Bayesian Hybrid Matrix Factorisation for Data Integration

Supplementary Materials

20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017).

Contents

1 Models **2**
- 1.1 Matrix factorisation models . . . 2
- 1.2 Matrix factorisation with ARD and importance values . . . 8
- 1.3 Hybrid matrix factorisation model . . . 10
  - 1.3.1 Model definition . . . 10
  - 1.3.2 Gibbs sampler . . . 11

2 Model discussion **17**
- 2.1 Software . . . 17
- 2.2 Complexity . . . 17
- 2.3 Missing values and predictions . . . 17
- 2.4 Initialisation . . . 17
- 2.5 Relation to tensor decomposition . . . 18

3 Datasets and preprocessing **21**
- 3.1 Drug sensitivity datasets . . . 21
- 3.2 Preprocessing drug sensitivity values . . . 22
- 3.3 Features . . . 22
- 3.4 Methylation and gene expression datasets . . . 25

4 Additional experiments **27**
- 4.1 Run-time comparison . . . 27
- 4.2 Model selection . . . 28
- 4.3 Initialisation . . . 28
- 4.4 Importance value . . . 31
- 4.5 Negativity constraints . . . 35
- 4.6 Factorisation types . . . 37

1 Models

In the model for Hybrid Matrix Factorisation (HMF), we combine models for Bayesian matrix factorisation, tri-factorisation, and tri-factorisation of similarity kernels. For each model, there are three versions: nonnegative, semi-nonnegative, and real-valued. In this section, we give the details for each of these nine models, as well as the Gibbs sampling algorithms.

In addition, for the semi-nonnegative and real-valued versions we can either use a univariate Gaussian (resulting in individual draws, but those can be drawn in parallel per column), or a multivariate Gaussian as the posterior (resulting in row-wise draws of new values, but each row can be drawn in parallel at the same time). We give Gibbs sampling algorithms for both options. For the nonnegative models, where the posterior is a truncated normal, this is technically also possible. However, we did not succeed in finding an efficiently implemented library for multivariate truncated normal draws, and therefore do not support it.

We first introduce each of the models separately in Section 1.1, extending them with Automatic Relevance Determination in Section 1.2, and finally explain how we combine them into one to form the HMF model in Section 1.3.

1.1 Matrix factorisation models

The matrix factorisation models are:

- Bayesian matrix factorisation (BMF), nonnegative matrix factorisation (BNMF), semi-nonnegative matrix factorisation (BSNMF). We decompose $\boldsymbol{D} \approx \boldsymbol{F} \cdot \boldsymbol{G}^T$. The Gibbs samplers are given in Table 2.
- Bayesian matrix tri-factorisation (BMTF), nonnegative matrix tri-factorisation (BNMTF), and semi-nonnegative matrix tri-factorisation (BSNMTF). We decompose $\boldsymbol{R} \approx \boldsymbol{F} \cdot \boldsymbol{S} \cdot \boldsymbol{G}^T$. The Gibbs samplers are given Table 3.
- Bayesian similarity matrix tri-factorisation (BSMTF), nonnegative similarity matrix tri-factorisation (BNSMTF), and semi-nonnegative similarity matrix tri-factorisation (BSNSMTF). We decompose $\boldsymbol{C} \approx \boldsymbol{F} \cdot \boldsymbol{S} \cdot \boldsymbol{F}^T$. The Gibbs sampler are given in Table 4.

The model priors and Gibbs sampling posteriors are given Table 1. Here,

- $\boldsymbol{R}, \boldsymbol{D}, \boldsymbol{C} \in \mathbb{R}^{I \times J}$, with $i = 1..I$, $j = 1..J$.
- $\boldsymbol{F} \in \mathbb{R}^{I \times K}$, with $i = 1..I$, $k = 1..K$.
- $\boldsymbol{S} \in \mathbb{R}^{K \times L}$ (matrix tri-factorisation), or $\boldsymbol{S} \in \mathbb{R}^{K \times K}$ (similarity matrix tri-factorisation), with $k = 1..K$, $l = 1..L$.
- $\boldsymbol{G} \in \mathbb{R}^{J \times K}$ (matrix factorisation), or $\boldsymbol{G} \in \mathbb{R}^{J \times L}$ (matrix tri-factorisation), with $j = 1..J$, $k = 1..K$, $l = 1..L$.
- $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Table 1: Model definitions, and Gibbs sampling posteriors.

Model name	Likelihood	Priors	Posteriors	Table
BMF	$D_{ij} \sim \mathcal{N}(D_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{N}(F_{ik}\|0, \lambda_F^{-1})$ $G_{jk} \sim \mathcal{N}(G_{jk}\|0, \lambda_G^{-1})$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{N}(F_{ik}\|\mu_{ik}^F, (\tau_{ik}^F)^{-1})$ $\mathcal{N}(G_{jk}\|\mu_{jk}^G, (\tau_{jk}^G)^{-1})$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	2
BMF (multivariate)			$\mathcal{N}(\boldsymbol{F}_i\|\boldsymbol{\mu_i^F}, \boldsymbol{\Sigma_i^F})$ $\mathcal{N}(\boldsymbol{G}_j\|\boldsymbol{\mu_j^G}, \boldsymbol{\Sigma_j^G})$	2
BNMF	$D_{ij} \sim \mathcal{N}(D_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $G_{jk} \sim \mathcal{E}(G_{jk}\|\lambda_G)$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{TN}(G_{jk}\|\mu_{jk}^G, \tau_{jk}^G)$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	2
BSNMF	$D_{ij} \sim \mathcal{N}(D_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $G_{jk} \sim \mathcal{N}(G_{jk}\|0, \lambda_G^{-1})$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{N}(G_{jk}\|\mu_{jk}^G, (\tau_{jk}^G)^{-1})$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	2
BSNMF (multivariate)			$\mathcal{N}(\boldsymbol{G}_j\|\boldsymbol{\mu_j^G}, \boldsymbol{\Sigma_j^G})$	2
BMTF	$R_{ij} \sim \mathcal{N}(R_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{N}(F_{ik}\|0, \lambda_F^{-1})$ $S_{kl} \sim \mathcal{N}(S_{kl}\|0, \lambda_S^{-1})$ $G_{jl} \sim \mathcal{N}(G_{jl}\|0, \lambda_G^{-1})$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{N}(F_{ik}\|\mu_{ik}^F, (\tau_{ik}^F)^{-1})$ $\mathcal{N}(S_{kl}\|\mu_{kl}^S, (\tau_{kl}^S)^{-1})$ $\mathcal{N}(G_{jl}\|\mu_{jl}^G, (\tau_{jl}^G)^{-1})$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	3
BMTF (multivariate)			$\mathcal{N}(\boldsymbol{F}_i\|\boldsymbol{\mu_i^F}, \boldsymbol{\Sigma_i^F})$ $\mathcal{N}(\boldsymbol{S}_k\|\boldsymbol{\mu_k^S}, \boldsymbol{\Sigma_k^S})$ $\mathcal{N}(\boldsymbol{G}_j\|\boldsymbol{\mu_j^G}, \boldsymbol{\Sigma_j^G})$	3
BNMTF	$R_{ij} \sim \mathcal{N}(R_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $S_{kl} \sim \mathcal{E}(S_{kl}\|\lambda_S)$ $G_{jl} \sim \mathcal{E}(G_{jl}\|\lambda_G)$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{TN}(S_{kl}\|\mu_{kl}^S, \tau_{kl}^S)$ $\mathcal{TN}(G_{jl}\|\mu_{jl}^G, \tau_{jl}^G)$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	3
BSNMTF	$R_{ij} \sim \mathcal{N}(R_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $S_{kl} \sim \mathcal{N}(S_{kl}\|0, \lambda_S^{-1})$ $G_{jl} \sim \mathcal{E}(G_{jl}\|\lambda_G)$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{N}(S_{kl}\|\mu_{kl}^S, (\tau_{kl}^S)^{-1})$ $\mathcal{TN}(G_{jl}\|\mu_{jl}^G, \tau_{jl}^G)$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	3
BSNMTF (multivariate)			$\mathcal{N}(\boldsymbol{S}_k\|\boldsymbol{\mu_k^S}, \boldsymbol{\Sigma_k^S})$	3
BSMTF	$C_{ij} \sim \mathcal{N}(C_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{F}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{N}(F_{ik}\|0, \lambda_F^{-1})$ $S_{kl} \sim \mathcal{N}(S_{kl}\|0, \lambda_S^{-1})$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{N}(F_{ik}\|\mu_{ik}^F, (\tau_{ik}^F)^{-1})$ $\mathcal{N}(S_{kl}\|\mu_{kl}^S, (\tau_{kl}^S)^{-1})$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	4
BSMTF (multivariate)			$\mathcal{N}(\boldsymbol{F}_i\|\boldsymbol{\mu_i^F}, \boldsymbol{\Sigma_i^F})$ $\mathcal{N}(\boldsymbol{S}_k\|\boldsymbol{\mu_k^S}, \boldsymbol{\Sigma_k^S})$	4
BNSMTF	$C_{ij} \sim \mathcal{N}(C_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{F}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $S_{kl} \sim \mathcal{E}(S_{kl}\|\lambda_S)$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{TN}(S_{kl}\|\mu_{kl}^S, \tau_{kl}^S)$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	4
BSNSMTF	$C_{ij} \sim \mathcal{N}(C_{ij}\|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{F}_j, \tau^{-1})$	$F_{ik} \sim \mathcal{E}(F_{ik}\|\lambda_F)$ $S_{kl} \sim \mathcal{N}(S_{kl}\|0, \lambda_S^{-1})$ $\tau \sim \mathcal{G}(\tau\|\alpha, \beta)$	$\mathcal{TN}(F_{ik}\|\mu_{ik}^F, \tau_{ik}^F)$ $\mathcal{N}(S_{kl}\|\mu_{kl}^S, (\tau_{kl}^S)^{-1})$ $\mathcal{G}(\tau\|\alpha^*, \beta^*)$	4
BSNSMTF (multivariate)			$\mathcal{N}(\boldsymbol{S}_k\|\boldsymbol{\mu_k^S}, \boldsymbol{\Sigma_k^S})$	4

The Gibbs sampling posteriors can be obtained using Bayes' theorem, for example for BN-MTF:

$$
\begin{aligned}
&p(F_{ik}|\tau, \boldsymbol{F}_{-ik}, \boldsymbol{S}, \boldsymbol{G}, \boldsymbol{R}, h) \\
&\quad \propto p(\boldsymbol{R}|\tau, \boldsymbol{F}, \boldsymbol{S}, \boldsymbol{G}) \times p(F_{ik}|\lambda_F) \\
&\quad \propto \prod_{j \in \Omega_i^1} \mathcal{N}(R_{ij}|\boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j, \tau^{-1}) \times \mathcal{E}(F_{ik}|\lambda_F) \\
&\quad \propto \exp\left\{ \frac{\tau}{2} \sum_{j \in \Omega_i^1} (R_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j)^2 \right\} \times \exp\left\{-\lambda_F F_{ik}\right\} \times u(x) \\
&\quad \propto \exp\left\{ \frac{F_{ik}^2}{2} \left[\tau \sum_{j \in \Omega_i^1} (\boldsymbol{S}_k \cdot \boldsymbol{G}_j)^2 \right] \right. \\
&\qquad \left. + F_{ik} \left[-\lambda_F + \tau \sum_{j \in \Omega_i^1} (R_{ij} - \sum_{k' \neq k} F_{ik'} (\boldsymbol{S}_{k'} \cdot \boldsymbol{G}_j)) (\boldsymbol{S}_k \cdot \boldsymbol{G}_j) \right] \right\} \times u(x) \\
&\quad \propto \exp\left\{ \frac{\tau_{ik}^F}{2} (F_{ik} - \mu_{ik}^F)^2 \right\} \times u(x) \\
&\quad \propto \mathcal{TN}(F_{ik}|\mu_{ik}^F, \tau_{ik}^F)
\end{aligned}
$$

where $h = \{\lambda_F, \lambda_G, \lambda_S, \alpha, \beta\}$ are the hyperparameters to the model, $u(x)$ is the unit step function, and $\Omega_i^1 = \{j \ |(i,j) \in \Omega\}$, $\Omega_j^2 = \{i \ |(i,j) \in \Omega\}$ indicate the observed entries per row and column, respectively.

All the parameter values can be found in Tables 2-4. We use $\otimes$ to denote the outer product, $\boldsymbol{I}$ for the identity matrix, and $\boldsymbol{S}_{\cdot,l}$ for the lth column of matrix $\boldsymbol{S}$.

For the similarity matrix factorisation we could have also decided to decompose $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^T + \boldsymbol{E}$, without the intermediate matrix $\boldsymbol{S}$. Ding et al. [2005] includes a good discussion of the benefits to our approach. For this decomposition we do not consider diagonal entries of $\boldsymbol{C}$ (in other words, $(i,i) \notin \Omega$, $i = 1..I$) as this leads to third and fourth order terms in the posteriors and makes Gibbs sampling impossible. See Zhang and Yeung [2012] for a non-probabilistic approach that does consider these elements, leading to a very complicated optimisation problem.

Table 2: Gibbs Samplers for Bayesian Matrix Factorisation (BMF, BNMF, BSNMF).

PARAM	UPDATE (GAUSSIAN PRIOR)	UPDATE (EXPONENTIAL PRIOR)
α^*	$\alpha + \frac{\lvert\Omega\rvert}{2}$	$\alpha + \frac{\lvert\Omega\rvert}{2}$
β^*	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(D_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{G}_j)^2$	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(D_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{G}_j)^2$
τ_{ik}^F	$\lambda_F + \tau \sum_{j\in\Omega_i^1} G_{jk}^2$	$\tau \sum_{j\in\Omega_i^1} G_{jk}^2$
μ_{ik}^F	$\frac{1}{\tau_{ik}^F}\left(\tau \sum_{j\in\Omega_i^1}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})G_{jk}\right)$	$\frac{1}{\tau_{ik}^F}\left(-\lambda_F + \tau \sum_{j\in\Omega_i^1}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})G_{jk}\right)$
τ_{jk}^G	$\lambda_G + \tau \sum_{i\in\Omega_j^2} F_{ik}^2$	$\tau \sum_{i\in\Omega_j^2} F_{ik}^2$
μ_{jl}^G	$\frac{1}{\tau_{jl}^G}\left(\tau \sum_{i\in\Omega_j^2}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})F_{ik}\right)$	$\frac{1}{\tau_{jl}^G}\left(-\lambda_G + \tau \sum_{i\in\Omega_j^2}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})F_{ik}\right)$
$\boldsymbol{\Sigma_i^F}$	$\left(\lambda^F \boldsymbol{I} + \tau \sum_{j\in\Omega_i^1} \boldsymbol{G}_j \otimes \boldsymbol{G}_j\right)^{-1}$	-
$\boldsymbol{\mu_i^F}$	$\boldsymbol{\Sigma_i^F} \cdot \left(\tau \sum_{j\in\Omega_i^1} D_{ij}\boldsymbol{G}_j\right)$	-
$\boldsymbol{\Sigma_j^G}$	$\left(\lambda^G \boldsymbol{I} + \tau \sum_{i\in\Omega_j^2} \boldsymbol{F}_i \otimes \boldsymbol{F}_i\right)^{-1}$	-
$\boldsymbol{\mu_j^G}$	$\boldsymbol{\Sigma_j^G} \cdot \left(\tau \sum_{i\in\Omega_j^2} D_{ij}\boldsymbol{F}_i\right)$	-

Table 3: Gibbs Samplers for Bayesian Matrix Tri-Factorisation (BMTF).

