entropy

MDPI

Article

From Bilinear Regression to Inductive Matrix Completion: A Quasi-Bayesian Analysis

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology,
7034 Trondheim, Norway; the.t.mai@ntnu.no

Abstract: In this paper, we study the problem of bilinear regression, a type of statistical modeling that deals with multiple variables and multiple responses. One of the main difficulties that arise in this problem is the presence of missing data in the response matrix, a problem known as inductive matrix completion. To address these issues, we propose a novel approach that combines elements of Bayesian statistics with a quasi-likelihood method. Our proposed method starts by addressing the problem of bilinear regression using a quasi-Bayesian approach. The quasi-likelihood method that we employ in this step allows us to handle the complex relationships between the variables in a more robust way. Next, we adapt our approach to the context of inductive matrix completion. We make use of a low-rankness assumption and leverage the powerful PAC-Bayes bound technique to provide statistical properties for our proposed estimators and for the quasi-posteriors. To compute the estimators, we propose a Langevin Monte Carlo method to obtain approximate solutions to the problem of inductive matrix completion in a computationally efficient manner. To demonstrate the effectiveness of our proposed methods, we conduct a series of numerical studies. These studies allow us to evaluate the performance of our estimators under different conditions and provide a clear illustration of the strengths and limitations of our approach.

Keywords: bilinear regression; matrix completion; low-rank model; PAC-Bayesian bound; Langevin Monte Carlo

1. Introduction

In this paper, we investigate the bilinear regression model, a statistical method that assumes a linear relationship between a set of multiple response variables and two sets of covariates. This model, also known as the growth curve model or generalized multivariate analysis model, is commonly used for analyzing longitudinal data, as shown in previous studies such as [1–6]. However, these studies only cover the scenario in which the response matrix is fully observed.

Recently, the bilinear regression model with incomplete response has been introduced and studied as the so-called inductive matrix completion, which is a generalization of the matrix completion problem [7,8]. This problem has attracted significant attention in various fields, such as drug repositioning [9], collaborative filtering [10], and genomics [7]. Inductive matrix completion is a challenging problem that arises when some of the entries in the response matrix are missing, which makes it difficult to infer the underlying relationship between the variables.

In this work, we explore the problem of bilinear regression and inductive matrix completion under a low-rank constraint on the coefficient matrix. Most existing approaches for these problems are frequentist methods, such as maximum likelihood estimation [2] or penalized optimization [7]. These methods are effective in providing point estimates for the parameters of the model but lack the ability to provide a full probabilistic characterization of the uncertainty. Recently, Bayesian approaches have been considered for these problems. For example, the paper [11] proposed a Bayesian approach for bilinear regression, and a

check for updates

Citation: Mai, T.T. From Bilinear Regression to Inductive Matrix Completion: A Quasi-Bayesian Analysis. *Entropy* **2023**, *25*, 333. https://doi.org/10.3390/e25020333

Academic Editors: Augustine Wong and Xiaoping Shi

Received: 27 January 2023
Revised: 8 February 2023
Accepted: 9 February 2023
Published: 11 February 2023

Bayesian method was proposed for inductive matrix completion in the work [9]. However, unlike frequentist approaches, the statistical properties of the Bayesian approach for these models have not been fully explored yet.

The aim of this paper is to address an existing gap in the understanding of the bilinear regression and inductive matrix completion problems. To achieve this goal, we propose a novel approach that combines elements of Bayesian statistics with a quasi-likelihood method. Specifically, we start by addressing the problem of bilinear regression using a quasi-Bayesian approach, where a quasi-likelihood is employed. We then generalize this approach to the problem of inductive matrix completion. To ensure that our method is adaptive to the rank of the coefficient matrix, we use a spectral scaled Student prior distribution, which allows us to prove that the posterior mean satisfies a tight oracle inequality. This result demonstrates that our method is able to accurately estimate the parameters of the model, even when the rank of the coefficient matrix is unknown. Additionally, we also prove the contraction properties of the posteriors, which further enhances the performance of our method.

The proposed method in this paper, the quasi-Bayesian approach, is an extension of the traditional Bayesian approach and is becoming increasingly popular in statistics and machine learning as a technique for generalized Bayesian inference, as noted in studies such as [12–14]. This approach allows for more flexibility in the modeling assumptions by replacing the likelihood function with a more general notion of risk or quasi-likelihood.

To provide theoretical guarantees for our proposed quasi-posteriors, we make use of the PAC-Bayesian technique [15–17]. This technique provides bounds on the generalization error of a learned estimator, and has been widely used in the literature as described in recent reviews and introductions, such as [18,19]. The PAC-Bayes bounds have been successfully applied in the context of matrix estimation problems as shown in studies such as [20–23]. Interestingly, by using the PAC-Bayesian technique for inductive matrix completion, we do not need to make any assumptions about the distribution of the missing entries in the response matrix. This is in contrast with previous works on matrix completion, such as [24–29], which typically require assumptions about the missing data. This makes our method more versatile and applicable to a wider range of problems.

The proposed method in this paper makes use of a spectral scaled Student prior, which is a specific choice of prior distribution. This choice is motivated by recent works in which it has been shown to lead to optimal rates in a variety of problems, including high-dimensional regression [30], image denoising [31] and reduced rank regression [23]. Although this prior is not conjugate to our problems, it allows for the convenient implementation of gradient-based sampling methods, which makes it computationally efficient.

To compute the proposed estimators and sample from the quasi-posterior, we employ a Langevin Monte Carlo (LMC) method. This method is a widely used algorithm for approximating complex distributions and allows for efficient computation of the proposed estimators. The LMC method allows us to obtain approximate solutions to the problem of bilinear regression and inductive matrix completion in a computationally efficient manner.

Furthermore, we use numerical studies to demonstrate the effectiveness of our proposed methods. These studies enable us to evaluate the performance of our estimators in various scenarios and provide insight into the capabilities and limitations of our approach. By conducting a thorough evaluation, we can gain a better understanding of how our method performs in different settings and make any necessary adjustments to improve its performance. We also compare our method with the ordinary least squared method. This comparison allows us to demonstrate the superiority of our method over the traditional approach in certain scenarios, such as when the response matrix contains missing data. Overall, the numerical studies serve as a valuable tool for assessing the effectiveness of our proposed method and provide a clear illustration of its strengths and weaknesses, as well as its improvement over the traditional method.

The remainder of this paper is organized as follows: In Section 2, we present the problem of bilinear regression, and introduce the low-rank promoting prior distribution that we use to address this problem. In Section 3, we extend our approach to the problem of inductive matrix completion. In Section 4, we discuss the Langevin Monte Carlo method used for the computation of the estimators, and present numerical studies to demonstrate the effectiveness of our proposed methods. Finally, we conclude our work and provide a summary of our findings in Section 5. The technical proofs are provided in Appendix A for the interested readers.

Notation 1. *Let $\mathbb{R}^{n_1 \times n_2}$ denote the set of $n_1 \times n_2$ matrices with real elements. Let $A^\intercal \in \mathbb{R}^{n_2 \times n_1}$ denote the transpose of A. For any $A \in \mathbb{R}^{n_1 \times n_2}$ and $I = (i,j) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$, we denote by $A_I = A_{(i,j)} = A_{i,j}$ the i-th row and j-th column elements of A. The matrix in $\mathbb{R}^{n_1 \times n_2}$ with all entries equal to 0 is denoted by $\mathbf{0}_{n_1 \times n_2}$. For a matrix $B \in \mathbb{R}^{n_1 \times n_1}$, we let $\mathrm{Tr}(B)$ denote its trace. The identity matrix in $\mathbb{R}^{n_1 \times n_1}$ is denoted by $\mathbf{I}_{n_1}$. For $A \in \mathbb{R}^{n_1 \times n_2}$, we define its sup-norm $\|A\|_\infty = \max_{i,j} |A_{i,j}|$; its Frobenius norm $\|A\|_F$ is defined by $\|A\|_F^2 = \mathrm{Tr}(A^\intercal A) = \sum_{i,j} A_{i,j}^2$ and $\mathrm{rank}(A)$ its rank.*

2. Bilinear Linear Regression

2.1. Model

Let $Y \in \mathbb{R}^{n \times q}$ consist of n independent response vectors, $X \in \mathbb{R}^{n \times p}$ be a given between-individuals design matrix and $Z \in \mathbb{R}^{k \times q}$ be a known within-individuals design matrix. Consider the bilinear regression model as follows:

$$Y = XM^*Z + E, \tag{1}$$

where $M^* \in \mathbb{R}^{p \times k}$ is the unknown parameter matrix. The random noise matrix E is assumed to have zero mean, $\mathbb{E}(E) = 0$. The main assumption here is the low-rank restriction on the model parameter that $\mathrm{rank}(M^*) < \min(p, k)$.

The model presented in Equation (1) is a bilinear regression model, which can be seen as a generalization of the reduced rank regression problem. In the case where $k = q$ and $Z = \mathbf{I}_q$, the model simplifies to the traditional reduced rank regression problem, which has been well-studied in the literature, such as [32,33]. However, in this paper, we consider the more general case where the matrix Z contains additional explanatory variables.

The low-rank assumption in this model can be interpreted as indicating the presence of a latent process that affects the response data, not only through the "between-individuals" structure of the model but also through the "within-individuals" structure. This model is often referred to as the growth curve model or the generalized multivariate analysis model (GMANOVA) and has been studied in depth in the literature, such as [2].

Assumption 1. *There is a known constant $C < +\infty$ such that $\|XM^*Z\|_\infty \leq C$.*

From Assumption 1, it is not reliable to return predictions XMZ with entries that are outside of interval $[-C, C]$. However, for computational reasons, it is extremely convenient to employ an unbounded prior for M. Therefore, we propose to use unbounded distributions for M but to use, as a predictor, a truncated version of XMZ rather than M itself. For a matrix A, let

$$\Pi_C(A) = \arg \min_{\|B\|_\infty \leq C} \|A - B\|_F$$

be the orthogonal projection of A on matrices with entries bounded by C. Note that B is simply obtained by replacing entries of A larger than C by C, and entries smaller than $-C$ by $-C$.

For a matrix $M \in \mathbb{R}^{p \times k}$, we denote by $r(M)$ the empirical risk of M as

$$r(M) = \frac{1}{nq} \|Y - \Pi_C(XMZ)\|_F^2$$

and its expectation is denoted by

$$R(M) = \mathbb{E}[r(M)] = \mathbb{E}\left[\left(Y_{11} - (\Pi_C(XMZ))_{11}\right)^2\right].$$

The focus of our work in this paper is on the predictive aspects of the model, that is, a matrix M predicts almost as well as M^* if $R(M) - R(M^*)$ is small. Under the assumption that E_{ij} has a finite variance, using the Pythagorean theorem, we have

$$R(M) - R(M^*) = \frac{1}{nq}\|\Pi_C(XMZ) - XM^*Z\|_F^2 \tag{2}$$

for any M, which means that our results can also be interpreted in terms of the Frobenius norm.

Let π be a prior distribution on $\mathbb{R}^{p\times k}$ (see Section 2.2). For any $\lambda > 0$, we define the quasi-posterior

$$\widehat{\rho}_\lambda(dM) \propto \exp(-\lambda r(M))\pi(dM).$$

It is worth noting that for a specific choice of $\lambda = nq/(2\sigma^2)$, the posterior distribution obtained corresponds to the case where the noise term E_{ij} is assumed to be Gaussian distributed with a mean of 0 and a variance of σ^2. However, our theoretical results hold under a more general class of noise distributions. It is known that a small enough λ is sufficient when the model is misspecified [14]. Additionally, even in the case of Gaussian noise, in high-dimensional settings, a smaller value of λ than $n/(2\sigma^2)$ leads to better adaptation properties [30,34]. The precise choice of λ in our method will be further discussed below.

