Variability Reduction with Optimal Transport

Kai M. Hung[1], Andrew Lipnick[2], Ryan Shìjié Dù[2], Nina Mortensen[2], Esteban G. Tabak[2]

[1]Rice University
[2]Courant Institute, New York University

1 Introduction

As our society generates increasingly complex datasets and utilize them for high-stake decision making systems, we are in need of a procedure to filter out the effects of unwanted features, or factors, from our data distributions. A few examples of said unwanted features may include features that correlate with sensitivity attributes resulting in discriminate predictions, batch effects that induce variability within clinical experiment data that does not inform the differential effectiveness between treatment policies, and uncontrollable factors that does not inform policy decision-making.

While there are many attempts to address these undesirable effects through regularizers on machine learning objective that discourages predictions correlating with said unwanted variables [5] or induce sparsity to encourage simplicity in features selected [8], there are currently, to the best of our knowledge, no method that combat the aforementioned challenges by directly mitigating the effect of said variables from the data distribution itself. In this paper, we present a method to reduce the effect of unwanted variables from a data distribution via optimal transport (OT). Formally, we define the effect of a variable as the variability induced by said variable. It follows that our solution is an optimal transport-driven approach to reduce the variability of the data distribution with respect to select variables. In fact, the problem may be reduced to solving for an optimal transport barycenter of the conditional distributions of the observed data given the select factor(s).

Building upon this OT formulation of the variability reduction problem, we also explore a scenario in which not all values of the unwanted variable is known. The relevance of this problem formulation is significant, particularly in some of our highest stake decisions. In the clinical setting, oftentimes data is incomplete due to fragmentation in data collection such as difference in equipment capability and documentation requirements across hospitals. On the other hand, despite our best attempt at fair decision-making systems, sensitivity attributes may not always be present due to individual incentives to withheld said information (such as voluntary identifications for job applications), yet they remain crucial for training and validation of a model that adhere to some notion of fairness across observation groups.

Our Contributions In this paper, we propose a flexible method for the removal of variability induced by select features through solving an optimal transport barycenter problem. Furthermore, we provide an extension to the framework under the problem setting in which the exact value for the select feature is not known across the distribution, a semi-supervised regime for variability reduction. Lastly, we conclude with discussing the extensions of this method for factor discovery and data augmentation.

2 Outline

We begin with formulating the problem, which includes a discussion on notation and an introduction of the necessary background in optimal transport. For the bulk of the paper, we discuss approaches for solving the barycenter problem and contrast the various methods of optimization. Next, we provide numerical results of our data on simulated and real-world datasets. Lastly, we provide a discussion of possible extensions of this method.

3 Problem Formulation

Let x be the observed data and z be the factor of interest such that we wish the filter out the effect of z from x. We propose that one can address this variability reduction problem by finding a new representation of x, call it y, that minimizes the deformation from x yet is independent from z. Formally, we would like to *optimally transport* the distribution of x to y subject to the condition that y is independent of z.

A Refresher on Optimal Transport The optimal transport (OT) problem dates back to the French mathematician Gaspard Monge in 1781 where his objective was to find an optimal plan for transporting mass from a pile of sand into a pit [9]. Later on, the problem is generalized to the transportation of the probability distributions and led to a proposed metric on the space of probability measures known as the Wasserstein distance. Today, the optimal transport problem is garnering attention from researchers in the mathematics, statistics, and machine learning community due to its theoretical elegance and practical application to data-driven problems which led to efficient solvers or reformulations of the optimal transport problem under different problem settings and successes of its integration to application areas such as generative modeling [1], domain adaptation [3, 7], computer graphics [6], single-cell sequencing [2], and etc.

OT Variability Reduction Let us denote the conditional probability distribution of x given z by $\rho(x|z)$ and the probability distribution of z as $\gamma(z)$. We can quantify the data deformation between x and the transformation y using the 1-Wasserstein distance

$$\min_{y=T(x,z)} \int c(x,y)\rho(x|z)\gamma(z)dxdy \ \text{ s.t. } \ y \perp\!\!\!\perp z \quad (1)$$

where T denotes a transport map from x to y and $c(\cdot,\cdot)$ is a ground cost function between x and y. Under this formulation, the distribution of y, $\mu(y)$, is the OT barycenter of the set of $\rho(x|z)$ for all values of z.

