5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–11

Independent Mechanism Analysis in GPLVMs

Patrik Reizinger* PATRIK.REIZINGER@UNI-TUEBINGEN.DE
University of Tübingen, Germany
Max Planck Institute for Intelligent Systems, Tübingen, Germany
International Max Planck Research School for Intelligent Systems (IMPRS-IS)
European Laboratory for Learning and Intelligent Systems (ELLIS)

Han-Bo Li* HBL26@CAM.AC.UK
Aditya Ravuri* AR847@CAM.AC.UK
Neil D Lawrence† NDL21@CAM.AC.UK
Ferenc Huszár† FH277@CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, United Kingdom

Abstract

Independence is a common assumption for modeling generative processes. Independent Mechanism Analysis (IMA) relies on the Independent Causal Mechanisms (ICM) principle to formulate non-statistical independence by measuring the decoder Jacobian's column-orthogonality. This work is based on observations of the same column-orthogonality in GPLVMs and shows how, e.g., additive and stationary kernels in GP priors give rise to independent mechanisms in expectation. To handle the stochasticity of the decoding function in GPLVMs, we upper bound the orthogonality measure under specific kernel conditions. We believe that the connection between IMA and GPLVMs highlights a useful inductive bias in GPLVMs for recovering the true latent factors, which we will study as part of future work.

1. Introduction

Independence is a prevalent assumption in machine learning, with most works relying on statistical independence of latent factors (Hyvärinen and Pajunen, 1999; Kingma and Welling, 2014). The ICM principle (Peters et al., 2018) conceptually formulates independence of "modules", corresponding to nature's mechanisms, and was translated into a non-statistical independence notion in Independent Mechanism Analysis (IMA) (Gresele et al., 2021). IMA defines independent mechanisms in generative models (or decoders, i.e., the map from latent factors to observations) by having a column-orthogonal Jacobian. The proposed regularized log-likelihood objective provably rules out spurious solutions when learning the true latent factors. Buchholz et al. (2022) proved identifiability for conformal maps, where the Jacobian columns have equal norms, and showed local identifiability for the IMA function class. Reizinger et al. (2022) connected IMA to Variational Autoencoders (VAEs) (Kingma and Welling, 2014) in the limiting case of isotropic observation noise with diminishing variance. Gaussian Process Latent Variable Models (GPLVMs) (Lawrence, 2005) also often possess the same column orthogonality, which we connect to IMA. Our **contributions** are:

* Equal contribution. Code available at: https://github.com/rpatrik96/gp-ima
† Joint senior authors

- We describe the kernel properties necessary to satisfy the non-statistical independence of IMA in GPLVM function priors;
- Similar to Ghosh (2022), we upper bound IMA's orthogonality measure and show that under specific kernel choices, the IMA principle holds approximately;
- We illustrate in toy experiments how the learned latent space in GPLVMs resembles the true one.

2. Background

Notation We focus on generative models with decoder maps $\boldsymbol{f} : \mathcal{Z} \to \mathcal{X}$ and their Jacobians $\mathbf{J}_{\boldsymbol{f}} = \mathbf{J}_{\boldsymbol{f}}(\boldsymbol{z})$, with $\mathcal{Z}$ and $\mathcal{X}$ being the latent and observation spaces with dimensions d and D. We denote $\mathbf{J}_{\boldsymbol{f}} := {}^{\partial \boldsymbol{f}}/_{\partial \boldsymbol{z}} \in \mathbb{R}^{D \times d}$ such that $(\mathbf{J}_{\boldsymbol{f}})_{ij} := {}^{\partial \boldsymbol{f}_i^k}/_{\partial z_j^k}$, i.e., the i^{th} row and j^{th} column gives the partial derivative of the i^{th} dimension of $\boldsymbol{f}$ for the latent dimension j (for a given sample k, which we will omit for brevity). The observations are denoted by matrix $\mathbf{X} \in \mathbb{R}^{n \times D}$ and the corresponding latent variables as $\mathbf{Z} \in \mathbb{R}^{n \times d}$. The i^{th} dimension of the k^{th} observation is given by x_i^k.

