A Appendix / supplemental material

The Appendix contains additional results and figures to supplement the main body of text.

Theoretical results used within the manuscript are presented in Section A.1. This is followed by a simplification of the KL term when orthogonality constraints are assumed which is outlined in Section A.2. The model architecture and training details implemented in the numerical experiments are detailed in Sections A.3 and A.4 respectively. Lastly, additional results from the conducted numerical experiments can be found in Section A.5. All results throughout this manuscript are given to 4 significant figures, with standard deviations reported to 2 significant figures.

A.1 Theoretical results

Theoretical results utilised within the paper are now presented.

A.1.1 Block matrix inverse

The inverse of a block matrix, which is needed in Section 2.5 to determine the inverse of $\boldsymbol{\Sigma}_C$ in order to derive the KL term, is now outlined.

For a matrix $\boldsymbol{A}$ with block matrix structure

$$\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \tag{13}$$

the following result holds [21, p. 18]:

Lemma A.1. *Assuming all relevant inverses exist, the inverse of $\boldsymbol{A}$ as defined in Eq. 13 is given by:*

$$\begin{pmatrix} (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} - \boldsymbol{A}_{22})^{-1} \\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} - \boldsymbol{A}_{11})^{-1} & (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \end{pmatrix}. \tag{14}$$

Applying this with $\boldsymbol{A}_{ii} = \boldsymbol{I}_i$, $\boldsymbol{A}_{12} = \boldsymbol{C}^T$ and $\boldsymbol{A}_{21} = \boldsymbol{C}$ gives the result in the text.

A.1.2 Weyl's inequality

The following inequality is used within the proof of Theorem 2.1 and concerns the sequence of eigenvalues of matrices. In contrast to the rest of the paper, the eigenvalues are indexed in non-increasing order, not non-increasing order of magnitude. For matrix $\boldsymbol{M}$ these are denoted by $\{\hat{\lambda}_j(\boldsymbol{M})\}_{j=1}^n$.

Weyl's inequality [21, p. 242, Corrollary 4.3.15] says:

Lemma A.2. *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric matrices. The following inequality holds*

$$\hat{\lambda}_j(\boldsymbol{A}) + \hat{\lambda}_1(\boldsymbol{B}) \leq \hat{\lambda}_j(\boldsymbol{A}+\boldsymbol{B}) \leq \hat{\lambda}_j(\boldsymbol{A}) + \hat{\lambda}_n(\boldsymbol{B}), \quad j = 1, \ldots, n. \tag{15}$$

Notice that $\lambda_k(\boldsymbol{I}_1) = \hat{\lambda}_j(\boldsymbol{I}_1) = 1$ for all j, k, and that as $-\boldsymbol{C}^T\boldsymbol{C}$ is negative semi-definite all eigenvalues are non-positive. This means $\hat{\lambda}_j(-\boldsymbol{C}^T\boldsymbol{C}) = \lambda_{n-j}(-\boldsymbol{C}^T\boldsymbol{C})$. Applying this lemma with $\boldsymbol{A} = -\boldsymbol{C}^T\boldsymbol{C}$ and $\boldsymbol{B} = \boldsymbol{I}_1$ gives Eq. 10.

A.2 KL term simplification

Under the orthogonality assumption on $\boldsymbol{C}$, the KL term derived in Eq. 9 can be simplified. Specifically, if $\boldsymbol{C}\boldsymbol{C}^T = \boldsymbol{C}^T\boldsymbol{C} = \alpha^2\boldsymbol{I}$ for some $\alpha \in (0,1)$ then $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \gamma\boldsymbol{I}$ where $\boldsymbol{I} = \boldsymbol{I}_1 = \boldsymbol{I}_2$ and $\gamma = 1/(1-\alpha^2)$. Therefore Eq. 9 reduces to:

$$\begin{aligned} D_{KL}(q_{\phi_1}(\cdot|\boldsymbol{x}_1)q_{\phi_2}(\cdot|\boldsymbol{x}_1)||p_{\boldsymbol{C}}) = \frac{1}{2}&\left[\gamma\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 - \ln|\boldsymbol{\Sigma}_1| - n + \gamma\,\mathrm{tr}\left\{\boldsymbol{\Sigma}_1\right\}\right] \\ &+ \frac{1}{2}\left[\gamma\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 - \ln|\boldsymbol{\Sigma}_2| - n + \gamma\,\mathrm{tr}\left\{\boldsymbol{\Sigma}_2\right\}\right] \\ &- \frac{1}{2}\left[n\ln|\gamma| + 2\gamma\boldsymbol{\mu}_1^T\boldsymbol{C}\boldsymbol{\mu}_2\right]. \end{aligned} \tag{16}$$

