Multiple Environments Can Reduce Indeterminacy in VAEs

Quanhan Xi
Department of Statistics
The University of British Columbia
Vancouver, BC, Canada
johnny.xi@stat.ubc.ca

Benjamin Bloem-Reddy
Department of Statistics
The University of British Columbia
Vancouver, BC, Canada
benbr@stat.ubc.ca

Abstract

Parameter and latent variable identifiability in variational autoencoders have received considerable attention recently, due to their empirical success in learning joint probabilities of complex data and their representations. Concurrently, modeling using multiple environments has been suggested for robust causal reasoning. We uncover additional theoretical benefits of multiple environments in the form of a strong identifiability result for a general class of additive noise models with (latent) covariate shift. We propose a novel representation learning algorithm that combines empirical Bayes and variational autoencoders, designed for identifiability of unique representations without compromising representative power, using multiple environments as a crucial technical and practical tool.

1 Introduction

Variational autoencoders (VAE) aim to learn the joint distribution of a latent code X and an observation Y in a Bayesian deep latent variable model, scaling to modern applications using variational inference. Theoretical analysis of Bayesian models typically assume that data are observed as independently and identically distributed (i.i.d.) random variables, and that future observations should also be predicted under this setting. However, it may be the case that data are only i.i.d. when conditioned on additional data, such as the physical environment or viewing angles of images.

Observing data from changing environments and learning invariant predictors has been used as a tool in causal discovery and robust inference [24, 17, 1]. Here, we assume that data may arise from a number of environments, indexed by c, differing by a latent *covariate shift* [22]. Specifically, we assume that the conditional distribution $P(Y|X)$ is invariant to shifts in X, but that the prior on X differs across environments, thus changing the respective joint distributions in each environment. We use covariate shift as a technical tool to obtain an identifiability result for a general class of additive noise models when applied in multiple environments, which we believe can be a valuable tool for causal representation learning [21]. Compared to past work, our contributions uncover a trade-off between model flexibility and identifiability in terms of the number of environments on which distributional assumptions are made. Crucially, we are able to characterize the exact source of non-identifiability in these models. We propose minimal assumptions to obtain strong identifiability of functional parameters in a general setting, and **unique representations** in a representation learning setting.

We examine the identifiability of additive noise models:

Workshop on Causal Inference & Machine Learning (WHY-21 at NeurIPS 2021).

$$
\begin{aligned}
X &\sim p_X \\
\epsilon &\sim g_\epsilon \\
Y &= f(X) + \epsilon \, .
\end{aligned} \tag{1}
$$

This describes a VAE if X is unobserved, i.e., latent, and of much lower dimension. VAEs are the main application in this paper, and so f will be referred to as the decoder. However, our results are directly applicable to the identification of additive noise causal mechanisms, where each environment represents a soft intervention on all parent nodes [18]. Note that this perspective of our work transcends latent variable modeling or VAEs—the X's may be observed, and not necessarily lower dimensional, and environments may represent different experimental settings. In this setting, our work says that if we are able to sample from the "right" interventional SCMs, or experimental settings, then a causal mechanism may be correctly identified.

In general, many decoder functions f can result in the same marginal distribution for Y, and so the model is unidentifiable and thus non-unique latent codes can give rise to the same observed data. The non-identifiability makes it difficult to establish statistical properties such as consistency, and the non-uniqueness stands in the way of interpreting latent dimensions separately and meaningfully (sometimes referred to as disentanglement [12]). A number of works, beginning with [8], have developed extensions of VAE to obtain various identifiability claims [9, 10, 15]. The problem of identifiabilty has also received some attention recently for deep (nonlinear) latent variable models in general [4, 6, 19], and more classically for ICA [3, 5].

Those familiar with factor analysis might recognize the VAE as resembling a non-linear, multi-environment form of factor analysis. Within our notation, a type of factor analysis model (with a single environment) is recovered when g_ϵ and p_X are both zero-mean Normal, and f is a linear transformation represented by a matrix F. A well-known case of non-identifiability in this model arises from the fact that an orthonormal transformation on X leaves its prior invariant [11]. If R is an orthonormal matrix, then setting $X^* = RX$ and $F^* = FR^\top$ results in an identical marginal density, but $F^* \neq F$.

The non-identifiability in a standard VAE follows a similar logic—there are many non-linear transformations that leave the prior invariant, even if it is not zero-mean Normal. Such a non-linearity can be undone in the same way as the factor model by changing f, thus violating identifiability [8]. While factor analysis typically avoids this problem by seeking an optimal rotation (according to a specific A) of its factors (see [13] for a modern VAE example), it is much more difficult to optimize over the general class of p_X-preserving measure automorphisms.

Our main contribution is a novel identifiablility analysis of the model described by Equation 1, formulated in terms of *measure automorphisms*. Specifically, A is said to be a ν-measure automorphism if $\nu \circ A = \nu \circ A^{-1} = \nu$ [2]. The following Theorem motivates the rest of our analysis and model formulation.

Theorem 1.1 (Informal Statement of Theorem A.1.5)**.** *Two injective decoders f_a and f_b resulting in the same marginal distribution for Y in a VAE must be related as*

$$
f_a(x) = f_b(A(x)) \, ,
$$

where A is a measure automorphism of the prior measures on the latent space.

This Theorem formalizes the idea that VAE non-identifiability is not only a result of invariant prior transformations [8], but in fact is completely characterized by such a transformation on the latent space. Furthermore, we see that the level of non-identification is exactly related to the class of prior-preserving measure automorphisms; if its only element is the identity, then Theorem 1.1 says that f is uniquely identified point-wise.

By fixing a subset of priors in multiple environments, we obtain strong identifiability results by constraining A to be of a simple form and applying the above Theorem. As a practical implementation of these ideas, Sections 2 and 3 use an appropriate set of transformed exponential family priors to obtain a simple characterization of A. Our model is otherwise a straightforward application of environment invariance to VAEs, precisely targeting strong latent variable identifiability while retaining the maximum amount of flexibility.

2 A Multiple VAE Model

Let $(\mathbb{R}^d, B(\mathbb{R}^d))$ be the measurable space representing the latent variables (X), and $(\mathbb{R}^m, B(\mathbb{R}^m))$ be the measurable space representing the observations (Y). Though our main application, i.e., representation learning, assumes $d << m$, our analysis is fully general. Let $\mathcal{Y}$ be a subset of $\mathbb{R}^m$, and $\mathcal{F}$ be a family of injective, Borel-measurable functions with common image $\mathcal{Y}$. We will refer to $\mathcal{F}$ as the family of decoder functions. All probability measures on either space will be assumed to have a density with respect to the Lebesgue measure.