2.1. Independent Mechanism Analysis (IMA)

Observed data $\boldsymbol{x}$ are often modeled as a mixture of latent factors, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z})$. Representation learning then learns an unmixing $\boldsymbol{g}$ such that the recovered components $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ recover the true ones up to tolerable ambiguities such as permutations or scalings (Bengio et al., 2013; Khemakhem et al., 2020)—which is impossible for nonlinear $\boldsymbol{f}$ without further constraints (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). IMA (Gresele et al., 2021) restricts the mixing function class, postulating that latent components influence the observations "independently" through the partial derivatives ${}^{\partial \boldsymbol{f}}/_{\partial z_k}$. This is equivalent to an orthogonality condition on the decoder Jacobian's columns. While identifiability only holds for a subset of this model class (Buchholz et al., 2022), IMA provably rules out the most common counterexamples to identifiability such as the Darmois construction[1] (Darmois, 1951) and helps recover the ground-truth latent factors in practice (Gresele et al., 2021; Sliwa et al., 2022; Reizinger et al., 2022). IMA optimizes the regularized log-likelihood: $\mathcal{L}_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z}) := \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \lambda \cdot c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z})$ where the regularizer $c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z})$ encourages column-orthogonality of $\mathbf{J}_{\boldsymbol{f}}(\boldsymbol{z})$. $c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z})$ and its expectation, w.r.t. the latent distribution $p(\boldsymbol{z})$), $C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z}))$ are defined as

$$c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z}) = \sum_k \log \left\| \frac{\partial \boldsymbol{f}}{\partial z_k}(\boldsymbol{z}) \right\| - \log |\mathbf{J}_{\boldsymbol{f}}(\boldsymbol{z})|; \quad C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z})) = \mathbb{E}_{p(\boldsymbol{z})}[c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z})]. \quad (1)$$

$c_{\text{IMA}}(\boldsymbol{f}, \boldsymbol{z})$ measures the deviation of the decoder Jacobian columns from being orthogonal: the product of the column norms is the volume of the hyperrectangle spanned by the Jacobian columns, which equals only the determinant (the volume of the parallelepiped) if the columns are orthogonal.

1. The Darmois construction recursively applies the conditional cumulative distribution function transform to yield independent and uniform latent factors, but these latents will be, in general, a nonlinear mixture of the true latent factors

2.2. Gaussian Process Latent Variable Models (GPLVMs)

Gaussian Processes (GPs) are stochastic processes over real-valued functions (Rasmussen and Williams, 2006), which provide a Bayesian non-parametric framework for inference. GPLVMs (Lawrence, 2005) model the mapping between latent variables $\boldsymbol{z}$ and observed data $\boldsymbol{x}$ with GP priors. Independent GPs model each of the D observed dimensions as $x_i^k = f_i(\boldsymbol{z}^k) + \varepsilon_i^k$, where $\varepsilon_i^k \sim \mathcal{N}\left(0; \beta^{-1}\mathbf{I}_d\right)$. When a GP prior is placed over $\mathbf{X}$ and when $\boldsymbol{f} = \{\boldsymbol{f}_i\}_{i=1}^D : \boldsymbol{f}_i := \boldsymbol{f}_i(\boldsymbol{z}) + \varepsilon_i$ is marginalized, the likelihood for $\mathbf{X}$ is given by:

$$p(\mathbf{X}|\mathbf{Z}) = \prod_{i=1}^{D} p(\boldsymbol{x}_i|\mathbf{Z}); \qquad p(\boldsymbol{x}_i|\mathbf{Z}) = \mathcal{N}(\boldsymbol{x}_i|\mathbf{0}, \mathcal{K} + \beta^{-1}\mathbf{I}_n), \tag{2}$$

where $\mathcal{K} \in \mathbb{R}^{n\times n}$ is the GP's covariance matrix, given by the kernel $[\mathcal{K}]_{ij} := k\left(\boldsymbol{z}^i; \boldsymbol{z}^j\right)$.