A.3 Model architecture

All encoders and decoders for the JPVAE, as well as the neural network used for classification, consist of two dense layers with 512 units each. The latent distribution layers consists of two 20 unit dense layers which each parameterize the mean and the log of the variances of 20 normal random variables. The decoders output layer parameterizes the distribution of a Bernouilli random variable for each pixel. The classifier output layer parameterizes the categorical distribution with 10 outcomes .All activation functions were ReLu.

A.4 Training details

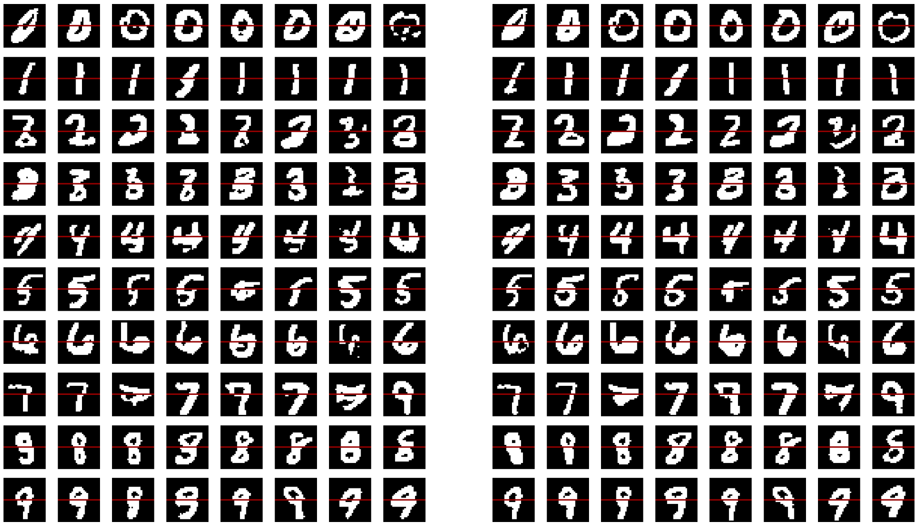
The JPVAE used an Adam optimiser [29] with learning rate 0.001, trained for 30 epochs with a batch size of 32 and used binary cross entropy as the reconstruction error. The cyclical KL annealing schedule as introduced by Fu et al. [16] is implemented, with $M = 30$.

The classifier is trained for 50 epochs with a batch size of 32, step size of 0.01, except for on original data where it is trained for 15 epochs to prevent overfitting. Cross entropy loss is used as the loss function.

A.5 Additional results

Additional results are presented in this section.

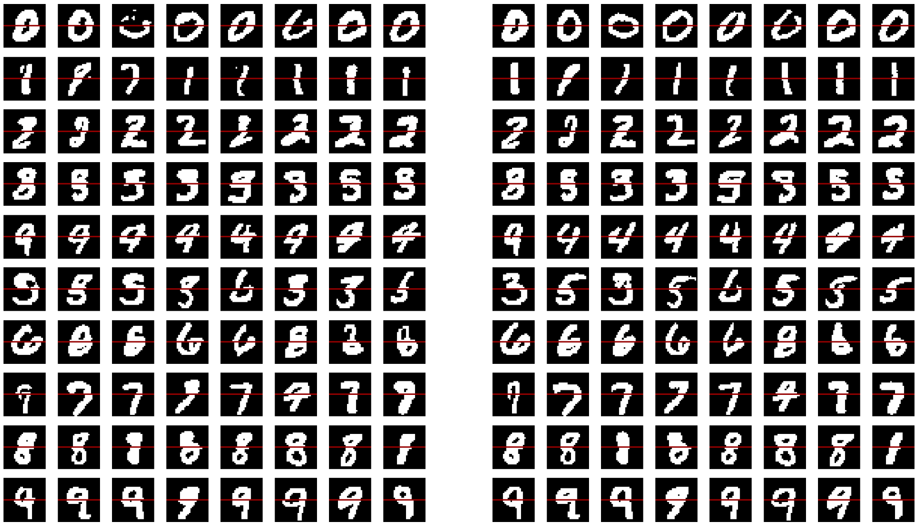
Figure 5 illustrates examples of reconstructing view 2 from view 1, with and without learning correlation natively, for a model trained with a different random seed to that displayed in Figure 1. Figures 6 and 7 illustrate examples of reconstructing view 1 from view 2, again with and without learning correlation natively, for models trained with different random seeds. As in Figure 1, the imputation of the missing view obtained when correlation of the latent spaces is learnt natively is of higher quality than when no correlation is learnt. To quantitively evaluate this superior performance, a simple multi-layer perceptron classifier is trained the concatenation of the reconstructed views. It is then tested on a combination of the reconstructed and imputed views. Explicitly, for datasets where *e.g.* view 2 is imputed from view 1, the classifier is trained on (the training split of) $[\tilde{\boldsymbol{X}}_1; \tilde{\boldsymbol{X}}_2]$ and tested on (the testing split of) $[\tilde{\boldsymbol{X}}_1; \tilde{\boldsymbol{X}}_{2|1}]$. Results of this classification task for various test datasets are displayed in Figure 8. Results for the classification task described in the main body of the text, and presented in Table 3, are illustrated in Figure 9 with standard deviation incorporated via error bars.



