What Happens to a Dataset Transformed by a Projection-based Concept Removal Method?

Richard Johansson
Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology
richard.johansson@gu.se

Abstract

We investigate the behavior of methods that use linear projections to remove information about a concept from a language representation, and we consider the question of what happens to a dataset transformed by such a method. A theoretical analysis and experiments on real-world and synthetic data show that these methods inject strong statistical dependencies into the transformed datasets. After applying such a method, the representation space is highly structured: in the transformed space, an instance tends to be located near instances of the opposite label. As a consequence, the original labeling can in some cases be reconstructed by applying an anti-clustering method.

Keywords: language representation, concept removal, projection, method analysis

1. Introduction

While most research in representation learning for NLP focuses on what information *is* encoded in a representation, in several scenarios it is important to be able to control what *is not* encoded. Most of the discussion in the NLP community has focused on demographic attributes (Bolukbasi et al., 2016). Another area of application is in domain adaptation: intuitively, if representations are uninformative about which domain a data point was sampled from, learned predictors based on those representations should generalize more robustly across domains (Ganin and Lempitsky, 2015).

A wide range of methods have been developed to learn a transformation of representations to try to enforce invariance with respect to a given concept while training machine learning models. While early approaches were mainly based on adversarial training (Ganin and Lempitsky, 2015), a number of recent methods have used *linear projections* for concept removal. For instance, the Iterative Nullspace Projection (INLP) method (Ravfogel et al., 2020) projects into a nullspace orthogonal to a set of linear models trained to predict the concept we wish to remove. Compared to adversarial methods, projection-based methods are mathematically more stable, more efficient, easier to implement, and have performed better in comparative evaluations.

We focus here on the use case where we want to transform the representations in a given *dataset* to make a given concept impossible to recover. For instance, for a given set of word embeddings, we may want to create a transformed set where the gender variable is impossible to predict, and then distribute this transformed set to the public. Other application areas include those where we want to carry out a statistical analysis on a dataset and ensure that some concept does not influence the analysis; for instance, Daoud et al. (2022) discuss the challenges to text-based causal inference methods caused by *treatment leakage*, that is when texts are contaminated by the treatment variable. Naively, a user could think that the direct application of projection-based concept removal would lead to a processed dataset resembling one sampled from a distribution where the concept is statistically independent of the representations: projection-based methods are claimed to “remove the linear information” about the undesired concept. To what extent is it actually true that the information about the concept is removed from the dataset?

In this paper, we investigate properties of datasets where a projection-based concept removal method has been applied to a dataset as a whole. The main takeaway is that the transformed representation space is highly structured: the assumption of independent and identically distributed (i.i.d.) instances does not hold after applying the method. Instead of resulting in statistical independence between the representation and the concept, we show that the concept is reflected in dependencies between rows (instances) in the transformed datasets. This injected row-wise dependence is present even in cases where there was no statistical dependence between the representations and the concept in the first place. We discuss the technical reasons for why this is the case, and then carry out a series of experiments to investigate the consequences of this observation. Our findings include the following:

- Cross-validation accuracies for predicting the removed concept in transformed datasets are lower than chance.
- The distribution of prediction probabilities for

cross-validated classifiers trained on projected representations are significantly different from those trained on i.i.d. data.

- In the transformed dataset, instances tend to be near those of the opposite category.
- The original labels can sometimes be decoded from the transformed dataset by applying anticlustering methods.

We finally discuss the implications of these findings for practitioners using projection-based concept removal methods to process datasets.

2. Concept Removal Methods

Most early work on methods that remove a concept was based on adversarial methods originally developed for learning domain-invariant representations (Ganin and Lempitsky, 2015). Adversarial methods have, among other use cases, been applied for the removal of demographic attributes (Raff and Sylvester, 2018; Li et al., 2018; Barrett et al., 2019).