2.1 Model Details

Suppose the existence of distinct environments, each of which carries a different marginal distribution for Y under the model. Denote the prior density of X and marginal density of Y under environment c to be $p_X^{(c)}$ and $p_Y^{(c)}$ respectively. We assume that the environment of an observation is observed deterministically (i.e., there is no need to infer it), and so each environment may be analyzed separately of others.

For each environment, we assume that the data arises independently as in a VAE,

$$\begin{aligned} &\text{for } c \in \mathcal{I} \\ &X \sim p_X^{(c)} \\ &\epsilon \sim g_\epsilon \\ &Y = f(X) + \epsilon \ . \end{aligned} \tag{2}$$

Here, $\mathcal{I}$ is an arbitrary index set, which is not necessarily finite. g_ϵ is the noise distribution denoted by its density, and $f \in \mathcal{F}$ is the shared decoder. Note this also implies the conditional density, which is the same across environments,

$$p_f(y|x) = g_\epsilon(y - f(x)) \ . \tag{3}$$

Hence, the differences between environments is driven entirely by the prior distributions, while the decoder and noise terms are shared. To improve the expressiveness of our priors, we will also assume that $p_X^{(c)}$ are partially parametrized by a function, also fixed across all environments. Details of this parametrization are given in Section 2.3, but it is useful to think of a normalizing flow for now [16]. Given an environment c and mild conditions on g_ϵ, the maringal density of Y is given by

$$p_Y^{(c)}(y) = \int g_\epsilon(y - f(x)) dP_X^{(c)} \ . \tag{4}$$

2.2 Model Identifiability

Identifiability of model parameters is often seen as a prerequisite to statistical inference. Roughly, observed data should not be explained equivalently well under different parameter values of the model. If two parameters lead to an identical fit, then the optimization landscape may have multiple modes, and the parameters fit can be difficult to interpret. Furthermore, any notion of strong consistency in the limit of infinite sample size is lost. Formally speaking, let θ_i be a pair of parameter values and denote the marginal density for environment c under the model parametrized by θ_i to be $p_{Y,\theta_i}^{(c)}$. The model is said to be strongly identifiable if

$$p_{Y,\theta_a}^{(c)} = p_{Y,\theta_b}^{(c)} \text{ for all } c \implies \theta_a = \theta_b \ . \tag{5}$$

In practice, we will say that the model is strongly identified if the parametrization is unique for all observed environments—that is, we replace "for all c" by "for all observed c" above. This is a stronger statement since it weakens the hypothesis, i.e., "for all c" $\implies$ "for all observed c". Note that we treat θ_i as functional parameters for identifiability, and $\theta_a = \theta_b$ means that, for example, $f_a = f_b$ almost everywhere on their domain. In practice, they may be parametrized by the weights and biases of some neural networks, or as normalizing flows.

Though technically distinct to parameter identifiability, we also refer to the latent code as being strongly identified if a unique code X can be associated to an observation Y for any given $p_{Y,\theta}$. In

the context of causal representation learning, this means that we may trace our observations back to semantically meaningful causes, and justifies intervening on latent codes. Recent work identifies these codes up to affine transformations [8]. However, even if identified up to just a permutation and translation, each "cause" can still take on infinitely many values for the same observational distribution. It is unclear how interventions, especially hard interventions, should be interpreted under this type of identifiability.

2.3 Fixed Subset of Transformed Exponential Family Priors

We will now describe the key structural assumption required for our model to be strongly identifiable. Suppose a subset of environment priors are from the same exponential family parametrized by the natural parameter η,

$$p_{\text{base}}^{(c)}(x; \eta_c) = m(x) \exp(\eta_c^\top T(x) - a(\eta_c)) \ . \tag{6}$$

Here, T, h and a are fixed across the subset. An example of such a parametrization is simply a Normal with fixed covariance, where the natural parameter is the mean vector. We refer to these as "fixed environments", noting that the remaining prior distributions can be arbitrary.

For a strictly more expressive model while maintaining strong identifiability of latent codes, we propose to transform the fixed priors jointly. To do this, we refer to Equation 6 as the "base" priors. By imposing a single transformation for every fixed environment, we are able to retain the shared structure required for identifiability. Specifically, let $\mathcal{H}$ be a family of smoothly invertible (C^1) functions from $\mathbb{R}^D$ to $\mathbb{R}^D$. The prior for each fixed environment is then transformed by $h \in \mathcal{H}$,

$$p_X^{(c)}(x) = p_{\text{base}}(h^{-1}(x); \eta_c) |\det \frac{\partial h}{\partial x}(x)|^{-1} \ . \tag{7}$$

Note that the priors only differ by their base natural parameter. In other words, each fixed environment has its own distinct starting point, and then proceeds on a shared trajectory. As part of our inference scheme, the priors are optimized by selecting the optimal h that maximizes the total marginal likelihood over all environments.

With this formulation, our probabilistic model is learnable via the functional parameters f and h. The fixed base priors are considered to be hyperparameters—this is analogous to fixing a family of base distributions, such as spherical Gaussians with environment-specific means. Note that the fixed priors are crucial for strong identifiability; we weaken this at the cost of a weaker identifiability result in the next section. Although the variational approximation of the posterior on X (i.e., the decoder q) is also learned simultaneously to f and h, we consider it external to our model and hence do not consider it for identifiability purposes.

We now state our main Theorem, a general result on the identifiability of f, h, and X assuming conditions only on a subset of base prior densities.

Theorem 2.1. *In the model described by Equation 2, let $\theta_a = (f_a, h_a)$ and $\theta_b = (f_b, h_b)$ denote two parametrizations and suppose that g_ϵ has a strictly positive characteristic function. Suppose a subset of $K + 1$ environments follow the prior described by Equations 6-7, and:*

1. *$m(x)$, shared for each base prior, is strictly positive.*
2. *K of the base priors have linearly independent natural parameter vectors, where K is the dimension of the base natural parameters and sufficient statistic.*

Then if $p_{Y,\theta_a}^{(c)} = p_{Y,\theta_b}^{(c)}$ for each observed c, we have $T(h_a^{-1}(f_a^{-1}(y))) = T(h_b^{-1}(f_b^{-1}(y)))$ almost everywhere in $\dagger$. If T is injective in at least one dimension, then $f_a(h_a(x)) = f_b(h_b(x))$ almost everywhere. If h_a and h_b are the identity, i.e., the priors are simply the base exponential families, then the decoder f is strongly identifiable.