GP derivatives. By the linearity of differentiation, the derivative of a GP is, in a mean-squared sense, a GP (Rasmussen and Williams, 2006, Sec. 9.4); i.e., $\mathcal{GP}\left(\mathbf{0}, \mathcal{K}\right)' = \mathcal{GP}\left(\mathbf{0}, \mathcal{K}''\right)$, provided that the kernel function is at least twice differentiable (Papoulis and Unnikrishna Pillai, 2002) (therefore, the covariance function must not have a noise component). For a given $\boldsymbol{z}$ (dropping k and using only $\boldsymbol{f}$ without subscript), the derivative GP's covariance function $\mathcal{K}''$, is given by a $d \times d$–dimensional matrix

$$\left[\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}} \middle| \boldsymbol{z}\right]\right]_{ij} = \operatorname{Cov}\left[\frac{\partial \boldsymbol{f}}{\partial z_i}, \frac{\partial \boldsymbol{f}}{\partial z_j}\right] = \frac{\partial^2 k\left(\boldsymbol{z}; \boldsymbol{z}\right)}{\partial z_i \partial z_j}. \tag{3}$$

This derivative process corresponds to the partial derivative $\partial \boldsymbol{f}/\partial \boldsymbol{z}$. This also holds for differentiable mean functions; however, w.l.o.g., we assume zero mean, which implies $\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}|\boldsymbol{z}\right] = \mathbf{0}_{D\times d}$ over the function priors.

3. Theoretical Results

Our observation is that the generative model in GPLVMs often has a Jacobian with orthogonal columns. IMA shows that this inductive bias is useful for learning the true latents. Thus, we use IMA to study the implications of kernel choice in GPLVMs and describe when the prior decoder will have column-orthogonal Jacobian in expectation. Our work also provides a causal perspective, since IMA was inspired by the ICM principle of the causality literature. A key insight for our analysis is that $c_{\mathrm{IMA}}(\boldsymbol{f}, \boldsymbol{z})$ is the left KL-measure of diagonality (Alyani et al., 2017) of the matrix $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}$, which depends on $\boldsymbol{z}$. $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}|\boldsymbol{z}$ has a Wishart distribution with D degrees of freedom, $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}|\boldsymbol{z} \sim \mathcal{W}_d\left(\mathbf{A}; D\right)$, where the scale matrix $\mathbf{A}$ needs to be diagonal for IMA to hold over the function priors. I.e., $\left[\mathbb{E}_{p(\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right]\right]_{i\neq j} = 0$. These off-diagonal entries correspond to the expectation of the cross-derivatives from (3), connecting IMA to the kernel—note that our claim is about the marginal variance, i.e., when $\boldsymbol{z} = \boldsymbol{z}' : k\left(\boldsymbol{z}; \boldsymbol{z}'\right)$. Since $\mathbb{E}_{p(\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right] = \mathbb{E}_{p(\boldsymbol{z})}\left[\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}} \middle| \boldsymbol{z}\right]\right]$, the inner expectation for element ij yields

$$\left[\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}} \middle| \boldsymbol{z}\right]\right]_{ij} = \sum_{k=1}^{D} \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\left(\mathbf{J}_{\boldsymbol{f}}\right)_{ki}\left(\mathbf{J}_{\boldsymbol{f}}\right)_{kj} \middle| \boldsymbol{z}\right] = \sum_{k=1}^{D} \operatorname{Cov}\left[\frac{\partial f_k}{\partial z_i}, \frac{\partial f_k}{\partial z_j}\right] = D\frac{\partial^2 k\left(\boldsymbol{z}; \boldsymbol{z}\right)}{\partial z_i \partial z_j} \tag{4}$$

Thus, $\left[\mathbb{E}_{p(\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right]\right]_{i\neq j} = 0$ when the expectation of the cross-derivative of $k(\boldsymbol{z};\boldsymbol{z})$ is zero. We leverage this insight in two steps: first, we show how to design GPLVMs such that the above expectation is diagonal: by choosing first-order additive or stationary kernels (Prop. 1). Then, we bound $C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z}))$ to show that it is possible to (approximately) satisfy the IMA principle (Prop. 2).

Proposition 1 *[$\partial^2_{ij} k(\boldsymbol{z};\boldsymbol{z}) = 0 \iff k(\boldsymbol{z};\boldsymbol{z}) \equiv \sum_i g_i(z_i; z_i)$] The cross-derivatives of (4) are zero if and only if the marginal variance can be decomposed into a sum of terms that depend at most on one latent coordinate z_i. Covariance functions that are stationary also admit zero cross-derivatives as the marginal variance is a constant w.r.t. the latent position.*

Prop. 1 characterizes the kernel family that fulfills the necessary requirement for the IMA principle (if the kernel is also twice-differentiable). See App. B for proof details. However, our result only implies that the expected value of $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}$ is diagonal. Since a GPLVM defines a distribution over $\boldsymbol{f}$ and $\mathbf{J}_{\boldsymbol{f}}$, this might imply that though $\mathbb{E}_{p(\boldsymbol{z})}\left[\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right]$ is diagonal, IMA is not satisfied for each $\boldsymbol{f}$ drawn from the $\mathcal{GP}$.

