Big Data: What are they and where are we headed with them

Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health

ACCORDS Big Data Workshop
Big Data: Wikipedia definition

- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."
“Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing.”
Although use of the term 'data science' has exploded in business environments, many academics and journalists see no distinction between data science and statistics. Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced 'business analytics' in contexts such as graduate degree programs. In the question-and-answer section of his keynote address at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, ‘I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn’t berate the term statistician’.”
Other References

- D. Donoho paper
Google Trends, part I
Google Trends, Part II

Interest over time

![Google Trends Chart for Data Science, Big Data, Machine Learning, Statistics](chart.png)
Big Data Characteristics: the 4 V’s (from http://ibmbigdatahub.com)

**Volume**
- Scale of Data
  - 40 ZETTABYTES (143 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005
  - 6 BILLION PEOPLE own 6 BILLION phones
  - WORLD POPULATION: 7 BILLION

**Velocity**
- Analysis of Streaming Data
  - By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS - almost 2.5 connections per person on earth

**Variety**
- Different Forms of Data
  - 30 BILLION PIECES OF CONTENT are shared on Facebook every month
  - 4 BILLION+ HOURS OF VIDEO are watched on YouTube each month
  - 400 MILLION TWEETS are sent per day by about 200 million monthly active users

**Veracity**
- Uncertainty of Data
  - 27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

**4 by 4 V’s**
- **Volume**: Scale of Data
- **Velocity**: Analysis of Streaming Data
- **Variety**: Different Forms of Data
- **Veracity**: Uncertainty of Data

**The FOUR V’s of Big Data**

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.3 million in the United States

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES (1,600 BILLION GIGABYTES)

By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Sources: McKinsey Global Institute, Twitter, Coca, Gartner, EMC, SAS, IBM, MEPTEC, QAS
Unique aspects of big data (IBM)

Extracting business value from the 4 V’s of big data

Volume: Scale of data
- Every day we create 2.5 quintillion bytes of data
- 90% of today’s data has been created in just the last 2 years
- (enough to fill 10 million Blu-ray discs)

Velocity: Speed of data
- Every 60 seconds there are:
  - 72 hours of footage uploaded to YouTube
  - 216,000 Instagram posts
  - 204,000,000 emails sent
- 50,000 GB/second is the estimated rate of global Internet traffic by 2018

Veracity: Certainty of data
- 1 in 3 business leaders don’t trust the information they use to make decisions

 Variety: Diversity of data
- 80% of data growth is video, images and documents
- 90% of generated data is "unstructured"
- This includes tweets, photos, customer purchase histories and customer service calls

The fifth “V”?

-$3.1 trillion is the estimated amount of money that poor data quality costs the US economy per year

[Scale of data: 1 in 3 business leaders don’t trust the information they use to make decisions]

$3.1 trillion is the estimated amount of money that poor data quality costs the US economy per year

[Scale of data: 1 in 3 business leaders don’t trust the information they use to make decisions]

90% of generated data is "unstructured"

This includes tweets, photos, customer purchase histories and customer service calls

[Velocity: Every 60 seconds there are:
- 72 hours of footage uploaded to YouTube
- 216,000 Instagram posts
- 204,000,000 emails sent
- 50,000 GB/second is the estimated rate of global Internet traffic by 2018]

Unlock the value of your big data.
Start here: ibm.co/technologyplatform
A conceptual framework for big data (Fayyad et al., KDD, 1996)
Some opinions

- Context (data use case) matters!!!!
- Really big data cannot be processed on a desktop straightforwardly
- Often a dynamic aspect that makes things computationally intensive
- Fundamentally, many of the questions we want big data to solve are causal in nature
- Identifying sources of variation and bias are key!!!
An example: Google Flu Trends models

- Idea: monitor Google users “health behaviors” online in order to predict where flu might be happening
- As a concrete example, if users in Nebraska are querying “flu” and related terms, we use that as a data point in the analysis
- Use logistic regression to predict the likelihood of a flu-related illness in a given geographic region based on a normalized measure of queries in that area
- Big data aspects: Google is crunching trillions of time series points in a dynamic manner to build the predictive model.
Google Flu Trends models: Predictions from Ginsburg et al. (2014, Nature)
Google Flu Trends models: issues

- Ethics: Is the privacy of individuals maintained with big data?
- Letter from Electronic Privacy Center to Google CEO:

Dear Dr. Schmidt,

We are writing to you regarding the Google Flu Trends initiative, a new web tool that may make it possible to detect flu outbreaks before they would otherwise be reported. This is an important use of user-generated data and could substantially support public health efforts in the United States and around the world.