If T is designed to be injective in at least one dimension, the result in the original parametrization means that x is a strongly identified base encoding, and $x_1 = h_1(x)$, $x_2 = h_2(x)$ are both acceptable encodings of y. From this perspective, x is recoverable by inverting h_1 or h_2

$$x = h_1^{-1}(x_1) = h_2^{-1}(x_2) \tag{8}$$

This perspective is highly practical. We can obtain a base encoding x which uniquely identifies the data. Meanwhile, we retain the ability to generate accurate samples from the model by passing it through h, despite $h(x)$ not being strongly identified.

2.3.1 Alternative Parametrization

Combined with Equations 6 and 7, our model for the $K+1$ fixed environments can also be written as

$$\begin{aligned} &\text{for } c \in \{c_1, \dots, c_{K+1}\} \\ &\quad X \sim p_{\text{base}}^{(c)} \\ &\quad \epsilon \sim g_\epsilon \\ &\quad Y = f(h(X)) + \epsilon \,, \end{aligned} \tag{9}$$

which is an equivalent model to Equation 2 for these environments, except that the decoder is now $f \circ h$, with h being a C^1 function from the latent space to itself. Note that Theorem 2.1 can be equivalently understood as strong identifiability of the reparametrized $f \circ h$, which also implies strong identification of X.

2.4 Fully Flexible Priors

Previously, we fixed a subset of priors to obtain strong identifiability. In this section, we give results that recover Theorem 1 of [8], with a novel interpretation under our framework. First, we state a generalized version of Theorem 1.1. The complete technical story can be found in Appendix A.2.

Theorem 2.2 (Informal Statement of Theorem A.2.1)**.** *Two injective decoders f_a and f_b with priors ν_a and ν_b resulting in the same marginal distribution for Y in an additive noise model must be related as*

$$f_a(x) = f_b(A(x)) \,,$$

where A transports between the priors ν_a and ν_b, and the process is invertible using A^{-1}.

This Theorem says that, for any decoder-prior pair in a VAE that matches in the marginal distribution of Y, there must exist a uniquely determined, invertible transport map between the two priors. Furthermore, the indeterminacy in the decoder, or equivalently in the latent code, is exactly this mapping. Of course, if $\nu_a = \nu_b$, we recover Theorem 1.1. Importantly, the identity map cannot transport between two distinct priors, and so Theorem 2.2 also implies that strong identifiability of the decoder, or the latent code, is impossible when none of the environment priors are fixed. The minimal constraints to obtain strong identifiability is hence exactly fixing the right priors such that their shared automorphisms can only be the identity function (our main results are a specific instance of this).

We apply Theorem 2.2 to recover the identifiability result in [8]. Suppose that our model (i.e., the marginals of Y) has exponential family priors as in Equation 6 parametrized by $\theta = (f, m, T, (\eta_c)_{c \in \mathcal{I}})$, where f, m, T are functional parameters, and η_c is a vector arbitrarily depending on c (e.g., a neural network mapping environment metadata to η).

Theorem 2.3. *In the model with a subset of $K+1$ environments with priors parametrized by $\theta_a = (f_a, m_a, T_a, \{\eta_c^{(a)}\}_{c \in \mathcal{I}}), \theta_b = (f_b, m_b, T_b, \{\eta_c^{(b)}\}_{c \in \mathcal{I}})$, suppose that g_ϵ has a strictly positive characteristic function and assume the following conditions on the priors:*

1. *$m^{(a)}(x), m^{(b)}(x)$ are strictly positive.*
2. *The same K base priors for θ_a and θ_b have linearly independent natural parameter vectors, where K is the dimension of the base natural parameters and sufficient statistic.*

Then if $P_{Y,\theta_a}^{(c)} = P_{Y,\theta_b}^{(c)}$ for each observed c, we have

$$T_b(A(x)) = L^\top T_a(x) + \mathbf{d} \quad \lambda - a.e. \text{ on } \mathbb{R}^d \,, \tag{10}$$

where $A(x) = f_b^{-1}(f_a(x))$, L is an invertible $K \times K$ matrix and $\mathbf{d}$ is some K-dimensional vector.

Note taking $y = f_a^{-1}(x)$ recovers Theorem 1 of [8]. To help interpret this result, view A as the transformation on the latent space that connects two possible encodings. Then, as perceived by the respective sufficient statistics, this transformation is affine. As an example, consider the case where we fit Gaussian priors with fixed covariance, i.e., we fix $T_a(x) = T_b(x) = x$. Then, the theorem says that $A(x) = L^\top x + \mathbf{d}$, i.e., any two encodings must be related by an invertible affine transformation.

Algorithm 1: Two-Step EB-VAE Optimization.

Input : Data (y_{obs}), hyperparameters, and initializations $f^{(0)}, q^{(0)}, h^{(0)}, \psi_c^{(0)}$, an index set of anchor environments $\dagger_c \subset \mathcal{I}$, with $|\dagger_c| = K + 1$.

Output : Functional parameters f, q, h, ψ_c.

Unsupervised representation learning $c \to \eta_c$ for $c \in \dagger_c$, such that $\text{span}(\{\eta_c\}) = \mathbb{R}^K$. **while** *not converged* **do**

  Sample minibatch from the prior and take an SGD step for h and ψ_c ;

  $h^{(t+1)}, \psi_c^{(t+1)} = \arg\max_{h,\psi_c} \sum_{c=1}^{C} \mathbb{E}_{p_X^{(c)}}[\log p_f(y^{(c)}|x)]$ given $f^{(t)}, q^{(t)}$ as in Eqn. 11 ;

  Sample minibatch from the approximate posterior and take an SGD step for f and q;

  $f^{(t+1)}, q^{(t+1)} = \arg\max_{f \in \mathcal{F}, q \in \mathcal{Q}} \text{ELBO}_{\text{total}}(y_{\text{obs}})$ given $h^{(t+1)}, \psi_c^{(t+1)}$ as in Eqn. 12

end

3 Inference in Multiple Environments

Empirical bayes (EB) describes a family of procedures in which a Bayesian prior distribution is treated as an estimated parameter [14]. The standard optimization objective is to maximize the marginal likelihood of the observed data—an objective that the VAE model also approximately optimizes [27].