Proposition 2 **($D \to \infty \implies C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z})) \to 0$)** *If a GPLVM has a kernel that is twice-differentiable and has zero cross-derivatives, then $D \to \infty \implies C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z})) \to 0$ for a given d[2]. Note that this statement assumes the prior distribution over $\boldsymbol{f}$ (i.e. before conditioning on any data points or inducing points).*

Proof We formulate $C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z}))$ as the expected left KL-measure of diagonality of $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}$ (Alyani et al., 2017):

$$C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z})) = \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[\log\left|\operatorname{diag}\left(\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right)\right| - \log\left|\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right|\right] \tag{5}$$

$$= \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\log\left|\operatorname{diag}\left(\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right)\right| - \log\left|\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}\right| \,\middle|\, \boldsymbol{z}\right]\right], \tag{6}$$

where $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}$ has a Wishart distribution when conditioned on the latents, $\mathcal{W}_d(\mathbf{A}; D)$, where $\mathbf{A}$ is the matrix with elements $a_{ij} = \partial^2_{ij} k(\boldsymbol{z};\boldsymbol{z})$. Together with the expression for the expectation of the log-determinant of a Wishart distribution (Bishop and Nasrabadi, 2006, App. B, p. 693), the observation that the diagonals of $\mathbf{J}_{\boldsymbol{f}}^\top \mathbf{J}_{\boldsymbol{f}}$ are $a_{ii} \cdot \chi^2_D$–distributed (Rao et al., 1973, 8b.2, p. 535), and that of the expectation of a $\log(\chi^2_D)$ distribution (Pav, 2015); $C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z}))$ becomes

$$= \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[\sum_{i=1}^{d}\left[\psi\left(\frac{D}{2}\right) - \psi\left(\frac{D+1-i}{2}\right)\right] + \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{z})}\left[\log\frac{|\operatorname{diag}(\mathbf{A})|}{|\mathbf{A}|}\middle|\boldsymbol{z}\right]\right], \tag{7}$$

where digamma $\psi(k)$ is $\psi(k) = \frac{d}{dk}\log(\Gamma(k)) = \frac{\Gamma'(k)}{\Gamma(k)}$.

2. Note that d does not have to be fixed, as long as it is restricted to a function of D such that the upper bound collapses

$\mathbf{A}$ is diagonal, when $\partial^2_{ij} k(\boldsymbol{z}; \boldsymbol{z}) = 0$ for $i \neq j$, which means that the last two terms cancel, the expectation is of a constant, and we can bound the expression with

$$C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z})) \leq \frac{d}{2}\left[\psi\left(\frac{D}{2}\right) - \psi\left(\frac{D+1-d}{2}\right)\right], \tag{8}$$

since $\psi(k)$ is strictly increasing. This upper bound goes to zero for a fixed d as $D \to \infty$ due to the continuity of ψ. This can also be seen through the limiting behaviour of the digamma function for large real arguments, i.e., $\psi(k) \approx \log(k) - 1/2k$, yielding

$$\approx \frac{d}{2}\left[\log\left(\frac{D}{D+1-d}\right) + \frac{1-d}{D(D+1-d)}\right]. \tag{9}$$

■

Notably, Prop. 2 aligns with the result of Ghosh (2022), which shows that as D grows for a given d, $C_{\text{IMA}}(\boldsymbol{f}, p(\boldsymbol{z}))$ can be upper bounded—Ghosh (2022) assumes that the columns of $\mathbf{J}_{\boldsymbol{f}}$ are drawn i.i.d. from a spherically invariant distribution. Since the prerequisite for Prop. 2 is to have a kernel with zero cross-derivatives, our result might suggest an unexpected consequence of kernel choice: since IMA is beneficial for recovering the true latent factors, GPLVMs might also have such properties, at least for high-dimensional observations. Lastly, given the kernel choice, the upper bound does not depend on any kernel hyper-parameters.

