ISA-VAE: INDEPENDENT SUBSPACE ANALYSIS WITH VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We provide a natural approach to encourage interpretable representations in variational autoencoders. Building on previous work on Independent Component Analysis (ICA) and Independent Subspace Analysis (ISA) we show that simply choosing an appropriate prior for the latent distribution breaks the degeneracy of the standard prior and leads to a interpretable laten representations. The proposed family of prior distributions leads to a factorized representation and its rotational asymmetry allows interpretability of individual latent dimensions. In our experiments we show that the commonly used higher weight (e.g. β-VAE) of the Kullback-Leibler divergence term in the evidence lower bound (ELBO) amplifies two biases of variational inference: over-pruning and over-orthogonalization. Extensive quantitative experiments demonstrate the performance of the proposed approach for disentanglement in variational autoencoders.

1 INTRODUCTION

Recently there has been an increased interest in unsupervised learning of disentangled representations. The term *disentangled* usually describes two main objectives: First, a factorized latent representation, and second, interpretability of the latent factors. Schmidhuber (1992); Ridgeway (2016); Achille & Soatto (2017) Most of this recent work is inspired by the β-VAE concept introduced in Higgins et al. (2016), which proposes to re-weight different parts of the evidence lower bound (ELBO) objective. In Higgins et al. (2016) a higher weight for the Kullback-Leibler divergence (KL) term is proposed, and putative mechanistic explanations for the effects of this modification are studied in Burgess et al. (2017); Chen et al. (2018). An alternative decomposition of the ELBO leads to the recent variant of β-VAE called β-TCVAE Chen et al. (2018), which shows the highest scores on recent disentanglement benchmarks.

These modifications of the evidence lower bound however lead to a tradeoff between disentanglement and reconstruction loss and therefore the quality of the learned model. This tradeoff is direcly encoded in the modified objective: by increasing the β-weight of the KL-term, the relative weight of the reconstruction loss term is more and more decreased. Therefore, optimization of the modified ELBO will lead to latent encodings which have a lower KL-divergence from the prior, but at the same time lead to a higher reconstruction loss. Furthermore, we discuss in section 2.4 that using a higher weight for the KL-term amplifies existing biases of variational inference.

There is a foundational contradiction in many approaches to disentangling deep generative models (DGMs): the standard model employed is not identifiable as it employs a standard normal prior which then undergoes a linear transformation. Any rotation of the latent space can be absorbed into the linear transform and is therefore statistically indistinguishable. If interpretability is desired, the modelling choices are setting us up to fail.

We make the following contributions:

- We show that the current state of the art approaches employ a tradeoff between reconstruction loss and disentanglement of the latent representation.
- In section 2.3 we show that variational inference techniques are biased towards estimating an over-orthogonal components and towards over-pruning the variance of the latent distribution.

- We provide a novel description of the origin of disentanglement in β-VAE and demonstrate in section 2.4 that increasing the weight of the KL term increases the over-pruning bias of variational inference.
- To mitigate these drawbacks of existing approaches, we propose a family of rotationally asymmetric distributions for the latent prior, which prevents the ambiguity of the latent encoding for arbitrary rotations of the latent space. This approach resembles independent component analysis (ICA) for variational autoencoders.
- We propose to use a prior which allows a decomposition of the latent space using independent subspace analysis (ISA) and demonstrate that this prior leads to disentangled representations even for the unmodified ELBO objective. This removes the tradeoff between disentanglement and reconstruction loss of existing approaches.
- An even higher disentanglement of the latent space can be achieved by incorporating the proposed prior distribution into the existing approaches β-VAE and β-TCVAE, thus exploiting their orthogonalization bias. Because the prior distribution already favours a disentangled representation, a high disentanglement score with at the same time low reconstruction loss can be achieved.

2 BACKGROUND

We briefly discuss previous work on variational inference in deep generative models and two modifications of the learning objective that have been proposed to learn a disentangled representation. We discuss characteristic biases of variational inference and how the modifications of the learning objective actually accentuate these biases.

2.1 DISENTANGLED REPRESENTATION LEARNING

Variational Autoencoder Kingma & Welling (2014) introduce a latent variable model that combines a generative model, the decoder, with an inference network, the encoder. Training is performed by optimizing the *evidence lower bound* (ELBO) averaged over the empirical distribution:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\right] - D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\,, \tag{1}$$

where the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and encoder $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ are deep learning models with parameters θ and each $\boldsymbol{z}^l$ is sampled from $\boldsymbol{z}^l \sim q_\theta(\boldsymbol{z}|\boldsymbol{x})$. When choosing appropriate families of distributions, gradients through the samples $\boldsymbol{z}^l$ can be estimated using the *reparameterization trick*. The approximate posterior $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ is usually modelled as a multivariate Gaussian with diagonal covariance matrix and the prior $p(\boldsymbol{z})$ is typically the standard normal distribution.

β-VAE Higgins et al. (2016) propose to modify the evidence lower bound objective and penalize the KL-divergence of the ELBO:

$$\mathcal{L}_{\beta\text{-ELBO}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\right] - \beta D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\,, \tag{2}$$

where $\beta > 1$ is a free parameter that should encourage a disentangled representation. In Burgess et al. (2017) the authors provide further thoughts on the mechanism that leads to these disentangled representations. However we will show in the following that this parameter introduces a trade-off between reconstruction loss and disentanglement. Furthermore, we show in section 2.4 that this parameter amplifies biases of variational inference towards orthogonalization and pruning.

β-TCVAE In Chen et al. (2018) the authors propose an alternative decomposition of the ELBO, tjat leads to the recent variant of β-VAE called β-TCVAE. They demonstrate that β-TCVAE allows to learn representations with higher MIG score than β-VAEHiggins et al. (2016), InfoGAN Chen et al. (2016) and FactorVAE Kim & Mnih (2018). The authors propose to decompose the KL-term in the ELBO objective into three parts and to weight them independently:

$$\mathbb{E}_{q(\boldsymbol{x})}\left[D_{\text{KL}}(q(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\right] = D_{\text{KL}}(q(\boldsymbol{z}|\boldsymbol{x})\|q(\boldsymbol{z})p(\boldsymbol{x})) + D_{\text{KL}}(q(\boldsymbol{z})\|\prod_j q(\boldsymbol{z}_j)) + D_{\text{KL}}(q(\boldsymbol{z}_j)\|p(\boldsymbol{z}_j))\,. \tag{3}$$

The first term is the index-code mutual information, the second term is the total correlation and the third term the dimension-wise KL-divergence. Because the index-code mutual information can be

viewed as an estimator for the mutual information between $p(\boldsymbol{x})$ and $q(\boldsymbol{z})$, the authors propose to exclude this term when reweighting the KL-term with the β weight. In addition to the improved objective, the authors propose a quantitative evaluation score for disentanglement, the mutual information gap (MIG). They propose to first estimate the mutual information between a latent factor and an underlying generative factor of the dataset. The mutual information gap is then defined as the difference of the mutual information between the highest and second highest correlated underlying factor.