We consider the following posterior mean of XMZ, given by

$$\widehat{XMZ}_\lambda = \int \Pi_C(XMZ)\widehat{\rho}_\lambda(dM). \tag{3}$$

It is worth noting that from the simulation experiments, it is observed that using reasonable values for C, the Monte Carlo algorithm never samples matrices M such that $\Pi_C(XMZ) \neq XMZ$. In other words, the boundedness constraint has very little impact on practice, and it is mainly necessary for technical proofs. If one is interested in obtaining an estimator of M^* instead of an estimator of XM^*Z, when $X^\intercal X$ and $ZZ^\intercal$ are invertible, one can consider the estimator $\widehat{M}_\lambda = (X^\intercal X)^{-1}X^\intercal\widehat{XMZ}_\lambda Z^\intercal(ZZ^\intercal)^{-1}$ and note that $X\widehat{M}_\lambda Z = \widehat{XMZ}_\lambda$. This estimator can be used to obtain the estimator of M^* in case the inverses of $X^\intercal X$ and $ZZ^\intercal$ are computationally feasible.

The quasi-posterior distribution investigated in this paper is often referred to as the "Gibbs posterior" in the PAC-Bayes approach. This terminology is used in the literature such as [17–19,35,36]. Additionally, the estimator $\widehat{M}_\lambda$ is sometimes referred to as the Gibbs estimator or the exponentially weighted aggregate (EWA) in the literature, such as [34,37,38].

2.2. Prior Specification

We consider, in this paper, the following spectral scaled Student prior distribution, with parameter $\tau > 0$:

$$\pi(M) \propto \det(\tau^2\mathbf{I}_m + MM^\intercal)^{-(p+m+2)/2}. \tag{4}$$

This prior distribution is designed to promote low-rankness by placing more probability mass on matrices with smaller singular values, which leads to sparse solutions. This prior has a similar form to the one used in related works such as [39,40], and it is known to lead to good performance in different problems, such as high-dimensional regression and image denoising. It can be verified that

$$\pi(M) \propto \prod_{j=1}^{m} (\tau^2 + s_j(M)^2)^{-(p+m+2)/2},$$

where $s_j(M)$ denotes the j^{th} largest singular value of M. The above expression is a scaled Student distribution evaluated at $s_j(M)$, which can be seen as a way to approximate sparsity on $s_j(M)$ [30]. The log-sum function $\sum_{j=1}^{m} \log(\tau^2 + s_j(M)^2)$ used by [39,40] is also known to enforce approximate sparsity on the singular values of $s_j(M)$. This means that under this prior, most of the $s_j(M)$ are close to 0, which implies that M is well approximated by a low-rank matrix. Therefore, it has the ability to promote the low-rankness of M.

As previously stated, this prior is not conjugate to the problem at hand. However, it is particularly convenient to implement gradient-based sampling algorithms, such as the Langevin Monte Carlo method, which will be discussed in more detail in Section 4. This is because the gradient of the log-posterior can be computed efficiently, and it allows for an efficient implementation of the LMC algorithm.

2.3. Theoretical Results

We assume the sub-exponential distribution assumption on the noise.

Assumption 2. *The entries $E_{i,j}$ of E are independent. There exist two known constants $\sigma > 0$ and $\xi > 0$ such that*

$$\forall k \geq 2,\ \mathbb{E}(|E_{i,j}|^k) \leq \sigma^2 k! \xi^{k-2}/2.$$

Let us put

$$C_1 = 8(\sigma^2 + C^2);\ C_2 = 64C \max(\xi, C);\ \tau^* = \sqrt{C_1(k+p)/(nkq\|X\|_F^2\|Z\|_F^2)}.$$

The statistical properties of mean estimator are given in the following theorem, where we propose a non-asymptotic analysis for our mean estimator.

Theorem 1. *Let Assumptions 1 and 2 be satisfied. Fix the parameter $\tau = \tau^*$ in the prior. Fix $\delta > 0$ and define $\lambda^* := nq \min(1/(2C_2), \delta/[C_1(1+\delta)])$. Then, for any $\varepsilon \in (0,1)$, we have, with probability at least $1-\varepsilon$ on the sample,*

$$\left\|\widehat{XMZ}_{\lambda^*} - XM^*Z\right\|_F^2 \leq \inf_{0 \leq r \leq pk} \inf_{\substack{\bar{M} \in \mathbb{R}^{p\times k} \\ \mathrm{rank}(\bar{M}) \leq r}} \Bigg\{ (1+\delta)\|X\bar{M}Z - XM^*Z\|_F^2 + \frac{C_1(1+\delta)^2}{\delta}\left[4r(k+p+2)\log\left(1 + \frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{nkq}{r(k+p)}}\right) + k + p + 2\log\frac{2}{\varepsilon}\right]\Bigg\}.$$

The choice of $\lambda = \lambda^*$ is determined by optimizing an upper bound on the risk R (as shown in the proof of this theorem). However, it is important to note that this choice may not necessarily be the best choice in practice, even though it gives a good estimate of the order of magnitude for λ. To ensure optimal performance, the user can use cross-validation to properly adjust the temperature parameter. Additionally, it is worth noting that $\mathrm{rank}(\bar{M}) \neq 0$ is not a requirement in the above formula. If $\mathrm{rank}(\bar{M}) = 0$, then $\bar{M} = 0$ and we interpret $0 \log(1 + 0/0)$ as 0.

The proof of this theorem is based on the PAC-Bayes theory, and it is provided in the Appendix A. By taking $\bar{M} = M^*$, we can obtain an upper bound on the infimum, leading to the following result.

Corollary 1. *Under the same assumptions and the same* τ, λ^* *as in Theorem 1, let* $r^* = \mathrm{rank}(M^*)$*. Then, for any* $\varepsilon \in (0,1)$*, we have, with probability at least* $1-\varepsilon$ *on the sample,*

$$\left\|\widehat{XMZ}_{\lambda^*} - XM^*Z\right\|_F^2 \leq$$
$$\frac{C_1(1+\delta)^2}{\delta}\left[4r^*(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|M^*\|_F}{\sqrt{C_1}}\sqrt{\frac{nkq}{r(k+p)}}\right)+k+p+2\log\frac{2}{\varepsilon}\right].$$

Theorem 1 provides an understanding of the statistical properties of the posterior mean. However, it is also important to understand the contraction properties of the quasi-posterior distribution. In the following theorem, we aim to provide a result that demonstrates this aspect of the proposed method.

Theorem 2. *Under the assumptions for Theorem 1, let* ε_n *be any sequence in* $(0,1)$ *such that* $\varepsilon_n \to 0$ *when* $n \to \infty$*. Define*

$$\mathcal{M}_n = \left\{ M \in \mathbb{R}^{p\times k} : \left\|\widehat{XMZ}_{\lambda^*} - XM^*Z\right\|_F^2 \leq \inf_{0\leq r\leq pk} \inf_{\substack{\bar{M}\in\mathbb{R}^{p\times k}\\ \mathrm{rank}(\bar{M})\leq r}} \left\{ (1+\delta)\|X\bar{M}Z - XM^*Z\|_F^2 + \right.\right.$$
$$\left.\left.\frac{C_1(1+\delta)^2}{\delta}\left[4r(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{nkq}{r(k+p)}}\right)+k+p+2\log\frac{2}{\varepsilon_n}\right]\right\}\right\}.$$

Then,

$$\mathbb{E}\Big[\mathbb{P}_{M\sim\hat{\rho}_\lambda}(M\in\mathcal{M}_n)\Big] \geq 1-\varepsilon_n \xrightarrow[n\to\infty]{} 1.$$

The proof of this theorem is provided in Appendix A.

3. Inductive Matrix Completion

3.1. Model and Method

In the context of inductive matrix completion, given two side information matrices X and Z, we assume that only a random subset Ω of the entries of Y in model (1) is observed. More precisely, we assume that we observe m independent and identically random pairs $(\mathcal{I}_1, Y_1), \ldots, (\mathcal{I}_m, Y_m)$ given by

$$Y_i = (XM^*Z)_{\mathcal{I}_i} + \mathcal{E}_i, \quad i = 1, \ldots, m \tag{5}$$

where $M^* \in \mathbb{R}^{p\times k}$ is the unknown parameter matrix expected to be low-rank and observation sample size is assumed that $m < nq$. The noise variables $\mathcal{E}_i$ are assumed to be independent with $\mathbb{E}(\mathcal{E}_i) = 0$. The variables $\mathcal{I}_i$ are independent and identical copies of a random variable $\mathcal{I}$ having distribution Π on the set $\{1,\ldots,n\}\times\{1,\ldots,q\}$, we denote $\Pi_{x,y} := \Pi(\mathcal{I} = (x,y))$.

Our goal in this paper is to investigate the problem of bilinear regression and also address the case where the response matrix contains missing data, a problem known as inductive matrix completion. In particular, when $p = n, k = q$ and $X = \mathbf{I}_n, Z = \mathbf{I}_q$ are the identity matrices, the problem reduces to the traditional matrix completion problem, which has been well studied in the literature [26]. Similarly, when $k = q$ and $Z = \mathbf{I}_q$ is the identity matrix, the problem becomes the reduced rank regression problem with incomplete response, which has also been studied in recent works, such as [23,41]. However, in the context of inductive matrix completion, we focus on the more general case where X and Z contain additional explanatory variables; in other words, we consider the side information from the n users and the q items in our model [8].

It has been acknowledged that there are two different ways to model the observed values of Y, either by including or excluding the possibility of observing the same entry multiple times. Previous studies have examined both of these methods, such as the examination of matrix completion without replacement in [25] and with replacement in [26]. Both methods have practical uses and use similar techniques for estimation. This particular study focuses on the scenario where the variables $\mathcal{I}_i$ are independently and identically distributed, meaning that it is possible to observe the same entry multiple times. Additionally, it is important to note that, according to the findings presented in Section 6 of [42], the results of this study can also be applied to the scenario of sampling without replacement, as long as the sampling is performed uniformly and there is no observation noise.

We are now adapting the quasi-Bayesian approach for bilinear regression in Section 2 to the context of inductive matrix completion. For a probability distribution P on $\{1,\ldots,n_1\} \times \{1,\ldots,n_2\}$, we generalize the Frobenius norm by $\|A\|_{F,P}^2 = \sum_{i,j} P[(i,j)]A_{i,j}^2$; note that when P is the uniform distribution, then $\|A\|_{F,P}^2 = \|A\|_F^2/(n_1 n_2)$.

For a matrix $M \in \mathbb{R}^{p\times k}$, we denote the empirical risk of M, $r'(M)$, and its expected risk $R'(M)$ respectively as

$$r'(M) = \frac{1}{m}\sum_{i=1}^{m}\big(Y_i - (\Pi_C(XMZ))_{\mathcal{I}_i}\big)^2,$$

$$R'(M) = \mathbb{E}[r'(M)] = \mathbb{E}[\big(Y_1 - (\Pi_C(XMZ))_{\mathcal{I}_1}\big)^2].$$

As in Section 2, we will focus on the predictive aspects of the model, that is, a matrix M predicts almost as well as M^* if $R'(M) - R'(M^*)$ is small. Under the assumption that $\mathcal{E}_i$ has a finite variance, based on the Pythagorean theorem, we have

$$R'(M) - R'(M^*) = \|\Pi_C(XMZ) - XM^*Z\|_{F,\Pi}^2 \tag{6}$$

for any M, which means that our results can also be interpreted in terms of an estimation of M^* with respect to a generalized Frobenius norm.

Here, the prior π is the low-rank inducing prior specified in the Section 2.2 above. For any $\lambda > 0$, we define the quasi-posterior

$$\hat{\rho}'_\lambda(dM) \propto \exp(-\lambda r'(M))\pi(dM).$$

We will actually specify our choice of λ below.

The truncated posterior mean of XMZ is given by

$$\widehat{XMZ}_\lambda = \int \Pi_C(XM)\hat{\rho}'_\lambda(dM). \tag{7}$$

Here, for the same technical reasons as in the context of bilinear regression, this truncation has a very little impact in practice for reasonable values of C.

