Cellij: A Modular Factor Model Framework for Interpretable and Accelerated Multi-Omics Data Integration

Martin Rohbeck[*1 2 3] **Arber Qoku**[*1 4 5] **Tim Treis**[*1 6] **Fabian J Theis**[6 7 8] **Britta Velten**[1 3]
Florian Buettner[1 4 5 9] **Oliver Stegle**[1 2 8]

Abstract

The analysis and integration of multi-omics datasets requires flexible modelling choices to faithfully capture the underlying biological processes that are active in one or multiple omics layers. Factor analysis is among the most successful approaches for this task, yet adapting this model class to specific biological questions and datasets is a time consuming step that has resulted in "reinventing the wheel". Here, we present *Cellij*, a versatile factor analysis framework for rapidly building and training a wide range of factor analysis models for multi-omics data. By demonstrating how the framework unifies dozens of previously distinct factor analysis models, *Cellij* enables to perform objective benchmarks, which we use to present a study of alternative sparsity assumptions for the first time. Finally, we illustrate how *Cellij* integrates covariates through Gaussian Processes on a real-world transcriptomic dataset – enhancing the interpretability of the resulting latent factors.

1. Introduction

Multi-omics experimental designs, whereby the same samples or biological specimen are assayed using multiple omics modalities, are increasingly deployed across different fields and questions, ranging from basic biology to translation. However, analysing and interpreting the resulting datasets remains a major challenge due to their high-dimensional nature, extensive missing values, batch effects and the need to model shared and modality-specific variation. Factor Analysis (FA) models have proven to be particularly effective in modelling such data, allowing for the identification of underlying factors that explain the shared and modality-specific variation between multiple omics layers. This model class avoids the need to impose strong assumptions such as a discrete clustering structure, yet is amenable to a direct mechanistic interpretation owing to their (local) linear structure.

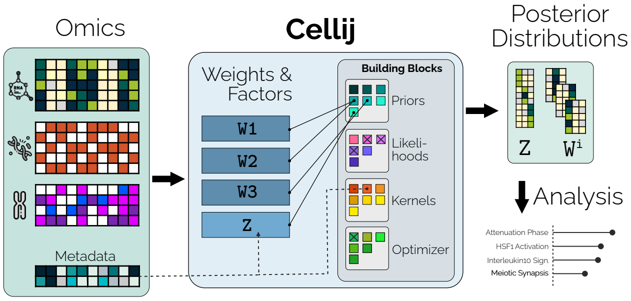

Figure 1. Cellij, a modular framework to rapidly build and train factor analysis models on multi-omics data. *Cellij* jointly models multiple omics modalities, and can be flexibly configured to employ modality-specific sparsity assumptions, modality-specific likelihoods and latent-space prior for metadata. Once trained, *Cellij* enables diverse downstream analyses and model inspection.

However, in order to maximise the utility and accuracy of FA models, appropriate choices on sparsity, model regularisation and noise distributions are indispensable. Even subtle changes in the statistical assumptions of FA models can dramatically affect the analysis outcomes. Consequently, a vast number of specialised methods that build on FA have been proposed across diverse biomedical use-cases and data modalities. These existing models are implemented using dedicated software and inference schemes. Hence, it remains a complex and challenging task to adapt such models to a specific problem at hand, or compare models that differ in a subset of their assumptions. Moreover, known covariates information such as experimental design, or the temporal (1D) or spatial (2D) arrangement of the samples remain underused in existing implementations.
A promising direction to ameliorate this fragmentation are

[*]Equal contribution [1]German Cancer Research Center (DKFZ), Germany [2]European Molecular Biology Laboratory (EMBL), Germany [3]Heidelberg University, Germany [4]German Cancer Consortium (DKTK), Germany [5]Goethe University Frankfurt, Germany [6]Helmholtz Center Munich, Germany [7]Technical University of Munich, Germany [8]Wellcome Sanger Institute, UK [9]Frankfurt Cancer Institute (FCI), Germany. Correspondence to: Oliver Stegle <o.stegle@dkfz.de>, Florian Buettner <florian.buettner@dkfz.de>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

meta models implemented using probabilistic programming frameworks, i.e. a template to define various specific models on-the-fly, which would allow for rapid model building and comparison of alternative statistical assumptions, while providing both explainability and expressiveness.

To address this, we here present *Cellij*, a flexible FA-based framework that enables researchers to rapidly explore and adapt models to specific use-cases (c.f. Figure 1). *Cellij* builds upon a Bayesian FA skeleton that is designed to provide wide-ranging customizability at all levels, ranging from likelihoods and optimisation procedures to sparsity-inducing priors. The framework is designed based on the following principles:
Rapid Prototyping: *Cellij* is designed for rapid prototyping of custom FA models, allowing users to efficiently define new models in an iterative fashion.
Interpretability: Due to its inherent linear structure *Cellij* allows (i) a straightforward interpretation of the unravelled biological processes and (ii) a more accountable assessment of the results by providing uncertainty estimates for the inferred parameters. By imposing structured sparsity priors on the latent decomposition, *Cellij* models uncover inherently interpretable latent factors.
Integration of Covariates: *Cellij* can incorporate metadata, such as spatial or temporal dependencies between the samples, using Gaussian Processes to structure and align the latent space.