We first introduce some additional notation specific to inference. The data are denoted by $y_{\text{obs}} = (Y^{(c)})_{c=1}^{C}$, where $y^{(c)}$ is understood to represent the (possibly many) observations from each environment. Let $c \in \{1, \ldots, C\}$ denote the observed training environments. Then, the total marginal likelihood is

$$\log p_{\text{total}}(y_{\text{obs}}) = \sum_{c=1}^{C} \log p_Y^{(c)}(y^{(c)}) \geq \sum_{c=1}^{C} \mathbb{E}_{p_X^{(c)}}[\log p_f(y^{(c)}|x)] \; . \tag{11}$$

The lower bound on the RHS is hence the main objective to maximize in our model.

3.1 The Total ELBO

The standard ELBO objective for training f in a VAE also approximately maximizes the marginal likelihood. Let $\mathcal{Q}$ be the variational family, and let the encoder or approximate posterior for the given data, under environment c be denoted $q(x|y^{(c)}) \in \mathcal{Q}$. The total ELBO objective is defined to be

$$\text{ELBO}_{\text{total}}(y_{\text{obs}}) = \sum_{c=1}^{C} \mathbb{E}_{q(x|y^{(c)})}[\log p_f(y^{(c)}|x)] - D_{KL}(q(x|y^{(c)}) \| p_X^{(c)}(x)) \; . \tag{12}$$

It is well known that the summands, which are the ELBOs for each environment, are lower bounds to the log-marginal likelihoods $\log p_Y^{(c)}(Y^{(c)})$. Hence, Equation 12 is also a lower bound to Equation 11, and so maximizing the total ELBO objective also approximately maximizes the total marginal likelihood.

3.2 Two-Stage Inference

We propose a generic procedure described by Algorithm 1 to obtain the identifiability proved in Theorem 2.1. As a pre-processing step, we select representative "anchor environments" and embed their metadata into a basis of natural parameters of a fixed exponential family. This amounts to unsupervised learning of an environment representation $c \to \eta_c$ in natural parameter space, while enforcing a basis constraint. Then, we train the decoder f, the diffeomorphism h, and ψ_c, representing an arbitrary functional parametrization of the remaining priors. The resulting latent codes are strongly identified once the anchor embedding is fixed. As this algorithm proceeds, the marginal likelihood is approximately maximized across the two steps.

4 Related Works

Our specific model described by Equations 2, 6 and 7 is most closely related to [8], which recently derived the first identifiable VAE (iVAE) using exponential family priors conditioned on an auxiliary observed variable u, which can represent a time index or class label. In fact, they applied their identifiability results to causal discovery as well, though not under the multiple environments framework. Specifically, their model assumes that each latent dimension follows factorized exponential family distributions parametrized by their sufficient statistics and natural parameters, written as functions of u, also parametrized as neural networks. Their theoretical analysis then establishes the identifiability of all these functional parameters up to some equivalence class.

Our technical contributions generalize the results in [8]. The conditional priors in iVAE can also be understood as covariate shift in multiple environments, where each environment is indexed by u. In this sense, the same factors are driving the identifiability results in both models—in fact, we assume a similar prior diversity condition (linear independence of the natural parameters) and are able to obtain an analogous main result (Theorem 2.3). Our analysis however is more general, does not require differentiability of any components, and does not require factorized or purely exponential family priors. The notion of disentanglement has been defined in various ways in the literature, but our identifiability is able to handle recent notions of "causal disentanglement", which generalizes beyond pure statistical independence [23, 26].

Our main contribution in this paper is to provide an alternative perspective on non-identifiability via Theorems 1.1 and 2.2, and to characterize the minimal required constraints for strong identifiability in an additive noise model. We believe that strong identifiability is of independent interest compared to the equivalence classes derived by [8]. In deep latent variable modeling for example, recent work associates a notion of latent variable non-identifiability to posterior mode collapse in VAEs, which is avoided under our modeling assumptions [25]. In causal representation learning, unique representations and mechanisms can be recovered once the DAG is assumed, which enables robust reasoning about interventions and causal effects on latent variables.

5 Limitations and Future Work

As a general limitation, we also remark that we have not yet empirically evaluated our algorithm. Hence, in this section, we will remark on possible limitations brought by our modeling set-up required for strong identifiability. We will discuss our perceived significance of the possible limitations, and propose ideas to remedy them.

Our set-up makes minimal assumptions, and so we believe the non-identifiability characterizations in Theorems 1.1 and 2.2 apply quite generally in deep latent variable modeling. In view of this, strong identifiability cannot be achieved unless at least some environment priors are fixed prior to training. Of course, this is standard in classical Bayesian inference within the context where both input and output are observed. However, the input space lacks real-world context in representation learning, and so subjective priors are difficult to formulate.

Hence, we believe our set-up is most suitable for cases where we are able to sample from a set of environments that are roughly equidistant, or in which the environment metadata are uninformative. For example, data that are collected from separate experiments or by different experimenters, but where the experimental conditions or individual habits are unknown. These environments can simply be given a canonical basis for a suitable exponential family, and serve as reference points for other environments. If this is not the case, we may also be able to compute similarity indices between environments, and then impose this within the hyperparameter vectors, so that environments have similar prior locations. For now, we leave these design choices for future empirical work.

References

[1] P. Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2018.

[2] V. I. Bogachev. *Isomorphisms of measure spaces*, pages 275–276. Springer, Berlin, 2007.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[4] A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, pages 3765–3773, 2016.

[5] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429439, 1999.

[6] A. Hyvärinen, H. Sasaki, and R. E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22st International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, pages 859–868, 2018.

[7] A. S. Kechris. *Borel Injections and Isomorphisms*, pages 89–93. Springer, New York, 1995.

[8] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, pages 2207–2217, 2020.

[9] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[10] A. Kumar and B. Poole. On implicit regularization in β-vaes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pages 5480–5490, 2020.

[11] D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 12(3):209–229, 1962.

[12] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 4114–4124, 2019.

[13] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 9403–9413, 2019.

[14] J. Maritz and T. Lwin. *Empirical Bayes Methods*. Routledge, 1989.

[15] G. Mita, M. Filippone, and P. Michiardi. An identifiable double VAE for disentangled representations. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pages 7769–7779, 2021.

[16] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[17] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[18] J. Peters, D. Janzig, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

[19] G. Roeder, L. Metz, and D. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pages 9030–9039, 2021.

[20] R. L. Schilling. *The Radon-Nikodým theorem and other applications of martingales*, page 202225. Cambridge University Press, 2005.

[21] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[22] M. Sugiyama and M. Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

[23] R. Suter, D. Miladinović, B. Schölkopf, and S. Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 6056–6065, 2019.

[24] J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI 2001*, pages 512–521, 2001.