Data-Based OT Variability Reduction In the real world, we often do not have access to the distributions $\rho(x|z)$ and $\gamma(z)$. Instead, we have a sample $\{x_i, z_i\}_i$ of the observation and their associated factors. Under this setting, we can rewrite the variability reduction problem as follows (albeit with a slight abuse of notation)

$$\min_{y_i=T(x_i,z_i)} \sum_i c(x_i, y_i) \ \text{ s.t. } \ y_i \perp\!\!\!\perp z_i. \quad (2)$$

Now that we have formulated the problem in terms of data samples, let us discuss how to solve the above constrained optimization problem.

4 Methods

Satisfying Independence Perhaps a question regarding the proposed formulation is: how do we ensure that y is independent of z? To do this, we propose the use of some *test function*, $F(y,z)$, that evaluates to 0 when our independence condition is met and strictly positive otherwise. Incorporating this test function with the Lagrangian multiplier method transforms our constrained optimization into the following unconstrained optimization problem

$$\min_{y_i=T(x_i,z_i)} \max_{\lambda} \sum_i \{c(x_i, y_i) + \lambda F(y_i, z_i)\} \quad (3)$$

where λ is the Lagrangian multiplier.

In terms of the choice for the test function F, one can adapt many proxies of independence. For example, one can utilize correlation to test for linear independence or for the conditional expectation of y, $\bar{y}(z)$, to be independent of z. The choice of the test function dictates the form of relaxation from the general independence condition that our algorithm actually guarantees, which determines the type of variability reduction that we are able of performing.

Among many choices, we propose the use of *mutual information* between y and z as our test function. In particular, the mutual information is defined as

$$I(y,z) := D_{KL}(\pi(y,z)\|\mu(y)\gamma(z)) \quad (4)$$

where $\pi(y,z)$ is the joint probability between y and z. Essentially the mutual information is the

Kullback-Leibler divergence between the joint distribution and the product distribution of y and z.

Under our data-based formulation, we can rewrite the mutual information term as

$$\sum_i \log\left(\frac{\pi(y_i, z_i)}{\mu(y_i)\gamma(z_i)}\right) \tag{5}$$

where we will approximate the distributions of π, μ, and γ using kernel density estimation. In particular, let there be kernels K_α^y and K_β^z for y and z respectively with bandwidths α and β. It follows that our test function is

$$F(y,z) := \log\left(\frac{\sum_l K_\alpha^y(y, y_l) K_\beta^z(z, z_l)}{\sum_l K_\alpha^y(y, y_l) \sum_l K_\beta^z(z, z_l)}\right) \tag{6}$$

Flow-Based Optimization While there are many methods for solving the combined objective of Eq. (3) and Eq. (6), we propose a flow-based approach. Instead of solving for the transport map T as a part of the optimization procedure, we initialize y to be equivalent to x and allow them to move around freely, like a flow, in a gradient descent scheme. Additionally, instead of solving a min-max optimization procedure, we propose a slow increase in λ over the gradient descent iteration. The specific choice of our optimization procedure is based primarily on intuition rather than a theoretical justification, as there are many other comparable approaches out there for solving such unconstrained minimax optimization problems with a multitude of hyperparameters. However, we would like to highlight why our choices were made this way and add that this algorithmic choice shows promising performance in practice.