2.2 Related Work

Recent work has shown an increased interest into learning of interpretable representations. In addition to the related work mentioned already, we briefly review some of the influential papers: Chen et al. (2016) present a variant of a GAN that encourages an interpretable latent representation by maximizing the mutual information between the observation and a small subset of latent variables. The approach relies on optimizing a lower bound of the intractable mutual information. Kim & Mnih (2018) propose a learning objective equivalent to β-TCVAE, and train it with the density ration trick Sugiyama et al. (2012). Kumar et al. (2017) introduce a regularizer of the KL-divergence between the approximate posterior and the prior distribution. A parallel line of research proposes not to train a perfect generative model but instead to find a simpler representation of the data (Vedantam et al. (2017); Hinton et al. (2011b)). A similar strategy is followed in semi-supervised approaches that require implicit or explicit knowledge about the true underlying factors of the data Kulkarni et al. (2015); Kingma et al. (2014); Reed et al. (2014); Baydin et al. (2017); Hinton et al. (2011a); Zhu et al. (2017); Goroshin et al. (2015); Hsu et al. (2017); Denton et al. (2017)).

2.3 Orthogonalization and Pruning in Variational Inference

There have been several interpretations of the behaviour of the β-VAE (Chen et al., 2018; Burgess et al., 2017). Here we provide a complementary perspective: that it enhances well known statistical biases in VI (Turner & Sahani, 2011) to produce disentangled, but not necessarily useful, representations. The form of these biases can be understood by considering the variational objective when written as an explicit lower-bound: the log-likelihood of the parameters minus the KL divergence between the approximate posterior and the true posterior

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\right] - D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \tag{4}$$

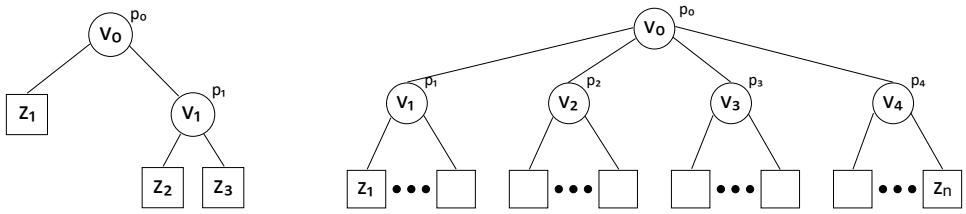
From this form it is clear that VI's estimates of the parameters θ will be biased away from the maximum likelihood solution (the maximizer of the first term) in a direction that reduces the KL between the approximate and true posteriors. When factorized approximating distributions are used, VI will therefore be biased towards settings of the parameters that reduce the dependence of the latent variables in the posterior. For example, this will bias learned components towards orthogonal directions in the output space as this reduces explaining away (e.g. in the factor analysis model, VI breaks the degeneracy of the maximum-likelihood solution finding the orthogonal PCA directions, see appendix). Moreover, these biases often cause components to be pruned out (in the sense that they have no effect on the observed variables) since then their posterior sits at the prior, which is typically factorized (e.g. in an over-complete factor analysis model VI prunes out components to return a complete model, see appendix). For simple linear models these effects are not pathological: indeed VI is arguably selecting from amongst the degenerate maximum likelihood solutions in a sensible way. However, for more complex models the biases are more severe: often the true posterior of the underlying model has significant dependencies (e.g. due to explaining away) and the biases can prevent the discovery of some components. For example, VAEs are known to over-prune (Burda et al., 2015; Cremer et al., 2018).

2.4 β-VAE Emphasizes Orthogonalization and Pruning

What happens to these biases in the β-VAE generalization when $\beta > 1$?

$$\mathcal{L}_{\beta\text{-ELBO}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\right] - \beta D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \tag{5}$$

The short answer is that they grow. This can be understood by considering coordinate ascent of the modified objective. With θ fixed, optimising q finds a solution that is closer to the prior distribution



(a) Tree corresponding to eq. 10

(b) Tree visualization of a L^p-nested ISA model.

than VI due to the upweighting of the KL term in 4. With q fixed, optimization over θ returns the same solution as VI (since the prior does not depend on the parameters θ and so the value of β is irrelevant). However, since q is now closer to the prior than before, the KL bias in 5 will be greater. These effects are shown in the ICA example below [or in the appendix]. VI ($beta = 1$) learns components that are more orthogonal than the underlying ones, but $\beta = 5$ prunes out one component entirely and sets the other two to be orthogonal. This is disentangled, but arguably leads to incorrect interpretation of the data. This happens even though both methods are initialised at the true model. Arguably, the β-VAE is enhancing a statistical bug in VI and leveraging this as a feature. We believe that this can be dangerous, preventing the discovery of the underlying model.

3 Latent Prior Distributions for Unsupervised Factorization

In this section we describe an approach for unsupervised learning of disentangled representations. Instead of modifying the ELBO-objective, we propose to use certain families of prior distributions $p(\boldsymbol{z})$, that lead to identifiable and interpretable models. In contrast to the standard normal distribution, the proposed priors are not rotationally invariant, and therefore allow interpretability of the latent space.

3.1 Independent Component Analysis

Independent Component Analysis (ICA) seeks to factorize a distribution into non-Gaussian factors. In order avoid the ambiguities of latent space rotations, a non-Gaussian distribution (e.g. Laplace or Student-t distribution) is used as prior for the latent variables.

Generalized Gaussian Distribution A generalized version of ICALee & Lewicki (2000); Zhang et al. (2004); Lewicki (2002); Sinz & Bethge (2010) uses a prior from the family of *exponential power distributions* of the form

$$p(\boldsymbol{z}) \propto \exp\left(-\tau ||\boldsymbol{z}||_p^p\right) \tag{6}$$

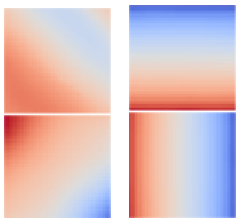
also called *generalized Gaussian*, *generalized Laplacian* or p*-generalized normal* distribution. Using $p = 2/(1+\kappa)$ the parameter κ is a measure of kurtosis.Box & Tiao (1973) This family of distributions generalizes the normal ($\kappa = 0$) and the Laplacian ($\kappa = 1$) distribution. In general we get for $\kappa > 0$ *leptokurtic* and for $\kappa < 0$ *platykurtic* distributions.

Obviously the normal distribution is a special instance of the class of L^p-spherically symmetric distributions, and the normal distribution is the only L^2-spherically symmetric distribution with independent marginals. Equivalently Sinz et al. (2009a) showed that this also generalizes to arbitrary values of p. The marginals of the p-generalized normal distribution are independent, and it is the only factorial model in the class of L^p-spherically symmetric distributions.

We investigate the behaviour of L^p-spherically symmetric distributions as prior distributions for $p(\boldsymbol{z})$ in the experiments in section 4.