2. Related Work

FA is a widely-used approach for integrating and analysing omics datasets (Thurstone, 1931). Conventional FA is limited to modelling observations from a single data view. In order to cope with multi-view datasets, extensions such as canonical correlation analysis (CCA) (Hotelling, 1936), (Klami et al., 2013) and group factor analysis (GFA) (Klami et al., 2015) have been introduced. These methods simultaneously model paired observations across multiple view, capturing their linear dependencies. Factor loadings play a crucial role in model interpretation, encoding the structure of each factor. They allow for the introduction of statistical assumptions, such as sparsity. GFA extends the automatic relevance determination technique (MacKay, 1994) to quantify the association between view and factor.
Non-Bayesian approaches commonly handle sparsity by introducing additional terms to the optimisation objective, such as the L1 penalty in LASSO (Tibshirani, 1996). Bayesian approaches, on the other hand, achieve sparse solutions through sparsity-inducing priors. Examples include the double exponential or Laplace prior, which is the Bayesian counterpart of the LASSO (Park & Casella, 2008b), and the discrete spike-and-slab (SnS) prior (Mitchell & Beauchamp, 1988), a mixture of a Dirac delta distribution and a normal distribution. Recently, the spike-and-slab LASSO (SSL) (Ročková & George, 2018) has emerged as a combination of two Laplace distributions, emulating both the spike and the slab components. Other shrinkage priors, such as the horseshoe prior (Carvalho et al., 2009) (HS), offer a continuous relaxation of the spike-and-slab approach. Several Bayesian approaches (Engelhardt & Stephens, 2010; Lan et al., 2014; Buettner et al., 2017) successfully combine latent variable models with sparsity-inducing priors. In the multi-view setting, Zhao et al. (2016) propose a hierarchical Bayesian GFA with structured sparsity, facilitated by a cascading three-parameter Beta prior (Armagan et al., 2011). This addition supports column-wise sparsity for inferring associations between views and element-wise sparsity for feature selection within factors. Multi-omics factor analysis (MOFA) (Argelaguet et al., 2018) assumes similar structured sparsity levels by combining automatic relevance determination (ARD) (MacKay, 1994) with a spike-and-slab prior. Tab. B compares *Cellij* to previously published methods.

3. Methods

As shown in (Argelaguet et al., 2020), we can extend traditional FA to a multi-view and multi-group setting. First we can divide the input $\mathbf{X} \in \mathbb{R}^{D \times N}$ into M views with N samples each, $\mathbf{X}^m \in \mathbb{R}^{D_m \times N}$. Each view consists of non-overlapping features that represent different assays. Applying the same reasoning, we can break the assumption of independent samples by separating the N samples into G groups of size N_g, obtaining $\mathbf{X}^{mg} \in \mathbb{R}^{N_g \times D_m}$.

3.1. Factor Analysis

Under a distribution Φ_θ, FA models the expectation of the data as a linear decomposition of K unobserved factors

$$\mathbf{X}^{mg} \sim \Phi_\theta(\mathbf{X}^{mg} \mid \mathbf{W}^m \mathbf{Z}^g) \tag{1}$$

where $\mathbf{Z}^g \in \mathbb{R}^{K \times N^g}$ denotes the latent factors, and $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ corresponds to the factor loadings. For instance, under the assumption of a normal distribution $\Phi_\theta = \mathcal{N}(\mathbf{\Lambda}, \mathbf{\Sigma})$, $\mathbf{\Lambda} = \mathbf{W}^m \mathbf{Z}^g$ and $\theta = \mathbf{\Sigma}$. The number of factors is K with $K \ll D^m \, \forall \, m$. The choice of the likelihood is data-specific and can vary across views.

3.2. Sparsity and Shrinkage Priors for Feature Selection

Sparsity-inducing priors enforce zero values for many model parameters, resulting in models with fewer active variables. Shrinkage priors, on the other hand, reduce the magnitude of the estimated parameters towards zero, but may not enforce exact sparsity. *Cellij* can impose sparsity both on $\mathbf{W}^m$ and $\mathbf{Z}^g$ leading to more interpretable results. This reduces the number of potential solutions to the optimisation problem and helps select an appropriate number of factors (see 4.1).

Despite their benefits, both prior approaches also pose challenges, such as the choice of the prior hyperparameters and the need for efficient inference algorithms. For a more fluent reading, we use the term sparsity priors for both sparsity-inducing and shrinkage priors from now.
We provide an overview of *Cellij*'s priors in App. F.

3.3. Inducing Structure in the Latent Space using GPs

Most FA models assume independence among samples, which is not true when there is a temporal or spatial arrangement to them. MEFISTO (Velten et al., 2022) allows modelling high-dimensional data in the presence of known spatial or temporal dependencies, enabling spatio-temporally informed dimensionality reduction, interpolation, and differentiation between smooth and non-smooth patterns of variation. In a similar fashion, we use Gaussian Processes (GP) for structuring the latent space based on these known covariates. We adopt the MEFISTO approach, i.e. one can associate any number of covariates with each factor k, where $\mathbf{C}^g \in \mathbb{R}^{C \times N_g}$ is the matrix of covariates, resulting in

$$z^g_{kn} = f_k(\mathbf{c}^g_n) + \epsilon^g_{kn} \qquad \text{with } f_k \sim \mathrm{GP}(0, \kappa_k) \qquad (2)$$

The user can chose a suitable kernel function from a predefined set or provide his own, e.g., to mix kernels.
Cellij extend the generic group factor analysis models in a module and flexible manner, enabling to mix and match custom-defined data likelihood, sparsity-inducing priors as well as imposing structured priors on the latent factors.