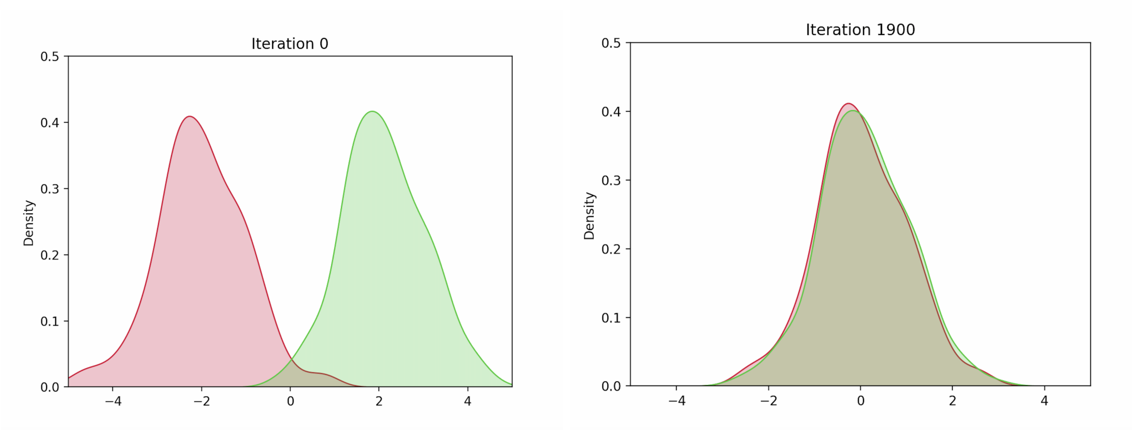
On the matter of initializing y as x as opposed to random initialization, this is because *minimizing data deformation* is one of the two central goals of our variability reduction problem (the other being ensure that the resulting representation of the data y is independent from the unwanted variable z). By initializing y as x, we ensure that the initialization does not stray too far from the original distribution of the data, which could otherwise have the adverse effect of inducing large enough deformation that lead to divergence or convergence at a sub-optimal point.

We make the choice of scaling λ as the iteration increases instead of solving for a min-max optimization problem for the following reasons. Firstly, min-max optimization solvers typically require second-order information for convergence, which is costly. Secondly, the nature of our optimization problem does not put y and z in complete adversarial positions. As long as y and z are independent, λ should not need to further increase and y can move freely to minimize the data deformation subject to the large value of λ that is sufficient for asserting independence. In other words, we have reasonable belief that λ need not to continually increase after reaching a sufficiently large value in practice since the worse offense of the independence condition occurs at initialization with $x = y$ and λ may grow large enough to remove the y from x as the algorithm carries on. Thirdly, a direct extension of the variability reduction problem, factor discovery, involves an additional maximization over the space of z which would effectively turn this problem into a max-min-max problem if we insist solving the current problem as a min-max optimization method. A max-min-max problem is even nastier to solve in practice, and for the sake of reducing technical challenges of the factor discovery problem, we adapt a scaling growth of λ in our algorithm instead of solving this as a min-max problem, i.e. we are solving for

$$\min_{y_i = T(x_i, z_i)} \sum_i \{c(x_i, y_i) + \lambda_t F(y_i, z_i)\} \tag{7}$$

such that $\lambda_t \to \infty$ as $t \to \infty$.

Semi-Supervised OT Variability Reduction Suppose that now the identify of some factors are unknown, denoted as $z^?$. Formally, we have samples $\{x_i, z_i\}_i$ such that there exists $z_i = z^?$ and let z_i take on a finite set of values, i.e. K discrete classes. Our proposal is to split each observation y_i into K versions of itself, $\{y_{i,k}, z_k\}_{k=1}^K$ and to assign each of them a probability p_{ik} for belonging the particular class.



(a) Initial KDE of y given z

(b) Converged KDE of y given z

Figure 1: The kernel density estimation (KDE) of the y_i given $z \in \{z_1, z_2\}$ in the Gaussian experiment over the optimization procedure. On the left, (a) shows that the y_i is initialized to reflect the x_i and hence the KDE approximates $\rho(x|z_1)$ and $\rho(x|z_2)$ respectively. On the right, (b) shows the approximations of the y that are now independent of z yet still preserve the shape of the original data distributions.

In this case, the variability reduction problem becomes

$$\min_{y_{i,k}=T(x_i,z_k)} \max_{\lambda} \sum_{i,k} p_{ik}\{c(x_i,y_i)+\lambda F(y_{i,k},z_k)\} \tag{8}$$

such that the probability p_{ik} is updated as follows

$$p_{ik}^{t+1} = \frac{p_{ik}^{t}\mu(y_{i,k})}{\sum_{k'} p_{ik'}^{t}\mu(y_{i,k'})}$$

where t represents the iteration.