3.2. Theoretical Results

In this section, we derive the statistical properties of the posterior $\hat{\rho}'_\lambda$ and the mean estimator $\widehat{XMZ}_\lambda$ for the context of inductive matrix completion. Let us first state our assumptions on this model.

Assumption 3. *The noise variables $\mathcal{E}_1,\ldots,\mathcal{E}_m$ are independent of $\mathcal{I}_1,\ldots,\mathcal{I}_m$. There exist two known constants $\sigma' > 0$ and $\xi' > 0$ such that*

$$\forall k \geq 2,\, \mathbb{E}(|\mathcal{E}_i|^k) \leq \sigma'^2 k!\xi'^{k-2}/2.$$

Assumptions 1 and 3 are both standard; they have been used in [41] for theoretical analysis of reduced rank regression and in [26] for trace regression and matrix completion.

Let us put

$$C_1' = 8(\sigma'^2 + C^2);\ C_2' = 64C\max(\xi', C);\ \tau^* = \sqrt{C_1'(k+p)/(mkp\|X\|_F^2\|Z\|_F^2)}.$$

Theorem 3. *Let Assumptions 1 and 3 be satisfied. Fix the parameter $\tau = \tau^*$ in the prior. Fix $\delta > 0$ and define $\lambda'^* := m\min(1/(2C_2'), \delta/[C_1'(1+\delta)])$. Then, for any $\varepsilon \in (0,1)$, we have, with probability at least $1-\varepsilon$ on the sample,*

$$\left\|\widehat{XMZ}_{\lambda'^*} - XM^*Z\right\|_{F,\Pi}^2 \leq \inf_{0\leq r\leq pk} \inf_{\substack{\bar{M}\in\mathbb{R}^{p\times k}\\ \operatorname{rank}(\bar{M})\leq r}} \left\{(1+\delta)\|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \frac{C_1'(1+\delta)^2}{\delta}\frac{\left(4r(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{mkp}{r(k+p)}}\right)+k+p+2\log\frac{2}{\varepsilon}\right)}{m}\right\}.$$

Similar to the context of bilinear regression, the choices of $\lambda = \lambda^*, \tau = \tau^*$ come from the optimization of an upper bound on the risk R (in the proof of this theorem). Therefore, these choices may not be necessarily the best choice in practice, even though it gives a good order of magnitude for tuning these parameters. The user could use cross-validation to properly tune them in practice. Note again that $\operatorname{rank}(\bar{M}) \neq 0$ is not required in the above formula, if $\operatorname{rank}(\bar{M}) = 0$ then $\bar{M} = 0$ and we interpret $0\log(1+0/0)$ as 0. The proof of this theorem is provided in the Appendix A. In particular, we can upper bound the infimum on $\bar{M}$ by taking $\bar{M} = M^*$, which leads to the following result.

Corollary 2. *Under the assumptions that Theorem 3 holds, let $r^* = \operatorname{rank}(M^*)$. Put*

$$R_{\delta,m,p,k,r^*,\varepsilon} := \frac{C_1'(1+\delta)^2}{\delta}\frac{\left(4r(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{mkp}{r(k+p)}}\right)+k+p+2\log\frac{2}{\varepsilon}\right)}{m},$$

then

$$\left\|\widehat{XMZ}_{\lambda'^*} - XM^*Z\right\|_{F,\Pi}^2 \leq R_{\delta,m,p,k,r^*,\varepsilon}$$

and in particular, if the sampling distribution Π is uniform,

$$\frac{\|\widehat{XMZ}_{\lambda'^*} - XM^*Z\|_F^2}{nq} \leq R_{\delta,m,p,k,r^*,\varepsilon}.$$

Remark 1. *Up to a log-term, our error rate $r(k+p)/m$ is similar to the best known up-to-date rate derived in [8].*

While Theorem 3 is about the finite sample convergence rate of the posterior mean, it is actually possible to prove that the quasi-posterior $\hat{\rho}_\lambda'$ contracts around M^* at the same rate.

Theorem 4. *Under the same assumptions for Theorem 3, and the same definition for τ and λ^*, let ε_m be any sequence in $(0,1)$ such that $\varepsilon_m \to 0$ when $m \to \infty$. Define*

$$\Omega_m = \Bigg\{ M \in \mathbb{R}^{p\times k} : \|\Pi_C(XMZ) - XM^*Z\|_{F,\Pi}^2 \leq \inf_{1\leq r\leq pk} \inf_{\substack{\bar{M}\in\mathbb{R}^{p\times k}\\ \mathrm{rank}(\bar{M})\leq r}} \Bigg[(1+\delta)\|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \frac{C_1'(1+\delta)^2}{\delta} \frac{\left(4r(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{mkp}{r(k+p)}}\right) + k + p + 2\log\frac{2}{\varepsilon_m}\right)}{m} \Bigg] \Bigg\}.$$

Then

$$\mathbb{E}\Big[\mathbb{P}_{M\sim\hat{\rho}_\lambda'}(M\in\Omega_m)\Big] \geq 1-\varepsilon_m \xrightarrow[m\to\infty]{} 1.$$

The proof of this theorem is provided in Appendix A.

4. Numerical Studies

4.1. Langevin Monte Carlo Implementation

In this section, we propose to sample from the (quasi) posterior, in Sections 2 and 3, by a suitable version of the Langevin Monte Carlo (LMC) algorithm, a gradient-based sampling method. We propose to use a constant step-size unadjusted LMC algorithm; see [43] for more details. The algorithm is given by an initial matrix M_0 and the recursion

$$M_{k+1} = M_k - h\nabla\log\widehat{\rho}_\lambda(M_k) + \sqrt{2h}\,N_k \qquad k = 0,1,\ldots \tag{8}$$

where $h > 0$ is the step-size, $\widehat{\rho}_\lambda$ is the (quasi) posterior and $N_0, N_1, \ldots$ are independent random matrices with independent and identical standard Gaussian entries. We provide a pseudo-code for LMC in Algorithm A1. For small values of the step-size h, the output of Algorithm A1, $\widehat{M}$, is very close to the integral (3) of interest. However, for some h that may not be small enough, the sum can explode [44]. In such cases, we consider to include a Metropolis–Hastings correction in the algorithm. Another possible choice is to take a smaller h and restart the algorithm; although it slows down the algorithm, we keep some control over its time of execution. On the other hand, the Metropolis–Hastings approach ensures the convergence to the desired distribution; however, the algorithm is greatly slowed down because of an additional acceptance/rejection step at each iteration.

Next, we propose a Metropolis–Hasting correction to the LMC algorithm. It guarantees the convergence to the (quasi) posterior, and it also provides a useful way for choosing h. More precisely, we consider the update rule in (8) as a proposal for a new candidate:

$$\tilde{M}_{k+1} = M_k - h\nabla\log\widehat{\rho}_\lambda(M_k) + \sqrt{2h}\,N_k, \qquad k = 0,1,\ldots, \tag{9}$$

Note that the matrix $\tilde{M}_{k+1}$ is normally distributed with mean $M_k - h\nabla\log\widehat{\rho}_\lambda(M_k)$ and the covariance matrices equal to $2h$ times the identity matrices. This proposal is then accepted or rejected according to the Metropolis–Hastings algorithm, where the proposal is accepted with probability:

$$A_{MALA} := \min\left\{1, \frac{\widehat{\rho}_\lambda(\tilde{M}_{k+1})q(M_k|\tilde{M}_{k+1})}{\widehat{\rho}_\lambda(M_k)q(\tilde{M}_{k+1}|M_k)}\right\}, \tag{10}$$

where

$$q(x'|x) \propto \exp\left(-\frac{1}{4h}\|x' - x + h\nabla\log\widehat{\rho}_\lambda(x)\|_F^2\right)$$

is the transition probability density from x to x'. The details of the Metropolis-adjusted Langevin algorithm (denoted by MALA) are presented in Algorithm A2. Compared to the random-walk Metropolis–Hastings, MALA usually proposes moves into regions of higher probability, which are then more likely to be accepted.

We note that the step-size h for MALA is chosen such that the acceptance rate is approximately 0.5 following [45], while the step-size for LMC in the same setting should be smaller than the one for MALA [46].

4.2. Simulation Studies for Biliear Regression

We perform some numerical studies on simulated data to assess the performance of our proposed algorithms. All simulations were conducted using the `R` statistical software [47].

For fixed dimensions $q = 10, k = 20$ of the data, we vary $n = 100$ and $n = 1000$ to check the effect of the samples, whereas the dimensions of the coefficient matrix are varied by $p = 10$ and $p = 100$. The entries of the design matrices X and Z are independently simulated from the standard Gaussian $\mathcal{N}(0, 1)$. Then, given a matrix M^*, we simulate the response matrix Y from model (1) whose entries of the noise matrix E are independent and identically sampled from $\mathcal{N}(0, 1)$. We consider the following setups for the true coefficient matrix:

- Model I: The true coefficient matrix M^* is a rank-2 matrix that is generated as $M^* = B_1 B_2^\top$ where $B_1 \in \mathbb{R}^{p\times 2}, B_2 \in \mathbb{R}^{k\times 2}$ and all entries in B_1 and B_2 are independent and identically sampled from $\mathcal{N}(0, 1)$.
- Model II: An approximate low-rank set up is studied. This series of simulations is similar to the Model I, except that the true coefficient is no longer rank 2, but it can be well approximated by a rank 2 matrix:

$$M^* = 2 \cdot B_1 B_2^\top + U,$$

where U is a matrix whose entries are independent and identically sampled from $\mathcal{N}(0, 0.1)$.

We compare our approaches denoted by LMC and MALA against the (generalized) ordinary least square [2], denoted by OLS. The OLS is defined as follows:

$$\hat{M}_{\text{OLS}} = (X^\top X)^\dagger X^\top Y Z^\top (ZZ^\top)^\dagger$$

where $A^\dagger$ denotes the Moore–Penrose inverse of matrix A. We fixed $\lambda = nq, \tau = 1$, and the LMC and MALA methods are initiated at the OLS estimator and are run with 10,000 iterations, where the first 1000 steps are removed as burn-in periods.

The evaluations are performed by using the mean squared estimation error (Est) and the normalized (relative) mean square error (Nmse):

$$\text{Est} := \|\hat{M} - M^*\|_F^2/(pk), \quad \text{Nmse} := \|\hat{M} - M^*\|_F^2/\|M^*\|_F^2,$$

and the prediction error (Pred) as

$$\text{Pred} := \|X(\hat{M} - M^*)Z\|_F^2/(nq),$$

where $\hat{M}$ here is one of the estimators for LMC, MALA or OLS. We report the averages and the standard deviation of these errors over 100 data replications.

The results of our study are presented in Tables 1 and 2. From the tables, it can be observed that our proposed methods perform similarly to the OLS method. However, the estimation method obtained from the MALA algorithm often results in smaller prediction errors, particularly in high-dimensional settings. This advantage is even more pronounced when the method is applied in the context of inductive matrix completion, as discussed in the next subsection.

4.3. Simulation Studies for Inductive Matrix Completion

The simulation settings for inductive matrix completion are similar to the settings for bilinear regression, Section 4.2. However, after obtaining the response matrix Y, we remove uniformly at random $\kappa = 10\%$ and $\kappa = 30\%$ of the entries of Y. Here, κ denotes the missing rate. We denote the response matrix with missing entries by Y_{miss}.

As in the context of inductive matrix completion, we only observe the response matrix with missing entries, Y_{miss}, and thus we cannot construct the OLS estimator as in the case of bilinear regression. For this purpose, we first impute the missing entries in Y_{miss} by using the `R` package `softImpute` [48], where the rank of M^* is specified as the true rank for matrix Y_{miss}. We denote the resulting imputed matrix by Y_{imp}.