PARAM	UPDATE (GAUSSIAN PRIOR)	UPDATE (EXPONENTIAL PRIOR)
α^*	$\alpha + \frac{\lvert\Omega\rvert}{2}$	$\alpha + \frac{\lvert\Omega\rvert}{2}$
β^*	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(R_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j)^2$	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(R_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{G}_j)^2$
τ_{ik}^F	$\lambda_F + \tau \sum_{j\in\Omega_i^1} (\boldsymbol{S}_k \cdot \boldsymbol{G}_j)^2$	$\tau \sum_{j\in\Omega_i^1} (\boldsymbol{S}_k \cdot \boldsymbol{G}_j)^2$
μ_{ik}^F	$\frac{1}{\tau_{ik}^F}\left(\tau \sum_{j\in\Omega_i^1}(R_{ij} - \sum_{k'\neq k} F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{G}_j))(\boldsymbol{S}_k \cdot \boldsymbol{G}_j)\right)$	$\frac{1}{\tau_{ik}^F}\left(-\lambda_F + \tau \sum_{j\in\Omega_i^1}(R_{ij} - \sum_{k'\neq k} F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{G}_j))(\boldsymbol{S}_k \cdot \boldsymbol{G}_j)\right)$
τ_{kl}^S	$\lambda_S + \tau \sum_{(i,j)\in\Omega} F_{ik}^2 G_{jl}^2$	$\tau \sum_{(i,j)\in\Omega} F_{ik}^2 G_{jl}^2$
μ_{kl}^S	$\frac{1}{\tau_{kl}^S}\left(\tau \sum_{(i,j)\in\Omega}(R_{ij} - \sum_{(k',l')\neq(k,l)} F_{ik'} S_{k'l'} G_{jl'}) F_{ik} G_{jl}\right)$	$\frac{1}{\tau_{kl}^S}\left(-\lambda_S + \tau \sum_{(i,j)\in\Omega}(R_{ij} - \sum_{(k',l')\neq(k,l)} F_{ik'} S_{k'l'} G_{jl'}) F_{ik} G_{jl}\right)$
τ_{jl}^G	$\lambda_G + \tau \sum_{i\in\Omega_j^2} (\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l})^2$	$\tau \sum_{i\in\Omega_j^2} (\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l})^2$
μ_{jl}^G	$\frac{1}{\tau_{jl}^G}\left(\tau \sum_{i\in\Omega_j^2}(R_{ij} - \sum_{l'\neq l} G_{jl'}(\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l'}))(\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l})\right)$	$\frac{1}{\tau_{jl}^G}\left(-\lambda_G + \tau \sum_{i\in\Omega_j^2}(R_{ij} - \sum_{l'\neq l} G_{jl'}(\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l'}))(\boldsymbol{F}_i \cdot \boldsymbol{S}_{\cdot,l})\right)$
$\boldsymbol{\Sigma}_{\boldsymbol{i}}^{\boldsymbol{F}}$	$\left(\lambda^F \boldsymbol{I} + \tau \sum_{j\in\Omega_i^1} (\boldsymbol{S} \cdot \boldsymbol{G}_j) \otimes (\boldsymbol{S} \cdot \boldsymbol{G}_j)\right)^{-1}$	-
$\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{F}}$	$\boldsymbol{\Sigma}_{\boldsymbol{i}}^{\boldsymbol{F}} \cdot \left(\tau \sum_{j\in\Omega_i^1} R_{ij}(\boldsymbol{S} \cdot \boldsymbol{G}_j)\right)$	-
$\boldsymbol{\Sigma}_{\boldsymbol{k}}^{\boldsymbol{S}}$	$\left(\lambda^S \boldsymbol{I} + \tau \sum_{(i,j)\in\Omega} F_{ik}(\boldsymbol{G}_j \otimes \boldsymbol{G}_j)\right)^{-1}$	-
$\boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{S}}$	$\boldsymbol{\Sigma}_{\boldsymbol{k}}^{\boldsymbol{S}} \cdot \left(\tau \sum_{(i,j)\in\Omega}(R_{ij} - \sum_{k'\neq k} F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{G}_j)) F_{ik} \boldsymbol{G}_j\right)$	-
$\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\boldsymbol{G}}$	$\left(\lambda^G \boldsymbol{I} + \tau \sum_{i\in\Omega_j^2} (\boldsymbol{F}_i \cdot \boldsymbol{S}) \otimes (\boldsymbol{F}_i \cdot \boldsymbol{S})\right)^{-1}$	-
$\boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{G}}$	$\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\boldsymbol{G}} \cdot \left(\tau \sum_{i\in\Omega_j^2} R_{ij}(\boldsymbol{F}_i \cdot \boldsymbol{S})\right)$	-

Table 4: Gibbs Samplers for Bayesian Similarity Matrix Tri-Factorisation (BSMTF).

PARAM	UPDATE (GAUSSIAN PRIOR)	UPDATE (EXPONENTIAL PRIOR)
α^*	$\alpha + \frac{\lvert\Omega\rvert}{2}$	$\alpha + \frac{\lvert\Omega\rvert}{2}$
β^*	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(C_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{F}_j)^2$	$\beta + \frac{1}{2}\sum_{(i,j)\in\Omega}(C_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{S} \cdot \boldsymbol{F}_j)^2$
τ_{ik}^F	$\lambda_F + \tau\left[\sum_{j\in\Omega_i^1}(\boldsymbol{S}_k \cdot \boldsymbol{F}_j)^2 + \sum_{i'\in\Omega_i^2}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,k})^2\right]$	$\tau\left[\sum_{j\in\Omega_i^1}(\boldsymbol{S}_k \cdot \boldsymbol{F}_j)^2 + \sum_{i'\in\Omega_i^2}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,k})^2\right]$
μ_{ik}^F	$\frac{1}{\tau_{ik}^F}\left(\tau\sum_{j\in\Omega_i^1}(C_{ij} - \sum_{k'\neq k}F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{F}_j))(\boldsymbol{S}_k \cdot \boldsymbol{F}_j)\right.$ $\left. +\tau\sum_{i'\in\Omega_i^2}(C_{i'i} - \sum_{l\neq k}F_{il}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,l}))(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,k})\right)$	$\frac{1}{\tau_{ik}^F}\left(-\lambda_F + \tau\sum_{j\in\Omega_i^1}(C_{ij} - \sum_{k'\neq k}F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{F}_j))(\boldsymbol{S}_k \cdot \boldsymbol{F}_j)\right.$ $\left. +\tau\sum_{i'\in\Omega_i^2}(C_{i'i} - \sum_{l\neq k}F_{il}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,l}))(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}_{\cdot,k})\right)$
τ_{kl}^S	$\lambda_S + \tau\sum_{(i,j)\in\Omega}F_{ik}^2F_{jl}^2$	$\tau\sum_{(i,j)\in\Omega}F_{ik}^2F_{jl}^2$
μ_{kl}^S	$\frac{1}{\tau_{kl}^S}\left(\tau\sum_{(i,j)\in\Omega}(C_{ij} - \sum_{(k',l')\neq(k,l)}F_{ik'}S_{k'l'}F_{jl'})F_{ik}F_{jl}\right)$	$\frac{1}{\tau_{kl}^S}\left(-\lambda_S + \tau\sum_{(i,j)\in\Omega}(C_{ij} - \sum_{(k',l')\neq(k,l)}F_{ik'}S_{k'l'}F_{jl'})F_{ik}F_{jl}\right)$
$\boldsymbol{\Sigma_i^F}$	$\left(\lambda^F\boldsymbol{I} + \tau\left[\sum_{j\in\Omega_i^1}(\boldsymbol{S} \cdot \boldsymbol{F}_j) \otimes (\boldsymbol{S} \cdot \boldsymbol{F}_j) + \sum_{i'\in\Omega_i^2}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S}) \otimes (\boldsymbol{F}_{i'} \cdot \boldsymbol{S})\right]\right)^{-1}$	-
$\boldsymbol{\mu_i^F}$	$\boldsymbol{\Sigma_i^F} \cdot \left(\tau\sum_{j\in\Omega_i^1}C_{ij}(\boldsymbol{S} \cdot \boldsymbol{F}_j) + \tau\sum_{i'\in\Omega_i^2}C_{i'i}(\boldsymbol{F}_{i'} \cdot \boldsymbol{S})\right)$	-
$\boldsymbol{\Sigma_k^S}$	$\left(\lambda^S\boldsymbol{I} + \tau\sum_{(i,j)\in\Omega}F_{ik}(\boldsymbol{F}_j \otimes \boldsymbol{F}_j)\right)^{-1}$	-
$\boldsymbol{\mu_k^S}$	$\boldsymbol{\Sigma_k^S} \cdot \left(\tau\sum_{(i,j)\in\Omega}(C_{ij} - \sum_{k'\neq k}F_{ik'}(\boldsymbol{S}_{k'} \cdot \boldsymbol{F}_j))F_{ik}\boldsymbol{F}_j\right)$	-

1.2 Matrix factorisation with ARD and importance values

We will now explain how to extend matrix factorisation with Automatic Relevance Determination (ARD) and importance values.

ARD We change the model definition to the following.

$$\begin{aligned} D_{ij} &\sim \mathcal{N}(D_{ij}|\boldsymbol{F}_i \cdot \boldsymbol{G}_j, \tau^{-1}) \\ F_{ik} &\sim \mathcal{N}(F_{ik}|0, (\lambda_k)^{-1}) \qquad \text{or} \qquad \mathcal{E}(F_{ik}|\lambda_k) \\ G_{jk} &\sim \mathcal{N}(G_{jk}|0, (\lambda_k)^{-1}) \qquad \text{or} \qquad \mathcal{E}(G_{jk}|\lambda_k) \\ \tau &\sim \mathcal{G}(\tau|\alpha, \beta) \end{aligned}$$

Notice that the main difference is replacing λ_F and λ_G by the parameter λ_k, for $k = 1..K$. We place a Gamma prior over these variables,

$$\lambda_k \sim \mathcal{G}(\lambda_k|\alpha_0, \beta_0).$$

This can similarly be done for matrix tri-factorisation, by placing one ARD over $\boldsymbol{F}$ (using λ_k^F) and another over $\boldsymbol{G}$ (using λ_l^G).

The Gibbs sampling algorithms remain largely the same, simply replacing λ_F, λ_G in the updates by λ_k. For the multivariate posteriors, we replace $\lambda^F \boldsymbol{I}$ by $\text{diag}(\boldsymbol{\lambda})$, a diagonal matrix where the kth diagonal element is given by λ_k. The posterior Gibbs sampling distribution for λ_k itself can be derived to be another Gamma distribution,

$$p(\lambda_k|\boldsymbol{D}, \boldsymbol{F}, \boldsymbol{G}, \tau) = \mathcal{G}(\lambda_k|\alpha_0^*, \beta_0^*)$$

where

$$\alpha_0^* = \alpha_0 + \frac{I}{2} + \frac{J}{2} \qquad \beta_0^* = \beta_0 + \frac{1}{2}\sum_{i=1}^{I} F_{ik}^2 + \frac{1}{2}\sum_{j=1}^{J} G_{jk}^2$$

for the real-valued (Gaussian prior) model, and

$$\alpha_0^* = \alpha_0 + I + J \qquad \beta_0^* = \beta_0 + \sum_{i=1}^{I} F_{ik} + \sum_{j=1}^{J} G_{jk}$$

for the nonnegative (exponential prior) model. Note that each λ_k therefore depends only on the values in the kth column of $\boldsymbol{F}$ and $\boldsymbol{G}$, and those values in turn depend on the value of λ_k: if most values in that column are high, λ_k gets a small value (indicating that the factor is active); and if λ_k has a high value, the kth column will get a Gibbs sampling posterior around 0, resulting in low values (pushing the other values for this factor down, since it is inactive).

Importance value As discussed in the paper, we modify the likelihood with an importances values $\alpha^n, \alpha^l, \alpha^m$ as follows.

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{R},\boldsymbol{D},\boldsymbol{C}) \propto & \prod_{t=1}^{T} p(\boldsymbol{F}^t|\boldsymbol{\lambda}^t) \prod_{n=1}^{N} p(\tau^n) \prod_{l=1}^{L} p(\tau^l) \prod_{m=1}^{M} p(\tau^m) \\ & \times \prod_{n=1}^{N} p(\boldsymbol{R}^n|\boldsymbol{F}^{t_n},\boldsymbol{S}^n,\boldsymbol{F}^{u_n},\tau^n)^{\alpha^n} \\ & \times \prod_{l=1}^{L} p(\boldsymbol{D}^l|\boldsymbol{F}^{t_l},\boldsymbol{G}^l,\tau^l)^{\alpha^l} \\ & \times \prod_{m=1}^{M} p(\boldsymbol{C}^m|\boldsymbol{F}^{t_m},\boldsymbol{S}^m,\tau^m)^{\alpha^m} \end{aligned}$$

where $\boldsymbol{\theta}$ is the set of model parameters $(\boldsymbol{F}^t, \boldsymbol{S}^n, \boldsymbol{S}^m, \boldsymbol{G}^l, \boldsymbol{\lambda}^t, \tau^n, \tau^l, \tau^m)$. We effectively repeat the dataset $\boldsymbol{R}^n$ α^n times, and similarly for $\boldsymbol{D}^l, \boldsymbol{C}^m$, requiring a better fit to the data. The Gibbs samplers remain largely the same, with the addition of several α values in the updates. This is illustrated below for a single dataset $\boldsymbol{D} \approx \boldsymbol{F} \cdot \boldsymbol{G}^T$ with nonnegative factors, importance value α, and without ARD.

$$\begin{aligned} \tau \sim \mathcal{G}(\tau|\alpha_*,\beta_*) \qquad & \alpha_* = \alpha_\tau + \alpha\frac{|\Omega|}{2} \\ & \beta_* = \beta_\tau + \alpha\frac{1}{2}\sum_{(i,j)\in\Omega^l}(D_{ij} - \boldsymbol{F}_i \cdot \boldsymbol{G}_j)^2 \\ F_{ik} \sim \mathcal{TN}(F_{ik}|\mu_{ik},\tau_{ik}) \qquad & \mu_{ik} = \frac{1}{\tau_{ik}}\left(-\lambda_F + \alpha\tau\sum_{j\in\Omega_i^1}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})G_{jk}\right) \\ & \tau_{ik} = \alpha\tau\sum_{j\in\Omega_i^1} G_{jk}^2 \\ G_{jk} \sim \mathcal{TN}(G_{jk}|\mu_{jk},\tau_{jk}) \qquad & \mu_{jk} = \frac{1}{\tau_{jk}}\left(-\lambda_G + \alpha\tau\sum_{i\in\Omega_j^2}(D_{ij} - \sum_{k'\neq k} F_{ik'}G_{jk'})F_{ik}\right) \\ & \tau_{jk} = \alpha\tau\sum_{i\in\Omega_j^2}(F_{ik})^2 \end{aligned}$$

Similar derivations can be done for matrix tri-factorisation, and real-valued versions.