At the same time, there is an obvious privacy concern. Search histories reveal personal information, and medical inquiries are particularly sensitive. A search for “flu symptoms” could as easily be a search for “AIDS symptoms” or “Ritalin” or “Paxil.”

In the aggregate, the data reveals useful trends and should be available for appropriate uses. But if disclosed and linked to a particular user, there could be adverse consequences for education, employment, insurance, and even travel. The disclosure of such information could also have a chilling effect on Internet users who may be reluctant to seek out important medical information online if they are concerned that their search histories will be revealed to others.
More issues: Lazer et al. (2015, Science)

Google estimates more than double CDC estimates

Google starts estimating high 100 out of 108 weeks
What happened in July 2011?

1. Google expanded scope of search terms related to the flu; this might induce bias
2. Other thing: Google modified search algorithm, and that might also be affecting the predictive model
3. Because Google is a private company, there is no possibility of a completely transparent and **reproducible** manner for generating the results
Recap of key points of Lazer et al., 2015

- Emphasis on need for replicability with big data
- Benefit versus cost of big data (simple model for flu prediction, so need to consider costs versus benefits of using big data)
- There is still a role for ‘small’ data in the big data era.
What is data science?

- From http://drewconway.com/

![Venn Diagram](image-url)
Inference versus Prediction: a big divide

- One staple of data science: machine learning
- Machine learning represents a class of algorithms in which the goal is to make a prediction
- Q&A with Jeremy Howard of Kaggle.
  Q: “Can you see any downside to the data-driven, black-box approach that dominates on Kaggle?”
  A: “Some people take the view that you don’t end up with a richer understanding of the problem. But that’s just not true: The algorithms tell you what’s important and what’s not. You might ask why those things are important, but I think that’s less interesting. You end up with a predictive model that works. There is not too much to argue about there.”
Now for the inference point of view:

Quote from authors of Doing Data Science (p. 115) “The science in data science is - given raw data, constraints, and a problem statement - how to navigate through that maze and make the best choices. Every design choice you make can be formulated as a hypothesis, against which you will use rigorous testing and experimentation to either validate or refute.”

Cool paper: The “two cultures paper” by Leo Breiman in Statistical Science
Future of Statistics? (Chamandy et al., 2015, *American Statistician*)

![Diagram of data processing and analysis]
A conceptual framework for big data (Fayyad et al., KDD, 1996)
Steps in the KDD process

1. “First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint.”

2. “Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.”

3. “Third is data cleaning and preprocessing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.”
Steps in the KDD process (cont’d.)

1. “Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task.”
2. “Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method.”
3. “Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns.”
4. “Seventh: data mining”
5. “Eighth: interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration.”
6. “Ninth: acting on discovered knowledge”
Types of tasks that arise in Data Science

- Regression model fitting
- Machine Learning:
  1. Supervised learning: using information to predict an outcome
  2. Unsupervised learning: using information to identify subgroups
  3. Recommendation system:
  4. Anomaly detection: identification of observations that deviate from the norm
Supervised Learning

- Goal: Given a training dataset with $X$’s and $Y$’s, build a model to predict $Y$ given $X$
- Then, take a new $X$ that wasn’t in the dataset and predict $Y$ for that person
- Many public challenges do this now (e.g., Kaggle, DREAM, Innocentive)
- This can be done using regression
- ”Fancier” machine learning tools: support vector/kernel machine methods, random forests/ensemble methods, K-nearest neighbors, tree methods
Supervised Learning: issues

- Key assumption: test observation comes from the same distribution as the training data
- This gets violated in many situations
- What is the implication for predictions from supervised learning algorithms?
Unsupervised Learning

