Data Science and the Policy Completion Problem

Sanjay Chawla
University of Sydney
Sydney, NSW, Australia
sanjay.chawla@sydney.edu.au

Federico Girosi
University of Western Sydney
Sydney, NSW, Australia
f.girosi@uws.edu.au

Fei Wang
University of Sydney
Sydney, NSW, Australia
brucefeiwang@gmail.com

ABSTRACT

The link between policy analysis and data science is more delicate than it may appear. A new policy, by definition, will change the underlying data generating model, rendering classification or supervised learning inapplicable. Perhaps eliciting causal relations from observational data is the correct framework for estimating policy impact. However, there are substantial gaps between the theory, practice and feasibility of causal models. In this paper we argue that transduction, a form of inference where we reason from specific training instances to specific test instances, may provide an appropriate framework for evidence-based policy analysis. In particular, we will demonstrate that the matrix completion problem, introduced in the data science community for making predictions in recommendation systems, can be a powerful tool for both predicting and evaluating the impact of new policy changes.

1. INTRODUCTION

Health care costs in the United States account for nearly eighteen percent of the US GDP valued at approximately eighteen trillion dollars at the end of 2013. Thus US health expenditure is bigger than the entire GDP of France [11]. Yet when the US Congress passed the Affordable Care Act (ACA) in 2010, its likely impact on the economy, and in particular on unemployment rates, was unknown. Even now, as the law is in effect and is being implemented across the country, it is still hard to provide an answer to the seemingly simple question: "Did the ACA lead to net job loss ?".

There are least three important challenges in both the simulation of future policies and the evaluation of current ones:

Identification of causal structure: A well known challenge in the use of data to infer causation is that when we observe a relationship $A \rightarrow B$ it is not clear whether there exists an unknown hidden variable H such that $H \rightarrow A$ and $H \rightarrow B$. Another common challenge is the presence of selection effects, that lead to acquire data that are not randomly sampled and that may also confound the causal relationship. In specific setting we can use randomized trials to control for these threats to valid causal inference, but in most cases setting up a randomized trial may not be feasible or practical.

KDD'14 Data Science for Social Good, August 24–27, 2014, New York, NY, USA.

Statistical Inference: Ultimately one needs to estimate certain quantities based on a finite set of data. In this regard not all statistical inference frameworks are born equal and some may be more advantageous than others.

Missing data: little can be done when a variable of interest is simply never observed. However it is often the case that some variables are observed on a subset of the population, possibly not at random. Whether and how this issue is tied into the statistical inference framework may have profound implications for the analysis.

The rest of the paper is as follows. In Section 2 we discuss the roles of causality calculus and of supervised learning in addressing the challenges listed above. In Section 3 we briefly introduce the transduction principle as an alternative form of inference that might be more advantageous in the study of policy impacts. In Section 4 we focus on the matrix completion problem (MCP) as a framework to perform transductive inference and to deal with missing data at the same time. A case study based on data from a global environment performance index is used to show the efficacy of the MCP in section 5. In Section 6 we discuss ways of using the transduction principle to evaluate the gap between policy intent and implementation. We conclude in Section 7 with a summary.

2. DATA GENERATING MECHANISM

From a data science perspective, a phenomenon or law can be seen as an unknown data generating mechanism (M). We observe the projection of M in the form of data and we query ($Q(M)$) about the mechanism using data and our hypothesis of the model. When a policy change is introduced, the model changes to M' with a corresponding change in the data projection. However discerning either the original intent of the policy change or the unintended consequences and feedback from the policy may take a long time to discern.

2.1 Learning and Generalization

While computer science in general deals with the issue of space and time complexity of algorithms, the primary focus of machine learning has been on *sample* complexity. For example how many sample instances (x_i, y_i) of an underlying data generating mechanism which relates X to Y, are required to learn a function $f : X \rightarrow Y$ up to a required degree of confidence under the assumption that the *data generating mechanism remains invariant* ? Having learnt the function f, we have generalized from the data and can now apply f

on previously unknown x^{new} and predict the answer y^{new} with a high degree of confidence.

Valiant, who introduced the concept of Probably Approximately Correct (PAC) learning, postulates two assumptions under which learning is (practically) possible [10]:

1. `Invariance:` The generalization (f) cannot be applied in a situation which is fundamentally different from which it was learnt.

2. `Learnable Regularity:` In order to have a constructive learning algorithm (i.e., beyond just existence), it must be learnable from a sample size which is a polynomial function of the accuracy desired. For example, if we want to learn f, up to accuracy ϵ then the number of samples required to learn f should be bounded by $poly(\frac{1}{\epsilon})$.