We compare our approaches denoted by LMC and MALA against the (imputed and generalized) ordinary least square, denoted by OLS_imp. The OLS_imp is defined as follows:

$$\hat{M}_{\text{OLS\_imp}} = (X^\top X)^\dagger X^\top Y_{\text{imp}} Z^\top (Z Z^\top)^\dagger$$

where $A^\dagger$ denotes the Moore–Penrose inverse of matrix A. The LMC and MALA methods are initiated at the OLS_imp estimator and are run with 10,000 iterations, where the first 1000 steps are removed as burn-in periods.

As previously discussed in Section 4.2, we present the averages and the standard deviation of the mean squared estimation error (Est), the normalized (relative) mean square error (Nmse), and the prediction error (Pred) over 100 data replications in our results.

The results are detailed in Tables 3 and 4. It is evident from these tables that the results obtained from our MALA method surpass those of the other methods in terms of prediction error in most of the settings considered. This advantage becomes more pronounced as the missing rate in the response matrix increases. Additionally, it is worth noting that our MALA method is robust and performs well in the approximate low-rank setting (model II), while the OLS and LMC methods do not.

Table 1. Simulation results on simulated data in Model I in bilinear regression, with fixed q = 10, k = 20, for different methods, with their standard error in parentheses over 100 replications. (Est: average of estimation errors; Pred: average of prediction errors; Nmse: average of normalized estimation errors).

		Errors	LMC	MALA	OLS
n = 100	$p = 10$	Est	1.0053 (0.5480)	1.0342 (0.5559)	1.0052 (0.5478)
		Pred	0.1138 (0.0171)	0.0985 (0.0151)	0.1014 (0.0154)
		Nmse	0.4931 (0.1178)	0.5100 (0.1207)	0.4930 (0.1178)
	$p = 100$	Est	1.3544 (0.5867)	1.3384 (0.5836)	1.3544 (0.5867)
		Pred	1.0066 (0.0430)	0.8761 (0.0756)	1.0030 (0.0424)
		Nmse	0.7049 (0.2944)	0.6963 (0.2927)	0.7049 (0.2944)
n = 1000	$p = 10$	Est	1.0776 (0.5671)	1.0900 (0.5670)	1.0776 (0.5671)
		Pred	0.0099 (0.0013)	0.0099 (0.0013)	0.0099 (0.0013)
		Nmse	0.5185 (0.1198)	0.5264 (0.1219)	0.5185 (0.1198)
	$p = 100$	Est	0.9662 (0.3240)	0.9688 (0.3244)	0.9662 (0.3240)
		Pred	0.0999 (0.0051)	0.0989 (0.0049)	0.0998 (0.0051)
		Nmse	0.4961 (0.1183)	0.4976 (0.1191)	0.4961 (0.1183)

Table 2. Simulation results on simulated data in Model II (approximate low-rank) in bilinear regression, with fixed $q = 10, k = 20$, for different methods, with their standard error in parentheses over 100 replications. (Est: average of estimation errors; Pred: average of prediction errors; Nmse: average of normalized estimation errors).

		Errors	**LMC**	**MALA**	**OLS**
n = 100	$p = 10$	Est	4.0731 (1.828)	4.0989 (1.821)	4.0731 (1.828)
		Pred	0.1090 (0.0160)	0.0969 (0.0140)	0.0987 (0.0145)
		Nmse	0.5119 (0.1226)	0.5162 (0.1241)	0.5118 (0.1226)
	$p = 100$	Est	4.6047 (1.812)	4.6038 (1.813)	4.6047 (1.812)
		Pred	1.0062 (0.0462)	1.0597 (0.0495)	1.0006 (0.0469)
		Nmse	0.5801 (0.1942)	0.5800 (0.1941)	0.5801 (0.1942)
n = 1000	$p = 10$	Est	3.6733 (1.606)	3.6884 (1.606)	3.6733 (1.606)
		Pred	0.0098 (0.0015)	0.0098 (0.0015)	0.0098 (0.0015)
		Nmse	0.4812 (0.1271)	0.4835 (0.1260)	0.4813 (0.1271)
	$p = 100$	Est	3.9972 (1.375)	3.9986 (1.376)	3.9972 (1.375)
		Pred	0.1000 (0.0043)	0.1032 (0.0057)	0.0999 (0.0043)
		Nmse	0.5013 (0.1061)	0.5014 (0.1063)	0.5013 (0.1062)

Table 3. Simulation results on simulated data in Model I in inductive matrix completion, with fixed $q = 10, k = 20$, for different methods, with their standard error in parentheses over 100 replications. (κ is the missing rate; Est: average of estimation errors; Pred: average of prediction errors; Nmse: average of normalized estimation errors).

		Errors	**LMC**	**MALA**	**OLS_imp**
n = 100 κ = 10%	$p = 10$	Est	1.0559 (0.5060)	1.0803 (0.5122)	1.0559 (0.5060)
		Pred	0.1028 (0.0193)	0.1082 (0.0143)	0.1020 (0.0197)
		Nmse	0.4986 (0.1116)	0.5139 (0.1197)	0.4986 (0.1116)
	$p = 100$	Est	1.4008 (0.8555)	1.3987 (0.8542)	1.4009 (0.8555)
		Pred	1.2250 (0.4568)	1.4468 (0.4137)	1.2252 (0.4570)
		Nmse	0.7148 (0.3591)	0.7136 (0.3581)	0.7148 (0.3591)
n = 100 κ = 30%	$p = 10$	Est	1.0432 (0.4963)	1.0917 (0.5085)	1.0432 (0.4963)
		Pred	0.2402 (0.2705)	0.1447 (0.0204)	0.2446 (0.2780)
		Nmse	0.5242 (0.1257)	0.5538 (0.1335)	0.5242 (0.1257)
	$p = 100$	Est	1.6242 (0.8179)	1.6224 (0.8169)	1.6242 (0.8179)
		Pred	9.8879 (14.11)	10.807 (13.84)	9.8901 (14.11)
		Nmse	0.7993 (0.3340)	0.7985 (0.3334)	0.7993 (0.3340)
n = 1000 κ = 10%	$p = 10$	Est	0.9810 (0.4532)	0.9882 (0.4478)	0.9810 (0.4532)
		Pred	0.0114 (0.0033)	0.0112 (0.0015)	0.0114 (0.0033)
		Nmse	0.4933 (0.1076)	0.4984 (0.1075)	0.4933 (0.1076)
	$p = 100$	Est	1.0063 (0.3465)	1.0088 (0.3471)	1.0063 (0.3465)
		Pred	0.1902 (0.1758)	0.1116 (0.0049)	0.1902 (0.1759)
		Nmse	0.5069 (0.1049)	0.5082 (0.1050)	0.5069 (0.1049)
n = 1000 κ = 30%	$p = 10$	Est	1.0110 (0.4886)	1.0223 (0.4872)	1.0110 (0.4886)
		Pred	0.0539 (0.0599)	0.0141 (0.0019)	0.0540 (0.0599)
		Nmse	0.5129 (0.1030)	0.5206 (0.1043)	0.5129 (0.1030)
	$p = 100$	Est	1.0291 (0.3567)	1.0312 (0.3555)	1.0291 (0.3567)
		Pred	1.7529 (1.914)	0.1475 (0.0078)	1.7530 (1.913)
		Nmse	0.5054 (0.1055)	0.5067 (0.1053)	0.5054 (0.1055)

Table 4. Simulation results on simulated data in Model II (approximate low-rank) in inductive matrix completion, with fixed $q = 10, k = 20$, for different methods, with their standard error in parentheses over 100 replications. (κ is the missing rate; Est: average of estimation errors; Pred: average of prediction errors; Nmse: average of normalized estimation errors).

		Errors	**LMC**	**MALA**	**OLS_imp**
n = 100 imis 10%	$p = 10$	Est	3.8319 (1.691)	3.8749 (1.719)	3.8319 (1.690)
		Pred	0.1604 (0.1271)	0.1092 (0.0153)	0.1598 (0.1322)
		Nmse	0.5116 (0.1154)	0.5169 (0.1147)	0.5116 (0.1155)
	$p = 100$	Est	5.9500 (2.834)	5.9452 (2.835)	5.9500 (2.834)
		Pred	4.7640 (5.272)	4.6964 (5.515)	4.7658 (5.275)
		Nmse	0.7313 (0.3454)	0.7307 (0.3455)	0.7313 (0.3454)
n = 100 imis 30%	$p = 10$	Est	4.1838 (1.850)	4.2535 (1.859)	4.1839 (1.850)
		Pred	0.7221 (0.7562)	0.1498 (0.0183)	0.7371 (0.7741)
		Nmse	0.5182 (0.1128)	0.5283 (0.1147)	0.5182 (0.1128)
	$p = 100$	Est	7.1589 (4.084)	7.1558 (4.083)	7.1589 (4.084)
		Pred	39.899 (52.40)	40.233 (51.76)	39.908 (52.41)
		Nmse	0.8998 (0.3821)	0.8994 (0.3820)	0.8998 (0.3821)
n = 1000 imis 10%	$p = 10$	Est	3.9618 (1.678)	3.9788 (1.677)	3.9618 (1.678)
		Pred	0.0409 (0.0269)	0.0110 (0.0015)	0.0409 (0.0269)
		Nmse	0.4968 (0.1196)	0.4989 (0.1195)	0.4968 (0.1196)
	$p = 100$	Est	4.1153 (1.295)	4.1163 (1.294)	4.1153 (1.295)
		Pred	1.0250 (0.9988)	0.1135 (0.0051)	1.0250 (0.9988)
		Nmse	0.5060 (0.1096)	0.5062 (0.1096)	0.5060 (0.1096)
n = 1000 imis 30%	$p = 10$	Est	4.1647 (1.990)	4.1836 (1.995)	4.1647 (1.990)
		Pred	0.4615 (0.3497)	0.0141 (0.0017)	0.4616 (0.3498)
		Nmse	0.4905 (0.1157)	0.4933 (0.1171)	0.4905 (0.1157)
	$p = 100$	Est	4.0578 (1.400)	4.0565 (1.397)	4.0578 (1.400)
		Pred	8.5608 (6.419)	0.1538 (0.0069)	8.5609 (6.419)
		Nmse	0.4944 (0.1184)	0.4943 (0.1180)	0.4944 (0.1184)

5. Discussion and Conclusions

In this paper, we focus on the problem of bilinear regression and its extension, the problem of inductive matrix completion, where the response matrix contains missing data. We propose a novel approach that combines elements of Bayesian statistics with a quasi-likelihood method. Our proposed method first addresses the problem of bilinear regression using a quasi-Bayesian approach and then adapts this approach to the problem of inductive matrix completion. By making use of a low-rankness assumption and leveraging the powerful PAC-Bayes bound technique, we provide statistical properties for our proposed estimators and for the quasi-posteriors.

Our proposed method includes an efficient gradient-based sampling algorithm that is designed to sample from the (quasi) posterior distribution. This algorithm allows for the approximate computation of mean estimators. These methods, referred to as LMC and MALA, were tested in various simulation studies and were found to perform well when compared to the ordinary least squared method. The ability to accurately sample from the (quasi) posterior distribution and compute mean estimators makes these methods a valuable tool for data analysis and modeling.

There are still some unresolved issues that require further investigation. One of these is the presence of missing data in the covariate matrices X and Z. This can have a significant impact on the analysis and may lead to biased results. Another area that needs further exploration is the assumption of independence and identically distributed data. In some cases, this assumption may not hold, and alternative models that allow for a dispersion matrix may be needed. These are potential topics for future research to address and further our understanding of these issues.

Funding: TTM is supported by the Norwegian Research Council grant number 309960 through the Centre for Geophysical Forecasting at NTNU.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The R codes used in the numerical experiments are available at: https://github.com/tienmt/blr_imc (accessed on 10 February 2023).

Acknowledgments: The author would like to thank the anonymous referees for their useful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

The main technique for our proofs is the oracle-type PAC-Bayes bounds, in the spirit of [35]. We start with a few preliminary lemmas.

