Efficient MCMC Sampling for Bayesian Matrix Factorization by Breaking Posterior Symmetries

Saibal De
Department of Mathematics
University of Michigan
Ann Arbor, MI 48109
saibalde@umich.edu

Hadi Salehi
Department of Aerospace Engineering
University of Michigan
Ann Arbor, MI 48109
hsalehi@umich.edu

Alex Gorodetsky
Department of Aerospace Engineering
University of Michigan
Ann Arbor, MI 48109
goroda@umich.edu

Abstract

Bayesian low-rank matrix factorization techniques have become an essential tool for relational data analysis and matrix completion. A standard approach is to assign zero-mean Gaussian priors on the columns or rows of factor matrices to create a conjugate system. This choice of prior leads to symmetries in the posterior distribution and can severely reduce the efficiency of Markov-chain Monte-Carlo (MCMC) sampling approaches. In this paper, we propose a simple modification to the prior choice that *provably* breaks these symmetries and maintains/improves accuracy. Specifically, we provide conditions that the Gaussian prior mean and covariance must satisfy so the posterior does not exhibit invariances that yield sampling difficulties. For example, we show that using non-zero linearly independent prior means significantly lowers the autocorrelation of MCMC samples, and can also lead to lower reconstruction errors.

1 Introduction

In many natural and social science applications, data is often relational—it describes links between entities. We can efficiently represent relational data involving only one or two groups of entities as matrices. Examples include the similarity matrix of structure between two chemical compounds (Hattori et al., 2003), or user ratings of movies (Bennett et al., 2007).

Low-rank matrix factorization methods provide one of the simplest means of discovering structure from such relational data. The main idea behind this approach, posed in context of movie ratings, is as follows: a user's rating of a movie can be modeled by inner product of the user's and movie's hidden feature vectors. Thus, we can recover a $m \times n$ ratings matrix from knowing the $r \times m$ feature matrix of all users and the $r \times n$ feature matrix of all movies:

$$\mathbf{X} \in \mathbb{R}^{m \times n}, \quad \mathbf{X} = \mathbf{A}\mathbf{B}^\top, \quad \mathbf{A} \in \mathbb{R}^{m \times r}, \mathbf{B} \in \mathbb{R}^{n \times r} \tag{1}$$

Here r is referred to as the rank of this decomposition, and ideally $r \ll \min\{m, n\}$. This low-rank factorization can also be seen as a compression task—we can recover $\mathcal{O}(mn)$ entries of $\mathbf{X}$ by only storing $\mathcal{O}(r(m + n))$ entries of the factor matrices $\mathbf{A}$ and $\mathbf{B}$.

When all entries of the matrix $\mathbf{X}$ are known, classical low-rank factorization techniques such as the singular value decomposition (Strang, 1980) or the skeleton decomposition (Goreinov et al., 1997) can

Preprint. Under review.

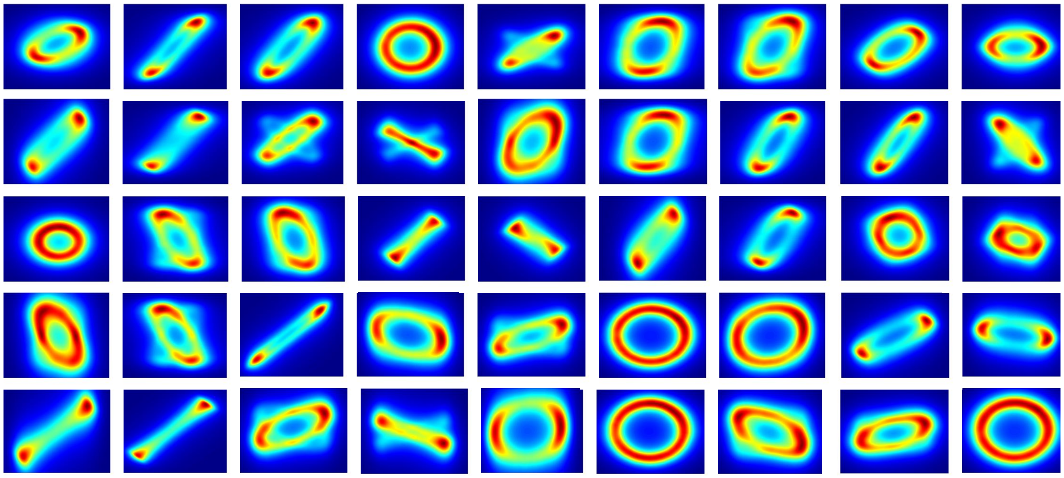


Figure 1: Joint posterior between some components of the factor matrices. Results obtained using Hamiltonian Monte Carlo with zero mean priors.

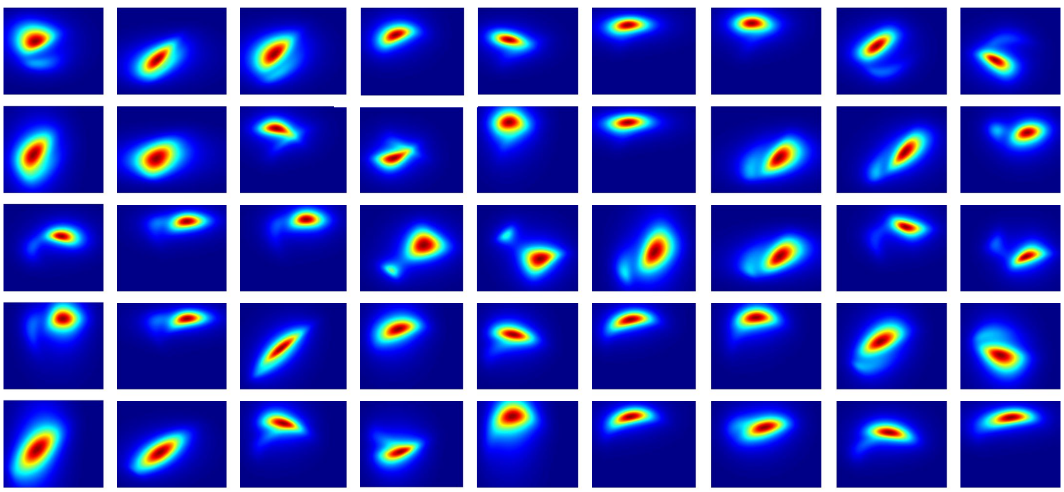


Figure 2: Joint posterior between some components of the factor matrices (same as in Figure 1). Results obtained using Hamiltonian Monte Carlo, but this time with non-zero mean priors.

be used to construct the low rank factorization (1). However in many practical applications, especially when the dimensions of the matrix are very large, we only have access to sparse observations. Optimization based approaches for reconstructing the unobserved entries (the *matrix completion* problem) have been well studied. In particular, Candès and Recht (2009) and Candès and Tao (2010) provide theoretical guarantees for recovering a matrix from few observations, and Candès and Plan (2010) extends these results in the context of noisy observations.

Bayesian inference techniques provide another avenue to solve this problem. Lim and Teh (2007) and Raiko et al. (2007) use variational approximation techniques to construct matrix factorization frameworks, and Nakajima and Sugiyama (2011), Nakajima et al. (2013) provide theoretical analysis on this subject. But as Salakhutdinov and Mnih (2008) points out, these variational methods are often based on an assumption of independence between the two factor matrices $\mathbf{A}$ and $\mathbf{B}$, which is not usually observed in practice. As a result, fully Bayesian models based on Markov chain Monte-Carlo (MCMC) sampling methods often outperform these variational models. Recently, Rai et al. (2014), Zhao et al. (2015a), Zhao et al. (2015b) have generalized this MCMC sampling approach to low-rank tensor factorizations in context of relational data with more than two groups of entities.

A key challenge for MCMC sampling algorithms in the context of matrix completion is the non-identifiability of the low-rank factorization (1)—given any $r \times r$ non-singular matrix $\mathbf{W}$ we can construct the new factors $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{W}$ and $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{W}^{-\top}$ such that $\mathbf{X} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^{\top}$. The effect of this *invertible invariance* shows up in the Bayes posteriors in the form of symmetries, especially when we impose zero-mean Gaussian priors on the columns or rows of $\mathbf{A}$ and $\mathbf{B}$ (seemingly one of the most popular choices made in the fully-Bayesian matrix completion literature). This phenomenon is illustrated in Figure 1, where we plot the joint posterior between various components of the factor matrices, constructed from Hamiltonian Monte-Carlo (HMC) samples, for the fully-observed 4×4 rank-2 matrix described in Example 1 (Section 4). Effective MCMC sampling from distributions with such wide varying multi-modal and non-connected geometries is, in general, a very difficult task.

In this paper, we propose a simple modification in the prior specification—we choose a linearly independent system of non-zero prior means. This leads to significant improvement in the sampling procedure. We theoretically prove that this choice breaks the symmetries arising from the invertible invariance, thus countering the issue of non-identifiability in matrix factorization affecting Bayesian inference. An illustration is given in Figure 2 where we plot the same set of joint posteriors as in Figure 1, but now constructed from HMC samples obtained with non-zero mean priors. Later in the paper, we demonstrate that this symmetry breaking ultimately leads to better performance of the MCMC sampling algorithms. We observe up to an order of magnitude decrease in the autocorrelations of generated samples, and corresponding improvement in reconstruction errors of the complete matrix, in several numerical examples.

The rest of the paper is structured as follows. In Section 2, we formally introduce the Bayesian inference setup and quantify the symmetries arising from zero-mean Gaussian priors. In Section 3, we show that choosing the priors means to be non-zero in a systematic fashion breaks the invertible invariance. In Section 4, we present the numerical experiments.

2 Notations and Bayesian inference setup

2.1 Notations

A vector $\boldsymbol{x}$ is always represented as a column, a row-vector is represented as $\boldsymbol{x}^\top$. The vector $\boldsymbol{e}_i$ is the i-th standard basis of appropriate (and inferable from context) size—all but its i-th entries are zero, and the non-zero entry is one.