1.3 Hybrid matrix factorisation model

The hybrid matrix factorisation (HMF) model combines all the ideas presented in the previous two sections. Recall we are given three types of datasets:

1. Main datasets $\boldsymbol{R} = \{\boldsymbol{R}^1, .., \boldsymbol{R}^N\}$, relating two different entity types. Each dataset $\boldsymbol{R}^n \in \mathbb{R}^{I_{t_n} \times I_{u_n}}$ relates entity types E_{t_n}, E_{u_n}. We use matrix tri-factorisation to decompose it into two entity type factor matrices $\boldsymbol{F}^{t_n}, \boldsymbol{F}^{u_n}$, and a dataset-specific matrix $\boldsymbol{S}^n \in \mathbb{R}^{K_{t_n} \times K_{u_n}}$.
$$\boldsymbol{R}^n = \boldsymbol{F}^{t_n} \boldsymbol{S}^n (\boldsymbol{F}^{u_n})^T + \boldsymbol{E}^n.$$

2. Feature datasets $\boldsymbol{D} = \{\boldsymbol{D}^1, .., \boldsymbol{D}^L\}$, representing features for an entity type. Each dataset $\boldsymbol{D}^l \in \mathbb{R}^{I_{t_l} \times J_l}$ relates an entity type E_{t_l} to J_l features. We use matrix factorisation to decompose it into one entity type factor matrix $\boldsymbol{F}^{t_l}$, and a dataset-specific matrix $\boldsymbol{G}^l \in \mathbb{R}^{J_l \times K_{t_l}}$.
$$\boldsymbol{D}^l = \boldsymbol{F}^{t_l} (\boldsymbol{G}^l)^T + \boldsymbol{E}^l.$$

3. Similarity datasets $\boldsymbol{C} = \{\boldsymbol{C}^1, .., \boldsymbol{C}^M\}$, giving similarities between entities of the same entity type. Each dataset $\boldsymbol{C}^m \in \mathbb{R}^{I_{t_m} \times I_{t_m}}$ relates an entity type E_{t_m} to itself. We use matrix tri-factorisation to decompose it into a entity type factor matrix $\boldsymbol{F}^{t_m}$, a dataset-specific matrix $\boldsymbol{S}^m \in \mathbb{R}^{K_{t_m} \times K_{t_m}}$, and $\boldsymbol{F}^{t_m}$ again.
$$\boldsymbol{C}^m = \boldsymbol{F}^{t_m} \boldsymbol{S}^m (\boldsymbol{F}^{t_m})^T + \boldsymbol{E}^m.$$

Observed entries are given by the sets $\Omega^n = \{(i,j)|\ R^n_{ij} \text{ observed}\}$, $\Omega^l = \{(i,j)|\ D^l_{ij} \text{ observed}\}$, $\Omega^m = \{(i,j)|\ C^m_{ij} \text{ observed}\}$, respectively.

1.3.1 Model definition

The model likelihood functions are

$$\begin{aligned}
R^n_{ij} &\sim \mathcal{N}(R^n_{ij} | \boldsymbol{F}^{t_n}_i \cdot \boldsymbol{S}^n \cdot \boldsymbol{F}^{s_n}_j, (\tau^n)^{-1}) \\
D^m_{ij} &\sim \mathcal{N}(D^l_{ij} | \boldsymbol{F}^{t_l}_i \cdot \boldsymbol{G}^l_j, (\tau^l)^{-1}) \\
C^m_{ij} &\sim \mathcal{N}(C^m_{ij} | \boldsymbol{F}^{t_m}_i \cdot \boldsymbol{S}^m \cdot \boldsymbol{F}^{t_m}_j, (\tau^m)^{-1}),
\end{aligned}$$

with Bayesian priors

$$\begin{aligned}
&\tau^n, \tau^l, \tau^m \sim \mathcal{G}(\tau^* | \alpha_\tau, \beta_\tau) \\
&F^t_{ik} \sim \mathcal{E}(F^t_{ik} | \lambda^t_k) && \text{or} \quad F^t_{ik} \sim \mathcal{N}(F^t_{ik} | 0, (\lambda^t_k)^{-1}) \\
&G^l_{jk} \sim \mathcal{E}(G^l_{jl} | \lambda^{t_l}_k) && \text{or} \quad G^l_{jk} \sim \mathcal{N}(G^l_{jl} | 0, (\lambda^{t_l}_k)^{-1}) \\
&S^n_{kl} \sim \mathcal{E}(S^n_{kl} | \lambda^n_S) && \text{or} \quad S^n_{kl} \sim \mathcal{N}(S^n_{kl} | 0, (\lambda^n_S)^{-1}) \\
&S^m_{kl} \sim \mathcal{E}(S^m_{kl} | \lambda^m_S) && \text{or} \quad S^m_{kl} \sim \mathcal{N}(S^m_{kl} | 0, (\lambda^m_S)^{-1}). \\
&\lambda^t_k \sim \mathcal{G}(\lambda^t_k | \alpha_0, \beta_0).
\end{aligned}$$

Finally, we add an importance value for each of the $\boldsymbol{R}^n, \boldsymbol{D}^l, \boldsymbol{C}^m$ datasets, respectively $\alpha_n, \alpha_l, \alpha_m$.

1.3.2 Gibbs sampler

The Gibbs sampling algorithm has updates that combine the parameter values given in Tables 2, 3, and 4, for the single-dataset matrix factorisations, matrix tri-factorisations, and similarity matrix tri-factorisations. Because there are so many different parts of the models involved, careful notational definition is essential.

The datasets relating a given entity type E_t are indicated by the following sets,

$$\begin{aligned} U_1^t &= \{n \mid \boldsymbol{R}^n \in \boldsymbol{R} \ \wedge t_n = t \ \} \\ U_2^t &= \{n \mid \boldsymbol{R}^n \in \boldsymbol{R} \ \wedge u_n = t \ \} \\ V^t &= \{l \mid \boldsymbol{D}^l \in \boldsymbol{D} \ \wedge t_l = t \ \} \\ W^t &= \{m \mid \boldsymbol{C}^m \in \boldsymbol{C} \ \wedge t_m = t \ \}. \end{aligned}$$

Since the updates for the ARD can be different if a feature dataset is decomposed using negative factors or real-valued factors for $\boldsymbol{G}^l$, we also introduce the sets

$$V_+^t = \{l \in V^t \mid \boldsymbol{G}^l \text{ is nonnegative}\} \qquad V_-^t = \{l \in V^t \mid \boldsymbol{G}^l \text{ is real-valued}\}.$$

Observed entries per row i and column j are given by

$$\begin{aligned} \Omega_i^{n1} &= \{j \mid (i,j) \in \Omega^n\} & \Omega_j^{n2} &= \{i \mid (i,j) \in \Omega^n\} \\ \Omega_i^{l1} &= \{j \mid (i,j) \in \Omega^l\} & \Omega_j^{l2} &= \{i \mid (i,j) \in \Omega^l\} \\ \Omega_i^{m1} &= \{j \mid (i,j) \in \Omega^m\} & \Omega_j^{m2} &= \{i \mid (i,j) \in \Omega^m\}. \end{aligned}$$

We obtain the following posterior distributions and parameter values:

Noise parameters

$$\tau^n \sim \mathcal{G}(\tau^n|\alpha_*^n, \beta_*^n) \qquad \alpha_*^n = \alpha_\tau + \alpha^n \frac{|\Omega^n|}{2} \qquad \beta_*^n = \beta_\tau + \alpha^n \frac{1}{2} \sum_{(i,j)\in\Omega^n} (R_{ij}^n - \boldsymbol{F}_i^{t_n} \cdot \boldsymbol{S}^n \cdot \boldsymbol{F}_j^{u_n})^2$$

$$\tau^l \sim \mathcal{G}(\tau^l|\alpha_*^l, \beta_*^l) \qquad \alpha_*^l = \alpha_\tau + \alpha^l \frac{|\Omega^l|}{2} \qquad \beta_*^l = \beta_\tau + \alpha^l \frac{1}{2} \sum_{(i,j)\in\Omega^l} (D_{ij}^l - \boldsymbol{F}_i^{t_l} \cdot \boldsymbol{G}_j^l)^2$$

$$\tau^m \sim \mathcal{G}(\tau^m|\alpha_*^m, \beta_*^m) \qquad \alpha_*^m = \alpha_\tau + \alpha^m \frac{|\Omega^m|}{2} \qquad \beta_*^m = \beta_\tau + \alpha^m \frac{1}{2} \sum_{(i,j)\in\Omega^m} (C_{ij}^m - \boldsymbol{F}_i^{t_m} \cdot \boldsymbol{S}^m \cdot \boldsymbol{F}_j^{t_m})^2$$

ARD If $\boldsymbol{F}^t$ contains nonnegative factors:

$$\lambda_k^t \sim \mathcal{G}(\lambda_k^t|\alpha_k^t, \beta_k^t) \qquad \alpha_k^t = \alpha_0 + I_t + \sum_{l\in V_+^t} I_{t_l} + \sum_{l\in V_-^t} \frac{I_{t_l}}{2}$$

$$\beta_k^t = \beta_0 + \sum_{i=1}^{I_t} F_{ik} + \sum_{l\in V_+^t} \sum_{j=1}^{J_l} G_{jk} + \sum_{l\in V_-^t} \frac{1}{2} \sum_{j=1}^{J_l} G_{jk}^2.$$

If $\boldsymbol{F}^t$ contains real-valued factors:

$$\lambda_k^t \sim \mathcal{G}(\lambda_k^t|\alpha_k^t,\beta_k^t) \qquad \alpha_k^t = \alpha_0 + \frac{I_t}{2} + \sum_{l\in V_+^t} I_{t_l} + \sum_{l\in V_-^t} \frac{I_{t_l}}{2}$$

$$\beta_k^t = \beta_0 + \frac{1}{2}\sum_{i=1}^{I_t} F_{ik}^2 + \sum_{l\in V_+^t}\sum_{j=1}^{J_l} G_{jk} + \sum_{l\in V_-^t}\frac{1}{2}\sum_{j=1}^{J_l} G_{jk}^2.$$

Dataset-specific factor matrices If $\boldsymbol{G}^l, \boldsymbol{S}^n, \boldsymbol{S}^m$ contain nonnegative factors:

$$G_{jk}^l \sim \mathcal{TN}(G_{jk}^l|\mu_{jk}^l,\tau_{jk}^l) \qquad \mu_{jk}^l = \frac{1}{\tau_{jk}^l}\left(-\lambda_k^{t_l} + \tau^l\alpha^l \sum_{i\in\Omega_j^{l2}} (D_{ij}^l - \sum_{k'\neq k} F_{ik'}^{t_l} G_{jk'}^l) F_{ik}^{t_l}\right)$$

$$\tau_{jk}^l = \tau^l\alpha^l \sum_{i\in\Omega_j^{l2}} (F_{ik}^{t_l})^2$$

$$S_{kl}^n \sim \mathcal{TN}(S_{kl}^n|\mu_{kl}^n,\tau_{kl}^n) \qquad \mu_{kl}^n = \frac{1}{\tau_{kl}^n}\left(-\lambda_S^n + \tau^n\alpha^n \sum_{(i,j)\in\Omega^n} (R_{ij}^n - \sum_{(k',l')\neq(k,l)} F_{ik'}^{t_n} S_{k'l'}^n F_{jl'}^{u_n}) F_{ik}^{t_n} F_{jl}^{u_n}\right)$$

$$\tau_{kl}^n = \tau^n\alpha^n \sum_{(i,j)\in\Omega^n} (F_{ik}^{t_n})^2 (F_{jl}^{u_n})^2$$

$$S_{kl}^m \sim \mathcal{TN}(S_{kl}^m|\mu_{kl}^m,\tau_{kl}^m) \qquad \mu_{kl}^m = \frac{1}{\tau_{kl}^m}\left(-\lambda_S^m + \tau^m\alpha^m \sum_{(i,j)\in\Omega^m} (C_{ij}^m - \sum_{(k',l')\neq(k,l)} F_{ik'}^{t_m} S_{k'l'}^m F_{jl'}^{t_m}) F_{ik}^{t_m} F_{jl}^{t_m}\right)$$

$$\tau_{kl}^m = \tau^m\alpha^m \sum_{(i,j)\in\Omega^m} (F_{ik}^{t_m})^2 (F_{jl}^{t_m})^2$$

If $\boldsymbol{G}^l, \boldsymbol{S}^n, \boldsymbol{S}^m$ contain real-valued factors:

$$G^l_{jk} \sim \mathcal{N}(G^l_{jk}|\mu^l_{jk}, (\tau^l_{jk})^{-1}) \qquad \mu^l_{jk} = \frac{1}{\tau^l_{jk}} \left(\tau^l \alpha^l \sum_{i \in \Omega^{l2}_j} (D^l_{ij} - \sum_{k' \neq k} F^{t_l}_{ik'} G^l_{jk'}) F^{t_l}_{ik} \right)$$

$$\tau^l_{jk} = \lambda^{t_l}_k + \tau^l \alpha^l \sum_{i \in \Omega^{l2}_j} (F^{t_l}_{ik})^2$$

$$\boldsymbol{G}^l_j \sim \mathcal{N}(\boldsymbol{G}^l_j|\boldsymbol{\mu}^l_j, \boldsymbol{\Sigma}^l_j) \qquad \boldsymbol{\mu}^l_j = \boldsymbol{\Sigma}^l_j \cdot \left(\tau^l \alpha^l \sum_{i \in \Omega^{l2}_j} D^l_{ij} \boldsymbol{F}^{t_l}_i \right)$$

$$\boldsymbol{\Sigma}^l_j = \left(\mathrm{diag}(\boldsymbol{\lambda}^t) + \tau^l \alpha^l \sum_{i \in \Omega^{l2}_j} \boldsymbol{F}^{t_l}_i \otimes \boldsymbol{F}^{t_l}_i \right)^{-1}$$

$$S^n_{kl} \sim \mathcal{N}(S^n_{kl}|\mu^n_{kl}, (\tau^n_{kl})^{-1}) \qquad \mu^n_{kl} = \frac{1}{\tau^n_{kl}} \left(\tau^n \alpha^n \sum_{(i,j) \in \Omega^n} (R^n_{ij} - \sum_{(k',l') \neq (k,l)} F^{t_n}_{ik'} S^n_{k'l'} F^{u_n}_{jl'}) F^{t_n}_{ik} F^{u_n}_{jl} \right)$$

$$\tau^n_{jk} = \lambda^n_S + \tau^n \alpha^n \sum_{(i,j) \in \Omega^n} (F^{t_n}_{ik})^2 (F^{u_n}_{jl})^2$$

$$\boldsymbol{S}^n_k \sim \mathcal{N}(\boldsymbol{S}^n_k|\boldsymbol{\mu}^n_k, \boldsymbol{\Sigma}^n_k) \qquad \boldsymbol{\mu}^n_k = \boldsymbol{\Sigma}^n_k \cdot \left(\tau^n \alpha^n \sum_{(i,j) \in \Omega^n} (R^n_{ij} - \sum_{k' \neq k} F^{t_n}_{ik'} (\boldsymbol{S}^n_{k'} \cdot \boldsymbol{F}^{u_n}_j)) F^{t_n}_{ik} \boldsymbol{F}^{u_n}_j \right)$$

$$\boldsymbol{\Sigma}^n_k = \left(\lambda^n_S \boldsymbol{I} + \tau^n \alpha^n \sum_{(i,j) \in \Omega^n} F^{t_n}_{ik} (\boldsymbol{F}^{u_n}_j \otimes \boldsymbol{F}^{u_n}_j) \right)^{-1}$$

$$S^m_{kl} \sim \mathcal{N}(S^m_{kl}|\mu^m_{kl}, (\tau^m_{kl})^{-1}) \qquad \mu^m_{kl} = \frac{1}{\tau^m_{kl}} \left(\tau^m \alpha^m \sum_{(i,j) \in \Omega^m} (C^m_{ij} - \sum_{(k',l') \neq (k,l)} F^{t_m}_{ik'} S^m_{k'l'} F^{t_m}_{jl'}) F^{t_m}_{ik} F^{t_m}_{jl} \right)$$

$$\tau^m_{kl} = \lambda^m_S + \tau^m \alpha^m \sum_{(i,j) \in \Omega^m} (F^{t_m}_{ik})^2 (F^{t_m}_{jl})^2$$

$$\boldsymbol{S}^m_k \sim \mathcal{N}(\boldsymbol{S}^m_k|\boldsymbol{\mu}^m_k, \boldsymbol{\Sigma}^m_k) \qquad \boldsymbol{\mu}^m_k = \boldsymbol{\Sigma}^m_k \cdot \left(\tau^m \alpha^m \sum_{(i,j) \in \Omega^m} (C^m_{ij} - \sum_{k' \neq k} F^{t_m}_{ik'} (\boldsymbol{S}^m_{k'} \cdot \boldsymbol{F}^{t_m}_j)) F^{t_m}_{ik} \boldsymbol{F}^{t_m}_j \right)$$

$$\boldsymbol{\Sigma}^m_k = \left(\lambda^m_S \boldsymbol{I} + \tau^m \alpha^m \sum_{(i,j) \in \Omega^m} F^{t_m}_{ik} (\boldsymbol{F}^{t_m}_j \otimes \boldsymbol{F}^{t_m}_j) \right)^{-1}$$

