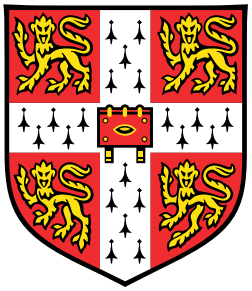
Statistical framework for the integration of single-cell multi-omics data sets



Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2019

Abstract

Single-cell profiling techniques have provided an unprecented opportunity to study cellular heterogeneity at multiple molecular levels. The maturation of single-cell RNA-sequencing technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment. This represents a remarkable advance over traditional bulk sequencing methods, particularly for the study of complex and heterogeneous biological processes, including the immune system, embryonic development and cancer. However, the accompanying epigenetic changes and the role of other molecular layers in driving cell fate decisions remains poorly understood. The profiling the epigenome at the single-cell level is receiving increasing attention. However, without associated transcriptomic readouts, the conclusions that can be extracted from epigenetic measurements are limited.
More recently, technological advances enabled multiple biological layers to be probed in parallel at the single-cell level, unveling a powerful approach for investigating regulatory relationships. Such single-cell multi-modal technologies can reveal multiple dimensions of cellular heterogeneity and uncover how this variation is coupled between the different molecular layers, hence enabling a more profound mechanistic insight than can be inferred by analysing a single data modality in separate cells. Yet, multi-modal sequencing protocols face multiple challenges, both from the experimental and the computational front.

In this thesis we propose an experimental methodology and a computational framework for the integrative study of multiple omics in single cells.
The first contribution of this thesis is Nucleosome, Methylome and Transcriptome sequencing (scNMT-seq), a multi-modal single-cell sequencing protocol for profiling RNA expression, DNA methylation and chromatin accessibility in single cells. scNMT-seq provides genome-wide epigenetic readouts at a base-pair resolution, hence expanding our ability to investigate the dynamics of the epigenome across cell fate transitions.
The second contribution of this thesis is Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of large-scale multi-omics data sets. MOFA aims at discovering the principal sources of variation while disentangling the axes of heterogeneity that are shared across multiple modalities from those specific to individual data modalities. This framework enables the unbiased interrogation of large (single-cell) data sets simultaneously across multiple data modalities and across different experiments or conditions.
The third contribution of this thesis is generation of an epigenetic roadmap of mouse gastrulation, resulting from the combined use of scNMT-seq and MOFA. Notably, we show that regulatory elements associated with the formation of the three germ layers are either epigenetically primed or epigenetically remodelled prior to overt cell fate decisions, providing the molecular logic for a hierarchical emergence of the primary germ layers.

0.1 Multi-Omics Factor Analysis: a framework for unsupervised integration of multi-omics data sets

The work described in this chapter results from a collaboration with the Multi-omics and statistical computing group lead by Wolfgang Huber at the EMBL (Heidelberg, Germany). It has been peer-reviewed and published in Argelaguet & Velten et al [**Argelaguet2018**].

The method was conceived by Florian Buettner, Oliver Stegle and me. I performed most of the mathematical derivations and implementation, but with significant contributions from Damien Arnol and Britta Velten. The single-cell application was led by me whereas the CLL data application was led by Britta Velten, but with joint contributions in either cases. Florian Buettner, Wolfgang Huber and Oliver Stegle supervised the project.
The article was jointly written by Britta Velten, Florian Buettner, Wolfgang Huber, Oliver Stegle and me.

0.1.1 Model description

MOFA is a multi-view generalisation of traditional Factor Analysis to M input matrices (or views) $\mathbf{Y}^m \in \mathbb{R}^{N\times D_m}$ based on the framework of Group Factor Analysis (discussed in Section X).
The input data consists on M views with non-overlapping features that often represent different assays. However, there is flexibility in the definition of views and they can be tailored to address different hypothesis. Formally, the input data is factorised as:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\varepsilon}^m \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times K}$ is a matrix that contains the factor values and $\mathbf{W}^m \in \mathbb{R}^{D_m\times K}$ are M matrices that contain the loadings that relate the high-dimensional space to the low-dimensional latent representation. Finally, $\boldsymbol{\varepsilon}^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic:

$$p(\boldsymbol{\varepsilon}_d^m) = \mathcal{N}\left(\boldsymbol{\varepsilon}_d^m \,|\, 0, 1/\tau_d^m\right) \tag{2}$$

Non-gaussian noise models can also be defined and is discussed in Section XX. Unless otherwise stated, we will always assume Gaussian noise.
Altogether, this results in the following likelihood:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{Z},\mathrm{T}) = \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{n=1}^{N}\mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_n^T\mathbf{w}_d^m, 1/\tau_d^m\right) \tag{3}$$

Interpretation of the factors

Each factor ordinates cells along a one-dimensional axis centered at zero. Samples with different signs indicate opposite phenotypes, with higher absolute value indicating a stronger effect. For example, if the k-th factor captures the variability associated with cell cycle, we could expect cells in the Mitosis state to be at one end of the factor (irrespective of the sign, only the relative positioning being of importance). In contrast, cells in G1 phase are expected to be at the other end of the factor. Cells with intermediate phenotype, or with

no clear phenotype (i.e. no cell cycle genes profiled), are expected to be located around zero, as specified by the prior distribution.

Interpretation of the loadings

The loadings provide a score for each gene on each factor, and are interpreted in a similar way as the factors. Genes with no association with the factor are expected to have values close to zero, as specified by the prior. In contrast, genes with strong association with the factor are expected to have large absolute values. The sign of the loading indicates the direction of the effect: a positive loading indicates that the feature is more active in the cells with positive factor values, and viceversa.
Following the cell cycle example from above, we expect genes that are upregulated in the M phase to have large positive loadings, whereas genes that are downregulated in the M phase (or, equivalently, upregulated in the G1 phase) are expected to have large negative loadings.

Interpretation of the noise

The use of a probabilistic framework allows the model to explicitly disentangle the signal (i.e. the explained variance) from the noise (i.e. unexplained variance). Large values of τ_d^m indicate high certainity on the observations for the feature d in view m, as predicted by the latent variables. In contrast, small values of τ_d^m are indicative of low predictive power by the latent variables.

Missing values

The probabilistic formalism naturally accounts for incomplete data matrices, as missing observations do not intervene in the likelihood.
In practice, we implement this using memory-efficient binary masks $\mathscr{O}^m \in \mathbb{R}^{N \times D_m}$ for each view m, such that $\mathscr{O}_{n,d} = 1$ when feature d is observed for sample n, 0 otherwise.

Prior distributions for the factors and the loadings

The key determinant of the model is the regularization used on the prior distributions of the factors and the weights.
For the factors, we follow common practice [**XX**] and define an isotropic Gaussian prior:

$$p(z_{nk}) = \mathscr{N}(z_{nk} \,|\, 0, 1) \tag{4}$$

For the weights we encode two levels of sparsity, a (1) view- and factor-wise sparsity and (2) an individual feature-wise sparsity. The aim of the factor- and view-wise sparsity is to disentangle the activity of factors to the different views, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not explain any variation in view m.
In addition, we place a second layer of sparsity which encourages inactive weights on each individual feature. Mathematically, we express this as a combination of an Automatic Relevance Determination (ARD) prior [**Mackay1996**] for the view- and factor-wise sparsity and a spike-and-slab prior [25] for the feature-wise sparsity: However, this formulation of the spike-and-slab prior contains a Dirac delta function, which makes

the inference procedure troublesome. To solve this we introduce a re-parametrization of the weights w as a product of a Gaussian random variable $\hat{w}$ and a Bernoulli random variable s, [36] resulting in the following prior: In this formulation α_k^m controls the activity of factor k in view m and θ_k^m controls the corresponding fraction of active loadings (i.e. the sparsity levels).

Finally, we define conjugate priors for θ and α:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{5}$$

$$p(\alpha_k^m) = \mathscr{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right), \tag{6}$$

with hyper-parameters $a_0^\theta, b_0^\theta = 1$ and $a_0^\alpha, b_0^\alpha = 1e^{-3}$ to get uninformative priors.
Posterior values of θ_k^m close to 0 implies that most of the weights of factor k in view m are shrinked to 0 (sparse factor). In contrast, a value of θ_k^m close to 1 implies that most of the weights are non-zero (non-sparse factor). A small value of α_k^m implies that factor k is active in view m. In contrast, a large value of α_k^m implies that factor k is inactive in view m.

All together, the joint probability density function of the model is given by

$$\begin{aligned}
p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \theta, \alpha, \tau) = & \prod_{m=1}^{M}\prod_{n=1}^{N}\prod_{d=1}^{D_m} \mathscr{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right) \\
& \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} \mathscr{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \text{Ber}(s_{d,k}^m | \theta_k^m) \\
& \prod_{n=1}^{N}\prod_{k=1}^{K} \mathscr{N}\left(z_{nk} \,|\, 0, 1\right) \\
& \prod_{m=1}^{M}\prod_{k=1}^{K} \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \\
& \prod_{m=1}^{M}\prod_{k=1}^{K} \mathscr{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \\
& \prod_{m=1}^{M}\prod_{d=1}^{D_m} \mathscr{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right).
\end{aligned} \tag{7}$$

and the corresponding graphical model is shown in Figure 2. This completes the definition of the MOFA model.

0.1.2 Downstream analysis

Once trained, the MOFA model can be queried for a set of downstream analysis:

- **Variance decomposition**: calculate the variance explained (R^2) by each factor in each view.
- **Ordination of the samples in the latent space**: scatterplots or beeswarm plots of factors, colored or shaped by sample covariates can reveal the main drivers of sample heterogeneity.

- **Inspection of loadings**: the weights (or loadings) can be interpreted as an activity score for each gene on each factor. Hence, inspecting the top loadings reveals the genes (or other genomic features) that underlie each factor.

- **Imputation**: MOFA generates a condensed and denoised low-dimensional representation of the data without missing values. As discussed in Section X, the data can be reconstructed from the latent space by a simple matrix multiplication: $\hat{\mathbf{Y}} = \mathbf{ZW}^T$.

- **Feature set enrichment analysis**: when a factor is difficult to characterise based only on the inspection of the top loadings, one can compute a statistical test for enrichment of biological pathways using predefined gene-set annotations. The statistical tests that we implemented are outlined in Section X.

The downstream functionalities implemented in MOFA are highlighted in Figure 1.