Given a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, we use x_{ij} to denote its (i,j)-th entry. Additionally, $\bar{\boldsymbol{x}}_i \in \mathbb{R}^n$ and $\boldsymbol{x}_j \in \mathbb{R}^m$ denotes i-th row and the j-th column of the matrix; thus

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = [\boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_n] = \begin{bmatrix} \bar{\boldsymbol{x}}_1^\top \\ \vdots \\ \bar{\boldsymbol{x}}_m^\top \end{bmatrix}$$

The matrix $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

We adapt MATLAB's notation for indexing—given index sets $\Lambda_r \subseteq \{1,\dots,m\}$ and $\Lambda_c \subseteq \{1,\dots,n\}$ we use $\mathbf{X}[\Lambda_r,\Lambda_c]$ to denote the intersection of the rows of $\mathbf{X} \in \mathbb{R}^{m\times n}$ indexed by Λ_r and columns indexed by Λ_c. If one of the index sets is singleton, e.g. if $\Lambda_r = \{i\}$, then we denote $\mathbf{X}[i,\Lambda_c] = \mathbf{X}[\{i\},\Lambda_c]$. If one of the index sets is full, e.g. if $\Lambda_c = \{1,\dots,n\}$, then we denote $\mathbf{X}[\Lambda_r,:] = \mathbf{X}[\Lambda_r,\{1,\dots,n\}]$.

2.2 Prior and likelihood models

Given a low-rank factorization (1), we impose independent Gaussian priors on the columns of the factor matrices:

$$\boldsymbol{a}_k \sim \mathcal{N}(\boldsymbol{\mu}_{a,k}, \tau_{a,k}^{-1}\mathbf{I}_m) \quad \text{and} \quad \boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{\mu}_{b,k}, \tau_{b,k}^{-1}\mathbf{I}_n).$$

The joint prior is then given by

$$p(\mathbf{A},\mathbf{B}) = \prod_{k=1}^{r} \mathcal{N}(\boldsymbol{a}_k \mid \boldsymbol{\mu}_{a,k}, \tau_{a,k}^{-1}\mathbf{I}_m)\mathcal{N}(\boldsymbol{b}_k \mid \boldsymbol{\mu}_{b,k}, \tau_{b,k}^{-1}\mathbf{I}_n).$$

We use a fairly standard observation model, namely that the entries of the matrix $\mathbf{X}$ are corrupted by Gaussian noise:

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \quad \mathbf{N} = [\eta_{ij}]_{m\times n}, \quad \eta_{ij} \sim \mathcal{N}(0, \tau_\eta^{-1})$$

We only have access to a subset of the observed entries, $\boldsymbol{y} = (y_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \Lambda)$, characterized by an index set Λ. Each element of this index set, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \Lambda$ with $1 \leq \lambda_1 \leq m$, $1 \leq \lambda_2 \leq n$, indicates the location of the observed matrix element. The likelihood is then given by

$$p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B}) = \prod_{\boldsymbol{\lambda} \in \Lambda} \mathcal{N}(y_{\boldsymbol{\lambda}} \mid \bar{\boldsymbol{a}}_{\lambda_1}^{\top} \bar{\boldsymbol{b}}_{\lambda_2}, \tau_{\eta}^{-1}) \tag{2}$$

The effect of invertible invariance is immediately obvious:

Proposition 1 (Likelihood invertible invariance)**.** *The likelihood $p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B})$ defined in (2) is invariant under invertible transformations. In particular, if $\mathbf{W} \in \mathbb{R}^{r \times r}$ is an invertible matrix, then*

$$p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B}) = p(\boldsymbol{y} \mid \mathbf{A}\mathbf{W}, \mathbf{B}\mathbf{W}^{-\top}) \quad \textit{for all} \quad \mathbf{A} \in \mathbb{R}^{m \times r} \textit{ and } \mathbf{B} \in \mathbb{R}^{n \times r}$$

Proof. Let $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{W}$ and $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{W}^{-\top}$. Then we note

$$\tilde{\mathbf{A}}\tilde{\mathbf{B}}^{\top} = (\mathbf{A}\mathbf{W})(\mathbf{B}\mathbf{W}^{-\top})^{\top} = \mathbf{A}\mathbf{W}\mathbf{W}^{-1}\mathbf{B}^{\top} = \mathbf{A}\mathbf{B}^{\top}$$

Since all the matrix entries appearing in (2) are the same for both $(\mathbf{A}, \mathbf{B})$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$, it follows that the likelihood is invariant. □

2.3 The posterior and its symmetries

Using Bayes rule, the posterior can be computed as

$$p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) \propto p(\mathbf{A}, \mathbf{B}) p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B}) = p(\mathbf{A}) p(\mathbf{B}) p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B})$$

We can write the negative log posterior explicitly as:

$$\begin{aligned} -\ln p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = {} & \frac{mr}{2} \ln(2\pi) + \frac{m}{2} \sum_{k=1}^{r} \ln \frac{1}{\tau_{a,k}} + \frac{1}{2} \sum_{k=1}^{r} \tau_{a,k} \|\boldsymbol{a}_k - \boldsymbol{\mu}_{a,k}\|^2 \\ & + \frac{nr}{2} \ln(2\pi) + \frac{n}{2} \sum_{k=1}^{r} \ln \frac{1}{\tau_{b,k}} + \frac{1}{2} \sum_{k=1}^{r} \tau_{b,k} \|\boldsymbol{b}_k - \boldsymbol{\mu}_{b,k}\|^2 \\ & + \frac{|\Lambda|}{2} \ln(2\pi) + \frac{|\Lambda|}{2} \ln \frac{1}{\tau_{\eta}} + \frac{\tau_{\eta}}{2} \sum_{\boldsymbol{\lambda} \in \Lambda} (y_{\boldsymbol{\lambda}} - \bar{\boldsymbol{a}}_{\lambda_1}^{\top} \bar{\boldsymbol{b}}_{\lambda_2})^2 + \text{const.} \end{aligned} \tag{3}$$

Here the first two lines correspond to the priors on $\mathbf{A}$ and $\mathbf{B}$, the first three terms of the final line to the likelihood ($|\Lambda|$ represents the cardinality of the observation index set Λ). The constant term corresponds to the evidence $p(\boldsymbol{y})$ of the observations, and can generally be ignored. We have already seen that the likelihood term is invariant under invertible transformations of the form $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{A}\mathbf{W}, \mathbf{B}\mathbf{W}^{-\top})$. Now we investigate any effect the prior terms might have. We immediately note the following:

Proposition 2. *The posterior corresponding to* (3) *is invariant under invertible transformation $\mathbf{W} \in \mathbb{R}^{r \times r}$, i.e.*

$$p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = p(\mathbf{A}\mathbf{W}, \mathbf{B}\mathbf{W}^{-\top} \mid \boldsymbol{y}) \quad \textit{for all} \quad \mathbf{A} \in \mathbb{R}^{m \times r} \textit{ and } \mathbf{B} \in \mathbb{R}^{n \times r}$$

if and only if the terms

$$\begin{aligned} f_1(\mathbf{A}) &= \sum_{k=1}^{r} \tau_{a,k} \|\boldsymbol{a}_k\|^2, & f_2(\mathbf{A}) &= \sum_{k=1}^{r} \tau_{a,k} \boldsymbol{\mu}_{a,k}^{\top} \boldsymbol{a}_k, \\ f_3(\mathbf{B}) &= \sum_{k=1}^{r} \tau_{b,k} \|\boldsymbol{b}_k\|^2, & f_4(\mathbf{B}) &= \sum_{k=1}^{r} \tau_{b,k} \boldsymbol{\mu}_{b,k}^{\top} \boldsymbol{b}_k \end{aligned}$$

are individually invariant under the $\mathbf{A} \mapsto \mathbf{A}\mathbf{W}$ and $\mathbf{B} \mapsto \mathbf{B}\mathbf{W}^{-\top}$ transformations.

We can establish this result by essentially using the homogeneity of the f_1, f_2, f_3 and f_4 terms (a formal proof is presented in Appendix A). Clearly, the addition of prior imposes further restrictions on $\mathbf{W}$ for invertible invariance of the posterior (compared to the invariance of the likelihood). In particular, with zero-mean priors, the invariance exhibited by the likelihood under the transformation $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{A}\mathbf{W}, \mathbf{B}\mathbf{W}^{-\top})$ holds for the posterior only when $\mathbf{W}$ is restricted to a very particular subclass of invertible matrices:

Theorem 3 (Posterior invertible invariance with zero mean priors)**.** *Let $\boldsymbol{\mu}_{a,k} = \mathbf{0}$ and $\boldsymbol{\mu}_{b,k} = \mathbf{0}$ for all $1 \leq k \leq r$ and denote the diagonal matrices of the precision of the priors on the columns as*

$$\mathbf{T}_a = \operatorname{diag}(\tau_{a,1}, \ldots, \tau_{a,r}), \quad \mathbf{T}_b = \operatorname{diag}(\tau_{b,1}, \ldots, \tau_{b,r})$$

Let $\{\Lambda_1, \ldots, \Lambda_q\}$ be a partition of $\{1, \ldots, r\}$ defined by the following:

$$k, k' \in \Lambda_\ell \iff \tau_{a,k}\tau_{b,k} = \tau_{a,k'}\tau_{b,k'}$$

Then the posterior corresponding to (3) *is invariant under transformation $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{AW}, \mathbf{BW}^{-\top})$ with invertible $\mathbf{W} \in \mathbb{R}^{r \times r}$ if and only if we can decompose*

$$\mathbf{W} = \mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2} = \mathbf{T}_b^{-1/2}\mathbf{Q}\mathbf{T}_b^{1/2}$$

where $\mathbf{Q}$ is orthogonal and block diagonal w.r.t. the partition $\{\Lambda_1, \ldots, \Lambda_q\}$, i.e. the sub-matrices

$$\mathbf{Q}[\Lambda_{\ell_1}, \Lambda_{\ell_2}] \textit{ are } \begin{cases} \textit{orthogonal} & \textit{if } \ell_1 = \ell_2 \\ \textit{zero} & \textit{if } \ell_1 \neq \ell_2 \end{cases}$$

We note that one direction of this theorem (that the above form of $\mathbf{W}$ is sufficient for the invariance of posterior) appears in Nakajima and Sugiyama (2011). We claim this structure of the matrix $\mathbf{W}$ is also necessary for invertible invariance. Two extreme cases of invariance with zero mean priors can be derived immediately from this theorem:

Corollary 4. *Let $\boldsymbol{\mu}_{a,k} = \mathbf{0}$ and $\boldsymbol{\mu}_{b,k} = 0$, $1 \leq k \leq r$. Further suppose $\tau_{a,1} = \cdots = \tau_{a,r}$ and $\tau_{b,1} = \cdots = \tau_{b,r}$. Then the posterior corresponding to* (3) *is invariant under invertible transformation $\mathbf{W}$ if and only if $\mathbf{W}$ is orthogonal.*

Corollary 5. *Let $\boldsymbol{\mu}_{a,k} = \mathbf{0}$ and $\boldsymbol{\mu}_{b,k} = 0$, $1 \leq k \leq r$. Further suppose $\tau_{a,1}\tau_{b,1}, \ldots, \tau_{a,r}\tau_{b,r}$ are all distinct. Then the posterior corresponding to* (3) *is invariant under invertible transformation $\mathbf{W}$ if and only if $\mathbf{W}$ is diagonal with non-zero entries ± 1.*

Proofs of all three claims are presented in Appendix B.