Here, μ is once again approximated using the kernel density estimation with a slight modification

$$\mu(y_{i,k}) \approx \sum_{l}\sum_{k'} p_{lk'} K_{\alpha}^{y}(y_{i,k}, y_{l,k'}) \tag{9}$$

such that every center is weighted by its probability. Note that this modification applies to instances of kernel density estimation invoked for F as suggested in Eq. (6). Naturally, this formulation will place less emphasis on version of y_i that are associated with less probable values of z. For observations (y_i, z_i) such that the factor is known, we assign $p_{i,k}$ to be 1 if $k = i$ and 0 otherwise. For observations $(y_i, z^?)$ such that the factor is unknown, we can initialize $p_{i,k}$ to be $1/K$ with some noise and then normalize.

There exists a hidden challenge with this formulation. Without additional guardrails, it is possible for the probability update to occur too quickly and for an observation to be incorrectly associated with a class before the independence assertion takes effect, i.e. before λ becomes sufficiently large. To mitigate this, we recommend two additional procedures. First, we can perform a slow update to the probabilities p_{ik} governed by some learning rate η

$$\hat{p}_{ik}^{t+1} = \hat{p}_{ik}^{t} + \eta(p_{ik}^{t+1} - \hat{p}_{ik}^{t})$$

where we would use $\hat{p}$ instead of p in Eq. (8) and Eq. (9). Second, we can begin to update the probabilities only after λ becomes sufficiently large.

5 Results

To validate our method, we conducted some preliminary numerical experiments on toy examples. All the code is publicly available at github.com/KataTech/OTFactorDiscovery.

Gaussian For our first simulation, we adapt the simple case that $z_i \in \{z_1, z_2\}$, $\rho(x|z)$ are Gaussian distributions with the mean is -2 if $z_i = z_1$ and 2 otherwise along with a variance of 1, and $\gamma(z)$ is a discrete uniform distribution. In other words, we have data points x_i that comes from one of two Gaussian distributions depending on their z_i values. As for the choice of kernel, we use the isotropic Gaussian kernel with $\sigma = 1$. Figure (1) shows the initial distributions of y given their corresponding z and the final distribution of y after the variability induced by z has been eliminated. As we can see from the resulting distribution, it is no longer possible to discern which Gaussian distributions generated the original points and their overall distributions is relatively preserved, satisfying both the independence and the preservation of the data conditions.

Iris Moving beyond experimenting on simulated data, we also apply our method on the slightly more complex case of the popular iris dataset from the UCI Machine Learning repository. The iris dataset is a collection of 4-feature observations from different flowers, with a total of 3 classes. In our experiment, we attempt to remove the variability induced by the flower identify from the 4 features.

For ease of visualization, we project the data to its first two principle components computed using the pre-transformed, standard iris data. From Figure (2), we can see that our algorithm effectively removes the variability induced by the floral classes.

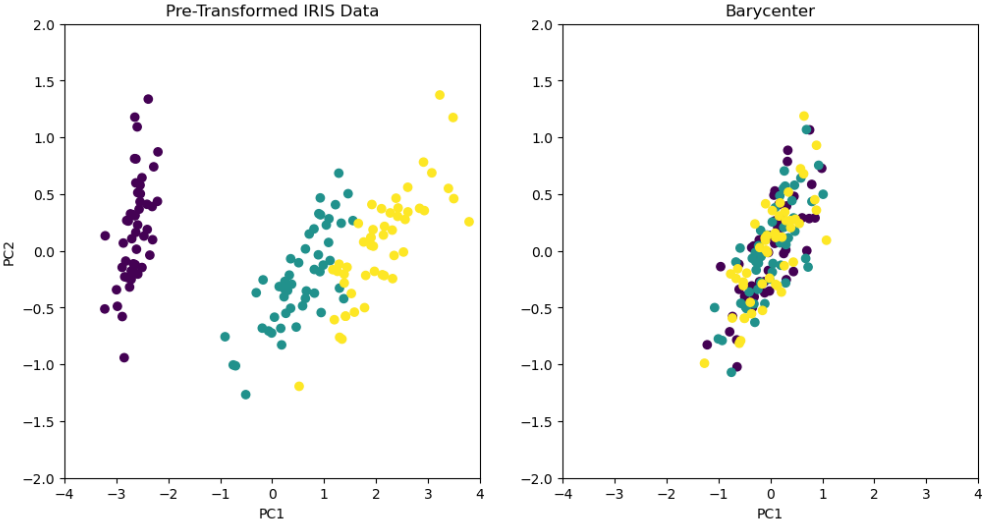


Figure 2: Visualization of the Iris dataset features on the first two principle component axis before and after reducing the variability induced by the floral classes. Each color represents a unique floral class.