[25] Y. Wang, D. Blei, and J. P. Cunningham. Posterior collapse and latent variable non-identifiability. In *Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 2021.

[26] Y. Wang and M. I. Jordan. Desiderata for representation learning: A causal perspective. *arXiv:2109.03795*, 2021.

[27] Y. Wang, A. C. Miller, and D. M. Blei. Comment: Variational Autoencoders as Empirical Bayes. *Statistical Science*, 34(2):229 – 233, 2019.

A Appendix

The appendix contains a mostly self-contained technical report consisting of proofs and intermediate results to the main Theorems of the paper. Notation is inherited from those defined in the main text.

A.1 Theoretical Development

Our theoretical contributions resolves this non-identifiability at its source, by analyzing the transformations such that the latent priors are invariant. Specifically, we show that the indeterminacy in a VAE can be characterized by a measure automorphism—an invertible transformation such that both the forward and inverse directions leave the base prior invariant (Theorem A.1.5). Then, we go on to characterize measure automorphisms, and uncover that under multiple environments, these transformations must be of a simple form (Theorem A.1.10). Finally, these technical results are combined to form our main identifiability result.

We will now provide the logic and proofs of our results. Recall that our model, described by Equations 2, 6 and 7 is parametrized by $\theta = (f, h)$, and that the associated marginal densities are $P^{(c)}_{Y,\theta}$. We will refer to the measure corresponding to the base prior in each environment as $\nu^{(c)}$, noting that the actual prior is then $\nu^{(c)} \circ h^{-1}$. To set the stage, denote a pair of decoders and prior transformations as $\theta_a = (f_a, h_a)$ and $\theta_b = (f_b, h_b)$. Measurable functions and sets are referred to as Borel functions and Borel sets (distinguishing between subsets of $\mathbb{R}^d$ or $\mathbb{R}^m$ when necessary) respectively. Recall that the family of decoder functions $\mathcal{F}$ are injective and Borel, with common image $\mathcal{Y}$. Recall that all inverse functions are defined only with respect to $\mathcal{Y}$. The following lemmas about Borel functions can be very useful.

Lemma A.1.1 ([7], Corollary 15.2)**.** *For f injective and Borel, if B is a Borel set then $f(B)$ is also a Borel set.*

Lemma A.1.2. *For any Borel set B in $\mathbb{R}^d$, if f_a, f_b are Borel then $f_a^{-1}(f_b(B))$ is also a Borel set in $\mathbb{R}^d$.*

Proof. From Lemma A.1.1, $f_b(B)$ is Borel. It then follows from the definition of measurable functions that $f_a^{-1}(f_b(B))$ is Borel. □

We will now state some preliminary definitions. An invertible transformation A on $\mathbb{R}^d$ is said to be a ν-preserving measure automorphism if

$$\nu \circ A = \nu \circ A^{-1} = \nu \tag{1}$$

as measures. If X is a random variable with distribution ν, it is equivalent to say that

$$X \overset{d}{=} A(X) \overset{d}{=} A^{-1}(X) \ . \tag{2}$$

The pushforward σ-algebra of a function $f : \mathbb{R}^d \to \mathbb{R}^m$ on $\mathcal{Y}$ is defined as

$$\sigma(f) = \{C \subset \mathcal{Y}; f^{-1}(C) \text{ is Borel}\} \ . \tag{3}$$

We will assume that all decoder functions also induce the same $\sigma(f)$. Note that it can be easily shown that $\sigma(f)$ is a σ-algebra, and that f is measurable with respect to it. We now prove a useful property for $f \in \mathcal{F}$.

Lemma A.1.3. *Suppose f is Borel and injective, then $\sigma(f)$ contains only Borel sets.*

Proof. Let $C \in \sigma(f)$. By definition, $f^{-1}(C)$ is Borel. Since f is injective, we have $f(f^{-1}(C)) = C$. By Lemma A.1.1, C must then be Borel since f is Borel. □

This lemma is important, as any measures that are equal on $\mathbb{B}(\mathbb{R}^m)$ will also be equal on $\sigma(f)$. From now on, when measures are referred to as being equivalent, we mean that their values agree for every Borel set in the applicable space. We now establish that equivalence of marginal distributions implies the equivalence of image measures that depend only on θ_a and θ_b.

Lemma A.1.4. *Suppose the characteristic function of g_ϵ is everywhere non-zero. Let ν_a, ν_b be the prior probability measures corresponding to h_a and h_b. If $P^{(c)}_{Y,\theta_a} = P^{(c)}_{Y,\theta_b}$, then $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$.*

Proof. Recall that

$$Y = Z + \epsilon \,, \tag{4}$$

with $Z = f(X)$. Let $\psi_\epsilon(t)$ be the characteristic function of g_ϵ and ψ_a, ψ_b be the characteristic functions of Z under θ_a and θ_b. Suppose $P^{(c)}_{Y,\theta_a} = P^{(c)}_{Y,\theta_b}$, then, their characteristic functions coincide and so

$$\begin{aligned} &\psi_a(t)\psi_\epsilon(t) = \psi_b(t)\psi_\epsilon(t) \\ &\implies \psi_a(t) = \psi_b(t) \,, \end{aligned} \tag{5}$$

which implies that $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$. □

The above Lemma reduces the identifiability analysis to the implications of $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$.

This definition sets up our first main Theorem, a version of which with h_a, h_b both the identity is stated in the main text as Theorem 1.1.

Theorem A.1.5. *Suppose that θ_a and θ_b are such that $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$. Then,*

$$f_a(h_a(x)) = f_b(h_b(A(x))) \quad \forall x \in \mathbb{R}^d \,, \tag{6}$$

where $A = h_b^{-1} \circ f_b^{-1} \circ f_a \circ h_a$ is a ν-preserving measure automorphism on $\mathbb{R}^d$.