3 Breaking the symmetries with non-zero mean priors

It is clear from Proposition 2 and Theorem 3 that the secret to completely breaking the symmetries can only hide in the f_2 and f_4 terms which involves the prior means. This leads to our main result:

Theorem 6 (Breaking posterior invertible invariance)**.** *Let $\mathbf{T}_a$, $\mathbf{T}_b$ and $\{\Lambda_1, \ldots, \Lambda_q\}$ be as defined in the statement of Theorem 3. Define the prior mean matrices*

$$\mathbf{M}_a = [\boldsymbol{\mu}_{a,1} \quad \cdots \quad \boldsymbol{\mu}_{a,r}] \quad \textit{and} \quad \mathbf{M}_b = [\boldsymbol{\mu}_{b,1} \quad \cdots \quad \boldsymbol{\mu}_{b,r}]$$

Then the posterior $p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y})$ is not invariant under the $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{AW}, \mathbf{BW}^{-\top})$ transformation for any non-identity invertible $r \times r$ matrix $\mathbf{W}$ if and only if the matrices

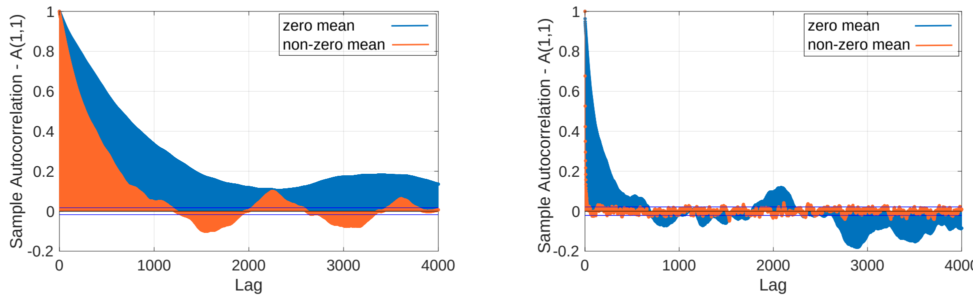
$$\begin{bmatrix} \mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\ \mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2} \end{bmatrix} \tag{4}$$

have full column rank for all $1 \leq \ell \leq q$.

We postpone a formal proof until Appendix C; for now, we present an immediate corollary: it provides an extremely simple way to ensure full rank matrices (4). It essentially states that if the means are chosen to be linearly independent (e.g. the entries of the mean matrix are sampled i.i.d. from a zero-mean Gaussian distribution—the columns would be independent with probability 1), then the invariance is broken.

Corollary 7. *Suppose either $\{\boldsymbol{\mu}_{a,k} : 1 \leq k \leq r\}$ or $\{\boldsymbol{\mu}_{b,k} : 1 \leq k \leq r\}$ form a linearly independent set in $\mathbb{R}^m$ or $\mathbb{R}^n$. Then the posterior $p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y})$ is not invariant under any non-identity invertible transformations.*

Thus, that by carefully choosing non-zero means for the priors on matrix factors, we can ensure that for any $r \times r$ invertible matrix $\mathbf{W} \neq \mathbf{I}$, the posteriors $p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y})$ and $p(\mathbf{AW}^\top, \mathbf{BW}^{-\top} \mid \boldsymbol{y})$ are distinct. In other words, with this choice of prior distributions, we can keep the identifiability issue from affecting Bayesian inference.



(a) Gibbs (b) HMC

Figure 3: Autocorrelation for factor $a_{1,1}$ in Example 1 with zero and non-zero mean priors, computed using 10th chain of Gibbs and HMC samplers.

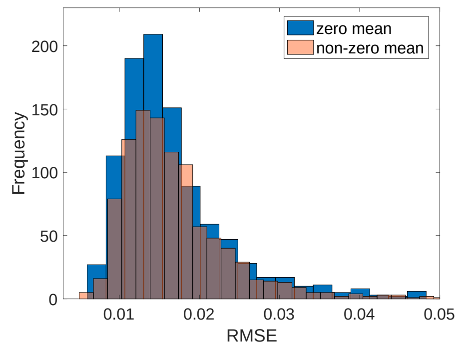


Figure 4: Reconstruction RMSE in Example 1 with zero and non-zero mean priors.

4 Numerical results

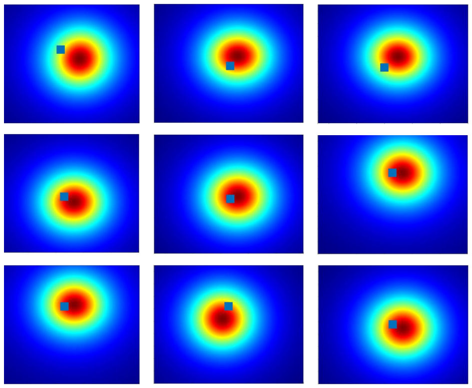
We now demonstrate that symmetry breaking improves MCMC sampling, both in terms of efficiency (by decreasing autocorrelation) and accuracy (by reducing reconstruction error) via four numerical experiments. Examples 1 and 2 apply Bayesian matrix factorization with synthetic data, and Examples 3 and 4 work with real-world data. For each example, the entries of the non-zero prior mean matrices are sampled from a uniform distribution. We use the root mean squared error (RMSE) as a measure of error between the true matrix and its reconstruction.

Example 1 (Fully observed synthetic matrix)**.** We contrast the results obtained from running the HMC and Gibbs samplers on a rank-2 4×4 matrix

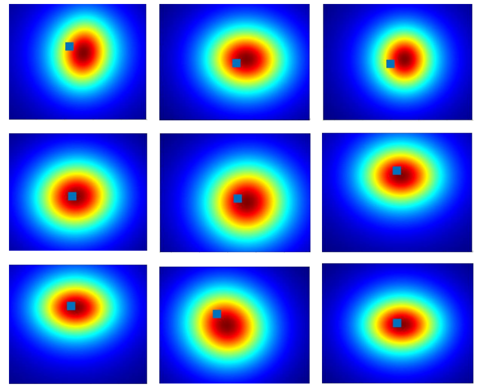
$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 5 \\ 2 & -1 & 1 & 4 \\ 4 & -1 & 3 & 14 \\ 3 & -1 & 2 & 9 \end{bmatrix}$$

with zero and non-zero mean priors. We observe the full matrix with noise precision $\tau_\eta = 10^4$. MCMC results obtained using both Gibbs and HMC samples that use 10 chains, each with 20000 samples. We also assume that the precision is unknown, and follow the standard procedure of hierarchical Bayes by inferring the precision with prior $\tau_\eta \sim \text{Ga}(3, 10^{-2})$. This leads to a conditionally conjugate posterior on τ_η (Alquier et al. (2014)).[1] In Figure 1 and Figure 2, we plot various joint posteriors, obtained using the samples generated using HMC, corresponding to the zero and non-zero mean priors. We can clearly see that the symmetries of the posterior in Figure 1 (corresponding to the zero mean priors) are not observed when using non-zero mean priors in Figure 2. This symmetry-breaking leads to better performance for MCMC samplers. For example, in Figure 3, we plot the autocorrelations for factor $a_{1,1}$ of the samples generated from both HMC and Gibbs

[1]Even though our theory was constructed assuming τ_η is constant, it generalizes in a very straightforward manner for arbitrary prior on τ_η.



(a) Zero mean priors



(b) Non-zero mean priors

Figure 5: Posterior predictions for some elements in Example 2 with Gibbs sampler. Blue marks indicate truth.

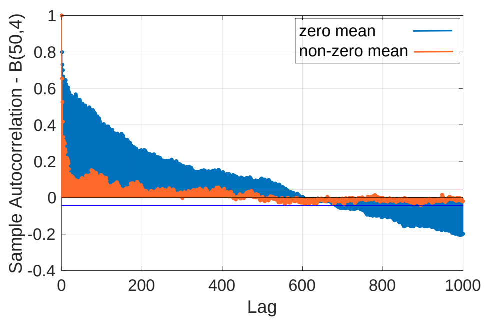


Figure 6: Autocorrelation for factor $b_{50,4}$ in Example 2 corresponding to zero and non-zero mean priors, computed with 5th chain of Gibbs sampler.

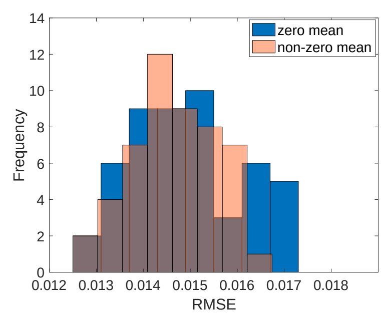


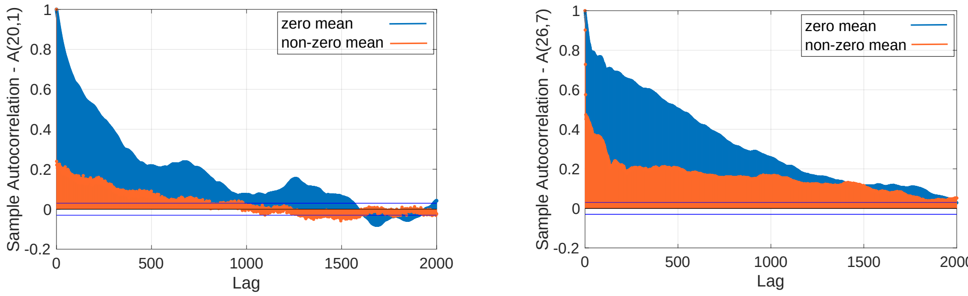
Figure 7: Reconstruction RMSE in Example 2 with zero and non-zero mean priors.

samplers. The autocorrelations are significantly lower, for both samplers, when non-zero mean priors are used. This indicates using non-zero mean priors leads to better/faster mixing and smaller autocorrelations—regardless of algorithm used because the geometry of the posteriors are more favorable to the types of structures exploited by virtually all MCMC techniques. Further, in Figure 4, we plot the histogram of RMSE values of the matrix reconstruction for 1000 repetitions of the data gathering procedure—each experiment differs due to the noise realization and randomly sampled prior mean. For each of the repetitions we run 10 chains of the Gibbs sampler. Here the RMSE is on the order of 10^{-2} for both samplers—this is exactly what was to be expected because it is the order of the noise standard deviation. We can clearly see that, for this example, there is little difference in the RMSE based on the prior mean.