Adversarial training is often unstable in practice and can be difficult to train because of the minimax objective. A mathematically more straightforward approach is to use a linear *projection*, originally introduced by Xu et al. (2017). Although recent progress in NLP highlights the importance of representations computed using nonlinear functions, it seems that in practice linear projections work well for concept removal even when nonlinear predictors are used. Ravfogel et al. (2020) proposed the Iterative Nullspace Projection (INLP) method that is one of the methods we consider in this paper. INLP iteratively trains a linear classifier to predict the concept, and then projects into the subspace orthogonal to the normal vector of the classifier's separating hyperplane.

More recently, a range of methods intended to improve over INLP have been developed. Ravfogel et al. (2022) unified the projection-based and adversarial families, and presented a method called R-LACE that finds a projection adversarially. Belrose et al. (2023) presented a theoretical formalization of conditions for linear guardedness and an approach to finding optimal projections. Haghighatkhah et al. (2022) described two variants of *mean projection* (MP), where the difference vector between the class centroids defines the projection, and they argued that this method is more effective and less intrusive than INLP.

3. Theoretical Analysis

We investigate the structure of datasets where projection-based concept removal methods have been applied, and we are interested in how such datasets differ from a normal dataset where the representation X is statistically independent of the concept Y. In this section, we take an analytical perspective and explain theoretically the structured arrangement of data points in the transformed space. In the next section, we show the results of empirical investigations complementing the theoretical analysis.

Our main result shows that instances after projection have an adversarial arrangement where each instance tends to be located close to those of the opposite label. For simplicity of analysis, we limit this analysis to MP (Haghighatkhah et al., 2022), which is equivalent to applying INLP with a nearest centroid classifier to find the projection vector. A full analysis of the general case is beyond the scope of this work because it depends on the data-generating distribution as well as the choice of method used to define the projections.

Theorem. *Let $X \in \mathbb{R}^{m,n}$ be a feature matrix and $Y \in \{0,1\}^m$ the class labels. MP is then applied to X with respect to Y and we refer to the result as X_{MP}. We carry out a leave-one-out cross-validation in the transformed dataset where we set a single instance x_i, y_i aside and train a nearest-centroid classifier on the remaining data. In this case, x_i cannot be classified with a positive margin by this classifier: that is, x_i is either misclassified or exactly on the classifier's decision boundary.*

Proof. In MP, the vector used to define the projection is equal to the difference between the class centroids in the original dataset X, Y. This means that in the projected dataset X_{MP}, the two class centroids c_0 and c_1 are identical. For ease of exposition, assume that $y_i = 0$ and that the number of instances in class 0 is $n_0 > 1$. The centroids of the leave-one-out classifier are $c_0' = \frac{n_0 c_0 - x_i}{n_0 - 1}$ and $c_1' = c_1$. Now, we have one of two cases. If x_i is identical to c_0, then $c_0' = c_0 = c_1$ so this instance is exactly on the classifier's decision boundary. Otherwise, the removal of x_i shifts the center of mass of c_0' in the direction away from x_i, so x_i is closer to c_1' than to c_0' and the instance is misclassified. □

This shows that the transformed dataset is fundamentally different from one where the representation and the label are statistically independent: if that were the case, the probability of an instance being classified correctly should correspond to the prior probability of its class.

It should be stressed that the first case (that the instance is exactly on the decision boundary) happens only in the theoretical case that the instance coincides exactly with its class centroid. In reality, the probability of this to occur is small in practice and we have never observed it experimentally: all LOO cross-validation accuracies we have seen with MP have been exactly 0. We imagine that the corner case may occur more frequently in datasets where many instances are identical.

4. Experiments

In the following, we carry out a set of experiments illustrating the consequences of the observations described in §3. We focus on INLP here to complement the theoretical analysis of MP in the previous section. Tentative experiments indicate that the tendencies are similar when applying R-LACE, but we do not investigate this algorithm thoroughly because it is computationally more demanding.