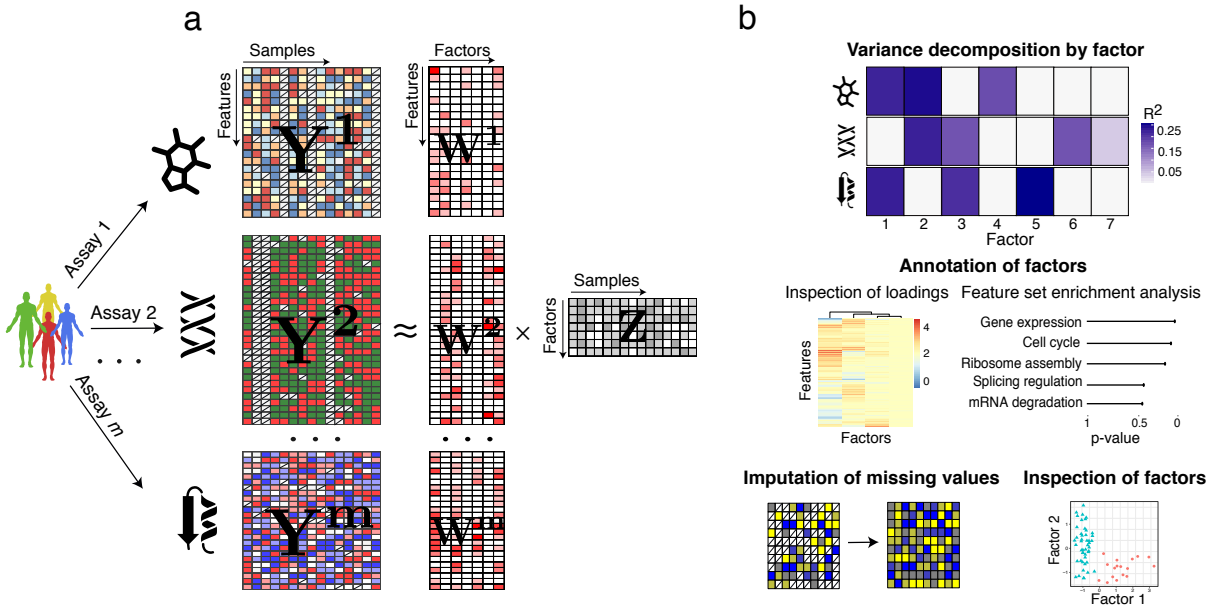


Fig. 1 MOFA overview. The model takes M data matrices as input ($\mathbf{Y}^1, \cdots, \mathbf{Y}^M$), one or more from each data modality, with co-occurrent samples but features that are not necessarily related and can differ in numbers. MOFA decomposes these matrices into a matrix of factors ($\mathbf{Z}$) and M weight matrices, one for each data modality ($\mathbf{W}^1, \cdots, \mathbf{W}^M$). White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denotes missing values. The fitted MOFA model can be queried for different downstream analyses, including a variance decomposition to assess the proportion of variance explained by each factor in each data modality.

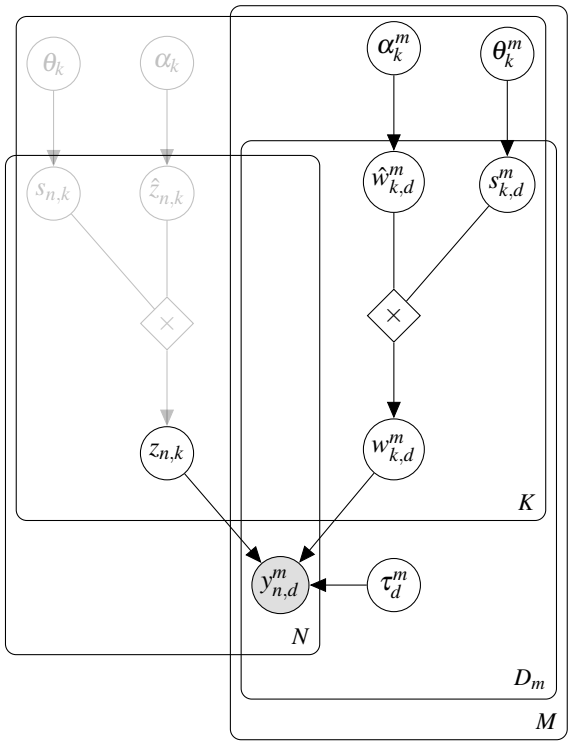


Fig. 2 Graphical model for MOFA. The white circles represent hidden variables that are infered by the model, whereasa the grey circles represent the observed variables. There are a total of four plates, each one representing a dimension of the model: M for the number of views, N for the number of samples, K for the number of factors and D_m for the number of features in view m

Inference

To make the model scalable to large data sets we adopt a Variational inference framework with a structured mean field approximation. A detailed overview is given in section XX, and details on the variational updates for the MOFA model are given in Appendix XX.
To enable efficient inference for non-Gaussian likelihoods we employ local bounds [**XX**]. This is described in detail in Section X

0.1.3 Monitoring convergence

An attractive property of Variational inference is that the objective function, the Evidence Lower Bound (ELBO), increases monotonically at every iteration. This provides a simple way of monitoring convergence Figure 3. This is indeed one of the reasons why we opted for this inference framework over Expectation Propagation or sampling-based approaches.

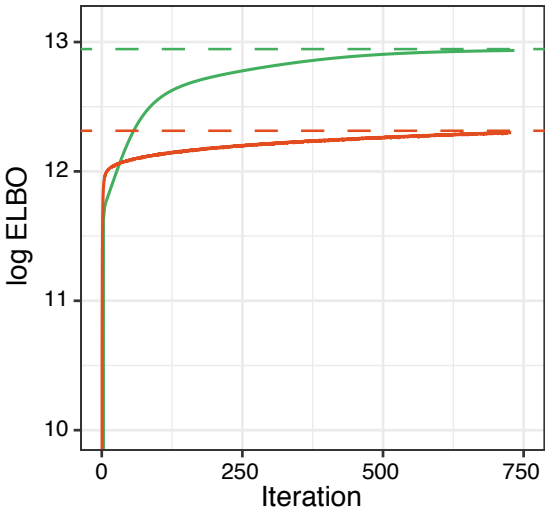


Fig. 3 Training curve for two diffent initialisations of MOFA. The y-axis displays the log of the ELBO, with higher values indicating a better fit. The x-axis displays the iteration number. The horizontal dash lines mark the value of the ELBO upon convergence.

0.1.4 Model selection and consistency across random initilizations

The variational optimisation problem in MOFA is not convex and the posterior distributions will vary depending on the initialisation of the model. Thus, it becomes mandatory to perform model selection and assess the consistency of the factors across different trials.

The strategy we adopted in this work is to train several MOFA models (e.g. 10 trials) under different parameter initialisations, and after training we select the model with the highest ELBO for downstream analysis Figure 4. In addition, we evaluate the robustness of the factors by plotting the Pearson correlations between factors across all trials. Figure 4.

A similar strategy has also been proposed in [**Hore2016**, **Hore2015-thesis**].

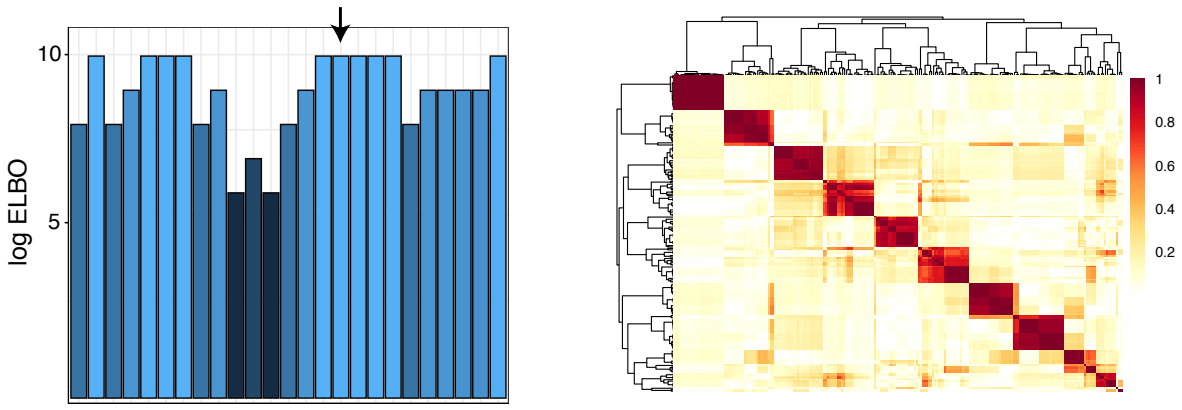


Fig. 4 Model selection and robustness analysis in MOFA. The left plot the log ELBO (y-axis) for 25 model instances (x-axis). The arrow indicates the model with the highest ELBO that would be selected for downstream analysis. The right plot displays the absolute value of the Pearson correlation coefficient between pairwise combinations of all factors across the 25 model instances. A block-diagonal matrix indicates that factors are robustly estimated regardless of the initialisation.

0.1.5 Learning the number of factors

As described in section X, the use of an ARD prior allows factors to be actively prunned by the model if their variance explained is negligible. In the implementation we control the prunning of factors by a hyperparameter that defines a threshold on the minimum fraction of variance explained by a factor (across all views).

Additionally, because of the non-convexity of the optimisation problem, different model instances can potentially yield solutions with different number of active factors (**??**). Thus, the optimal number of factors can be selected by the model selection strategy outlined in Section 0.1.4.

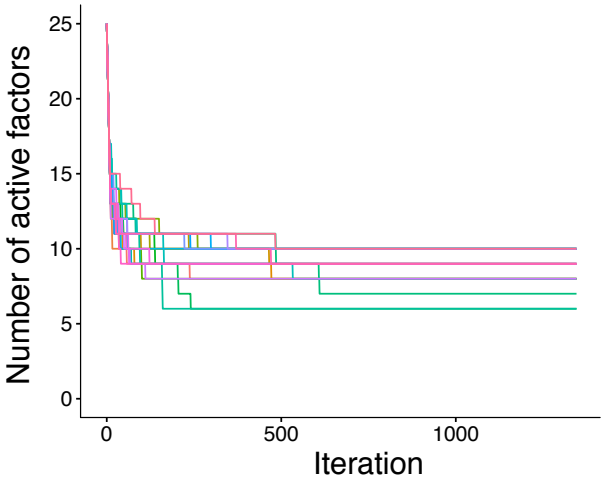


Fig. 5 Training curve for the number of active factors across 25 different model instances. The y-axis displays the number of active factors. The x-axis displays the iteration number. Different lines denote different model instances.

0.1.6 Modelling and inference with non-Gaussian data

To implement efficient variational inference in conjunction with a non-Gaussian likelihood we adapt prior work from [**seeger**] using local variational bounds. The key idea is to dynamically approximate non-Gaussian data by Gaussian pseudo-data based on a second-order Taylor expansion. To make the approximation justifiable we need to introduce variational parameters that are adjusted alongside the updates to improve the fit.