Appendix A.1. Preliminary Lemmas

First, we state a version of Bernstein's inequality from Proposition 2.9 page 24 in [49].

Lemma A1 (Bernstein's inequality)**.** *Let $U_1, \dots, U_n$ be independent real valued random variables. Let us assume that there are two constants v and w such that $\sum_{i=1}^{n} \mathbb{E}[U_i^2] \leq v$ and for all integers $k \geq 3$, $\sum_{i=1}^{n} \mathbb{E}\left[(U_i)^k\right] \leq v\frac{k!w^{k-2}}{2}$. Then, for any $\zeta \in (0, 1/w)$,*

$$\mathbb{E} \exp\left[\zeta \sum_{i=1}^{n} [U_i - \mathbb{E}(U_i)]\right] \leq \exp\left(\frac{v\zeta^2}{2(1 - w\zeta)}\right).$$

Another basic tool to derive the PAC-Bayes bounds is Donsker and Varadhan's variational inequality, see Lemma 1.1.3 in Catoni [17] for a proof (among others). From now, for any $\Theta \subset \mathbb{R}^{n_1 \times n_2}$, we let $\mathcal{P}(\Theta)$ denote the set of all probability distributions on Θ equipped with the Borel σ-algebra. For $(\mu, \nu) \in \mathcal{P}(\Theta)^2$, the Kullback–Leibler divergence is defined by $\mathcal{K}(\nu, \mu) = \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\theta)\right)\nu(\mathrm{d}\theta)$ if ν admits a density $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ with respect to μ, and $\mathcal{K}(\nu, \mu) = +\infty$ otherwise.

Lemma A2 (Donsker and Varadhan's variational formula)**.** *Let $\mu \in \mathcal{P}(\Theta)$. For any measurable, bounded function $h : \Theta \to \mathbb{R}$, we have*

$$\log \int \mathrm{e}^{h(\theta)} \mu(\mathrm{d}\theta) = \sup_{\rho \in \mathcal{P}(\Theta)} \left[\int h(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \mu)\right].$$

Moreover, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure μ_h defined by its density with respect to μ

$$\mathrm{d}\mu_h(\theta) = \frac{\mathrm{e}^{h(\theta)} \mathrm{d}\mu}{\int \mathrm{e}^{h(\vartheta)} \mu(\mathrm{d}\vartheta)}. \tag{A1}$$

These two lemmas are the only tools we need to prove Theorems 1 and 2. Their proofs are quite similar, with a few differences. For the sake of simplicity, we will state the common parts of the proofs as a separate result in Lemma A3. Note that the proof of this lemma will use Lemmas A1 and A2.

Lemma A3. *Under Assumptions 1 and 2, put*

$$\alpha = \left(\lambda - \frac{\lambda^2 C_1}{2nq(1-\frac{C_2\lambda}{nq})}\right) \quad \textit{and} \quad \beta = \left(\lambda + \frac{\lambda^2 C_1}{2nq(1-\frac{C_2\lambda}{nq})}\right). \tag{A2}$$

Then, for any $\varepsilon \in (0,1)$, and $\lambda \in (0, nq/C_2)$,

$$\mathbb{E}\left[\int \exp\left\{\alpha\Big(R(M) - R(M^*)\Big) + \lambda\Big(-r(M) + r(M^*)\Big) - \log\left[\frac{d\widehat{\rho}_\lambda}{d\pi}(M)\right] - \log\frac{2}{\varepsilon}\right\}\widehat{\rho}_\lambda(dM)\right] \leq \frac{\varepsilon}{2} \tag{A3}$$

and

$$\mathbb{E} \sup_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})} \exp\left[\beta\Big(-\int R d\rho + R(M^*)\Big) + \lambda\Big(\int r d\rho - r(M^*)\Big) - \mathcal{K}(\rho,\pi) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2}. \tag{A4}$$

Proof of Lemma A3. We prove the first inequality (A10) as follows. Fix any M with $\|XMZ\|_\infty \leq C$ and put

$$T_{ij} = \big(Y_{ij} - (XM^*Z)_{ij}\big)^2 - \big(Y_{ij} - (\Pi_C(XMZ))_{ij}\big)^2.$$

Note that the random variables T_{ij} with $i = 1,\ldots,n; j = 1,\ldots,q$ are independent by construction. We have

$$\begin{aligned}
\sum_{i=1}^n\sum_{j=1}^q \mathbb{E}[T_{ij}^2] &= \sum_{i=1}^n\sum_{j=1}^q \mathbb{E}\Big[\big(2Y_{ij} - (XM^*Z)_{ij} - \Pi_C(XMZ)_{ij}\big)^2\big((XM^*Z)_{ij} - \Pi_C(XMZ)_{ij}\big)^2\Big] \\
&= \sum_{i=1}^n\sum_{j=1}^q \mathbb{E}\Big[\big(2E_{ij} + (XM^*Z)_{ij} - \Pi_C(XMZ)_{ij}\big)^2\big((XM^*Z)_{ij} - \Pi_C(XMZ)_{ij}\big)^2\Big] \\
&\leq \sum_{i=1}^n\sum_{j=1}^q \mathbb{E}\Big[8\Big[E_{ij}^2 + C^2\Big]\left[(XM^*Z)_{ij} - \Pi_C(XMZ)_{ij}\right]^2\Big] \\
&\leq \sum_{i=1}^n\sum_{j=1}^q 8\Big[\sigma^2 + C^2\Big]\mathbb{E}\left[(XM^*Z)_{ij} - (XMZ)_{ij}\right]^2 \\
&\leq 8nq(\sigma^2 + C^2)[R(M) - R(M^*)] \\
&= nqC_1[R(M) - R(M^*)] =: v(M,M^*).
\end{aligned}$$

Next we have, for any integer $k \geq 3$, that

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[(T_{ij})^k\right]\\
\leq&\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[\left|2Y_{ij}-(XM^*Z)_{ij}-\Pi_C(XMZ)_{ij}\right|^k\left|(XM^*Z)_{ij}-\Pi_C(XMZ)_{ij}\right|^k\right]\\
\leq&\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[2^{k-1}\left[|2E_{ij}|^k+(2C)^k\right]\left|(XM^*Z)_{ij}-\Pi_C(XMZ)_{ij}\right|^k\right]\\
\leq&\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[2^{2k-1}\left(|E_{ij}|^k+C^k\right)(2C)^{k-2}\left|(XM^*Z)_{ij}-\Pi_C(XMZ)_{ij}\right|^2\right]\\
\leq&2^{2k-1}\left[\sigma^2k!\xi^{k-2}+C^k\right](2C)^{k-2}\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left|(XM^*Z)_{ij}-(XMZ)_{ij}\right|^2\\
\leq&\frac{2^{3k-3}\left[\sigma^2k!\xi^{k-2}+C^k\right]C^{k-2}}{8(\sigma^2+C^2)}v(M,M^*)\\
\leq&\frac{2^{3k-6}\left[\sigma^2\xi^{k-2}+C^k\right]C^{k-2}}{(\sigma^2+C^2)}k!v(M,M^*)\\
\leq&2^{3k-5}\left[\xi^{k-2}+C^{k-2}\right]C^{k-2}k!v(M,M^*)\\
\leq&2^{3k-4}\max(\xi,C)^{k-2}C^{k-2}k!v(M,M^*)\\
=&[2^3\max(\xi,C)C]^{k-2}2^2k!v(M,M^*)
\end{aligned}
$$

and use the fact that, for any $k\geq 3$, $2^2\leq 2^{3(k-2)}/2$ to obtain

$$\sum_{i=1}^{n}\sum_{j=1}^{q}\mathbb{E}\left[(T_{ij})^k\right]\leq\frac{[2^6\max(\xi,C)C]^{k-2}k!v(M,M^*)}{2}=v(M,M^*)\frac{k!C_2^{k-2}}{2}.$$

Thus, we can apply Lemma A1 with $U_i:=T_i$, $v:=v(M,M^*)$, $w:=C_2$ and $\zeta:=\lambda/nq$. We obtain, for any $\lambda\in(0,nq/w)=(0,nq/C_2)$,

$$
\begin{aligned}
\mathbb{E}\exp\left[\lambda\Big(R(M)-R(M^*)-r(M)+r(M^*)\Big)\right]&\leq\exp\left[\frac{v\lambda^2}{2(nq)^2(1-\frac{w\lambda}{nq})}\right]\\
&=\exp\left[\frac{C_1[R(M)-R(M^*)]\lambda^2}{2nq(1-\frac{C_2\lambda}{nq})}\right].
\end{aligned}
$$

Rearranging terms, and using the definition of α in (A2),

$$\mathbb{E}\exp\left[\alpha\Big(R(M)-R(M^*)\Big)+\lambda\Big(-r(M)+r(M^*)\Big)\right]\leq 1.$$

Multiplying both sides by $\varepsilon/2$ and then integrating with respect to the probability distribution $\pi(.)$, we obtain

$$\int\mathbb{E}\left[\exp\left\{\alpha\Big(R(M)-R(M^*)\Big)+\lambda\Big(-r(M)+r(M^*)\Big)-\log\frac{2}{\varepsilon}\right\}\right]\pi(dM)\leq\frac{\varepsilon}{2}.$$

Next, Fubini's theorem gives

$$\mathbb{E}\left[\int \exp\left[\alpha\Big(R(M)-R(M^*)\Big)+\lambda\Big(-r(M)+r(M^*)\Big)-\log\frac{2}{\varepsilon}\right]\pi(dM)\right]\leq\frac{\varepsilon}{2}.$$

and note that for any measurable function h,

$$\int \exp[h(M)]\pi(dM)=\int \exp\left[h(M)-\log\frac{\mathrm{d}\widehat{\rho}_\lambda}{\mathrm{d}\pi}(M)\right]\widehat{\rho}_\lambda(dM)$$

to obtain (A3).

Let us now prove (A4). Here again, we start with an application of Lemma A1, but this time with $U_i := -T_i$ (we keep $v := v(M, M^*)$, $w := C_2$ and $\zeta := \lambda/nq$). We obtain, for any $\lambda \in (0, nq/C_2)$,

$$\mathbb{E}\exp\left[\lambda\Big(r(M)+r(M^*)-R(M)+R(M^*)\Big)\right]\leq\exp\left[\frac{C_1[R(M)-R(M^*)]\lambda^2}{2nq(1-\frac{C_2\lambda}{nq})}\right].$$

Rearranging terms, using the definition of β in (A2) and multiplying both sides by $\varepsilon/2$, we obtain

$$\mathbb{E}\exp\left[\beta(-R(M)+R(M^*))+\lambda(r(M)-r(M^*))-\log\frac{2}{\varepsilon}\right]\leq\frac{\varepsilon}{2}.$$

We integrate with respect to π and use Fubini to obtain

$$\mathbb{E}\left[\int\exp\left[\beta(-R(M)+R(M^*))+\lambda(r(M)-r(M^*))-\log\frac{2}{\varepsilon}\right]\pi(\mathrm{d}M)\right]\leq\frac{\varepsilon}{2}.$$

Here, we use a different argument from the proof of the first inequality: we use Lemma A2 on the integral, and this gives directly (A4). □

Finally, in both proofs, we will use quite often distributions $\rho \in \mathcal{P}(\mathbb{R}^{p\times k})$ that will be defined as translations of the prior π. We introduce the following notation.

Definition A1. *For any matrix $\bar{M} \in \mathbb{R}^{p\times k}$, we define $\rho_{\bar{M}} \in \mathcal{P}(\mathbb{R}^{p\times k})$ by*

$$\rho_{\bar{M}}(M)=\pi(\bar{M}-M).$$

The following technical lemmas from [31] will be useful in the proofs.