Shared factor matrices If $\boldsymbol{F}^t$ contains nonnegative factors:

$$F^t_{ik} \sim \mathcal{TN}(F^t_{ik}|\mu^t_{ik}, \tau^t_{ik})$$

$$\begin{aligned}
\mu^t_{ik} =& \frac{1}{\tau^t_{ik}} \Bigg(-\lambda^t_k + \sum_{n \in U^t_1} \tau^n \alpha^n \sum_{j \in \Omega^{n1}_i} (R^n_{ij} - \sum_{k' \neq k} F^t_{ik'} (\boldsymbol{S}^n_{k'} \cdot \boldsymbol{F}^{u_n}_j))(\boldsymbol{S}^n_k \cdot \boldsymbol{F}^{u_n}_j) \\
&+ \sum_{n \in U^t_2} \tau^n \alpha^n \sum_{i' \in \Omega^{n2}_i} (R^n_{i'i} - \sum_{l \neq k} F^t_{il} (\boldsymbol{F}^{t_n}_{i'} \cdot \boldsymbol{S}^n_{\cdot,l}))(\boldsymbol{F}^{t_n}_{i'} \cdot \boldsymbol{S}^n_{\cdot,k}) \\
&+ \sum_{l \in V^t} \tau^l \alpha^l \sum_{j \in \Omega^{l1}_i} (D^l_{ij} - \sum_{k' \neq k} F^t_{ik'} G^l_{jk'}) G^l_{jk} \Bigg) \\
&+ \sum_{m \in W^t} \tau^m \alpha^m \Bigg[\sum_{j \in \Omega^{m1}_i} (C^m_{ij} - \sum_{k' \neq k} F^t_{ik'} (\boldsymbol{S}^m_{k'} \cdot \boldsymbol{F}^t_j))(\boldsymbol{S}^m_k \cdot \boldsymbol{F}^t_j) \\
&\qquad + \sum_{i' \in \Omega^{m2}_i} (C^m_{i'i} - \sum_{l \neq k} F^t_{il} (\boldsymbol{F}^t_{i'} \cdot \boldsymbol{S}^m_{\cdot,l}))(\boldsymbol{F}^t_{i'} \cdot \boldsymbol{S}^m_{\cdot,k}) \Bigg] \Bigg)
\end{aligned}$$

$$\begin{aligned}
\tau^t_{ik} =& \sum_{n \in U^t_1} \tau^n \alpha^n \sum_{j \in \Omega^{n1}_i} (\boldsymbol{S}^n_k \cdot \boldsymbol{F}^{u_n}_j)^2 + \sum_{n \in U^t_2} \tau^n \alpha^n \sum_{i' \in \Omega^{n2}_i} (\boldsymbol{F}^{t_n}_{i'} \cdot \boldsymbol{S}^n_{\cdot,k})^2 \\
&+ \sum_{l \in V^t} \tau^l \alpha^l \sum_{j \in \Omega^{l1}_i} (G^l_{jk})^2 \\
&+ \sum_{m \in W^t} \tau^m \alpha^m \Bigg[\sum_{j \in \Omega^{m1}_i} (\boldsymbol{S}^m_k \cdot \boldsymbol{F}^t_j)^2 + \sum_{i' \in \Omega^{m2}_i} (\boldsymbol{F}^t_{i'} \cdot \boldsymbol{S}^m_{\cdot,k})^2 \Bigg]
\end{aligned}$$

If $\boldsymbol{F}^t$ contains real-valued factors:

$$F_{ik}^t \sim \mathcal{N}(F_{ik}^t|\mu_{ik}^t,(\tau_{ik}^t)^{-1}) \qquad \mu_{ik}^t = \frac{1}{\tau_{ik}^t}\Bigg(\sum_{n\in U_1^t}\tau^n\alpha^n\sum_{j\in\Omega_i^{n1}}(R_{ij}^n-\sum_{k'\neq k}F_{ik'}^t(\boldsymbol{S}_{k'}^n\cdot\boldsymbol{F}_j^{u_n}))(\boldsymbol{S}_k^n\cdot\boldsymbol{F}_j^{u_n})$$
$$+\sum_{n\in U_2^t}\tau^n\alpha^n\sum_{i'\in\Omega_i^{n2}}(R_{i'i}^n-\sum_{l\neq k}F_{il}^t(\boldsymbol{F}_{i'}^{t_n}\cdot\boldsymbol{S}_{.,l}^n))(\boldsymbol{F}_{i'}^{t_n}\cdot\boldsymbol{S}_{.,k}^n)$$
$$+\sum_{l\in V^t}\tau^l\alpha^l\sum_{j\in\Omega_i^{l1}}(D_{ij}^l-\sum_{k'\neq k}F_{ik'}^tG_{jk'}^l)G_{jk}^l$$
$$+\sum_{m\in W^t}\tau^m\alpha^m\Bigg[\sum_{j\in\Omega_i^{m1}}(C_{ij}^m-\sum_{k'\neq k}F_{ik'}^t(\boldsymbol{S}_{k'}^m\cdot\boldsymbol{F}_j^t))(\boldsymbol{S}_k^m\cdot\boldsymbol{F}_j^t)$$
$$+\sum_{i'\in\Omega_i^{m2}}(C_{i'i}^m-\sum_{l\neq k}F_{il}^t(\boldsymbol{F}_{i'}^t\cdot\boldsymbol{S}_{.,l}^m))(\boldsymbol{F}_{i'}^t\cdot\boldsymbol{S}_{.,k}^m)\Bigg]\Bigg)$$

$$\tau_{ik}^t = \lambda_k^t+\sum_{n\in U_1^t}\tau^n\alpha^n\sum_{j\in\Omega_i^{n1}}(\boldsymbol{S}_k^n\cdot\boldsymbol{F}_j^{u_n})^2+\sum_{n\in U_2^t}\tau^n\alpha^n\sum_{i'\in\Omega_i^{n2}}(\boldsymbol{F}_{i'}^{t_n}\cdot\boldsymbol{S}_{.,k}^n)^2$$
$$+\sum_{l\in V^t}\tau^l\alpha^l\sum_{j\in\Omega_i^{l1}}(G_{jk}^l)^2$$
$$+\sum_{m\in W^t}\tau^m\alpha^m\Bigg[\sum_{j\in\Omega_i^{m1}}(\boldsymbol{S}_k^m\cdot\boldsymbol{F}_j^t)^2+\sum_{i'\in\Omega_i^{m2}}(\boldsymbol{F}_{i'}^t\cdot\boldsymbol{S}_{.,k}^m)^2\Bigg]$$

$$\boldsymbol{F}_i^t \sim \mathcal{N}(\boldsymbol{F}_i^t|\boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t) \qquad \boldsymbol{\mu}_i^t = \boldsymbol{\Sigma}_i^t \cdot \Bigg(\sum_{n \in U_1^t} \tau^n \alpha^n \sum_{j \in \Omega_i^{n1}} R_{ij}^n (\boldsymbol{S}^n \cdot \boldsymbol{F}_j^{u_n}) + \sum_{n \in U_2^t} \tau^n \alpha^n \sum_{i' \in \Omega_i^{n2}} R_{i'i}^n (\boldsymbol{F}_{i'}^{t_n} \cdot \boldsymbol{S}^n)$$
$$+ \sum_{l \in V^t} \tau^l \alpha^l \sum_{j \in \Omega_i^{l1}} D_{ij}^l \boldsymbol{G}_j^l$$
$$+ \sum_{m \in W^t} \tau^m \alpha^m \left[\sum_{j \in \Omega_i^{m1}} C_{ij}^m (\boldsymbol{S}^m \cdot \boldsymbol{F}_j^t) + \sum_{i' \in \Omega_i^{l2}} C_{i'i}^m (\boldsymbol{F}_{i'}^t \cdot \boldsymbol{S}^m) \right] \Bigg)$$

$$\boldsymbol{\Sigma}_i^t = \Bigg(\mathrm{diag}(\boldsymbol{\lambda}_k^t) + \sum_{n \in U_1^t} \tau^n \alpha^n \sum_{j \in \Omega_i^{n1}} (\boldsymbol{S}^n \cdot \boldsymbol{F}_j^{u_n}) \otimes (\boldsymbol{S}^n \cdot \boldsymbol{F}_j^{u_n})$$
$$+ \sum_{n \in U_2^t} \tau^n \alpha^n \sum_{i' \in \Omega_i^{n2}} (\boldsymbol{F}_{i'}^{t_n} \cdot \boldsymbol{S}^n) \otimes (\boldsymbol{F}_{i'}^{t_n} \cdot \boldsymbol{S}^n)$$
$$+ \sum_{l \in V^t} \tau^l \alpha^l \sum_{j \in \Omega_i^{l1}} (\boldsymbol{G}_j^l \otimes \boldsymbol{G}_j^l)$$
$$+ \sum_{m \in W^t} \tau^m \alpha^m \left[\sum_{j \in \Omega_i^{m1}} (\boldsymbol{S}^m \cdot \boldsymbol{F}_j^t) \otimes (\boldsymbol{S}^m \cdot \boldsymbol{F}_j^t) \right.$$
$$\left. + \sum_{i' \in \Omega_i^{m2}} (\boldsymbol{F}_{i'}^t \cdot \boldsymbol{S}^m) \otimes (\boldsymbol{F}_{i'}^t \cdot \boldsymbol{S}^m) \right] \Bigg)^{-1}$$

2 Model discussion

2.1 Software

We have provided an open-source Python implementation of all models discussed in the paper, available at `https://github.com/ThomasBrouwer/HMF`. We furthermore provide all datasets, preprocessing scripts, and Python code for the experiments. Please refer to the README in the Github project.

2.2 Complexity

The updates for the Gibbs sampler for Bayesian matrix tri-factorisation have time complexity $\mathcal{O}(IJK^2L)$, compared to $\mathcal{O}(IJK^2)$ for Bayesian matrix factorisation. For HMF the complexity becomes $\mathcal{O}((N+M+L)I^2K^3)$ where $I = \max_t I_t$ and $K = \max_t K_t$. Notice that **our model scales linearly in the number of observed datasets**. Furthermore, the random draws for columns of the factor matrices are independent of each other, and therefore the parameter updates can be formulated as efficient joint matrix operations and new values drawn in parallel. Alternatively, the draws can be done per row by using a multivariate posterior, and then all these row-wise draws can be done in parallel as well.

2.3 Missing values and predictions

Missing values can be indicated to the model through the mask sets $\Omega^n, \Omega^l, \Omega^m$. Note that this also means that if specific feature values are missing for one of the entities, these features can still be included for the other entities, simply by marking them as unobserved when we do not know their value. This is much better than imputing those values, for example using the row or column average, as the model will still fit to those imputed values.

The missing values can then be predicted, by using the posterior draws of the Gibbs sampler (after burn-in and thinning) to estimate the posteriors of the factor matrices. For example, if we wish to predict missing values in the matrix $\boldsymbol{D}^l$, we estimate $\boldsymbol{F}^{t_l}$ and $\boldsymbol{G}^l$, and predictions for the missing entries are given by $\boldsymbol{F}^{t_l} \cdot (\boldsymbol{G}^l)^T$.

2.4 Initialisation

Gibbs sampling can easily get stuck in a local minimum of posterior likelihood, and therefore initialisation of the random variables is essential to obtain a good solution. There are two obvious ways to do this. Since the user specifies the values of the hyperparameters, α_0, β_0, α_τ, β_τ, λ_S^n, λ_S^m, we can use the model definition to initialise the variables $\boldsymbol{F}^t$, $\boldsymbol{G}^l$, $\boldsymbol{S}^n$, $\boldsymbol{S}^m$, τ^n, τ^l, τ^m, λ_k^t either using the expectation of the prior model distribution, or by randomly drawing their value according to that distribution.

Alternatively, we can initialise the entity type factor matrices $\boldsymbol{F}^t$ using K-means clustering, as suggested by Ding et al. [2006], and initialise the dataset-specific matrices $\boldsymbol{S}^n, \boldsymbol{S}^m, \boldsymbol{F}^l$ using least squares. This can be done using the Moore-Penrose pseudo-inverse ($^+$), as long as the dataset-specific matrices $(\boldsymbol{S}^n, \boldsymbol{S}^m, \boldsymbol{G}^l)$ are real-valued. For example,

$$\boldsymbol{S}^n = (\boldsymbol{F}^{t_n})^+ \cdot \boldsymbol{R}^n \cdot ((\boldsymbol{F}^{u_n})^T)^+.$$

If the datasets are not real-valued, we can still initialise $\boldsymbol{S}^n$ or the other factor matrices in this way, but then set all values below zero to zero. We measure the effectiveness of the different initialisation methods in Section 4.3, which shows that this combination of K-means and least squares initialisation generally gives the fastest convergence.

2.5 Relation to tensor decomposition

Multiple matrix factorisation and tri-factorisation methods are closely linked with tensor decomposition. Here, we explore some of these connections. In particular, we show that the CANDECOMP/PARAFAC (CP, Harshman [1970]) method is a less general version of the multiple matrix tri-factorisation (MMTF) part of our HMF model; and furthermore that the Tucker Decomposition (TD, Tucker [1966]), without its orthogonality constraints, is equivalent to MMTF. All three decompositions are illustrated in Figure 1, and we define them mathematically below.