Denoting the parameters in the MOFA model as $\mathbf{X} = (\mathbf{Z}, \mathbf{W}, \alpha, \tau, \theta)$, recall that the variational framework approximates the posterior $p(\mathbf{X}|\mathbf{Y})$ with a distribution $q(\mathbf{X})$, which is indirectly optimised by optimising a lower bound of the log model evidence. The resulting optimization problem can be re-written from **??** as

$$\min_{q(\mathbf{X})} -\mathscr{L}(\mathbf{X}) = \min_{q(\mathbf{X})} \mathbb{E}_q\big[-\log p(\mathbf{Y}|\mathbf{X})\big] + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})].$$

Expanding the MOFA model to non-Gaussian likelihoods we now assume a general likelihood of the form $p(\mathbf{Y}|\mathbf{X}) = p(\mathbf{Y}|\mathbf{C})$ with $\mathbf{C} = \mathbf{Z}\mathbf{W}^T$, that can write as

$$-\log p(\mathbf{Y}|\mathbf{X}) = \sum_{n=1}^{N}\sum_{d=1}^{D} f_{nd}(c_{nd})$$

with $f_{nd}(c_{nd}) = -\log p(y_{nd}|c_{nd})$. We dropped the view index m to keep notation uncluttered.
Extending [**seeger**] to our heteroscedastic noise model, we require $f_{nd}(c_{nd})$ to be twice differentiable and bounded by κ_d, such that $f''_{nd}(c_{nd}) \leq \kappa_d \, \forall n,d$. This holds true in many important models as for example the Bernoulli and Poisson case. Under this assumption a lower bound on the log likelihood can be constructed using Taylor expansion,

$$f_{nd}(c_{nd}) \leq \frac{\kappa_d}{2}(c_{nd} - \zeta_{nd})^2 + f'(\zeta_{nd})(c_{nd} - \zeta_{nd}) + f_{nd}(\zeta_{nd}) := q_{nd}(c_{nd}, \zeta_{nd}),$$

where $\zeta = \zeta_{nd}$ are additional variational parameters that determine the location of the Taylor expansion and have to be optimised to make the lower bound as tight as possible. Plugging the bounds into above optimization problem, we obtain:

$$\min_{q(\mathbf{X}),\zeta} \quad \sum_{d=1}^{D}\sum_{n=1}^{N} \mathbb{E}_q[q_{nd}(c_{nd}, \zeta_{nd})] + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})]$$

The algorithm propsed in [**seeger**] then alternates between updates of ζ and q(Θ). The update for ζ is given by

$$\zeta \leftarrow \mathbb{E}[\mathbf{W}]\mathbb{E}[\mathbf{Z}]^T$$

where the expctations are taken with respect to the corresponding q distributions.
On the other hand, the updates for $q(\mathbf{X})$ can be shown to be identical to the variational Bayesian updates with a conjugate Gaussian likelihood when replacing the observed data $\mathbf{Y}$ by a pseudo-data $\hat{\mathbf{Y}}$ and the precisions τ_{nd} (which were treated as random variables) by the constant terms κ_d introduced above.
The pseudodata is given by

$$\hat{y}_{nd} = \zeta_{nd} - f'(\zeta_{nd})/\kappa_d.$$

Depending on the log likelihoods $f(\cdot)$ different κ_d are used resulting in different pseudo-data updates. Two special cases implemented in MOFA are the Poisson and Bernoulli likelihood described in the following.

Bernoulli likelihood for binary data

When the observations are binary, $y \in \{0,1\}$, they can be modelled using a Bernoulli likelihood:

$$\mathbf{Y}|\mathbf{Z},\mathbf{W} \sim \mathrm{Ber}(\sigma(\mathbf{Z}\mathbf{W}^T)),$$

where $\sigma(a) = (1+e^{-a})^{-1}$ is the logistic link function and $\mathbf{Z}$ and $\mathbf{W}$ are the latent factors and weights in our model, respectively.
In order to make the variational inference efficient and explicit as in the Gaussian case, we aim to approximate the Bernoulli data by a Gaussian pseudo-data as proposed in [**seeger**] and described above which allows to recycle all the updates from the model with Gaussian views. While [**seeger**] assumes a homoscedastic approximation with a spherical Gaussian, we adopt an approach following [**Jaakkola**], which allows for heteroscedaticity and provides a tighter bound on the Bernoulli likelihood.

Denoting $c_{nd} = (\mathbf{Z}\mathbf{W}^T)_{nd}$ the Jaakkola upper bound [**Jaakkola**] on the negative log-likelihood is given by

$$\begin{aligned} -\log\left(p(y_{nd}|c_{nd})\right) &= -\log\left(\sigma\left((2y_{nd}-1)c_{nd}\right)\right) \\ &\leq -\log(\zeta_{nd}) - \frac{(2y_{nd}-1)c_{nd}-\zeta_{nd})}{2} + \lambda(\zeta_{nd})\left(c_{nd}^2-\zeta_{nd}^2\right) \\ &=: b_J(\zeta_{nd}, c_{nd}, y_{nd}) \end{aligned}$$

with λ given by $\lambda(\zeta) = \frac{1}{4\zeta}\tanh\left(\frac{\zeta}{2}\right)$.
This can easily be derived from a first-order Taylor expansion on the function $f(x) = -\log(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = \frac{x}{2} - \log(\sigma(x))$ in x^2 and by the convexity of f in x^2 this bound is global as discussed in [**Jaakkola**].
In order to make use of this tighter bound but still be able to re-use the variational updates from the Gaussian case we re-formulate the bound as a Gaussian likelihood on pseudo-data $\hat{\mathbf{Y}}$.
As above we can plug this bound on the negative log-likelihood into the variational optimization problem to obtain

$$\min_{q(\mathbf{X}),\zeta} \quad \sum_{d=1}^{D}\sum_{n=1}^{N} \mathbb{E}_q b_J(\zeta_{nd}, c_{nd}, y_{nd}) + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})].$$

This is minimized iteratively in the variational parameter ζ_{nd} and the variational distribution of Z,W:
Minimizing in the variational parameter ζ this leads to the updates given by

$$\zeta_{nd}^2 = \mathbb{E}[c_{nd}^2]$$

as described in [**Jaakkola**], [**bishop2006pattern**].
For the variational distribution $q(\mathbf{Z},\mathbf{W})$ we observe that the Jaakkola bound can be re-written as

$$b_J(\zeta_{nd}, c_{nd}, y_{nd}) = -\log\left(\varphi\left(\hat{y}_{nd}; c_{nd}, \frac{1}{2\lambda(\zeta_{nd})}\right)\right) + \gamma(\zeta_{nd}),$$

where $\varphi(\cdot;\mu,\sigma^2)$ denotes the density function of a normal distribution with mean μ and variance σ^2 and γ is a term only depending on ζ. This allows us to re-use the updates for $\mathbf{Z}$ and $\mathbf{W}$ from a setting with Gaussian likelihood by considering the Gaussian pseudo-data

$$\hat{y}_{nd} = \frac{2y_{nd}-1}{4\lambda(\zeta_{nd})}$$

updating the data precision as $\tau_{nd} = 2\lambda(\zeta_{nd})$ using updates generalized for sample- and feature-wise precision parameters on the data.

Poisson likelihood for count data

When observations are a natural numbers, such as count data $y \in \mathbb{N} = \{0,1,\cdots\}$, they can be modelled using a Poisson likelihood:

$$p(y|c) = \lambda(c)^y e^{-\lambda(c)}$$

where $\lambda(c) > 0$ is the rate function and has to be convex and log-concave in order to ensure that the likelihood is log-concave.
As done in [**seeger**], here we choose the following rate function: $\lambda(c) = \log(1+e^c)$.

Then an upper bound of the second derivative of the log-likelihood is given by

$$f''_{nd}(c_{nd}) \leq \kappa_d = 1/4 + 0.17 * \max(\mathbf{y}_{:,d}).$$

The pseudodata updates are given by

$$\hat{y}_{nd} = \zeta_{nd} - \frac{\mathrm{S}(\zeta_{nd})(1 - y_{nd}/\lambda(\zeta_{nd}))}{\kappa_d}.$$

0.1.7 Model validation with simulated data

We used simulated data from the generative model to systematically test the technical capabilities of MOFA.

Recovery of the latent space

First, we tested the ability of MOFA to recover simulated factors, varying the number of views, the number of features, the number of factors and the fraction of missing values.
For every simulation scenario we initialised a model with a high number of factors ($K = 100$), and inactive factors were automatically dropped during model training using a threshold of 1% variance explained. In addition, to test the robustness under different initialisations, ten models were trained for every simulation scenario.
We observe that in most settings the model accurately recovers the correct number of factors (Figure 6). Exceptions occur when the dimensionality of the latent space is too large (more than 50 factors) or when an excessive amount of missing values (more than 80%) is present in the data.

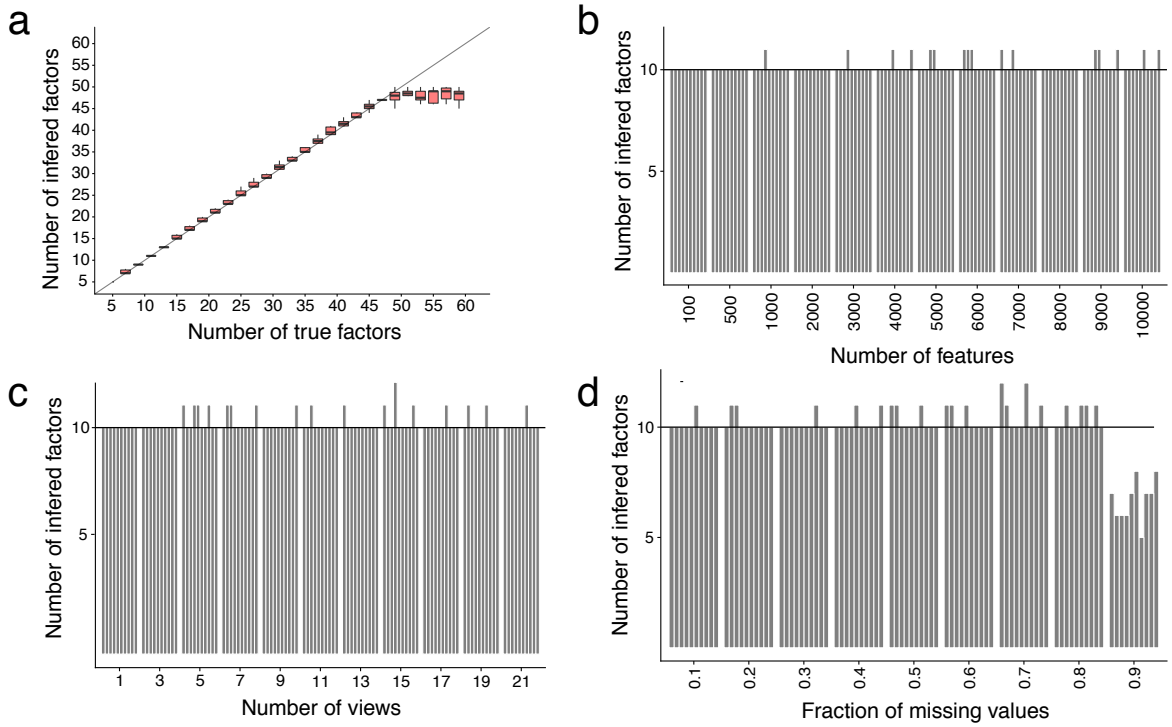


Fig. 6 Assessing the ability of MOFA to recover simulated latent spaces. In all plots the y-axis displays the number of infered factors. (a) x-axis displays the number of true factors, and boxplots summarise the distribution across 10 model instances. For (c-d) the true number of factors was set to $K = 10$ and each bar corresponds to a different model instance. (b) x-axis displays the number of features, (c) x-axis displays the number of views, (d) x-axis displays fraction of missing values.