4. Experiments

Setup. We generate 500 samples via the Möbius transform using (Stimper et al., 2021), where $\boldsymbol{z}$ is uniformly distributed in $[0; 1]^d$. For visualization, we set $d = D = 2$ (the Möbius transform is a bijective map). We use (GPy, 2012), Radial Basis Function (RBF) kernel, the L-BFGS (Liu and Nocedal, 1989) optimizer with 5 seeds, 5 restarts, and monitor C_{IMA}.
Results. Figure 1 replicates (Gresele et al., 2021, Fig. 4 (Top)) to visualize the relation of the inferred and true latent factors (the estimated latents are transformed by the cumulative distribution function of the uniform distribution for the figure). From left to right, we plot the true latent factors, the observations, and the reconstructed latents[3], respectively (we plot the representations of the models with the lowest C_{IMA}). We evaluate our bound quantitatively in Appx. A.

5. Discussion

Limitations. To visualize the latent space structure, our experiments are restricted to $d = D = 2$, which is a practically less relevant, though theoretically widely-employed scenario when studying the recovery of the true latent factors (Hyvärinen and Pajunen, 1999; Gresele et al., 2021). However, our results suggest that when the IMA condition (approximately) holds in GPLVM, then they elicit beneficial properties for recovering the true latent factors, though this requires further investigation.
Conclusion. Our work connects kernel choice in GPLVMs to Independent Mechanism

3. We used a sparse GPLVMs for Fig. 1, as that resulted in better reconstruction

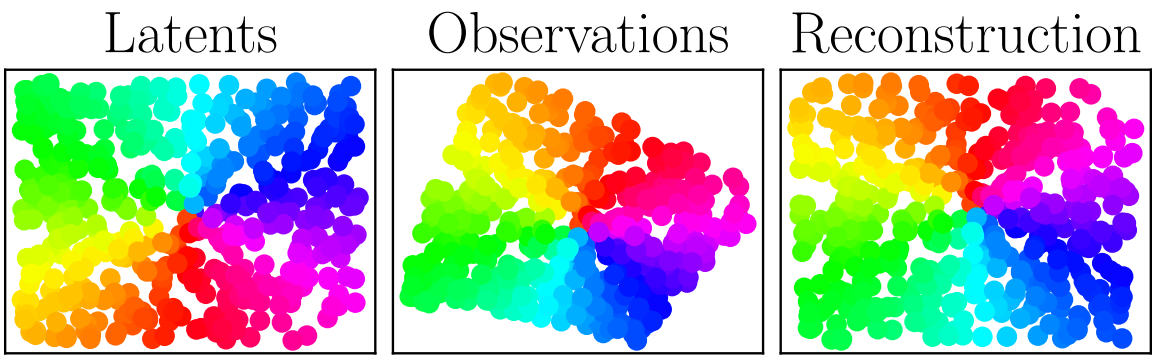

Figure 1: **From left to right:** true latent factors; observations generated by the Möbius transform; reconstructed latents reconstructed latents (we used a sparse GPLVM)

Analysis (IMA) (Gresele et al., 2021) by stating that zero cross-derivatives—such as in first-order additive and stationary kernels—are necessary for the IMA condition to hold in the prior over the functions of the GPLVM. Furthermore, we upper bound IMA's (non-statistical) independence measure based on the observation and latent dimensions and show that for a given latent dimensionality, increasing the observation dimension makes the bound go to zero. The connection between IMA and GPLVMs possibly explains the observation of practitioners that the decoder Jacobian in GPLVMs can have orthogonal columns. Our experimental results on synthetic data suggest that the orthogonality of the Jacobian columns is beneficial for learning the true latent factors in GPLVMs.

Acknowledgments

The authors thank Luigi Gresele for poising the question of whether the connection between IMA and variational inference goes beyond VAEs and for detailed suggestions on both qualitative and quantitative evaluations. The authors would like to thank Francisco Vargas for detailed and insightful comments on an earlier form of the paper. In addition, the authors would like to thank Michel Besserve, Bernhard Schölkopf, and Viacheslav Borovitskiy for fruitful discussions. This work was supported by a Turing AI World-Leading Researcher Fellowship G111021. Patrik Reizinger thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support and acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. Hanbo Li would like to thank the Cambridge Trust for their support of his studies. Aditya Ravuri would like to thank the Accelerate Programme for Scientific Discovery for support relating to the PhD.

