

Conformalized Tensor Completion with Riemannian Optimization

Hu Sun

Department of Statistics, University of Michigan, Ann Arbor
and

Yang Chen*

Department of Statistics and Michigan Institute for Data Science
University of Michigan, Ann Arbor

May 2, 2024

Abstract

Tensor data, or multi-dimensional array, is a data format popular in multiple fields such as social network analysis, recommender systems, and brain imaging. It is not uncommon to observe tensor data containing missing values and tensor completion aims at estimating the missing values given the partially observed tensor. Sufficient efforts have been spared on devising scalable tensor completion algorithms but few on quantifying the uncertainty of the estimator. In this paper, we nest the uncertainty quantification (UQ) of tensor completion under a split conformal prediction framework and establish the connection of the UQ problem to a problem of estimating the missing propensity of each tensor entry. We model the data missingness of the tensor with a tensor Ising model parameterized by a low-rank tensor parameter. We propose to estimate the tensor parameter by maximum pseudo-likelihood estimation (MPLE) with a Riemannian gradient descent algorithm. Extensive simulation studies have been conducted to justify the validity of the resulting conformal interval. We apply our method to the regional total electron content (TEC) reconstruction problem.

Keywords: Tensor Completion; Uncertainty Quantification; Conformal Prediction; Riemannian Gradient Descent; Binary Tensor Decomposition.

*Email: ychenang@umich.edu

1 Introduction

Tensor, or multi-dimensional array, has become a popular data format in several applications such as collaborative filtering (Bi et al. 2018), financial time series modeling (Li & Xiao 2021), hypergraph networks analysis (Ke et al. 2019), neuroimaging study (Li et al. 2018) and astrophysics imaging analysis (Sun, Manchester, Jin, Liu & Chen 2023). Tensor gains this popularity due to its efficient representation of structural high-dimensional data. For example, in collaborative filtering (Bi et al. 2018), the rating data is naturally embedded in a 3-way tensor with user \times item \times context with each entry being the rating by a user on a certain item under a specific context. In neuroimaging analysis (Wei et al. 2023), as another example, each brain voxel in the 3-way tensor is identified by its coordinate in the 3-D Euclidean space.

Tensor completion (Yuan & Zhang 2016, Xia et al. 2021, Cai, Li, Poor & Chen 2022) is a technique that provides an estimator of the tensor when missing values are present. Typically, given only one tensor sample with missingness, tensor completion aims at finding a low-rank tensor that best imputes the missing entries. Various optimization techniques (Kressner et al. 2014, Yuan & Zhang 2016, Lee & Wang 2020, Cai, Li, Poor & Chen 2022) have been proposed for computationally efficient tensor completion and the statistical error of tensor completion has also been carefully investigated (Xia et al. 2021).

However, given the progress above, very little work has been done on the uncertainty quantification of tensor completion. Existing work on the uncertainty quantification of matrix completion (Chen, Fan, Ma & Yan 2019) and tensor completion (Cai, Poor & Chen 2022) typically relies on asymptotic analysis of the estimator by a specific completion algorithm and assumes that data is missing uniformly at random. In this paper, we aim to devise a data-driven approach that does not rely on a specific choice of the completion algorithm nor assume the data is missing uniformly at random, which is more adaptive to real application scenarios.

Conformal prediction (Vovk et al. 2005) is a model-agnostic approach for uncertainty quantification. Recently, Gui et al. (2023) applies the idea of conformal prediction to matrix completion under the assumption that data is missing independently. The method requires one to estimate the missing propensity of each matrix entry and weigh them accordingly to construct well-calibrated confidence regions. In this paper, we generalize this idea to tensor completion. The generalization is non-trivial as one cannot simply reshape the tensor back to a matrix for the conformal prediction without significantly increasing the dimensionality of the nuisance parameter. We keep the tensor structure and leverage low-rank tensor representations for dimension reduction. Furthermore, we do not assume data is missing independently but allow for locally-dependent missingness. We capture such correlatedness of missingness by a novel low-rank tensor Ising model, which could be of independent interest. Finally, we propose a Riemannian gradient descent algorithm (Kressner et al. 2014) for scalable computation, which is necessary since tensor data is typically high-dimensional.

The key insight of the method is that one puts higher weight on the tensor entries with a higher probability of missing, which can be considered as “nearest neighbors” of the missing entries. Such a weighted conformal prediction approach (Tibshirani et al. 2019) is also seen in spatial conformal prediction (Mao et al. 2022) and localized conformal prediction (Guan 2023) where higher weights are put on neighbors in the Euclidean or feature space. However, our method is significantly different in that we estimate the weights by using the entire tensor and determine the weights of all tensor entries altogether while other methods determine the weight of each data locally and thus can be slow under the tensor setting.

The remainder of the paper is organized as follows. We outline the notations used in the paper in Section 1.1. Section 2 describes the conformalized tensor completion (CTC) method and the probabilistic model for the data missingness. Section 3 is dedicated to the

computational algorithm of the CTC. We validate the performance of our proposed CTC using extensive simulations in Section 4 and a real data application to a geophysics dataset in Section 5. Section 6 concludes. The supplemental material contains technical proofs and additional details and results of the simulation and data application.

1.1 Notation

Throughout this paper, we use calligraphic boldface letters (e.g. \mathcal{A}, \mathcal{B}) for tensors with at least three modes, boldface uppercase letters (e.g. \mathbf{X}, \mathbf{Y}) for matrices, boldface lowercase letters (e.g. \mathbf{u}, \mathbf{v}) for vectors and blackboard boldface letters (e.g. \mathbb{S}, \mathbb{T}) for sets. To index a tensor/matrix/vector, we use square brackets with subscripts such as $[\mathcal{A}]_{i_1 \dots i_K}, [\mathbf{X}]_{ij}, [\mathbf{u}]_i$ and will ignore the square brackets when it is clear from the context. For a positive integer n , we denote its index set $\{1, \dots, n\}$ as $[n]$. For a K -mode tensor with size $d_1 \times \dots \times d_K$, we use \mathbb{S} to denote $[d_1] \times \dots \times [d_K]$, namely the indices of all tensor entries, and we often use a single index such as i, j, s instead of a K -tuple to denote elements from \mathbb{S} for notational brevity.