6 Conclusion

We propose a general framework for reducing the variability of specified features from a data distribution using optimal transport. From the general framework, we then introduced a data-driven formulation and relaxation for inducing independence between the transformed data distribution and the specified features for practical applications. In the context of this framework, we highlight optimization tricks in practice and showcased the effectiveness of our method on simulated datasets.

Limitations There are two particularly strong limitations of the current method that comes to mind. Firstly, the flexibility of this method also requires significant trial-and-error with different hyperparameters in practice. In particular, there is no clear "best" choice of a sufficiently large λ, the kernels, and their corresponding bandwidths for KDE. Secondly, the semi-supervised variability reduction regime can be used for classification of the unknown observations by design, where we seek to predict the values of z. However, preliminary experiments show that the accuracy is subpar compared to most simple classifiers on relatively simple datasets.

Future Works Although our discussion focuses on reduction of variability from unwanted variable, our framework provides an intuitive measure of factor importance implicitly. Specifically, one could quantify the important of a particular factor by measuring the reduction in variability as a result of filter out that factor. By searching over the space of z to maximize this reduction in variability, our work naturally extends into a solution for a *hidden factor discovery* problem.

On the other hand, our work also lays the groundwork for a new data augmentation technique: suppose that there are few observations x_i associated with a factor of interest z^*, we can create more samples of observations associated with z^*, call it x_i^*, by pushing all the observations x_i to the barycenter y_i and then reversing them with using the inverse map $T_{z^*}^{-1}$. Although our formulation does not include a direct computation of the transport map, we believe it is feasible to perform the reversal by reformulating Eq. (3).

Another promising direction is to utilize our method as an unsupervised pre-training procedure on the data distribution prior to downstream supervised training, which has been demonstrated to be helpful empirically [4]. Since our method can reduce the variability induced by unwanted variables, we are hopeful that this may result in more robust results in unison with deep neural networks, which are known to overfit to noises in the supplied data. In the context of machine learning fairness, it may be better to reduce the data variability associated with sensitivity attributes prior to the supervised training procedure as opposed to adding a regularization to penalize prediction dependence on sensitivity attributes. This intuition stems from the fact that with our method, the model only has access to a data set with the dependencies to sensitivity attributes removed a priori, and therefore should not use information related to them during the learning process.

7 Acknowledgements

This research is supported by the National Science Foundation under grant no. DMS-1646339 as part of the Applied Math Summer Undergraduate Research Experience hosted by the Courant Institute of Mathematical Sciences at New York University. The authors would like to thank the support of Aleksandar Donev, Fortino Garcia, and Olivia Pomerenk throughout the program.

8 References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 214–223.

[2] Charlotte Bunne et al. "Learning Single-Cell Perturbation Responses using Neural Optimal Transport". In: *Nature Methods* (Aug. 2023).

[3] Nicolas Courty et al. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

[4] Dumitru Erhan et al. "Why Does Unsupervised Pre-training Help Deep Learning?" In: *Journal of Machine Learning Research* 11.19 (2010), pp. 625–660.

[5] Toshihiro Kamishima et al. "Fairness-Aware Classifier with Prejudice Remover Regularizer". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50. isbn: 978-3-642-33486-3.

[6] Z. Paul, D. Smirnov, and J. Solomon. "Wassersplines for neural vector-field controlled animation". In: *Symposium on Computer Animation (SCA)*. 2022.

[7] Ievgen Redko et al. "Optimal Transport for Multi-source Domain Adaptation under Target Shift". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 849–858.

[8] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. issn: 00359246. (Visited on 07/27/2023).

[9] J. Zhang and S. Y. Philip. *Optimal transport: Old and new*. Springer Science Business Media, 2008.