Example 2 (Partially observed synthetic matrix)**.** As we pointed out earlier, we often do not have access to observations of all entries of a matrix. To demonstrate the utility of our theory in context of partial observations, we use the Gibbs sampler to reconstruct a rank-5 100×100 matrix from observing only 20% of its entries. Sampling is performed with 5 chains, each with 5000 samples. The prior on the precision is the same as the previous example. We show the posterior predictions for some elements in Figure 5 corresponding to zero and non-zero mean priors (results are presented for same elements in Figure 5(a) and Figure 5(b)). We can clearly see that using non-zero mean priors results in better MCMC performance. The corresponding order-of-magnitudes improvement in autocorrelation of the factor $b_{50,4}$ by choosing non-zero mean priors is showcased in Figure 6. The RMSE histogram constructed from 50 repetitions of the experiment is shown in Figure 7—here we see that using non-zero mean priors leads to better reconstruction errors.

Example 3 (Impaired driving dataset)**.** We now apply our theory to reconstruct the Impaired Driving Death Rate by Age and Gender data set available publicly from CDC.[2] We assume 60% of the data

[2]https://data.cdc.gov/Motor-Vehicle/Impaired-Driving-Death-Rate-by-Age-and-Gender-2012/ebbj-sh54



(a) Factor $a_{20,1}$

(b) Factor $a_{26,7}$

Figure 8: Autocorrelation for some factors in Example 3 corresponding to zero and non-zero mean priors, computed from 3rd chain of Gibbs sampler.

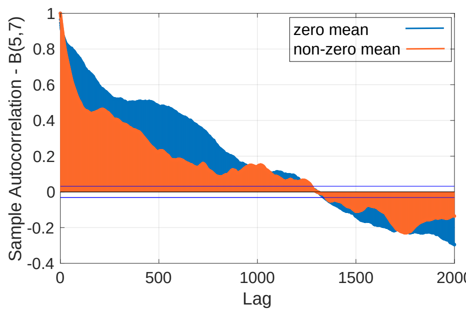


Figure 9: Autocorrelation for factor $b_{5,7}$ in Example 4 corresponding to zero/non-zero means, computed from 4th chain of Gibbs sampler.

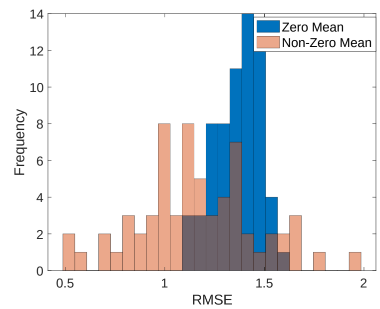


Figure 10: Reconstruction RMSE in Example 4 from 64 Gibbs sampling experiments with zero and non-zero mean priors.

is observed. Gibbs sampling is initiated with 5 chains, each with 10000 samples. Identical setup as in previous examples is used for this experiment. We demonstrate that introducing non-zero mean priors improves the autocorrelation for the $a_{20,1}$ and $a_{26,7}$ in Figure 8. Again we see significant improvement in efficiency in the same MCMC sampler.

Example 4 (Mice protein expression dataset)**.** Finally, we apply our theory to the mice-protein expression dataset (Higuera et al., 2015) available from the UCI Machine Learning Repository. [3] The dataset consists of 77 protein expressions, measured in terms of nuclear fractions, from 1080 mice specimens. For our experiments, we randomly sub-sampled this 1080×77 matrix, creating a 50×50 sub-matrix. We then constructed a rank 10 factorization of the form (1) using Gibbs sampling while observing only 50% of the entries. Sampling was initiated for 4 chains, each with 10000 samples. We set the parameter values $\tau_{a,1} = \cdots = \tau_{a,10} = \tau_{b,1} = \cdots = \tau_{b,10} = 25$ and $\tau_\eta = 10^2$, the first 1000 samples from each chain were discarded as burn-in. We observe that using non-zero mean priors leads to improvement of sample autocorrelation for the $b_{5,7}$ factor in Figure 9. We also see improved reconstruction errors in Figure 10—the RMSE histograms were computed from 64 independent repetitions of the experiment described above, with ten-fold sample thinning.

5 Conclusion

We have presented a full theoretical treatment of the symmetries of posteriors that arise from non-identifiability in Bayesian low-rank matrix factorization due to the standard choice of Gaussian priors on the matrix factors. We established that using a carefully chosen set of prior means, we can eliminate these symmetries, leading to better performance of MCMC sampling algorithms both with synthetic and real-world data. In future, we intend to extend this framework to address similar non-identifiability issues for low-rank tensor factorizations.

[3] https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

Acknowledgments and Disclosure of Funding

S. De acknowledges the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-19-2-0001 U.S. Army CCDC Ground Vehicle Systems Center (GVSC) Warren, MI.

A. Gorodetsky and H. Salehi acknowledge support from the Department of Energy Office of Scientific Research, ASCR under grant DE-SC0020364.

References

Pierre Alquier, Vincent Cottet, Nicolas Chopin, and Judith Rousseau. Bayesian matrix completion: prior specification. *arXiv preprint arXiv:1406.1440*, 2014.

James Bennett, Stan Lanning, et al. The Netflix prize. In *Proceedings of KDD Cup and Workshop*, 2007.

Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2010.

Sergei A Goreinov, Eugene E Tyrtyshnikov, and Nickolai L Zamarashkin. A theory of pseudoskeleton approximations. *Linear algebra and its applications*, 261, 1997.

Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125, 2003.

Clara Higuera, Katheleen J Gardiner, and Krzysztof J Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, 10, 2015.

Yew Jin Lim and Yee Whye Teh. Variational bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*, 2007.

Shinichi Nakajima and Masashi Sugiyama. Theoretical analysis of bayesian matrix factorization. *Journal of Machine Learning Research*, 12, 2011.

Shinichi Nakajima, Masashi Sugiyama, S. Derin Babacan, and Ryota Tomioka. Global analytic solution of fully-observed variational bayesian matrix factorization. *Journal of Machine Learning Research*, 14, 2013.

Piyush Rai, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson, and Lawrence Carin. Scalable bayesian low-rank decomposition of incomplete multiway tensors. In *International Conference on Machine Learning*, 2014.

Tapani Raiko, Alexander Ilin, and Juha Karhunen. Principal component analysis for large scale problems with lots of missing values. In *European Conference on Machine Learning*, 2007.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, 2008.

Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press, New York, 2nd edition, 1980.

Qibin Zhao, Liqing Zhang, and Andrzej Cichocki. Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE transactions on pattern analysis and machine intelligence*, 37, 2015a.

Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-Ichi Amari. Bayesian robust tensor factorization for incomplete multiway data. *IEEE transactions on neural networks and learning systems*, 27, 2015b.

A Proof of proposition 2

Proposition. *The posterior corresponding to*

$$-\log p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = \frac{1}{2}\sum_{k=1}^{r} \tau_{a,k}\|\boldsymbol{a}_k - \boldsymbol{\mu}_{a,k}\|^2 + \frac{1}{2}\sum_{k=1}^{r} \tau_{b,k}\|\boldsymbol{b}_k - \boldsymbol{\mu}_{b,k}\|^2 + \frac{\tau_\eta}{2}\sum_{\boldsymbol{\lambda}\in\Lambda}(y_{\boldsymbol{\lambda}} - \bar{\boldsymbol{a}}_{\lambda_1}^\top \bar{\boldsymbol{b}}_{\lambda_2})^2 + \textit{const.}$$

is invariant under invertible transformation $\mathbf{W} \in \mathbb{R}^{r\times r}$*, i.e.*

$$p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = p(\mathbf{AW}, \mathbf{BW}^{-\top} \mid \boldsymbol{y}) \quad \textit{for all} \quad \mathbf{A} \in \mathbb{R}^{m\times r} \textit{ and } \mathbf{B} \in \mathbb{R}^{n\times r}$$

if and only if the terms

$$f_1(\mathbf{A}) = \sum_{k=1}^{r} \tau_{a,k}\|\boldsymbol{a}_k\|^2, \qquad f_2(\mathbf{A}) = \sum_{k=1}^{r} \tau_{a,k}\boldsymbol{\mu}_{a,k}^\top \boldsymbol{a}_k,$$
$$f_3(\mathbf{B}) = \sum_{k=1}^{r} \tau_{b,k}\|\boldsymbol{b}_k\|^2, \qquad f_4(\mathbf{B}) = \sum_{k=1}^{r} \tau_{b,k}\boldsymbol{\mu}_{b,k}^\top \boldsymbol{b}_k$$

are individually invariant under the $\mathbf{A} \mapsto \mathbf{AW}$ *and* $\mathbf{B} \mapsto \mathbf{BW}^{-\top}$ *transformations.*

Proof. Note that

$$\|\boldsymbol{a}_k - \boldsymbol{\mu}_{a,k}\|^2 = \|\boldsymbol{a}_k\|^2 - 2\boldsymbol{\mu}_{a,k}^\top \boldsymbol{a}_k + \text{const.}$$

and similarly

$$\|\boldsymbol{b}_k - \boldsymbol{\mu}_{b,k}\|^2 = \|\boldsymbol{b}_k\|^2 - 2\boldsymbol{\mu}_{b,k}^\top \boldsymbol{b}_k + \text{const.}$$

Thus, we can rewrite the negative log posterior as

$$-\log p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = \frac{1}{2}f_1(\mathbf{A}) - f_2(\mathbf{A}) + \frac{1}{2}f_3(\mathbf{B}) - f_4(\mathbf{B}) - \log p(\boldsymbol{y} \mid \mathbf{A}, \mathbf{B}) + \text{const.} \tag{5}$$

From Proposition 1, we see that the likelihood term is already invariant under the invertible transformation. It follows invariance of f_1, f_2, f_3 and f_4 is sufficient for invariance of the posterior.

