Matrix Tri-factorisation

We want to decompose a matrix into three non-negative matrices, so that their product approximates the original matrix. This can give us some insights about the dataset, and allows us to predict missing values.

\[
\begin{pmatrix}
R \\
I \times J
\end{pmatrix} \approx
\begin{pmatrix}
F \\
I \times K \\
S \\
K \times L \\
G^T \\
L \times J
\end{pmatrix}
\]

We use a Bayesian probabilistic model, treating the three matrices as unobserved (latent) variables, and trying to infer their distributions after observing the data \( R \).

Bayesian Model

We assume Exponential priors over the matrices, and Gaussian noise of precision \( \tau \). Our model’s parameters are \( \theta = \{ F, S, G, \tau \} \). Exact inference is intractable, so we approximate the posterior using Gibbs sampling or variational inference.

\[
R_{ij} \sim N(R_{ij}|F_i \cdot S \cdot G_j, \tau^{-1}) \\
F_{ik} \sim E(F_{ik}|\lambda^F_{ik}) \\
S_{kl} \sim E(S_{kl}|\lambda^S_{kl}) \\
G_{jl} \sim E(G_{jl}|\lambda^G_{jl})
\]

Gibbs sampling can be used to approximate the posterior, \( p(\theta|D) \). We draw new values for each of the parameter \( \theta_i \) from the distribution \( p(\theta_i|\theta_{-i}, D) \), conditional on the data \( D \) and the current values of the other parameters \( \theta_{-i} \). This is guaranteed to converge to draws of the true posterior.

\[
p(\tau|F, S, G, R) = G(\tau|\alpha^*, \beta^*) \\
p(F_{ik}|F_{-ik}, S, G, \tau, R) = \mathcal{TN}(F_{ik}|\mu^F_{ik}, \tau^F_{ik}) \\
p(S_{kl}|F, S_{-kl}, G, \tau, R) = \mathcal{TN}(S_{kl}|\mu^S_{kl}, \tau^S_{kl}) \\
p(G_{jl}|F, S, G_{-jl}, \tau, R) = \mathcal{TN}(G_{jl}|\mu^G_{jl}, \tau^G_{jl})
\]

\( \mathcal{TN}(x|\mu, \tau) \) is a truncated normal – a Gaussian with zero density below \( x = 0 \). Note that our model does not have conjugacy, but we can still sample from the posteriors.

Variational Bayesian Inference

Gibbs sampling can be slow – variational inference is often much faster. Instead, we introduce a distribution \( q(\theta) \) that we will tune to be as similar to the posterior \( p(\theta|D) \) as possible, by minimising the KL-divergence. To make inference tractable we assume all variables are independent in \( q \). We use the same form of posteriors as we did with Gibbs sampling.

\[
q(\tau) = G(\tau|\alpha^*, \beta^*) \\
q(F_{ik}) = \mathcal{TN}(F_{ik}|\mu^F_{ik}, \tau^F_{ik}) \\
q(S_{kl}) = \mathcal{TN}(S_{kl}|\mu^S_{kl}, \tau^S_{kl}) \\
q(G_{jl}) = \mathcal{TN}(G_{jl}|\mu^G_{jl}, \tau^G_{jl})
\]

Results

We test our model by predicting missing values. We generate data from our model’s assumptions with with \( I = 50, J = 50, K = 10, L = 5, \lambda^F_{ik} = \lambda^S_{kl} = \lambda^G_{jl} = 1 \). We mark a fraction of the values as unobserved and measure the model’s performance at predicting them, for varying levels of noise. We use Gibbs sampling for inference.

We can see that our matrix tri-factorisation model can predict missing values accurately even if we observe a very small fraction of the data, whereas matrix factorisation starts to perform poorly much earlier.

\[
\text{MSE}
\]

\[
\text{PD}
\]

\[
\text{Fraction missing}
\]