For any tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, we use $\text{vec}(\mathcal{X}), \text{vec}(\mathcal{Y})$ to denote the corresponding vectorized tensors, where all entries are aligned in such an order that the first index changes the fastest. We use $\langle \mathcal{X}, \mathcal{Y} \rangle$ to denote tensor inner product and basically $\langle \mathcal{X}, \mathcal{Y} \rangle = \text{vec}(\mathcal{X})^\top \text{vec}(\mathcal{Y})$. Tensor Frobenius norm $\|\mathcal{X}\|_{\text{F}}$ is defined as $\sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$ and tensor max-norm $\|\mathcal{X}\|_\infty$ is defined as $\max_{s \in \mathbb{S}} |\mathcal{X}_s|$. For any tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ and any matrix $\mathbf{U} \in \mathbb{R}^{J \times d_k}$, the k -th mode tensor-matrix product, denoted as $\mathcal{X} \times_k \mathbf{U}$, is a tensor of size $d_1 \times \dots \times d_{k-1} \times J \times d_{k+1} \times \dots \times d_k$ that satisfies:

$$[\mathcal{X} \times_k \mathbf{U}]_{i_1 \dots i_{k-1} j i_{k+1} \dots i_K} = \sum_{i_k=1}^{d_k} [\mathcal{X}]_{i_1 \dots i_k \dots d_K} [\mathbf{U}]_{j i_k}.$$

More preliminaries on tensor notations and the related algebra will be covered in later sections and we refer our readers to Kolda & Bader (2009) for more references on the

related tensor algebra. In this paper, when referring to a tensor that is a random variable, we add a tilde over the top of the tensor such as $\tilde{\mathcal{W}}, \tilde{\mathcal{X}}$ and use the raw version \mathcal{W}, \mathcal{X} to denote concrete samples. We add an asterisk to the superscript such as $\mathcal{X}^*, \mathcal{B}^*$ to denote the non-random, ground truth parameters.

2 Method

Suppose we have a K -mode random tensor $\tilde{\mathcal{X}}$ of size $d_1 \times \cdots \times d_K$. Further, suppose that one obtains a sample \mathcal{X} for $\tilde{\mathcal{X}}$ with part of the entries in \mathcal{X} missing. To encode the missingness in \mathcal{X} , we define the binary missingness tensor $\mathcal{W} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ and set $\mathcal{W}_s = 1$ when \mathcal{X}_s is observed and $\mathcal{W}_s = -1$ when \mathcal{X}_s is missing. We assume that the missingness \mathcal{W} is a sample of a random binary tensor $\tilde{\mathcal{W}}$ whose likelihood is $p(\cdot)$.

The tensor completion problem (Yuan & Zhang 2016, Xia et al. 2021, Cai, Li, Poor & Chen 2022) deals with estimating the values in \mathcal{X} where $\mathcal{W}_s = -1$, i.e. where data is missing. Although the main framework of our paper does not rely on a specific choice of the tensor completion algorithm, it is beneficial to provide one example here which is also the algorithm we will be using in our numerical experiments and data application.

Since one only has one sample \mathcal{X} of $\tilde{\mathcal{X}}$, estimating the missing values in \mathcal{X} is impossible without imposing additional parsimony over the estimator. Following the literature on tensor completion (Kressner et al. 2014, Xia et al. 2021, Cai, Li & Xia 2022b), we assume that the estimator has a low tensor rank and solve for the estimator by the following constrained least-square problem:

$$\min_{\mathcal{A}: \text{rank}(\mathcal{A}) \leq r} \quad \frac{1}{2} \sum_{s: \mathcal{W}_s = 1} (\mathcal{X}_s - \mathcal{A}_s)^2, \quad (1)$$

where the notion of tensor rank will be introduced later. We denote the minimizer of (1) as $\hat{\mathcal{X}}$. The goal of the paper is to quantify the uncertainty for $\hat{\mathcal{X}}$ by constructing confidence interval $C(\hat{\mathcal{X}})$ around $\hat{\mathcal{X}}$ to cover \mathcal{X} with a pre-specified level of confidence. The framework,

called conformalized tensor completion, will be introduced next.

2.1 Conformalized Tensor Completion (CTC)

Conformal prediction (Vovk et al. 2005) is a model-agnostic, distribution-free approach for predictive uncertainty quantification. To put in the context of the tensor completion problem, we utilize specifically the *split conformal prediction* (Papadopoulos et al. 2002) approach for its simplicity and scalability to complex data structures such as tensor data. We leave the discussion of *full conformal prediction* (Shafer & Vovk 2008) to future work.

Split conformal prediction starts by partitioning all observed entries in \mathcal{X} , whose indices are denoted as \mathbb{S}_{obs} , randomly into a training set \mathbb{S}_{tr} and a calibration set \mathbb{S}_{cal} . One first provides a tensor completion estimator $\widehat{\mathcal{X}}$ using the training set *only*, say by solving for (1) using entries in \mathbb{S}_{tr} . Then one calculates the *non-conformity score* over the calibration set by a score function $\mathcal{S}(\mathcal{X}_s, \widehat{\mathcal{X}}_s)$ such as $\mathcal{S}(\mathcal{X}_s, \widehat{\mathcal{X}}_s) = |\mathcal{X}_s - \widehat{\mathcal{X}}_s|$. To quantify the uncertainty of $\widehat{\mathcal{X}}_{s^*}$ at any missing entry $s^* \in \mathbb{S}_{miss}$, where \mathbb{S}_{miss} includes the indices of all missing entries, the canonical conformal interval at $(1 - \alpha)$ confidence level is constructed as $C_{1-\alpha,s^*}(\widehat{\mathcal{X}}) = \{x \in \mathbb{R} | \mathcal{S}(x, \widehat{\mathcal{X}}_{s^*}) \leq \widehat{q}\}$, with \widehat{q} defined as:

$$\widehat{q} = \mathcal{Q}_{1-\alpha} \left(\frac{1}{|\mathbb{S}_{cal}| + 1} \cdot \sum_{s \in \mathbb{S}_{cal}} \delta_{\mathcal{S}(\mathcal{X}_s, \widehat{\mathcal{X}}_s)} + \frac{1}{|\mathbb{S}_{cal}| + 1} \cdot \delta_{+\infty} \right), \quad (2)$$

where δ_a is a point mass at $x = a$ and $\mathcal{Q}_\tau(\cdot)$ extracts the $(100\tau)^{\text{th}}$ quantile of a CDF. The validity of such a conformal interval $C_{1-\alpha,s^*}(\widehat{\mathcal{X}})$ relies on the assumption of *data exchangeability* (Lei et al. 2018). To put it in the context of tensor completion, we re-label $\mathbb{S}_{cal} \cup \{s^*\}$ as $\{s_1, \dots, s_{n+1}\}$, with $n = |\mathbb{S}_{cal}|$ and $s_{n+1} = s^*$ and define event \mathcal{E}_0 as:

$$\mathcal{E}_0 = \left\{ \widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_{tr} \cup \mathbb{S}_{cal}, \mathbb{S}_{cal} \cup \{s^*\} = \{s_1, \dots, s_{n+1}\} \text{ and } \widetilde{\mathcal{W}}_s = -1 \text{ o.w.} \right\}. \quad (3)$$

Then data exchangeability is equivalent to saying that the probability:

$$P \left[\widetilde{\mathcal{W}}_{s_k} = -1 \text{ and } \widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_k \middle| \mathcal{E}_0 \right]$$

is equal for all $k = 1, \dots, n+1$, where $\mathbb{S}_k = \{s_1, \dots, s_{n+1}\} \setminus \{s_k\}$. Equivalently, this states that conditioning on observing data only from \mathbb{S}_{tr} and n out of $n+1$ entries from $\{s_1, \dots, s_{n+1}\}$, it is equally likely to observe any n entries from $\{s_1, \dots, s_{n+1}\}$. This assumption will hold when data are missing independently with the same probability, a common assumption made in the literature on matrix/tensor completion uncertainty quantification (Chen, Fan, Ma & Yan 2019, Cai, Poor & Chen 2022). However, this assumption might not hold when the data missingness is dependent and is surely violated when the missingness is independent but with heterogeneous probabilities. Therefore, it is necessary to account for more general data missing patterns when conducting uncertainty quantification.