References

Khaled Alyani, Marco Congedo, and Maher Moakher. Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem. *Linear Algebra and its Applications*, 528:290–320, September 2017. ISSN 0024-3795. doi: 10.1016/j.laa.2016.08.031. URL https://www.sciencedirect.com/science/article/pii/S0024379516303834.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.2013.50. URL https://doi.org/10.1109/tpami.2013.50.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. page 25, 2022. URL https://arxiv.org/abs/2208.06406.

George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

Shubhangi Ghosh. Independent Mechanism Analysis in High-Dimensional Observation Spaces. Master's thesis, June 2022. URL https://www.research-collection.ethz.ch/handle/20.500.11850/591418?show=full. Accepted: 2023-01-11T08:54:06Z.

GPy. GPy: A gaussian process framework in python. http://github.com/SheffieldML/GPy, 2012.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *arXiv:2106.05200 [cs, stat]*, June 2021. URL http://arxiv.org/abs/2106.05200. arXiv: 2106.05200.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, June 2020. URL http://proceedings.mlr.press/v108/khemakhem20a.html. ISSN: 2640-3498.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(11), 2005.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, May 2019. URL http://proceedings.mlr.press/v97/locatello19a.html. ISSN: 2640-3498.

Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables and stochastic processes*. 2002.

Steven E Pav. Moments of the log non-central chi-square distribution. *arXiv preprint arXiv:1503.06266*, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. *Journal of Statistical Computation and Simulation*, 88(16):3248–3248, November 2018. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2018.1505197. URL https://www.tandfonline.com/doi/full/10.1080/00949655.2018.1505197.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9. OCLC: ocm61285753.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the Gap: VAEs Perform Independent Mechanism Analysis. October 2022. URL https://openreview.net/forum?id=G4GpqX4bKAH.

Joanna Sliwa, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf. Probing the Robustness of Independent Mechanism Analysis for Representation Learning, July 2022. URL http://arxiv.org/abs/2207.06137. arXiv:2207.06137 [cs, stat].

Vincent Stimper, Luigi Gresele, Joanna Sliwa, and Adrián Javaloy. Independent Mechanism Analysis repository. https://github.com/lgresele/independent-mechanism-analysis, 2021.

Appendix A. Additional Experiments

To show quantitatively that our bound (8) holds, we generate data with the Möbius transformation ($d = D$), for $\{2; 3; 5; 8; 10\}$ and plot the IMA contrast before fitting the data and after fitting the data—here we use a vanilla GPLVM, not a sparse one; all other parameters are the same as in § 4. As the left plot in Fig. 2 shows, both quantities are below the bound[4]. Interestingly, c_{IMA} is much lower for the posterior than the prior. We will investigate in the future whether the bound can be tightened for the posterior. The trained GPLVMs reconstruct the true latent factors reasonably well (Fig. 2, right).

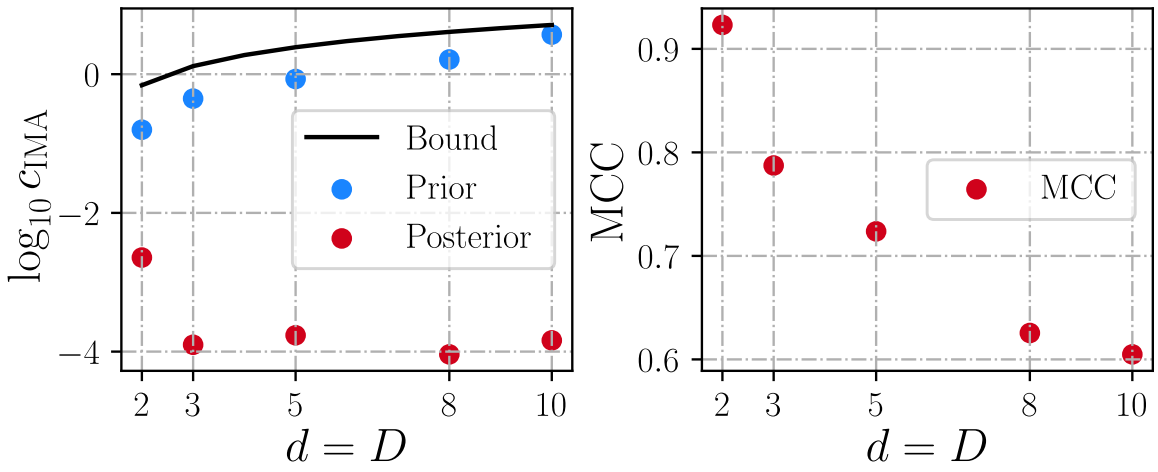

Figure 2: Quantitative evaluation for Moebius transformation data ($d = D$). **Left:** logarithm of c_{IMA} for the GP prior (before training) and posterior (after training) compared to our bound from (8). **Right:** Mean Correlation Coefficient (MCC) for quantifying the reconstruction quality of the true latent factors (higher is better, range is $[0; 1]$)