3.4. Optimisation

To approximate the posterior distributions over variables, we use stochastic variational inference (Hoffman et al., 2013) to optimise for the evidence lower bound. In addition, since *Cellij* is a Bayesian framework, we estimate uncertainties in our model parameters – indicating how reliable the results are. The framework is developed in Python, relying on Pyro (Bingham et al., 2019) and GPyTorch (Gardner et al., 2018), and will be open-sourced on GitHub.

4. Results

Leveraging the flexibility of *Cellij*, we applied the framework to different synthetic and real-data benchmarks.

4.1. Benchmarking Sparsity and Shrinkage Priors

The choice of the prior distribution for variable selection can greatly impact the accuracy and interpretability of the resulting estimates. Here, we assess common sparsity priors to regularise the FA models. Full details on the specific sparsity prior distributions can be found in App. F.

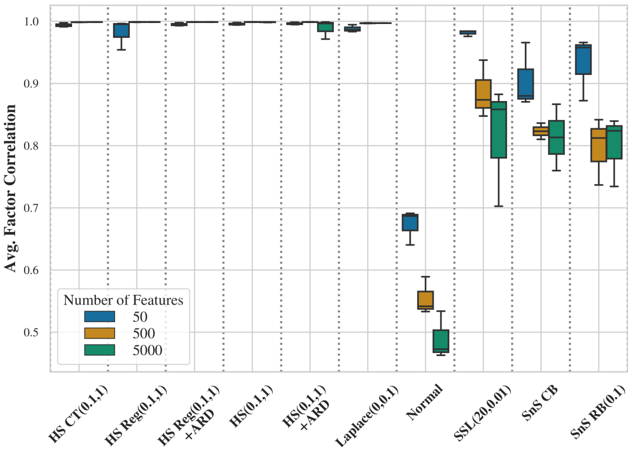

Figure 2. Assessment of FA with alternative sparsity priors in terms of the accuracy of latent factor recovery. Shown are average correlation of top 10 factors with ground truth data across multiple experiments. For details on the sparsity prior employed, see App.F.

Reconstruction of simulated latent factors: We assess the ability of FA models to reconstruct the underlying latent factors on synthetic data (c.f. App. D), by quantifying the correlation coefficient between the simulated and the inferred factor scores. We train several FA models using $K = 20$ factors, where only 10 are active according to the data generation process. Next, we select the top 10 inferred factors with the highest correlation to the 10 ground truth factors (Fig. 2). Overall, the HS prior and its derivatives offered the most robust recovery performance across all feature sizes. The best overall performance was achieved by a HS prior with a constant value of $\tau^m = 0.1$. The SnS priors performed worse compared to the HS priors and show a decreasing performance with growing feature sizes.
Surprisingly, Laplace priors shows a strong performance across large feature dimensions. Dense models without any sparsity assumption achieved the lowest performance.

Estimation of the true number of latent factors: As a second evaluation metric, we consider the selection of the most appropriate number of active factors. Using the same setup as in the previous experiment, we rank the relevance of individual learnt factors using the ℓ_2-norm as a proxy for it's relevance. The results are summarised in Fig. 4. The SnS flavours and the Regularised Horseshoe provide a clear cut-off in factor activity. Again, the Laplace prior exhibits good results as well. As expected, the baseline Normal prior results in a large number of active features across all factors. On the other hand, the Horseshoe priors deactivate factors reliably for small feature sizes, but suffer on higher dimensions. Including a factor-specific variable (+ARD) deteriorates the results, potentially due to a much harder inference.

Modelling the underlying sparsity of factor loadings We evaluate how well each prior models the true underlying sparsity, while maintaining a low reconstruction error. We first compute binary scores between the inferred and the true factor loadings, based on an optimal threshold for each prior assumption. Next, we assess the quality of the decomposition by computing the RMSE between the observed and the reconstructed data. We report the main results in Tab. 4.1. For a more comprehensive summary, see App. C. Similar to previous results, the HS derivatives and the Laplace prior exhibit a significantly higher F1 score compared to the SnS priors. As expected, the Normal prior is unable to capture the underlying sparse structure. However, all methods perform equally well in terms of faithfully modelling the data, as supported by nearly identical RMSE.

Table 1. Performance comparison on synthetic data.

PRIOR	F1	RMSE
HS CT(0.1,1)	0.993 ± 0.001	0.304 ± 0.0021
HS REG(0.1,1)	0.994 ± 0.002	0.305 ± 0.0023
HS(0.1,1)	0.993 ± 0.002	0.305 ± 0.0021
LAPLACE(0,0.1)	0.989 ± 0.003	0.304 ± 0.0019
NORMAL	0.485 ± 0.036	0.302 ± 0.0023
SNS CB	0.668 ± 0.069	0.306 ± 0.0009
SNS RB(0.1)	0.706 ± 0.120	0.307 ± 0.0020
SSL RB(20,0.01,0.1)	0.690 ± 0.054	0.322 ± 0.0020

Assessment of data imputation on a CLL dataset: Finally, we study the influence of alternative priors on the model performance for reconstructing missing data. We applied *Cellij* to a chronic lymphocytic leukaemia (CLL) dataset, which combined ex vivo drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays (Dietrich et al., 2018). We then introduced missing values at random across four views and predicted the missing drug response data. The results indicate that models trained with a HS prior provide the best predictive performance, followed by the Laplace and the Normal prior, and with a larger gap the SnS priors.

Table 2. RMSE on drug response prediction.