We modify the canonical conformal prediction to accommodate more general data missing patterns by re-weighting each calibration entry using the weighted exchangeability framework (Tibshirani et al. 2019). The result is summarized in Proposition 2.1.

Proposition 2.1. *For any testing entry $s^* \in \mathbb{S}_{miss}$, let $\mathbb{S}_{cal} \cup \{s^*\} = \{s_1, \dots, s_{n+1}\}$ and $\mathbb{S}_k = \{s_1, \dots, s_{n+1}\} \setminus \{s_k\}$, then define p_k as:*

$$p_k = P\left(\widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_{tr} \cup \mathbb{S}_k, \widetilde{\mathcal{W}}_s = -1 \text{ o.w.}\right), \quad (4)$$

for $k = 1, \dots, n+1$. Let $\widehat{\mathcal{X}}$ be the output of any tensor completion method using entries only from \mathbb{S}_{tr} and define \widehat{q}_{s^*} as:

$$\widehat{q}_{s^*} = \mathcal{Q}_{1-\alpha} \left(\sum_{i=1}^n \omega_{s_i} \cdot \delta_{\mathcal{S}(\mathcal{X}_s, \widehat{\mathcal{X}}_s)} + \omega_{s_{n+1}} \cdot \delta_{+\infty} \right), \quad \text{where } \omega_k = \frac{p_k}{\sum_{i=1}^{n+1} p_i}, \quad (5)$$

and construct the $(1-\alpha)$ -level conformal interval as $C_{1-\alpha, s^*}(\widehat{\mathcal{X}}) = \{x \in \mathbb{R} | \mathcal{S}(x, \widehat{\mathcal{X}}_{s^*}) \leq \widehat{q}_{s^*}\}$, then given the definition of \mathcal{E}_0 in (3), we have:

$$P\left(\mathcal{X}_{s^*} \in C_{1-\alpha, s^*}(\widehat{\mathcal{X}}) \mid \mathcal{E}_0\right) \geq 1 - \alpha. \quad (6)$$

We provide the detailed proof in Appendix A.1. Proposition 2.1 indicates that as long

as one can properly weight each calibration entry in proportion to p_k as defined in (4), one can obtain the conditional coverage guarantee in (6). A similar result to Proposition 2.1 has been established for conformalized matrix completion (Gui et al. 2023), where the data is assumed to be missing independently. In our paper, we do not assume independent missingness but provide a more general statement that requires one to weight each calibration and testing entry by directly evaluating the likelihood of $\widetilde{\mathcal{W}}$ under $n + 1$ different missingness, where each time we set 1 out of $n + 1$ entries as missing. In Section 2.2, we will formally introduce the likelihood of the binary tensor $\widetilde{\mathcal{W}}$ that nests the independent missingness as a special case.

2.2 Missing Propensity Model

The key to constructing the conformal interval with coverage guarantee is to properly weight each calibration sample by p_k in (4), which requires the knowledge of the likelihood of $\widetilde{\mathcal{W}}$. In practice, one does not have access to such knowledge but needs to estimate the likelihood of $\widetilde{\mathcal{W}}$, given a single sample \mathcal{W} , and then plug in (4) to get an estimator \hat{p}_k . Previous works (Chen, Fan, Ma & Yan 2019, Cai, Poor & Chen 2022, Gui et al. 2023) assume that all matrix/tensor entries are missing independently, potentially with heterogeneous probabilities. This assumption, however, is not general enough. For example, for spatio-temporal tensors, data might miss together if located close in space or time. As another example, hypergraph adjacency tensor (Ke et al. 2019) may have data missing together if two entries share a group of nodes in the network.

Accounting for the dependencies of binary random variables turns out to be even more challenging in our context because all the binary random variables in $\widetilde{\mathcal{W}}$ are embedded in a tensor grid with ultra-high dimensionality. Fortunately, the Ising model (Cipra 1987) provides one way of modeling dependent binary random variables on a lattice grid. The binary random variables here are the indicators of data missingness instead of atomic spins

in ferromagnetism but a similar idea applies to our modeling context.

To start with, the Ising model prescribes a Boltzmann distribution for $\widetilde{\mathcal{W}}$: $p(\widetilde{\mathcal{W}}) \propto \exp[-\beta \mathcal{H}(\widetilde{\mathcal{W}})]$, where $\beta > 0$ is the inverse temperature parameter and $\mathcal{H}(\widetilde{\mathcal{W}})$ is the *Hamiltonian* of $\widetilde{\mathcal{W}}$, describing the “energy” of $\widetilde{\mathcal{W}}$. In our paper, we extend the richness of this model by augmenting $p(\widetilde{\mathcal{W}})$ with an unknown tensor parameter $\mathcal{B} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ such that:

$$p(\widetilde{\mathcal{W}}|\mathcal{B}) \propto \exp\{-\mathcal{H}(\widetilde{\mathcal{W}}|\mathcal{B})\} \quad (7)$$

$$\mathcal{H}(\widetilde{\mathcal{W}}|\mathcal{B}) = -\frac{1}{2} \sum_{i \sim j} g(\mathcal{B}_i, \mathcal{B}_j) \widetilde{\mathcal{W}}_i \widetilde{\mathcal{W}}_j - \sum_i h(\mathcal{B}_i) \widetilde{\mathcal{W}}_i, \quad (8)$$

where $i, j \in [d_1] \times \dots \times [d_K]$, $g(\cdot, \cdot)$ is a symmetric bi-variate function and $h(\cdot)$ is a univariate function with the inverse temperature parameter β being incorporated into $g(\cdot, \cdot)$ and $h(\cdot)$, and $i \sim j$ means that the two entries indexed by i and j are “neighbors”. For brevity, we often denote $g(\mathcal{B}_i, \mathcal{B}_j)$ as g_{ij} and $h(\mathcal{B}_i) = h_i$ for any i, j . We call (7) together with (8) as the missing propensity model.

One can interpret the unknown parameter \mathcal{B} as a 1-dimensional feature of each tensor entry. Each neighboring pair of entries i and j contribute to the Hamiltonian via their “co-missingness” $\widetilde{\mathcal{W}}_i \widetilde{\mathcal{W}}_j$ and the interaction of their features $\mathcal{B}_i, \mathcal{B}_j$ through $g(\mathcal{B}_i, \mathcal{B}_j)$. The function $g(\cdot, \cdot)$ describes the tendency of neighboring entries to be observed or missing together. Every entry i also contributes individually to the Hamiltonian via h_i , commonly known as the “external magnetic field” when modeling ferromagnetism. In our context, the function $h(\cdot)$ describes the tendency of each entry to be observed or missing. We provide two concrete examples here to provide the interpretation of the model.

