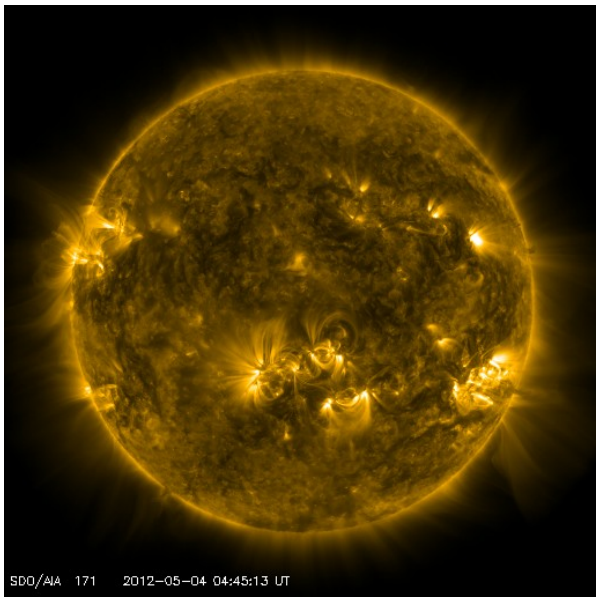

Image credit: NASA/SDO

Massive data

- Preprocessing ubiquitous in massive-data astrophysics
- Two solar observatories: SDO and ATST

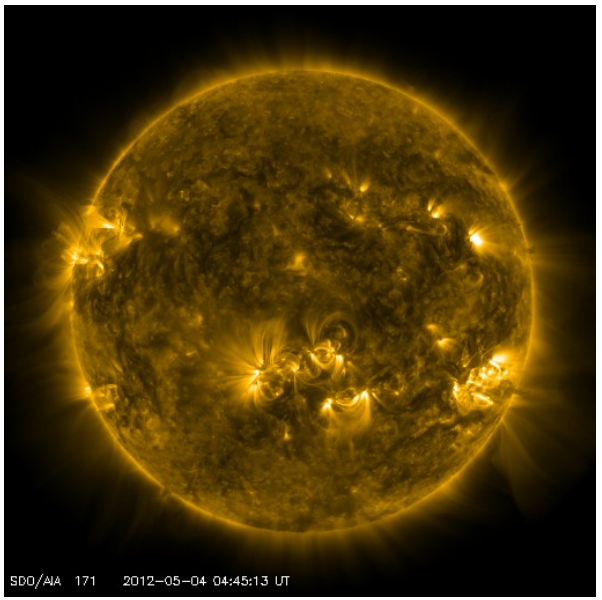

Image credit: NASA/SDO

Massive data

- Preprocessing ubiquitous in massive-data astrophysics
- Two solar observatories: SDO and ATST
- Terabytes of data per day

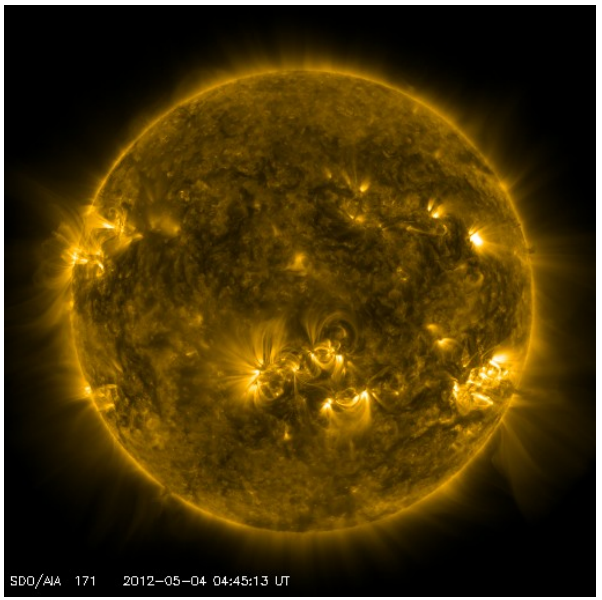

Image credit: NASA/SDO

Massive data

- Preprocessing ubiquitous in massive-data astrophysics
- Two solar observatories: SDO and ATST
- Terabytes of data per day
- Raw data inaccessible (SDO) or completely unavailable (ATST)

