Prediction under Latent Subgroup Shifts with High-Dimensional Observations

William I. Walker [1] **Arthur Gretton** [1] **Maneesh Sahani** [1]

Abstract

We introduce a new approach to prediction in graphical models with latent-shift adaptation, i.e., where source and target environments differ in the distribution of an unobserved confounding latent variable. Previous work has shown that as long as "concept" and "proxy" variables with appropriate dependence are observed in the source environment, the latent-associated distributional changes can be identified, and target predictions adapted accurately. However, practical estimation methods do not scale well when the observations are complex and high-dimensional, even if the confounding latent is categorical. Here we build upon a recently proposed probabilistic unsupervised learning framework, the recognition-parametrised model (RPM), to recover low-dimensional, discrete latents from image observations. Applied to the problem of latent shifts, our novel form of RPM identifies causal latent structure in the source environment, and adapts properly to predict in the target. We demonstrate results in settings where predictor and proxy are high-dimensional images, a context to which previous methods fail to scale.

1. Introduction

In real world prediction problems, challenges often arise owing to shifts between training and test distributions. For example, suppose we want to predict language exam performance (Y) given past written essays (X) from the same students. A naive model trained on students from one school may fail to accurately predict the performance of students from another where the mix of socio-economic backgrounds differs, as these contextual factors can affect not only the marginal distributions of prior and future scores, but also the relationship between them. Attempts to mitigate the impact of such distribution changes on performance of tasks such as classification depend on assumptions about which distributions have changed and which probabilities are conserved. In covariate shift (Shimodaira, 2000), the distribution of the observed covariate X changes but the conditional probability of $Y|X$ remains the same. In the label shift setting (Gart & Buck, 1966), the distribution of the predicted variable Y shifts but the conditional dependence of $X|Y$ is preserved. However, in our school example, both the distributions on X and Y as well as their conditionals may shift, and these assumptions are not general enough.

Latent shift assumes a change in the distribution of an unobserved confounding latent subgroup variable U that influences both X and $Y|X$, while the conditional dependence of X, Y and any other observed variables on U are preserved. Alabdulmohsin et al. (2023) introduce a graph with proxy and concept variables observable in the source environment, which allows the identification of the distribution of U and its confounding effects. They show that this model can be used to adapt predictions in the target environment to shifts in the distribution of U.

However, the learning algorithm they propose has difficulty scaling to settings where the observed covariates are high-dimensional and complex, such as images or documents. This is a setting in which the Recognition-Parametrised Model (RPM) (Walker et al., 2023) is helpful. The RPM makes it possible to learn a tractable underlying latent graphical structure from high dimensional observations, without needing an explicit generative model. Here we propose a novel adaptation of the RPM in which factors of the graph involving high dimensional observations are parametrised in recognition form, while keeping generative factors for categorical variables including the predicted variable Y. We refer to this model as a "partial" RPM. We show that it can be used to identify the confounding latent within a source distribution P with complex observed variables, and to adapt for latent subgroup shift in a target distribution Q, outperforming previous models as the dimensionality of the observation increases. We demonstrate an application in which the observations are images from the CIFAR-10 dataset, and quantify predictive performance after shift for different observed sample sizes in source and target.

[1]Gatsby Computational Neuroscience Unit, UCL, London, UK. Correspondence to: William I. Walker <william.walker.18@ucl.ac.uk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

2. Previous Work

Many recent studies have sought to learn models that adapt to shifts in distribution from source P to target Q, with different assumptions on which conditional probabilities are preserved. Covariate shift (Shimodaira, 2000; Zadrozny, 2004; Huang et al., 2006; Gretton et al., 2008; Bickel et al., 2009; Sugiyama & Kawanabe, 2012; Chen et al., 2016; Schneider et al., 2020) considers the case in which $P(X) \neq Q(X)$ but the conditional probability $P(Y|X) = Q(Y|X)$ stays the same. This is usually solved by reweighting the classifier loss by $Q(X)/P(X)$ in an attempt to make the source data look like it was drawn from the target data. Similarly, label shift (Manski & Lerman, 1977; Rosenbaum & Rubin, 1983; Saerens et al., 2002; Forman, 2008; Storkey, 2009; Plessis & Sugiyama, 2013; Zhang et al., 2013; Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020; Garg et al., 2020; Tachet des Combes et al., 2020; Wu et al., 2021) considers the case where $P(Y) \neq Q(Y)$ but the conditional probabilities $P(X|Y) = Q(X|Y)$ remain the same. Again, this shift is usually mitigated by reweighting the classifier loss, this time with $Q(Y)/P(Y)$. Both the problem settings above have been framed as assumptions on the causal structure of the generation of the data (Schölkopf et al., 2012).