3.2 Independent Subspace Analysis

ICA can be generalized to include indepdence between subspaces, but dependencies within them, by using a more general prior, the family of L^p-nested symmetric distributions. Hyvärinen & Hoyer (2000); Hyvärinen & Köster (2007); Sinz et al. (2009b); Sinz & Bethge (2010)



(a) $p = 1.0$ (b) $p = 2.4$

Figure 2: Leptokurtic and platykurtic priors encourage different orientations of the encoding of the X and Y coordinate in the dSprites dataset. A leptokurtic distribution (here the Laplace distribution) expects most of the probability mass around 0. Because the X and Y coordinates in dSprites are distributed in a square, the projection of the positions of the samples onto the diagonal fit better to the Laplace prior. A platykurtic distribution however fits better to the uniform distribution of x and y along the axes.

To start, let's consider functions of the form

$$f(\boldsymbol{z}) = \left(|z_1|^{p_0} + (|z_2|^{p_1} + |z_3|^{p_1})^{\frac{p_0}{p_1}}\right)^{\frac{1}{p_0}} , \tag{7}$$

with $p_0, p_1 \in \mathbb{R}$. This function is a cascade of two L^p-norms. To give a better intuition we provide a visualization of this distribution in figure 1a, which depicts equation 10 as a tree that visualizes the nested structure of the norms. We call the class of functions which employ this structure *L^p-nested*.

L^p-nested Distribution Given a L^p-nested function f and a radial density $\phi : \mathbb{R} \mapsto \mathbb{R}^+$ we define the *L^p-nested symmetric distribution* following Fernandez et al. (1995) as

$$\rho(\boldsymbol{z}) = \frac{\phi(f(\boldsymbol{x}))}{f(\boldsymbol{x})^{n-1}\mathcal{S}_f(1)} , \tag{8}$$

where $\mathcal{S}_f(1)$ is the surface area of the L^p-nested sphere. This surface area can be obtained by using the gamma function:

$$\mathcal{S}_f(R) = R^{n-1}2^n \prod_{i \in I} \frac{\prod_{k=1}^{l_1} \Gamma\left[\frac{n_{i,k}}{p_i}\right]}{p_i^{l_i-1}\Gamma\left[\frac{n_i}{p_i}\right]} , \tag{9}$$

where l_i is the number of children of a node i, n_i is the number of leaves in a subtree under the node i, and $n_{i,k}$ is the number of leaves in the subtree of the k-th children of node i. For further details we refer the reader to the excellent work of Sinz & Bethge (2010).

Independent Subspace Analysis The family of L^p-nested distibutions allow a generalization of ICA called independent subspace analysis (ISA). ISA uses a subclass of L^p-distributions, which are defined by functions of the form

$$f(\boldsymbol{z}) = \left(\left(\sum_{i=1}^{n_1} |z_i|^{p_1}\right)^{\frac{p_0}{p_1}} + \cdots + \left(\sum_{i=n_1+\cdots+n_{l-1}+1}^{n} |z_i|^{p_l}\right)^{\frac{p_0}{p_l}}\right)^{\frac{1}{p_0}} , \tag{10}$$

and correspond to trees of depth two. The tree structure of this subclass of functions is visualizaed in figure 1b. Sinz & Bethge (2010) showed that the subspaces $v_1, \ldots, v_{l_0}$ become independent when using the radial distribution

$$\phi_0(v_0) = \frac{p_0 v_0^{n-1}}{\Gamma\left[\frac{n}{p_0}\right] s^{\frac{n}{p_0}}} \exp\left(-\frac{v_0^{p_0}}{s}\right) \tag{11}$$

which we can interpret as a generalization of the Chi-distribution.

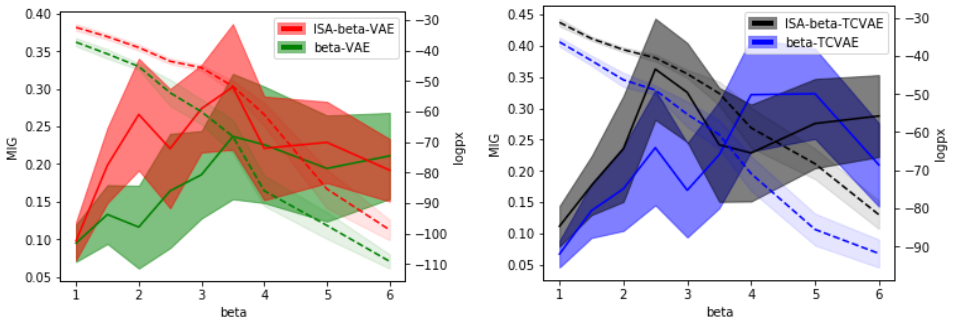


Figure 3: Comparison of the proposed ISA-VAE model with β-VAE and β-TCVAE. Evaluation is performed with respect to the quality of the disentanglement (MIG score, solid line) and reconstruction quality (dashed line) for different values of β. The proposed models ISA-VAE and ISA-TCVAE reach a disentangled representation already for small values of β which allows better reconstructions due to the trade-off between the β-weight of the KL-term and the reconstruction loss in the modified ELBO. Shaded regions depict 90% confidence intervals. Evaluated on the dSprites dataset with $n = 16$ for each value of β. L^p-layout: $p_0 = 2.1$, $p_1 = 2.2$, $l_0 = 5$, $l_1 = 4$.

We propose to choose the latent prior $p(\boldsymbol{z})$ from this family of distributions, which allows us to define independent subspaces in the latent space. Experiments in the following section demonstrate that priors of this form support unsupervised learning of disentangled representation even for the unmodified ELBO objective ($\beta = 1$).

4 Experiments

In our experiments, we evaluate the influence of the proposed prior distribution on disentanglement and on the quality of the reconstruction. To provide a quantitative evaluation of the disentanglement we compute the disentanglement metric *Mutual Information Gap* (MIG) that was proposed in Chen et al. (2018). In this publication the authors demonstrate that β-TCVAE, a modification of β-VAE, enables learning of representations with higher MIG score than β-VAEHiggins et al. (2016), InfoGAN Chen et al. (2016) and FactorVAE Kim & Mnih (2018). Therefore we choose β-TCVAE and β-VAE for comparison in our experiments.

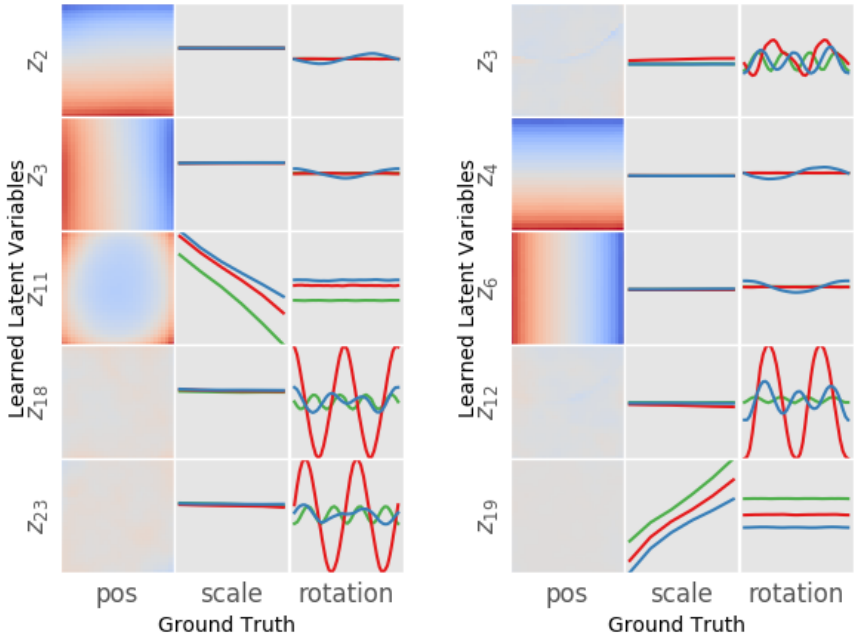
4.1 Support of the Prior to Learn Disentangled Representations

First, we investigate the ability of the prior to support unsupervised learning of disentangled representations for the unmodified ELBO-objective. We compare two DGMs after training with the unmodified ELBO-objective: the standard VAE with standard normal prior $p(\boldsymbol{z})$, and ISA-VAE, the combination of the standard VAE with a prior $p(\boldsymbol{z})$ from the family of L^p-nested ISA distributions.