- Goal: Find a subgroup of subjects based on X’s
- This is done if the goal is to treat subgroups differently
- Machine learning tools: spectral clustering, manifold clustering, k-means, mixture models, hierarchical clustering, singular value decomposition
Unsupervised Learning output (stackoverflow.com)
Unsupervised Learning: issues

- How many clusters?
- How to define distance between observations?
- Lots of variants on machine learning (i.e., the topics in the Ph.D. projects)
Recommendation engines

- These are ubiquitous!
- Given your purchases, you get recommended things based on customers who have made similar purchases
- Concept: given your purchase history, identify similar customers and “recommend” their purchases that you have not made
- Terminology: Collaborative/content-based filtering
Recommendation engines (Figure 8-1 from DDS)

Figure 8-1. Bipartite graph with users and items (television shows) as nodes
Recommendation engines: Issues

- Curse of dimensionality
- Overfitting
- Sparseness
- Measurement error
- Missing data
- Computational cost
Causal inference has recently received intense interest in biomedical studies.

In these settings, the treatment is not randomly assigned and is subject to self-selection/confounding.

A very important modelling strategy was proposed 30 years ago by Rosenbaum and Rubin (1983), termed the propensity score.

In words, the propensity score is defined as the probability of receiving treatment, given covariates.

Conditional on propensity score, one achieves “covariate balance” on observed covariates.
Potential Outcomes: Notation and Assumptions

- Let $T \in \{0, 1\}$ denote the treatment
- Let $\{Y(0), Y(1)\}$ denote the potential outcomes for $Y$ under each of the treatments
- Targets of estimation:
  \[
  ACE = E[Y_i(1) - Y_i(0)]
  \]
- Strong Ignorability of Treatment Assignment Assumption (SITA):
  \[
  T \perp \{Y(0), Y(1)\} | X
  \]
  where $X$ are covariates
Potential Outcomes (cont’d.)

- Propensity score: \( e(X) = P(T = 1|X) \)
- SITA implies:
  \[ T \perp \{Y(0), Y(1)\} | e(X) \]
- Estimate propensity score using
  - logistic regression
  - machine learning methods
Questions addressed today

**Q1:** Why is the model/variable selection problem different for causal inference problems?

**Q2:** How can machine learning methods be utilized for causal inference problems?
Different from

Targeted Learning
Causal Inference for Observational and Experimental Data
There are two models: the outcome model and the propensity score model.

The outcome model defines the scientific estimand and can be defined in terms of potential outcomes.

Propensity score model:
- We do not seek to interpret the model for the propensity score.
- We only use the estimated probabilities from the fitted model.
For example, ACE corresponds to

$$Y(1) - Y(0) = \tau + \epsilon,$$

which is different from

$$Y(1) - Y(0) = \tau^* + \delta'X + \epsilon^*$$

- $\tau^*$ in the latter model defines an $X$–specific causal effect
- This is all about scientific **estimands**, not estimates
Propensity score model

- One main goal is overlap of treatment and control groups

**Figure:** Distribution of covariate $X$ for treatment and control groups. The blue line denotes the kernel density estimation for $X$ in the $T = 1$ group, while the magenta line represents the kernel density estimate for $X$ in the $T = 0$ group. The bars represent the histogram of $X$ regardless of the treatment groups.
Modelling: remarks

- Two models which serve different goals
- One goal of propensity score model: balance in covariate distribution between treatment and control groups
- This runs counter to most algorithms that try to separate groups
- Optimizing covariate balance leads to different modelling procedures (Zhu et al., 2015; Imai and Ratkovic, 2014).
What variables to include in a propensity score model?