We now establish that invariance of f_1, f_2, f_3 and f_4 is necessary for invariance of the posterior. To show this, first note that f_1, f_3 are homogeneous of degree two, and f_2, f_4 are homogeneous of degree one. More explicitly, for all $t, s \in \mathbb{R}$ we have

$$f_1(t\mathbf{A}) = t^2 f_1(\mathbf{A}) \qquad f_2(t\mathbf{A}) = t f_2(\mathbf{A})$$
$$f_3(s\mathbf{B}) = s^2 f_3(\mathbf{B}) \qquad f_4(s\mathbf{B}) = s f_4(\mathbf{B})$$

Now, fix $\mathbf{A}$ and $\mathbf{B}$, then for invariance we must have $p(t\mathbf{A}, s\mathbf{B} \mid \boldsymbol{y}) = p(t\mathbf{AW}, s\mathbf{BW}^{-\top} \mid \boldsymbol{y})$ for all $s, t \in \mathbb{R}$. Expanding this out using (5) and applying the homogeneity properties, we obtain

$$\frac{t^2}{2}f_1(\mathbf{A}) - tf_2(\mathbf{A}) + \frac{s^2}{2}f_3(\mathbf{B}) - sf_4(\mathbf{B}) = \frac{t^2}{2}f_1(\mathbf{AW}) - tf_2(\mathbf{AW}) + \frac{s^2}{2}f_3(\mathbf{BW}^{-\top}) - sf_4(\mathbf{BW}^{-\top})$$

where the likelihood term cancels out. Comparing the coefficients of like-powered terms on the both sides, we conclude that we must have

$$f_1(\mathbf{A}) = f_1(\mathbf{AW}) \qquad f_2(\mathbf{A}) = f_2(\mathbf{AW})$$
$$f_3(\mathbf{B}) = f_3(\mathbf{BW}^{-\top}) \qquad f_4(\mathbf{B}) = f_4(\mathbf{BW}^{-\top})$$

Since $\mathbf{A}$ and $\mathbf{B}$ are arbitrary, it follows that f_1, f_2, f_3 and f_4 must be individually invariant under the $\mathbf{A} \mapsto \mathbf{AW}$ and $\mathbf{B} \mapsto \mathbf{BW}^{-\top}$ transformations. □

B Proof of theorem 3 and corollaries

B.1 A key lemma

Lemma 8. *Let* $\mathbf{Q} \in \mathbb{R}^{r\times r}$ *be a matrix satisfying* $\|\mathbf{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|$ *for all* $\boldsymbol{x} \in \mathbb{R}^r$. *Then* $\mathbf{Q}$ *is orthogonal.*

Proof. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two arbitrary vectors. Then

$$\|\mathbf{Q}\boldsymbol{x}_1 + \mathbf{Q}\boldsymbol{x}_2\|^2 = \|\mathbf{Q}\boldsymbol{x}_1\|^2 + \|\mathbf{Q}\boldsymbol{x}_2\|^2 + 2\langle \mathbf{Q}\boldsymbol{x}_1, \mathbf{Q}\boldsymbol{x}_2\rangle$$

and

$$\|\boldsymbol{x}_1 + \boldsymbol{x}_2\|^2 = \|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2 + 2\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$$

Using the facts $\|\mathbf{Q}\boldsymbol{x}_1 + \mathbf{Q}\boldsymbol{x}_2\| = \|\mathbf{Q}(\boldsymbol{x}_1 + \boldsymbol{x}_2)\| = \|\boldsymbol{x}_1 + \boldsymbol{x}_2\|$, $\|\mathbf{Q}\boldsymbol{x}_1\| = \|\boldsymbol{x}_1\|$ and $\|\mathbf{Q}\boldsymbol{x}_2\| = \|\boldsymbol{x}_2\|$ we obtain

$$\langle \mathbf{Q}\boldsymbol{x}_1, \mathbf{Q}\boldsymbol{x}_2\rangle = \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \quad \text{for all} \quad \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^r$$

Setting $\boldsymbol{x}_1 = \boldsymbol{e}_i$ and $\boldsymbol{x}_2 = \boldsymbol{e}_j$, we obtain

$$\langle \boldsymbol{q}_i, \boldsymbol{q}_j\rangle = \langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle$$

i.e. $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r\}$ form an orthonormal system in $\mathbb{R}^r$, hence $\mathbf{Q}$ is orthogonal. □

B.2 Proof of theorem 3

For convenience, we restate Theorem 3 slightly differently than earlier:

Theorem. *Let us denote*

$$\mathbf{T}_a = \operatorname{diag}(\tau_{a,1}, \ldots, \tau_{a,r}), \quad \mathbf{T}_b = \operatorname{diag}(\tau_{b,1}, \ldots, \tau_{b,r})$$

Let $\{\Lambda_1, \ldots, \Lambda_q\}$ *be a partition of* $\{1, \ldots, r\}$ *defined by the equivalence relation*

$$k, k' \in \Lambda_\ell \iff \tau_{a,k}\tau_{b,k} = \tau_{a,k'}\tau_{b,k'}$$

Then the posterior given by

$$-\log p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = \frac{1}{2}\sum_{k=1}^{r} \tau_{a,k}\|\boldsymbol{a}_k\|^2 + \frac{1}{2}\sum_{k=1}^{r} \tau_{b,k}\|\boldsymbol{b}_k\|^2 + \frac{\tau_\eta}{2}\sum_{\boldsymbol{\lambda}\in\Lambda} (y_{\boldsymbol{\lambda}} - \bar{\boldsymbol{a}}_{\lambda_1}^\top \bar{\boldsymbol{b}}_{\lambda_2})^2 + const. \tag{6}$$

is invariant under transformation $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{A}\mathbf{W}, \mathbf{B}\mathbf{W}^{-\top})$ *with invertible* $\mathbf{W} \in \mathbb{R}^{r\times r}$ *if and only if we can decompose*

$$\mathbf{W} = \mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2} = \mathbf{T}_b^{-1/2}\mathbf{Q}\mathbf{T}_b^{1/2} \tag{7}$$

where $\mathbf{Q}$ *is orthogonal and block diagonal w.r.t. the partition* $\{\Lambda_1, \ldots, \Lambda_q\}$, *i.e. the sub-matrices*

$$\mathbf{Q}[\Lambda_{\ell_1}, \Lambda_{\ell_2}] \textit{ are } \begin{cases} \textit{orthogonal} & \textit{if } \ell_1 = \ell_2 \\ \textit{zero} & \textit{if } \ell_1 \neq \ell_2 \end{cases} \tag{8}$$

Proof. From Proposition 2, invertible invariance (6) holds if and only if the terms

$$f_1(\mathbf{A}) = \sum_{k=1}^{r} \tau_{a,k}\|\boldsymbol{a}_k\|^2, \quad f_3(\mathbf{B}) = \sum_{k=1}^{r} \tau_{b,k}\|\boldsymbol{b}_k\|^2$$

are invariant under the $\mathbf{A} \mapsto \mathbf{A}\mathbf{W}$ and $\mathbf{B} \mapsto \mathbf{B}\mathbf{W}^{-\top}$ transformations.

We divide the rest of the proof into three steps:

- Step 1 derives the following condition on $\mathbf{W}$ necessary for invertible invariance:

$$\mathbf{Q}_a = \mathbf{T}_a^{-1/2}\mathbf{W}\mathbf{T}_a^{1/2} \quad \text{and} \quad \mathbf{Q}_b = \mathbf{T}_b^{1/2}\mathbf{W}\mathbf{T}_b^{-1/2} \tag{9}$$

  must be orthogonal matrices.

- Step 2 establishes $\mathbf{Q}_a = \mathbf{Q}_b$, and derives the block diagonal structure (8) on this common orthogonal matrix $\mathbf{Q}$. This establishes (7) as a necessary condition for invertible invariance.

- Step 3 proves that (7) is in fact sufficient for invertible invariance.

Step 1. Let us choose

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & \cdots & \alpha_r \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{m\times r} \implies \mathbf{AW} = \begin{bmatrix} \boldsymbol{\alpha}^\top \\ \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \end{bmatrix} [\boldsymbol{w}_1 \quad \cdots \quad \boldsymbol{w}_r] = \begin{bmatrix} \boldsymbol{\alpha}^\top \boldsymbol{w}_1 & \cdots & \boldsymbol{\alpha}^\top \boldsymbol{w}_r \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{m\times r}$$

for arbitrary $\boldsymbol{\alpha} \in \mathbb{R}^r$. Then we have

$$f(\mathbf{A}) = \sum_{k=1}^{r} \tau_{a,k} \alpha_k^2 = \|\mathbf{T}_a^{1/2} \boldsymbol{\alpha}\|^2$$
$$f(\mathbf{AW}) = \sum_{k=1}^{r} \tau_{a,k} (\boldsymbol{\alpha}^\top \boldsymbol{w}_k)^2 = \sum_{k=1}^{r} \tau_{a,k} (\boldsymbol{w}_k^\top \boldsymbol{\alpha})^2 = \|\mathbf{T}_a^{1/2} \mathbf{W}^\top \boldsymbol{\alpha}\|^2$$

These equalities follow from the following observations:

$$\mathbf{T}_a^{1/2} \boldsymbol{\alpha} = \begin{bmatrix} \tau_{a,1}^{1/2} \alpha_1 \\ \vdots \\ \tau_{a,r}^{1/2} \alpha_r \end{bmatrix}, \quad \mathbf{T}_a^{1/2} \mathbf{W}^\top \boldsymbol{\alpha} = \mathbf{T}_a^{1/2} \begin{bmatrix} \boldsymbol{w}_1^\top \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{w}_r^\top \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \tau_{a,1}^{1/2} \boldsymbol{w}_1^\top \boldsymbol{\alpha} \\ \vdots \\ \tau_{a,r}^{1/2} \boldsymbol{w}_r^\top \boldsymbol{\alpha} \end{bmatrix}$$

Thus, invariance of f_1 under $\mathbf{A} \mapsto \mathbf{AW}$ transformation requires

$$\|\mathbf{T}_a^{1/2} \boldsymbol{\alpha}\| = \|\mathbf{T}_a^{1/2} \mathbf{W}^\top \boldsymbol{\alpha}\| \quad \text{for all} \quad \boldsymbol{\alpha} \in \mathbb{R}^r$$

Set $\boldsymbol{\beta} = \mathbf{T}_a^{1/2} \boldsymbol{\alpha}$, then by the invertibility of $\mathbf{T}_a$ we can rewrite this condition as

$$\|\boldsymbol{\beta}\| = \|\underbrace{\mathbf{T}_a^{1/2} \mathbf{W}^\top \mathbf{T}_a^{-1/2}}_{\mathbf{Q}_a^\top} \boldsymbol{\beta}\| \quad \text{for all} \quad \boldsymbol{\beta} \in \mathbb{R}^r$$

Lemma 8 then implies that $\mathbf{Q}_a^\top$, and consequently $\mathbf{Q}_a$ (as defined in (9)), must be orthogonal. Proceeding in a similar manner, we can show that the invariance of f_3 under the $\mathbf{B} \mapsto \mathbf{BW}^{-\top}$ transformation would require $\mathbf{Q}_b$ (again, as defined in (9)) to be orthogonal.