Latent shift was framed in the causal view by Yue et al. (2021), where a shift in distribution of an unobserved latent confounder changes the distribution on both X and Y. This work finds a proxy for X and one for Y, both caused by U, and finds an invariant bridge function to remove the effect of the latent shift. However, as pointed out by Alabdulmohsin et al. (2023), this approach does not guarantee that the mapping from U to either proxy is identifiable, a condition which is necessary to correctly adapt to the shift in distribution of U. Alabdulmohsin et al. (2023) solve the identifiability problem of U by introducing proxy variable W and concept variable C as in the graph of Figure 1. Both C and W are observed in source P but not in target Q. They then introduce an algorithm to adapt to shifts in the latent distribution under the assumption that all probabilities conditioned on U are preserved between source and target. However, as we demonstrate here, this algorithm does not scale well as the dimensionality of X and W grow, although such high-dimensional covariates may be encountered in realistic settings. Our goal in this study is to develop a new approach to estimation and prediction in the latent-shift setting that scales to real-world dimensionalities

3. Latent Shift Adaptation

We follow Alabdulmohsin et al. (2023) by assuming the graphical model of Figure 1, where U is the unobserved latent, X is the observed covariate, Y the variable to be predicted, W is the proxy variable, and C is a concept variable.

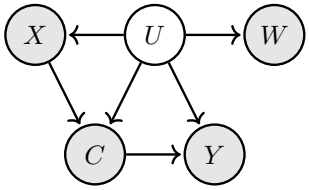

Figure 1. Graphical model of observations and confounding latent U. X is the observed variable, C is the concept variable, W is the proxy variable, and Y is the predicted variable. All of $\{X, C, W, Y\}$ are observed in the source, but only X is available in the target.

The concept C ensures that $Y \perp\!\!\!\perp X|C, U$, while the proxy W provides sufficient additional information about U to make it identifiable. Further discussion about the inclusion of C and W can be found in (Alabdulmohsin et al., 2023). We take C, Y, and U to be discrete variables whereas X and W may be discrete or continuous, possibly vector-valued and high-dimensional. Our goal is to predict Y given observations X from the target distribution Q, having trained on observations $\{X, C, W, Y\}$ from the source distribution P. We assume the distribution of the unobserved latent U can shift between P and Q, but that all probabilities conditioned on U are preserved across the source and target.

3.1. The Partial Recognition-Parametrised Model

Let X and W be (discrete or continuous) vector-valued, observed random variables and C and Y be discrete observed random variables. We want to identify the underlying discrete latent variables U giving rise to the conditional independencies shown in the graph of Figure 1. The conditional dependence structure implies the factorisation

$$p(U,C,Y,X,W) = p(X|U)p(W|U)p(C|X,U)p(Y|C,U)p(U)$$

The Recognition-Parametrised Model (RPM; Walker et al., 2023) makes it possible to learn a flexible recognition model (e.g. a neural network) without an explicit generative model for the observations. This is appropriate to the current setting, where there is no need to generate either X or W in either source or target setting. Therefore, we introduce RPM-like terms in place of the conditional dependencies $p(X|U)$ and $p(W|U)$

$$\begin{aligned} p(X|U) &\rightarrow \frac{P_{0x}(X) f_{\theta_{\mathbf{x}}}(U|X)}{F_{\theta_{\mathbf{x}}}(U)} \\ p(W|U) &\rightarrow \frac{P_{0w}(W) f_{\theta_{\mathbf{w}}}(U|W)}{F_{\theta_{\mathbf{w}}}(U)} \end{aligned} \tag{1}$$

$P_{0x}(X) = \frac{1}{N}\sum_n \delta(X - X^{(n)})$ is the empirical measure with atoms at the N data points $X^{(n)}$. Similarly, $P_{0w}(W) = \frac{1}{N}\sum_n \delta(W - W^{(n)})$. The factor $f_{\theta_{\mathbf{x}}}(U|X)$ is a parametrised recognition distribution and $F_{\theta_{\mathbf{x}}}(U) =$