Experience has shown that insisting on a polynomial bound on constructive learning is a practical requirement. However, assuming invariance is too strong an assumption in policy settings. While laws of physics are immutable, societies and the human condition and context evolves. For example, consider a simple query: "What will be the average productivity of a person working at minimum-wage who has access to health insurance coverage which is not tied to an employer ?" Such a counterfactual query cannot be answered under the Invariance assumption, pointing at the fact that some thought needs to be given before blindly applying supervised learning techniques in the policy setting.

2.2 Causality Calculus

It is not clear whether causal or counterfactual queries can be answered using the existing suite of supervised learning algorithms. Stripped of all complexity, learning algorithms which attempt to learn a function $f : X \rightarrow Y$ are primarily trying to exploit the correlation between X and Y. However, the use of correlation is undermined by the well known Simpson's Paradox where it can be shown that two random variables X and Y can be positively correlated but a third random variable Z can be constructed such that both (X, Z) and Y and $(X, \neg Z)$ and Y are negatively correlated. Pearl has introduced a causality calculus to overcome Simpson's paradox and answer causal and counterfactual queries from observational data [7, 8]. Whether causality calculus

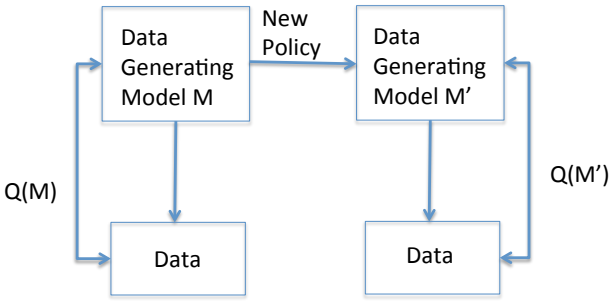


Figure 1: Data generating mechanism is transformed from an unknown M to another unknown M' under a new policy change. Queries on both M and M' using observational data can lead to confounding errors.

is at a stage where it can be applied in a practical world of incremental policy changes and missing and messy data is still a subject of vigorous debate. Furthermore even at the conceptual level there are some basic issues with causal modeling that need to be resolved. A famous example of an identifiability problem with causal modelling is shown in Figure 2. If the causal query is "does smoking cause lung cancer ?" then there is a possibility that there is a hidden variable (perhaps a genotype) that causes a person to smoke and also causes cancer. However, because of the topology of the graph, the left model is not identifiable, i.e., cannot be used to make causal inference, while the right model where the relationship between smoking and cancer is mediated through a third variable is identifiable [6].

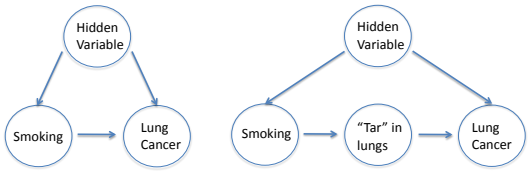


Figure 2: The left model is not identifiable, while the right model is. While there are tools to detect identifiability it is still not completely clear why it should be like that.

3. TRANSDUCTION

In the standard supervised and unsupervised learning framework one typically learns a "rule" (such as a function or a probability distribution) from a set of sparse data points and then performs prediction by applying the learnt rule on new data points. If all one needs is the evaluations of the rule at specific points, however, this paradigm violates an important principle attributed to V. Vapnik: *when trying to solve some problem, one should not solve a more difficult problem as an intermediate step.* In this case the intermediate step is learning a rule, that is a complex object possibly belonging to an infinite dimensional space. An alternative to this paradigm is *transduction* [2], a form of inference where one goes from the particular (values of the function on the data) to the particular (values of the function on specified new data points), as opposed to inductive inference where one goes from the particular to the general (the "rule"). Transduction seems particularly useful in a framework where underlying relationships may be shifted by a policy and one does not want to relying on a specific assumptions on the class of rules one considers.

As an example of transduction suppose we want to formulate a policy for agriculture with the objective of improving the overall state of the ecosystem. The policy is articulated in terms of agriculture subsidies and use of pesticide. In a transductive framework we can learn from the experience of some countries and transduce (generalize) to a specific list of other countries. This is appealing since it may not be feasible (or required) to learn a general function that can be applied to an arbitrary country.

4. MATRIX COMPLETION: A POWERFUL TRANSDUCTIVE ALGORITHM

We now describe the matrix completion problem (MCP) which has emerged as a general solution paradigm for many

practical problems across data mining, machine learning, statistics and information theory [1, 3, 9]. The MCP is an instantiation of the transductive principle[4] and can be used as a design pattern to address several common problems in data science including:

1. Estimate missing data. For example in policy work, researchers often work with survey data where the resulting tables tend to have a significant number of missing values. The estimation of missing values can be framed using MCP.