Example 2.2 (Independent Bernoulli Model). Suppose that $g(\cdot, \cdot) = 0$, and let $h(x) = 0.5 \cdot \log f(x)/[1 - f(x)]$, where $f(\cdot)$ is an inverse link function (e.g. sigmoid function), then

the missing propensity model indicates that for every $s \in [d_1] \times \cdots \times [d_K]$:

$$\widetilde{\mathbf{W}}_s = \begin{cases} 1, & p = f(\mathbf{B}_s) \\ -1, & p = 1 - f(\mathbf{B}_s), \end{cases} \quad (9)$$

and all $\widetilde{\mathbf{W}}_s$ are independent. This independent Bernoulli model nests the previous works that assume missing uniformly at random as a special case.

Example 2.3 (Ising Model). Suppose that $h(\cdot) = 0$, and let $g(x, y) = xy$. Under this scenario, the conditional distribution of $\widetilde{\mathbf{W}}_s$, given all other entries in $\widetilde{\mathbf{W}}$ as $\widetilde{\mathbf{W}}_{-s}$, is:

$$p(\widetilde{\mathbf{W}}_s = 1 | \mathbf{B}, \widetilde{\mathbf{W}}_{-s}) = \frac{\exp\left[2\mathbf{B}_s \sum_{j \in \mathcal{N}(s)} \widetilde{\mathbf{W}}_j \mathbf{B}_j\right]}{1 + \exp\left[2\mathbf{B}_s \sum_{j \in \mathcal{N}(s)} \widetilde{\mathbf{W}}_j \mathbf{B}_j\right]} = f(\mathbf{B}_s | \sigma_s), \quad (10)$$

where $\mathcal{N}(s) = \{j \in [d_1] \times \cdots \times [d_K] | s \sim j\}$, and $f(x|\sigma) = [1 + \exp(-x/\sigma)]^{-1}$ is the sigmoid function with scale parameter σ . This model is similar to the Bernoulli model in (9) but has entry-specific scale parameter $\sigma_s = (2 \sum_{j \in \mathcal{N}(s)} \widetilde{\mathbf{W}}_j \mathbf{B}_j)^{-1}$ that depends on the missingness and feature of the neighboring entries.

Given the missing propensity model in (7) and (8), we can compute the p_k according to (4) and obtain the conformal weight ω_k as:

$$\omega_k = \frac{p_k}{\sum_{i=1}^{n+1} p_i} = \frac{\exp\left[-2 \sum_{s_j \in \mathcal{N}(s_k)} g(\mathbf{B}_{s_k}, \mathbf{B}_{s_j}) \widetilde{\mathbf{W}}_{s_j} - 2h(\mathbf{B}_{s_k})\right]}{\sum_{i=1}^{n+1} \exp\left[-2 \sum_{s_j \in \mathcal{N}(s_i)} g(\mathbf{B}_{s_i}, \mathbf{B}_{s_j}) \widetilde{\mathbf{W}}_{s_j} - 2h(\mathbf{B}_{s_i})\right]}, \quad (11)$$

with $\widetilde{\mathbf{W}}_s = 1$ for any $s \in \mathbb{S}_{tr} \cup \mathbb{S}_{cal} \cup \{s^*\}$ and $\widetilde{\mathbf{W}}_s = -1$ otherwise. Unfortunately, computing ω_k is still slow in this way because for each $s^* \in \mathbb{S}_{miss}$, we have to temporarily set $\widetilde{\mathbf{W}}_{s^*} = 1$ to compute all the weights. To speed up the computation, we approximate the weight in (11) by using the observed binary tensor \mathbf{W} instead, and thus the only difference is that now we have $\mathbf{W}_{s^*} = -1$ for all $s^* \in \mathbb{S}_{miss}$. This approximation makes very little difference since it will only affect those calibration entries in the neighborhood of s^* and the total number of calibration entries is much larger. In the simulation section, we also demonstrate that this approximation has a very negligible impact on the coverage of the

conformal interval.

With this approximation, the conformal weight ω_k is now proportional to $(1 - \tilde{p}_{s_k})/\tilde{p}_{s_k}$, where $\tilde{p}_s = p(\widetilde{\mathcal{W}}_s = 1 | [\widetilde{\mathcal{W}}]_{s'} = [\mathcal{W}]_{s'}, \forall s' \neq s)$ is the full conditional probability of entry s being observed given all other entries. The only problem remaining is to estimate the tensor parameter \mathbf{B} using a single sample \mathcal{W} , which we will discuss next.

3 Estimating Algorithm

In this section, we discuss the details of estimating \mathbf{B} based on a single binary tensor sample \mathcal{W} drawn from the missing propensity model specified by (7) and (8). More specifically, we attempt to estimate \mathbf{B} using $\mathcal{W}_{\mathbb{S}_{tr}}$, the binary tensor with 1 only in the training set \mathbb{S}_{tr} . We describe the estimation framework in Section 3.1 and the algorithm in Section 3.2.

3.1 Low-rank MPLE Framework

Since we only have access to one sample \mathcal{W} and the tensor parameter \mathbf{B} is of the same dimensionality as \mathcal{W} , it is infeasible to obtain an estimator $\widehat{\mathbf{B}}$ without imposing additional constraints over \mathbf{B} . Similar to previous literature (Wang & Li 2020, Cai, Li & Xia 2022a), we assume that the tensor \mathbf{B} has low tensor rank.

In this paper, we assume that the tensor \mathbf{B} has a low Tensor-Train (TT) rank (Oseledets 2011). A low TT-rank tensor \mathcal{A} can be represented by a series of 3-mode TT factor tensors $\mathcal{T}_k \in \mathbb{R}^{r_{k-1} \times d_k \times r_k}, k = 1, \dots, K, r_0 = r_K = 1$, where for every entry of \mathcal{A} , one has:

$$[\mathcal{A}]_{i_1, \dots, i_K} = [\mathcal{T}_1]_{:i_1:} [\mathcal{T}_2]_{:i_2:} \cdots [\mathcal{T}_K]_{:i_K:}, \quad (12)$$

and the right-hand side is a series of matrix multiplication. We say $\mathbf{r} = (r_1, \dots, r_{K-1})$ is the TT-rank of \mathcal{A} and compactly, we write $\mathcal{A} = [\mathcal{T}_1, \dots, \mathcal{T}_K]$ and $\text{rank}^{\text{tt}}(\mathcal{A}) = \mathbf{r}$. As compared to the more commonly used Tucker rank (Kolda & Bader 2009), the Tensor-Train rank ensures that the number of parameters representing a low-rank tensor scales linearly with

K , the number of modes, making the low TT-rank tensors more efficient for representing high-order tensors.