Though the prior, for which our bound holds, is data-agnostic, we need samples at which we evaluate the Jacobian. We observe that the calculated c_{IMA} is highly influenced by the number of data points used. We demonstrate this data-dependence in Fig. 3 for multiple dimensions (color-coded) and number of samples (marker-coded). For 8 and 10 dimensions, the bound is only satisfied for sufficiently high data points. A possible explanation is that since C_{IMA} in (1) takes the expectation w.r.t. $p(\boldsymbol{z})$, we need more samples for higher dimensions to "cover" the space.

Appendix B. Proofs

B.1. Proof of Prop. 1

Proposition 3 *$[\partial^2_{ij} k(\boldsymbol{z}; \boldsymbol{z}) = 0 \iff k(\boldsymbol{z}; \boldsymbol{z}) \equiv \sum_i g_i(z_i; z_i)$] The cross-derivatives of (4) are zero if and only if the marginal variance can be decomposed into a sum of terms that depend at most on one latent coordinate z_i. Covariance functions that are stationary also admit zero cross-derivatives as the marginal variance is a constant w.r.t. the latent position.*

4. The bound is *increasing*, since we increase both d and D

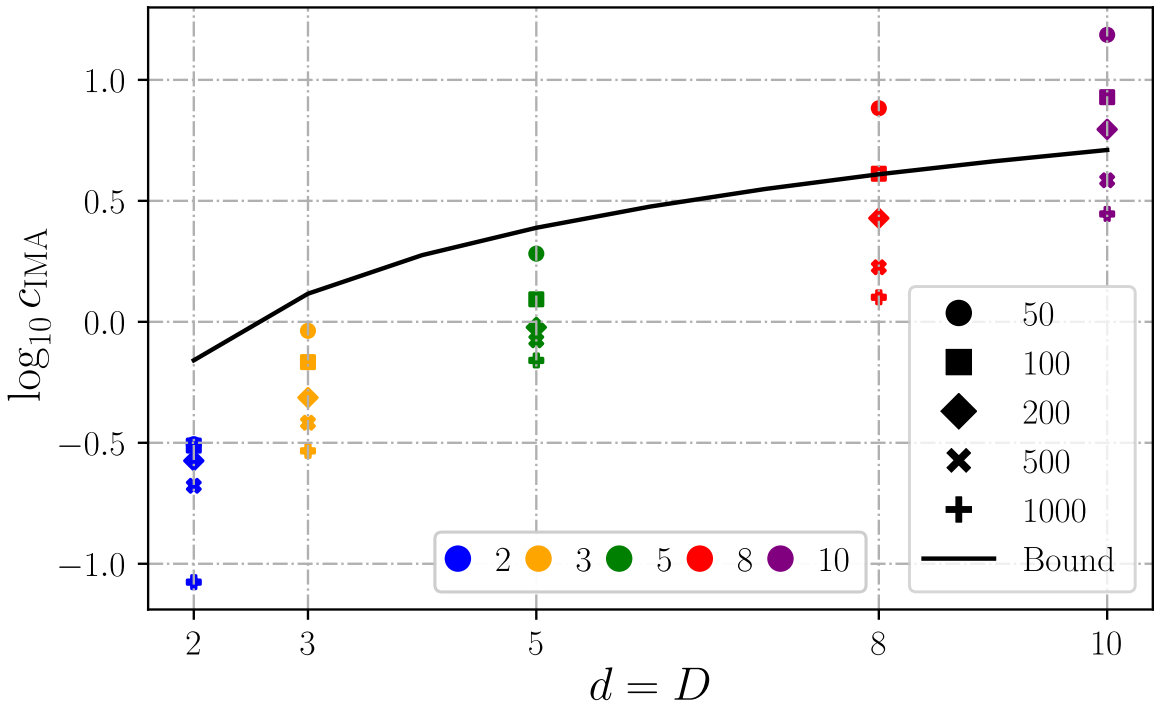

Figure 3: Data-dependence of calculating c_{IMA} for the prior for Möbius transformation data ($d = D$). The dimension is color-coded, different marker stand for different number of samples for calculating the Jacobian