- Advice from Rosenbaum and Rubin: include everything associated with outcome $Y$
- Advice from Pearl: think about a graph structure for relating the variables, do NOT include “colliders”
- Many simulation studies reported in the literature, e.g., Brookhart et al. (2006)
- Recent innovation: Bayes model averaging (Wang et al., 2012; Zigler and Dominici, 2013; Lefevbre et al., 2015) and SuperLearner (Polley et al., 2007)
### Causal Inference as Missing Data Problem

- **Data visualization**

<table>
<thead>
<tr>
<th>$Y(0)$</th>
<th>$Y(1)$</th>
<th>$T$</th>
<th>$X_1$</th>
<th>⋯</th>
<th>$X_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>$y_1$</td>
<td>1</td>
<td>$x_{11}$</td>
<td>⋯</td>
<td>$x_{1p}$</td>
</tr>
<tr>
<td>$y_2$</td>
<td>?</td>
<td>0</td>
<td>$x_{21}$</td>
<td>⋯</td>
<td>$x_{2p}$</td>
</tr>
<tr>
<td>?</td>
<td>$y_3$</td>
<td>1</td>
<td>$x_{31}$</td>
<td>⋯</td>
<td>$x_{3p}$</td>
</tr>
<tr>
<td>?</td>
<td>$y_4$</td>
<td>1</td>
<td>$x_{41}$</td>
<td>⋯</td>
<td>$x_{4p}$</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>$y_n$</td>
<td>?</td>
<td>0</td>
<td>$x_{n1}$</td>
<td>⋯</td>
<td>$x_{np}$</td>
</tr>
</tbody>
</table>
The missing data mechanism and SITA assumption

\[ \{ Y(0), Y(1) \} \perp T \mid X \]

suggests the following two-step algorithm

1. Fill in missing responses (imputation)
2. Perform variable selection on the ‘complete’ data
Inherently, variable selection for causal inference involves a multivariate response variable.

Our approach: impute potential outcomes, take difference in potential outcomes and run a LASSO.

This is an application of the predictive LASSO idea of Tran et al. (2012).

Combines missing data and LASSO techniques.

Imputation step is critical.

Best approach: do multiple imputation and LASSO.
Data Application: RHC study

- Data from Connors et al. (1996)
- Goal: study effect of right-heart catherization on survival
- Response: 30-day survival
- 74 variables in the dataset, but here, we consider 21 variables
Potential Outcomes Framework

Data Application (cont.’d)
Table: Estimated Coefficients by LASSO and the Bootstrap Standard Errors

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Estimated Coefficient</th>
<th>Bootstrap SE</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>$-3.86 \times 10^{-4}$</td>
<td>$1.86 \times 10^{-4}$</td>
<td>$(-7.73 \times 10^{-4}, -1.34 \times 10^{-4})$</td>
</tr>
<tr>
<td>hema</td>
<td>$-6.13 \times 10^{-2}$</td>
<td>$2.78 \times 10^{-2}$</td>
<td>$(-1.23 \times 10^{-1}, -2.63 \times 10^{-2})$</td>
</tr>
<tr>
<td>card</td>
<td>$7.84 \times 10^{-3}$</td>
<td>$7.56 \times 10^{-3}$</td>
<td>$(-7.33 \times 10^{-3}, 2.53 \times 10^{-2})$</td>
</tr>
</tbody>
</table>
### Table: Causal Risk Difference Estimates in Different Strata

<table>
<thead>
<tr>
<th>Strata</th>
<th>bp &lt; 78.2, card=Yes</th>
<th>bp &lt; 78.2, card=No</th>
<th>bp ≥ 78.2, card=Yes</th>
<th>bp ≥ 78.2, card=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>hema &lt; 30, ph &lt; 7.4</td>
<td>0.160(0.083)</td>
<td>0.089(0.053)</td>
<td>-0.159(0.127)</td>
<td>-0.003 (0.102)</td>
</tr>
<tr>
<td></td>
<td>0.056</td>
<td>0.093</td>
<td>0.217</td>
<td>0.973</td>
</tr>
<tr>
<td>hema &lt; 30, ph ≥ 7.4</td>
<td>0.047 (0.088)</td>
<td>-0.052(0.058)</td>
<td>0.062(0.136)</td>
<td>0.003(0.081)</td>
</tr>
<tr>
<td></td>
<td>0.597</td>
<td>0.368</td>
<td>0.649</td>
<td>0.971</td>
</tr>
<tr>
<td>hema ≥ 30, ph &lt; 7.4</td>
<td>0.018(0.069)</td>
<td>-0.001(0.065)</td>
<td>0.139(0.108)</td>
<td>0.094(0.162)</td>
</tr>
<tr>
<td></td>
<td>0.792</td>
<td>0.991</td>
<td>0.200</td>
<td>0.560</td>
</tr>
<tr>
<td>hema ≥ 30, ph ≥ 7.4</td>
<td>0.042(0.074)</td>
<td>0.025(0.077)</td>
<td>-0.152(0.092)</td>
<td>-0.051 (0.147)</td>
</tr>
<tr>
<td></td>
<td>0.574</td>
<td>0.740</td>
<td>0.104</td>
<td>0.730</td>
</tr>
</tbody>
</table>
Matching