$\int dX\, P_{0x}(X) f_{\theta_{\mathbf{x}}}(U|X)$ is the mixture with respect to P_{0x}. The factors $f_{\theta_{\mathbf{w}}}$ and $F_{\theta_{\mathbf{w}}}$ are defined analogously. Together, these terms define a normalised joint model

$$\mathsf{P}_{\theta,\mathbb{X}^{(N)},\mathbb{W}^{(N)}}(U,C,Y,X,W) = P_\phi(C|X,U)P_\psi(Y|C,U)$$
$$P_{\theta_P}(U)\frac{P_{0x}(X)f_{\theta_{\mathbf{x}}}(U|X)}{F_{\theta_{\mathbf{x}}}(U)}\frac{P_{0w}(W)f_{\theta_{\mathbf{w}}}(U|W)}{F_{\theta_{\mathbf{w}}}(U)} \quad (2)$$

where θ_x, θ_w, θ_p, ϕ, and ψ are learned parameters. $P_{\theta_P}(U)$ is a normalised distribution on the latent. The observed datasets $\mathbb{X}^{(N)} = \{\mathcal{X}^{(1)} \dots \mathcal{X}^{(N)}\}$ and $\mathbb{W}^{(N)}$ appear in the subscript to highlight that the model parametrisation depends on these data through P_{0x} and P_{0w} and the mixtures.

3.2. RPM Latent Adaptation

We observe data $\{X, Y, C, W\}$ from source P, and $\{X\}$ from target Q and seek to estimate $Q(Y|X)$. The steps for learning and adaptation in our model are detailed below.

Step 1: Learn RPM for source P. The parameters $\{\theta_x, \theta_w, \theta_u, \phi, \psi\}$ of the source model of Equation (2) are learnt by maximizing the free energy using Expectation-Maximization (EM).

$$\begin{aligned}\mathcal{F} \underset{+C}{=} &\langle \log f_{\theta_{\mathbf{x}}}(U|X)\rangle_{\eta(U)} - \langle \log F_{\theta_{\mathbf{x}}}(U)\rangle_{\eta(U)} \\ &+ \langle \log f_{\theta_{\mathbf{w}}}(U|W)\rangle_{\eta(U)} - \langle \log F_{\theta_{\mathbf{w}}}(U)\rangle_{\eta(U)} \\ &+ \langle \log P_\phi(C|X,U)\rangle_{\eta(U)} + \langle \log P_\psi(Y|C,U)\rangle_{\eta(U)} \\ &+ \langle \log P_{\theta_P}(U)\rangle_{\eta(U)} + H[\eta(U)] \quad (3)\end{aligned}$$

where $\eta(U)$ is the variational distribution and H is the entropy function. The expectation step to update $\eta(U)$ in EM is exact for discrete U (Walker et al., 2023) and so EM will converege to a (local) maximum of the likelihood.

Step 2: Learn the shifted prior $Q_{\theta_Q}(U)$ on target. Now only using $\{X\}$ from target Q, we learn the joint $Q(U, X)$ by maximizing the free energy. Note that any conditional probability dependent on U remains the same as in P. Only the prior changes: $P(U) \to Q(U)$. Thus, we can reuse parameters learned from the source distribution. We then have:

$$\begin{aligned}Q(U|X) &= \frac{Q(X|U)Q_{\theta_Q}(U)}{Q(X)} \\ &= \frac{P(X|U)Q_{\theta_Q}(U)}{Q(X)} \\ &\propto \frac{f_{\theta_{\mathbf{x}}}(U|X)}{F_{\theta_{\mathbf{x}}}(U)}Q_{\theta_Q}(U) \quad (4)\end{aligned}$$

Now we maximize the free energy to learn the parameter θ_Q while keeping all other parameters fixed

$$\begin{aligned}\mathcal{F} \underset{+C}{=} &\langle \log f_{\theta_{\mathbf{x}}}(U|X)\rangle_{\nu(U)} - \langle \log F_{\theta_{\mathbf{x}}}(U)\rangle_{\nu(U)} \\ &+ \langle \log Q_{\theta_Q}(U)\rangle_{\nu(U)} + H[\nu(U)] \quad (5)\end{aligned}$$