The combination of the L^p-nested ISA prior with the standard VAE approach learns reasonably disentangled representations even for the unmodified ELBO objective($\beta = 1$). Figure 4a depicts the structure of the latent representation for ISA-VAE after training. Even for the unmodified objective the latent space becomes interpretable, which we did not observe in our experiments for the standard VAE with standard normal prior.

4.2 Quantitative Comparisons of the Different Approaches

The proposed L^p-nested prior distribution can be easily combined with the existing approaches for unsupervised disentangled representation learning β-VAE Higgins et al. (2016) and β-TCVAE Chen et al. (2018). We compare the performance of the four different approaches in figure 5: β-VAE and ISA-β-VAE, a combination of β-VAE with the proposed L^p-nested ISA prior, and equivalently β-TCVAE and ISA-β-TCVAE. Clearly both approaches benefit from the proposed prior, and higher disentanglement scores can be achieved with ISA-β-VAE and ISA-β-TCVAE in comparison to the original approaches.



(a) $\beta = 1$, MIG: 0.29

(b) $\beta = 2.5$, MIG: 0.54

Figure 4: Disentangled representations with high MIG-scores that we observed in our experiments for ISA-VAE and β-ISA-TCVAE on the dSprites dataset. The ISA-VAE model is able to find a disentangled representation even without a modification of the ELBO ($\beta = 1.0$). When combining the ISA-model with β-TCVAE, a model with a high disentanglement score of MIG $= 0.54$ can be reached. This is the so far highest score reported for dSprites in the literature. a) ISA-VAE with $p_0 = 2.1$, $p_1 = \frac{4}{3}$, $l_0 = l_1 = 5$, MIG-score: 0.29 b) β-ISA-TCVAE for $\beta = 2.5$ and $p_0 = 2.1$, $p_1 = 2.2$, $l_0 = 5$, $l_1 = 4$, MIG-score: 0.54

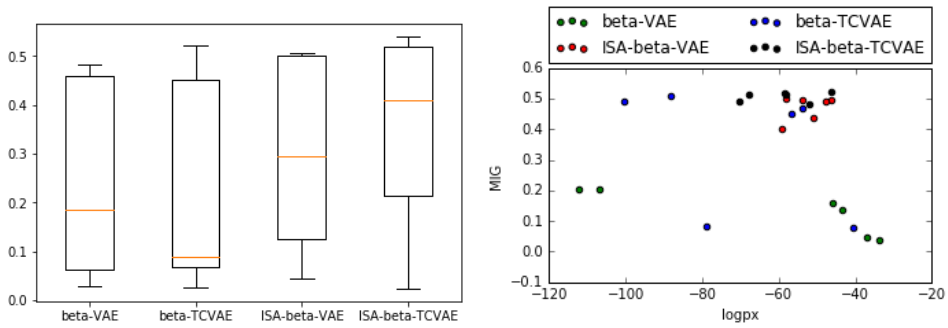
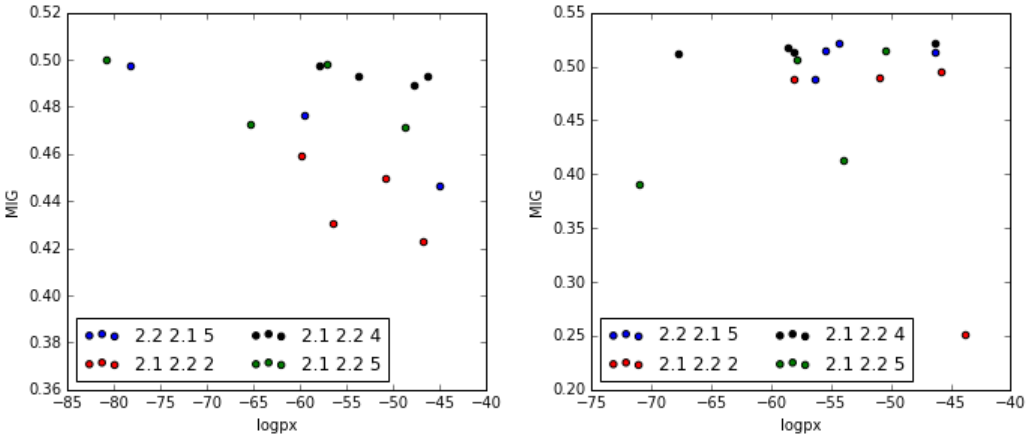


Figure 5: Left: Comparison of the MIG score of the proposed ISA-VAE model with β-VAE and β-TCVAE for the individually best values of β. Right: The proposed models ISA-VAE and ISA-TCVAE reach a disentangled representation already for small values of β which allows better reconstructions due to the trade-off between the β-weight of the KL-term and the reconstruction loss in the modified ELBO. L^p-layout of the ISA model: $p_0 = 2.1$, $p_1 = 2.2$, $l_0 = 5$, $l_1 = 4$.



(a) ISA-VAE

(b) ISA-TCVAE

Figure 6: Disentanglement score (MIG) with respect to the reconstruction quality (logpx) for different layouts for the ISA subspace and different values of $\beta \in [1, 4]$. The number of subspaces $l_0 = 5$ is constant throughout all experiments. The parameters p_0, p_1 and l_0 are varied and for each configuration the four highest MIG scores for different values of β are shown. For this dataset, the confguration $p_0 = 2.1$, $p_1 = 2.2$ and $l_0 = 4$ (denoted in black) is most appropriate as it achieves high MIG scores while maintaining a good reconstruction quality, both for the ISA-TCVAE and the ISA-VAE model.

4.3 Trade-off between Disentanglement and Reconstruction Loss

Since the prior facilitates learning of disentangled representations, not only a higher disentanglement score can be reached, but also higher scores are reached for smaller values of β, when compared to the original approaches. This leads to a clear improvement of the trade-off between disentanglement and reconstruction loss. The improvement of this trade-off is demonstrated in figure 5, where we plot both the disentanglement score and the reconstruction loss for varying values of β. ISA-β-VAE and ISA-β-TCVAE reach high values of the disentanglement score for smaller values of β which and at the same time preserves a higher quality of the reconstruction than the respective original approaches. At the same time we observe that with the proposed prior the quality of the reconstruction decreases at a smaller rate than for the original approaches.