Lemma A4 (Lemma 1 in [31])**.** *We have $\int \|M\|_F^2\pi(\mathrm{d}M)\leq pk\tau^2$.*

Lemma A5 (Lemma 2 in [31])**.** *For any $\bar{M} \in \mathbb{R}^{p\times k}$, we have*

$$\mathcal{K}(\rho_{\bar{M}},\pi)\leq 2\mathrm{rank}(\bar{M})(k+p+2)\log\left(1+\frac{\|\bar{M}\|_F}{\tau\sqrt{2\mathrm{rank}(\bar{M})}}\right)$$

with the convention $0\log(1+0/0)=0$.

Appendix A.2. Proof of Theorem 1

Proof of Theorem 1. An application of Jensen's inequality on inequality (A3) yields

$$\mathbb{E}\exp\left[\alpha\left(\int Rd\widehat{\rho}_\lambda - R(M^*)\right) + \lambda\left(-\int rd\widehat{\rho}_\lambda + r(M^*)\right) - \mathcal{K}(\widehat{\rho}_\lambda,\pi) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2}.$$

Using the standard Chernoff's trick to transform an exponential moment inequality into a deviation inequality, that is: $\exp(x) \geq \mathbf{1}_{\mathbb{R}_+}(x)$, we obtain

$$\mathbb{P}\left\{\left[\alpha\left(\int Rd\widehat{\rho}_\lambda - R(M^*)\right) + \lambda\left(-\int rd\widehat{\rho}_\lambda + r(M^*)\right) - \mathcal{K}(\widehat{\rho}_\lambda,\pi) - \log\frac{2}{\varepsilon}\right] \geq 0\right\} \leq \frac{\varepsilon}{2} \tag{A5}$$

Using (2) we have

$$\begin{aligned}\int Rd\widehat{\rho}_\lambda - R(M^*) &= \frac{1}{nq}\int \|\Pi_C(XMZ) - XM^*Z\|_F^2\widehat{\rho}_\lambda(\mathrm{d}M) \\ &\geq \frac{1}{nq}\left\|\int \Pi_C(XMZ)\widehat{\rho}_\lambda(\mathrm{d}M) - XM^*Z\right\|_F^2 \\ &\geq \frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2\end{aligned}$$

where we used Jensen's inequality in the second line, and the definition of $\widehat{XMZ}_\lambda$ from the second to the third line. Plugging this into our probability bound (A5), and dividing both sides by α, we obtain

$$\mathbb{P}\left\{\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \leq \frac{\int rd\widehat{\rho}_\lambda - r(M^*) + \frac{1}{\lambda}\left[\mathcal{K}(\widehat{\rho}_\lambda,\pi) + \log\frac{2}{\varepsilon}\right]}{\frac{\alpha}{\lambda}}\right\} \geq 1 - \frac{\varepsilon}{2}$$

under the additional condition that λ is such that $\alpha > 0$, which we will assume from now (note that this is satisfied by λ^*). Using Lemma A2, we can rewrite this as

$$\mathbb{P}\left\{\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})}\frac{\int rd\rho - r(M^*) + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\frac{\alpha}{\lambda}}\right\} \geq 1 - \frac{\varepsilon}{2}. \tag{A6}$$

We consider now the consequences of the second inequality in Lemma A3, that is (A4). With Chernoff's trick and rearranging terms a little, we obtain

$$\mathbb{P}\left\{\forall\rho\in\mathcal{P}(\mathbb{R}^{p\times k}), \int rd\rho - r(M^*) \leq \frac{\beta}{\lambda}\left[\int Rd\rho - R(M^*)\right] + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]\right\} \geq 1 - \frac{\varepsilon}{2}.$$

which we can rewrite as, $\forall\rho\in\mathcal{P}(\mathbb{R}^{p\times k})$, with probability at least $1 - \frac{\varepsilon}{2}$,

$$\int rd\rho - r(M^*) \leq \frac{\beta}{\lambda}\int \frac{1}{nq}\|\Pi_C(XMZ) - XM^*Z\|_F^2\rho(\mathrm{d}M) + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]. \tag{A7}$$

Combining (A7) and (A6) with a union bound argument gives the following bound, with probability of at least $1 - \varepsilon$,

$$\begin{aligned}&\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \\ &\qquad\leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})}\frac{\beta\int\frac{1}{nq}\|\Pi_C(XMZ) - XM^*Z\|_F^2\rho(\mathrm{d}M) + 2\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha}.\end{aligned}$$

Noting that, for any (i,j), $(XM^*Z)_{i,j} \in [-C,C]$ implies that

$$|(\Pi_C(XMZ))_{i,j} - (XM^*Z)_{i,j}| \leq |(XMZ)_{i,j} - (XM^*Z)_{i,j}|$$

and thus

$$\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})} \frac{\beta\int \frac{1}{nq}\|XMZ - XM^*Z\|_F^2\rho(\mathrm{d}M) + 2\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha}.$$

The end of the proof consists in making the right-hand side in the inequality more explicit. In order to do so, we restrict the infimum bound above to the distributions given by Definition A1:

$$\mathbb{P}\left\{\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \right. \\ \left. \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}} \frac{\beta\int \frac{1}{nq}\|XMZ - XM^*Z\|_F^2\rho_{\bar{M}}(\mathrm{d}M) + 2\left[\mathcal{K}(\rho_{\bar{M}},\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha}\right\} \geq 1-\varepsilon. \quad \text{(A8)}$$

We see immediately that Dalalyan's lemma will be extremely useful for that. First, Lemma A5 provides an upper bound on $\mathcal{K}(\rho_{\bar{M}},\pi)$. Moreover,

$$\int \|XMZ - XM^*Z\|_F^2\rho_{\bar{M}}(\mathrm{d}M) \\ \leq \int \|X\bar{M}Z - XM^*Z - XMZ\|_F^2\pi(\mathrm{d}M) \\ = \|X\bar{M}Z - XM^*Z\|_F^2 - 2\int\sum_{i,j}(X\bar{M}Z - XM^*Z)_{j,i}(XMZ)_{i,j}\pi(\mathrm{d}M) + \int \|XMZ\|_F^2\pi(\mathrm{d}M).$$

The second term in the right-hand side is null because π is centered, and thus

$$\begin{aligned}\int \|XMZ - XM^*Z\|_F^2\rho_{\bar{M}}(\mathrm{d}M) &\leq \|X\bar{M}Z - XM^*Z\|_F^2 + \int \|XMZ\|_F^2\,\pi(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_F^2 + \|X\|_F^2\|Z\|_F^2\int \|M\|_F^2\,\pi(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_F^2 + \|X\|_F^2\|Z\|_F^2 pk\tau^2\end{aligned}$$

where we used elementary properties of the Frobenius norm, and Lemma A4 in the last line. We can now plug this (and Lemma A5) back into (A8) to obtain

$$\mathbb{P}\left\{\frac{1}{nq}\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|_F^2 \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}}\left[\frac{\beta}{\alpha}\frac{1}{nq}\|X\bar{M}Z - XM^*Z\|_F^2 + \frac{\beta}{\alpha}\frac{1}{nq}\|X\|_F^2\|Z\|_F^2 pk\tau^2 \right.\right. \\ \left.\left. + \frac{1}{\alpha}\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1 + \frac{\|\bar{M}\|_F}{\tau\sqrt{2\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon}\right)\right]\right\} \geq 1-\varepsilon.$$

We are now making the constants explicit. First, if $\lambda \leq nq/(2C_2)$, then $2nq(1 - C_2\lambda/nq) \geq np$ and thus

$$\frac{\beta}{\alpha} = \frac{1 + \frac{\lambda C_1}{2nq(1-\frac{C_2\lambda}{nq})}}{1 - \frac{\lambda C_1}{2nq(1-\frac{C_2\lambda}{nq})}} \leq \frac{1 + \frac{\lambda C_1}{nq}}{1 - \frac{\lambda C_1}{nq}}.$$

Then, $\lambda \leq \frac{nq\delta}{C_1(1+\delta)}$ leads to $\frac{\beta}{\alpha} \leq (1+\delta)$.

Note that $\lambda^* = nq \min(1/(2C_2), \delta/[C_1(1+\delta)])$ satisfies these two conditions, so from now, $\lambda = \lambda^*$. We also use the following:

$$\frac{1}{\alpha} = \frac{1}{\lambda^*\left(1 - \frac{\lambda^* C_1}{2nq(1-C_2\lambda^*/nq)}\right)} \leq \frac{\beta}{\lambda^*\alpha} \leq \frac{(1+\delta)}{nq\min(1/(2C_2), \delta/[C_1(1+\delta)])} \leq \frac{C_1(1+\delta)^2}{nq\delta}.$$

So far the bound is:

$$\mathbb{P}\left\{\frac{1}{nq}\left\|\widehat{XMZ}_{\lambda^*} - XM^*Z\right\|_F^2 \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}}\left[\frac{(1+\delta)}{nq}\|X\bar{M}Z - XM^*Z\|_F^2 + \frac{(1+\delta)}{nq}\|X\|_F^2\|Z\|_F^2 pk\tau^2 + \frac{C_1(1+\delta)^2\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1+\frac{\|\bar{M}\|_F}{\tau\sqrt{2\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon}\right)}{nq\delta}\right]\right\} \geq 1-\varepsilon.$$

In particular, with probability at least $1-\varepsilon$, the choice $\tau^2 = C_1(k+p)/(nkq\|X\|_F^2\|Z\|_F^2)$ gives

$$\frac{1}{nq}\left\|\widehat{XMZ}_{\lambda^*} - XM^*Z\right\|_F^2 \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}}\left[\frac{(1+\delta)}{nq}\|X\bar{M}Z - XM^*Z\|_F^2 + \frac{C_1(1+\delta)(k+p)}{nq} + \frac{C_1(1+\delta)^2\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1+\frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1}}\sqrt{\frac{nkq}{(k+p)\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon}\right)}{nq\delta}\right].$$

□

Appendix A.3. Proof of Theorem 2

Proof of Theorem 2. We also start with an application of Lemma A3, and focus on (A3), applied to $\varepsilon := \varepsilon_n$, that is:

$$\mathbb{E}\left[\int \exp\left\{\alpha\left(R(M) - R(M^*)\right) + \lambda\left(-r(M) + r(M^*)\right) - \log\left[\frac{d\widehat{\rho}_\lambda}{d\pi}(M)\right] - \log\frac{2}{\varepsilon_n}\right\}\widehat{\rho}_\lambda(dM)\right] \leq \frac{\varepsilon_n}{2}.$$

Using Chernoff's trick, this gives

$$\mathbb{E}\left[\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M \in \mathcal{A}_n)\right] \geq 1 - \frac{\varepsilon_n}{2}$$

where

$$\mathcal{A}_n = \left\{M : \alpha\left(R(M) - R(M^*)\right) + \lambda\left(-r(M) + r(M^*)\right) \leq \log\left[\frac{d\widehat{\rho}_\lambda}{d\pi}(M)\right] + \log\frac{2}{\varepsilon_n}\right\}.$$

Using the definition of $\widehat{\rho}_\lambda$, for $M \in \mathcal{A}_n$ we have

$$\begin{aligned}\alpha\Big(R(M)-R(M^*)\Big) &\leq \lambda\Big(r(M)-r(M^*)\Big)+\log\left[\frac{d\widehat{\rho}_\lambda}{d\pi}(M)\right]+\log\frac{2}{\varepsilon_n}\\ &\leq -\log\int\exp[-\lambda r(M)]\pi(\mathrm{d}M)-\lambda r(M^*)+\log\frac{2}{\varepsilon_n}\\ &=\lambda\Big(\int r(M)\widehat{\rho}_\lambda(\mathrm{d}M)-r(M^*)\Big)+\mathcal{K}(\widehat{\rho}_\lambda,\pi)+\log\frac{2}{\varepsilon_n}\\ &=\inf_\rho\Big\{\lambda\Big(\int r(M)\rho(\mathrm{d}M)-r(M^*)\Big)+\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\Big\}.\end{aligned}$$