The CP method decomposes a given tensor $\boldsymbol{R} \in \mathbb{R}^{I \times J \times N}$ into the sum of K rank-1 tensors. This is effectively a generalisation of matrix factorisation to three (rather than two) dimensions, with each dimension getting its own factor matrix: $\boldsymbol{F}^1 \in \mathbb{R}^{I \times K}, \boldsymbol{F}^2 \in \mathbb{R}^{J \times K}, \boldsymbol{S} \in \mathbb{R}^{N \times K}$. Overall, we perform the factorisation $\boldsymbol{R} = \boldsymbol{F}^1 \otimes \boldsymbol{F}^2 \otimes \boldsymbol{S}$, where $\otimes$ denotes the matrix outer product. Each individual entry in $\boldsymbol{R}$ is decomposed as follows:

$$R_{ijn} = \sum_{k=1}^{K} F^1_{ik} \cdot F^2_{jk} \cdot S_{nk}. \tag{1}$$

The Tucker decomposition is defined similarly, but in addition to the three factor matrices, we also get a core tensor $\boldsymbol{G} \in \mathbb{R}^{K \times L \times Q}$, and the factor matrices have its own number of latent factors K, L, Q; $\boldsymbol{F}^1 \in \mathbb{R}^{I \times K}, \boldsymbol{S} \in \mathbb{R}^{J \times L}, \boldsymbol{F}^2 \in \mathbb{R}^{N \times Q}$. We now factorise $\boldsymbol{R} = \boldsymbol{G} \bullet_1 \boldsymbol{F}^1 \bullet_2 \boldsymbol{F}^2 \bullet_3 \boldsymbol{S}$, where $\bullet_i$ denotes the matrix dot product using the ith dimension of tensor $\boldsymbol{G}$. Individual entries in $\boldsymbol{R}$ are decomposed as:

$$R_{ijn} = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{q=1}^{Q} F^1_{ik} \cdot F^2_{jl} \cdot S_{nq} \cdot G_{klq}. \tag{2}$$

Now consider the multiple matrix tri-factorisation part of our HMF model. Say we are given N datasets $\boldsymbol{R}^n$, all spanning the same two entity types E_1, E_2, with I rows, J columns, K row factors (for entity type E_1), and L row factors (for entity type E_2). Performing MMTF on these datasets can be seen as concatenating the N matrices into one big tensor. Each

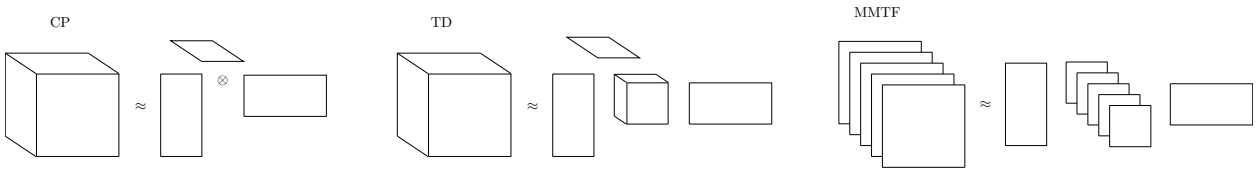


Figure 1: Overview of the CANDECOMP/PARAFAC (CP, left), Tucker Decomposition (TD, middle), and multiple matrix tri-factorisation (MMTF, right) methods. CP uses the outer product ($\otimes$), whereas TD and MMTF use the matrix product.

entry is decomposed as:

$$R_{ij}^n = \sum_{k=1}^{K} \sum_{l=1}^{L} F_{ik}^1 \cdot F_{jl}^2 \cdot S_{kl}^n. \tag{3}$$

Firstly, compare this with the TD formulation. If we define a new matrix $H_{kl}^n = \sum_{q=1}^{Q} S_{nq} \cdot G_{klq} = \boldsymbol{S}_n \cdot \boldsymbol{G}_{kl}$, we can rewrite the TD expression as:

$$R_{ijn} = \sum_{k=1}^{K} \sum_{l=1}^{L} F_{ik}^1 \cdot F_{jl}^2 \cdot H_{kl}^n. \tag{4}$$

Note that this is now equivalent to our MMTF expression in Equation 3, with $F_{ik}^1 \leftrightarrow F_{ik}^1, F_{jl}^2 \leftrightarrow F_{jl}^2, S_{kl}^n \leftrightarrow H_{kl}^n$. Since both the $\boldsymbol{S}$ and $\boldsymbol{G}$ matrices have to be inferred by our model, we can in fact merge them into one – as long as no further constraints are placed on them individually. Often in the TD method the three factor matrices $(\boldsymbol{F}^1, \boldsymbol{F}^2, \boldsymbol{S})$ have orthogonality constraints placed on them. If these constraints are dropped, TD is equivalent to MMTF.

Moving on to CP, consider constraining our MMTF model to have only diagonal entries in the $\boldsymbol{S}^n$ matrices (so that $S_{kl}^n = 0$ for $k \neq l$). Equation 3 then becomes:

$$R_{ij}^n = \sum_{k=1}^{K} F_{ik}^1 \cdot F_{jk}^2 \cdot S_{kk}^n. \tag{5}$$

This is equivalent to the CP formulation in Equation 1, with $F_{ik}^1 \leftrightarrow F_{ik}^1, F_{jk}^2 \leftrightarrow F_{jk}^2, S_{kk}^n \leftrightarrow S_{nk}$, showing that CP is a constrained version of MMTF, where the middle factor matrices S_{kl}^n are constrained to be diagonal.

Finally, we wanted to validate that the more general formulation offered by our HMF model is necessary to obtain good predictive performances, by comparing it with the CP method. We constrained our HMF model to have diagonal $\boldsymbol{S}^n$ matrices (we call this method HMF CP), but otherwise the exact same Bayesian priors and settings. The results are given in Tables 5 and 6. We can see that in the in-matrix prediction setting CP still does fairly well, although our unconstrained MMTF model (HMF D-MTF) outperforms it on all four datasets. However, in the out-of-matrix prediction setting CP does not manage to give sensible predictions, barely doing better than the gene average baseline.

Table 5: Mean squared error (MSE) of 10-fold in-matrix cross-validation results on the drug sensitivity datasets. The best performances are highlighted in bold.

Method	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
HMF D-MF	0.0775	0.0919	0.0592	**0.1062**
HMF D-MTF	**0.0768**	**0.0908**	**0.0558**	0.1073
HMF CP	0.0796	0.0913	0.0560	0.1104

Table 6: Mean squared error (MSE) of 10-fold out-of-matrix cross-validation results on the methylation datasets. The best performances are highlighted in bold.

Method	GM, PM to GE	GE, GM to PM	GE, PM to GM
Gene average	1.009	1.008	1.009
HMF D-MF	**0.788**	**0.735**	**0.602**
HMF D-MTF	0.850	0.798	0.640
HMF S-MF	0.820	0.794	0.672
HMF CP	1.006	0.972	0.968

3 Datasets and preprocessing

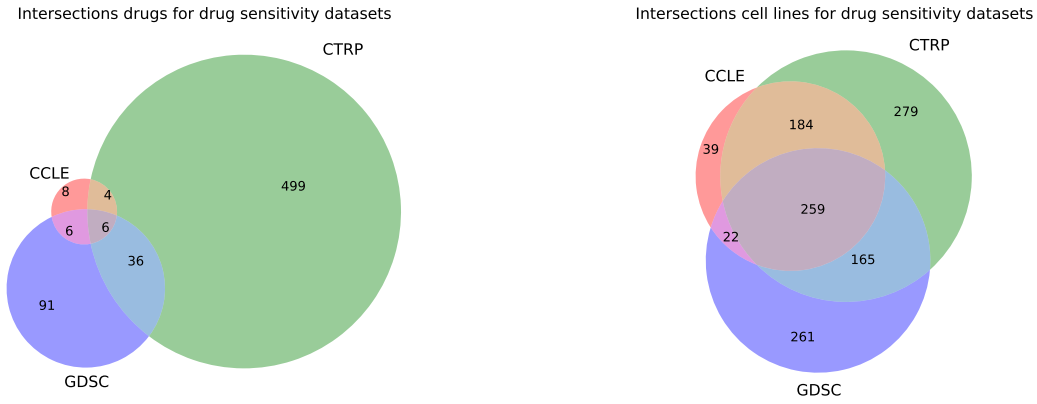
3.1 Drug sensitivity datasets

We will now describe the preprocessing steps undertaken for the drug sensitivity datasets used in the paper. We used four different datasets:

- Genomics of Drug Sensitivity in Cancer (GDSC v5.0, Yang et al. [2013]) – giving the natural log of IC_{50} values for 139 drugs across 707 cell lines, with 80% observed entries.
- Cancer Therapeutics Response Portal (CTRP v2, Seashore-Ludlow et al. [2015]) – giving EC_{50} values for 545 drugs across 887 cell lines, with 80% observed entries.
- Cancer Cell Line Encyclopedia (CCLE, Barretina et al. [2012]) – giving both IC_{50} and EC_{50} values for 24 drugs across 504 cell lines, with 96% and 63% observed entries respectively.

IC_{50} values indicate the required drug concentration needed to reduce the activity of a given cell line (cancer type in a tissue) by half. We thus measure when an undesired effect has been inhibited by half. With EC_{50} values we measure the maximal (desired) effect a drug can have on a cell line, and then measure the concentration of the drug where we achieve half of this value. In both cases, a lower value is better.

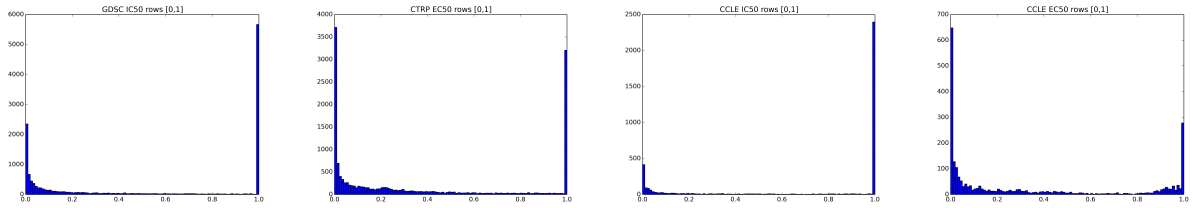
In this paper we are most interested in enhancing predictive power by integrating different datasets. Therefore we focus on drugs and cell lines for which at least two of the four datasets have values available, giving 52 drugs and 630 cell lines. Venn diagrams displaying the overlaps between drugs and cell lines are given in Figures 2a and 2b, respectively. The CTRP dataset contains a large number of small molecule probes (311) causing very little intersection with the other datasets. We also filtered out cell lines with no features available, as discussed in Subsection 3.3.



(a) Drugs

(b) Cell lines

Figure 2: Venn diagrams of the drugs and cell lines in the four datasets.



(a) GDSC IC_{50} (b) CTRP EC_{50} (c) CCLE IC_{50} (d) CCLE EC_{50}

Figure 3: Plots of the distribution of values in the drug sensitivity datasets, after capping the extremely high values in the CTRP EC_{50} and GDSC IC_{50} datasets to 20.

3.2 Preprocessing drug sensitivity values

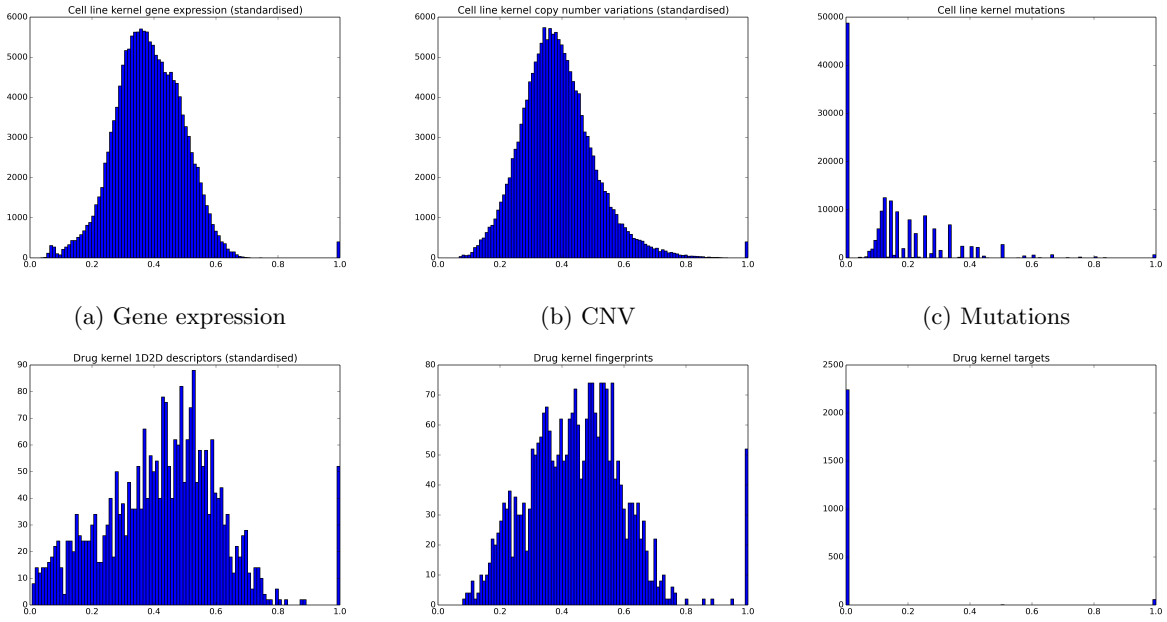
The CCLE and CTRP datasets all give the drug concentration levels, but the GDSC dataset gives the natural log transform of these values. We undo this transform by taking the exponent of each value. The drug sensitivity values for the CCLE IC_{50} and EC_{50} datasets lie in the range [0,8] and [0,10], but the other two datasets sporadically have extremely large values. This is a result of the curve fitting procedure used to approximate IC_{50} and EC_{50}, and in those cases it indicates an inefficient drug for the cell line. We cap all values above 20 to 20 to resolve this issue, and as a result obtain a similar shape of distribution of values to the CCLE datasets. Finally, we map the values in each row (per cell line) to the range [0,1]. This is shown in Figure 3, where we see that the data tends to be bimodal.

3.3 Features

We also want to incorporate feature information, which is readily provided by the GDSC dataset for 399 of the 630 cell lines. This gave us gene expression information (13321 genes, positive values), copy number variations (CNV; 426 features, count data), and mutations (82, binary data). We filtered out cell lines without this information.

For the drugs we used the PubChem Identifier Exchange Service (`https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi`) to obtain the PubChem identifiers for all the drugs. Where there were multiple, we used the one in the GDSC database, or otherwise the first one. We then used the PaDeL-Descriptor software (`http://www.yapcwsoft.com/dd/padeldescriptor/`) to extract 1D and 2D descriptors, as well as Pubchem fingerprints of functional groups in the drugs. The GDSC dataset has drug target information available for 48 out of the 52 drugs, which we extracted from their website and encoded as a binary dataset. For the four remaining targets we mark the entries as unknown using the mask matrix in the feature dataset and kernel. We removed features with the same value across all drugs.

For each of the feature datasets we constructed a similarity kernel. For binary data we used a Jaccard kernel, and for real-valued data we first standardised each feature to have zero mean and unit variance, and then used a Gaussian kernel to compute similarities, with as variance parameter the number of features. The resulting distributions of kernel similarity values are given in Figure 4. We found that adding the similarity kernels in our HMF model



(a) Gene expression (b) CNV (c) Mutations

(d) 1D 2D descriptors (e) Fingerprints (f) Targets

Figure 4: Plots of the distribution of similarity values in the kernel matrices based on the feature datasets. The similarity scores based on drug fingerprints, drug targets, and cell line mutations were constructed using a Jaccard kernel. The scores for gene expression data, copy number variations, and 1D and 2D descriptors were computed using a Gaussian kernel after standardising the values per feature (zero mean, unit variance) and using the number of features as the variance parameter.

did not improve the predictions. This is because adding dissimilar datasets makes it very hard (if not impossible) to find a good solution, and instead converges to a bad one.