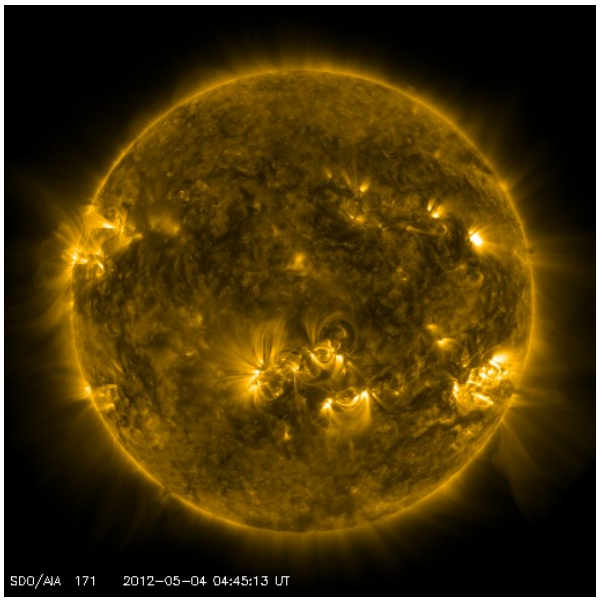

Image credit: NASA/SDO

Massive data

- Preprocessing ubiquitous in massive-data astrophysics
- 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