Now, let us define

$$\mathcal{B}_n=\left\{\forall\rho:\beta\Big(-\int Rd\rho+R(M^*)\Big)+\lambda\Big(\int rd\rho-r(M^*)\Big)\leq\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\right\}.$$

Using (A4), we have that

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{B}_n}\Big]\geq 1-\frac{\varepsilon_n}{2}.$$

We will now prove that, if λ is such that $\alpha>0$,

$$\mathbb{E}\Big[\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\in\mathcal{M}_n)\Big]\geq\mathbb{E}\Big[\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\in\mathcal{A}_n)\mathbf{1}_{\mathcal{B}_n}\Big]$$

which, together with

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\in\mathcal{A}_n)\mathbf{1}_{\mathcal{B}_n}\Big]&=\mathbb{E}\Big[(1-\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\notin\mathcal{A}_n))(1-\mathbf{1}_{\mathcal{B}_n^c})\Big]\\ &\geq\mathbb{E}\Big[1-\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\notin\mathcal{A}_n)-\mathbf{1}_{\mathcal{B}_n^c}\Big]\\ &\geq 1-\varepsilon_n\end{aligned}$$

will bring

$$\mathbb{E}\Big[\mathbb{P}_{M\sim\widehat{\rho}_\lambda}(M\in\mathcal{M}_n)\Big]\geq 1-\varepsilon_n.$$

In order to do so, assume that we are on the set $\mathcal{B}_n$, and let $M\in\mathcal{A}_n$. Then,

$$\begin{aligned}\alpha\Big(R(M)-R(M^*)\Big)&\leq\inf_\rho\Big\{\lambda\Big(\int r(M)\rho(\mathrm{d}M)-r(M^*)\Big)+\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon_n}\Big\}\\ &\leq\inf_\rho\Big\{\beta\Big(\int R(M)\rho(\mathrm{d}M)-R(M^*)\Big)+2\mathcal{K}(\rho,\pi)+2\log\frac{2}{\varepsilon_n}\Big\}\end{aligned}$$

that is,

$$R(M)-R(M^*)\leq\inf_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})}\frac{\beta[\int Rd\rho-R(M^*)]+2\big[\mathcal{K}(\rho,\pi)+\log\frac{2}{\varepsilon}\big]}{\alpha}$$

or, rewriting it in terms of norms,

$$\|\Pi_C(XMZ)-XM^*Z\|_F^2\leq\inf_{\bar{M}\in\mathbb{R}^{p\times k}}\frac{\beta\int\|XMZ-XM^*Z\|_F^2\rho_{\bar{M}}(\mathrm{d}M)+2\big[\mathcal{K}(\rho_{\bar{M}},\pi)+\log\frac{2}{\varepsilon}\big]}{\alpha}.$$

We upper-bound the right-hand side exactly as in the proof of Theorem 1, which gives $M\in\mathcal{M}_n$. □

Appendix A.4. Proof of Theorem 3

Lemma A6. *Under Assumptions 1 and 3, put*

$$\alpha' = \left(\lambda - \frac{\lambda^2 C_1'}{2m(1-\frac{C_2'\lambda}{m})}\right) \quad \text{and} \quad \beta' = \left(\lambda + \frac{\lambda^2 C_1'}{2m(1-\frac{C_2'\lambda}{m})}\right). \tag{A9}$$

Then for any $\varepsilon \in (0,1)$, and $\lambda \in (0, m/C_2')$,

$$\mathbb{E}\left[\int \exp\left\{\alpha'\Big(R'(M) - R'(M^*)\Big) + \lambda\Big(-r'(M) + r'(M^*)\Big) - \log\left[\frac{d\widehat{\rho}_\lambda'}{d\pi}(M)\right] - \log\frac{2}{\varepsilon}\right\}\widehat{\rho}_\lambda'(dM)\right] \le \frac{\varepsilon}{2} \tag{A10}$$

and

$$\mathbb{E}\sup_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})}\exp\left[\beta'\Big(-\int R'd\rho + R'(M^*)\Big) + \lambda\Big(\int r'd\rho - r'(M^*)\Big) - \mathcal{K}(\rho,\pi) - \log\frac{2}{\varepsilon}\right] \le \frac{\varepsilon}{2}. \tag{A11}$$

Proof of Lemma A6. The inequality (A10) is proved in a similar way to the proof of Lemma A3. That is, we apply Lemma A1 to the following independent random variables

$$V_i = (Y_i - (XM^*Z)_i)^2 - (Y_i - (\Pi_C(XMZ))_i)^2, i = 1,\ldots,m.$$

The proof of the inequality (A11) is processed similar in the proof of Lemma A3 in which we apply Lemma A1 to the independent random variables $-V_i, i = 1,\ldots,m$. □

Proof of Theorem 3. Similar to the proof of Theorem 1, until the (A6), and noting that using (6), we have

$$\begin{aligned}\int R'd\widehat{\rho}_\lambda - R'(M^*) &= \int \|\Pi_C(XMZ) - XM^*Z\|^2_{F,\Pi}\widehat{\rho}_\lambda'(\mathrm{d}M) \\ &\ge \left\|\int \Pi_C(XMZ)\widehat{\rho}_\lambda'(\mathrm{d}M) - XM^*Z\right\|^2_{F,\Pi} \\ &\ge \left\|\widehat{XMZ}_\lambda - XM^*Z\right\|^2_{F,\Pi},\end{aligned}$$

and thus, we obtain

$$\mathbb{P}\left\{\left\|\widehat{XMZ}_\lambda - XM^*Z\right\|^2_{F,\Pi} \le \inf_{\rho\in\mathcal{P}(\mathbb{R}^{p\times k})}\frac{\int r'd\rho - r'(M^*) + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\frac{\alpha'}{\lambda}}\right\} \ge 1 - \frac{\varepsilon}{2}. \tag{A12}$$

We consider now the consequences of inequality (A11) in Lemma A6. With Chernoff's trick and rearranging terms a little, we obtain $\forall\rho \in \mathcal{P}(\mathbb{R}^{p\times k})$, with probability at least $1-\frac{\varepsilon}{2}$,

$$\int r'd\rho - r'(M^*) \le \frac{\beta'}{\lambda}\int \|\Pi_C(XMZ) - XM^*Z\|^2_{F,\Pi}\rho(\mathrm{d}M) + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]. \tag{A13}$$

Combining (A13) and (A12) with a union bound argument gives the bound and noting that for any (i,j), $(XM^*Z)_{i,j} \in [-C, C]$ implies that $|(\Pi_C(XMZ))_{i,j} - (XM^*Z)_{i,j}| \leq |(XMZ)_{i,j} - (XM^*Z)_{i,j}|$ and thus

$$\mathbb{P}\left\{ \left\| \widehat{XMZ}_\lambda - XM^*Z \right\|_{F,\Pi}^2 \right. \\ \left. \leq \inf_{\rho \in \mathcal{P}(\mathbb{R}^{p\times k})} \frac{\beta' \int \|XMZ - XM^*Z\|_{F,\Pi}^2 \rho(\mathrm{d}M) + 2\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha'} \right\} \geq 1 - \varepsilon.$$

We are now making the right-hand side in the inequality more explicit. In order to do so, we restrict the infimum bound above to the distributions given by Definition A1:

$$\mathbb{P}\left\{ \left\| \widehat{XMZ}_\lambda - XM^*Z \right\|_{F,\Pi}^2 \leq \right. \\ \left. \inf_{\bar{M} \in \mathbb{R}^{p\times k}} \frac{\beta' \int \|XMZ - XM^*Z\|_{F,\Pi}^2 \rho_{\bar{M}}(\mathrm{d}M) + 2\left[\mathcal{K}(\rho_{\bar{M}},\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha'} \right\} \geq 1 - \varepsilon. \quad \text{(A14)}$$

We see immediately that Dalalyan's lemma will be extremely useful for that. First, Lemma A5 provides an upper bound on $\mathcal{K}(\rho_{\bar{M}},\pi)$.

Moreover,

$$\begin{aligned} &\int \|XMZ - XM^*Z\|_{F,\Pi}^2 \rho_{\bar{M}}(\mathrm{d}M) \\ &\leq \int \|X\bar{M}Z - XM^*Z - XMZ\|_{F,\Pi}^2 \pi(\mathrm{d}M) \\ &= \|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 - 2\int \sum_{i,j} \Pi_{i,j}(X\bar{M}Z - XM^*Z)_{j,i}(XMZ)_{i,j}\pi(\mathrm{d}M) + \\ &\int \|XMZ\|_{F,\Pi}^2 \pi(\mathrm{d}M). \end{aligned}$$

The second term in the above right-hand side is null because π is centered, and thus

$$\begin{aligned} &\int \|XMZ - XM^*Z\|_{F,\Pi}^2 \rho_{\bar{M}}(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \int \|XMZ\|_{F,\Pi}^2 \pi(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \int \|XMZ\|_F^2\, \pi(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \|X\|_F^2\|Z\|_F^2 \int \|M\|_F^2\, \pi(\mathrm{d}M) \\ &\leq \|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \|X\|_F^2\|Z\|_F^2 pk\tau^2 \end{aligned}$$

where we used the elementary properties of the Frobenius norm, and Lemma A4 in the last line. We can now plug this (and Lemma A5) back into (A14) to obtain:

$$\mathbb{P}\left\{ \left\| \widehat{XMZ}_\lambda - XM^*Z \right\|_{F,\Pi}^2 \leq \inf_{\bar{M} \in \mathbb{R}^{p\times k}} \left[\frac{\beta'}{\alpha'}\|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \frac{\beta'}{\alpha'}\|X\|_F^2\|Z\|_F^2 pk\tau^2 \right. \right. \\ \left. \left. + \frac{1}{\alpha'}\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1 + \frac{\|\bar{M}\|_F}{\tau\sqrt{2\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon} \right) \right] \right\} \geq 1 - \varepsilon.$$

We are now making the constants explicit. First, if $\lambda \leq m/(2C_2')$, then $2m(1 - C_2'\lambda/m) \geq m$ and thus

$$\frac{\beta'}{\alpha'} = \frac{1 + \frac{\lambda C_1'}{2m(1-\frac{C_2'\lambda}{m})}}{1 - \frac{\lambda C_1'}{2m(1-\frac{C_2'\lambda}{m})}} \leq \frac{1 + \frac{\lambda C_1'}{m}}{1 - \frac{\lambda C_1'}{m}}.$$

Then, $\lambda \leq \frac{m\delta}{C_1'(1+\delta)}$ leads to $\frac{\beta'}{\alpha'} \leq (1+\delta)$.