The datasets are summarised in Table 7, and represented graphically in Figure 5. Of the remaining datasets (52 drugs by 399 cell lines), 95.1% of the entries have a value in at least one of the four datasets, and 62.9% have an entry in two or more. GDSC IC_{50} has 67.9% observed entries, CTRP EC_{50} has 72.3%, CCLE IC_{50} has 18.8%, and CCLE EC_{50} has 11.4%. This is also shown in Figure 5. Individually, the GDSC dataset contains entries for 48 drugs and 399 cell lines (73.6% observed), CTRP spans 46 drugs and 379 cell lines (86.0% observed), and finally CCLE IC_{50} and EC_{50} have 16 drugs and 253 cell lines (96.4% and 58.6% observed).

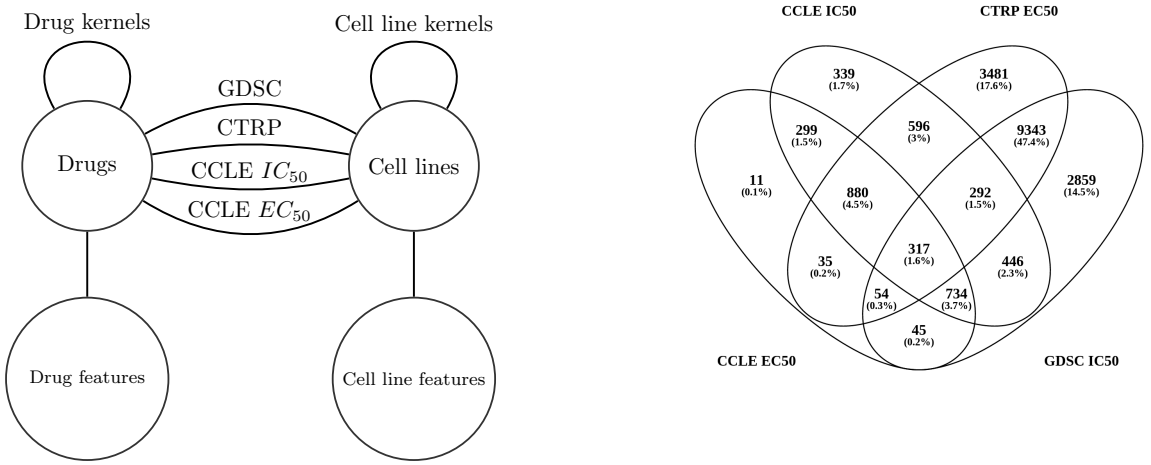


Figure 5: The datasets used for drug sensitivity prediction on the left; and the overlap of values between the four datasets on the right.

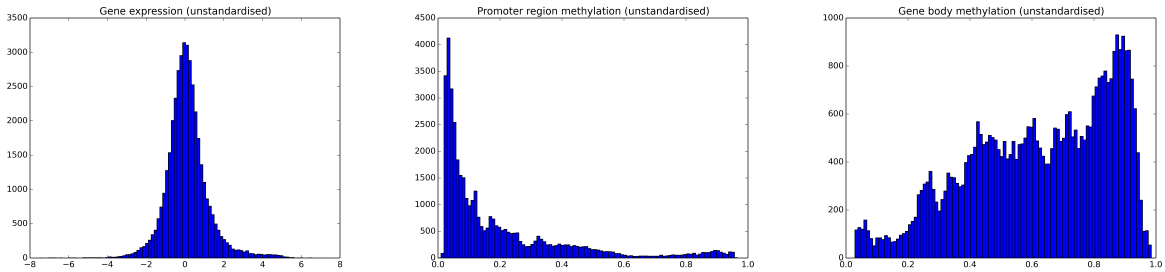
Table 7: Drug sensitivity datasets and feature datasets, summarising the entities they relate, the sizes of the datasets, and the fraction of observed entries.

Dataset	**Entities**	**Size**	**Observed**
GDSC IC_{50}	Cell lines, drugs	399, 52	67.9%
CTRP EC_{50}	Cell lines, drugs	399, 52	72.3%
CCLE IC_{50}	Cell lines, drugs	399, 52	18.8%
CCLE EC_{50}	Cell lines, drugs	399, 52	11.4%
Targets	Drugs, targets	52, 53	92.3%
1D2D	Drugs, 1D 2D descriptors	52, 1160	100%
FP	Drugs, fingerprints	52, 495	100%
Targets kernel	Drugs, drugs	52, 52	85.2%
1D2D kernel	Drugs, drugs	52, 52	100%
FP kernel	Drugs, drugs	52, 52	100%
CNV	Cell lines, CNVs	399, 426	100%
Mutations	Cell lines, mutations	52, 82	100%
Gene expression kernel	Cell lines, cell lines	399, 399	100%
CNV kernel	Cell lines, cell lines	399, 399	100%
Mutations kernel	Cell lines, cell lines	399, 399	100%

3.4 Methylation and gene expression datasets

The preprocessing for the methylation and gene expression datasets is much simpler. We obtained three datasets from the The Cancer Genome Atlas (TCGA, Koboldt et al. [2012]): promoter-region methylation (PM), gene body methylation (GM), and gene expression (GE) profiles for 254 breast cancer patients. This dataset originally spanned 13966 genes, but we focused on 160 breast cancer driver genes, given by the IntOGen database (Gonzalez-Perez et al. [2013]). We standardised each of the datasets to have zero mean and unit standard deviation per gene. Plots of the datasets containing the 160 genes, before and after standardising, can be found in Figure 6.

To construct the similarity kernels for the HMF S-MF model, giving the similarity of samples, we used a Gaussian kernel with $\sigma^2 =$ no. genes. The kernel value distributions are plotted in Figure 7. Notice the Gaussian-like distribution of values between [0,1]. If we generate values randomly from our HMF model's probabilistic definition (for example using value 1 for all hyperparameters, and then randomly sampling values from the prior distributions), we obtain a similar distribution of values. σ^2 was chosen in such a way to match the model definition more closely.



(a) Gene expression, unstandardised

(b) Promoter-region methylation, unstandardised

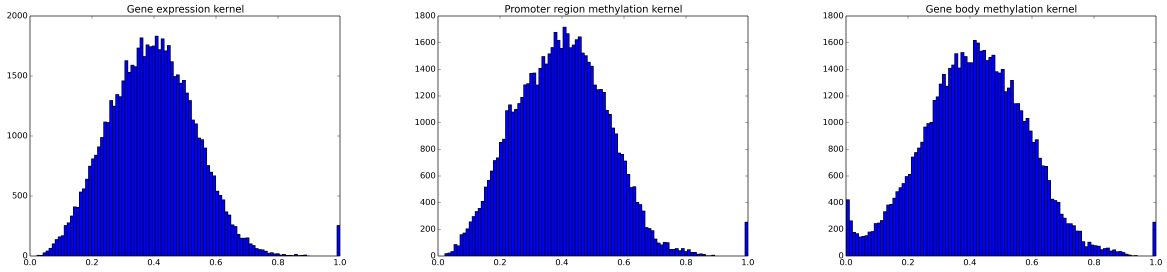
(c) Gene body methylation, unstandardised

(d) Gene expression, standardised

(e) Promoter-region methylation, standardised

(f) Gene body methylation, standardised

Figure 6: Plots of the distribution of the methylation datasets, before standardising (top row) and after (bottom row).



(a) Gene expression kernel (b) Promoter-region methylation kernel (c) Gene body methylation kernel

Figure 7: Plots of the distribution of the methylation kernels.

4 Additional experiments

We performed several additional experiments on the drug sensitivity and methylation datasets, to explore the advantages and limits of our models. In particular, we looked at: the run-time performance of the models; the effectiveness of automatic model selection using the Automatic Relevance Determination (ARD) Bayesian prior; the best initialisation approach; the best values to use for the dataset importances α; and the trade-offs of our hybrid choices: effects of factorisation types and nonnegativity constraints on predictive performance.

4.1 Run-time comparison

In this section we give a rough indication of the run-time performances of the models we compared in our experiments. We compare the time it takes for most of the matrix factorisation methods (HMF D-MF, HMF D-MTF, BNMF, BNMTF, NMF, NMTF, and Multiple NMF) on the four drug sensitivity datasets. In particular, we give the number of iterations we used in our experiments for each model, the total run-time it took to train each model, and the time per iteration of the model updates. We use $K_t = 10, K = 10, L = 10$ for all models.

The run-time performances on the drug sensitivity datasets are given in Table 8. Often the matrix factorisation variants are faster than the matrix tri-factorisation versions (up to six times, in the case of NMF and NMTF). The non-probabilistic methods give the fastest run-time (both per iteration, and total), but our experiments show that their predictive performance is worse and they are more prone to overfitting. Our HMF models incur a slightly higher run-time than the other Bayesian models (BNMF, BNMTF), due to having to consider all four datasets at the same time, but only by a factor of roughly two (despite there being four datasets in total).

Note that when we use cross-validation to measure the predictive performances, our HMF models do not need to run nested cross-validation to choose the dimensionality (K_t), whereas the other matrix factorisation models (BNMF, BNMTF, NMF, NMTF, Multiple NMF) do. This actually makes our models faster in cross-validation.

Table 8: Run-time performances of the matrix factorisation models on each of the four drug sensitivity datasets, giving the total number of iterations (It) used to train each model in our experiments (the same on each dataset), and then for each dataset the average number of seconds per iteration (s / it), and the total time in seconds to train a single model (Total).

		GDSC IC_{50}		CTRP EC_{50}		CCLE IC_{50}		CCLE EC_{50}	
Model	Iterations	s / it	Total	s / it	Total	s / it	Total	s / it	Total
HMF D-MF	200	0.120	23.9	0.119	23.8	0.058	11.6	0.052	10.5
HMF D-MTF	200	0.303	60.6	0.293	58.6	0.148	29.5	0.148	29.6
BNMF	1000	0.061	61.2	0.059	58.8	0.034	34.3	0.037	37.0
BNMTF	500	0.112	56.1	0.107	53.5	0.052	26.1	0.056	28.1
NMF	1000	0.007	6.9	0.007	6.6	0.002	2.3	0.002	2.0
NMTF	1000	0.037	36.6	0.033	33.4	0.009	9.3	0.010	9.8
Multiple NMF	1000	0.017	16.9	0.017	17.0	0.012	11.6	0.011	11.0

All run-time experiments were conducted on a MacBook Pro laptop, with 2.2 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 memory, and an Intel Iris Pro 1536 MB Graphics card.

4.2 Model selection

Our model employs the ARD prior to perform automatic model selection. In this experiment, we verify how effective this is. We repeat the in-matrix cross-validation experiments on the four drug sensitivity experiments, for the HMF D-MF (multiple matrix factorisation) and HMF D-MTF (multiple matrix tri-factorisation) models. We vary the values for K_t, using the values [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30].

Usually when we add more factors to a matrix factorisation models, it gives the model more freedom to fit well to the data, eventually leading to overfitting. Hence, the average cross-validation performance should initially go down, and then go up again as it starts overfitting. If the ARD works as desired, and we add more factors (K_t increases), the ARD will turn them off and use a similar number of factors to before. This should result in less overfitting than the equivalent model with no ARD, resulting in a flatter curve going back up as the number of factors increase.

We compare the effects of adding ARD to the HMF model in Figure 8. Here, we clearly see that adding ARD to the model consistently reduces overfitting on all four drug sensitivity datasets. ARD is not perfect, and we can still see that the curve goes up as the values for K_t increase, but this overfitting is significantly less severe than the models without ARD.

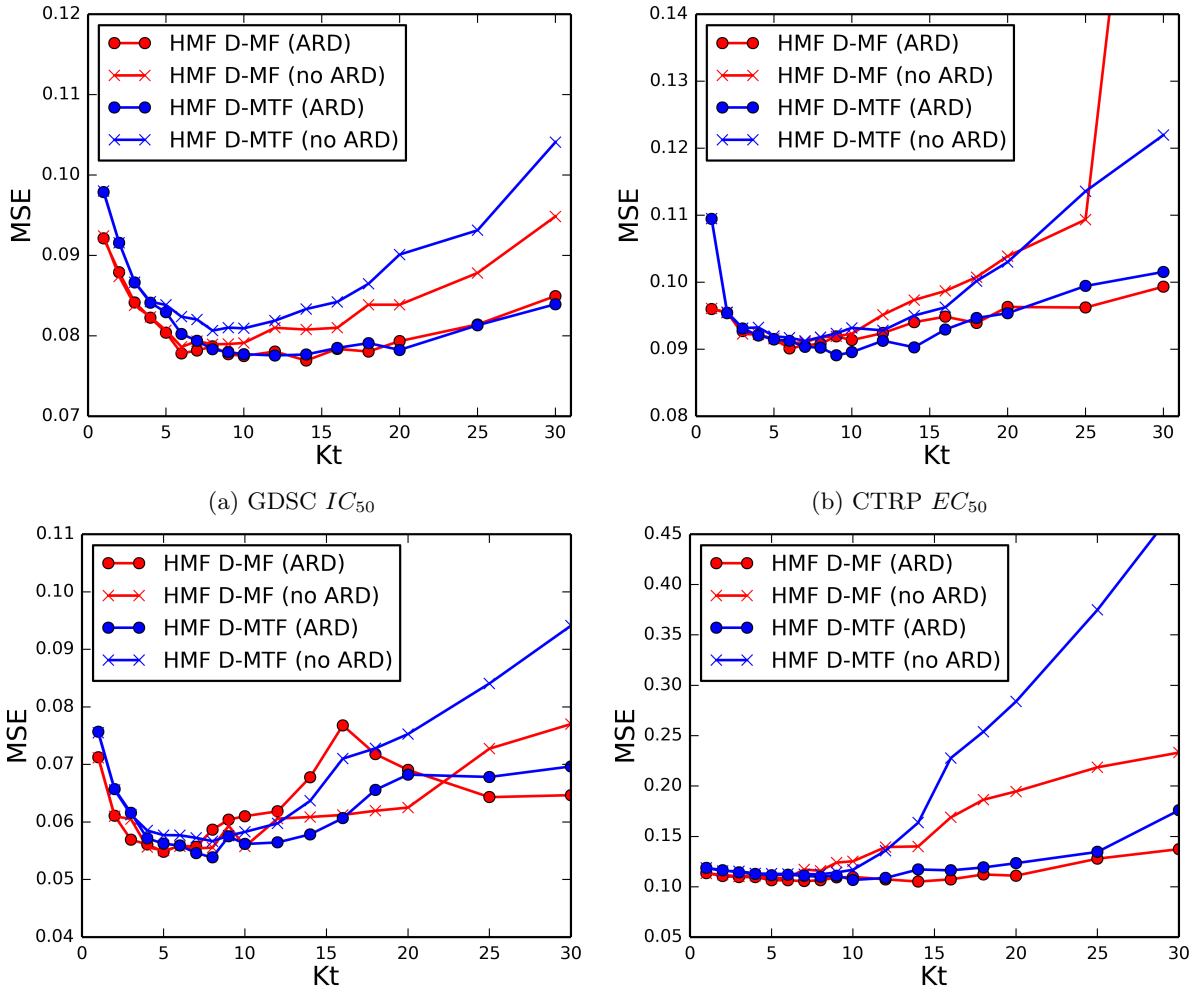
Usage recommendations Always use ARD to reduce overfitting in the model. Even though ARD does not always entirely eliminate the need for model selection (there is still some overfitting as K_t becomes very large), it generally makes it much less critical to try a large range of dimensionalities to find the best one. Instead, trying one or a couple will prove just as effective.