- Causal inference is a multi-stage modelling process:
  1. Model the propensity score $P(T|X)$
  2. Match on the propensity score
  3. Check for balance between treatment groups in matched sample
  4. Estimate average causal effect
- Our focus is on step 3).
- One state of the art balance approach: genetic algorithms (Sekhon), very computationally expensive.
Covariate Balance

- Fundamentally a comparison of $X|T = 0$ and $X|T = 1$
- Theory: equal percent bias reduction of Rubin and collaborators
- In practice: two-sample $t$-tests typically are done before and after matching
- Recent innovation: CBPS of Imai and Ratkovic (2014) for propensity scores that achieve balance.
Our proposal: probability metrics

- Use

\[ \gamma(P, Q) = \sup_{f \in \mathcal{F}} \left| \int f dP - \int f dQ \right|, \quad (1) \]

where \( \mathcal{F} \) is a class of functions, to assess balance between \( P \) (probability law for \( X|T = 0 \)) and \( Q \) (probability law for \( X|T = 1 \)).

- If \( \gamma(P, Q) = 0 \), then \( P \) and \( Q \) induce equivalent probability spaces.

- If \( \mathcal{F} \) corresponds to an RKHS (next slide), then \( \gamma(P, Q) \) will have a simple empirical estimator that is evaluable in closed form.

- We term (1) the kernel distance and use as our statistic for evaluating balance.
Mathematics: theory of RKHS

- Let $T$ be a general index set
- RKHS: Hilbert space of real-valued functions $h$ on $T$ with the property that for each $t \in T$, there exists an $M = M_t$ such that
  \[ |h(t)| \leq M \|h\|_H \]
- 1-1 correspondence between positive definite functions $K$ defined on $T \times T$ with RKHS of real-valued functions on $T$ with $K$ as its reproducing kernel ($H_K$)
RKHS (cont’d.)

- If $f(x) = \beta_0 + h(x)$, then the estimate of $h$ is obtained by minimizing
  \[
g(y, f(x)) + \lambda \|h\|^2_{H_K},
\]
  where $g(\cdot)$ is a loss function, and $\lambda > 0$ is the smoothing parameter.
- RKHS theory guarantees minimizer has form
  \[
f_\lambda(x) = \beta_0 + \sum_{i=1}^{n} \beta_i K(x, x_i).
\]
  and
  \[
  \|h\|^2_{H_K} \equiv \sum_{i,j=1}^{n} \beta_i \beta_j K(x_i, x_j).
  \]
RKHS – remarks

- Subject of this 2014 JSM Fisher Lecture by Grace Wahba
- A lot of complicated theory
- However, what is captured by $K$ are similarities between pairs of individuals
- By focusing on individuals, we change the dimension of the problem from $p$ to $n$.
- 1-1 correspondence between $K$ in RKHS and covariance functions for Gaussian Processes
Some simulation studies