Step 2. Using (9) we can compute

$$\mathbf{Q}_a^{-1} = \mathbf{T}_a^{-1/2} \mathbf{W}^{-1} \mathbf{T}_a^{1/2} \implies \mathbf{Q}_a^{-\top} = \mathbf{T}_a^{1/2} \mathbf{W}^{-\top} \mathbf{T}_a^{-1/2}$$

But since $\mathbf{Q}_a$ is orthogonal we have

$$\mathbf{Q}_a = \mathbf{Q}_a^{-\top} \implies \mathbf{T}_a^{-1/2} \mathbf{W} \mathbf{T}_a^{1/2} = \mathbf{T}_a^{1/2} \mathbf{W}^{-\top} \mathbf{T}_a^{-1/2} \implies \mathbf{W} \mathbf{T}_a = \mathbf{T}_a \mathbf{W}^{-\top} \tag{10}$$

Similarly, from (9) and orthogonality of $\mathbf{Q}_b$, we can derive

$$\mathbf{T}_b \mathbf{W} = \mathbf{W}^{-\top} \mathbf{T}_b \tag{11}$$

Using (10) and (11) along with associativity of matrix multiplication, we obtain

$$\mathbf{T}_a \mathbf{T}_b \mathbf{W} = \mathbf{T}_a (\mathbf{T}_b \mathbf{W}) \overset{(11)}{=} \mathbf{T}_a (\mathbf{W}^{-\top} \mathbf{T}_b) = (\mathbf{T}_a \mathbf{W}^{-\top}) \mathbf{T}_b \overset{(10)}{=} (\mathbf{W} \mathbf{T}_a) \mathbf{T}_\mathbf{b} = \mathbf{W} \mathbf{T}_a \mathbf{T}_b$$

Now, equating (i, j)-th entries of the two boundary matrices in the above chain (which are easy to compute given diagonal $\mathbf{T}_a$ and $\mathbf{T}_b$), we get

$$\tau_{a,i} \tau_{b,i} w_{ij} = w_{ij} \tau_{a,j} \tau_{b,j} \quad \text{for all} \quad 1 \le i, j \le r$$

Clearly, if $\tau_{a,i} \tau_{b,i} \neq \tau_{a,j} \tau_{b,j}$ for some pair of indices (i, j), then we must have $w_{ij} = 0$. This leads us to the block-diagonal structure of $\mathbf{W}$ w.r.t. partition $\{\Lambda_1, \ldots, \Lambda_q\}$, i.e. $\mathbf{W}[\Lambda_{\ell_1}, \Lambda_{\ell_2}]$ is non-zero only if $\ell_1 = \ell_2$. Using this with the diagonal nature of $\mathbf{T}_a$ and $\mathbf{T}_b$ in (9), we can conclude $\mathbf{Q}_a$ and $\mathbf{Q}_b$ have the same block-diagonal structure.

Next, for each $1 \leq \ell \leq q$ we have $\tau_{a,i}\tau_{b,i} = \tau_{a,j}\tau_{b,j}$ for all $i, j \in \Lambda_\ell$. Let us call this common value c_ℓ, then we have

$$\begin{aligned}
\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell] &= \operatorname{diag}(\tau_{a,i} : i \in \Lambda_\ell) \operatorname{diag}(\tau_{b,i} : i \in \Lambda_\ell) \\
&= \operatorname{diag}(\tau_{a,i}\tau_{b,i} : i \in \Lambda_\ell) \\
&= \operatorname{diag}(c_\ell : i \in \Lambda_\ell) \\
&= c_\ell \mathbf{I}_{r_\ell}
\end{aligned}$$

where we denote $r_\ell = |\Lambda_\ell|$. We conclude

$$\begin{aligned}
\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell] &= c_\ell \mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{-1} \\
\implies \mathbf{Q}_b[\Lambda_\ell, \Lambda_\ell] &= \mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2}\mathbf{W}[\Lambda_\ell, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{-1/2} \\
&= c_\ell^{1/2}\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{-1/2}\mathbf{W}[\Lambda_\ell, \Lambda_\ell]c_\ell^{-1/2}\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\
&= \mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{-1/2}\mathbf{W}[\Lambda_\ell, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\
&= \mathbf{Q}_a[\Lambda_\ell, \Lambda_\ell]
\end{aligned}$$

Combining these for all the blocks, we obtain $\mathbf{Q}_a = \mathbf{Q}_b$. We call this common value $\mathbf{Q}$, and (7) is trivially satisfied.

Step 3. This part of the proof is taken from Nakajima and Sugiyama (2011, Appendix G.3):

Note that we can write

$$f_1(\mathbf{A}) = \sum_{k=1}^{r} \tau_{a,k}\|\boldsymbol{a}_k\|^2 = \operatorname{trace}((\mathbf{T}_a\mathbf{A}^\top)\mathbf{A}) = \operatorname{trace}(\mathbf{A}(\mathbf{T}_a\mathbf{A}^\top)) = \operatorname{trace}(\mathbf{A}\mathbf{T}_a\mathbf{A}^\top)$$

where the second equality follows from the following observation:

$$\begin{aligned}
(\mathbf{T}_a\mathbf{A}^\top)\mathbf{A} &= \left(\begin{bmatrix} \tau_{a,1} & & \\ & \ddots & \\ & & \tau_{a,r} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_r^\top \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_r \end{bmatrix} \\
&= \begin{bmatrix} \tau_{a,1}\boldsymbol{a}_1^\top \\ \vdots \\ \tau_{a,r}\boldsymbol{a}_r^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_r \end{bmatrix} \\
&= \begin{bmatrix} \tau_{a,1}\boldsymbol{a}_1^\top\boldsymbol{a}_1 & \cdots & \tau_{a,1}\boldsymbol{a}_1^\top\boldsymbol{a}_r \\ \vdots & \ddots & \vdots \\ \tau_{a,1}\boldsymbol{a}_r^\top\boldsymbol{a}_1 & \cdots & \tau_{a,1}\boldsymbol{a}_r^\top\boldsymbol{a}_r \end{bmatrix}
\end{aligned}$$

Now, using $\mathbf{W} = \mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2}$ from (7), we get

$$\begin{aligned}
f_1(\mathbf{AW}) &= \operatorname{trace}((\mathbf{AW})\mathbf{T}_a(\mathbf{AW})^\top) \\
&= \operatorname{trace}(\mathbf{AW}\mathbf{T}_a\mathbf{W}^\top\mathbf{A}^\top) \\
&= \operatorname{trace}(\mathbf{A}(\mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2})\mathbf{T}_a(\mathbf{T}_a^{-1/2}\mathbf{Q}^\top\mathbf{T}_a^{1/2})\mathbf{A}^\top) \\
&= \operatorname{trace}(\mathbf{A}\mathbf{T}_a^{1/2}\mathbf{Q}(\mathbf{T}_a^{-1/2}\mathbf{T}_a\mathbf{T}_a^{-1/2})\mathbf{Q}^\top\mathbf{T}_a^{1/2}\mathbf{A}^\top) \\
&= \operatorname{trace}(\mathbf{A}\mathbf{T}_a^{1/2}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{T}_a^{1/2}\mathbf{A}^\top) \\
&= \operatorname{trace}(\mathbf{A}(\mathbf{T}_a^{1/2}\mathbf{T}_a^{1/2})\mathbf{A}^\top) \\
&= \operatorname{trace}(\mathbf{A}\mathbf{T}_a\mathbf{A}^\top) \\
&= f_1(\mathbf{A})
\end{aligned}$$

We can similarly prove that $f_3(\mathbf{B}\mathbf{W}^{-\top}) = f_3(\mathbf{B})$ with $\mathbf{W} = \mathbf{T}_b^{-1/2}\mathbf{Q}\mathbf{T}_b^{1/2}$. Thus f_1 and f_3 satisfy the desired invariance property. $\square$

B.3 Proof of corollary 4

Corollary. *Let $\boldsymbol{\mu}_{a,k} = \mathbf{0}$ and $\boldsymbol{\mu}_{b,k} = 0$, $1 \leq k \leq r$. Further suppose $\tau_{a,1} = \cdots = \tau_{a,r}$ and $\tau_{b,1} = \cdots = \tau_{b,r}$. Then the posterior corresponding to*

$$-\log p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = \frac{1}{2}\sum_{k=1}^{r} \tau_{a,k} \|\boldsymbol{a}_k\|^2 + \frac{1}{2}\sum_{k=1}^{r} \tau_{b,k} \|\boldsymbol{b}_k\|^2 + \frac{\tau_\eta}{2} \sum_{\boldsymbol{\lambda} \in \Lambda} (y_{\boldsymbol{\lambda}} - \bar{\boldsymbol{a}}_{\lambda_1}^\top \bar{\boldsymbol{b}}_{\lambda_2})^2 + \mathit{const.}$$

is invariant under invertible transformation $\mathbf{W}$ if and only if $\mathbf{W}$ is orthogonal.

