# Prior Specification for Bayesian Matrix Factorization via Prior Predictive Matching

HELSINGIN YLIOPISTO

Eliezer de Souza da Silva[1] Tomasz Kuśmierczyk[2] Marcelo Hartmann[2] Arto Klami[2]

(1) Norwegian University of Science and Technology
(2) University of Helsinki

NTNU

## Overview

- **Challenge:** Selecting hyperparameters for Bayesian models is hard and computationally expensive.
- **Solution:** Match moments of prior predictive distributions.
  - **Approach 1:** Closed-form prior predictive moments obtained from model analysis and the application of Laws of total variation, total covariance and total expectation
  - **Approach 2:** Black-box gradient-based optimization of statistics derived from the prior predictive moment.
- **Benefit:** Immediate (fast) and good hyperparameters.

## Method

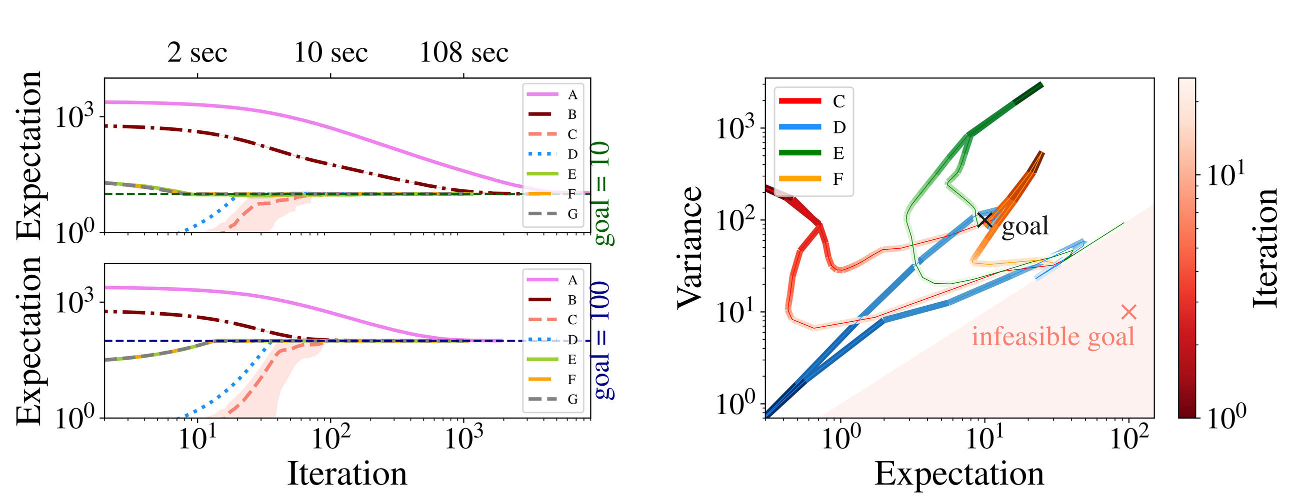
1. **Prior Predictive Distribution (PPD)**: Define the prior predictive distribution, which integrates out the model parameters:
$$p(Y;\lambda) = \int p(Y|Z;\lambda)p(Z;\lambda)dZ,$$
where $\lambda$ denotes the hyperparameters, $Y$ represents the data, and $Z$ represents the latent variables and model parameters.
2. **Virtual Statistics Calculation**: Calculate virtual statistics $\hat{T}_\lambda$ from the PPD (used to inform or adjust the hyperparameters without direct reliance on the observed data).
3. **Target Statistics**: Determine target statistics $T^*$, which could be provided by domain experts or estimated from a subset of the actual data. These statistics represent expected values that the model should reproduce.
4. **Solution**: Find $\lambda$ so that the virtual statistics derived from the PPD match the targets:
$$T^* = \hat{T}_\lambda$$
5. **Validation**: Validate (with e.g. posterior predictive checks) whether the chosen hyperparameters are appropriate, ensuring the model predictions align well with actual data characteristics.

## Black-box gradient-based optimization

1. Forward-sample $Y$ from PPD to estimate $\hat{T}(\mathbb{E}[g(Y);\lambda])$
2. Obtain gradients $\nabla_\lambda \hat{T}(\mathbb{E}[g(Y);\lambda])$ using automatic differentiation
3. Optimize the hyperparameters $\lambda$ so that the virtual statistics derived from the PPD match the target statistics as closely as possible:
$$\text{argmin}_\lambda \quad d(T^*, \hat{T}_\lambda)$$

The stochastic algorithm can be used for richer model families, but takes time, may fail to converge and is more computationally expensive.

**PMF example:**




convergence to target expectations (left) and failure for infeasible $T^*$ (right)

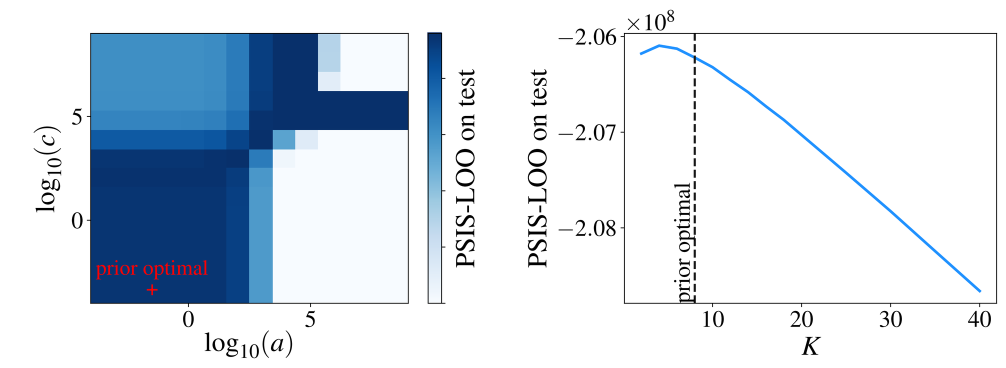
## Example: Poisson Matrix Factorization (PMF)