5 Conclusion

We presented a structured prior for unsupervised learning of disentangled representations in deep generative models. We choose the prior from the family of *L^p-nested symmetric distributions* which enables defining of a hierarchy of independent subspaces in the latent space. In contrast to the standard normal prior that is often used in training of deep generative models the proposed prior is not rotationally invariant and therefore enhances the interpretability of the latent space. We demonstrate in our experiments, that a combination of the proposed prior with existing approaches for unsupervised learning of disentangled representations allows a significant improvement of the trade-off between disentanglement and reconstruction loss.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–153, 2017.

George EP Box and George C Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading,Mass., 1973.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. In *Learning Disentangled Representations: From Perception to Control Workshop*, 2017.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4414–4423, 2017.

Carmen Fernandez, Jacek Osiewalski, and Mark FJ Steel. Modeling and inference with υ-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340, 1995.

Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pp. 4086–4093, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pp. 44–51. Springer, 2011a.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pp. 44–51. Springer, 2011b.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pp. 1878–1889, 2017.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

Aapo Hyvärinen and Urs Köster. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100, 2007.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL `https://openreview.net/forum?id=33X9fd2-9FyZd`.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

Tejas D Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2539–2547. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5851-deep-convolutional-inverse-graphics-network.pdf.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Te-won Lee and Michael Lewicki. The generalized gaussian mixture model using ica. 08 2000.

Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356, 2002.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pp. 1431–1439, 2014.

Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Fabian Sinz and Matthias Bethge. Lp-nested symmetric distributions. *Journal of Machine Learning Research*, 11(Dec):3409–3451, 2010.

Fabian Sinz, Sebastian Gerwinn, and Matthias Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820, 2009a.

Fabian H Sinz, Eero P Simoncelli, and Matthias Bethge. Hierarchical modeling of local image features through L^p-nested symmetric distributions. In *Advances in neural information processing systems*, pp. 1696–1704, 2009b.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, T. Cemgil, and S. Chiappa (eds.), *Bayesian Time series models*, chapter 5, pp. 109–130. Cambridge University Press, 2011.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

Liqing Zhang, Andrzej Cichocki, and Shun-ichi Amari. Self-adaptive blind source separation based on activation functions adaptation. *IEEE Transactions on Neural Networks*, 15(2):233–244, 2004.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

Supplementary Material

1 Toy examples showing biases in VI and β-VI

This section provides the details of the toy examples that reveal the biases in variational methods.
First we will consider the factor analysis model showing that VI breaks the degeneracy of the maximum-likelihood solution to 1) discover orthogonal weights that lie in the PCA directions, 2) prune out extra components in over-complete factor analysis models, even though there are solutions with the same likelihood that preserve all components. We also show that in these examples the β-VI returns identical model fits to VI regardless of the setting of β.
Second, we consider an over-complete ICA model and initialize using the true model. We show that 1) VI is biased away from the true component directions towards more orthogonal directions, and 2) β-VI with a modest setting of $\beta = 5$ prunes away one of the components and finds orthogonal directions for the other two. That is, it finds a disentangled representation, but one which does not reflect the underlying components.

1.1 Background

The β-VAE optimizes the modified free-energy, $\mathcal{F}_\beta(q(z_{1:N}), \theta)$, with respect to the parameters θ and the variational approximation $q(z_{1:N})$,

$$\mathcal{F}_\beta(q(z_{1:N}), \theta) = \mathbb{E}_{q(z_{1:N})}(\log p(x_{1:N}|z_{1:N}, \theta)) - \beta \mathrm{KL}(q(z_{1:N})||p(z_{1:N})).$$

Consider the case where $M = \frac{1}{\beta}$ is a positive integer, $M \in \mathbb{N}$, we then have

$$\mathcal{F}_\beta(q(z_{1:N}), \theta) = \sum_{n=1}^{N} \left[\mathbb{E}_{q(z_n)}(M(\beta) \log p(x_n|z_n, \theta)) - \mathrm{KL}(q(z_n)||p(z_n))\right]$$

In this case, the β-VAE can be thought of as attaching M replicated observations to each latent variable z_n and then running standard variational inference on the new replicated dataset. This can equivalently be thought of as raising each likelihood $p(x_n|z_n, \theta)$ to the power M.
Now in real applications β will be set to a value that is greater than one. In this case, the effect of β is the opposite: it is to reduce the number of effective data points per latent variable to be less than one $M < 1$. Or equivalently we

raise each likelihood term to a power M that is less than one. Standard VI is then run on these modified data (e.g. via joint optimization of q and θ).
Although this view is mathematically straightforward, the perspective of the β-VAE i) modifying the dataset, and ii) applying standard VI, is useful as it will allow us to derive optimal solutions for the variational distribution $q(z)$ in simple cases like the factor analysis model considered next.

1.2 Factor analysis

Consider the factor analysis generative model. Let $\mathbf{x} \in \mathbb{R}^L$ and $\mathbf{z} \in \mathbb{R}^K$.

$$\begin{aligned} &\text{for } \ n = 1...N \\ &\quad \mathbf{z}_n \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathrm{I}), \\ &\quad \mathbf{x}_n \sim \mathcal{N}(\mathbf{x}; W\mathbf{z}, D) \ \text{ where } \ D = [\sigma_1^2, ..., \sigma_{\mathrm{D}}^2] \end{aligned} \tag{1}$$

The true posterior is a Gaussian $p(\mathbf{z}_n|\mathbf{x}_n, \theta) = \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$ where

$$\mu_{\mathbf{z}|\mathbf{x}} = \Sigma_{\mathbf{z}|\mathbf{x}} W^\top D^{-1}\mathbf{x} \ \text{ and } \ \Sigma_{\mathbf{z}|\mathbf{x}} = (W^\top D^{-1} W + \mathrm{I})^{-1}. \tag{2}$$

The true log-likelihood of the parameters is

$$\begin{aligned} \log p(\mathbf{x}_{1:N}|\theta) &= \sum_{n=1}^{N} \log \mathcal{N}(\mathbf{x}_n, \mathbf{0}, WW^\top + D) \\ &= -\frac{N}{2} \log\det(2\pi(WW^\top + D)) - \frac{1}{2}\sum_{n=1}^{N} \mathbf{x}_n^\top (WW^\top + D)^{-1}\mathbf{x}_n \\ &= -\frac{1}{2}N\left[\log\det(2\pi(WW^\top + D)) + \mathrm{trace}((WW^\top + D)^{-1}(\mu_{\mathbf{x}}\mu_{\mathbf{x}}^\top + \Sigma_{\mathbf{x}}))\right] \end{aligned}$$