(a) No correlation learnt, but estimated empirically after training.

(b) Correlation learnt natively.

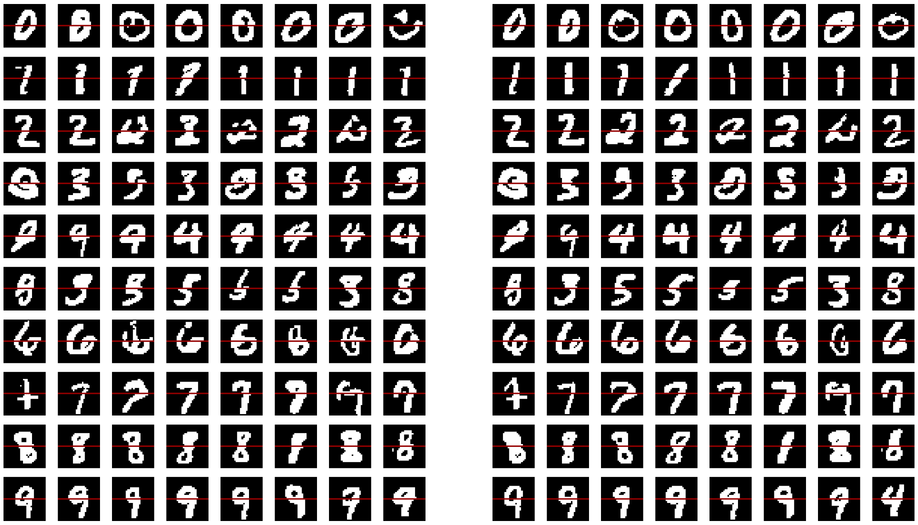
Figure 5: Additional realisation of an imputation of the bottom half of MNIST digits (view 2 of the data) using the top half of the image (view 1) on a JPVAE model trained with (a) independent priors (completely separate VAEs) and (b) a joint prior with learnt correlation structure between latent spaces. The cross entropy loss between true bottom half of image and imputation is 111.9 in (a) and 101.5 in (b).



(a) No correlation learnt, but estimated empirically after training.

(b) Correlation learnt natively.

Figure 6: Imputation of the top half of MNIST digits (view 1 of the data) using the bottom half of the image (view 2) on a JPVAE model trained with (a) independent priors (completely separate VAEs) and (b) a joint prior with learnt correlation structure between latent spaces. The cross entropy loss between true top half of image and imputation is 117.8 in (a) and 100.2 in (b).



(a) No correlation learnt, but estimated empirically after training.

(b) Correlation learnt natively.

Figure 7: Additional realisation of an imputation of the top half of MNIST digits (view 1 of the data) using the bottom half of the image (view 2) on a JPVAE model trained with (a) independent priors (completely separate VAEs) and (b) a joint prior with learnt correlation structure between latent spaces. The cross entropy loss between true top half of image and imputation is 114.3 in (a) and 104.1 in (b).