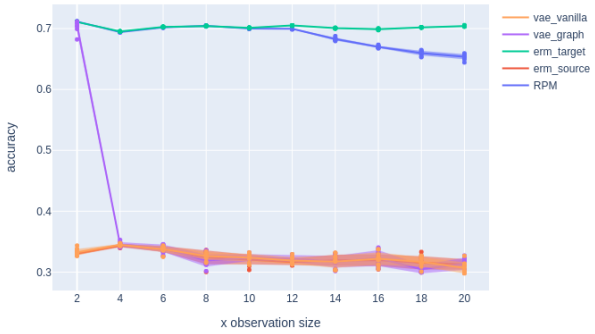

Figure 2. Accuracy in predicting Y given X from the target distribution, as a function of the dimensionality of X. erm_source predicts using a model trained on source distribution X and Y, giving a lower bound on prediction performance. erm_target predicts using an oracle model trained on target distribution X and Y, giving an upper bound on prediction performance.

where $\nu(U)$ is the variational distribution which can be computed analytically in the expectation step of EM.

Step 3: Make predictions $Q(Y|X)$. Using all the factors learned in the previous steps, make predictions

$$\begin{aligned}Q(Y|X) &= \sum_{U,C} Q(Y|C,U)Q(C|U,X)Q(U|X) \\ &= \sum_{U,C} P_\psi(Y|C,U)P_\phi(C|U,X)Q(U|X) \quad (6)\end{aligned}$$

where again we use the fact that by conditioning on U, we can use the probabilities learned on the source P.

4. Experiments

4.1. Scaling with observation dimensionality

To show how the partial RPM is able to handle complex observations, we used a simulated numerical setting with varying dimensionalities of X.

We constructed data with binary U, Y, W, and C but continuous X. In the source distribution P, U was drawn with $P(U = 1) = 0.9$ and in the target Q, $Q(U = 1) = 0.1$. The first two dimensions of X were drawn from a Gaussian with mean dependent on U. All the other dimensions of X were simply $\{10, -10\}$ with probability 0.5 and had no relevance to the prediction of Y. This structure was chosen based on the hypothesis that previous methods using the structured Wassertein Autoencoder (WAE) or Variational Autoencoder (VAE) (Kingma & Welling, 2014), which must learn to reconstruct the observation X, would be challenged as irrelevant dimensions of X were added and so learn a worse representation of U. Further details of the generative model are given in Appendix A.

We compared our approach to several baselines. Empirical risk minimization is a neural network (NN) trained on source data (erm_source) to predict Y from X. It is then simply applied to target data X and serves as a lower bound to prediction. Conversely, erm_target is an oracle network trained on *target* data X and Y and is an upper bound to prediction performance. Latent shift adaptation using the structured WAE (vae_graph) (Alabdulmohsin et al., 2023) is trained with a recognition model $\hat{p}_e(\tilde{U}|X, C, Y, W)$ and a decoder that respects the conditional independence structure of the model and reconstructs all observations by $\hat{Y} = f_y(C, U)$, $\hat{C} = f_c(X, U)$, $\hat{X} = f_x(U)$, and $\hat{W} = f_w(U)$. Details of its training can be found in the Appendix A. Latent shift adaptation using the vanilla VAE (vae_vanilla) (Alabdulmohsin et al., 2023) is trained with the same recognition model to U $\hat{p}_e(\tilde{U}|X, C, Y, W)$ but the decoder has no structure $\hat{p}_d(X, C, Y, W|\tilde{U})$. The partial RPM recognition model from X to U is a multilayer perceptron (MLP) with a single hidden layer of 100 units and ReLU activations, the same recognition model used in vae_graph and vae_vanilla, and the same as the NN in erm_source and erm_target. The partial RPM recognition model has binary U as in the true generative model, unlike vae_graph and vae_vanilla where U is allowed 5 categories. This is needed in the VAE models to capture more structure in X, and using binary U in those cases results in worse performance.