4.1. Datasets

We carry out the experiments on synthetic and real-world natural language datasets. The synthetic data was used for investigating how INLP behaves when applied to data that does *not* contain any signal representing the concept. For the X variables, we generated instances from a standard isotropic multivariate Gaussian. The labels Y were balanced.

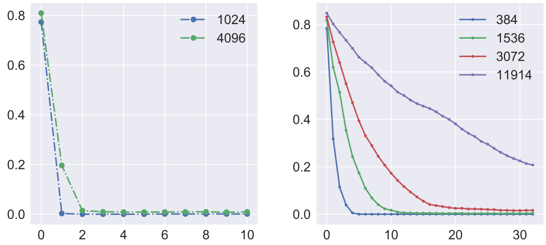
For experiments using natural data, we used six domains from the sentiment classification corpus collected by Blitzer et al. (2007). This corpus associates each document with a positive or negative polarity label; the label distribution is balanced. We did not use the domain information. In the experiments, the representations X were tfidf-weighted bag of words (BoW) and the output of a BERT model (Devlin et al., 2019) at the `[CLS]` token. We considered different values of the number n of instances, and set the number d of features to 2^{10} for random and BoW representations. (The overall picture is similar with other values of d.) For BERT, the number of dimensions is 768.

For all datasets, we applied the INLP algorithm. As described above, the algorithm iteratively trains a linear classifier, and we used a L_2-regularized logistic regression model for this purpose. We ran INLP for several iterations and the result after each iteration will be considered in the experiments.

4.2. Prediction Accuracy

We applied INLP to the datasets and computed 32-fold cross-validation accuracy scores for predicting the removed concept. In all experiments, we used a L_2-regularized logistic regression model ($C = 1$) applied to the L_2-normalized output of the INLP algorithm. Figure 1 shows the accuracies over the INLP iterations for the BoW and BERT representations. We show the results for different sizes n of the dataset.

Clearly, the behavior of the model is different from what would have been expected if the instances were i.i.d. and X independent of Y. Even after just a few iterations of INLP, accuracies fall far below the chance level. This tendency is strongest for the BoW representation, which falls to zero almost immediately. For BERT, we see the same overall picture although INLP requires more iterations, in particular when the dataset grows larger. Presumably, this is because the information represented by BERT is more difficult to express using a linear model.

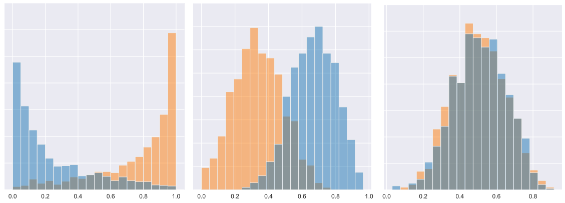


(a) Bag of words ($d = 2^{10}$). (b) BERT ($d = 768$).

Figure 1: Cross-validated accuracy scores for predicting the removed concept over INLP iterations. Each curve corresponds to a size n of the dataset.

4.3. Predicted Probabilities

To further illustrate the behavior of predictive models trained on projected representations, we considered how probabilities predicted by the models are distributed. Figure 2 shows the distributions of predicted probabilities for the sentiment dataset. We show the outputs of a model trained on the unprocessed BERT representations and on projected representations (10 iterations of INLP). To compare with a situation where representations are independent of the labels, we also include probabilities predicted by a model trained on random labels independent of the text, and we see a clear difference between the projected and the independent settings.

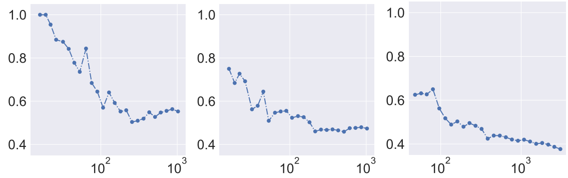


(a) Original. (b) Projected. (c) Independent.

Figure 2: Distribution of predicted probabilities for the positive (orange) and negative classes (blue).