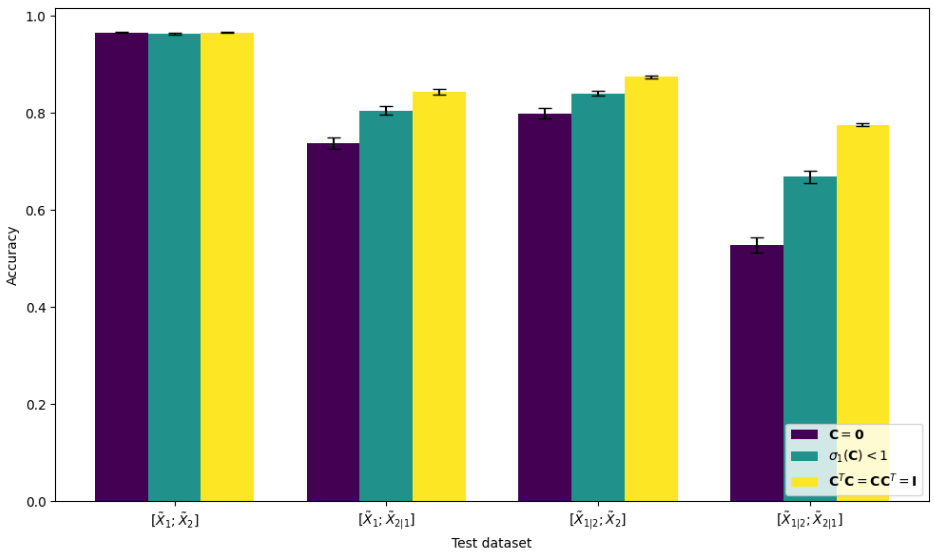


Figure 8: Results for $[\boldsymbol{Y}_1; \boldsymbol{Y}_2]$ represent classification accuracy % for model trained on the training split of $[\tilde{\boldsymbol{X}}_1; \tilde{\boldsymbol{X}}_2]$ and tested on the test split of $[\boldsymbol{Y}_1; \boldsymbol{Y}_2]$ (the column wise concatenation of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$). Accuracy for $[\boldsymbol{X}_1; \boldsymbol{X}_2]$ with standard deviation in brackets is 98.04% (0.074). Error bars present +/- one standard deviation.

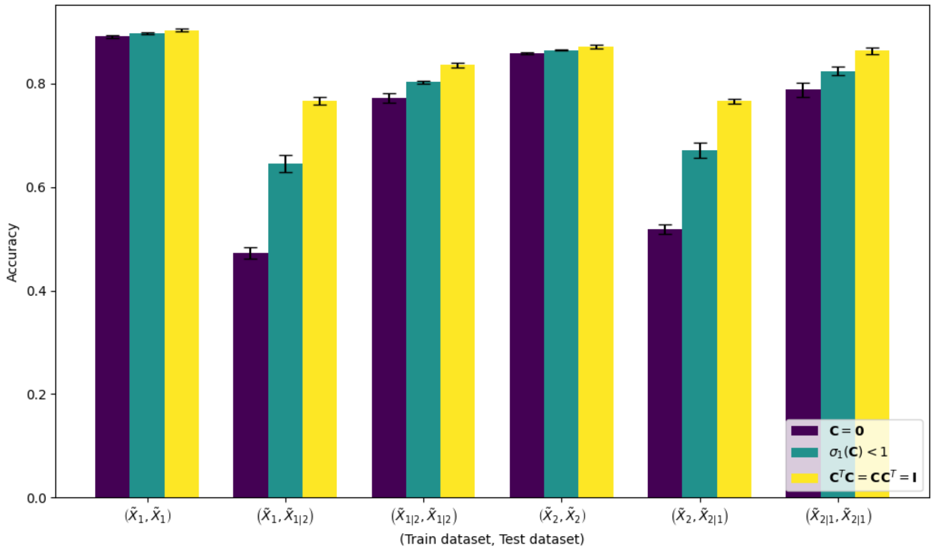


Figure 9: Results for $(\boldsymbol{Y}, \boldsymbol{Z})$ represent classification accuracy % for model trained on the training split of $\boldsymbol{Y}$ and tested on the test split of $\boldsymbol{Z}$. Accuracies for $(\boldsymbol{X}_1, \boldsymbol{X}_1)$ and $(\boldsymbol{X}_2, \boldsymbol{X}_2)$ with standard deviation in brackets are 93.59% (0.25) and 90.83% (0.23) respectively. Error bars present +/- one standard deviation.