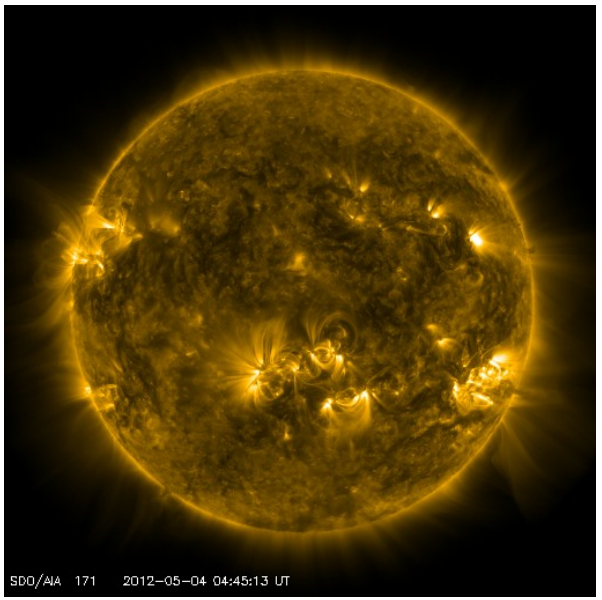

Image credit: NASA/SDO

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

Image credit: NASA/JPL

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

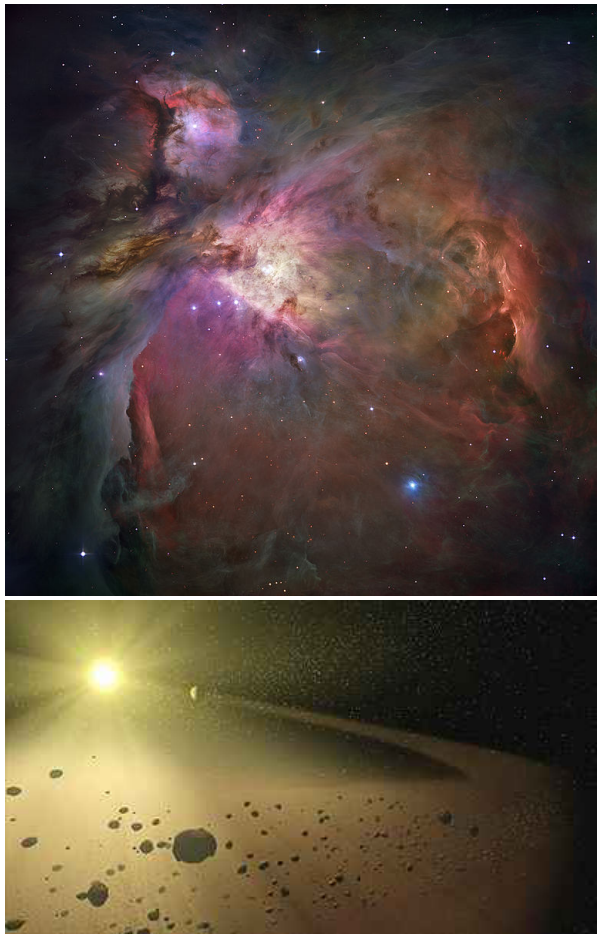

Image credit: NASA/JPL

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)

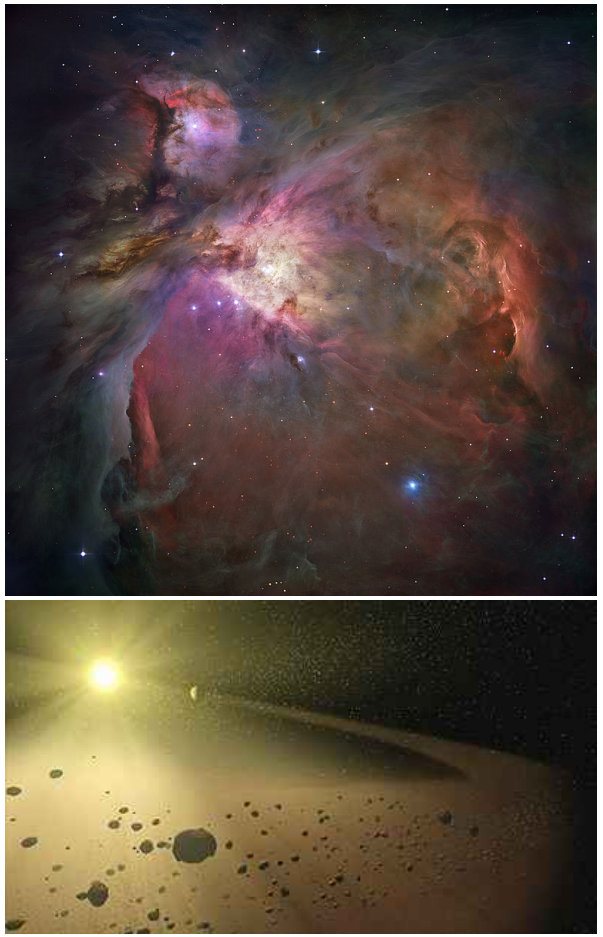

Image credit: NASA/JPL

Measuring gene expression

- Quantity of interest for many biological processes

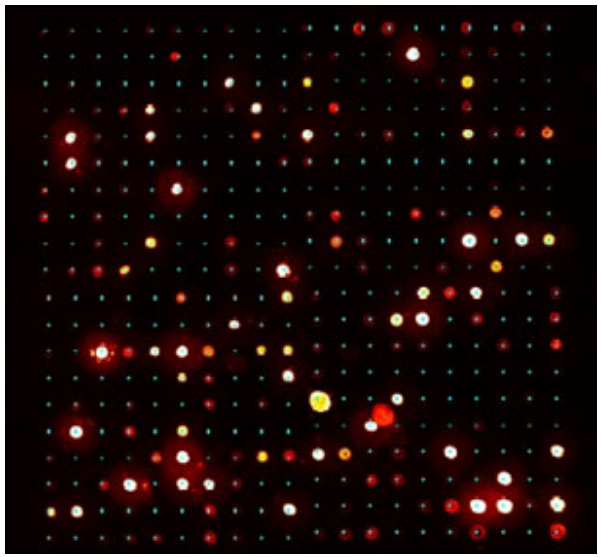

Image credit: PNL

Measuring gene expression

- Quantity of interest for many biological processes
- Want to measure relative expression between genes and experimental conditions