Proof. To ease notation, denote the functions $f_a \circ h_a = g_a$ and $f_b \circ h_b = g_b$, understanding that these functions remain bijective from $\mathbb{R}^d \to \mathcal{Y}$. We have $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$ as measures, which by definition implies that $\nu \circ g_a^{-1} = \nu \circ g_b^{-1}$. Then, we have for any Borel set,

$$\nu(B) = \nu(g_a^{-1}(g_a(B))) = \nu(g_b^{-1}(g_a(B))) \,. \tag{7}$$

The first equality is due to the injectivity of g_a. The second equality is due to the equality of image measures. The exact same argument with $g_b(h_b(B))$ yields

$$\nu(B) = \nu(g_b^{-1}(g_b(B))) = \nu(g_b^{-1}(g_b(B))) \,. \tag{8}$$

Notice that this defines two new measures $\nu^{(1,2)} = \nu \circ g_a^{-1} \circ g_b$ and $\nu^{(2,1)} = \nu \circ g_b^{-1} \circ g_a$ where g_a, g_b are understood to be set functions representing their respective images. The above equations imply that

$$\nu = \nu^{(1,2)} = \nu^{(2,1)} \tag{9}$$

as measures on $(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$. Now, notice that the image functions g_a, g_b are in fact the preimages (defined on $(\mathcal{Y}, \sigma(f))$) of their inverses g_a^{-1} and g_b^{-1} by injectivity. Hence, $\nu^{(1,2)}$ is in fact also an image measure, this time of $g_b^{-1} \circ g_a : \mathbb{R}^d \to \mathbb{R}^d$. In the exact same way, $\nu^{(2,1)}$ is the image measure of $g_a^{-1} \circ g_b$. The equivalence above hence implies that, viewing X as the random variable with distribution ν, we have

$$X \overset{d}{=} g_b^{-1}(g_a(X)) \overset{d}{=} g_a^{-1}(g_b(X)) \,. \tag{10}$$

Moreover, the functions $g_b^{-1} \circ g_a$ and $g_a^{-1} \circ g_b$ are inverses of each other. Assume the notation $g_b^{-1}(g_a(x)) = A(x)$ and similarly define $A^{-1}(x)$. The above equations hence imply that

$$X \overset{d}{=} A(X) \overset{d}{=} A^{-1}(X) \,. \tag{11}$$

Hence, A is a measure automorphism. Clearly, we have $g_a(x) = g_b(g_b^{-1}(g_a(x))) = g_b(A(x))$. Unrolling g_a and g_b yields $f_a(h_a(x)) = f_b(h_b(A(x)))$. Note that this equality is true everywhere, as it is simply a consequence of our definition of A. □

Note that this result does not use the multiple environments assumption, it is applicable to VAEs more generally. To make use of this result, we now move to analyzing the class of measure automorphisms.

A.1.1 Characterizations of Measure Automorphisms

To make sense of the implications of Theorem A.1.5, we must analyze the class of prior-preserving measure automorphisms. Since the function classes for f and h are so general, the only way to constrain the set of measure automorphisms to a manageable size is to manipulate the prior distribution. In this section, we uncover significant simplifications by requiring that A is an automorphism for multiple exponential family measures. Specifically, suppose that two exponential family base priors are parametrized by its natural parameters, η_a and η_b as in Equation 6.

To prove the next result, we need some intermediate results. Let λ denote the Lebesgue measure. A measure ν is said to be equivalent to λ if $\mu << \lambda$ and $\lambda << \mu$ where $<<$ denotes absolute continuity.

Lemma A.1.6. *Let $h \in \mathcal{H}$ be smoothly invertible on $\mathbb{R}^d$. If ν is equivalent to λ, then $\nu \circ h^{-1}$ is also equivalent to λ.*

Proof. Note that any measure has a strictly positive density almost everywhere if and only if it is equivalent to λ ([20], Problem 19.5). Since h has a smooth inverse, its Jacobian is non-zero. By the change of variables formula, if ν had strictly positive density w.r.t. λ, then so does the image measure $\nu \circ h^{-1}$, and hence it is also equivalent to λ. □

Lemma A.1.7. *Suppose that ν admits a strictly positive density p with respect to the Lebesgue measure λ. Suppose A is a ν-preserving measure automorphism. Then,*

$$p(A(x))k_A(x) = p(x) \qquad \lambda - a.e.\ , \tag{12}$$

where k only depends on A and is strictly positive.

Proof. Since A is a ν-preserving measure automorphism and ν is equivalent to λ, we have that for a Borel set B,

$$\lambda(B) = 0 \iff \nu(B) = 0 \iff \nu(A(B)) = 0 \iff \lambda(A(B))\ , \tag{13}$$

where the second implication is because A is a measure automorphism, and the first and third implications are because ν is equivalent to λ. Hence, $\lambda \circ A$ is equivalent to λ, and so it also has a strictly positive Radon-Nikodym derivative k_A. Then, by the Radon-Nikodym Theorem, we have for a Borel set B,

$$\nu(B) = \int_B \nu(dx) = \int_B p(x)\lambda(dx) = \nu(A(B)) = \int_{A(B)} \nu(dx) = \int_B p(A(x))\lambda(A(dx))\ , \tag{14}$$

by the standard change of variables formula for Lebesgue integration. Now, We have

$$\int_B p(x)\lambda(dx) = \int_B p(A(x))k_A(x)\lambda(dx)\ . \tag{15}$$

Since B was arbitrary, we have

$$p(A(x))k_A(x) = p(x) \qquad \lambda - a.e.\ , \tag{16}$$

where $k_A(x) > 0$ λ-a.e.. □

Corollary A.1.8. *For two strictly positive prior densities p_a and p_b such that A is a measure automorphism, we have*

$$\frac{p_a}{p_b}(x) = \frac{p_a}{p_b}(A(x)) \qquad \lambda - a.e.\ . \tag{17}$$

Proof. This follows immediately from the fact that k_A is strictly positive. □

Lemma A.1.9. *Suppose A is an automorphism for two exponential family densities parametrized by η_a, η_b. Suppose that $m(x)$ is strictly positive. Then, we have*

$$(\eta_a - \eta_b)^\top T(x) = (\eta_a - \eta_b)^\top T(A(x)) \qquad \lambda - a.e.\ . \tag{18}$$

Proof. The expression is a direct consequence of Corollary 17 when plugging in the exponential family densities $p(x;\eta_a)$ and $p(x;\eta_b)$. Taking logs on both sides, we have

$$\begin{aligned}\eta_a^\top T(x) - \eta_b^\top T(x) - A(\eta_a) + A(\eta_b) &= \eta_a^\top T(A(x)) - \eta_b^\top T(A(x)) - A(\eta_a) + A(\eta_b) \\ (\eta_a - \eta_b)^\top T(x) &= (\eta_a - \eta_b)^\top T(A(x)) \ . \end{aligned} \tag{19}$$

□

We can immediately apply this to obtain a powerful characterization of automorphisms of a carefully constructed set of base prior distributions.