4.3 Initialisation

The initialisation method can have a huge impact on performance. Initialise too well, and it leads to overfitting very quickly. Initialise poorly, and your model may not converge to a good solution.

As discussed in Section 2.4, there are several ways to initialise the Gibbs sampling parameter values. Here, we measure the convergence of the HMF D-MF and HMF D-MTF models (nonnegative shared factors, real-valued private factors) on the drug sensitivity datasets. We try the following initialisation approaches:

1. **Exp:** All parameters initialised using expectation.
2. **Random:** ARD initialised using expectation, all other parameters using random draws.



(a) GDSC IC_{50}

(b) CTRP EC_{50}

(c) CCLE IC_{50}

(d) CCLE EC_{50}

Figure 8: Graphs showing the cross-validation performance of in-matrix predictions on the drug sensitivity datasets, where we vary the dimensionality K_t for our HMF models (HMF D-MF in red, HMF D-MTF in blue), both for HMF with ARD (o), and without (x). Adding ARD clearly reduces overfitting as K_t increases.

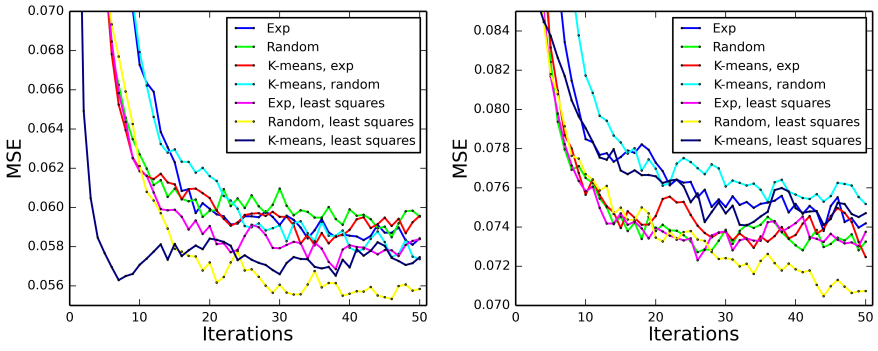
3. ***K*-means, exp:** Entity type factor matrices F^t initialised using K-means, all other parameters using expectation.

4. ***K*-means, random:** Entity type factor matrices F^t initialised using K-means, ARD initialised using expectation, and all other factor matrices using random draws.

5. **Exp, least squares:** Entity type factor matrices F^t and ARD initialised using expectation, and all other factor matrices using least squares.

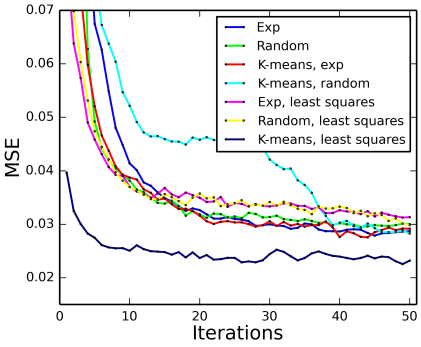
6. **Random, least squares:** Entity type factor matrices F^t initialised using random draws, ARD initialised using expectation, and all other factor matrices using least squares.

7. ***K*-means, least squares:** Entity type factor matrices F^t initialised using K-means, ARD initialised using expectation, and all other factor matrices using least squares.

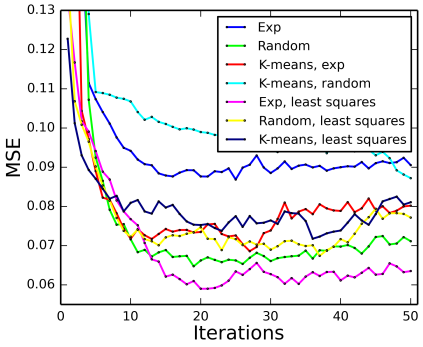


(a) HMF D-MF, GDSC IC_{50}

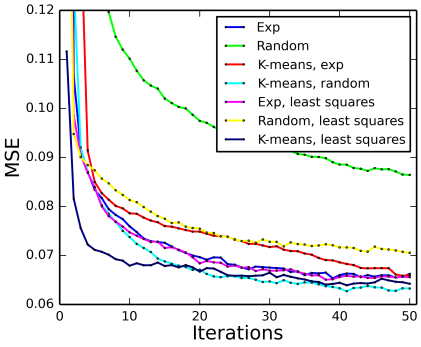
(b) HMF D-MF, CTRP EC_{50}



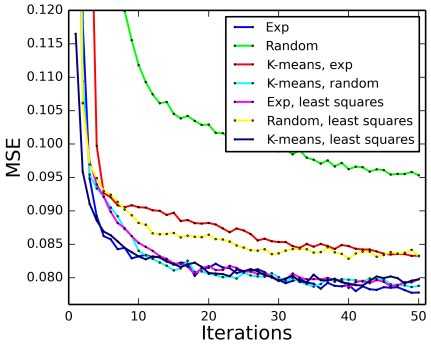
(c) HMF D-MF, CCLE IC_{50}



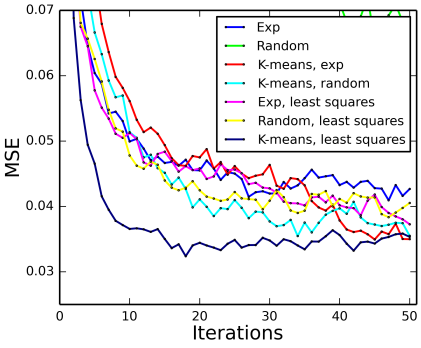
(d) HMF D-MF, CCLE EC_{50}



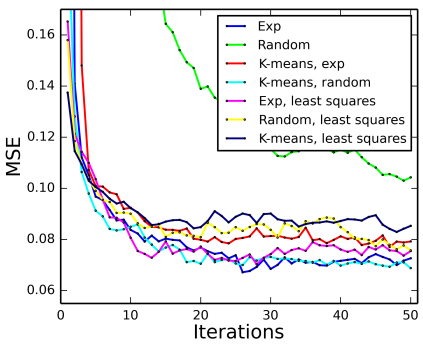
(e) HMF D-MTF, GDSC IC_{50}



(f) HMF D-MTF, CTRP EC_{50}



(g) HMF D-MTF, CCLE IC_{50}



(h) HMF D-MTF, CCLE EC_{50}

Figure 9: Graphs showing the convergence of the HMF D-MF (top two rows) and HMF D-MTF (bottom two rows) models on the four drug sensitivity datasets, for the seven different initialisation approaches.

The plots of convergence for HMF D-MF and HMF D-MTF are given in Figure 9. We can see that the K-means with least squares initialisation strategy (dark blue) provides the fastest convergence on half of the datasets, oftentimes significantly faster. Random for $\boldsymbol{F}^t$ and least squares also performs well. The other strategies sometimes provide faster convergence, but none of them do so consistently.

Recommendation The fastest convergence is generally provided by combining K-means initialisation for $\boldsymbol{F}^t$ with least squares for the other factor matrices. Random initialisation for $\boldsymbol{F}^t$ with least squares also works well.

4.4 Importance value

We experimented with different values for the importance values α, for the out-of-matrix prediction setting. Specifically, we consider the case where we predict one dataset (either gene expression, promoter-region methylation, or gene body methylation) using the other two datasets as additional datasets.

We vary the value of α for the dataset we are trying to predict (α_0) as well as for the two other datasets we are learning from (α_1, α_2), using the values $[0.25, 0.5, 1.0, 1.5, 2.0]$. We use $K_t = 10$ for all experiments, non-negative factors for the shared factor matrices, and real-valued for the private ones. For initialisation we use K-means for the shared factor matrices, and least squares for the private (real-valued) ones.

We perform 10-fold cross-validation, taking out 10% of the samples each time for the dataset we are trying to predict, and then measuring the mean squared error (MSE) of predictions. The average performances can be found in Table 9, for both HMF D-MF (multiple matrix factorisation) and HMF D-MTF (multiple matrix tri-factorisation).

For the HMF D-MF model (left column) we see that datasets with low predictivity (such as GE and PM – note the high MSE) have the best parameter values when the importance of the dataset to be predicted (α_0) is low, and the importance of the datasets to learn from (α_1, α_2) is high. This is presumably because higher importance values lead to a better fit to the data, and if there is low predictivity, we should not fit to the data too much (otherwise we might overfit). In contrast, when the predictivity is high (such as GM), the importance value for all datasets should not be too low, because this results in a poor fit to the data and hence poor predictions.

For the HMF D-MTF model (right column) we see a similar effect, in that if the predictivity is better, the best values for the importance increase. However, for this approach all of the importance values should generally be set low if the datasets are different.

We conducted the same experiment for the approach based on similarity kernels, where we use the same ones as used in the out-of-matrix predictions from the paper. For the dataset we are trying to predict we decompose it using matrix factorisation (HMF S-MF). Matrix tri-factorisation could have also been chosen, but because the third dataset is not shared in this case, it is equivalent to matrix factorisation (in fact, we get very similar tables as the

ones shown for HMF S-MF). As before, we use nonnegative factors for the shared matrices, real-valued factors for the private ones, and K-means and least squares for initialisation. The results can be found in Table 10. Here, we see that the importance value is much less important, as long as it is not lower than 1.0.

Recommendation If the datasets are very dissimilar and have low predictivity, use a low importance value for the main dataset for which we are trying to predict values (like 0.5). When using HMF D-MF, use a higher importance value for the other datasets (like 1.5), but when using HMF D-MTF, use a low importance value as well (like 0.5). If the datasets are more similar, use normal importance values (1.0) for all datasets. Finally, when using similarity kernels (HMF S-MF), use the normal importance value for the kernels (1.0).

Table 9: Performances of out-of-matrix cross-validation results for HMF D-MF (left column) and D-MTF (right column), where we vary the importance value for the dataset we are trying to predict (α_0), and for the other two datasets we are learning from (α_1, α_2). We have three different datasets (gene expression, GE; gene body methylation, GM; and promoter region methylation, PM). We therefore have three different prediction settings. We have highlighted the most promising parameter value areas in green, and the least promising in red.

HMF D-MF, GM + PM → GE					
	GM (α_1), PM (α_2)				
GE (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.878	0.843	0.835	0.831	0.832
0.5	0.845	0.849	0.829	0.832	0.829
1.0	0.848	0.916	1.195	1.322	0.831
1.5	0.871	0.948	1.288	1.407	1.470
2.0	0.896	0.966	1.392	1.711	1.782

HMF D-MTF, GM + PM → GE					
	GM (α_1), PM (α_2)				
GE (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.866	0.906	0.941	0.950	0.954
0.5	0.874	0.870	0.914	0.941	0.945
1.0	0.933	0.985	0.940	0.942	0.961
1.5	0.959	0.958	1.065	1.026	1.000
2.0	0.957	1.080	1.211	1.145	1.147

HMF D-MF, GE + GM → PM					
	GE (α_1), GM (α_2)				
PM (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.859	0.799	0.795	0.799	0.798
0.5	0.811	0.812	0.783	0.783	0.784
1.0	0.789	0.814	0.987	0.788	0.783
1.5	0.784	0.809	1.070	1.247	0.870
2.0	0.798	0.835	1.111	1.280	1.255

HMF D-MTF, GE + GM → PM					
	GE (α_1), GM (α_2)				
PM (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.862	0.893	0.943	0.946	0.947
0.5	0.841	0.885	0.898	0.943	0.944
1.0	0.852	0.848	1.012	0.885	0.904
1.5	0.876	0.866	0.900	1.080	0.904
2.0	0.884	0.881	0.900	1.096	1.099

HMF D-MF, GE + PM → GM					
	GE (α_1), PM (α_2)				
GM (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.790	0.708	0.703	0.697	0.701
0.5	0.724	0.670	0.687	0.688	0.685
1.0	0.698	0.657	0.659	0.670	0.685
1.5	0.698	0.655	0.667	0.657	0.666
2.0	0.689	0.668	0.671	0.676	0.665

HMF D-MTF, GE + PM → GM					
	GE (α_1), PM (α_2)				
GM (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.773	0.775	0.851	0.854	0.861
0.5	0.761	0.746	0.774	0.818	0.850
1.0	0.782	0.757	0.754	0.786	0.779
1.5	0.832	0.752	0.745	0.783	0.795
2.0	0.836	0.811	0.802	0.791	0.805

Table 10: Performances of out-of-matrix cross-validation results for HMF S-MF, where we vary the importance value for the dataset we are trying to predict (α_0), and for the other two datasets we are learning from (α_1, α_2). We have three different datasets (gene expression, GE; gene body methylation, GM; and promoter region methylation, PM). We therefore have three different prediction settings. We have highlighted the most promising parameter value areas in green, and the least promising in red.

HMF S-MF, GM + PM → GE					
	GM (α_1), PM (α_2)				
GE (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.873	0.870	0.871	0.880	0.879
0.5	0.852	0.850	0.853	0.852	0.849
1.0	0.839	0.835	0.846	0.843	0.842
1.5	0.872	0.839	0.837	0.842	0.840
2.0	0.945	0.849	0.834	0.845	0.833

HMF S-MF, GE + GM → PM					
	GM (α_1), PM (α_2)				
GE (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.858	0.858	0.866	0.859	0.854
0.5	0.829	0.830	0.823	0.832	0.828
1.0	0.824	0.813	0.815	0.814	0.814
1.5	0.853	0.804	0.806	0.812	0.808
2.0	0.898	0.858	0.810	0.807	0.809

HMF S-MF, GE + PM → GM					
	GM (α_1), PM (α_2)				
GE (α_0)	0.25	0.5	1.0	1.5	2.0
0.25	0.762	0.771	0.782	0.796	0.797
0.5	0.722	0.724	0.740	0.742	0.753
1.0	0.701	0.706	0.707	0.718	0.721
1.5	0.756	0.703	0.702	0.709	0.710
2.0	0.757	0.710	0.702	0.697	0.701

4.5 Negativity constraints

Negativity constraints have two advantages. Firstly, they make it easier to analyse the factor values after performing matrix factorisation, and for example identify clusters in the data. With real-valued factors this can be a lot more complicated. Secondly, the nonnegativity constraints can reduce overfitting on sparse or noisy datasets.

We measured the effects of nonnegativity constraints on our HMF models. On the drug sensitivity datasets we tried three variants: nonnegative (all factor matrices are nonnegative), real-valued (all factor matrices are real-valued), and semi-nonnegative (shared factor matrices are nonnegative, dataset-specific ones are real-valued). On the gene expression and methylation datasets we tried the real-valued and semi-nonnegative versions. For the variants containing real-valued factor matrices, we also see whether there is a difference between row-wise and column-wise posterior draws.