As seen in Figure 2, partial RPM accuracy remains high as more dimensions are incorporated into X, while the vae_graph performance falls to the erm_source baseline. This is consistent with our hypothesis that VAE models, unlike the RPM, require the latent U to capture enough information to reconstruct observation X and are thus affected by the noisy dimensions. The RPM avoids this by dispensing with generation, and so focusing only on the information in X that is shared with the other observed variables through U. In the next section, we consider the case where W is also continuous and high-dimensional.

4.2. Latent shift with CIFAR-10 image observations

To demonstrate close to real-world scaling, we allow X and W to be images from different sets of classes in the CIFAR-10 dataset. The data are generated with U, Y, C, $\tilde{X}$, and $\tilde{W}$ discrete assuming sizes of $k_u = 3$, $k_y = 2$, $k_c = 3$, $k_{\tilde{X}} = 2$, and $k_{\tilde{W}} = 3$ respectively. U is drawn from the prior according to $P(U)$ for the source and $Q(U)$ for the target. $P(Y|C, U)$, $P(C|\tilde{X}, U)$, $P(\tilde{X}|U)$, and $P(\tilde{W}|U)$ are all drawn according to probability matrices given in Appendix B. The variables $\tilde{X}$ and $\tilde{W}$ give the class identity of the CIFAR-10 image for X and W, which are then randomly drawn from the corresponding class. We chose X and W from a set of classes that did not overlap. Although X and W are similar here, this is not essential to the RPM model. As $f_{\theta_\mathbf{X}}(U|X)$ and $f_{\theta_\mathbf{W}}(U|W)$ are parameterized by two separate neural networks (NN), the observations can be completely different in dimensionality or statistical structure (Walker et al., 2023).

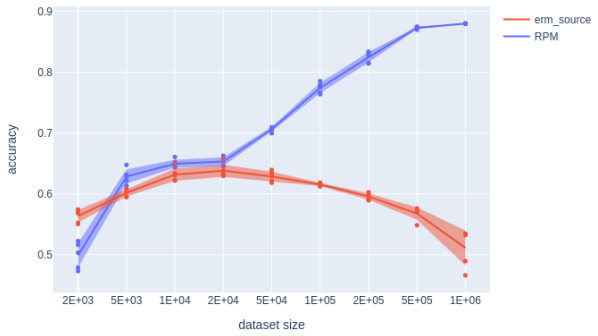

Figure 3. Accuracy in predicting Y given X from the target distribution. erm_source predicts using a model trained on source distribution X images and Y and does not adapt to latent shift.

In the partial RPM, $f_{\theta_\mathbf{X}}(U|X)$ and $f_{\theta_\mathbf{W}}(U|W)$ are convolutional NNs that output discrete probabilities on U and $P_\phi(C|X, U)$ is a convolutional NN to discrete probabilities on C. As all terms are discrete in $P_\psi(Y|C, U)$, its parameters are simply the corresponding probability matrices. Learning is achieved by maximizing the free energy in Equation (3), with adaptation following by the steps elaborated in Section 3.2. In Figure 3, the target adaptation performance is shown as a function of source and target dataset size. The partial RPM (blue) is compared to erm_source (red) with a NN that is the same size as the RPM recognition network for $f_{\theta_\mathbf{X}}(U|X)$ but with output size k_y. erm_source demonstrates that no matter the size of the dataset, simple prediction trained on source will not account for the shift and perform poorly on target. We see here that the partial RPM can adapt to the latent shift even for complex inputs such as CIFAR-10 images with sufficient training data.

5. Conclusion

We have introduced a new method for latent shift adaptation using the RPM. Our method scales to high-dimensional, structured data by avoiding the need to reconstruct that data. We have shown that previous models do not perform well as the dimensionality of the observed covariate is increased, even when the added dimensions only contain noise, while the RPM approach is still able to predict with these observations. Furthermore, we demonstrated that the observations can be highly structured as in CIFAR-10 dateset and the RPM still adapts to latent shifts with enough training data. This makes the RPM a strong model for real world applications.

References

Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M. J., Pfohl, S. R., Salaudeen, O., Schrouff, J., and Tsai, K. Adapting to latent subgroup shifts via concepts and proxies. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9637–9661. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/alabdulmohsin23a.html.

Alexandari, A., Kundaje, A., and Shrikumar, A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pp. 222–232. PMLR, 2020.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts, 2019.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10: 2137–2155, dec 2009. ISSN 1532-4435.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. Robust covariate shift regression. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1270–1279, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/chen16d.html.