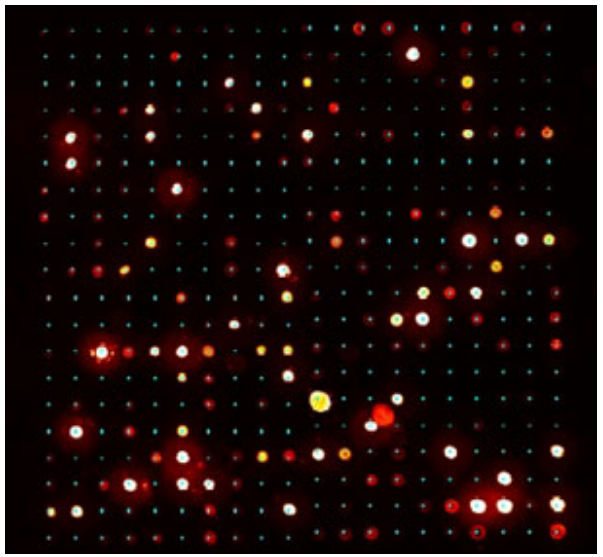

Image credit: PNL

Measuring gene expression

- Quantity of interest for many biological processes
- Want to measure relative expression between genes and experimental conditions
- Microarrays are popular technology for this

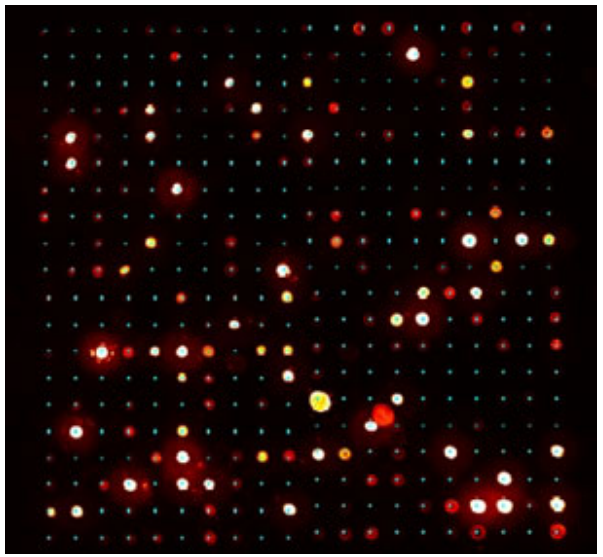

Image credit: PNL

Measuring gene expression

- Quantity of interest for many biological processes
- Want to measure relative expression between genes and experimental conditions
- Microarrays are popular technology for this
- Measure brightness of points on array; infer expression

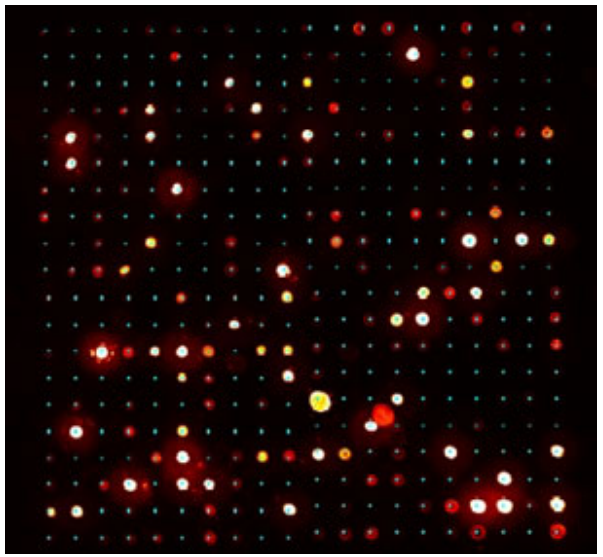

Image credit: PNL

Role of preprocessing

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 (e.g. SAM)
- Processed results also default output from archives (e.g. NIH's GEO)

Role of preprocessing

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 (e.g. SAM)
- Processed results also default output from archives (e.g. NIH's GEO)

Why not a joint model?