- Simulate nine covariates, mix of continuous and binary.
- Treatment variable $T$ is generated from Bernoulli($e(Z)$) where
  \[
  \logit(e(Z)) = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_4 + \alpha_4 Z_5 + \alpha_5 Z_7 + \alpha_6 Z_8 + \alpha_7 Z_2 Z_4 \\
  + \alpha_8 Z_2 Z_7 + \alpha_9 Z_7 Z_8 + \alpha_{10} Z_4 Z_5 + \alpha_{11} Z_1 Z_1 + \alpha_{12} Z_7 Z_7,
  \]
  and
  \[
  \alpha = (0, \log(2), \log(1.4), \log(2), \log(1.4), \log(2), \log(1.4), \log(1.2), \log(1.4), \log(1.6), \log(1.2), \log(1.4), \log(1.6)).
  \]
- The outcome variable $Y$ is generated from four different scenarios (A,B,C,D) which differ in terms of model complexity (taken from Stuart et al. (2013))
- True causal effect is a constant and equals 3.
Some simulation studies (cont’d.)

- Fit the same types of propensity score models as in Stuart et al. (2013), many of which will be misspecified.
- 1-1 matching, focus on the average causal effect of the treated (ACET)
- Misspecified propensity score → imbalance in covariates → biased causal effects
- Goal of balance statistics is to detect this imbalance
- Metric of evaluation: correlation between balance statistic and bias
- Choice of kernel: Gaussian kernel with $\sigma^2$ estimated by mean of all pairwise distances between observations
- Comparisons with average standardized mean difference (ASMD)-based balance statistics and Kolmogorov-Smirnov (KS).
### Results (1000 simulations)

#### Table: Mean and Standard Deviation of the Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Outcome A</th>
<th>Outcome B</th>
<th>Outcome C</th>
<th>Outcome D</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean ASMD</td>
<td>0.632 (0.134)</td>
<td>0.609 (0.155)</td>
<td>0.606 (0.152)</td>
<td>0.587 (0.156)</td>
</tr>
<tr>
<td>max ASMD</td>
<td>0.557 (0.176)</td>
<td>0.542 (0.196)</td>
<td>0.548 (0.185)</td>
<td>0.512 (0.193)</td>
</tr>
<tr>
<td>median ASMD</td>
<td>0.367 (0.214)</td>
<td>0.351 (0.212)</td>
<td>0.347 (0.221)</td>
<td>0.356 (0.213)</td>
</tr>
<tr>
<td>mean KS</td>
<td>0.372 (0.264)</td>
<td>0.358 (0.267)</td>
<td>0.348 (0.276)</td>
<td>0.333 (0.273)</td>
</tr>
<tr>
<td>mean t-statistic</td>
<td>0.634 (0.133)</td>
<td>0.609 (0.155)</td>
<td>0.610 (0.149)</td>
<td>0.588 (0.155)</td>
</tr>
<tr>
<td>kernel distance</td>
<td>0.797 (0.115)</td>
<td>0.773 (0.140)</td>
<td>0.788 (0.120)</td>
<td>0.759 (0.125)</td>
</tr>
</tbody>
</table>
Some simulation studies: Compare to genetic matching

- Genetic matching: iteratively updates the weight for each covariate while performing multivariate matching by minimizing/maximizing a certain balance metric.

- Our approach: compute kernel distance for a set of candidate propensity score models, create matchings based on the propensity scores, compute kernel distance for all matched datasets and choose the one with the smallest distance value.
Table: Comparison with Genetic Matching