Proof $\partial^2_{ij} k(\boldsymbol{z}; \boldsymbol{z})$ denotes the cross derivative of $k(\boldsymbol{z}; \boldsymbol{z})$ w.r.t. z_i, z_j, and A, B some functionals.
$\implies$: By assumption, the cross-derivative $\partial^2_{ij} k(\boldsymbol{z}; \boldsymbol{z})$ is zero. We integrate twice: by z_i, then by z_j, which gives rise to the additive functionals A, B. Integration by z_i would mean that the resulting additive functional can depend on any $z_{j \neq i}$. Since $i \neq j$ are arbitrary, z_i, z_j cannot be in the same functional, so only one latent component remains:

$$\partial_j k(\boldsymbol{z}; \boldsymbol{z}) = A'(z_j) \tag{10}$$

$$k(\boldsymbol{z}; \boldsymbol{z}) = \int A'(z_j) dz_j = A(z_j) + B(z_i), \tag{11}$$

which shows that $k(\boldsymbol{z}; \boldsymbol{z})$ factorizes into functions of single latent coordinates (or constant terms), i.e., when the kernel is the sum of base kernels depending on at most one coordinate.
$\impliedby$: By assumption, the kernel decomposes into base kernels depending on at most one latent coordinate, i.e., $k(\boldsymbol{z}; \boldsymbol{z}) = \sum_{i=1}^{d} k_i(\boldsymbol{z}; \boldsymbol{z})$. Differentiating by z_i and then by z_j yields

$$\partial_i k(\boldsymbol{z}; \boldsymbol{z}) = \sum_{i=1}^{d} \partial_i \; k_i(\boldsymbol{z}; \boldsymbol{z}) = \partial_i \; k_i(\boldsymbol{z}; \boldsymbol{z}) \tag{12}$$

$$\partial^2_{ij} k(\boldsymbol{z}; \boldsymbol{z}) = \partial_{ij} \; k_i(\boldsymbol{z}; \boldsymbol{z}) = 0, \tag{13}$$

which is zero since each k_i depends only on z_i. ■

Acronyms

IMA Independent Mechanism Analysis

GP Gaussian Process
GPLVM Gaussian Process Latent Variable Model

i.i.d. independent and identically distributed
ICM Independent Causal Mechanisms

KL Kullback-Leibler Divergence

MCC Mean Correlation Coefficient

RBF Radial Basis Function

VAE Variational Autoencoder

Nomenclature

Independent Mechanism Analysis
- $C_{\mathbf{IMA}}$ global IMA contrast
- $\boldsymbol{y}$ reconstructed sources
- $\mathcal{L}_{\mathbf{IMA}}$ IMA loss function
- $c_{\mathbf{IMA}}$ local IMA contrast

Gaussian Processes
- $\mathcal{GP}$ Gaussian Process
- $\mathcal{K}$ $n \times n$ covariance matrix of a GP
- k kernel function

Variational Autoencoders
- $\boldsymbol{\theta}$ parameters of the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$
- $\boldsymbol{g}$ inverse decoder
- $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ marginal likelihood
- $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ conditional distribution of the decoded samples of the VAE, mapping $\boldsymbol{z} \mapsto \boldsymbol{x}$, parametrized by $\boldsymbol{\theta}$

$\boldsymbol{f}$ decoder map $\mathcal{Z} \to \mathcal{X}$
f decoder map component
n number of samples

Algebra
- $\mathbf{0}$ a vector of zeros
- $\mathbf{I}$ identity matrix
- $\mathbf{J}$ Jacobian matrix

Latents
- $\boldsymbol{z}$ latent vector
- $\mathbf{Z}$ latent matrix of $\mathbb{R}^{n \times d}$
- $\mathcal{Z}$ latents
- d dimensionality of the latent space $\mathcal{Z}$
- z latent single component

Observations
- D dimensionality of the observation space $\mathcal{X}$
- $\boldsymbol{x}$ observation vector
- $\mathbf{X}$ observation matrix of $\mathbb{R}^{n \times D}$
- $\mathcal{X}$ observation space
- x observation single component

Probability theory
- $\mathcal{N}$ normal distribution
- $\mathcal{W}$ Wishart distribution