	MISSING FRACTION		
PRIOR	30%	50%	70%
HS CT(0.1,1)	0.108	0.112	0.117
LAPLACE(0,0.1)	0.110	0.113	0.119
NORMAL	0.109	0.117	0.121
SNS CB	0.117	0.119	0.124

4.2. Structuring the Latent Space with GPs

To illustrate the benefits of this, we have trained a *Cellij* model with 2 factors on mouse blastocyst developmental data from (Guo et al., 2010), encoding the cell divisons from the 1-cell stage to the 64-cell stage as a covariate. Fig. 3 shows the loadings of two factors. We can clearly see that utilising the covariate structures the latent space. This allows to investigate which features associate with the factors than reflect the covariate which aids interpretability.

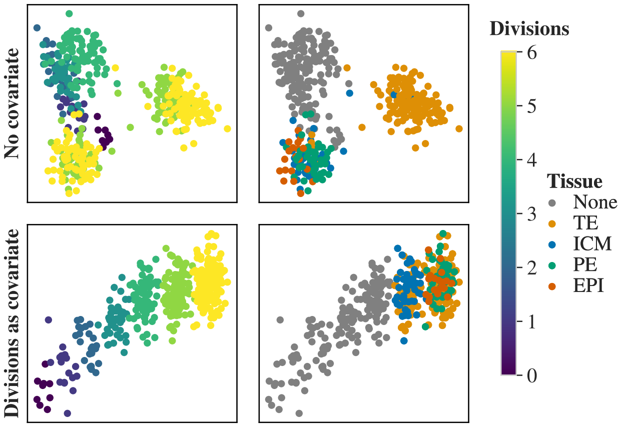

Figure 3. Loadings of the 2 latent factors plotted against each other coloured by divisions and tissue type. Top row: training without covariates, bottom row: training with time covariate.

5. Discussion and Future Work

This paper presents our flexible FA-based framework, *Cellij*. *Cellij* facilitates usability and rapid prototyping of scalable and interpretable FA models. Experiments on synthetic data using different shrinkage and sparsity priors reveal that HS priors consistently recover the underlying sparse structure, while maintaining a low reconstruction error. In addition, the HS priors provide the best results in the imputation task. On the other hand, the SnS priors robustly pinpoint relevant factors in the presence of redundancy. Finally, the Laplace prior performs surprisingly well across a large set of experiments and glances with its simplicity. Furthermore, we showed that *Cellij* can exploit additional covariate information via GPs – allowing to structure the latent space and associating covariates with gene sets. Moving forward, there are several areas that can be explored building upon the current state of the framework.

Domain Knowledge Integration One direction is to incorporate additional biological prior knowledge from graph structures, including gene sets and pathways. These functional relationships can be used in a similar vein as in MuVI (Qoku & Buettner, 2023), thereby extending the MuVI approach to a broader range of models.

Automation To reduce the amount of manual labour required to find the optimal setting for a problem-specific FA model, we plan generalise *Cellij* towards an *Automated Bioinformatician*, similar to ideas presented in the *Automated Statistician* (Steinruecken et al., 2019). Specifically, we aim to automate model configuration, e.g., number of latent factors, prior choices and parameter settings.

References

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. 14(6): e8124, 2018. ISSN 1744-4292. doi: 10.15252/msb.20178124. URL https://www.embopress.org/doi/full/10.15252/msb.20178124. Publisher: John Wiley & Sons, Ltd.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. 21(1):111, 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02015-1. URL https://doi.org/10.1186/s13059-020-02015-1.

Armagan, A., Clyde, M., and Dunson, D. Generalized beta mixtures of gaussians. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/ad972f10e0800b49d76fed33a21f6698-Abstract.html.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian statistics*, 7:733–742, 2003.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. The horseshoe+ estimator of ultra-sparse signals, 2015. URL http://arxiv.org/abs/1502.00560.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. 2019.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18: 1–13, 2017.

Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling sparsity via the horseshoe. 2009.

Carvalho, C. M., Polson, N. G., and Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika*, 97(2): 465–480, 2010.

Dietrich, S., Oleś, M., Lu, J., Sellner, L., Anders, S., Velten, B., Wu, B., Hüllein, J., Liberio, M. d. S., Walther, T., Wagner, L., Rabe, S., Ghidelli-Disse, S., Bantscheff, M., Oleś, A. K., Słabicki, M., Mock, A., Oakes, C. C., Wang, S., Oppermann, S., Lukas, M., Kim, V., Sill, M., Benner, A., Jauch, A., Sutton, L. A., Young, E., Rosenquist, R., Liu, X., Jethwa, A., Lee, K. S., Lewis, J., Putzker, K., Lutz, C., Rossi, D., Mokhir, A., Oellerich, T., Zirlik, K., Herling, M., Nguyen-Khac, F., Plass, C., Andersson, E., Mustjoki, S., Kalle, C. v., Ho, A. D., Hensel, M., Dürig, J., Ringshausen, I., Zapatka, M., Huber, W., and Zenz, T. Drug-perturbation-based stratification of blood cancer. 128(1):427–445, 2018. ISSN 0021-9738. doi: 10.1172/JCI93801. URL https://www.jci.org/articles/view/93801. Publisher: American Society for Clinical Investigation.

Engelhardt, B. E. and Stephens, M. Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. 6(9):e1001117, 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001117. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001117. Publisher: Public Library of Science.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. 2018.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4): 675–685, April 2010. ISSN 1878-1551. doi: 10.1016/j.devcel.2010.02.012.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. 14(40):1303–1347, 2013. ISSN 1533-7928. URL http://jmlr.org/papers/v14/hoffman13a.html.