4.4. Neighborhood Structure

To investigate the arrangement of instances in the projected feature space, we carried out an experiment where we look at how frequently the Euclidean nearest neighbors are of the opposite value of the target concept. Intuitively, one would expect that when X and Y are unrelated and the instances i.i.d., this proportion should be around 0.5, while it would be expected to be close to 0 if there is a strong association between X and Y.



(a) Random. (b) Bag of words. (c) BERT.

Figure 3: Proportion of instances whose nearest neighbor is of the opposite label, for different n.

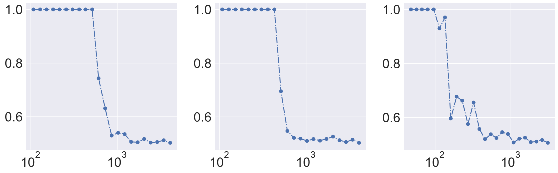
Figure 3 shows these proportions for different data set sizes n, and the tendency to place instances near those of the opposite label is clearly visible. This again illustrates the non-i.i.d. distribution of the projected representations. This tendency is most pronounced when $d \gg n$.

It is important to note that *group-based* statistical measures that quantify the strength of association between X and Y can be misleading because the effects discussed here are discernible for *individual* instances. To illustrate, we computed the MMD (Gretton et al., 2012) of BERT representations between the positive and negative groups, and we saw that the estimates steadily decrease as we apply INLP iterations, despite the projected dataset becoming *more* informative about the labels.

4.5. Recovering the Original Grouping

The theoretical result in §3 and the empirical observation from §4.4 that instances tend to be located close to instances of the opposite label gives an intuition for a procedure that recovers the groups defined by the original labeling. Intuitively, we can partition the data points into groups selected so that each instance is maximally dissimilar to the other instances in the same group. This reverses the logic of regular clustering models and has been referred to as *anti-clustering* (Späth, 1986). For instance, we can adapt Lloyd's algorithm for k-means clustering to the anti-clustering setup, simply by changing the algorithm to assign an instance to the cluster it is *least* similar to.

We applied the `anticlust` R package (Papenberg and Klau, 2021) using two clusters, the diversity criterion and 100 repetitions of the search method by Brusco et al. (2020). The clusters were then compared to the original labels of the datasets. Figure 4 shows the cluster purity scores.



(a) Random. (b) Bag of words. (c) BERT.

Figure 4: Cluster purity scores comparing the original labeling to the anti-clustering result.

We observe that the anti-clustering algorithm applied to the projected representations often perfectly reconstructs the grouping defined by the concept we wanted to remove, in particular when $d \gg n$. As we have already argued, projection inscribes the training labels into the data, and a reconstruction is possible even if the original dataset was random and unrelated to the training labels.

5. Related Work

This investigation falls into the category of work that analyzes the behavior of concept removal methods. Most of the early discussions focused on the pros and cons of adversarial methods. For instance, Elazar and Goldberg (2018) claimed that these methods leak information; their conclusions were later challenged by Barrett et al. (2019).

The work that is most similar in spirit to ours is arguably the investigation by Gonen and Goldberg (2019), which analyzed the geometric structure of word embedding models processed by gender debiasing methods. They argued that debiasing does not remove the gender information, but only stores it a less obvious way, and they showed that this information could be recovered by considering distances in the processed space.

6. Implications and Conclusion

How much does it matter in practice that instances in a projected dataset are not i.i.d.? Projection-based concept removal methods are useful for the purpose for which they were originally developed: transforming a dataset to make sure that a ML *model* trained on the transformed data does not rely on the target concept. However, a naive practictioner may get the misguided impression that the projection "removes information" about the concept from the *dataset* itself, when the opposite is in fact true. Clearly, one needs to be careful if we want to use projection for the purpose of scrubbing some signal from a dataset before distributing it. We should also stress that the effects discussed here are not problematic in case one can afford to set aside a subset of the data reserved for the purpose of training the projection: the case we focus on assumes that we want to use the *whole* dataset.