- Computational scale
- Complexity of measurement process
- Separation of knowledge and effort

Pitfalls and improvements

Missing pieces

- Measures of uncertainty and error models not used throughout analysis
- Other calibration steps irreversible
- Processed results often insufficient for follow-up

Pitfalls and improvements

Missing pieces

- Measures of uncertainty and error models not used throughout analysis
- Other calibration steps irreversible
- Processed results often insufficient for follow-up

Remedies

- Use model-based preprocessing
- Retain summaries of uncertainty
- Integrate into downstream analyses

A model for two phases, in two phases

Notation

- $Y \in \mathbb{R}^N$ are observed data
- $X \in \mathbb{R}^J$ are scientific variables of interest
- θ are parameters governing scientific process of interest

A model for two phases, in two phases

Notation

- $Y \in \mathbb{R}^N$ are observed data
- $X \in \mathbb{R}^J$ are scientific variables of interest
- θ are parameters governing scientific process of interest

Overall model

$$p(Y, X|\theta) = p_Y(Y \mid X) \cdot p_X(X \mid \theta)$$

- Conditional independence structure
- Separation of knowledge

A model for two phases, phase two

Scientific model

$$p_X(X \mid \theta)$$

- Encapsulates scientific process of interest
- Basis for idealized analysis

A model for two phases, phase two

Scientific model

$$p_X(X \mid \theta)$$

- Encapsulates scientific process of interest
- Basis for idealized analysis

Observation model

$$p_Y(Y \mid X)$$

- Encapsulates measurement processes, including nuisance parameters, and other knowledge
- Basis for preprocessing (but not usually enough)

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 θ 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 θ 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 θ 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

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

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

Ordering outputs

- Need to compare outputs under different constraints
- Start with deterministic dependence
- Let $T \preceq \tilde{T}$ indicate T is a deterministic function of $\tilde{T}$

Regulating procedures

Ordering outputs

- Need to compare outputs under different constraints
- Start with deterministic dependence
- Let $T \preceq \tilde{T}$ indicate T is a deterministic function of $\tilde{T}$

Definition (Procedure)

A *multiphase estimation procedure* $\mathcal{P}$ is a set of estimators $\{\hat{\theta}_k(T_k) : k \in C\}$ indexed by the set of constraints C, where T_k corresponds to output under the kth first-phase constraint.

Regulating procedures

Ordering outputs

- Need to compare outputs under different constraints
- Start with deterministic dependence
- Let $T \preceq \tilde{T}$ indicate T is a deterministic function of $\tilde{T}$

Definition (Procedure)

A *multiphase estimation procedure* $\mathcal{P}$ is a set of estimators $\{\hat{\theta}_k(T_k) : k \in C\}$ indexed by the set of constraints C, where T_k corresponds to output under the kth first-phase constraint.

- Extends analysis procedure definition of Meng (1994) beyond complete and missing data

Risk monotonicity

- Using given ordering, can restrict procedures based on coherence

Risk monotonicity

- Using given ordering, can restrict procedures based on coherence

Definition (Risk monotonicity)

A multiphase estimation procedure $\mathcal{P}$ is *risk monotone* with respect to a loss function L if, for all pairs of outputs $T, \tilde{T}$ such that $T \preceq \tilde{T}$ implies $R(\hat{\theta}(\tilde{T}), L) \leq R(\hat{\theta}(T), L)$.

Risk monotonicity

- Using given ordering, can restrict procedures based on coherence

Definition (Risk monotonicity)

A multiphase estimation procedure $\mathcal{P}$ is *risk monotone* with respect to a loss function L if, for all pairs of outputs $T, \tilde{T}$ such that $T \preceq \tilde{T}$ implies $R(\hat{\theta}(\tilde{T}), L) \leq R(\hat{\theta}(T), L)$.

- Similar concept to self-efficiency (Meng 1994)

Risk monotonicity

- Using given ordering, can restrict procedures based on coherence

Definition (Risk monotonicity)

A multiphase estimation procedure $\mathcal{P}$ is *risk monotone* with respect to a loss function L if, for all pairs of outputs $T, \tilde{T}$ such that $T \preceq \tilde{T}$ implies $R(\hat{\theta}(\tilde{T}), L) \leq R(\hat{\theta}(T), L)$.