Group-wise sparsity on the loadings

One of the most important statistical assumptions underlying MOFA (and other Group Factor Analysis methods) is the sparsity prior aimed at disentangling the activity of factors across views.

To evaluate this feature we simulated data from the generative model were the factors were clearly set to be active or inactive in specific views. We compared the performance with two other methods: the iCluster+ model [**Mo2013**] and a GFA implementation [**Leppaaho2017**].

the GFA implementation shares the same factor and view-wise sparsity as MOFA, and is therefore expected to show similar performance. On the other hand, iCluster is a model that is aimed at clustering and it only contains a sparsity constraint (in a penalised maximum likelihood setting) per factor, shared across all views. On simulated data, MOFA and GFA correctly infer the true activity pattern of the factors whereas iCluster infers incorrect sharedness of factors across views, especially with increasing dimensionality of the latent space.

Fig. 7 Evaluating the ability of MOFA, iCluster and GFA to recover sparse factor activity patterns across views. The leftmost plot displays the true activity pattern, with factors being (strongly) active in different subsets of views. The remaining three plots show, for each model, the fraction of variance explained (R^2) by each factor in each view.

Feature-wise sparsity on the loadings

A key aspect of MOFA is the use of a spike-and-slab prior distribution to enforce feature-wise sparsity on the loadings, which yields a more interpretable solution (see Section XX).

To assess the effect of the spike-and-slab prior we fit a group of models with a spike-and-slab prior and another group of models only with Automatic Relevance Determination prior. We further compared both solutions to a conventional Principal Component Analysis fit on the concatenated data set.

As expected, we observe that the spike-and-slab prior induces more zero-inflated weights, although the ARD prior provided a moderate degree of regularisation. The PCA solution was notably more dense than both bayesian models (Figure 8).

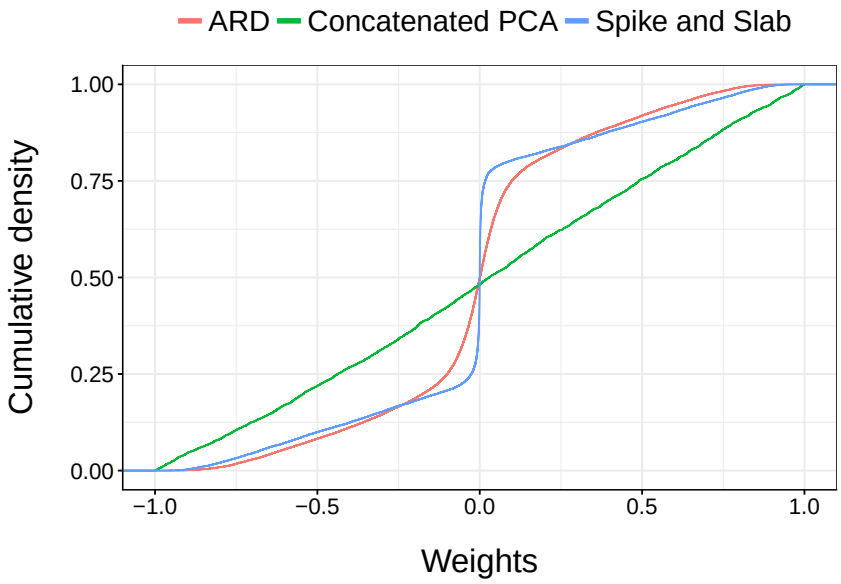


Fig. 8 Assessing sparsity on the loadings in MOFA. The plot shows the empirical cumulative density function of the loadings for an arbitrary factor in a single view. The loadings were simulated with a sparsity level of $\theta_k^m = 0.5$ (50% of active features.)

Non-gaussian likelihoods

A key improvement of MOFA with respect to previous methods is the use of non-Gaussian likelihoods to integrate multiple data modalities. As described in section XX, we implemented a Bernoulli likelihood to model binary data and a Poisson likelihood to model count data.

To validate both likelihood models, we simulated binary and count data using the generative model and we fit two sets of models for each data type: a group of models with a Gaussian likelihood and a group of models with a Bernoulli or Poisson likelihood, respectively.

We observe that although both likelihoods are able to recover the true number of factors, the models with the non-Gaussian likelihoods clearly result in a better fit to the data (**????**).

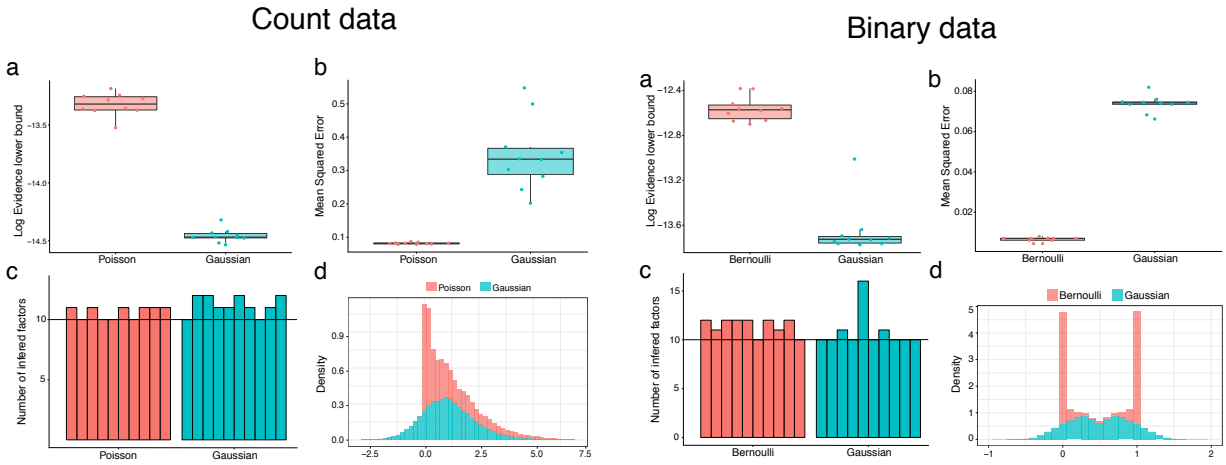


Fig. 9 Validation of the non-gaussian likelihood models implemented in MOFA on simulated data. The four plots on the left assess the Poisson and the Gaussian likelihoods applied to count data. The four plots on the right assess the Bernoulli and the Gaussian likelihoods applied to binary data. (a) The y-axis displays the ELBO for each model instance (x-axis). (b) The y-axis displays the mean reconstruction error for each model instance (x-axis). (c) The y-axis displays the number of estimated factrors for each model instance (x-axis). The horizontal dashed line marks the true number of factors $K = 10$. (d) Distribution of reconstructed data.

Scalability

Finally, we evaluated the scalability of the model when varying each of its dimensions independently (**??**), and we compared the speed with a Gibbs sampling implementation of GFA [**Leppaaho2017**] and iCluster+ [**Mo2013**].

Overall, we observe that MOFA scales linear with respect to all dimensions and is significantly faster than any of the three evaluated techniques.

As a real application showcase, the training on the CLL data Figure 11 required 25 minutes using MOFA, 34 hours with GFA and 5-6 days with iCluster.

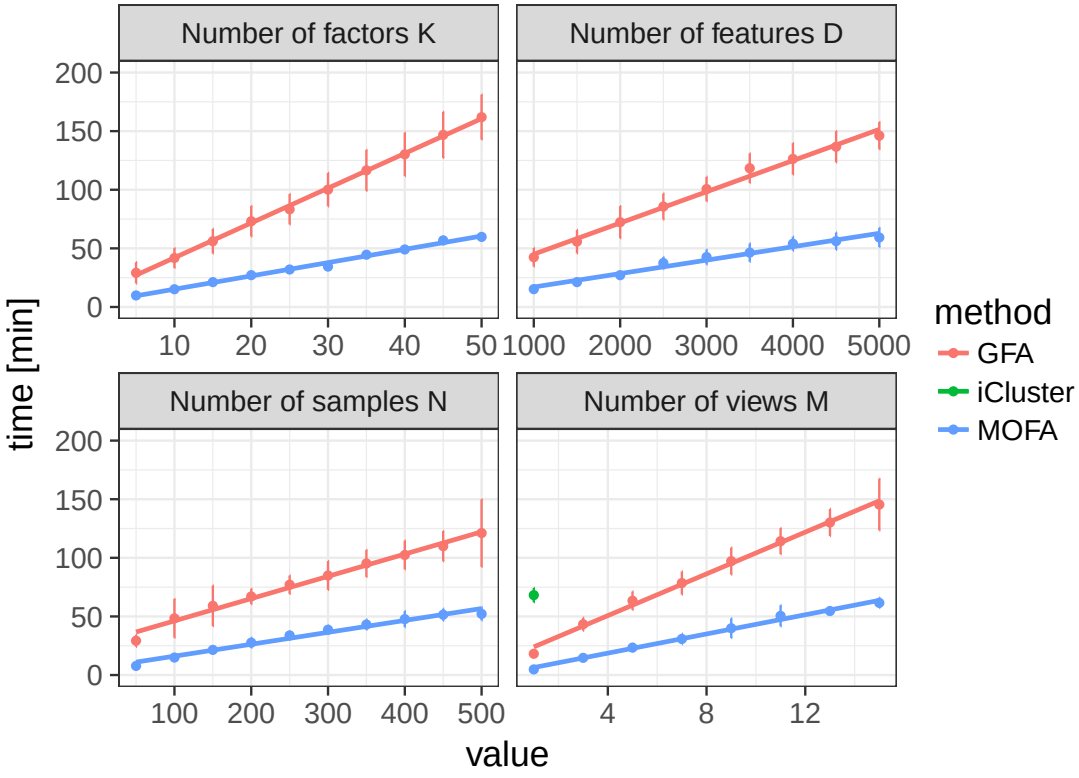


Fig. 10 Evaluation of speed and scalability in MOFA. The y-axis displays the time required for convergence. The x-axis displays the value of the dimension that was tested, either number of factors (K), number of features (D), number of samples (N) and number of views (M). Baseline parameters were $M = 3, K = 10, D = 1000, N = 100$. Each line represents a different model, GFA (red), MOFA (blue) and iCluster (green). Default convergence criteria where used for all methods. Each dot displays the average time across 10 trials with error bars denoting the standard deviation. iCluster is only shown for one value as all other settings required more than 200min for convergence.