Proposition A.1.10. *Suppose A is an automorphism for $K+1$ exponential family measures parametrized by η_c, $c \in \mathcal{I}$, where*

1. *$m(x)$, shared for each measure, is strictly positive.*
2. *K of the natural parameter vectors are linearly independent, where K is the dimension of the base natural parameters and sufficient statistic.*

Then, $T(A(x)) = T(x)$ almost everywhere, where T is the K-dimensional sufficient statistic. If T is injective, then A is the identity almost everywhere.

Proof. For any pair (i,j) such that $i \neq j$, Equation 18 holds. Fix j to be the natural parameter that is not linearly independent. Then, $\{\eta_i - \eta_j\}_{i\neq j}$ forms a basis. With some algebra, we may deduce that

$$\begin{aligned} & T(x)^\top(\eta_i - \eta_j) = T(A(x))^\top(\eta_i - \eta_j) \\ \implies & (T(x) - T(A(x)))^\top(\eta_i - \eta_j) = 0 \quad (i,j) \in P \ . \end{aligned} \tag{20}$$

Since $\{\eta_i - \eta_j\}_{(i,j)\in P}$ forms a basis, the representation of $(T(x) - T(A(x)))$ with respect to this basis and Equation 20 immediately implies $\|T(x) - T(A(x))\|^2 = 0$. Hence, $T(x) = T(A(x))$ almost everywhere. If T is injective, it follows immediately that $A(x) = x$ almost everywhere. □

In words, this result says that, if A is an automorphism for a diverse enough group of exponential family measures, then it is perceived as the identity function under the shared sufficient statistic. Furthermore, the Normal distribution has an injective sufficient statistic, and so the Theorem states that only the identity can preserve such a basis of Normal distributions.

A.1.2 Identifiability Analysis

Combining Theorems A.1.5 and A.1.10, we are now able to prove Theorem 2.1.

Proof of Theorem 2.1. Let the base prior for each environment be denoted $\nu^{(c)}$, and the corresponding priors by applying h_a and h_b by $\nu_a^{(c)}, \nu_b^{(c)}$. By Lemma A.1.4, it is sufficient to study the implication of

$$\nu_a^{(c)} \circ f_a^{-1} = \nu_b^{(c)} \circ f_b^{-1} \ .$$

By Theorem A.1.5, we have

$$f_a(h_a(x)) = f_b(h_b(A(x))) \ ,$$

where A is a $\nu^{(c)}$-preserving measure automorphism on $\mathbb{R}^d$ for each c. In particular, it is an automorphism for the subset of environments for which $\{\eta_i - \eta_j\}_{i\neq j}$ forms a basis of $\mathbb{R}^K$. Recall that A is uniquely determined by f_a, f_b, h_a, h_b, and so it is the same across environments. Now, from the identity $A = h_b^{-1} \circ f_b^{-1} \circ f_a \circ h_a$, we have $h_b^{-1}(f_b^{-1}(y)) = A(h_a^{-1}(f_a^{-1}(y)))$ for any $y \in \mathcal{Y}$. For almost every $h_a^{-1}(f_a^{-1}(y))$ and $h_b^{-1}(f_b^{-1}(y))$, we have, from Theorem A.1.10,

$$\begin{aligned} T(h_b^{-1}(f_b^{-1}(y))) &= T(A(h_a^{-1}(f_a^{-1}(y)))) \\ &= T(h_a^{-1}(f_a^{-1}(y))) \ . \end{aligned} \tag{21}$$

If T is injective in at least one dimension, then $h_a^{-1}(f_a^{-1}(y)) = h_b^{-1}(f_b^{-1}(y))$. Since f_a, f_b, h_a, h_b are all injective, this implies $f_a(h_a(x)) = f_b(h_b(x))$ almost everywhere. This concludes the proof. □

A.2 Alternative Result for Varying Priors, and Connection to iVAE

In this section we present a modified theory that allows for varying (i.e., concurrently optimized) prior distributions. Our results here essentially recover Theorem 1 in [8]. Consider the following modification of Theorem A.1.5, where $\theta_i = (\nu_i, f_i)$ and ν_i may be arbitrarily different.

Theorem A.2.1. *Suppose that θ_a and θ_b are such that $\nu_a \circ f_a^{-1} = \nu_b \circ f_b^{-1}$. Then,*

$$f_a(x) = f_b(A(x)) \quad \forall x \in \mathbb{R}^d \ , \tag{22}$$

where $A = f_b^{-1} \circ f_a$ is a Borel bijection on $\mathbb{R}^d$ such that $\nu_a \circ A^{-1} = \nu_b$ and $\nu_a = \nu_b \circ A$.

Proof. First note that by Lemma A.1.2, A is clearly Borel. By the injectivity of f_a and f_b, A is a bijection. Working on the space $(\mathcal{Y}, \sigma(f))$, f_b^{-1} is defined and f_b is understood to be its preimage. Then for any Borel set B in $\mathbb{R}^d$, we have

$$\nu_a(A^{-1}(B)) = \nu_a(f_a^{-1}(f_b(B))) = \nu_b(f_b^{-1}(f_b(B))) = \nu_b(B) \ , \tag{23}$$

where the last equality is due to the injectivity of f_b. This shows that $\nu_a \circ A^{-1} = \nu_b$, to show the other direction is simply to swap the roles of the indices 1 and 2. □

We will call such an A an invertible transport map. This Theorem essentially states that if two pushforwards are equal in the data space, then there exists a Borel automorphism in the latent space such that one arbitrary prior measure may be transported to another on $(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$, and that the process is invertible. Now, we place some structure on these priors to similarly restate Lemma A.1.7.

Lemma A.2.2. *Suppose that ν_a and ν_b admit strictly positive densities p_a and p_b with respect to the Lebesgue measure λ. Suppose A is an invertible transport map. Then,*

$$p_b(A(x))k_A(x) = p_a(x) \qquad \lambda - a.e. \ , \tag{24}$$

where k only depends on A and is strictly positive.