- Similar concept to self-efficiency (Meng 1994)
- Addresses data reduction. Side information (e.g. on models used) needs additional structure

Understanding risk monotonicity

- Seemingly minimal constraint — no worse with more information

Understanding risk monotonicity

- Seemingly minimal constraint — no worse with more information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information

Understanding risk monotonicity

- Seemingly minimal constraint — no worse with more information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information
- How to generate such procedures?

Understanding risk monotonicity

- Seemingly minimal constraint — no worse with more information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information
- How to generate such procedures?
  - Bayes rules from marginal model $P(Y|\theta)$
  - MLEs from such models (asymptotically)
  - More practically, Bayes rules from model $P(T|\theta)$ cover all functions of T

Understanding risk monotonicity

- Seemingly minimal constraint — no worse with more information
- Actually a strong constraint on mutual knowledge — no **misuse** of additional information
- How to generate such procedures?
  - Bayes rules from marginal model $P(Y|\theta)$
  - MLEs from such models (asymptotically)
  - More practically, Bayes rules from model $P(T|\theta)$ cover all functions of T
- Models with principled estimation as generators

Connection to multiple imputation

- Multiple imputation (MI) is clearly a multiphase procedure

Connection to multiple imputation

- Multiple imputation (MI) is clearly a multiphase procedure
- Complete data correspond to X, observed to Y

Connection to multiple imputation

- Multiple imputation (MI) is clearly a multiphase procedure
- Complete data correspond to X, observed to Y
- MI provides lower bound on multiphase performance

Connection to multiple imputation

- Multiple imputation (MI) is clearly a multiphase procedure
- Complete data correspond to X, observed to Y
- MI provides lower bound on multiphase performance
  - Passing m completed datasets of size J each
  - With congeniality, obtain estimator with usual variance (Rubin's rule)
  - Without congeniality, have results of Xie & Meng (2012)

Connection to multiple imputation

- Multiple imputation (MI) is clearly a multiphase procedure
- Complete data correspond to X, observed to Y
- MI provides lower bound on multiphase performance
  - Passing m completed datasets of size J each
  - With congeniality, obtain estimator with usual variance (Rubin's rule)
  - Without congeniality, have results of Xie & Meng (2012)
- Shared theoretical foundations, distinct objectives — reduction vs. expansion

Role of constraints

Constraints are key

- Multiphase theory hinges on procedural constraints
- Without tight constraints, many inquiries yield trivial results
- For example, optimal method simply computes optimal estimator with Y then passes it on

Role of constraints

Constraints are key

- Multiphase theory hinges on procedural constraints
- Without tight constraints, many inquiries yield trivial results
- For example, optimal method simply computes optimal estimator with Y then passes it on

MI highlights importance of constraints

- For MI, first phase passes predictive draws of X
- Downstream analyst repeats complete-data procedure and combines
- Pragmatic constraints enable deep theory

Two constraints for multiphase

Not too much more work

- Require that complete-data estimator is special case of multiphase estimator
- For example, weighted least-squares regression when complete data call for unweighted regression
- Tight constraint on multiphase procedures

Two constraints for multiphase

Not too much more work

- Require that complete-data estimator is special case of multiphase estimator
- For example, weighted least-squares regression when complete data call for unweighted regression
- Tight constraint on multiphase procedures

Spreading the load

- Require that first-phase procedures be distributable across researchers
- Base 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 θ?

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 θ?
- Conversely, for which models p_X do we preserve sufficiency with given preprocessing?