0.1.8 Application to chronic lymphocytic leukaemia

Personalised medicine is an attractive field for the use of multi-omics, as dissecting heterogeneity across patients is a major challenge in complex diseases, and requires integration of information from multiple biological layers [**Chen2013**, **Costello2014**, **Alyass2015**].

In most cases, predicting patient survival and response to a treatment is still not reliable due to a lack of predictive biomarkers and our incomplete understanding of the mechanisms underlying response heterogeneity. Identification of the main drivers of inter-patient variation and their molecular basis is an important step towards personalized treatment decisions.

To demonstrate the potential of the method, we applied MOFA to a study of 200 patient samples of chronic lymphocytic leukaemia (CLL) profiled for somatic mutations, RNA expression, DNA methylation and ex-vivo drug responses[**Dietrich2018**] Figure 11.

This data set was selected for five main reasons.

- The large number of cases to benchmark the speed of MOFA against other common integrative methods.

- The rich literature and numerous studies that have been published for this type of cancer [**XX**], which provides a good resource for the interpretation of the factors
- The complex missing data structure of the study, with nearly 40% samples having incomplete assays **??**, in addition to the missing values present within some assays. hence, this study is ideal to benchmark MOFA's capabilities to deal with missing entries. As described in Section X, the inference framework we implemented allows the model to cope with this setting by merely ignoring missing entries and, when possible, pooling information from other molecular layers in order to infer the factor values.
- The different data modalities: after data processing, three assays had continuous observations whereas for the somatic mutations the observations were binary. As described in section XX, MOFA can combine different likelihood models to integrate multiple data types.
- The existence of clinical covariates: after model fitting, the factors can be associated with additional covariates. This provides an excellent test to assess whether the MOFA factors can capture the clinical phenotypes better than other dimensionality reduction techniques.

Model overview

In this data set, MOFA recovered 10 factors explaining a minimum of 3% of variance. Among these, the first two factors (sorted by variance explained) were active across most views, indicating a strong effect across multiple molecular layers. Other factors such as Factor 3 or Factor 5 explained variation in two data modalities, whereas Factor 4 was only in the RNA expression data.
Overall, the then MOFA factors inferred explained 41% of variance in the drug response data, 38% in the mRNA expression, 24% in the DNA methylation and 24% in somatic mutations.
Inspection of the top weights in the somatic mutation view revealed that Factor 1 was strongly associated with the mutation status of the immunoglobulin heavy-chain variable (IGHV) region, while Factor 2 was aligned with a trisomy of chromosome 12 (Figure X). In a completely unsupervised fashion, MOFA identified the two major axes of molecular disease heterogeneity being indeed the two most important clinical markers in CLL [**Fabbri2016**, **Zenz2010**].
A scatterplot based on these factors shows a clear separation of patients by their IGHV status on the first factor and presence or absence of trisomy 12 on the second factor. Remarkably, the IGHV status of a fraction of patients was missing (grey dots). Yet, as the two factors were shared across multiple views, MOFA was able to pool information from the other molecular layers to map those samples to the latent space.

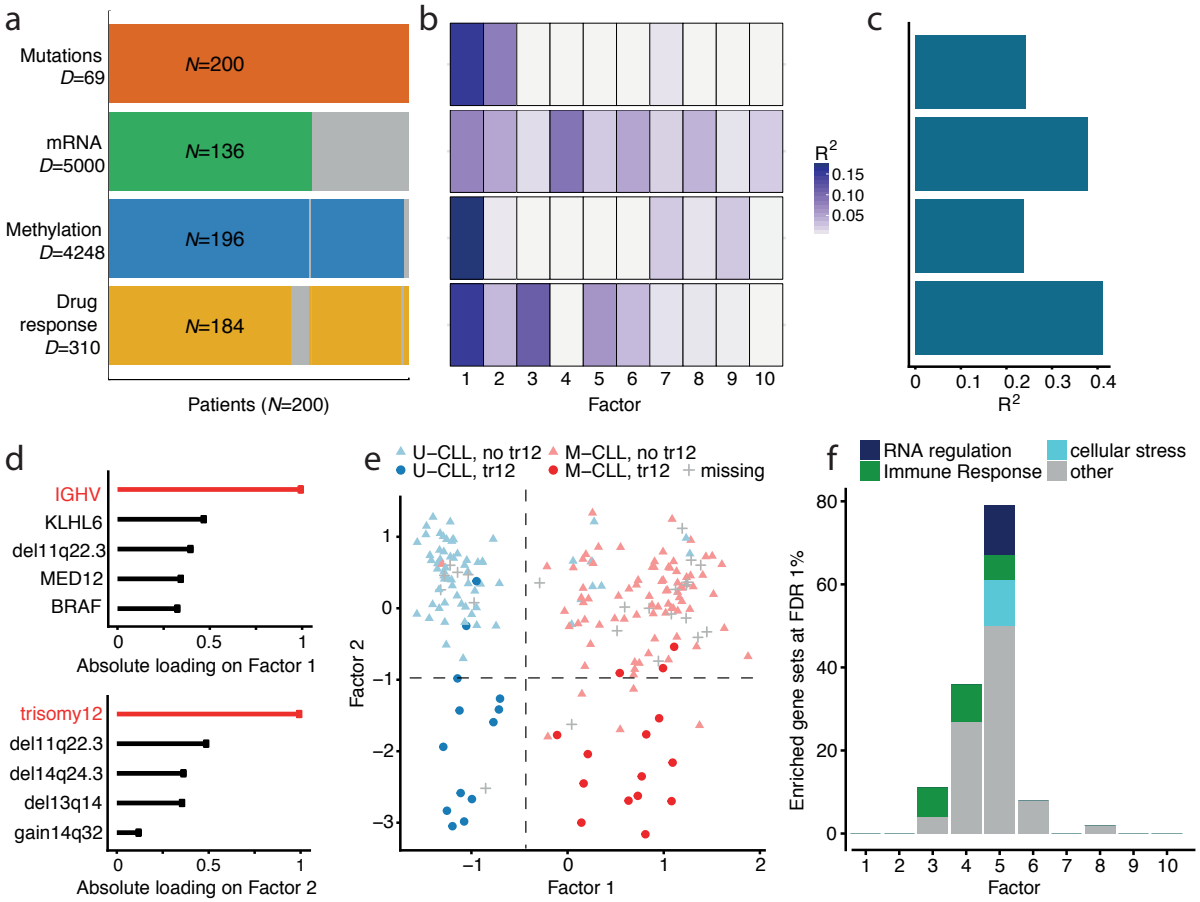


Fig. 11 XX

Characterisation of Factor 1

IGHV status is probably the most important prognostic marker in CLL and has routinely been used to distinguish between two distinct subtypes of the disease[**Fabbri2016**, **Bulian2017**, **Crombie2017**, **Damle1999**]. Molecularly, it is a surrogate of the level of activation of the B-cell receptor, which is in turn related to the differentiation state of the tumoral cells.
Multiple studies have associated mutated IGHV with a better resposne to chemoimmunotherapy, whereas unmutated IGHV patients have a worse prognosis [**Fabbri2016**, **Bulian2017**, **Crombie2017**, **Damle1999**].

In clinical practice, the IGHV status has been generally considered binary. However, our results suggest a more complex structure with at least three groups or a potential underlying continuum, as also suggested in [**Oakes2015**, **Queiros2014**].
Interestingly, there is some discrepancy between the IGHV status predicted by MOFA and the IGHV status reported in the clinical data. Out the 200 patients, MOFA classifies 176 in accordance with the clinical label, it classifies 12 patients that lacked the clinical marker and it re-classifies 12 patients to the opposite group. To validate the MOFA-based classification, we inspected the molecular profiles. sample-to-sample correlation matrices for the individual layers suggest that for 3 of the cases where the inferred factor disagrees with the clinical label, the molecular data supports the predicted label. The other 9 cases showed intermediate molecular sgnatures now well captured by the binary classification.

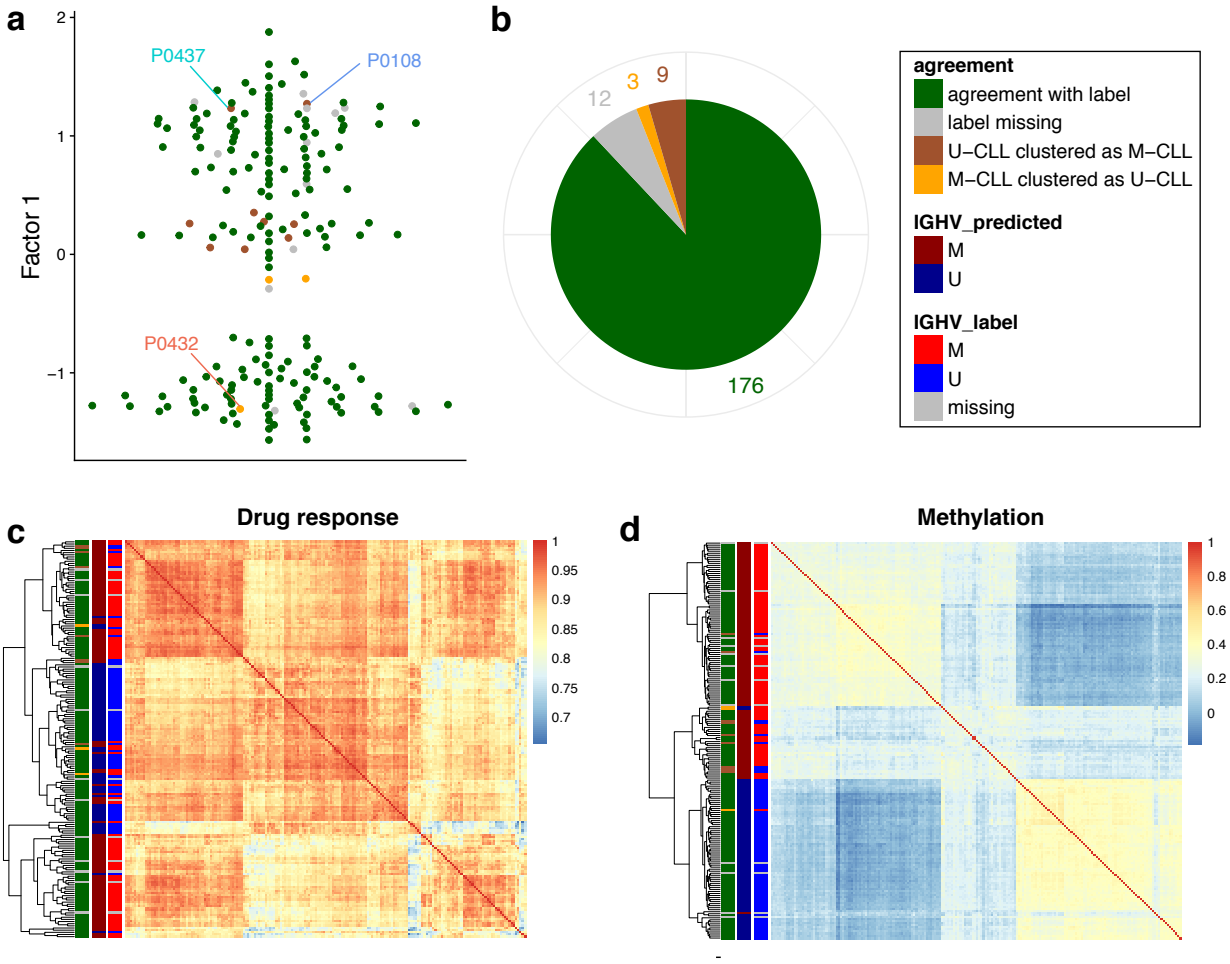


Fig. 12 XX

Finally, we characterised in more detail the molecular changes associated with IGHV status, as predicted by MOFA.