**Importance:** Bayesian Matrix Factorization (BMF) models are foundational in applications like recommendation systems.

**PMF model specfication:**

$$\theta_{ik} \sim F(\mu_\theta, \sigma_\theta^2), \quad \beta_{jk} \sim F(\mu_\beta, \sigma_\beta^2)$$
$$Y_{ij} \sim \text{Poisson}\left(\sum_{k=1}^{K} \theta_{ik}\beta_{jk}\right)$$

**Difficulty of selecting good priors:**




predictive quality on the `hetrec-lastfm` data

**Virtual Statistics:** derived from prior predictive distribution

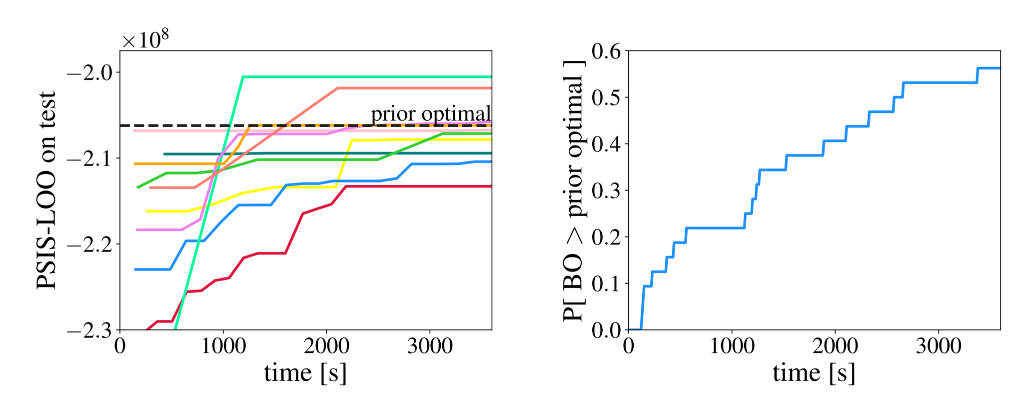
$$\mathbb{E}[Y_{ij}] = K\mu_\theta\mu_\beta, \quad \mathbb{V}[Y_{ij}] = K[\mu_\theta\mu_\beta + (\mu_\beta\sigma_\theta)^2 + (\mu_\theta\sigma_\beta)^2 + (\sigma_\theta\sigma_\beta)^2]$$
$$\rho_1[Y_{ij}, Y_{il}] = \frac{K(K\mu_\beta\sigma_\theta)^2}{\mathbb{V}[Y_{ij}]}, \quad \rho_2[Y_{ij}, Y_{tj}] = \frac{K(K\mu_\theta\sigma_\beta)^2}{\mathbb{V}[Y_{ij}]}$$

**Target Statistics:** $\mathbb{E}[Y_{ij}]$, $\mathbb{V}[Y_{ij}]$, $\rho_1$, $\rho_2$ provided by the user or estimated from data.

**Solution for number of latent factors:**

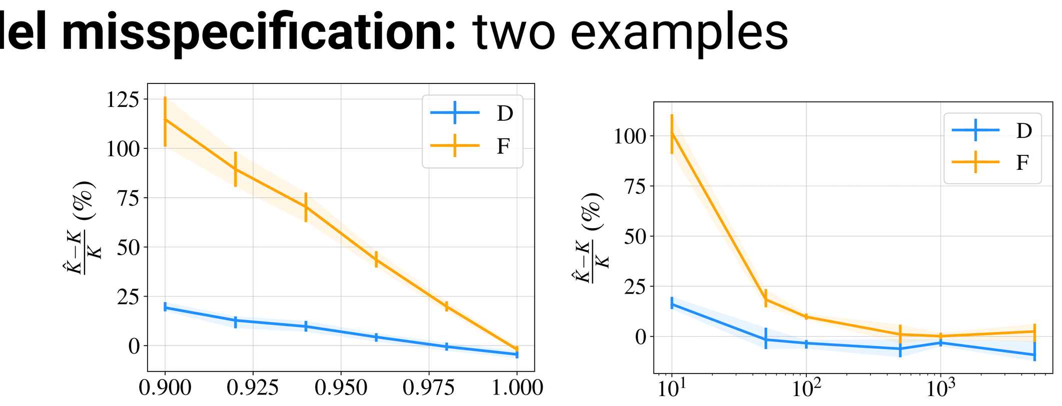
$$K = \frac{\tau\mathbb{V}[Y_{ij}] - \mathbb{E}[Y_{ij}]}{\rho_1\rho_2}\left(\frac{\mathbb{E}[Y_{ij}]}{\mathbb{V}[Y_{ij}]}\right)^2, \quad \tau = 1 - (\rho_1 + \rho_2)$$

**Performance compared to Bayesian Optimization (BO)**




BO runs (solid lines) vs the proposed method (dashed line)

**Sensitivity to model misspecification:** two examples




zero-inflated (left) and overdispersed (right) data

## Example: Compound Poisson Matrix Factorization (CPMF)

See the paper.

Reference: Eliezer de Souza da Silva, Tomasz Kuśmierczyk, Marcelo Hartmann, Arto Klami: *Prior Specification for Bayesian Matrix Factorization via Prior Predictive Matching.* Journal of Machine Learning Research, 24 (2023), 1-51.