Forman, G. Quantifying counts and costs via classification. *Data Min. Knowl. Discov.*, 17:164–206, 10 2008. doi: 10.1007/s10618-008-0097-y.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. A unified view of label shift estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3290–3300. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf.

Gart, J. J. and Buck, A. A. Comparison of a Screening Test and a Reference Test in Epidemiologic Studies: II. A Probabilistic Model for the Comparison of Diagnostic Tests. *American Journal of Epidemiology*, 83(3):593–602, 05 1966. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a120610. URL https://doi.org/10.1093/oxfordjournals.aje.a120610.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. 131Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press, 12 2008. ISBN 9780262170055. doi: 10.7551/mitpress/9780262170055.003.0008. URL https://doi.org/10.7551/mitpress/9780262170055.003.0008.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv*, arxiv:1312.6114, 2014.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Manski, C. F. and Lerman, S. R. The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica*, 45(8):1977–1988, November 1977. URL https://ideas.repec.org/a/ecm/emetrp/v45y1977i8p1977-88.html.

Plessis, M. and Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural networks : the official journal of the International Neural Network Society*, 50C:110–119, 11 2013. doi: 10.1016/j.neunet.2013.11.010.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural computation*, 14:21–41, 02 2002. doi: 10.1162/089976602753284446.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift

adaptation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11539–11551. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/85690f81aadc1749175c187784afc9ee-Paper.pdf.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2, 06 2012.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Storkey, A. *When Training and Test Sets Are Different: Characterizing Learning Transfer*, pp. 3–28. 01 2009. ISBN 9780262170055. doi: 10.7551/mitpress/9780262170055.003.0001.

Sugiyama, M. and Kawanabe, M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 03 2012. ISBN 9780262017091. doi: 10.7551/mitpress/9780262017091.001.0001. URL https://doi.org/10.7551/mitpress/9780262017091.001.0001.

Tachet des Combes, R., Zhao, H., Wang, Y.-X., and Gordon, G. J. Domain adaptation with conditional distribution matching and generalized label shift. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19276–19289. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/dfbfa7ddcfffeb581f50edcf9a0204bb-Paper.pdf.

Walker, W. I., Soulat, H., Yu, C., and Sahani, M. Unsupervised representation learning with recognition-parametrised probabilistic models. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4209–4230. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/walker23a.html.

Wu, R., Guo, C., Su, Y., and Weinberger, K. Q. Online adaptation to label distribution shift. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11340–11351. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5e6bd7a6970cd4325e587f02667f7f73-Paper.pdf.

Yue, Z., Sun, Q., Hua, X., and Zhang, H. Transporting causal mechanisms for unsupervised domain adaptation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8579–8588, 2021.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 114, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015425. URL https://doi.org/10.1145/1015330.1015425.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 819–827, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zhang13d.html.

A. Numerical Simulation Experiment

We describe the numerical experiments in which we change the dimensionality of the observation X. U, Y, C, and W are discrete while X is continuous. We let the dimensions of each of the variables to be $k_U = 2$, $k_C = 2$, $k_Y = 2$, and $k_W = 2$. The dimensionality of X is varied from 2 to 20 in steps of 2. Below is how each variable is generated, closely following (Alabdulmohsin et al., 2023). $\mathbf{o}(U)$ is the one-hot representation of the variable U. I_{k_X} is the identity matrix of size $k_X \times k_X$. Θ is the Heaviside function, $\Theta(z) = 1$ if $z > 0$ and 0 otherwise.

$$
\begin{aligned}
U &\sim \text{Categorical}(\boldsymbol{\pi}) \\
W|U=u &\sim \Theta(\mathcal{N}(\mathbf{o}(U)\boldsymbol{M}_{W|U}, 1)) \\
X[0:2]|U=u &\sim \mathcal{N}(\mathbf{o}(U)\boldsymbol{M}_{X|U}, \boldsymbol{I}_{k_X}) \\
X[2:] &\sim 10 \times (2 \times \text{Bernoulli}(0.5) - 1) \\
C|X=x, U=u &\sim \text{Bernoulli}(\text{logit}^{-1}(X\boldsymbol{M}_{C|X,U=u} + \mathbf{o}(U)\boldsymbol{M}_{C|U})) \\
Y|C=c, U=u &\sim \text{Bernoulli}(\text{logit}^{-1}(C\boldsymbol{M}_{Y|C,U=u} + \mathbf{o}(U)\boldsymbol{M}_{Y|U}))
\end{aligned}
$$