To ensure the identifiability of TT factors $\mathcal{T}_1, \dots, \mathcal{T}_K$ in (12), it is often required that $\mathcal{T}_1, \dots, \mathcal{T}_{K-1}$ being *left-orthogonal*. A 3-mode tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is left-orthogonal if $\mathbf{L}(\mathcal{T})^\top \mathbf{L}(\mathcal{T}) = \mathbf{I}_{d_3 \times d_3}$, where $\mathbf{L}(\cdot) : \mathbb{R}^{d_1 \times d_2 \times d_3} \mapsto \mathbb{R}^{(d_1 d_2) \times d_3}$ is the so-called left-unfolding operator. Finding the representation (12) of a low TT-rank tensor with left orthogonality constraint can be achieved by the TT-SVD algorithm (Oseledets 2011). For completeness, we restate the TT-SVD algorithm in Algorithm 1 and denote it as $\text{SVD}_\mathbf{r}^{\text{tt}}(\cdot)$.

Algorithm 1 Tensor-Train Singular Value Decomposition (TT-SVD)

Input: Tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, tensor-train rank $\mathbf{r} = (r_1, \dots, r_{K-1})$.
 $\mathcal{A} \leftarrow \mathcal{X}$, $r_0, r_K \leftarrow 1$.
for $k = 1, \dots, K-1$ **do**
 $\mathbf{A} \leftarrow \text{reshape}[\mathcal{A}, (r_{k-1} d_k, d_{k+1} \dots d_K)]$. % $\text{reshape}(\cdot, \cdot)$ from MATLAB
 Conduct SVD on \mathbf{A} and truncate at rank r_k : $\mathbf{A} \approx \mathbf{U} \mathbf{S} \mathbf{V}^\top$.
 $\mathcal{T}_k \leftarrow \text{reshape}[\mathbf{U}, (r_{k-1}, d_k, r_k)]$.
 $\mathcal{A} \leftarrow \mathbf{S} \mathbf{V}^\top$.
end for
 $\mathcal{T}_K \leftarrow \text{reshape}(\mathcal{A}, (r_{K-1}, d_K, 1))$.
Output: Tensor-Train representation $\widehat{\mathcal{X}} = [\mathcal{T}_1, \dots, \mathcal{T}_K]$ with $\text{rank}^{\text{tt}}(\widehat{\mathcal{X}}) \leq \mathbf{r}$.

Given the assumption that the tensor \mathcal{B} has low TT-rank $\mathbf{r} = (r_1, \dots, r_{K-1})$, we can re-formulate the MLE of \mathcal{B} as the solution of a low-rank tensor learning problem:

$$\widehat{\mathcal{B}} = \arg \min_{\mathcal{B}: \text{rank}^{\text{tt}}(\mathcal{B}) \leq \mathbf{r}} -\log p(\widetilde{\mathcal{W}} = \mathcal{W}_{\text{str}} | \mathcal{B}), \quad (13)$$

where $\text{rank}^{\text{tt}}(\mathcal{B}) = (r'_1, \dots, r'_{K-1}) \leq \mathbf{r}$ means that $r'_k \leq r_k$ for any $k = 1, \dots, K-1$.

However, the likelihood in (13) is incorrect since we did not account for the random sampling of the training set, and is also difficult to evaluate its normalizing constant. To circumvent these issues, we consider estimating \mathcal{B} by the maximum pseudo-likelihood estimator (MPLE), which is a common approach for the estimation and inference of Ising model (Ravikumar et al. 2010, Barber & Drton 2015, Bhattacharya & Mukherjee 2018).

Formally, for each entry i , define $\tilde{p}_i(\mathcal{B})$ as:

$$\begin{aligned}\tilde{p}_i(\mathcal{B}) &= p\left(\widetilde{\mathcal{W}}_i = 1 \mid [\widetilde{\mathcal{W}}]_s = [\mathcal{W}_{\mathbb{S}_{tr}}]_s, \forall s \neq i, \mathcal{B}\right) \\ &= \frac{\exp\left[2 \sum_{j \in \mathcal{N}(i)} g(\mathcal{B}_i, \mathcal{B}_j)[\mathcal{W}_{\mathbb{S}_{tr}}]_j + 2h(\mathcal{B}_i)\right]}{1 + \exp\left[2 \sum_{j \in \mathcal{N}(i)} g(\mathcal{B}_i, \mathcal{B}_j)[\mathcal{W}_{\mathbb{S}_{tr}}]_j + 2h(\mathcal{B}_i)\right]}. \end{aligned} \quad (14)$$

and we often write it directly as \tilde{p}_i . The low-rank MPLE of \mathcal{B} can now be written as:

$$\widehat{\mathcal{B}} = \arg \min_{\mathcal{B}: \text{rank}^{\text{tt}}(\mathcal{B}) \leq \mathbf{r}} \ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B}) = - \sum_{i: [\mathcal{W}_{\mathbb{S}_{tr}}]_i = 1} \log q\tilde{p}_i - \sum_{i: [\mathcal{W}_{\mathbb{S}_{tr}}]_i = -1} \log (1 - q\tilde{p}_i), \quad (15)$$

where we $q \in (0, 1)$ is the probability of selecting an observed entry into the training set.

We discuss the optimization algorithm for solving (15) next.

3.2 Riemannian Gradient Descent (RGrad) Algorithm

To solve for (15), a natural idea is to directly estimate the tensor-train factors $\mathcal{T}_1, \dots, \mathcal{T}_K$ for $\widehat{\mathcal{B}}$ one at a time, while keeping the others fixed, and iterate until convergence. Such an alternating minimization algorithm has been applied to low-rank binary tensor decomposition (Wang & Li 2020, Lee & Wang 2020). However, alternating minimization is computationally inefficient here as each step requires fitting a generalized linear model (GLM) with high-dimensional covariates. Another candidate approach for estimating $\widehat{\mathcal{B}}$ is the projected gradient descent (Chen, Raskutti & Yuan 2019), where in each iteration one updates \mathcal{B} along the gradient direction first and then projects it back to the low-rank tensor space with TT-SVD. This is also undesirable since the projection for a high-rank tensor can be very slow.

In this paper, we propose an optimization technique called Riemannian gradient descent (RGrad), motivated by the fact that rank- \mathbf{r} tensor-train tensors lie on a smooth manifold (Holtz et al. 2012), which we denote as $\mathbb{M}_{\mathbf{r}}$. As compared to the aforementioned methods, RGrad is faster because each step updates \mathcal{B} with a gradient along the tangent space of \mathcal{B} , avoiding fitting multiple high-dimensional GLMs. Also, the projection from the

tangent space back to the manifold \mathbb{M}_r is faster than the projected gradient descent since the tensors in the tangent space are also low-rank. RGrad has been extensively applied to tensor completion (Kressner et al. 2014, Steinlechner 2016, Cai, Li & Xia 2022b), generalized tensor learning (Cai, Li & Xia 2022a) and tensor regression (Luo & Zhang 2022). The current work, to the best of our knowledge, is the first to apply RGrad to the low TT-rank binary tensor decomposition.