On the RNA expression, inspection of the top weights pinpoint genes that have been previously associated to IGHV status, some of which have been proposed as clinical markers [**Vasconcelos2005**, **Maloum\GenericError {(inputenc) }{Package inputenc Error: Unicode char âĂŐ (U+200E)\MessageBrea \MessageBreak or <return> to continue without it.}2009**, **Trojani2011**, **Morabito2015**, **Plesingerova2017**]. Heatmaps of the RNA expression levels for these genes reveals clear differences between samples when ordered according to the corresponding Factor 1 values.

On the drug response data the loadings highlight kinase inhibitors targeting the B-cell receptor pathway. Splitting the patients into three groups based on k-means clustering shows clear separation in the drug response curves.

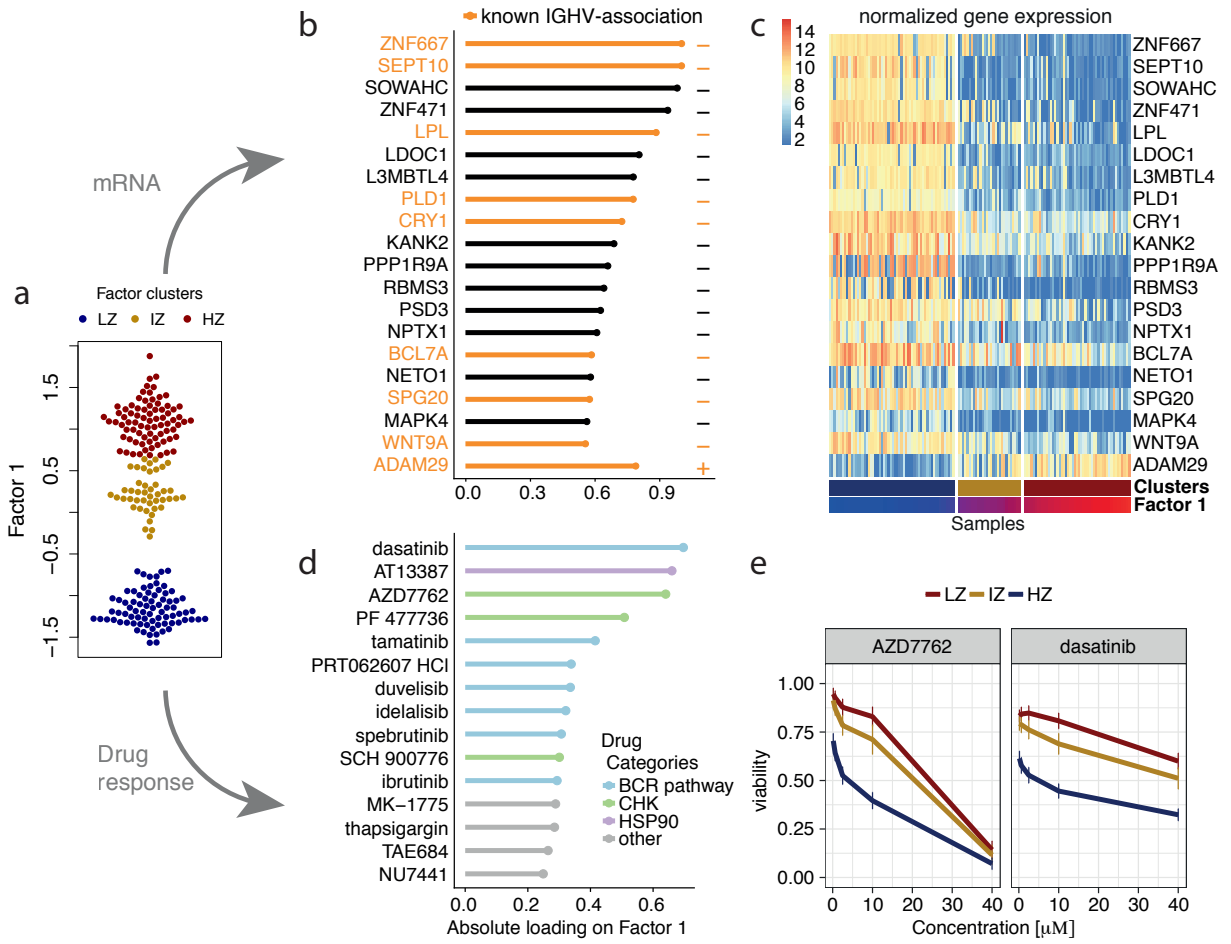


Fig. 13 XX

Characterisation of other Factors

Prediction of clinical outcome

Imputation of missing values

A promising application of MOFA is the imputation of missing values as well as entire missing assays, which could massively reduce experimental costs.
The principle of imputation in MOFA follows the same logic as simulating from the generative model: if the factors and weights are available one can reconstruct the data by a simple matrix multiplication:

$$\hat{\mathbf{Y}} = \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}]^T$$

where $\mathbb{E}[\mathbf{Z}]$ and $\mathbb{E}[\mathbf{W}]$ denote the expected values of the variational distributions for the factors and the loadings, respectively. Instead of point estimates, more advanced and fully Bayesian posterior predictive distribution could be obtained by propagating the uncertainity [**XX**]. Yet, given that the variance of the variational distributions tend to be heavily underestimated, we did not attempt this approach [**XX**].

To assess the imputation performance, we trained MOFA models on patients with complete measurements after masking parts of the drug response data. In a first experiment, we masked values at random, and in a

second experiment we masked the entire drug response data. We compared the results to some established imputation strategies, including imputation by feature-wise mean, SoftImpute [**Mazumder2010**], a k-nearest neighbour method [**Troyanskaya2001**].

For both imputation tasks, MOFA consistently yielded more accurate predictions, albeit the differences became less pronounced in the imputation of full assays, a more challenging task.

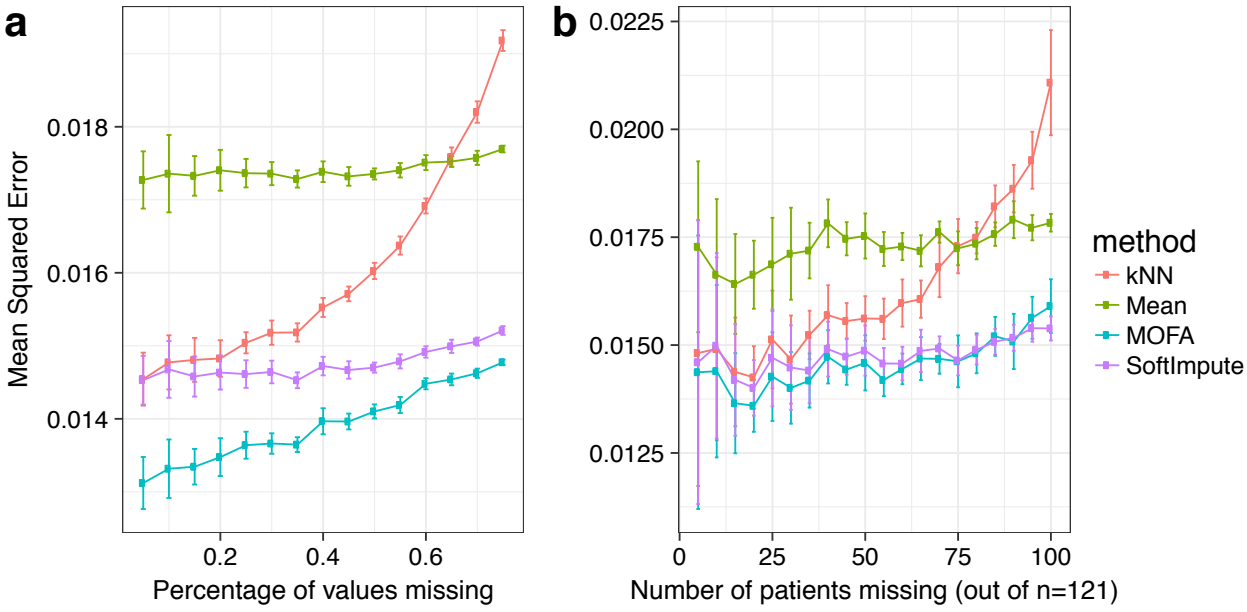


Fig. 14 Evaluation of imputation performance in the drug response assay of the CLL data. The y-axis shows averages of the mean-squared error across 15 trials for increasing fractions of missing data (x-axis). Two experiments were considered: (a) values missing at random and (b) entire assays missing at random. Error bars represent standard deviations.

0.1.9 Application to single-cell multi-omics

The emergence of single-cell multi-modal techniques has created open opportunities for the development of novel computational techniques that integrate data sets across multiple modalities [**Stuart2019**, **Colome-Tatche2018**, **Chappell2018**]. Here, we investigated the potential of MOFA to unravel the heterogeneity in one of the earliest single-cell multi-omics experiments [**Angermueller2016**].
The data set consists on 87 embryonic stem cells (ESCs) where RNA expression and DNA methylation were simultaneously measured using single-cell Methylation and Transcriptome sequencing (scM&T-seq). Two populations of ESCs were profiled: the first one contains 16 cells grown in 2i media, which induces a native pluripotency state with genome-wide DNA hypomethylation [**Ficz2013**]. the second population contains 71 cells grown in serum media, which triggers a primed pluripotency state poised for differentiation [**Tosolini2016**].