Here are each of the generative parameters:

$$
\begin{aligned}
\boldsymbol{M}_{W|U} = \begin{bmatrix}-3 & 3\end{bmatrix}^T, \boldsymbol{M}_{X|U} = \begin{bmatrix}-0.5 & 0.5 \\ 0.5 & -0.5\end{bmatrix}, \boldsymbol{M}_{C|U} = \begin{bmatrix}-1 & 1\end{bmatrix}^T \\
\boldsymbol{M}_{C|X,U=u_0} = \begin{bmatrix}-1 & 1\end{bmatrix}, \boldsymbol{M}_{C|X,U=u_1} = \begin{bmatrix}1 & -1\end{bmatrix} \\
\boldsymbol{M}_{Y|U} = \begin{bmatrix}-2 & 2\end{bmatrix}^T, \boldsymbol{M}_{Y|C,U=u_0} = \begin{bmatrix}-1 & 1\end{bmatrix}^T, \boldsymbol{M}_{Y|C,U=u_1} = \begin{bmatrix}1 & -1\end{bmatrix}^T
\end{aligned} \tag{7}
$$

For the source distribution, $\boldsymbol{\pi} = [0.1, 0.9]$ while in the target distribution $\boldsymbol{\pi} = [0.9, 0.1]$. We used 7×10^4 samples of both source and target distributions.

We chose these parameters such that any VAE model would struggle to reconstruct the data. To see this, notice that the latent dimensionality k_U is fixed at 2. The first two dimensions of X contain all the relevant information about U and all other dimensions are noise. However, the noisy dimensions have a large scale. Therefore, when the WAE reconstructs the data X, error in the noisy dimensions contribute significantly to the loss. Therefore, the WAE may learn to represent the noisy dimensions in the latent U and ignoring the parts of X that pertain to the latent shift. As the dimensionality of X grows, this problem only worsens. In the RPM, however, there is no reconstruction in the loss term and therefore no need to represent the noisy dimensions. The U that maximizes the loss is only that which makes X conditionally independent of the rest of the graph.

The ERM source and target models both were MLPs with one hidden layer of 100 units with ReLU activations. They were trained for 500 epochs with batch size of 128 and learning rate of 0.01 in stochastic gradient descent (SGD). This learning rate was rediced by a factor of 10 if the training loss did not decrease by 0.01 in the last 20 epochs. The minimum learning rate was set at 10^{-7}. Weight decay of 10^{-6} was used.

The WAE and VAE approaches used a latent of dimensionality 5. The encoder models both were MLPs with one hidden layer of 100 units with ReLU activations. The decoder for WAE used the factored form of the joint distribution implied by the graphical model of Figure 1. The separate decoder networks $\{f_Y, f_C, f_X, f_W\}$ all were MLPs with one hidden layer of 100 units with ReLU activations. Details on the loss function are found in (Alabdulmohsin et al., 2023). The WAE was fit with RMSprop for 500 epochs using 10^-4 learning rate. The annealing strategy was the same as the ERM models.

The partial RPM used a latent of dimensionality 2. The recognition model from X to U ($f_{\theta_\mathbf{X}}(U|X)$) and the generative model from X and U to C ($P_\phi(C|X,U)$) were MLPs with one hidden layer of 100 units with ReLU activations. $f_{\theta_\mathbf{W}}(U|W)$ and $P_\psi(Y|C,U)$ were simply probability matrices. All parameters were learned using Adam for 2000 epochs with learning rate of 10^{-3} and batch size of 1000. We also multiply the entropy in the loss by a factor of 5.0 which linearly reduces to 1.0 in 400 epochs.

We ran 10 random initializations of each model on the same generated data for every size of observation X.

B. CIFAR-10 Observation Experiment

In the CIFAR-10 experiments, we make observed X and W both CIFAR-10 images. The generative process for these observations is described here.