To summarize the RGrad algorithm, we break down the procedures into three steps.

Step I: Compute Vanilla Gradient. We first compute the vanilla gradient $\nabla \ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B})$ at the current iterative value \mathcal{B} . Formally, the vanilla gradient tensor \mathcal{G} satisfies:

$$[\mathcal{G}]_i = 2 \sum_{j \in \mathcal{N}(i)} (\mathcal{V}_i [\mathcal{W}_{\mathbb{S}_{tr}}]_j + \mathcal{V}_j [\mathcal{W}_{\mathbb{S}_{tr}}]_i) g_x(\mathcal{B}_i, \mathcal{B}_j) + 2h'(\mathcal{B}_i) \mathcal{V}_i, \quad (16)$$

where $g_x(\cdot, \cdot) = \partial g(\cdot, \cdot) / \partial x$ and $\mathcal{V}_i = (1 - \tilde{p}_i)(1 - q\tilde{p}_i)^{-1}(q\tilde{p}_i - \mathbb{1}_{\{\mathcal{W}_{\mathbb{S}_{tr}}\}_i=1})$, with \tilde{p}_i defined in (14).

Step II: Tangent Space Projected Gradient Descent. Suppose that the current iterative value \mathcal{B} has a tensor-train representation $\mathcal{B} = [\mathcal{T}_1, \dots, \mathcal{T}_K]$. Then any tensor \mathcal{A} within the tangent space \mathbb{T} at \mathcal{B} has an explicit form:

$$\mathcal{A} = \sum_{k=1}^K \mathcal{C}_k, \quad \mathcal{C}_k = [\mathcal{T}_1, \dots, \mathcal{T}_{k-1}, \mathcal{Y}_k, \mathcal{T}_{k+1}, \dots, \mathcal{T}_K], \quad (17)$$

with the constraint that $\mathbf{L}(\mathcal{Y}_k)^\top \mathbf{L}(\mathcal{T}_k) = \mathbf{O}_{r_k \times r_k}$ for all $k < K$, where \mathbf{O} is a zero matrix, and \mathcal{C}_k has the property that $\langle \mathcal{C}_i, \mathcal{C}_j \rangle = 0$ for all $i \neq j$. In this step, one projects the vanilla gradient \mathcal{G} from step I onto \mathbb{T} and obtains the projected gradient $\mathcal{P}_{\mathbb{T}}(\mathcal{G})$. Thanks to the orthogonality of different \mathcal{C}_k , the projection problem is solving:

$$\min_{\mathcal{Y}_k: \mathbf{L}(\mathcal{Y}_k)^\top \mathbf{L}(\mathcal{T}_k) = \mathbf{O}_{r_k \times r_k}} \frac{1}{2} \|\mathcal{G} - \mathcal{C}_k\|_F^2, \quad \text{s.t. } \mathcal{C}_k = [\mathcal{T}_1, \dots, \mathcal{T}_{k-1}, \mathcal{Y}_k, \mathcal{T}_{k+1}, \dots, \mathcal{T}_K], \quad (18)$$

for any $k \leq K - 1$ and \mathbf{y}_k is unconstrained if $k = K$. Solution to (18) is:

$$\mathbf{L}(\widehat{\mathbf{y}}_k) = [\mathbf{I}_{r_{k-1}d_k} - \mathbf{L}(\mathcal{T}_k)\mathbf{L}(\mathcal{T}_k)^\top] (\mathcal{B}^{\leq k-1} \otimes \mathbf{I}_{d_k})^\top \mathcal{G}^{<k>} (\mathcal{B}^{\geq k+1})^\top \left[\mathcal{B}^{\geq k+1} (\mathcal{B}^{\geq k+1})^\top \right]^{-1}, \quad (19)$$

for $k \leq K - 1$ and:

$$\mathbf{L}(\widehat{\mathbf{y}}_K) = (\mathcal{B}^{\leq K-1} \otimes \mathbf{I}_{d_K})^\top \mathcal{G}^{<K>}, \quad (20)$$

where \otimes is the matrix Kronecker product. In (19) and (20), $\mathcal{G}^{<k>}$ is the k -mode separation of tensor \mathcal{G} , which basically reshapes \mathcal{G} to a matrix of size $(\prod_{l \leq k} d_l) \times (\prod_{l > k} d_l)$. Any tensor \mathcal{B} has its k -mode separation as $\mathcal{B}^{<k>} = \mathcal{B}^{\leq k} \mathcal{B}^{\geq k+1}$, where $\mathcal{B}^{\leq k}, \mathcal{B}^{\geq k+1}$ are called the k -th left part and $(k+1)$ -th right part. Given that $\mathcal{B} = [\mathcal{T}_1, \dots, \mathcal{T}_K]$, one can recursively compute $\mathcal{B}^{\leq k}$ as $(\mathcal{B}^{\leq k-1} \otimes \mathbf{I}_{d_k})\mathbf{L}(\mathcal{T}_k)$ and $\mathcal{B}^{\geq k+1}$ as $\mathbf{R}(\mathcal{T}_{k+1})(\mathbf{I}_{d_{k+1}} \otimes \mathcal{B}^{\geq k+2})$ with the convention that $\mathcal{B}^{\leq 0} = \mathcal{B}^{\geq K+1} = 1$, where $\mathbf{R}(\cdot) : \mathbb{R}^{d_1 \times d_2 \times d_3} \mapsto \mathbb{R}^{d_1 \times d_2 d_3}$ is the right-unfolding operator.

After computing $\widehat{\mathbf{y}}_k$ with (19) and (20), one ends up with $\widehat{\mathcal{C}}_k = [\mathcal{T}_1, \dots, \widehat{\mathbf{y}}_k, \dots, \mathcal{T}_K]$ and thus the projected gradient $\mathcal{P}_{\mathbb{T}}(\mathcal{G}) = \sum_k \widehat{\mathcal{C}}_k$. This step is completed after one updates \mathcal{B} to $\widetilde{\mathcal{B}} = \mathcal{B} - \eta \mathcal{P}_{\mathbb{T}}(\mathcal{G})$, where η is a constant step size.

Step III: Retraction. As a property of low TT-rank tensors, the updated tensor $\widetilde{\mathcal{B}}$ has its TT-rank upper bounded by $2\mathbf{r}$. To enforce the rank constraint, the last step of RGrad is to retract $\widetilde{\mathcal{B}}$ back to the manifold $\mathbb{M}_{\mathbf{r}}$. We do so by applying TT-SVD to $\widetilde{\mathcal{B}}$: $\mathcal{B}' = \text{SVD}_{\mathbf{r}}^{\text{tt}}(\widetilde{\mathcal{B}})$, and \mathcal{B}' will be the value used for the next iteration.