The RNA expression data was processed using standard pipelines to obtain log normalised counts, followed by a selection of the top 5,000 most overdispersed genes [**Lun2016**].
The DNA methylation data was processed as described in section XXX. Briefly, for each CpG site, we calculated a binary methylation rate from the ratio of methylated read counts to total read counts. Next, CpG sites were classified by overlapping with genomic contexts, namely promoters, CpG islands and enhancers

(defined by the presence of distal H3K27ac marks). Finally, for each annotation we selected the top 5,000 most variable CpG sites with a minimum coverage of 10% across cells.
Each of the resulting matrices was defined as a separate view for MOFA.

GAUSS VS BINARY
In this data set, MOFA learnt 3 factors (minimum explained variance of 1%). Factor 1 captured the transition from naive to primed pluripotent states, which MOFA links to widespread coordinated changes between DNA methylation and RNA expression. Inspection of the gene loadings for Factor 1 pinpoints important pluripotency markers including Rex1/Zpf42 or Essrb [**Mohammed2017**]. As previously described both in vitro [**Angermueller2016**]) and in vivo [**Auclair2014**], the dynamics of DNA methylation are driven by a genome-wide increase in DNA methylation levels.
Factor 2 captured a second dimension of heterogeneity driven by the transition from a primed pluripotency state to a differentiated state, with RNA loadings enriched with canonical differentiation markers including keratins and annexins [**XX**].

Jointly, the combination of Factors 1 and 2 reconstruct the full differentiation trajectory from naive pluripotent cells to differentiated cells **??**. When applying popular integrative clustering algorithms [**Wang2014**, **Shen2009**, **Moi2013**], the trajectory is not recovered **??**, illustrating the importance of learning continuous latent spaces before applying clustering methods.

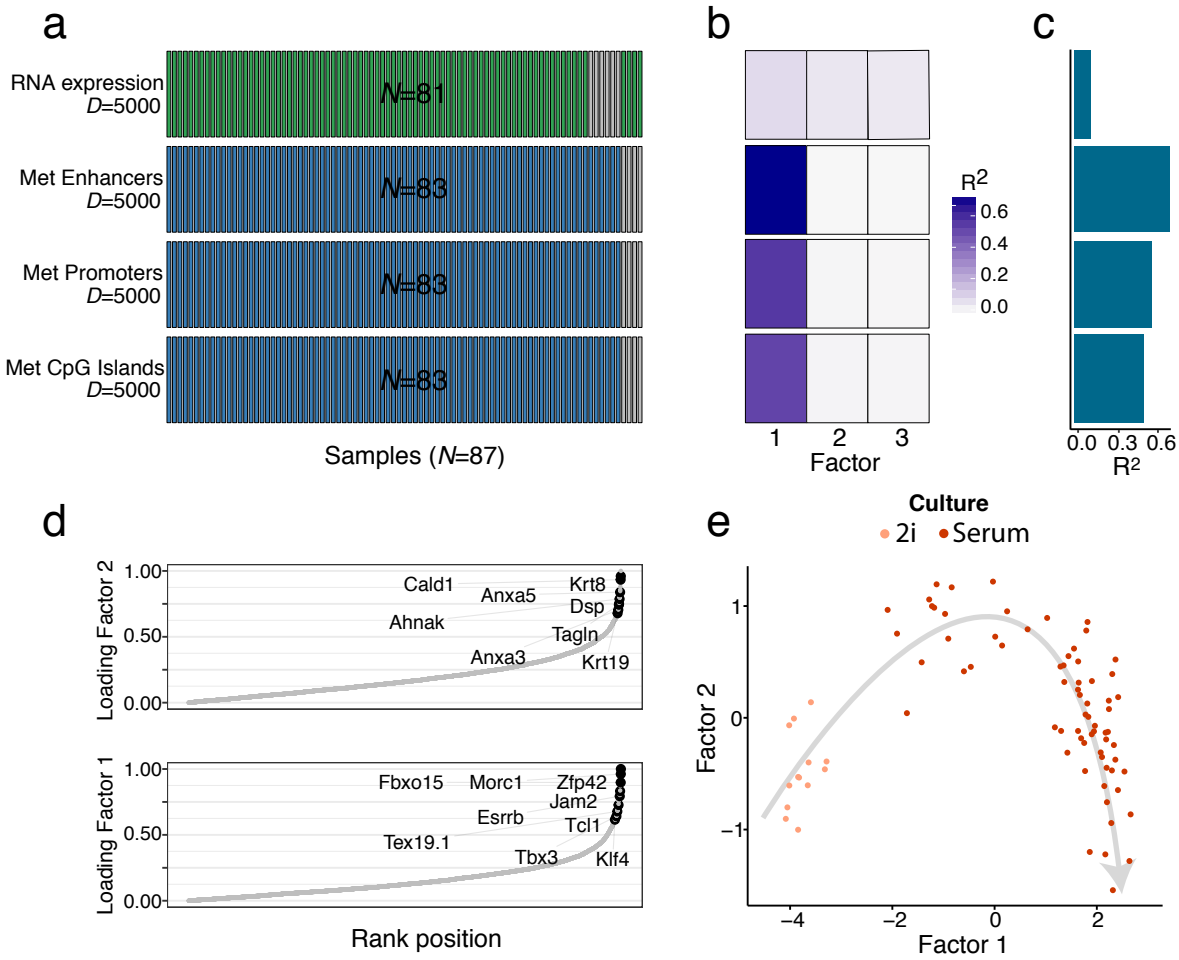

a
RNA expression
D=5000
N=81
Met Enhancers
D=5000
N=83
Met Promoters
D=5000
N=83
Met CpG Islands
D=5000
N=83
Samples (N=87)
b
R2
0.6
0.4
0.2
0.0
1
2
3
Factor
c
0.0 0.3 0.6
R2
d
Loading Factor 2
1.00
0.50
0.00
Cald1
Anxa5
Krt8
Dsp
Ahnak
Tagln
Anxa3
Krt19
Loading Factor 1
Fbxo15
Morc1
Zfp42
Jam2
Esrrb
Tcl1
Tex19.1
Tbx3
Klf4
Rank position
e
Culture
2i
Serum
Factor 2
1
0
−1
−4
−2
0
2
Factor 1

Fig. 15 XX

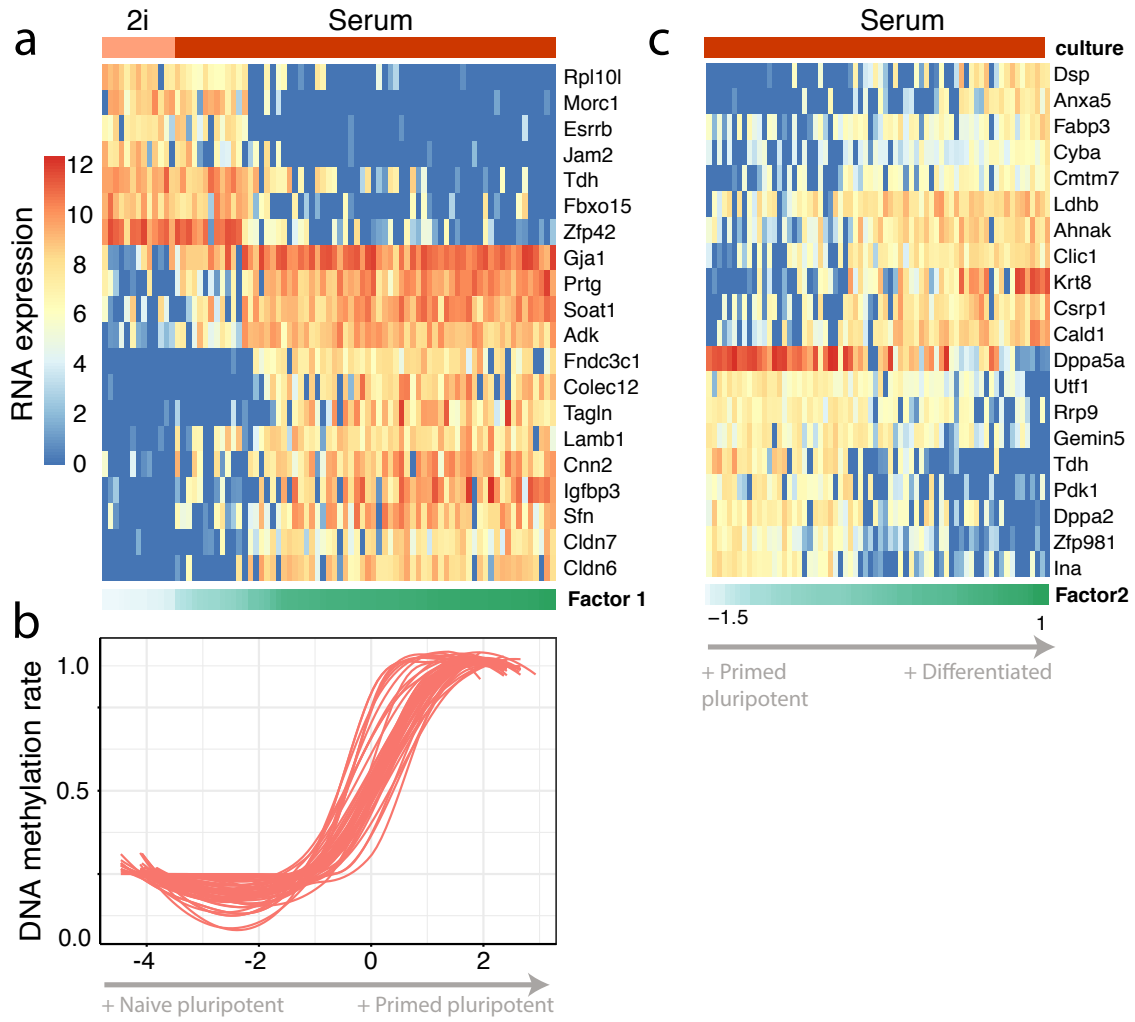


Fig. 16 XX

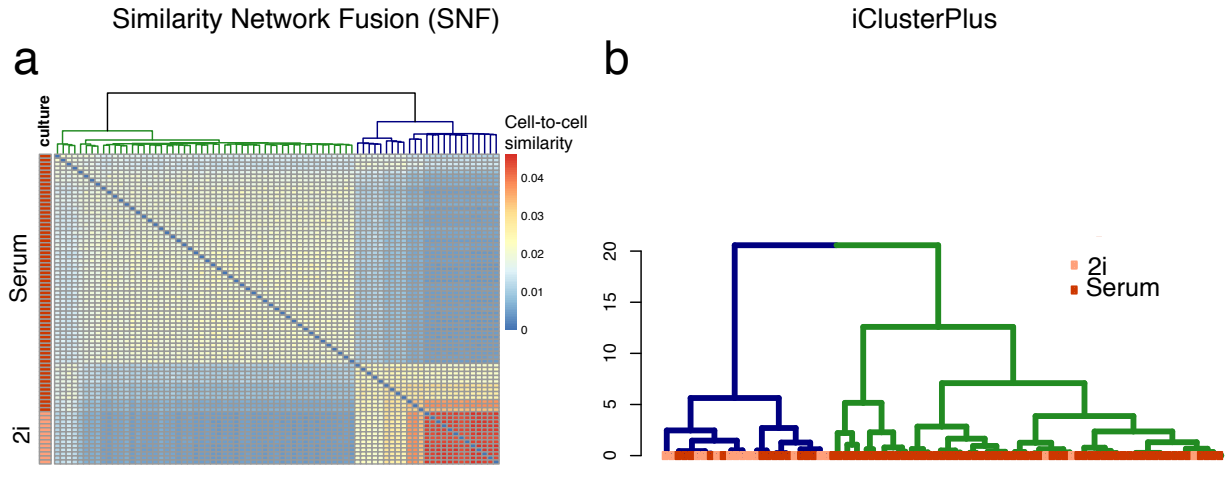


Fig. 17 XX

0.1.10 Open perspectives

MOFA addresses important challenges for the integrative analysis of multi-omics applications. Yet, MOFA is not free of limitations and there are multiple lines of improvement, some of which we followed in Chapter XX.