Here we have defined the empirical mean and covariance of the observations $\mu_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$ and $\Sigma_{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathbf{x}})(\mathbf{x}_n - \mu_{\mathbf{x}})^\top$ i.e. the sufficient statistics.
The true likelihood is invariant under orthogonal transformations of the latent variables: $\mathbf{z}' = R\mathbf{z}$ where $RR^\top = \mathrm{I}$.
Interpreting β-VI as running VI in a modified generative model (see previous section) we have the new generative process

$$\begin{aligned} &\text{for } \ n = 1...N \\ &\quad \mathbf{z}_n \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathrm{I}), \\ &\quad \text{for } \ m = 1...M(\beta) \\ &\qquad \mathbf{x}_{n,m} \sim \mathcal{N}(\mathbf{x}; W\mathbf{z}, D) \ \text{ where } \ D = [\sigma_1^2, ..., \sigma_{\mathrm{D}}^2] \end{aligned}$$

We now observe data and set $\mathbf{x}_{n,m} = \mathbf{x}_n$.
The posterior is again Gaussian $p(\mathbf{z}_n|\mathbf{x}_n, \theta, M(\beta)) = \mathcal{N}(\mathbf{z}; \tilde{\mu}_{\mathbf{z}|\mathbf{x}}(\beta), \tilde{\Sigma}_{\mathbf{z}|\mathbf{x}}(\beta))$ where

$$\tilde{\mu}_{\mathbf{z}|\mathbf{x}}(\beta) = \tilde{\Sigma}_{\mathbf{z}|\mathbf{x}}^{-1}(\beta) M(\beta) W^\top D^{-1}\mathbf{x} \ \text{ and } \ \tilde{\Sigma}_{\mathbf{z}|\mathbf{x}}(\beta) = (M(\beta)W^\top D^{-1}W + \mathrm{I})^{-1}$$

Here we have taken care to explicitly reveal all of the direct dependencies on β. Mean-field variational inference, $q(\mathbf{z}_n) = \prod_k q_{n,k}(z_{k,d})$, will return a diagonal Gaussian approximation to the true posterior with the same mean and matching diagonal precision,

$$q(\mathbf{z}_n|\mathbf{x}_n, \theta, M(\beta)) = \mathcal{N}\left(\mathbf{z}; \tilde{\mu}_{\mathbf{z}|\mathbf{x}}(\beta), \Sigma_q(\beta)\right) \quad \text{where} \quad \Sigma_q^{-1}(\beta) = \text{diag}\left(\tilde{\Sigma}_{\mathbf{z}|\mathbf{x}}^{-1}(\beta)\right)$$

We notice that the posterior mean is a linear combination of the observations $\tilde{\mu}_{\mathbf{z}|\mathbf{x}}(\beta) = R(\beta)\mathbf{x}$ where $R(\beta) = \tilde{\Sigma}_{\mathbf{z}|\mathbf{x}}(\beta)M(\beta)W^\top D^{-1}$ are recognition weights. Notice that the recognition weights and the posterior variances are the same for all data points: they do not depend on n. The free-energy is then

$$\mathcal{F}(q, \theta, \beta) = \mathbb{E}_{q(z)}(\log p(x|z)) - \text{KL}(q(z)|p(z))$$

with the reconstruction term being

$$\begin{aligned}
\mathbb{E}_{q(z)}(\log p(x|z)) &= -\frac{1}{2\beta}\sum_{n=1}^{N} \mathbf{x}_n^\top (D^{-1} - 2R^\top W^\top D^{-1} + R^\top W^\top D^{-1} W R)\mathbf{x}_n \\
&\quad - \frac{N}{2\beta}\log\det(2\pi D) - \frac{N}{2\beta}\text{trace}(W^\top D^{-1}\Sigma_q) \\
&= -\frac{N}{2\beta}\Big(\text{trace}\left((D^{-1} - 2R^\top W^\top D^{-1} + R^\top W^\top D^{-1} W R)(\Sigma_{\mathbf{x}} + \mu_{\mathbf{x}}\mu_{\mathbf{x}}^\top)\right) \\
&\quad + \log\det(2\pi D) + \text{trace}(W^\top D^{-1} W \Sigma_q)\Big) \qquad (3)
\end{aligned}$$

and the KL or regularization term being

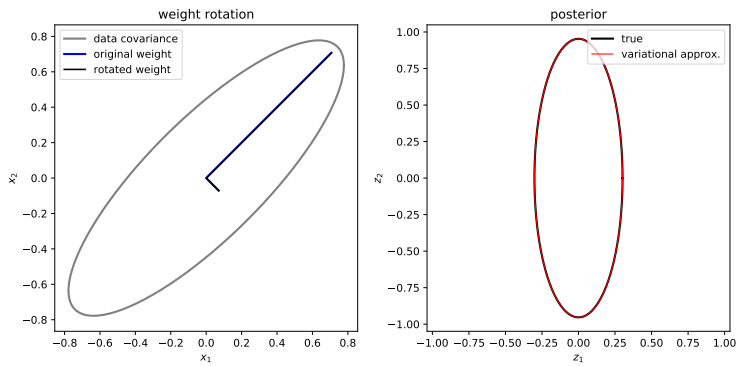
$$\begin{aligned}
\text{KL}(q(z)|p(z)) &= -\frac{NK}{2} - \frac{N}{2}\log\det(\Sigma_q) + \frac{N}{2}\text{trace}(\Sigma_q) + \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n^\top R^\top R \mathbf{x}_n \\
&= -\frac{N}{2}\left(K + \log\det(\Sigma_q) - \text{trace}(\Sigma_q) - \text{trace}(R^\top R(\Sigma_{\mathbf{x}} + \mu_{\mathbf{x}}\mu_{\mathbf{x}}^\top))\right).
\end{aligned}$$

We will now consider the objective functions and the posterior distributions in several cases to reason about the parameter estimates arising from the methods above.

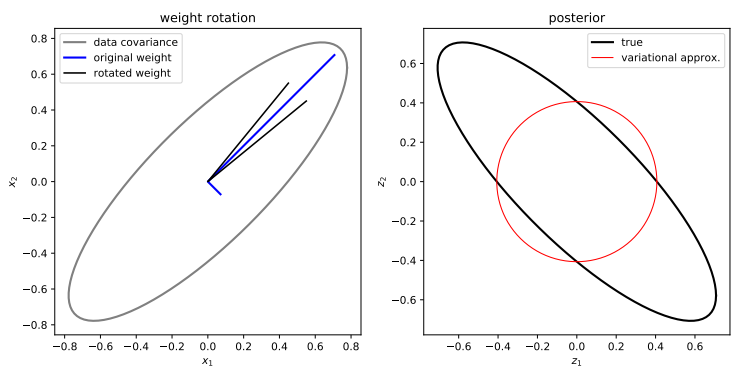
1.3 Experiment 1: mean field VI applied to factor analysis yields the PCA directions

Consider the situation where we know a maximum likelihood solution of the weights W_{ML}. For simplicity we select the solution W_{ML} which has orthogonal weights in the observation space. We then rotate this solution by an amount θ so that $W'_{\text{ML}} = R(\theta)W_{\text{ML}}$. The resulting weights are no longer orthogonal (assuming the rotation is not an integer multiple of $\pi/2$). We compute the log-likelihood (which will not change) and the free-energy (which will change) and