Here, U, Y, and C are discrete. Let $\tilde{X}$ and $\tilde{W}$ be discrete variables that determine the class from which CIFAR-10 images for X and W are randomly drawn. We let the dimensionalities of each of the variables be $k_U = 3$, $k_C = 3$, $k_Y = 2$, $k_{\tilde{X}} = 2$, and $k_{\tilde{W}} = 3$. $\mathbf{o}(v)$ is the one-hot representation of some variable $v \in V$.

$$
\begin{aligned}
U &\sim \text{Categorical}(\boldsymbol{\pi}) \\
\tilde{W}|U = u &\sim \text{softmax}(\boldsymbol{M}_{\tilde{W}|U}\mathbf{o}(U)) \\
\tilde{X}|U = u &\sim \text{softmax}(\boldsymbol{M}_{\tilde{X}|U}\mathbf{o}(U)) \\
C|\tilde{X} = \tilde{X}, U = u &\sim \text{softmax}(\boldsymbol{M}_{C|\tilde{X}}\mathbf{o}(\tilde{X}) + \boldsymbol{M}_{C|U}\mathbf{o}(U)) \\
Y|C = c, U = u &\sim \text{sigmoid}(\boldsymbol{M}_{Y|C}\mathbf{o}(C) + \boldsymbol{M}_{Y|U}\mathbf{o}(U))
\end{aligned}
$$

Here are each of the generative parameters:

$$
\boldsymbol{M}_{\tilde{W}|U} = \begin{bmatrix} 1e2 & -1e20 & -1e20 \\ -1e20 & 1e2 & -1e20 \\ -1e20 & -1e20 & 1e2 \end{bmatrix}, \boldsymbol{M}_{\tilde{X}|U} = \begin{bmatrix} 1e2 & 1e2 & -1e20 \\ 1e2 & -1e20 & 1e2 \end{bmatrix}, \boldsymbol{M}_{C|U} = \begin{bmatrix} 5.0 & 5.0 & 0.5 \\ 5.0 & 0.5 & 5.0 \\ 0.5 & 5.0 & 5.0 \end{bmatrix}
$$
$$
\boldsymbol{M}_{C|\tilde{X}} = \begin{bmatrix} 5.0 & 0.5 \\ 5.0 & 5.0 \\ 0.5 & 5.0 \end{bmatrix}, \boldsymbol{M}_{Y|U} = \begin{bmatrix} 5.0 & 5.0 & 0.5 \\ 0.5 & 5.0 & 5.0 \end{bmatrix}, \boldsymbol{M}_{Y|C} = \begin{bmatrix} 5.0 & 5.0 & 0.5 \\ 0.5 & 5.0 & 5.0 \end{bmatrix} \tag{8}
$$

For the source distribution, $\boldsymbol{\pi} = \text{softmax}([1.0, 0.1, 0.1])$ while in the target distribution $\boldsymbol{\pi} = \text{softmax}([0.1, 0.1, 1.0])$. The prediction accuracy of Y after latent shift was tested for 10 random initializations for each dataset, the size of which was chosen from $[2 \times 10^3, 5 \times 10^3, 10^4, 2 \times 10^4, 5 \times 10^4, 10^5, 2 \times 10^5, 5 \times 10^5, 10^6]$. The dataset size for both the source P and the target Q was the same.

The ERM source model was a convolutional NN. The first 2 layers were convolutional layers, with 3 input to 10 output channels and 10 input to 20 output channels. Each had kernel size 5 and the output was put through a 2D max pooling then ReLU function before the next layer. Then one linear layer of size 500 to 50 with ReLU activation and then one more linear layer to output. The output were logits on the discrete Y categories. It was trained for 5000 epochs with batch size of 2000 and learning rate of 0.001 using Adam.

The partial RPM used a recognition model for both X ($f_{\theta_\mathbf{x}}(U|X)$) and W ($f_{\theta_\mathbf{w}}(U|W)$) which was the same convolutional NN described for the ERM source model. The recognition model the generative model from X and U to C ($P_\phi(C|X, U)$) was also the same convolutional NN. $P_\psi(Y|C, U)$ was a probability matrix. All parameters were learned using Adam for 20000 epochs with learning rate of 10^{-3} for the recognition parameters, 10^{-4} for the generative parameters, and 10^{-3} for the prior parameters. The batch size was 2000. We also multiply the entropy in the loss by a factor of 5.0 which linearly reduces to 1.0 in 3000 epochs.