Note that $\lambda'^* = m\min(1/(2C_2'), \delta/[C_1'(1+\delta)])$ satisfies these two conditions, so from now $\lambda = \lambda'^*$. We also use the following:

$$\frac{1}{\alpha'} = \frac{1}{\lambda'^*\left(1 - \frac{\lambda'^* C_1'}{2m(1-C_2\lambda'^*/m)}\right)} \leq \frac{\beta'}{\lambda'^*\alpha'} \leq \frac{(1+\delta)}{m\min(1/(2C_2'), \delta/[C_1'(1+\delta)])} \leq \frac{C_1'(1+\delta)^2}{m\delta}.$$

So far the bound is:

$$\mathbb{P}\Bigg\{\left\|\widehat{XMZ}_{\lambda'^*} - XM^*Z\right\|_{F,\Pi}^2 \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}}\Bigg[(1+\delta)\|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \\ (1+\delta)\|X\|_F^2\|Z\|_F^2 pk\tau^2 + \\ \frac{C_1'(1+\delta)^2\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1 + \frac{\|\bar{M}\|_F}{\tau\sqrt{2\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon}\right)}{m\delta}\Bigg]\Bigg\} \geq 1 - \varepsilon.$$

In particular, with probability at least $1-\varepsilon$, the choice $\tau^2 = C_1'(k+p)/(mkp\|X\|_F^2\|Z\|_F^2)$ gives

$$\left\|\widehat{XMZ}_{\lambda'^*} - XM^*Z\right\|_{F,\Pi}^2 \leq \inf_{\bar{M}\in\mathbb{R}^{p\times k}}\Bigg[(1+\delta)\|X\bar{M}Z - XM^*Z\|_{F,\Pi}^2 + \frac{C_1'(1+\delta)(k+p)}{m} + \\ \frac{C_1'(1+\delta)^2\left(4\mathrm{rank}(\bar{M})(k+p+2)\log\left(1 + \frac{\|X\|_F\|Z\|_F\|\bar{M}\|_F}{\sqrt{C_1'}}\sqrt{\frac{mkp}{(k+p)\mathrm{rank}(\bar{M})}}\right) + 2\log\frac{2}{\varepsilon}\right)}{m\delta}\Bigg].$$

□

Appendix A.5. Proof of Theorem 4

Proof. The proof is proceeded completely similar to the proof of Theorem 2, in Appendix A.3. □

Appendix B. Comments on Algorithm Implementation

For the case of inductive matrix completion, we write the logarithm of the density of the posterior

$$\log\widehat{\rho}_\lambda(M) = -\frac{\lambda}{n}\sum_{i=1}^{n}(Y_i - (\Pi_C(XMZ))_i)^2 - \frac{p+m+2}{2}\log\det(\tau^2\mathbf{I}_m + MM^\intercal).$$

Let us now differentiate this expression in M. Note that the term $(Y_i - (\Pi_C(XMZ))_i)^2$ does actually not depend on M locally if $(XMZ)_i \notin [-C, C]$, in this case its differential with respect to M is $\mathbf{0}_{p\times m}$. Otherwise, $(Y_i - (\Pi_C(XMZ))_i)^2 = (Y_i - (XMZ)_i)^2$. In order to be able to differentiate the term $(XMZ)_i$, let us introduce a notation for the entries of $\mathcal{I}_i$: $\mathcal{I}_i = (a_i, b_i)$. Then $\nabla(XMZ)_i = D$ where the matrix $D \in \mathbb{R}^{p\times m}$ satisfies $D_{x,y} = \mathbf{1}_{\{x=b_j\}}X_{a_j,y}$. Then

$$\nabla \log \widehat{\rho}_\lambda(M) = \frac{2\lambda}{n}\sum_{i=1}^{n}(\nabla(XMZ)_i)(Y_i - (XMZ)_i)\mathbf{1}_{\{|(XMZ)_i|<C\}} - (p+m+2)(\tau^2\mathbf{I}_m + MM^\intercal)^{-1}M.$$

The above calculation still requires to calculate a $p \times p$ matrix inversion at each iteration; for very large p, this might be expensive and can slow down the algorithm. Therefore, we could replace this matrix inversion by its accurately approximation through a convex optimization. It is noted that the matrix $\mathbf{B} := (\tau^2\mathbf{I}_m + MM^\intercal)^{-1}M$ is the solution to the following convex optimization problem: $\min_{\mathbf{B}}\left\{\|\mathbf{I}_p - M^\top\mathbf{B}\|_F^2 + \tau^2\|\mathbf{B}\|_F^2\right\}$. The solution of this optimization problem can be conveniently obtained by using the package '`glmnet`' [50] (with the family option '`mgaussian`'). This avoids performing matrix inversion or other costly calculation. However, we note here that the LMC algorithm is being used with approximate gradient evaluation; theoretical assessment of this approach can be found in [51].

Algorithm A1 LMC

```
Input: The data.
Parameters: Positive real numbers $\lambda, \tau, h, T$.
Output: The matrix $\widehat{M}$
Initialize: $M_0, \widehat{M} = \mathbf{0}_{m\times p}$
for $k \leftarrow 1$ to $T$ do
    Sample $M_k$ from (8);
    $\widehat{M} \leftarrow \widehat{M} + M_k/T$
end for
```

Algorithm A2 MALA

```
Input: The data.
Parameters: Positive real numbers $\lambda, \tau, h, T$
Output: The matrix $\widehat{M}$
Initialize: $M_0; \widehat{M} = \mathbf{0}_{m\times p}$
for $k = 1$ to $T$ do
    Sample $\tilde{M}_k$ from (9).
    Set $M_k = \tilde{M}_k$ with probability $A_{MALA}$, from (10), otherwise $M_k = M_{k-1}$ .
    $\widehat{M} \leftarrow \widehat{M} + M_k/T$ .
end for
```

References

1. Rosen, D.V. Bilinear Regression with Rank Restrictions on the Mean and Dispersion Matrix. In *Methodology and Applications of Statistics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 193–211.
2. Von Rosen, D. *Bilinear regression analysis: An Introduction*; Lecture Notes in Statistics; Springer: Berlin/Heidelberg, Germany, 2018; Volume 220.
3. Potthoff, R.F.; Roy, S. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **1964**, *51*, 313–326.
4. Woolson, R.F.; Leeper, J.D. Growth curve analysis of complete and incomplete longitudinal data. *Commun. Stat.-Theory Methods* **1980**, *9*, 1491–1513.
5. Kshirsagar, A.; Smith, W. *Growth Curves*; CRC Press: Boca Raton, FL, USA, 1995; Volume 145.
6. Jana, S. Inference for Generalized Multivariate Analysis of Variance (GMANOVA) Models and High-Dimensional Extensions. Ph.D. Thesis, Mcmaster University, Hamilton, ON, Canada, 2017. Available online: http://hdl.handle.net/11375/22043 (accessed on 26 January 2023).
7. Natarajan, N.; Dhillon, I.S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* **2014**, *30*, i60–i68.
8. Zilber, P.; Nadler, B. Inductive Matrix Completion: No Bad Local Minima and a Fast Algorithm. In Proceedings of the 39th ICML Baltimore, MA, USA, 17–23 July 2022; PMLR; Volume 162, pp. 27671–27692.
9. Zhang, W.; Xu, H.; Li, X.; Gao, Q.; Wang, L. DRIMC: An improved drug repositioning approach using Bayesian inductive matrix completion. *Bioinformatics* **2020**, *36*, 2839–2847.

10. Hsieh, C.J.; Natarajan, N.; Dhillon, I. PU learning for matrix completion. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2445–2453.
11. Jana, S.; Balakrishnan, N.; Hamid, J.S. Bayesian growth curve model useful for high-dimensional longitudinal data. *J. Appl. Stat.* **2019**, *46*, 814–834.
12. Knoblauch, J.; Jewson, J.; Damoulas, T. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *J. Mach. Learn. Res.* **2022**, *23*, 1–109.
13. Bissiri, P.G.; Holmes, C.C.; Walker, S.G. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2016**, *78*, 1103–1130. https://doi.org/10.1111/rssb.12158.
14. Grünwald, P.; Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **2017**, *12*, 1069–1103.
15. McAllester, D. Some PAC-Bayesian theorems. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; ACM: New York, NY, USA, 1998; pp. 230–234.
16. Shawe-Taylor, J.; Williamson, R. A PAC analysis of a Bayes estimator. In Proceedings of the Tenth Annual Conference on Computational Learning Theory, Nashville, TN, USA, 6–9 July 1997; ACM: New York, NY, USA, 1997; pp. 2–9.
17. Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*; IMS Lecture Notes—Monograph Series, 56; Institute of Mathematical Statistics: Beachwood, OH, USA, 2007; p. xii+163.
18. Guedj, B. A primer on PAC-Bayesian learning. *arXiv* **2019**, arXiv:1901.05353.
19. Alquier, P. User-friendly introduction to PAC-Bayes bounds. *arXiv* **2021**, arXiv:2110.11216.
20. Mai, T.T.; Alquier, P. A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electron. J. Statist.* **2015**, *9*, 823–841. https://doi.org/10.1214/15-EJS1020.
21. Cottet, V.; Alquier, P. 1-Bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Mach. Learn.* **2018**, *107*, 579–603.
22. Mai, T.T.; Alquier, P. Pseudo-Bayesian quantum tomography with rank-adaptation. *J. Stat. Plan. Inference* **2017**, *184*, 62–76.
23. Mai, T.T.; Alquier, P. Optimal quasi-Bayesian reduced rank regression with incomplete response. *arXiv* **2022**, arXiv:2206.08619.
24. Jain, P.; Dhillon, I.S. Provable inductive matrix completion. *arXiv* **2013**, arXiv:1306.0626.
25. Candès, E.J.; Plan, Y. Matrix completion with noise. *Proc. IEEE* **2010**, *98*, 925–936.
26. Koltchinskii, V.; Lounici, K.; Tsybakov, A.B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **2011**, *39*, 2302–2329. https://doi.org/10.1214/11-AOS894.
27. Foygel, R.; Shamir, O.; Srebro, N.; Salakhutdinov, R. Learning with the weighted trace-norm under arbitrary sampling distributions. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 2133–2141.
28. Klopp, O. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **2014**, *20*, 282–303. https://doi.org/10.3150/12-BEJ486.
29. Negahban, S.; Wainwright, M.J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **2012**, *13*, 1665–1697.
30. Dalalyan, A.S.; Tsybakov, A.B. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. Syst. Sci.* **2012**, *78*, 1423–1443.
31. Dalalyan, A.S. Exponential weights in multivariate regression and a low-rankness favoring prior. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* **2020**, *56*, 1465–1483.
32. Anderson, T.W. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.* **1951**, *22*, 327–351.
33. Izenman, A.J. Modern multivariate statistical techniques. *Regres. Classif. Manifold Learn.* **2008**, *10*, 978.
34. Dalalyan, A.; Tsybakov, A.B. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **2008**, *72*, 39–61.
35. Catoni, O. Statistical learning theory and stochastic optimization. In *Saint-Flour Summer School on Probability Theory 2001*; Picard, J., Ed.; Lecture Notes in Mathematics; Springer: Berlin, Germany, 2004; Volume 1851, p. viii+272. https://doi.org/10.1007/b99352.
36. Alquier, P.; Ridgway, J.; Chopin, N. On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **2016**, *17*, 8374–8414.
37. Rigollet, P.; Tsybakov, A.B. Sparse estimation by exponential weighting. *Stat. Sci.* **2012**, *27*, 558–575.
38. Dalalyan, A.S.; Grappin, E.; Paris, Q. On the exponentially weighted aggregate with the Laplace prior. *Ann. Stat.* **2018**, *46*, 2452–2478.
39. Candes, E.J.; Wakin, M.B.; Boyd, S.P. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.* **2008**, *14*, 877–905.
40. Yang, L.; Fang, J.; Duan, H.; Li, H.; Zeng, B. Fast low-rank Bayesian matrix completion with hierarchical gaussian prior models. *IEEE Trans. Signal Process.* **2018**, *66*, 2804–2817.
41. Luo, C.; Liang, J.; Li, G.; Wang, F.; Zhang, C.; Dey, D.K.; Chen, K. Leveraging mixed and incomplete outcomes via reduced-rank modeling. *J. Multivar. Anal.* **2018**, *167*, 378–394.
42. Hoeffding, W. Probability Inequalities for Sums of Bounded Random Variables. *J. Am. Stat. Assoc.* **1963**, *58*, 13–30.
43. Durmus, A.; Moulines, E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **2019**, *25*, 2854–2882.

44. Roberts, G.O.; Stramer, O. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **2002**, *4*, 337–357.
45. Roberts, G.O.; Rosenthal, J.S. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1998**, *60*, 255–268.
46. Dalalyan, A.S. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2017**, *3*, 651–676.
47. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
48. Hastie, T.; Mazumder, R. *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD*, 2021. R Package Version 1.4-1. Available online: https://cran.r-project.org/package=softImpute (accessed on 26 January 2023).
49. Massart, P. *Concentration Inequalities and Model Selection*; Lecture Notes in Mathematics; Springer: Berlin, Germany, 2007; Volume 1896, p. xiv+337.
50. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
51. Dalalyan, A.S.; Karagulyan, A. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Process. Their Appl.* **2019**, *129*, 5278–5311.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.