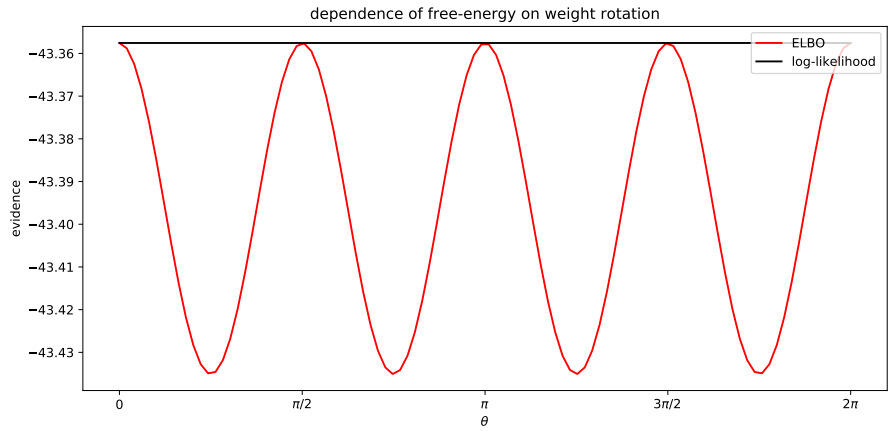
plot the true and approximate posterior covariance (which does not depend on the datapoint value x_n).
First here are the weights are aligned with the true ones. The log-likelihood and the free-energy is -17.82 nats.



Second, here are the weights rotated $\pi/4$ and the log-likelihood is -17.82 nats and the free-energy -57.16 nats.



When varying θ the plots above indicate that orthogonal settings of the weights ($\theta = m\pi/2$ where $m = 0, 1, 2, ...$) lead to factorized posteriors. In these cases the KL between the approximate posterior and the true posterior is zero and the free-energy is equal to the log-likelihood. This will be the optimal free-energy for any weight setting (due to the fact that it is equal to the true log-likelihood which is maximal, and the free-energy is a lower bound of this quantity.) For intermediate values of θ the posterior is correlated and the free-energy is not tight to the log likelihood.
Now let's plot the free-energy and the log-likelihood as θ is varied. This shows that the free-energy prefers orthogonal settings of the weights as this leads to factorized posteriors, even though the log-likelihood is insensitive to θ. So, variational inference recovers the same weight directions as the PCA solution.



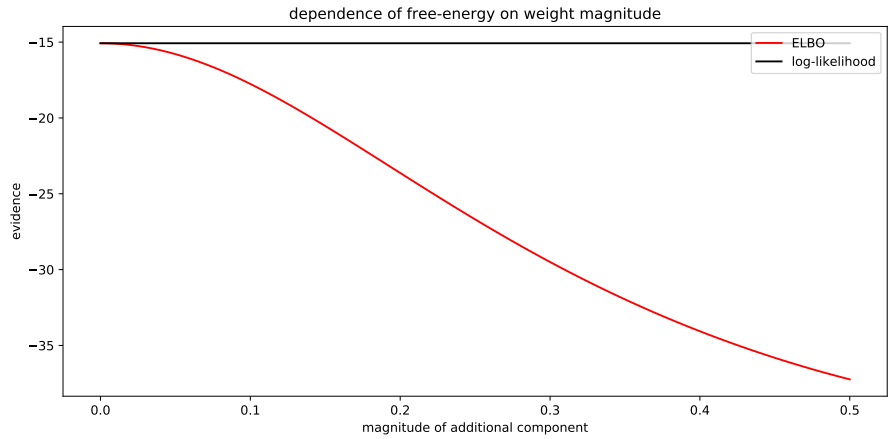
The above shows that the bias inherent in variational methods will cause them to break the symmetry in the log-likelihood and find orthorgonal latent components. This occurs because orthoginal components result in posterior distributions that are factorized. These are then well-modelled by the variational approximation and result in a small KL between the approximate and true posteriors.

1.4 Experiment 2: mean field VI applied to over-complete factor analysis prunes out the additional latent dimensions

A similar effect occurs if we model 2D data with a 3D latent space. Many settings of the weights attain the maximum of the likelihood, including solutions which use all three latent variables. However, the optimal solution for VI is to retain two orthogonal components and to set the magnitude of the third component to zero. This solution a) has weights maximise the likelihood, and b) has a factorised posterior distribution (the pruned component having a posterior equal to its prior) that therefore incurs no cost $\mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x},\theta)) = 0$. In this way the bound becomes tight.
Here's an example of this effect. We consider a model of the form:

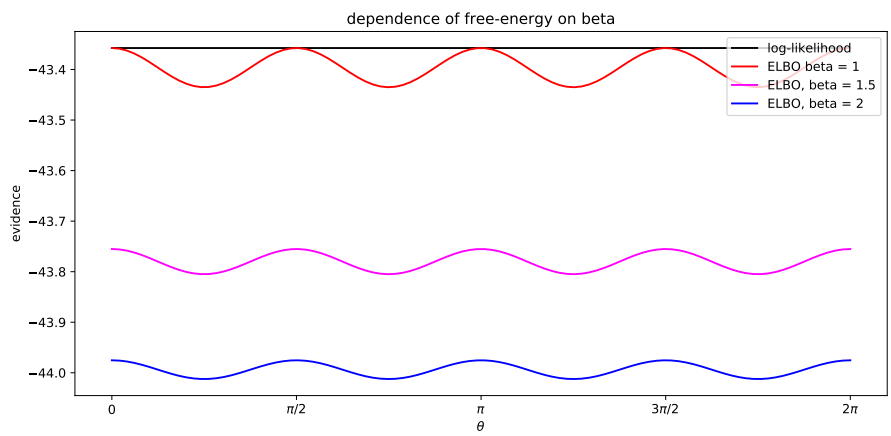
$$\mathbf{x} = \frac{\alpha}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} z_1 + \frac{\beta}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} z_2 + \frac{\rho}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} z_3 + \epsilon \tag{4}$$

We set $\alpha^2 + \beta^2 = 1$ so that all models imply the same covariance and set this to be the maximum likelihood covariance by construction. We then consider varying α from 0 to 1/2. The setting equal to 0 attains the maximum of the free-energy, even though it has the same likelihood as any other setting.



1.5 Experiment 3: The β-VAE also yields the PCA components, changing β has no effect on the direction of the estimated components in the FA model

How does the setting of β change things? Here we rerun experiment 1 for different values of β.



In this example, changing β in this example just reduces the amplitude of the fluctuations in the free-energy, but it does not change the directions found. A similar observation applies to the pruning experiment.
Increasing β will increase the uncertainty in the posterior as it is like reducing the number of observations (or increasing the observation noise, from the perspective of q).

1.6 Summary of Factor Analysis Experiments

The behaviours introduced by the β-VAE appear relatively benign, and perhaps even helpful, in the linear case: VI is breaking the degeneracy of the maximum likelihood solution in a sensible way: selecting amongst the maximum likelihood solutions to find those that have orthogonal components and removing spurious latent dimensions. This should be tempered by the fact that the β generalization recovered precisely the same solutions and so it was necessary to obtain the desired behaviour in the PCA case.