Hotelling, H. Relations between two sets of variates. 28(3):321–377, 1936. ISSN 0006-3444. doi: 10.2307/2333955. URL https://www.jstor.org/stable/2333955. Publisher: [Oxford University Press, Biometrika Trust].

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Klami, A., Virtanen, S., and Kaski, S. Bayesian canonical correlation analysis. 14(30):965–1003, 2013. ISSN 1533-7928. URL http://jmlr.org/papers/v14/klami13a.html.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. Group factor analysis. 26(9):2136–2147, 2015. ISSN 2162-2388. doi: 10.1109/TNNLS.2014.2376974.

Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. Sparse factor analysis for learning and content analytics. 15(57):1959–2008, 2014. ISSN 1533-7928. URL http://jmlr.org/papers/v15/lan14a.html.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. 401(6755):788–791. ISSN 1476-4687. doi: 10.1038/44565. URL https://www.nature.com/articles/44565. Number: 6755 Publisher: Nature Publishing Group.

Loaiza-Ganem, G. and Cunningham, J. P. The continuous bernoulli: fixing a pervasive error in variational autoencoders, 2019. URL http://arxiv.org/abs/1907.06845.

MacKay, D. J. C. Bayesian nonlinear modeling for the prediction competition. 1994. ISSN 0001-2505. URL https://www.osti.gov/biblio/33309. Number: CONF-9406105- Publisher: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA (United States).

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL http://arxiv.org/abs/1611.00712.

Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression. 83(404):1023–1032, 1988. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478694. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694.

Park, T. and Casella, G. The bayesian lasso. 103 (482):681–686, 2008a. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214508000000337. URL https://www.tandfonline.com/doi/full/10.1198/016214508000000337.

Park, T. and Casella, G. The bayesian lasso. 103 (482):681–686, 2008b. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214508000000337. URL https://www.tandfonline.com/doi/full/10.1198/016214508000000337.

Piironen, J. and Vehtari, A. Sparsity information and regularization in the horseshoe and other shrinkage priors. 11 (2), 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1337SI. URL http://arxiv.org/abs/1707.01694.

Qoku, A. and Buettner, F. Encoding domain knowledge in multi-view latent variable models: A bayesian approach with structured sparsity. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11545–11562. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/qoku23a.html.

Ročková, V. and George, E. I. The spike-and-slab LASSO. 113(521):431–444, 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2016.1260469. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1260469.

Steinruecken, C., Smith, E., Janz, D., Lloyd, J., and Ghahramani, Z. The automatic statistician. In Hutter, F., Kotthoff, L., and Vanschoren, J. (eds.), *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pp. 161–173. Springer International Publishing, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5_9. URL https://doi.org/10.1007/978-3-030-05318-5_9.

Thurstone, L. L. Multiple factor analysis. *Psychological review*, 38(5):406, 1931.

Tibshirani, R. Regression shrinkage and selection via the lasso. 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x.

Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. 19(2): 179–186, 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01343-9. URL https://www.nature.com/articles/s41592-021-01343-9. Number: 2 Publisher: Nature Publishing Group.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. Bayesian group factor analysis with structured sparsity. 17(196):1–47, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-472.html.

A. Abbreviations

Abbreviation	Description	Equation
HS(a, b)	Horseshoe$(\sigma_\tau = a, \sigma_\lambda = b)$	4
HS CT(a, b)	Horseshoe$(\sigma_\lambda = b)$ with const. τ, i.e. $\tau^m = a$	5
HS Reg(a, b)	Regularised Horseshoe$(\sigma_\tau = a, \sigma_\lambda = b)$	7
HS+(a, b)	Horseshoe-Plus$(\tau = a, \sigma_\eta = b)$	8
SnS CB	Spike-and-Slab + Continuous Bernoulli	11
SnS RB(t)	Spike-and-Slab + Relaxed Bernoulli with temperature a	
SSL(a, b, t)	Spike-and-Slab Lasso with inverse scales $\lambda_0 = a, \lambda_1 = b$ + Relaxed Bernoulli with temperature t	12

If the prior names contains an additional "+ARD", we add a factor-specific variable $\delta_k^m \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$ to the model, as described in Eq. 6. In case of the SnS RB and SSL prior, we often drop the parameter t in the abbreviation, which is a constant $t = 0.1$ across all experiments.

B. Existing Factor Analysis Models

Model	Likelihoods	Sparse Priors	Multi-Views	Multi-Groups	Linear	Missings
Cellij	Arbitrary	HS, SnS, NN, Lap	√	√	√	√
FISHFactor	Poisson point process	NN	×	√	√	×
MOFA	Gauss, Poi, Bern	SnS	√	×	√	√
MOFA+	Gauss, Poi, Bern	SnS	√	√	√	√
MEFISTO	Gauss, Poi, Bern	SnS	√	√	√	√
f-scLVM	Gauss, Poi, Bern	SnS	×	×	√	√
MuVI	Gauss, Poi, Bern	HS	√	×	√	√
ZIFA	Gauss	custom	×	×	√	×
ZINB-WaVE	Negbinom, Poi	custom	×	×	√	×
GLM-PCA	Negbinom, Poi	×	×	×	√	√
OJSNMF	Gauss, Poi, Exp	custom	√	×	√	×
NSF	Negbinom, Poi	NN	×	×	√	√