- Linearity: this is an assumption that is critical for obtaining interpretable feature loadings. Yet, there is often a trade-off between explanatory power and interpretability, particularly in computational biology where the drivers of variation result from the complex interaction between multiple components. As such, non-linear approaches, including deep neural networks or variational autoencoders have shown promising results when it comes to dimensionality reduction [**XX**], batch correction[**XX**], denoising [**XX**] or imputation [**XX**]. Notably, few non-linear multi-view factor analysis models exist [**Damianou2016**], and it could be an interesting line of research.

- Scalability: the size of biological datasets is rapidly increasing, particularly in the field of single cell sequencing, with some studies reporting more than a milion cells[**Svensson2018**, **Cao2019**].
  When compared to previous methods that make use of maximum likelihood (with grid-search) or sampling-based Bayesian methods, the variatonal framework implemented in MOFA has yield a notable improvement in scalability. Yet, in its vanilla form, variational inference becomes prohibitively slow with very large datasets [**XX**], hence motivating the development of even more efficient inference frameworks that potentially scale to milions of samples. This line of research is followed in Chapter XX, with the development of a stochastic version of the variational inference algorithm.

- Sample independence assumption and generalisations to multi-group structures: the sparsity assumptions in MOFA are based on the principle that features are structured into well-defined views, and that some factors may explain variability in only subsets of views, resulting in a structured sparsity as illustrated in **??**. Following the same logic, many studies contained structured samples, either as multiple experiments, conditions or data sets [**XX**]. The integration of multiple conditions or studies requires breaking the assumption of independent samples and introducing a prior that captures the existence of different groups, such that some factors are allowed to be active in subsets of groups. This line of research is followed in Chapter XX, with the introduction of a symmetric multi-group and multi-view sparsity prior.

- Tailored likelihoods for single-cell analysis: MOFA enables the modular extension to arbitrary non-gaussian likelihoods, provided that they can be locally bounded and integrated into the variational framework (see Section XX). New likelihood models such as zero-inflated negative binomial distributions [**Risso2018**] could make MOFA more suited to the analysis of single-cell data.

- Bayesian treatment of predictions: in the current implementation, after inference we extract point estimates for each variable, namely expectations. While convienient for plotting, this ignores the uncertainity associated with the estimates, one of the main strength of Bayesian methods. Future extensions could attempt a more comprehensive Bayesian treatment that propagates uncertainity in the downstream analyses, mainly when it comes to making predictions and imputation [**XX**].

- Incorporation of prior information: an unsupevised approach is appealing for discovering the principal axes of variation. however, sometimes this can yield challenges in the interpretation of factors. Future

extensions could exploit the rich information encoded in pathway databases, similar to the approach proposed in [**Buettner2017**].

Bibliography

[1] F. R. Bach and M. I. Jordan. *A probabilistic interpretation of canonical correlation analysis*. Tech. rep. 2005.

[2] C. Bishop. "Variational Principal Components". In: *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*. Vol. 1. IEE, Jan. 1999, pp. 509–514.

[3] C. M. Bishop. "Bayesian PCA". In: *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 382–388. ISBN: 0-262-11245-0.

[4] C. M. Bishop. "Pattern recognition". In: *Machine Learning* 128 (2006), pp. 1–58.

[5] K. Bunte et al. "Sparse group factor analysis for biclustering of multiple data sources". In: *Bioinformatics* 32.16 (2016), pp. 2457–2463.

[6] L. Dietz. *Directed Factor Graph Notation for Generative Models*. 2010.

[7] C. Gao, C. D. Brown, and B. E. Engelhardt. "A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects". In: *arXiv e-prints*, arXiv:1310.4792 (2013), arXiv:1310.4792. arXiv: 1310.4792 [stat.AP].

[8] Y. Guo et al. "Sufficient Canonical Correlation Analysis". In: *Trans. Img. Proc.* 25.6 (June 2016), pp. 2610–2619. ISSN: 1057-7149. DOI: 10.1109/TIP.2016.2551374.

[9] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2007, pp. 321–30. ISBN: 978-3-540-72243-4.

[10] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83. DOI: 10.1186/s13059-017-1215-1.

[11] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.

[12] H. Hotelling. "Relations between two sets of variates". In: *Biometrika* 28.3-4 (Dec. 1936), pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: http://oup.prod.sis.lan/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf.

[13] S. Huang, K. Chaudhary, and L. X. Garmire. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods". In: *Frontiers in Genetics* 8 (2017), p. 84. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00084.

[14] A. Ilin and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values". In: *J. Mach. Learn. Res.* 11 (Aug. 2010), pp. 1957–2000. ISSN: 1532-4435.

[15] S. A. Khan et al. “Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis”. In: *Bioinformatics* 30.17 (2014), pp. i497–i504.

[16] A. Klami and S. Kaski. “Probabilistic approach to detecting dependencies between data sets”. In: *Neurocomputing* 72.1 (2008), pp. 39–46.

[17] A. Klami et al. “Group factor analysis”. In: *IEEE transactions on neural networks and learning systems* 26.9 (2015), pp. 2136–2147.

[18] S. Komili and P. A. Silver. “Coupling and coordination in gene expression processes: a systems biology view”. In: *Nat. Rev. Genet.* 9 (Jan. 2008), p. 38.

[19] N. D. Lawrence et al. “Efficient inference for sparse latent variable models of transcriptional regulation”. In: *Bioinformatics* 33.23 (Aug. 2017), pp. 3776–3783. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx508. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/33/23/3776/25168082/btx508.pdf.

[20] J. T. Leek and J. D. Storey. “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis”. In: *PLoS Genet.* 3.9 (Sept. 2007), e161.

[21] X. Li, B. Xiao, and X.-S. Chen. “DNA Methylation: a New Player in Multiple Sclerosis”. In: *Molecular Neurobiology* 54.6 (Aug. 2017), pp. 4049–4059. ISSN: 1559-1182. DOI: 10.1007/s12035-016-9966-3.

[22] Y. Li, F.-X. Wu, and A. Ngom. “A review on machine learning principles for multi-view biological data integration”. In: *Briefings in Bioinformatics* 19.2 (Dec. 2016), pp. 325–340. ISSN: 1477-4054. DOI: 10.1093/bib/bbw113. eprint: http://oup.prod.sis.lan/bib/article-pdf/19/2/325/25524236/bbw113.pdf.

[23] S. D. McCabe, D.-Y. Lin, and M. I. Love. “MOVIE: Multi-Omics VIsualization of Estimated contributions”. In: *bioRxiv* (2018). DOI: 10.1101/379115. eprint: https://www.biorxiv.org/content/early/2018/07/29/379115.full.pdf.

[24] C. Meng et al. “Dimension reduction techniques for the integrative analysis of multi-omics data”. In: *Brief. Bioinform.* 17.4 (July 2016), pp. 628–641.

[25] T. J. Mitchell and J. J. Beauchamp. “Bayesian variable selection in linear regression”. In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[26] R. M. Neal. *Bayesian learning for neural networks*. 1995.

[27] M. Pilling. “Handbook of Applied Modelling: Non-Gaussian and Correlated Data”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4 (2018), pp. 1264–1265. DOI: 10.1111/rssa.12402. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12402.

[28] I. Pournara and L. Wernisch. “Factor analysis for gene regulatory networks and transcription factor activity profiles”. In: *BMC Bioinformatics* 8.1 (2007), p. 61.

[29] M. Rattray et al. “Inference algorithms and learning theory for Bayesian sparse factor analysis”. In: *Journal of Physics: Conference Series* 197 (Dec. 2009), p. 012002. DOI: 10.1088/1742-6596/197/1/012002.

[30] M. Ringnér. “What is principal component analysis?” In: *Nat. Biotechnol.* 26 (Mar. 2008), p. 303.

[31] M. D. Ritchie et al. “Methods of integrating data to uncover genotype–phenotype interactions”. In: *Nature Reviews Genetics* 16 (Jan. 2015),

[32] D. B. Rubin and D. T. Thayer. “EM algorithms for ML factor analysis”. In: *Psychometrika* 47.1 (1982), pp. 69–76.

[33] O. Stegle et al. “Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses”. en. In: *Nat. Protoc.* 7.3 (Feb. 2012), pp. 500–507.

[34] G. L. Stein-O’Brien et al. “Enter the Matrix: Factorization Uncovers Knowledge from Omics”. In: *Trends in Genetics* 34.10 (2018), pp. 790–805.

[35] M. Tipping and C. Bishop. “Probabilistic Principal Component Analysis”. In: *Journal of the Royal Statistical Society* 61(3) (1999), pp. 611–22.

[36] M. K. Titsias and M. Lázaro-Gredilla. “Spike and slab variational inference for multi-task and multiple kernel learning”. In: *Advances in neural information processing systems*. 2011, pp. 2339–2347.

[37] S. Virtanen et al. “Bayesian group factor analysis”. In: *Artificial Intelligence and Statistics*. 2012, pp. 1269–1277.

[38] C. Xu, D. Tao, and C. Xu. “A Survey on Multi-view Learning”. In: *arXiv e-prints*, arXiv:1304.5634 (Apr. 2013), arXiv:1304.5634. arXiv: 1304.5634 [cs.LG].

[39] I. S. L. Zeng and T. Lumley. “Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science)”. In: *Bioinformatics and Biology Insights* 12 (2019/03/16 2018), p. 1177932218759292.

[40] Z. Zhang et al. “Opening the black box of neural networks: methods for interpreting neural network models in clinical applications.” In: *Annals of translational medicine* 6 11 (2018), p. 216.

[41] S. Zhao et al. “Bayesian group factor analysis with structured sparsity”. In: *Journal of Machine Learning Research* 17.196 (2016), pp. 1–47.