Similar effects will occur in deep generative models, not least since these typically also have a Gaussian prior over latent variables, and these latents are initially linearly transformed, thereby resulting in a similar degeneracy to factor analysis.

However, the behaviours above benefited from the fact that maximum-likelihood solutions could be found in which the posterior distribution over latent variables factorized. In real world examples, for example in deep generative models, this will not be case. In such cases, these same effects will cause the variational free-energy and its β-generalization to **bias the estimated parameters far away from maximum-likelihood settings, toward those settings that imply factorized Gaussian posteriors over the latent variables**.

1.7 Independent Component Analysis

We now apply VI and the β free-energy method to ICA. We're interested the properties of the variational objective and the β-VI objective and so we 1. fit the data using the true generative model to investigate the biases in VI and β-VI 2. do not use amortized inference, just optimizing the approximating distributions for each data point (this is possible for these small examples)

The linear independent component analysis generative model we use is defined as follows. Let $\mathbf{x} \in \mathbb{R}^L$ and $\mathbf{z} \in \mathbb{R}^K$.

$$
\begin{aligned}
&\text{for } \; n = 1...N \\
&\quad \text{for } \; k = 1...K \\
&\qquad z_{n,k} \sim \text{Student-t}(0, \sigma, v), \\
&\quad \mathbf{x}_n \sim \mathcal{N}(W\mathbf{z}, D) \;\; \text{where} \;\; D = [\sigma_1^2, ..., \sigma_{\mathrm{D}}^2]
\end{aligned}
$$

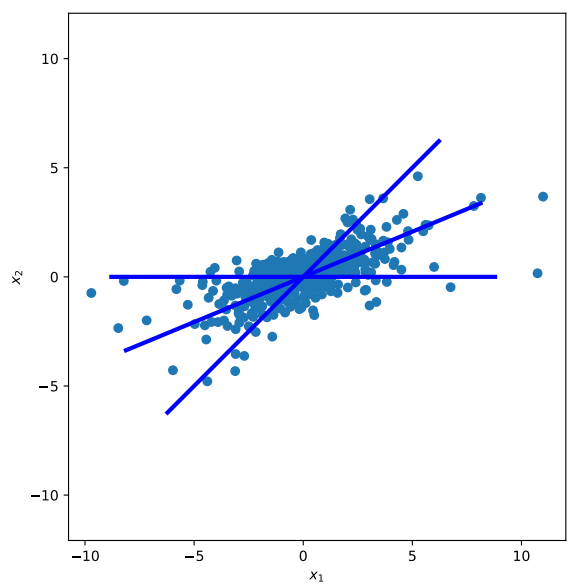
We apply mean-field variational inference, $q(\mathbf{z}_n) = \prod_k q_{n,k}(z_{k,d})$, and use Gaussian distributions for each factor $q_{n,k}(z_{n,k}) = \mathcal{N}(z_{n,k}; \mu_{n,k}, \sigma_{n,k}^2)$.

The free-energy is computed as follows: The reconstruction term is identical to PCA: an avergage of a quadratic form wrt to a Gaussian, which is analytic. The KL is broken down into the differential entropy of q which is also analytic and the cross-entropy with the prior which we evaluate by numerical integration (finite differences). There is a cross-entropy term for each latent variable which is one reason why the code is slow (requiring N 1D numerical integrations). The

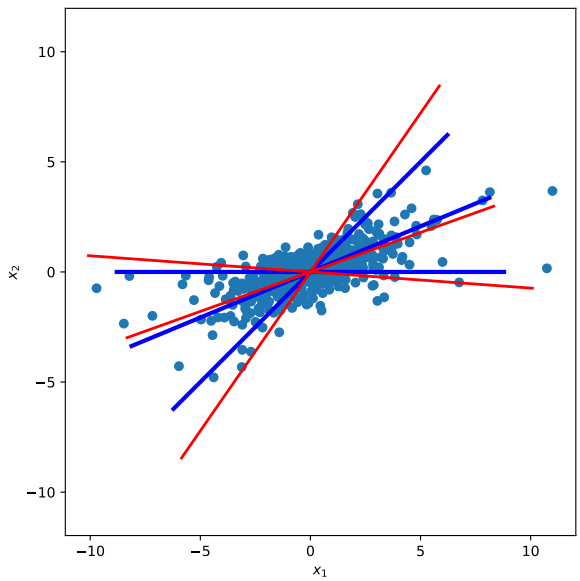
gradient of the free-energy wrt the parameters W and the means and variances of the Gaussian q distributions are computed using autograd.
In order to be as certain as possible that we are finding a global maximum of the free-energies, all experiments initialise at the true value of the parameters and then ensure that each gradient step improves the free-energy. Stochastic optimization or a procedure that accepted all steps regardless of the change in the objective would be faster, but they might also move us into the basis of attraction of a worse (local) optima.

1.8 Experiment 1: Learning in over-complete ICA

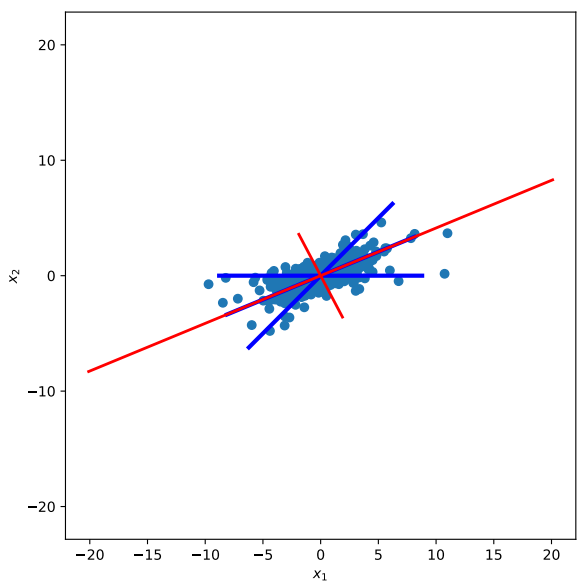
Now we define the dataset. We use a very sparse Student's t-distribtion with $v = 3.5$. For $v < 4$ the the kurtosis is undefined so the model is fairly simple to estimate model (it's a long way away from the degenerate factor analysis case which is recovered in the limit $v \to \infty$).
We use three latent components and a two dimensional observed space. The directions of the three weights are shown in blue below with data as blue circles.



First we run variational inference finding components (shown in red below) which are more orthogonal than the true directions. This bias is in this directions as this reduces the dependencies (explaining away) in the underlying posterior.



Second we run β-VI with $\beta = 5$. Two components are now found that are orthogonal with one component pruned from the solution.



In this case the bias is so great that the true component directions are not discovered. Instead the components are forced into the orthogonal setting regardless of the structure in the data.
Summary of ICA experiment

The ICA example illustrates that this approach – of relying on a bias inherent in VI to discover meaningful components – will sometimes return meaningful structure (e.g. in the PCA experiments above). However it does not seem to be a sensible way of doing so in general. For example, explaining away often means that the true components will be entangled in the posterior, as is the case in the ICA example, and the variational bias will then move us away from this solution.