We summarize the RGrad algorithm in Algorithm 2. To provide an initial estimator of \mathcal{B} , we apply TT-SVD to a randomly perturbed version of the binary tensor $\mathcal{W}_{\mathbb{S}_{tr}}$, which works quite well empirically. We typically set $\eta = 0.1$ and denote the output of Algorithm 2 as $\text{RGrad}(\mathcal{W}_{\mathbb{S}_{tr}}, \mathbf{r})$. By assuming $d_k = O(d), r_k = O(r), \forall k$ and $\max_s |\mathcal{N}(s)| = O(K)$, the computational complexity of RGrad is $O(K(d^K r^2 + dr^3))$ per iteration. See Steinlechner (2016) for more details on the computational complexity of RGrad.

Combining all the discussions in Section 2 and 3, we summarize the conformalized tensor completion (CTC) algorithm in Algorithm 3. We make several remarks for Algorithm 3.

Algorithm 2 MPLE of Low-rank Ising Model with Riemannian Gradient Descent

Input: Binary tensor $\mathcal{W}_{\mathbb{S}_{tr}}$, tensor-train rank $\mathbf{r} = (r_1, \dots, r_{K-1})$, step size η , train-calibration split probability q .

Initialize: let $\mathcal{E} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and $\widehat{\mathcal{B}} \leftarrow \text{SVD}_{\mathbf{r}}^{\text{tt}}(\mathcal{W}_{\mathbb{S}_{tr}} + \mathcal{E}) = [\widehat{\mathcal{T}}_1, \dots, \widehat{\mathcal{T}}_K]$ by Algorithm 1.

for $l = 1, \dots, l_{\max}$ **do**

 Compute the vanilla gradient \mathcal{G} using (16).

for $k = 1, \dots, K$ **do**

 Compute $\widehat{\mathcal{Y}}_k$ following (19) if $k < K$ and (20) if $k = K$.

$\widehat{\mathcal{C}}_k \leftarrow [\widehat{\mathcal{T}}_1, \dots, \widehat{\mathcal{T}}_{k-1}, \widehat{\mathcal{Y}}_k, \widehat{\mathcal{T}}_{k+1}, \dots, \widehat{\mathcal{T}}_K]$.

end for

$\mathcal{P}_{\mathbb{T}}(\mathcal{G}) \leftarrow \sum_{k=1}^K \widehat{\mathcal{C}}_k$.

$\widetilde{\mathcal{B}} \leftarrow \widehat{\mathcal{B}} - \eta \mathcal{P}_{\mathbb{T}}(\mathcal{G})$.

$\widehat{\mathcal{B}} \leftarrow \text{SVD}_{\mathbf{r}}^{\text{tt}}(\widetilde{\mathcal{B}}) = [\widehat{\mathcal{T}}_1, \dots, \widehat{\mathcal{T}}_K]$ by Algorithm 1.

end for

Output: Maximum Pseudo-Likelihood Estimator (MPLE) $\widehat{\mathcal{B}}$ with $\text{rank}^{\text{tt}}(\widehat{\mathcal{B}}) \leq \mathbf{r}$.

Algorithm 3 Conformalized Tensor Completion (CTC)

Input: Data tensor \mathcal{X} , tensor-train rank \mathbf{r} , train-calibration split probability $q \in (0, 1)$, target mis-coverage $\alpha \in (0, 1)$, arbitrary tensor completion algorithm \mathcal{A} .

$\mathbb{S} \leftarrow \{s \in [d_1] \times \dots \times [d_K] | \mathcal{X}_s \neq \text{NaN}\}$. % indices of entries that are observed

$\mathcal{W} \leftarrow 2 \times \mathbb{1}_{\{s \in \mathbb{S}\}} - 1$.

Randomly partition \mathbb{S} independently into $\mathbb{S}_{tr} \cup \mathbb{S}_{cal}$ with probability q and $1 - q$.

$\widehat{\mathcal{X}} \leftarrow \mathcal{A}(\mathcal{X}_{\mathbb{S}_{tr}})$. % $[\mathcal{X}_{\mathbb{S}_{tr}}] = \mathcal{X}_s$ if $s \in \mathbb{S}_{tr}$ and NaN otherwise

$\widehat{\mathcal{B}} \leftarrow \text{RGrad}(\mathcal{W}_{\mathbb{S}_{tr}}, \mathbf{r})$. % $\text{RGrad}(\cdot, \cdot)$ is Algorithm 2

for $s \in \mathbb{S}_{cal} \cup \mathbb{S}^c$ **do**

$\tilde{p}_s \leftarrow \left\{ 1 + \exp \left[-2 \sum_{j \in \mathcal{N}(s)} [\mathcal{W}_{\mathbb{S}_{tr}}]_{sj} g(\widehat{\mathcal{B}}_s, \widehat{\mathcal{B}}_j) - 2h(\widehat{\mathcal{B}}_s) \right] \right\}^{-1}$.

$\omega_s \leftarrow (1 - \tilde{p}_s) \tilde{p}_s^{-1}$.

end for

for $s^* \in \mathbb{S}^c$ **do** % See Remark 3.1

 Re-normalize $\omega_s, s \in \mathbb{S}_{cal}$ and ω_{s^*} s.t. $\sum_{s \in \mathbb{S}_{cal}} \omega_s + \omega_{s^*} = 1$.

$\widehat{q}_{s^*} \leftarrow \mathcal{Q}_{1-\alpha} \left(\sum_{s \in \mathbb{S}_{cal}} \omega_s \cdot \delta_{\mathcal{S}(\mathcal{X}_s, \widehat{\mathcal{X}}_s)} + \omega_{s^*} \cdot \delta_{+\infty} \right)$.

end for

Output: $(1 - \alpha)$ -level conformal interval $C_{1-\alpha, s^*}(\widehat{\mathcal{X}}) \leftarrow \{x \in \mathbb{R} | \mathcal{S}(x, \widehat{\mathcal{X}}_{s^*}) \leq \widehat{q}_{s^*}\}, \forall s^* \in \mathbb{S}^c$.

Remark 3.1 (Fast Entry-wise Quantile Computation). In the last step of Algorithm 3, we compute the empirical $(1 - \alpha)$ -quantile of the weighted eCDF of the non-conformity score of all calibration data. The for-loop looks slow superficially as one needs to evaluate the quantile for each testing entry s^* . However, \hat{q}_{s^*} can be computed faster via:

$$\hat{q}_{s^*} = \begin{cases} +\infty, & \text{if } \omega_{s^*} \geq \alpha \\ \mathcal{Q}_{\frac{1-\alpha}{1-\omega_{s^*}}} \left(\sum_{s \in \mathbb{S}_{cal}} \frac{\omega_s}{1-\omega_{s^*}} \cdot \delta_{\mathcal{S}(\mathbf{x}_s, \hat{\mathbf{x}}_s)} \right), & \text{if } \omega_{s^*} < \alpha. \end{cases}$$

which only requires evaluating the quantile of a fixed eCDF shared by all testing entries.