A consequence of the i.i.d violation is that any statistical analysis requiring strict i.i.d. assumptions is likely to be invalid if applied to representations computed by a projection-based method. For instance, text-based causal inference methods (Keith et al., 2020) involving the text representation and the removed concept may be affected if projection is applied: such causal inference methods typically rely on predicted probabilities or representation similarity, which as we have seen in §4.3 and §4.4 are strongly affected. Daoud et al. (2022) and Gui and Veitch (2023) highlight the problem for causal inference when the text encodes information about a variable of interest, and our results suggest that it could be risky to try to apply projection to remove this undesired information. Effects on predictions in cross-validations (§4.2) are visible already in moderately low-dimensional settings.

7. Limitations

There are a number of ways in which this work could be put on firmer ground theoretically. In §3, we limited the theoretical analysis to MP (or equivalently, INLP based on a nearest centroid classifier), and in future work we would like to find a more general formal justification for why the adversarial arrangement emerges. In the empirical section, we would also like to take a more general approach in the future and investigate additional concept removal methods, such as more recent projection-based methods as well as adversarial representation learning methods.

Furthermore, we do not have a clear understanding of the role played by the dimensionality d in relation to the dataset size n. The experiments (§4.4 and §4.5) indicate that that such effects play a role, but this is currently not taken into account in the theoretical analysis.

8. Ethical Discussion

Whether the behaviors investigated here matter in practice depend on the application, and as discussed above, the consequences are likely to be limited if the only purpose of the processed representations is for training a model. In other cases, in particular when the intention is for the projected dataset to be distributed, the effects may be more problematic. For instance, if projection is applied to a set of word embeddings in order to make them invariant to a demographic attribute, we may accidentally encode information about the attribute into the embedding geometry, so that it can later be decoded from representations.

Furthermore, the fact that in many cases the original groups can be reconstructed from the projected data (§4.5), even if the original dataset did not encode any information about the target concept, shows that projection-based methods should not be viewed as privacy-preserving (Coavoux et al., 2018). To be clear, the inventors of the methods we have considered did not claim that they are intended to ensure privacy,[1] but again it is important for users to understand that projection is not equivalent to information removal in a dataset.

9. Acknowledgements

The results presented here are a by-product of discussions with Adel Daoud. This research was supported by the projects *Interpreting and Grounding Pre-trained Representations for NLP* and *Representation Learning for Conversational AI*, both under the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the project *Countering Bias in AI Methods in the Social Sciences* under the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

10. Bibliographical References

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong. Association for Computational Linguistics.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063. Curran Associates, Inc.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual*

[1] In contrast, Xu et al. (2017) explicitly considered projection for privacy, but we have not investigated their method.

Meeting of the Association of Computational Linguistics, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Michael J. Brusco, J. Dennis Cradit, and Douglas Steinley. 2020. Combining diversity and dispersion criteria for anticlustering: A bicriterion approach. *British Journal of Mathematical and Statistical Psychology*, 73(3):375–396.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Adel Daoud, Connor Jerzak, and Richard Johansson. 2022. Conceptualizing treatment leakage in text-based causal inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5638–5645, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, United States. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Lin Gui and Victor Veitch. 2023. Causal estimation for text data with (apparent) overlap violations. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8395–8416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Martin Papenberg and Gunnar W. Klau. 2021. Using anticlustering to partition data sets into equivalent parts. *Psychological Methods*, 26(2):161–174.

Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D. Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

H. Späth. 1986. Anticlustering: Maximizing the variance criterion. *Control and Cybernetics*, 15(2):213–218.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208. Curran Associates, Inc.

Ke Xu, Tongyi Cao, Swair Shah, Crystal Maung, and Haim Schweitzer. 2017. Cleaning the null space: A privacy mechanism for predictors. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2789–2795, San Francisco, United States. AAAI Press.