Columns: (i) *Likelihoods* (ii) *Sparsity Priors*: all pre-implemented sparsity/shrinkage priors (HS: Horseshoe, SnS: Spike-and-Slab, NN: Non-Negativity, Lap: Laplace), (iII) *Multi-Views*: Support for multiple views, (iv) *Multi-Groups*: Support for multiple groups, (v) *Linear*: Linear model (vi) *Missings*: Can the model handle missing data

C. Performance on Synthetic Data

Prior	RMSE	F1	Precision	Recall
HS CT(0.1,1)	0.304 ± 0.0021	0.993 ± 0.001	0.996 ± 0.002	0.992 ± 0.001
HS Reg(0.1,1)	0.305 ± 0.0023	0.994 ± 0.002	1.000 ± 0.000	0.990 ± 0.002
HS Reg(0.1,1)+ARD	0.306 ± 0.0022	0.994 ± 0.001	0.999 ± 0.001	0.992 ± 0.000
HS(0.1,1)	0.305 ± 0.0021	0.993 ± 0.002	0.998 ± 0.001	0.990 ± 0.001
HS(0.1,1)+ARD	0.306 ± 0.0021	0.992 ± 0.003	0.994 ± 0.005	0.993 ± 0.001
Laplace(0,0.1)	0.304 ± 0.0019	0.989 ± 0.003	0.992 ± 0.003	0.986 ± 0.003
Normal	0.302 ± 0.0023	0.485 ± 0.036	0.386 ± 0.038	0.731 ± 0.040
SnS CB	0.306 ± 0.0009	0.668 ± 0.069	0.561 ± 0.078	0.826 ± 0.040
SnS RB(0.1)	0.307 ± 0.0020	0.706 ± 0.120	0.635 ± 0.148	0.826 ± 0.055
SSL(20,0.01,0.1)	0.322 ± 0.0020	0.690 ± 0.054	0.662 ± 0.081	0.731 ± 0.018

First value describes the average metric score, the second value its standard deviation.

D. Synthetic Data Generation Process

We compile several synthetic datasets of $N = 200$ samples across three views, each comprising between $D_m = 50$ and $D_m = 10,000$ features ($m \in \{1, 2, 3\}$), with the latter to further emphasise the $N \ll D$ settings (Bernardo et al., 2003), e.g. in gene expression data. The latent space $\mathbf{Z}$ consists of $K = 10$ factors, that are linearly transformed by a set of sparse factor loadings $\mathbf{W}^m$. Each weight $w_{dk}^m \sim \mathcal{N}(0, 1)$ is sampled independently from a standard normal distribution. To introduce sparsity we randomly set 85%-95% of the weights, as well as weight loadings with an absolute value of less than 0.1 to practically zero, i.e. to a very small random term $\epsilon \sim \mathcal{N}(0, 0.01)$.

E. Training Procedure

All models in the studies have been trained with at least three different seeds. We set a global maximum of 25'000 epochs and stopped the training earlier if the ELBO was not reduced by at least 0.01% within 500 epochs.
In experiment 4.1 we generated data with 10 active factors for each of the three views with 200 samples each. The number of features was varied between 50 and 10000. We trained each model using three different seeds and three different learning rates (0.1, 0.01, 0.001).

F. Sparsity and Shrinkage Priors

F.1. Horseshoe Priors

The Horseshoe prior was introduced in (Carvalho et al., 2010) and employs a bivariate-normal distribution. According to our model in Eq. 1, we have

$$p(w_{dk}^m \mid \tau^m, \lambda_{dk}^m) = \mathcal{N}(0, (\tau^m)^2 \cdot (\lambda_{dk}^m)^2), \tag{3}$$

where τ^m is a view-specific global shrinkage parameter defining the general level of sparsity in view m, and λ_{dk}^m is a local shrinkage parameter allowing each element in $\mathbf{W}^m$ to escape the global sparsity. In general λ_{dk}^m is sampled from a Half-Cauchy distribution $\lambda_{dk}^m \sim \mathcal{C}^+(0, \sigma_\lambda)$ whereas different approaches for τ^m exists. In the simplest form, we have $\tau^m = c^m$ with a view-specific constant c^m. In a different setting, one samples τ^m from a Half-Cauchy distribution $\tau^m \sim \mathcal{C}^+(0, \sigma_\tau)$. In summary we have the following HS variations

$$\tau^m \sim \mathcal{C}^+(0, 1), \qquad \lambda_{dk}^m \sim \mathcal{C}^+(0, 1) \tag{4}$$

or

$$\tau^m = \text{const.}(m), \qquad \lambda_{dk}^m \sim \mathcal{C}^+(0, 1), \tag{5}$$

where we set $\sigma_\lambda = \sigma_\tau = 1$. In addition to λ and θ, one can add a factor-specific variable to estimate the factor's general level of *activeness*. This renders Eq. 3 into

$$p(w_{dk}^m \mid \tau^m, \lambda_{dk}^m, \delta_k^m) = \mathcal{N}(0, (\tau^m)^2 \cdot (\delta_k^m)^2 \cdot (\lambda_{dk}^m)^2). \tag{6}$$

We use a fixed hyperprior on $\delta_k^m \sim \text{Beta}(0.5, 0.5)$. Low values of δ_k^m indicate that the factor k is not active in view m.