Mathematical structure

- Assume observation model factors by researcher (i)

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

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 $\eta = \{\eta_i\}$

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

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

$$p_W(Y_i|\eta_i) = \int p(Y_i|X_i) p_W(X_i|\eta_i) \, \mathrm{d}X_i$$

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

$$p_W(Y_i|\eta_i) = \int p(Y_i|X_i) p_W(X_i|\eta_i) \, \mathrm{d}X_i$$

- Each researcher preprocesses observations Y_i into T_i

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

$$p_W(Y_i|\eta_i) = \int p(Y_i|X_i) p_W(X_i|\eta_i) \, \mathrm{d}X_i$$

- Each researcher preprocesses observations Y_i into T_i
- Constraining T_i to be sufficient for η_i

Mixture condition

- When will working model preserve sufficiency for θ?
- Formally, when will sufficiency for η imply sufficiency for θ?

Mixture condition

- When will working model preserve sufficiency for θ?
- Formally, when will sufficiency for η imply sufficiency for θ?
- Not enough for working model to be (marginally) correct for each Y_i

Mixture condition

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

$$p_X(X|\theta) = \int_{\mathrm{H}} \prod_i p_W(X_i|\eta_i) \, \mathrm{d}P(\eta|\theta)$$

Implications and extensions

Applied guidance

- Not enough to reduce data based on correctly-specified model
- Need to look to other 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 correctly-specified model
- Need to look to other 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

Doing the best with what you get

For fixed preprocessing T, what determines performance?

- In large samples, fraction of missing information (FMI) determines upper bound
- Assume analyst uses MLE for θ based upon the correctly-specified model for Y

Doing the best with what you get

For fixed preprocessing T, what determines performance?

- In large samples, fraction of missing information (FMI) determines upper bound
- Assume analyst uses MLE for θ based upon the correctly-specified model for Y

Mathematical setup

- Decompose log-likelihood $\ell_Y(\theta) = \ell_T(\theta) + \ell_{Y|T}(\theta)$
- Define corresponding score and Fisher information for each part of decomposition
- Define fraction of missing information $F = I_{Y|T}/I_Y$

Formal result

Theorem (Upper bound)

Assuming standard regularity conditions for the log (conditional) likelihoods $\ell_Y(\theta)$, $\ell_T(\theta)$, *and* $\ell_{Y|T}(\theta)$,

$$\text{Var}\left(\hat{\theta}(T) - \hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to F \text{ and}$$

$$\text{Var}\left(\hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to 1 - F \text{ as } N \to \infty.$$

Formal result

Theorem (Upper bound)

Assuming standard regularity conditions for the log (conditional) likelihoods $\ell_Y(\theta)$, $\ell_T(\theta)$, *and* $\ell_{Y|T}(\theta)$,

$$\text{Var}\left(\hat{\theta}(T) - \hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to F \text{ and}$$

$$\text{Var}\left(\hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to 1 - F \text{ as } N \to \infty.$$

- Straightforward extension to multivariate case

Formal result

Theorem (Upper bound)

Assuming standard regularity conditions for the log (conditional) likelihoods $\ell_Y(\theta)$, $\ell_T(\theta)$, *and* $\ell_{Y|T}(\theta)$,

$$\text{Var}\left(\hat{\theta}(T) - \hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to F \text{ and}$$

$$\text{Var}\left(\hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to 1 - F \text{ as } N \to \infty.$$

- Straightforward extension to multivariate case
- Standard caveats on asymptotics with multilevel models

Formal result

Theorem (Upper bound)

Assuming standard regularity conditions for the log (conditional) likelihoods $\ell_Y(\theta)$, $\ell_T(\theta)$, *and* $\ell_{Y|T}(\theta)$,

$$\text{Var}\left(\hat{\theta}(T) - \hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to F \text{ and}$$

$$\text{Var}\left(\hat{\theta}(Y)\right) / \text{Var}\left(\hat{\theta}(T)\right) \to 1 - F \text{ as } N \to \infty.$$

- Straightforward extension to multivariate case
- Standard caveats on asymptotics with multilevel models
- Large-sample result; weak guidance for finite-sample settings

Giving all that you can

- Now, consider the reverse

Giving all that you can

- Now, consider the reverse
- Fix second-phase procedure $\hat{\theta}(T)$ and form of input

Giving all that you can

- Now, consider the reverse
- Fix second-phase procedure $\hat{\theta}(T)$ and form of input
- What does optimal preprocessing look like?