<table>
<thead>
<tr>
<th>Scenario</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qqmean.mean</td>
<td>1.368</td>
<td>1.556</td>
<td>1.947</td>
<td>2.492</td>
</tr>
<tr>
<td>qqmedian.median</td>
<td>2.358</td>
<td>2.676</td>
<td>2.876</td>
<td>3.575</td>
</tr>
<tr>
<td>qqmax.max</td>
<td>1.332</td>
<td>1.508</td>
<td>1.753</td>
<td>2.320</td>
</tr>
<tr>
<td>pvals</td>
<td>1.516</td>
<td>1.677</td>
<td>2.112</td>
<td>2.858</td>
</tr>
<tr>
<td>kernel distance</td>
<td>0.267</td>
<td>0.334</td>
<td>0.930</td>
<td>1.065</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qqmean.mean</td>
<td>0.414</td>
<td>0.587</td>
<td>0.324</td>
<td>0.497</td>
</tr>
<tr>
<td>qqmedian.median</td>
<td>0.719</td>
<td>0.866</td>
<td>0.606</td>
<td>0.831</td>
</tr>
<tr>
<td>qqmax.max</td>
<td>0.402</td>
<td>0.552</td>
<td>0.340</td>
<td>0.484</td>
</tr>
<tr>
<td>pvals</td>
<td>0.504</td>
<td>0.623</td>
<td>0.443</td>
<td>0.698</td>
</tr>
<tr>
<td>kernel distance</td>
<td>0.764</td>
<td>1.073</td>
<td>0.603</td>
<td>0.918</td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qqmean.mean</td>
<td>2.042</td>
<td>2.765</td>
<td>3.895</td>
<td>6.455</td>
</tr>
<tr>
<td>qqmedian.median</td>
<td>6.074</td>
<td>7.908</td>
<td>8.636</td>
<td>13.468</td>
</tr>
<tr>
<td>qqmax.max</td>
<td>1.936</td>
<td>2.579</td>
<td>3.188</td>
<td>5.616</td>
</tr>
<tr>
<td>pvals</td>
<td>2.550</td>
<td>3.201</td>
<td>4.659</td>
<td>8.656</td>
</tr>
<tr>
<td>kernel distance</td>
<td>0.626</td>
<td>1.261</td>
<td>1.229</td>
<td>1.975</td>
</tr>
</tbody>
</table>
Matching using Reproducing Kernel Hilbert Spaces

Intuition: a new framework for balance

- Covariate Balance Propensity Score (CBPS) proposal: estimate logistic regression so that

\[ E(X_i|T = 1) = E(X_i|T = 0) \]

for \( i \)th covariate, \( i = 1, \ldots, p \)

- Our approach: fit functions \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) and guarantee that

\[ f(X|T = 0) = f(X|T = 1) \]

for as rich a function class as possible.

- Gaussian kernel is a natural one to use because its corresponding function space is dense in \( L_2 \).
Conditional Independence Assumptions and Causal Inference

- Recall strong Ignorability of Treatment Assignment Assumption (SITA):
  \[ T \perp \{ Y(0), Y(1) \} | X \]
  
  where \( X \) are covariates

- Also called no unmeasured confounders (Robins)

- This allows for causal inference

- This is also a conditional independence assumption
For the average causal effect, we can relax SITA to

\[ T \perp Y(0)|X \]

and

\[ T \perp Y(1)|X \]

We can link this up to dimension reduction methodology (subject of 2005 Fisher lecture by D. R. Cook)

Idea behind dimension reduction: identify bases of subspaces that capture the essential information about regression relationships

Two objects in dimension reduction: central subspace and central mean subspace (CMS). Here, we will only need CMS.
Central mean subspace (CMS)

- Definition: subspace $S_{E(W|X)}$ spanned by $\beta$ ($p \times d$ matrix) such that
  
  $$E(W|X) \perp X|X'\beta$$

- Note that this trivially holds if $d = p$ and $\beta$ is the identity matrix

- Typically, CMS exists under relatively mild assumptions

- One of our key results: We have

  $$S_{E(Y(1) - Y(0)|X)} \subset S\{E(Y(1)|X), E(Y(0)|X)\}$$

  and that under weak ignorability and a common support condition for $X'\beta$, we have that

  $$S_{E(Y(1)|X)} = S_{E(Y|X, T=1)}$$

  and

  $$S_{E(Y(0)|X)} = S_{E(Y|X, T=0)}$$
Algorithm

- This results suggests the following algorithm:
  1. Estimate CMS for $T = 1$ and $T = 0$
  2. Regress the outcome on the reduced covariates within each treatment group separately
  3. Estimate the regression causal effect by taking the difference of the regression function

- Relaxes covariate overlap/treatment positivity assumption needed for causal inference

- We use MAVE method for CMS estimation.