The *regularized horseshoe prior* introduced in (Piironen & Vehtari, 2017) guarantees that the prior always shrinks the coefficients at least by a small amount towards zero. We have

$$\begin{aligned} \lambda_{dk}^m &= \sqrt{\frac{(\alpha^m)^2 (\tilde{\lambda}_{dk}^m)^2}{(\alpha^m)^2 + (\tau^m \cdot \tilde{\lambda}_{dk}^m)^2}} \\ \tilde{\lambda}_{dk}^m &\sim \mathcal{C}^+(0, 1) \\ \tau^m &\sim \mathcal{C}^+(0, 1) \\ a^m &\sim \text{Inv.-Gamma}(0.5, 0.5). \end{aligned} \tag{7}$$

Last, we want to introduce the Horseshoe-Plus prior (Bhadra et al., 2015), used to model ultra-sparse signals. It follows a hierarchical approach defined as

$$\begin{aligned} p(w_{dk}^m \mid \tau^m, \eta_{dk}^m, \lambda_{dk}^m) &= \mathcal{N}(0, (\lambda_{dk}^m)^2) \\ \lambda_{dk}^m &\sim \mathcal{C}^+(0, \tau^m \cdot \eta_{dk}^m) \\ \eta_{dk}^m &\sim \mathcal{C}^+(0, 1) \\ \tau^m &= \text{const.}(m), \end{aligned} \tag{8}$$

where we face an additional Half-Cauchy mixing variable $\eta_{dk}^m \sim \mathcal{C}^+(0, \sigma_\eta)$. One has the option of using either a standard Half-Cauchy prior or a Uniform(0, 1) prior for τ^m to obtain a comprehensive Bayesian framework, but we leave this for future work.

F.2. Spike-and-Slab Priors

In contrast to shrinkage priors, the Spike-and-Slab prior is a true sparsity prior allowing for variable selection by assigning either a zero or non-zero coefficient to each predictor. It has a two-component structure with a spike at zero, promoting sparsity, and a slab component with non-zero values, allowing for variable inclusion.

$$p(w_{dk}^m \mid \gamma_{dk}^m, \tau_k^m) = \gamma_k^m \delta_0(w_{dk}^m) + (1 - \gamma_k^m) N(w_{dk}^m \mid 0, 1/(\tau_k^m)^2) \tag{9}$$

where $\delta_0(\cdot)$ is a point mass at zero, $N(w_{dk}^m \mid 0, 1/(\tau_k^m)^2)$ is a normal distribution with zero mean and precision τ_k^m, γ_k^m is either an indicator variable or a weighting between spike and slab, i.e. between 0 and 1.

However, the Delta distribution at 0 makes inference less straightforward. To address this issue, we use a re-parameterisation technique as presented in (Argelaguet et al., 2018), where the weights are expressed as the product of two random variables following a Gaussian distribution and a Bernoulli distribution

$$p(\hat{w}_{dk}^m, \hat{s}_{dk}^m) = N(\hat{w}_{dk}^m \mid 0, 1/(\tau_k^m)^2) \text{Ber}(\hat{s}_{dk}^m \mid \theta_k^m) \tag{10}$$

Typically ones placesa Beta-prior on $\theta_k^m \sim \text{Beta}(\alpha_\theta^m, \beta_\theta^m)$ and a Gamma-prior on the precision $\tau_k^m \sim \text{Gamma}(\alpha_\tau^m, \beta_\tau^m)$.

To allow for a continuous optimisation, we relax the discrete Bernoulli distribution with two possible continuous approximations. On the one hand, we make use of the Continuous-Bernoulli as presented in (Loaiza-Ganem & Cunningham, 2019)

$$p(\hat{s}_{dk}^m \mid \lambda^m) = C(\lambda^m)(\lambda^m)^{\hat{s}_{dk}^m}(1 - \lambda^m)^{1-\hat{s}_{dk}^m} \qquad \text{with} \quad C(\lambda^m) = \begin{cases} 2 & \text{if } \lambda^m = \frac{1}{2} \\ \frac{2\tanh^{-1}(1-2\lambda^m)}{1-2\lambda^m} & \text{otherwise} \end{cases} \tag{11}$$

for $\lambda^m \in (0, 1)$. We place a hyperprior on λ^m with $\lambda^m \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$. On the other hand, we make use of the Relaxed-Bernoulli with a straight-through gradient estimator (RBSTG) (Maddison et al., 2017). The RBSTG can be seen as a bivariate Gumbel-Softmax distribution (Jang et al., 2016) with a temperature parameter that relaxes categorical random variables. In our setting we use a temperature parameter of $t = 0.1$.

The *Spike-and-Slab Lasso* (Ročková & George, 2018) prior is an approach to create a continuum between the penalised LASSO and the Bayesian point-mass spike-and-slab formulations. Compared to Eq. 9 the Delta and Normal distribution are replaced with two Laplace distributions, with $\mu = 0$ and different scales λ_0, λ_1. As used in the original paper, for the slab we use a scale value of $\lambda_0 = 20$ and for the spike $\lambda_1 = 0.01$ across all views.

$$p(w_{dk}^m \mid \gamma_{dk}^m, \lambda_0, \lambda_1) = \gamma_k^m \psi(w_{dk}^m \mid \lambda_0^m) + (1 - \gamma_k^m) \psi(w_{dk}^m \mid \lambda_1^m) \tag{12}$$

with $\psi(w_{dk}^m \mid \lambda^m) = \frac{\lambda^m}{2} e^{-\lambda^m |w_{dk}^m|}$. Again, we combine Eq. 12 with a relaxed Bernoulli distribution.