Proof. We have $\tau_{a,1}\tau_{b,1} = \cdots = \tau_{a,r}\tau_{b,r}$. Hence $\mathbf{Q}$ only has one block. Consequently invertible invariance holds if and only if $\mathbf{Q}$ is orthogonal (by Theorem 3). Additionally, we can write $\mathbf{T}_a = \tau_{a,1}\mathbf{I}_r$ where $\mathbf{I}_r$ is the $r \times r$ identity matrix. It follows that

$$\mathbf{W} = \mathbf{T}_a^{1/2} \mathbf{Q} \mathbf{T}_a^{-1/2} = \tau_{a,1}^{1/2} \mathbf{I}_r \mathbf{Q} \tau_{a,1}^{-1/2} \mathbf{I}_r = \mathbf{Q}$$

i.e. $\mathbf{Q}$ is orthogonal if and only if $\mathbf{W}$ is also orthogonal. □

B.4 Proof of corollary 5

Corollary. *Let $\boldsymbol{\mu}_{a,k} = \mathbf{0}$ and $\boldsymbol{\mu}_{b,k} = 0$, $1 \leq k \leq r$. Further suppose $\tau_{a,1}\tau_{b,1}, \ldots, \tau_{a,r}\tau_{b,r}$ are all distinct. Then the posterior corresponding to*

$$-\log p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = \frac{1}{2}\sum_{k=1}^{r} \tau_{a,k} \|\boldsymbol{a}_k\|^2 + \frac{1}{2}\sum_{k=1}^{r} \tau_{b,k} \|\boldsymbol{b}_k\|^2 + \frac{\tau_\eta}{2} \sum_{\boldsymbol{\lambda} \in \Lambda} (y_{\boldsymbol{\lambda}} - \bar{\boldsymbol{a}}_{\lambda_1}^\top \bar{\boldsymbol{b}}_{\lambda_2})^2 + \mathit{const.}$$

is invariant under invertible transformation $\mathbf{W}$ if and only if $\mathbf{W}$ is diagonal with non-zero entries ± 1.

Proof. It follows from Theorem 3 that $\mathbf{Q}$ must be block diagonal with block size 1, and the each block has to be ± 1 (these are the only 1×1 orthogonal matrices). Let us write $\mathbf{Q} = \operatorname{diag}(q_1, \ldots, q_r)$ with each $q_i = \pm 1$. Then we have

$$\mathbf{W} = \operatorname{diag}(\tau_{a,1}^{1/2}, \ldots, \tau_{a,r}^{1/2}) \operatorname{diag}(q_1, \ldots, q_r) \operatorname{diag}(\tau_{a,1}^{-1/2}, \ldots, \tau_{a,r}^{-1/2}) = \operatorname{diag}(q_1, \ldots, q_r)$$

It follows that invertible invariance holds if and only if $\mathbf{W}$ is diagonal with entries ± 1. □

C Proof of theorem 6 and corollary

C.1 A key lemma

Lemma 9. *Let $\mathbf{P} \in \mathbb{R}^{m \times n}$ with $m \geq n$. Then the matrix equation*

$$\mathbf{P}\mathbf{W} = \mathbf{P}, \quad \mathbf{W} \in \mathbb{R}^{n \times n} \text{ orthogonal}$$

has the unique solution $\mathbf{W} = \mathbf{I}$ if and only if $\mathbf{P}$ has full column rank.

Proof. Multiplying both sides of the matrix equation by $\mathbf{P}^\top$ we obtain $\mathbf{P}^\top \mathbf{P} \mathbf{W} = \mathbf{P}^\top \mathbf{P}$. If $\mathbf{P}$ has full column rank, then $\mathbf{P}^\top \mathbf{P}$ is invertible, and it follows that $\mathbf{W} = \mathbf{I}$ is the unique solution.

Conversely, suppose $\mathbf{P}$ is not full rank. Then there exists a non-zero $\boldsymbol{x} \in \mathbb{R}^n$ with unit norm such that $\mathbf{P}\boldsymbol{x} = \mathbf{0}$. Let $\mathbf{W} = \mathbf{I} - 2\boldsymbol{x}\boldsymbol{x}^\top$, then clearly

$$\mathbf{P}\mathbf{W} = \mathbf{P}(\mathbf{I} - 2\boldsymbol{x}\boldsymbol{x}^\top) = \mathbf{P} - 2(\mathbf{P}\boldsymbol{x})\boldsymbol{x}^\top = \mathbf{P}$$

and

$$\mathbf{W}^\top \mathbf{W} = (\mathbf{I} - 2\boldsymbol{x}\boldsymbol{x}^\top)^\top (\mathbf{I} - 2\boldsymbol{x}\boldsymbol{x}^\top) = \mathbf{I} - 2\boldsymbol{x}\boldsymbol{x}^\top - 2\boldsymbol{x}\boldsymbol{x}^\top + 4\boldsymbol{x}(\boldsymbol{x}^\top \boldsymbol{x})\boldsymbol{x}^\top = \mathbf{I}$$

since $\boldsymbol{x}^\top \boldsymbol{x} = 1$. Thus we have constructed a second solution to the matrix equation. □

C.2 Proof of theorem 6

Theorem. *Let $\mathbf{T}_a$, $\mathbf{T}_b$ and $\{\Lambda_1, \dots, \Lambda_q\}$ be as defined in the statement of Theorem 3. Define the prior mean matrices*

$$\mathbf{M}_a = [\boldsymbol{\mu}_{a,1} \quad \cdots \quad \boldsymbol{\mu}_{a,r}] \quad \textit{and} \quad \mathbf{M}_b = [\boldsymbol{\mu}_{b,1} \quad \cdots \quad \boldsymbol{\mu}_{b,r}]$$

Then the posterior $p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y})$ is not invariant under the $(\mathbf{A}, \mathbf{B}) \mapsto (\mathbf{AW}, \mathbf{BW}^{-\top})$ transformation for any non-identity invertible $r \times r$ matrix $\mathbf{W}$ if and only if the matrices

$$\mathbf{P}_\ell = \begin{bmatrix} \mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\ \mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2} \end{bmatrix}$$

have full column rank for all $1 \le \ell \le q$.

Proof. In this proof, we attempt to reduce the set of all possible $r \times r$ invertible matrices $\mathbf{W}$, for which invertible invariance

$$p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y}) = p(\mathbf{AW}, \mathbf{BW}^{-\top} \mid \boldsymbol{y}) \quad \text{for all} \quad \mathbf{A} \in \mathbb{R}^{m \times r}, \mathbf{B} \in \mathbb{R}^{n \times r}$$

holds, to the singleton $\{\mathbf{I}_r\}$. We will demonstrate that this reduction is possible if and only if the $\mathbf{P}_\ell$ matrices (as defined in the theorem statement) have full column rank. We achieve this as follows:

- We have already shown in Proposition 2 that invertible invariance of the posterior holds if and only if the f_1, f_2, f_3 and f_4 terms (as defined in the aforementioned proposition) are individually invariant under the $\mathbf{A} \mapsto \mathbf{AW}$ and $\mathbf{B} \mapsto \mathbf{BW}^{-\top}$ transformations.
- Theorem 3 established that in order for the f_1 and f_3 terms to invariant under the transformation above, $\mathbf{W}$ must have the structure

$$\mathbf{W} = \mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2} = \mathbf{T}_b^{-1/2}\mathbf{Q}\mathbf{T}_b^{1/2}$$

  where $\mathbf{Q}$ is block-diagonal w.r.t. partition $\{\Lambda_1, \dots, \Lambda_q\}$ as defined in the statement of the aforementioned theorem, and the nonzero diagonal blocks $\mathbf{Q}[\Lambda_\ell, \Lambda_\ell]$ are orthogonal for all $1 \le \ell \le q$.
- In Step 1 below, we consider the terms f_2 and f_4, and derive simpler and equivalent conditions on matrix $\mathbf{W}$ (more specifically, the matrix $\mathbf{Q}$) to ensure invariance under the $\mathbf{A} \mapsto \mathbf{AW}$ and $\mathbf{B} \mapsto \mathbf{BW}^{-\top}$ transformations. These conditions are formulated in terms of the prior means $\boldsymbol{\mu}_{a,k}$, $\boldsymbol{\mu}_{b,k}$ and precisions $\tau_{a,k}$, $\tau_{b,k}$ for invariance.
- In Step 2, we further analyze these simpler conditions and frame them as matrix equations on diagonal blocks of $\mathbf{Q}$.
- Finally, in Step 3, we will use Lemma 9 to demonstrate $\mathbf{W} = \mathbf{I}$ is the only solution of this matrix system if and only if the matrices $\mathbf{P}_\ell$ are full rank for $1 \le \ell \le q$.

Step 1. Let us explicitly write out the invariance of f_2: we have $f_2(\mathbf{AW}) = f_2(\mathbf{A})$, i.e.

$$\begin{aligned} \sum_{k'=1}^{r} \tau_{a,k'}\boldsymbol{\mu}_{a,k'}^\top(\mathbf{AW})_{k'} &= \sum_{k'=1}^{r} \tau_{a,k'}\boldsymbol{\mu}_{a,k'}^\top \boldsymbol{a}_{k'} \\ \implies \sum_{k'=1}^{r} \tau_{a,k'}\boldsymbol{\mu}_{a,k'}^\top \mathbf{AW}\boldsymbol{e}_{k'} &= \sum_{k'=1}^{r} \tau_{a,k'}\boldsymbol{\mu}_{a,k'}^\top \boldsymbol{a}_{k'} \end{aligned} \tag{12}$$