2. Make predictions. The same principle that is used for estimating missing values can be used for transductive prediction. For example, recommender systems, use data collected from user activity to recommend products. Estimating a general function which maps an arbitrary user to a ranked vector of interests is too complicated and perhaps not necessary. Recommendations are required for specific users based on their partially observed activities.

Before, we explain the technical content of MCP we situate the problem in a context which will underscore its far reaching impact.

Within the computer science community, several subcommunities have emerged which are focused on the computational analysis of specific data modalities. However, despite seemingly different modalities (text and image) there is an underlying commonality which can be exploited and which we formalize as follows:

Axiom of Data Science: Observational data, independent of modality, tends to be highly redundant.

When data is represented in a matrix form, then high redundancy in data is equivalent to the matrix having a low rank. The fact that most observational data, when represented in matrix form has low rank explains the widespread use of matrix factorization techniques in wide ranging applications including image and text retrieval and recommendation systems.

The MCP was inspired by a competition set up by Netflix, a US-based company that allows users to stream movies on their digital devices. Users can rate the movies they watch and Netflix uses these ratings to predict which movies a user may like, providing an effective recommendation system. The Netflix competition was an open challenge to come up with a recommendation algorithm which could beat the in-house algorithm.

The MCP problem can be stated as follows. Let $M \in R^{a \times b}$ be a partially filled matrix. Let Ω be the set of entries in the matrix which are observed, i.e., $(i,j) \in \Omega$ if M_{ij} is observed. Find X which satisfies the following optimization problem.

$$\begin{array}{lll} \text{minimize} & rank(X) & \\ \text{subject to} & X_{ij} = M_{ij} & (i,j) \in \Omega \\ & X \in R^{a \times b} & \end{array}$$

Notice how the MCP definition directly incorporates the low rank assumption of observational data. Thus MCP provides a general framework to reason across data modalities as long as data can be represented in a matrix form. While we are not aware of the use of MCP in the social science or economics literature, applications of MCP for climate change and in phylogeny reconstruction have been reported in the literature [3, 9].

5. ENVIRONMENTAL CASE STUDY

We demonstrate the matrix completion framework on a data set that is used to calculate the global Environmental Performance Index (EPI). The EPI rates two hundred and thirty two countries on twenty two environment, ecosystem and health indicators covering ten policy categories [12]. The EPI was designed to be an objective data-driven approach to measure global performance with respect to environmental policy goals and to complement the traditional Gross Domestic Product (GDP) measure.

5.1 Imputing Missing Values

Despite being a comprehensive data set where data is meticulously collected and curated, the EPI has to deal with the issue of missing data. To estimate missing values, the EPI uses the following two part approach: (i) if the data is missing in the interior of a time series, then it is linearly interpolated based on the neighboring data points, (ii) data set at the two ends of the time series is imputed based on the values of the nearest year.

However, the EPI data is highly redundant. As show in Figure 3(b), the first four singular values of the matrix can explain over 90% of the variance. We have carried out a very preliminary analysis on imputing the missing values using MCP and compared it to the simple strategy of estimating the missing value based on the average of the matrix column. The relative error results are shown in Figure 3(a), where the y-axis is on a log-scale. It is clear that the MCP approach leads to substantially better results in twenty one of the twenty two indicators[1]

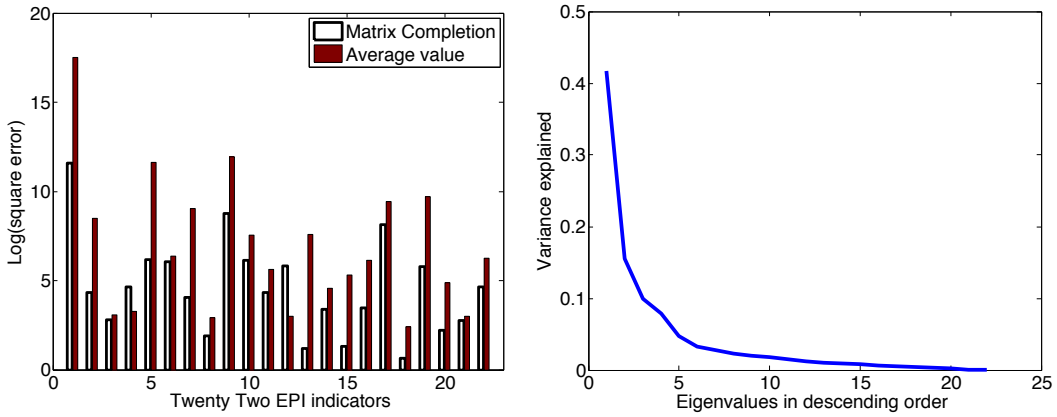
6. POLICY GAP ANALYSIS AND MCP

Matrix completion is a transductive framework with applications beyond missing value prediction. An invariant in policy making is that there are time and information gaps between a policy and its implementation, and policies are not always implemented in a manner which is consistent with the original intent. This is particularly relevant in the case of policies with no enforcement mechanism, such as the issue of guidelines in health care, which may result in uneven implementation across implementation units.