- Semiparametric version of G-computation algorithm of Robins (1986)

- Empirical influence function theory can be used to prove the consistency and asymptotic normality of the estimated average causal effect
One surprising finding

Under the sufficient dimension reduction assumption, if \( \text{var}(Y_t|X) \) is measurable with respect to \( X'\beta_t \) for \( t = 0, 1 \), then the asymptotic variance of the ACE estimator is less than or equal to the semiparametric efficiency bound for regular estimators, and the equality occurs if and only if the following conditions are satisfied:

(a) \( \pi(X) \in (\tau, 1 - \tau) \) on \( \Omega(X) \) for some \( \tau \in (0, 0.5) \).

(b) \( \pi(X) \) is measurable with respect to \( X'\beta_t \) for both \( t = 0, 1 \).
Simulations bear it out

Table 2. Compare estimators of average causal effect

<table>
<thead>
<tr>
<th>Models</th>
<th>GCL</th>
<th>GCQ</th>
<th>GAM</th>
<th>RF</th>
<th>IPW</th>
<th>Matching</th>
<th>DR</th>
<th>DR-oracle</th>
<th>TMLE</th>
<th>a-SORCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>.096</td>
<td>-.139</td>
<td>-.035</td>
<td>-.140</td>
<td>-.083</td>
<td>-.080</td>
<td>-.132</td>
<td>-.004</td>
<td>-.035</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>(.105)</td>
<td>(.158)</td>
<td>(.086)</td>
<td>(.069)</td>
<td>(.218)</td>
<td>(.085)</td>
<td>(.278)</td>
<td>(.089)</td>
<td>(.208)</td>
<td>(.049)</td>
</tr>
<tr>
<td>II</td>
<td>-.007</td>
<td>.008</td>
<td>.000</td>
<td>.009</td>
<td>.004</td>
<td>-.043</td>
<td>.000</td>
<td>-.007</td>
<td>-.016</td>
<td>-.008</td>
</tr>
<tr>
<td></td>
<td>(.113)</td>
<td>(.120)</td>
<td>(.126)</td>
<td>(.099)</td>
<td>(.163)</td>
<td>(.093)</td>
<td>(.129)</td>
<td>(.090)</td>
<td>(.139)</td>
<td>(.068)</td>
</tr>
<tr>
<td>II</td>
<td>-.005</td>
<td>.008</td>
<td>.005</td>
<td>.020</td>
<td>.043</td>
<td>-.016</td>
<td>-.013</td>
<td>-.008</td>
<td>-.020</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>(.123)</td>
<td>(.073)</td>
<td>(.094)</td>
<td>(.096)</td>
<td>(.181)</td>
<td>(.132)</td>
<td>(.152)</td>
<td>(.094)</td>
<td>(.169)</td>
<td>(.064)</td>
</tr>
<tr>
<td>IV</td>
<td>-.088</td>
<td>.010</td>
<td>-.007</td>
<td>-.019</td>
<td>-.004</td>
<td>-.090</td>
<td>.000</td>
<td>.001</td>
<td>-.052</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>(.346)</td>
<td>(.068)</td>
<td>(.073)</td>
<td>(.150)</td>
<td>(.318)</td>
<td>(.248)</td>
<td>(.070)</td>
<td>(.068)</td>
<td>(.217)</td>
<td>(.072)</td>
</tr>
<tr>
<td>V</td>
<td>-.001</td>
<td>.000</td>
<td>.001</td>
<td>.003</td>
<td>.021</td>
<td>-.001</td>
<td>.000</td>
<td>-.001</td>
<td>-.001</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>(.022)</td>
<td>(.022)</td>
<td>(.025)</td>
<td>(.057)</td>
<td>(.088)</td>
<td>(.081)</td>
<td>(.022)</td>
<td>(.023)</td>
<td>(.023)</td>
<td>(.023)</td>
</tr>
</tbody>
</table>
Why?

- Under assumptions of theorem, dimension reduction functions as a \textit{sparsity}-inducing assumption
- Thus, we can beat doubly-robust estimators for certain classes of models
- Tradeoff: our estimator can behave erratically under other model configurations
- Current work: see if we can alleviate efficiency issues using alternative two-stage approaches
"Big data is not about the data, it’s about the algorithms" – Gary King (Harvard)

Lots of interest in machine learning algorithms for big data, but ultimately, analytics will also have a substantial causal component as well.