for all $\mathbf{A}$. Note that we can write

$$\mathbf{A} = \sum_{i'=1}^{r} \boldsymbol{a}_{i'}\boldsymbol{e}_{i'}^\top$$

Substituting this expression on the left hand size of (12), we obtain

$$
\begin{aligned}
\sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \mathbf{A}\mathbf{W}\boldsymbol{e}_{k'} &= \sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \left(\sum_{i'=1}^{r} \boldsymbol{a}_{i'} \boldsymbol{e}_{i'}^\top \right) \mathbf{W}\boldsymbol{e}_{k'} \\
&= \sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \sum_{i'=1}^{r} \boldsymbol{a}_{i'} \boldsymbol{e}_{i'}^\top (\mathbf{T}_a^{1/2} \mathbf{Q} \mathbf{T}_a^{-1/2}) \boldsymbol{e}_{k'} \\
&= \sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \sum_{i'=1}^{r} \boldsymbol{a}_{i'} (\boldsymbol{e}_{i'}^\top \mathbf{T}_a^{1/2}) \mathbf{Q} (\mathbf{T}_a^{-1/2} \boldsymbol{e}_{k'}) \\
&= \sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \sum_{i'=1}^{r} \boldsymbol{a}_{i'} (\tau_{a,i'}^{1/2} \boldsymbol{e}_{i'}^\top) \mathbf{Q} (\tau_{a,k'}^{-1/2} \boldsymbol{e}_{k'}) \\
&= \sum_{k'=1}^{r} \tau_{a,k'}^{1/2} \boldsymbol{\mu}_{a,k'}^\top \sum_{i'=1}^{r} \tau_{a,i'}^{1/2} \boldsymbol{a}_{i'} (\boldsymbol{e}_{i'}^\top \mathbf{Q} \boldsymbol{e}_{k'}) \\
&= \sum_{k'=1}^{r} \tau_{a,k'}^{1/2} \boldsymbol{\mu}_{a,k'}^\top \sum_{i'=1}^{r} \tau_{a,i'}^{1/2} \boldsymbol{a}_{i'} q_{i',k'} \\
&= \sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'}^\top \sum_{i'=1}^{r} \tilde{\boldsymbol{a}}_{i'} q_{i',k'}
\end{aligned}
$$

where we denote

$$
\tilde{\boldsymbol{\mu}}_{a,k'} = \tau_{a,k'}^{1/2} \boldsymbol{\mu}_{a,k'}, \quad \tilde{\boldsymbol{a}}_{k'} = \tau_{a,k'}^{1/2} \boldsymbol{a}_{k'}, \quad k' \in \{1, \ldots, r\}
$$

The right hand side of (12) can be rewritten using this notation as

$$
\sum_{k'=1}^{r} \tau_{a,k'} \boldsymbol{\mu}_{a,k'}^\top \boldsymbol{a}_{k'} = \sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'}^\top \tilde{\boldsymbol{a}}_{k'}
$$

These two computations simplifies (12) to

$$
\sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'}^\top \sum_{i'=1}^{r} \tilde{\boldsymbol{a}}_{i'} q_{i',k'} = \sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'}^\top \tilde{\boldsymbol{a}}_{k'}
$$

Switching the order of summation on the left side, changing the summation index on the right side, and using $\tilde{\boldsymbol{a}}_{i'}^\top \tilde{\boldsymbol{\mu}}_{a,k'} = \tilde{\boldsymbol{\mu}}_{a,k'}^\top \tilde{\boldsymbol{a}}_{i'}$, we obtain

$$
\sum_{i'=1}^{r} \tilde{\boldsymbol{a}}_{i'}^\top \sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'} q_{i',k'} = \sum_{i'=1}^{r} \tilde{\boldsymbol{a}}_{i'}^\top \tilde{\boldsymbol{\mu}}_{a,i'} \tag{13}
$$

It has to hold for any arbitrary $\tilde{\boldsymbol{a}}_{i'} \in \mathbb{R}^m$ for $i' \in \{1, \ldots, r\}$ (since the columns $\boldsymbol{a}_{i'}$ of matrix $\mathbf{A}$ are arbitrary and $\tau_{a,i'}$ are positive reals). Let us fix $1 \le i \le r$ and assume all but the i-th of these vectors $\tilde{\boldsymbol{a}}_{i'}$ are zeros. Then (13) reduces to

$$
\tilde{\boldsymbol{a}}_{i}^\top \sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'} q_{i,k'} = \tilde{\boldsymbol{a}}_{i}^\top \tilde{\boldsymbol{\mu}}_{a,i}
$$

Since this holds for arbitrary $\tilde{\boldsymbol{a}}_i \in \mathbb{R}^m$, we conclude

$$
\sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{a,k'} q_{i,k'} = \tilde{\boldsymbol{\mu}}_{a,i} \tag{14}
$$

Conversely, if (14) holds for all $1 \le i \le r$, then (13) is trivially satisfied.

Let us pause for a moment and review our progress. Under the $\mathbf{A} \mapsto \mathbf{A}\mathbf{W}$ transformation, where $\mathbf{W}$ has the form defined in (7) and (8) (required for invariance of the f_1 and f_3 terms, c.f. Theorem 3), the term f_2 is invariant if and only if identity (12) holds if and only if identity (13) holds if and only if equation (14) is true.

Note that $\mathbf{W}^{-\top} = \mathbf{T}_b^{1/2}\mathbf{Q}^{-\top}\mathbf{T}_b^{-1/2} = \mathbf{T}_b^{1/2}\mathbf{Q}\mathbf{T}_b^{-1/2}$ where the last equality follows from orthogonality of $\mathbf{Q}$. We can now repeat the same process as above, and establish that f_4 is invariant under the $\mathbf{B} \mapsto \mathbf{B}\mathbf{W}^{-\top}$ transformation if and only if

$$\sum_{k'=1}^{r} \tilde{\boldsymbol{\mu}}_{b,k'} q_{i,k'} = \tilde{\boldsymbol{\mu}}_{b,i} \tag{15}$$

holds.

We combine these two arguments, and conclude f_2 and f_4 are invariant (after assuming the conditions (7) and (8) equivalent to invariances of f_1 and f_3) if and only if (14) and (15) holds.

Step 2. We now frame (14) and (15) as matrix equations for the diagonal blocks of the $\mathbf{Q}$ matrix. In (14), let us assume $i \in \Lambda_\ell$ for some $\ell \in \{1, \ldots, q\}$. Then, since $\mathbf{Q}$ is block-diagonal w.r.t. partitions $\{\Lambda_1, \ldots, \Lambda_\ell\}$, we have

$$q_{i,k'} = 0 \quad \text{for all} \quad k' \notin \Lambda_\ell$$

and (14) further reduces to

$$\sum_{k' \in \Lambda_\ell} \tilde{\boldsymbol{\mu}}_{a,k'} q_{i,k'} = \tilde{\boldsymbol{\mu}}_{a,i} \implies \sum_{k' \in \Lambda_\ell} \tau_{a,k'}^{1/2} \boldsymbol{\mu}_{a,k'} q_{i,k'} = \tau_{a,i}^{1/2} \boldsymbol{\mu}_{a,i}$$

This is a linear system with unknown $\mathbf{Q}[i, \Lambda_\ell]$; in matrix form, we can write it as

$$\mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2}\mathbf{Q}[i, \Lambda_\ell]^\top = \tau_{a,i}^{1/2}\boldsymbol{\mu}_{a,i} \tag{16}$$

We can similarly pose (15) as a matrix equation

$$\mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2}\mathbf{Q}[i, \Lambda_\ell]^\top = \tau_{b,i}^{1/2}\boldsymbol{\mu}_{b,i} \tag{17}$$

Combining (16) and (17) for all $i \in \Lambda_\ell$, we obtain the system

$$\begin{bmatrix} \mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\ \mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2} \end{bmatrix} \mathbf{Q}[\Lambda_\ell, \Lambda_\ell]^\top = \begin{bmatrix} \mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\ \mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2} \end{bmatrix}$$

Denoting the matrix on the right hand side as $\mathbf{P}_\ell$, we obtain

$$\mathbf{P}_\ell \mathbf{Q}[\Lambda_\ell, \Lambda_\ell]^\top = \mathbf{P}_\ell \tag{18}$$

In summary, given the block-diagonal structure of $\mathbf{Q}$ from invariance of f_1 and f_3 terms, we have derived matrix equation (18) which is equivalent to (14) and (15). These later two conditions are both necessary and sufficient for invariance of f_2 and f_4 to hold.

Step 3. By Lemma 9, the solution $\mathbf{Q}[\Lambda_\ell, \Lambda_\ell] = \mathbf{I}_{r_\ell}$ of (18) among orthogonal $r_\ell \times r_\ell$ matrices is unique if and only if the matrix $\mathbf{P}_\ell$ has full column rank. Collecting this result for all $\ell \in \{1, \ldots, q\}$, we conclude that $\mathbf{Q} = \mathbf{I}$ is the unique matrix generating the invertible invariance matrix $\mathbf{W}$ if and only if the matrices $\mathbf{P}_\ell$ are full rank.

Finally note that

$$\mathbf{Q} = \mathbf{I} \implies \mathbf{W} = \mathbf{T}_a^{1/2}\mathbf{Q}\mathbf{T}_a^{-1/2} = \mathbf{T}_a^{1/2}\mathbf{I}\mathbf{T}_a^{-1/2} = \mathbf{I}$$

and

$$\mathbf{W} = \mathbf{I} \implies \mathbf{Q} = \mathbf{T}_a^{-1/2}\mathbf{W}\mathbf{T}_a^{1/2} = \mathbf{T}_a^{-1/2}\mathbf{I}\mathbf{T}_a^{1/2} = \mathbf{I}$$

Thus $\mathbf{Q} = \mathbf{I}$ if and only if $\mathbf{W} = \mathbf{I}$ and we conclude our proof. □

C.3 Proof of corollary 7

Corollary. *Suppose either $\{\boldsymbol{\mu}_{a,k} : 1 \leq k \leq r\}$ or $\{\boldsymbol{\mu}_{b,k} : 1 \leq k \leq r\}$ form a linearly independent set in $\mathbb{R}^m$ or $\mathbb{R}^n$. Then the posterior $p(\mathbf{A}, \mathbf{B} \mid \boldsymbol{y})$ is not invariant under any non-identity invertible transformations.*

Proof. Suppose $\{\boldsymbol{\mu}_{a,k} : 1 \leq k \leq r\}$ is a linearly independent set in $\mathbb{R}^m$. Then $\mathbf{M}_a[:, \Lambda_\ell]$, and consequently $\mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2}$, are full rank for all ℓ. It follows that

$$\mathbf{P}_\ell = \begin{bmatrix} \mathbf{M}_a[:, \Lambda_\ell]\mathbf{T}_a[\Lambda_\ell, \Lambda_\ell]^{1/2} \\ \mathbf{M}_b[:, \Lambda_\ell]\mathbf{T}_b[\Lambda_\ell, \Lambda_\ell]^{1/2} \end{bmatrix}$$

has full rank for all ℓ, and Theorem 6 applies. A similar proof can be constructed when the prior means $\{\boldsymbol{\mu}_{b,k} : 1 \leq k \leq r\}$ form a linearly independent set in $\mathbb{R}^n$. □