We ran 10-fold cross-validation for in-matrix predictions of the drug sensitivity datasets. We use $K_t = 10$, K-means initialisation for the $\boldsymbol{F}^t$ matrices, and least squares initialisation for the other matrices. Results are given in Table 11, where we see that the entirely real-valued models consistently outperform all other versions, for both HMF D-MF and D-MTF. In addition, row-wise draws perform better than column-wise ones.

Similarly, we ran 10-fold cross-validation for out-of-matrix predictions of the drug sensitivity datasets. We ran the HMF D-MF, HMF D-MTF, and HMF S-MF models, again with $K_t = 10$, and K-means and least squares initialisation. For HMF D-MF we used importance value 0.5 for all three datasets. For HMF D-MTF, 1.5 for the main dataset we are trying to predict, and 0.5 for the others. For HMF S-MF we used 1.0 for all datasets. These results are given in Table 12, showing again that the real-valued version performs better for HMF D-MF, D-MTF, and S-MF. Column draws sometime do best, but generally row-wise draws are the best options.

Although nonnegativity can reduce the chance for overfitting, it comes at the cost of worse fitting to the data, as the nonnegativity makes convergence harder. This probably explains the lower predictive performance in this experiment for the nonnegative and semi-nonnegative models. The Bayesian nature of the models already reduces overfitting, potentially making the nonnegativity redundant. However, if the data is very sparse, and hence overfitting is more likely, the nonnegativity could be a great option.

To explore the advantages of nonnegativity in sparse settings, we repeat the experiments for sparse data predictions that were performed in the main paper on the CTRP drug sensitivity dataset, comparing the five different versions of HMF D-MF and D-MTF from the previous section. We vary the fraction of observed data, splitting the data randomly into train and test 10 times, and taking the average performance of predictions. This is shown for both methods in Figure 10. We do see that when the sparsity increases to very high levels like 90%, the nonnegative model (in dark blue) outperforms the real-valued model with row draws (in light blue). This is especially true for the HMF D-MTF model.

Table 11: Performances of in-matrix cross-validation results for HMF D-MF (top table) and D-MTF (bottom table) on the drug sensitivity datasets, where we vary the nonnegativity of the matrices. We try nonnegative, semi-nonnegative, and real-valued variants, and for the real-varied variants try both row-wise draws and column-wise draws. The best performances are highlighted in bold.

HMF D-MF		MSE			
$\boldsymbol{F}_t$	$\boldsymbol{G}_l$	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
Nonnegative	Nonnegative	0.0783	0.0907	0.0544	0.1083
Nonnegative	Real-valued (row)	0.0764	0.0899	0.0541	0.1069
Nonnegative	Real-valued (column)	0.0760	0.0903	0.0572	0.1050
Real-valued (row)	Real-valued (row)	**0.0758**	**0.0878**	**0.0519**	**0.1028**
Real-valued (column)	Real-valued (column)	0.0761	0.0889	0.0521	0.1029

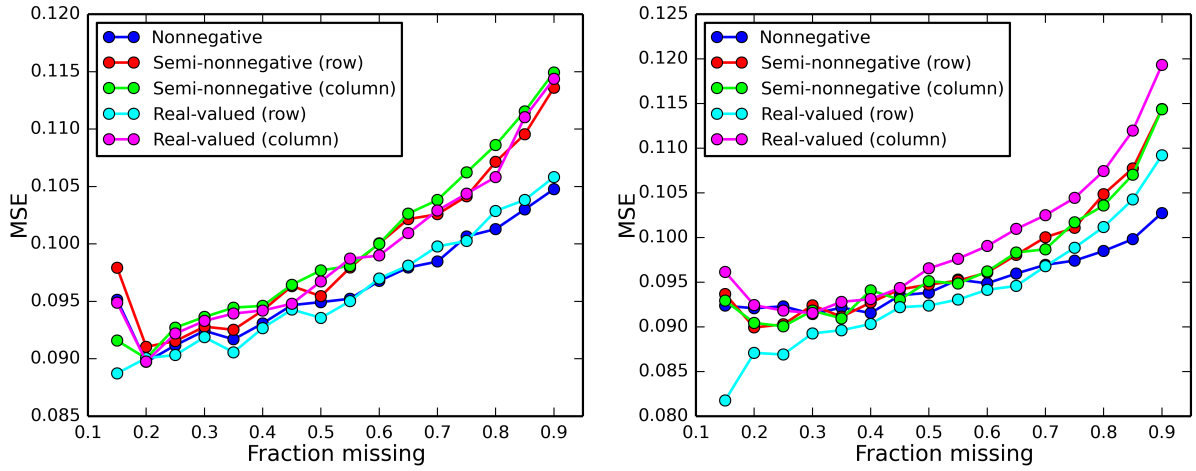
HMF D-MTF		MSE			
$\boldsymbol{F}_t$	$\boldsymbol{S}_n$	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
Nonnegative	Nonnegative	0.0802	0.0916	0.0573	0.1133
Nonnegative	Real-valued (row)	0.0773	0.0899	0.0558	0.1110
Nonnegative	Real-valued (column)	0.0780	0.0904	0.0549	0.1116
Real-valued (row)	Real-valued (row)	**0.0770**	**0.0883**	**0.0535**	**0.1011**
Real-valued (column)	Real-valued (column)	0.0799	0.0904	0.0572	0.1101

Table 12: Performances of out-of-matrix cross-validation results for HMF D-MF (top table), D-MTF (middle table), and S-MF (bottom table), on the gene expression and methylation datasets, where we vary the nonnegativity of the matrices. We try semi-nonnegative and real-valued variants, and also row-wise draws and column-wise draws. The best performances are highlighted in bold.

HMF D-MF		MSE		
$\boldsymbol{F}_t$	$\boldsymbol{G}_l$	GM, PM to GE	GE, GM to PM	GE, PM to GM
Nonnegative	Real-valued (row)	0.880	0.792	0.670
Nonnegative	Real-valued (column)	0.851	0.803	0.672
Real-valued (row)	Real-valued (row)	**0.834**	0.775	0.651
Real-valued (column)	Real-valued (column)	0.839	**0.769**	**0.647**

HMF D-MTF		MSE		
$\boldsymbol{F}_t$	$\boldsymbol{S}_n$	GM, PM to GE	GE, GM to PM	GE, PM to GM
Nonnegative	Real-valued (row)	0.959	0.864	0.755
Nonnegative	Real-valued (column)	0.985	0.869	0.777
Real-valued (row)	Real-valued (row)	**0.893**	**0.822**	0.756
Real-valued (column)	Real-valued (column)	0.909	0.837	**0.738**

HMF S-MF		MSE		
$\boldsymbol{F}_t$	$\boldsymbol{S}_n$	GM, PM to GE	GE, GM to PM	GE, PM to GM
Nonnegative	Real-valued (row)	0.846	0.818	0.713
Nonnegative	Real-valued (column)	0.836	0.811	0.723
Real-valued (row)	Real-valued (row)	**0.815**	**0.805**	**0.697**
Real-valued (column)	Real-valued (column)	0.849	0.823	0.721



(a) HMF D-MF

(b) HMF D-MTF

Figure 10: Graphs showing the performance for the different negativity options for the HMF D-MF (left) and HMF D-MTF (right) models, as the sparsity of the CTRP dataset increases. We plot the average mean squared error (MSE) across 10 random training and data splits.

Recommendation It is generally best use Gaussian priors for all factor matrices, resulting in entirely real-valued models. Row-wise draws most often give better performances than column-wise draws. In the case of very sparse matrices, nonnegative models can reduce overfitting.

4.6 Factorisation types

Finally, we experiment with the choice of factorisationt types. Recall that each dataset can be factorised either using matrix factorisation ($\boldsymbol{D}_l$) or matrix tri-factorisation ($\boldsymbol{R}_n$). For the drug sensitivity dataset, there are therefore a number of hybrid factorisation possibilities: using matrix factorisation for all (as used in the main paper; HMF D-MF), using matrix tri-factorisation for all (HMF D-MTF), using matrix factorisation on one dataset and tri-factorisation for the other three, or using matrix factorisation on two datasets and tri-factorisation on two as well. Note that applying matrix tri-factorisation on only one dataset is equivalent to using matrix factorisation on all four (since the second factor matrix is not shared with any other dataset). In this section we explore some of these choices.

Methylation data

We firstly consider the methylation datasets (GE, GM, and PM). We computed the Spearman correlation of values in each pair of these datasets, as given in Table 13. Here we see that GM and PM are (weakly) positively correlated, and GE and PM are (weakly negatively) correlated. We then performed 10-fold out-of-matrix cross-validation experiments (as before), varying the factorisation types on the three datasets. For simplicity we used $K_t = 10$ and $\alpha = 0.5$ for all datasets, K-means initialisation for the shared factor matrices ($\boldsymbol{F}_t$) and least squares for the private ones ($\boldsymbol{G}_l, \boldsymbol{S}_n$).

The results are given in Table 14, where the left three columns indicate the factorisation

Table 13: Spearman correlation between values of each of the dataset pairs.

	GE	GM	PM
GE	-	-0.07	-0.12
GM	-0.07	-	0.14
PM	-0.12	0.14	-

Table 14: Performances of out-of-matrix cross-validation results on the methylation datasets, where we vary the factorisation types on each of the three datasets. The best performances are highlighted in bold.

Factorisation type			Performance		
GE	GM	PM	GE	GM	PM
R	R	R	0.876	0.744	0.864
D	R	R	0.877	0.717	0.949
R	D	R	1.128	0.698	0.960
R	R	D	**0.860**	0.692	0.834
D	D	D	0.869	**0.663**	**0.799**

types, and the right three give the predictive performances on each of the datasets (each column corresponds to an experiment where we predict missing rows in that dataset). As can be seen, using multiple matrix factorisation ($\boldsymbol{D}$ for all matrices) gives the best performance most of the time, which is unsurprising since the three datasets are so weakly correlated. The best performance for GE (R, R, D) is most likely due to random variations of performance in the cross-validation procedure (the splitting of data into train and test sets is done randomly each time).

Drug sensitivity data

Similarly, we do this experiment for the four drug sensitivity datasets, for in-matrix predictions. These four datasets have much higher correlation (as they are repeated experiments – the same experiment conducted by different biological labs), as shown in Table 15, giving the Spearman correlation between the overlapping observed entries in each pair of the datasets. You can also find the overlaps between the datasets in Table 16, from the main paper. Here we see that:

- CCLE IC_{50} and CCLE EC_{50} are highly correlated, and have a big overlap.
- GDSC IC_{50} and CCLE IC_{50} are highly correlated, but few GDSC entries are also in CCLE. In contrast, many CCLE entries are in GDSC.
- GDSC IC_{50} and CCLE EC_{50} are (relatively) weakly correlated, but have a very small overlap. Overlap wise the same applies as for CCLE IC_{50}.
- Very few CTRP EC_{50} entries are in CCLE IC_{50} or EC_{50}, but many are in GDSC IC_{50}.

For the experiments we used $K_t = 10$ and $\alpha = 1.0$ for all datasets, K-means initialisation for the shared factor matrices ($\boldsymbol{F}_t$) and least squares for the private ones ($\boldsymbol{G}_l, \boldsymbol{S}_n$, and if we used matrix factorisation we shared the row factors (corresponding to drugs). The results are given in Table 17. There are a lot of results, so we will consider them one column at a time.

Table 15: Spearman correlation between values of each of the dataset pairs.

	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
GDSC IC_{50}	-	0.47	0.59	0.39
CTRP EC_{50}	0.47	-	0.44	0.45
CCLE IC_{50}	0.59	0.44	-	0.65
CCLE EC_{50}	0.39	0.45	0.65	-

Table 16: Overview of the four drug sensitivity dataset after preprocessing and filtering.

	Number	Number	Fraction	Overlap with other datasets			
Dataset	cell lines	drugs	observed	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
GDSC IC_{50}	399	48	73.57%	-	52.25%	9.34%	6.00%
CTRP EC_{50}	379	46	86.03%	57.39%	-	11.96%	7.37%
CCLE IC_{50}	253	16	96.42%	44.19%	51.51%	-	55.06%
CCLE EC_{50}	252	16	58.88%	28.52%	31.87%	55.28%	-

Table 17: Performances of in-matrix cross-validation results on the drug sensitivity datasets, where we vary the factorisation types on each of the four datasets. The best performances are highlighted in bold.

Factorisation type				Performance			
GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}	GDSC IC_{50}	CTRP EC_{50}	CCLE IC_{50}	CCLE EC_{50}
R	R	R	R	0.0767	**0.0889**	**0.0537**	0.1071
D	R	R	R	0.0765	0.0901	0.0546	0.1098
R	D	R	R	**0.0758**	0.0909	0.0543	0.1079
R	R	D	R	0.0765	0.0890	0.0584	**0.1055**
R	R	R	D	0.0768	0.0893	**0.0537**	0.1078
D	D	R	R	0.0763	0.0899	0.0569	0.1079
D	R	D	R	0.0766	0.0901	0.0553	0.1090
D	R	R	D	0.0769	0.0898	0.0554	0.1064
R	D	D	R	0.0765	0.0901	0.0566	0.1060
R	D	R	D	0.0772	0.0906	0.0543	0.1082
R	R	D	D	0.0771	0.0892	0.0542	0.1076
D	D	D	D	0.0776	0.0910	0.0562	0.1064

- **GDSC** IC_{50} (first column) – multiple matrix tri-factorisation (first row) achieves better performance than multiple matrix factorisation (last row), but a slight improvement can be achieved using a hybrid approach.
- **CTRP** EC_{50} (second column) – the best performance is achieved when using multiple matrix tri-factorisation on all datasets. Note that the CTRP dataset has a very similar correlation to all three other datasets.
- **CCLE** IC_{50} (third column) – the best performance is achieved when using multiple matrix tri-factorisation on all datasets, or when only CCLE EC_{50} is not decomposed as an R matrix, but using matrix factorisation (D) instead.
- **CCLE** EC_{50} (last column) – the best performances are achieved when the CCLE IC_{50} and EC_{50} are not both decomposed as an R matrix, but instead one or either is decomposed as a D matrix.

The last two items seem to imply that having a large overlap of entries and high correlation (such as CCLE IC_{50} and EC_{50}), does not necessarily mean we should use matrix tri-factorisation on both datasets. The best hybrid combination of factorisations is not so obvious, but by trying out multiple candidates a suitable hybridity can be found.

Recommendation For dissimilar datasets, with low correlation (like the methylation data), it is best to use multiple matrix factorisation. When the datasets are very similar, with high correlation (like the drug sensitivity data), matrix tri-factorisation can give better results. These two models will generally give very good performance already. One of the hybrid combinations of matrix factorisation and tri-factorisation can sometimes lead to even better results. Nested cross-validation can be used to find the best hybrid combination.

Bibliography

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–7, Mar. 2012.

C. Ding, X. He, and H. D. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *Proceedings of the fifth SIAM International Conference on Data Mining (SDM)*, pages 606–610, 2005.

C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 126, New York, New York, USA, Aug. 2006. ACM Press.

A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas. IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082, Sept. 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2642.

R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16(10):1–84, 1970.

D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael, L. L. Fulton, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Sept. 2012.

B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer discovery*, 5(11):1210–23, Nov. 2015.

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31 (3):279–311, Sept. 1966.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(Database issue):D955–61, Jan. 2013.

Y. Zhang and D.-Y. Yeung. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 606, New York, New York, USA, Aug. 2012. ACM Press.