Giving all that you can

- Now, consider the reverse
- Fix second-phase procedure $\hat{\theta}(T)$ and form of input
- What does optimal preprocessing look like?

Theorem (Optimal preprocessing)

Consider a smooth, strictly convex loss function L. Then, under appropriate regularity conditions, if $\hat{\theta}(T)$ is a smooth function of T, all admissible procedures for generating T are Bayes or generalized Bayes rules.

Giving all that you can

- Now, consider the reverse
- Fix second-phase procedure $\hat{\theta}(T)$ and form of input
- What does optimal preprocessing look like?

Theorem (Optimal preprocessing)

Consider a smooth, strictly convex loss function L. Then, under appropriate regularity conditions, if $\hat{\theta}(T)$ is a smooth function of T, all admissible procedures for generating T are Bayes or generalized Bayes rules.

- Extension of standard complete-class results

Practical and theoretical relevance

Difficult to apply

- In general, very difficult to compute such rules
- Specific to particular second-phase procedure
- Not a feasible target for applied use

Practical and theoretical relevance

Difficult to apply

- In general, very difficult to compute such rules
- Specific to particular second-phase procedure
- Not a feasible target for applied use

Theoretical guidance

- Shows difficulty of optimality even with stringent constraints
- Importance of pragmatic constraints
- Role of flexibility in second-phase procedures

Recap

Goal

- Building foundation for multiphase inference
- Motivated by statistical practice and pragmatic constraints

Recap

Goal

- Building foundation for multiphase inference
- Motivated by statistical practice and pragmatic constraints

A formal framework for multiphase theory

- Defined model and multiphase procedures
- Constraints crucial for theoretical development

Recap

Goal

- Building foundation for multiphase inference
- Motivated by statistical practice and pragmatic 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

Context — Fisher vs. Wald

- Wald advocated decision-theoretic statistics

Context — Fisher vs. Wald

- Wald advocated decision-theoretic statistics
- Fisher strongly objected. As Savage (1976) noted,

  In later works, he hinted that it might have its mundane applications for the slaves of Wall Street and the Kremlin (Fisher 1955 p70) but not for a free scientist in search of truth.

Context — Fisher vs. Wald

- Wald advocated decision-theoretic statistics
- Fisher strongly objected. As Savage (1976) noted,

  In later works, he hinted that it might have its mundane applications for the slaves of Wall Street and the Kremlin (Fisher 1955 p70) but not for a free scientist in search of truth.

- Contrasted decision-making with growing scientific knowledge

Reconciliation?

- Distinction is fundamental to multiphase inference

Reconciliation?

- Distinction is fundamental to multiphase inference
- Passing “optimal” estimates to later analysts is not a recipe for valid inference

Reconciliation?

- Distinction is fundamental to multiphase inference
- Passing “optimal” estimates to later analysts is not a recipe for valid inference
- Decision theory forms basis of multiphase

Reconciliation?

- Distinction is fundamental to multiphase inference
- Passing “optimal” estimates to later analysts is not a recipe for valid inference
- Decision theory forms basis of multiphase
- Can view Fisher’s objections as a rejection of myopic loss in multiphase settings

Scientific inference and logic

- Links to Dempster's thinking on statistical foundations and logic of inference

Scientific inference and logic

- Links to Dempster's thinking on statistical foundations and logic of inference
- Want to understand logic of multiphase inference — "what we can say" when inference is separated

Scientific inference and logic

- Links to Dempster's thinking on statistical foundations and logic of inference
- Want to understand logic of multiphase inference — "what we can say" when inference is separated
- Bringing the previously informal and unmodeled into a coherent framework

Scientific inference and logic

- Links to Dempster's thinking on statistical foundations and logic of inference
- Want to understand logic of multiphase inference — "what we can say" when inference is separated
- Bringing the previously informal and unmodeled into a coherent framework
- Role of (largely hidden) subjective assumptions in statistical inference

Extensions to come

Theory

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

Extensions to come

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
- Art Dempster
- Stephen Blyth
- Harvard Statistics Faculty
- Paula Griffin