To measure the gap between policy intent and implementation the following approach based on MCP could be used. Suppose we want to measure the impact of a new policy at regional levels: (i) Select two regions r_1 and r_2 and use administrative data, such as electronic health records (EHRs) or insurance claims, to build a database of sample residents in the two regions, possibly supplemented by socio-economics status (SES) information obtained linking the EHR to some survey data[2] (see Figure 1 for an example); (ii) Construct snapshots of the database before the new policy came into effect and another one after a time lag t. Let $D_i(0)$ and $D_i(t)$ be the snapshots for $i = 1, 2$. (iii) Use a

[1]The error is estimated by removing some known entries and comparing them with those estimated by the model.

[2]For example, in an Australian setting one could use the survey data from the 45 and Up study linked to Medicare claims and hospital data.



(a) Missing value imputation with MCP

(b) The EPI numerical rank after imputation

Figure 3: Preliminary results on the application of MCP on the EPI data set. (a) MCP leads to better imputation accuracy and (b) exploits the low rank of the data set.

	diagnosis			medication			lab		demographics		SES	
	d_1	d_2	d_3	m_1	m_2	m_3	l_1	l_2	dm_1	dm_2	ses_1	ses_2
patient 1	x		x	x	x		x			x	x	x
patient 2		x		x		x		x	x	x		x
patient 3		x		x		x			x	x	x	x
patient 4	x		x			x		x		x		
patient 5	x	x	x			x		x	x	x		
patient 6	x		x		x		x	x	x	x		

Table 1: Potential use of matrix completion for predicting the policy impact using using electronic health records [5]. Blank entries denote missing values. In this example the first three records have been linked to survey data and therefore SES information is available for (almost) all of them, while completely missing for the remaining records.

matrix completion algorithm to complete the database snapshots. (iv) Now query the completed databases and look for differences. For example, are more residents visiting their local doctors for preventative check ups? Has there been a substantial change in the rank of the completed matrices? For example if the relative rank of $D_2(t)$ is higher than that of $D_1(t)$, then it is a preliminary indicator that more change is occurring in region r_2 than region r_1.

7. CONCLUSION

The rise of data science as a discipline is an opportunity to provide a firmer footing for evidence-based policy analysis. However, since a new policy will *cause* a shift in the underlying data generating mechanism, classical machine learning tools like supervised learning may not be the most appropriate tools for evaluating the impact of new policy. We have argued that transduction, a form of inference where we learn from specific training instances to specific test instances, provides a promising alternative framework for policy analysis. In particular the matrix completion formulation is a practical method for measuring the impact of policy at a fine-grained level, that combines in one step the task of prediction and imputation of missing values. Very preliminary analysis on a real data set for measuring the impact of policy on global environment assessment suggests that the matrix completion approach is potentially promising and deserves further evaluation.

8. REFERENCES

[1] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[2] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[3] M. Ghafarianzadeh and C. Monteleoni. Climate prediction via matrix completion. In *AAAI (Late-Breaking Developments)*, 2013.

[4] A. B. Goldberg, X. Zhu, B. Recht, J.-M. Xu, and R. D. Nowak. Transduction with matrix completion: Three birds with one stone. In *NIPS*, pages 757–765, 2010.

[5] P. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records:towards better research applications and clinical care. *Nature Genetics*, 13:395–405, June 2012.

[6] M. Nielsen. If correlation doesn't imply causality then what does? http://www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/, 2012. [Online; accessed 12-July-2014].

[7] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[8] J. Pearl. The mathematics of causal inference. In *KDD*, page 5, 2011.

[9] H. Shan, J. Kattge, P. Reich, A. Banerjee, F. Schrodt,

and M. Reichstein. Gap filling in the plant kingdom - trait prediction using hierarchical probabilistic matrix factorization. In *ICML*, 2012.

[10] L. Valiant. *Probably Approximately Correct*. Basic Books, 2013.

[11] Wikipedia. List of countries by gdp (nominal)? http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal), 2014. [Online; accessed 12-July-2014].

[12] YCEP and CIESIN. 2012 environmental performance index and pilot trend environmental performance index. http://dx.doi.org/10.7927/H48913SG, 2012. [Online; accessed 12-July-2014].