Proof. Since ν_a, ν_b are equivalent to λ, we have that for a Borel set B,

$$\lambda(B) = 0 \iff \nu_a(B) = 0 \iff \nu_b(A(B)) = 0 \iff \lambda(A(B)) \ , \tag{25}$$

where the second implication is due to the hypothesis that A is an invertible transport map. Hence, $\lambda \circ A$ is equivalent to λ, and so it also has a strictly positive Radon-Nikodym derivative k_A. Then, by the Radon-Nikodym Theorem, we have for a Borel set B,

$$\nu_a(B) = \int_B \nu_a(dx) = \int_B p_a(x)\lambda(dx) = \nu_b(A(B)) = \int_{A(B)} \nu_b(dx) = \int_B p_b(A(x))\lambda(A(dx)) \ , \tag{26}$$

by the standard change of variables formula for Lebesgue integration. Now, We have

$$\int_B p_a(x)\lambda(dx) = \int_B p_b(A(x))k_A(x)\lambda(dx) \ . \tag{27}$$

Since B was arbitrary, we have

$$p_b(A(x))k_A(x) = p_a(x) \qquad \lambda - a.e. \ , \tag{28}$$

where $k_A(x) > 0$ λ-a.e.. □

We also obtain a similar corollary:

Corollary A.2.3. *For two pairs of strictly positive densities (p_1^a, p_2^a) and (p_1^b, p_2^b), such that A is an invertible transport map simultaneously for p_1^a to p_2^a and p_1^b to p_2^b, we have*

$$\frac{p_1^{(a)}}{p_2^{(a)}}(x) = \frac{p_1^{(b)}}{p_2^{(b)}}(A(x)) \qquad \lambda - a.e. \ . \tag{29}$$

If we assume exponential densities, we obtain the following version of Proposition A.1.10.

Proposition A.2.4. *Suppose A is an invertible transport map simultaneously between two sets of $K+1$ exponential family measures with strictly positive densities $(\nu_0^a, \dots, \nu_k^a)$ to $(\nu_0^b, \dots, \nu_k^b)$, where*

1. *$m^{(a)}(x), m^{(b)}(x)$ is fixed for each a, b and are strictly positive.*
2. *The same K natural parameters are linearly independent for both a and b, where K is the dimension of the base natural parameters and sufficient statistic.*

Then,

$$T_b(A(x)) = L^\top T_a(x) + \mathbf{d} \quad \lambda - a.e.\ , \tag{30}$$

where L is an invertible matrix and T_a, T_b are the sufficient statistics (shared for all $K+1$ measures).

Proof. Fix j to be the natural parameter index that is not linearly independent. From Corollary A.2.3 and taking logarithms, we have

$$\begin{aligned}&(\eta_i^{(a)})^\top T_a(x) - a(\eta_i^{(a)}) - ((\eta_j^{(a)})^\top T_a(x) - a(\eta_j^{(a)})) \\ &= (\eta_i^{(b)})^\top T_b(A(x)) - a(\eta_i^{(b)}) - ((\eta_j^{(b)})^\top T_b(A(x)) - a(\eta_j^{(a)}))\ ,\end{aligned} \tag{31}$$

which simplifies to

$$(\eta_i^{(a)} - \eta_j^{(a)})^\top T_a(x) - c_i^{(a)} = (\eta_i^{(b)} - \eta_j^{(b)})^\top T_b(A(x)) - c_i^{(b)} \quad (i,j) \in P \quad \lambda - a.e.\ . \tag{32}$$

Written in matrix form, we have

$$\begin{bmatrix} \eta_{i1}^{(a)} - \eta_{j1}^{(a)} \\ \vdots \\ \eta_{iK}^{(a)} - \eta_{jK}^{(a)} \end{bmatrix}^\top T_a(x) = \begin{bmatrix} \eta_{i1}^{(b)} - \eta_{j1}^{(b)} \\ \vdots \\ \eta_{iK}^{(b)} - \eta_{jK}^{(b)} \end{bmatrix}^\top T_b(A(x)) + \mathbf{c} \quad \lambda - a.e.\ , \tag{33}$$

where $\mathbf{c}$ is the vector of differences $c_i^{(a)} - c_i^{(b)}$. Following [8], we call these two matrices L_a and L_b, noting that they are invertible since their rows are linearly independent by assumption. Then, we obtain

$$L_a^\top T_a(x) - L_b^\top T_b(A(x)) = \mathbf{c} \tag{34}$$

$$\implies T_b(A(x)) = (L_b^{-1} L_a)^\top T_a(x) - (L_b^{-1} L_a)^\top \mathbf{c} \tag{35}$$

$$\implies T_b(A(x)) = L^\top T_a(x) + \mathbf{d} \quad \lambda - a.e\ , \tag{36}$$

where $L = L_b^{-1} L_a$ is invertible and $\mathbf{d} = -L^\top \mathbf{c}$. □

A.2.1 Identifiability

Re-stating and proving Theorem 2.1, our main identifiability result, is a simple application of the original proof technique. Let us describe the modified model for clarity. Recall that we have data arising from C environments. For each environment, we assume that the data arises independently:

$$\begin{aligned}&\text{for } c \in \{\mathcal{I}\} \\ &X \sim p_X^{(c)} \\ &\epsilon \sim g_\epsilon \\ &Y = f(X) + \epsilon\ ,\end{aligned} \tag{37}$$

where g_ϵ is the fixed noise distribution denoted by its density, and $f \in \mathcal{F}$ is the shared decoder. We further assume that the priors are each estimated by an exponential family distribution, differing between environments by its natural parameter,

$$p_X^{(c)}(x) = m(x) \exp(\eta_c^\top T(x) - a(\eta_c))\ . \tag{38}$$

Note that a is determined by m, η_c, and T. Our model (i.e., the marginals of Y) is hence parametrized by $\theta = (f, m, T, (\eta_c)_{c=1}^C)$. We refer to the marginal distributions of Y as $P_{Y,\theta}^{(c)}$.

Theorem A.2.5 (Restating Theorem 2.3)**.** *In the model with a subset of $K+1$ environments with priors parametrized by $\theta_a = (f_a, m_a, T_a, \{\eta_c^{(a)}\}_{c\in\mathcal{I}}), \theta_b = (f_b, m_b, T_b, \{\eta_c^{(b)}\}_{c\in\mathcal{I}})$, suppose that g_ϵ has a non-zero characteristic function and assume the following conditions on the priors:*

1. *$m^{(a)}(x), m^{(b)}(x)$ are strictly positive.*
2. *The same K base priors for θ_a and θ_b have linearly independent natural parameter vectors, where K is the dimension of the base natural parameters and sufficient statistic.*

Then if $P^{(c)}_{Y,\theta_a} = P^{(c)}_{Y,\theta_b}$ for each observed c, we have

$$T_b(A(x)) = L^\top T_a(x) + \mathbf{d} \quad \lambda - a.e. \text{ on } \mathbb{R}^d \,, \tag{39}$$

where $A(x) = f_b^{-1}(f_a(x))$, L is an invertible $K \times K$ matrix and $\mathbf{d}$ is some K-dimensional vector.

The proof of this Theorem is a direct consequence of Theorem A.2.1 and Proposition A.2.4, and is identical to the proof of Theorem 2.1.