F.3. Bayesian LASSO

The Bayesian Lasso (Least Absolute Shrinkage and Selection Operator) prior, first introduced in (Park & Casella, 2008a), has emerged as a powerful tool for variable selection and regularisation in Bayesian statistics. It introduces a Laplace (double exponential) distribution as a prior distribution for the coefficients

$$p(w_{dk}^m \mid \sigma, \lambda) = \frac{\lambda}{2\sigma} e^{-\lambda |\beta_{dk}^m| / \sigma}$$

Instead of imposing another hyperprior on σ, we choose a fixed $\sigma = 1$. However, instead of the true Laplace distribution, we make use of the Soft-Laplace, a smooth distribution with Laplace-like tail behavior, which is infinitely differentiable everywhere.

F.4. Non-Negativity

An additional way to achieve sparsity is using non-negativity constraints (Lee & Seung), i.e.

$$p(q_{dk}^m) = \mathcal{N}(\mu_p^m, (\sigma_p^m)^2)$$
$$p(w_{dk}^m) = g^{-1}(q_{dk}^m)$$

where g^{-1} is an inverse link function, such as softplus, ReLU or exponential. The mean and variance of the Normal distribution are constant across all elements in $\mathbf{W}$ in our experiments. For simplicity, we choose softplus and define $\mu_p = 0$ and $\sigma_p = 1$.

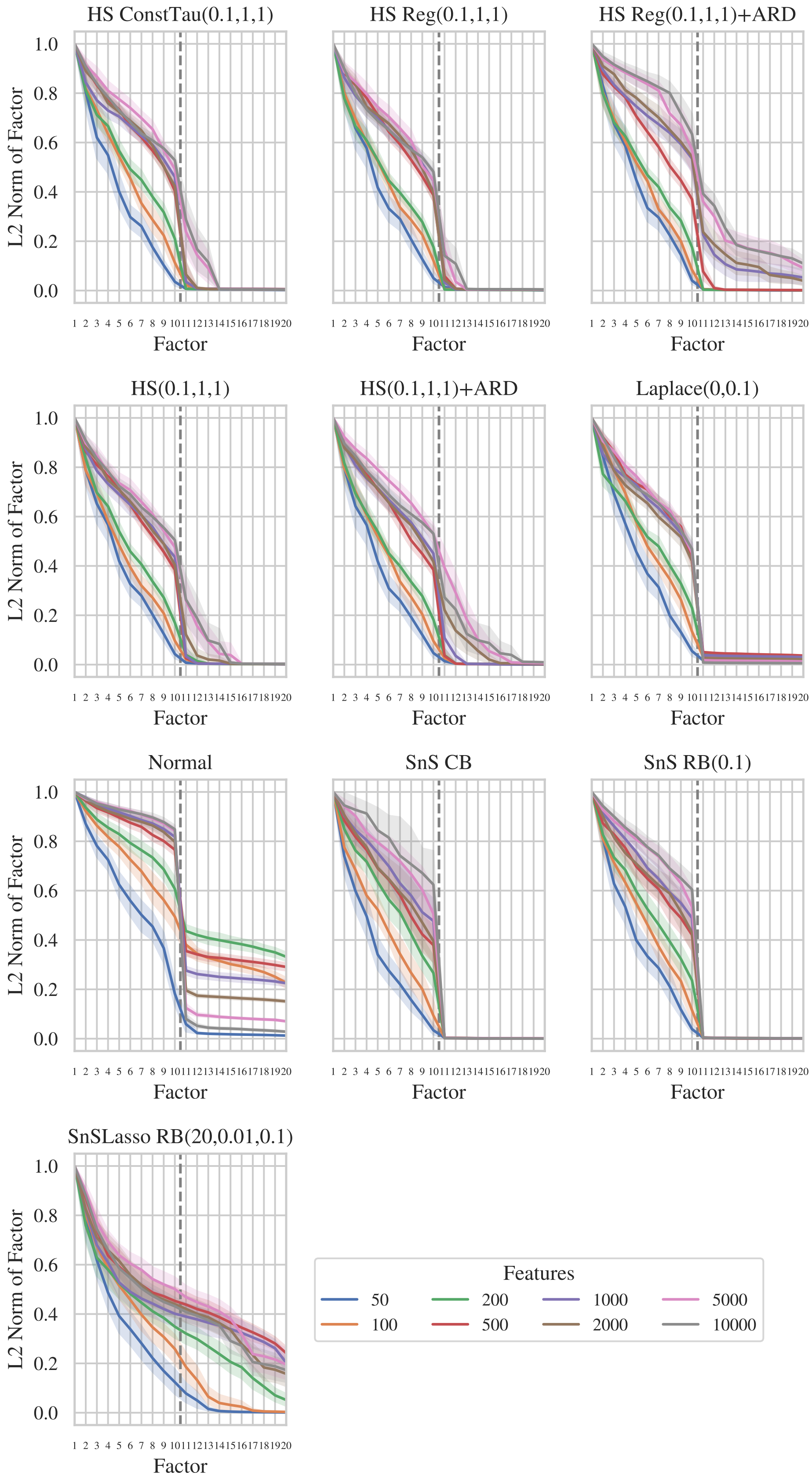

Figure 4. An overview of factor *activeness* measured using the ℓ_2-norm of each factor. Different colours represent different feature sizes on a dataset with three views and 200 samples. Data was generated with 10 active factors and estimated with 20 factors.