Remark 3.2 (Rank Selection). The implementation of the CTC algorithm requires a proper choice of the tensor-train rank \mathbf{r} for the low-rank Ising model. Typically in low-rank tensor learning literature (Wang & Li 2020, Cai, Li & Xia 2022a), either the Akaike Information Criterion (AIC) (Akaike 1973) or the Bayesian Information Criterion (BIC) (Schwarz 1978) is used for the rank selection. Unfortunately, they are not applicable here since we can only compute the pseudo-likelihood. According to previous literature on the model selection of Markov Random Fields (Ji & Seymour 1996, Csiszár & Talata 2006, Matsuda et al. 2021), one can replace the likelihood in AIC/BIC with pseudo-likelihood and obtain the Pseudo-AIC (P-AIC) and Pseudo-BIC (P-BIC), which are still consistent under some regularity conditions. The P-AIC and P-BIC are defined as:

$$\text{P-AIC}(r') = 2\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \hat{\mathcal{B}}) + 2 \left\{ \sum_{k=1}^{K-1} [d_k r'_{k-1} r'_k - (r'_k)^2] + d_K r'_{K-1} \right\}. \quad (21)$$

$$\text{P-BIC}(r') = 2\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \hat{\mathcal{B}}) + \left\{ \sum_{k=1}^{K-1} [d_k r'_{k-1} r'_k - (r'_k)^2] + d_K r'_{K-1} \right\} \log \left(\prod_{k=1}^K d_k \right). \quad (22)$$

Among all candidate ranks, we select the rank with the smallest P-AIC or P-BIC. In Section C.2 of the supplemental material, we provide empirical evidence on the consistency of P-AIC and the inconsistency of P-BIC.

Remark 3.3 (Estimation and Coverage Error Bound). In Section A.2 of the supplemental material, we derive theoretically the non-asymptotic bound for $\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F$ under the special

case where $g(x, y) = 0$ (i.e. the Bernoulli model) and with the same assumption we further derive the coverage probability lower bound of the CTC algorithm in Section A.3. It is a remarkable result that the estimating error, as well as the shortfall of the coverage from the target coverage, increases with $(r^* \bar{d}/d^*)^{1/2}$, where r^*, \bar{d}, d^* are $\prod_k r_k, \sum_k d_k, \prod_k d_k$ for \mathcal{B}^* , respectively. If one assumes that $d_k = O(d), r_k = O(r), \forall k$, then the estimation error and coverage shortfall scales with $(r/d)^{(K-1)/2}$. Higher r/d indicates that the data missing pattern is more complex and thus the uncertainty quantification is harder. Although we do not have the theoretical results when $g(x, y) \neq 0$, we show empirically in Section 4 that this tendency also holds for the Ising model.

4 Simulation Experiments

In this section, we validate the effectiveness of the proposed conformalized tensor completion algorithm via numerical simulations. We consider order-3 cubical tensor of size $d \times d \times d$ and summarize our simulation settings below. Additional details about the simulation setups and results are included in Section C of the supplemental material.

4.1 Simulation Setup

We simulate the $d \times d \times d$ true tensor parameter \mathcal{B}^* via the Gaussian tensor block model (TBM) (Wang & Zeng 2019), where $\mathcal{B}^* = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 + \mathcal{E}_1$ with $\mathcal{C} \in \mathbb{R}^{r \times r \times r}$ being a core tensor with i.i.d. entries from a Gaussian mixture model: $0.5 \cdot \mathcal{N}(1, 0.5) + 0.5 \cdot \mathcal{N}(-1, 0.5)$, and $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \in \{0, 1\}^{d \times r}$ with only a single 1 in each row and $\mathcal{E}_1 \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.2)$. We re-scale the simulated \mathcal{B}^* such that $\|\mathcal{B}^*\|_\infty = 2$. We enforce each column of $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ to have 1 in consecutive rows so that the simulated \mathcal{B}^* demonstrates a noisy “checker box” structure, as illustrated in Figure 1(a).

Given the simulated \mathcal{B}^* , we then simulate the binary data missingness tensor \mathcal{W} from the Ising model. Throughout this section, we suppose that two tensor entries i and j are

neighbors, i.e. $i \sim j$, if and only if their indices differ by 1 in just one mode. Consequently, for 3-way tensors, each non-boundary entry has six neighbors. We simulate \mathbf{W} from the missing propensity model specified by (7) and (8) with a block-Gibbs sampler and generate samples from a Monte Carlo Markov Chain (MCMC). The MCMC has 4×10^4 iterations with the first 10^4 samples burnt in and we take one sample every other 10^3 iterations to end up with $n = 30$ samples. In Figure 1(b), we visualize one simulated \mathbf{W} .

Lastly, the data tensor \mathbf{X} is generated from an additive noise model: $\mathbf{X} = \mathbf{X}^* + \mathbf{\mathcal{E}}$, which is similar to \mathbf{B}^* , with \mathbf{X}^* having a Tucker rank $(3, 3, 3)$. The noiseless tensor \mathbf{X}^* also possesses a ‘‘checker box’’ structure and is contaminated by the noise tensor $\mathbf{\mathcal{E}}$, whose distribution depends on the specific simulation setting described later. We re-scale \mathbf{X}^* to have $\|\mathbf{X}^*\|_\infty = 2$ and define the signal-to-noise ratio (SNR) of \mathbf{X} as $\|\mathbf{X}^*\|_\infty/\|\mathbf{\mathcal{E}}\|_\infty$ and re-scale $\mathbf{\mathcal{E}}$ such that SNR= 2. The data tensor \mathbf{X} is then masked by \mathbf{W} , as plotted in Figure 1(c).

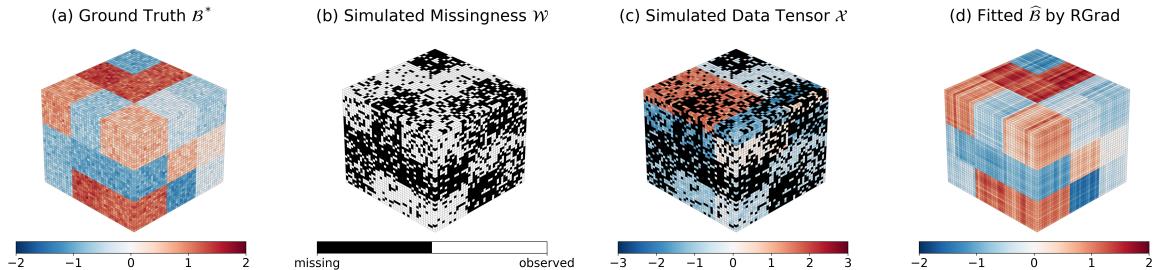


Figure 1: Visualizations of key tensors in the simulation setup. (a) Ising model parameter tensor \mathbf{B}^* with $d = 40, r = 3$. (b) Simulated binary tensor \mathbf{W} with $g(x, y) = xy/15, h(x) = x/2$. (c) Simulated data tensor \mathbf{X} masked by \mathbf{W} with $r_0 = 3, \text{SNR} = 2.0$ and $\mathbf{\mathcal{E}}$ having i.i.d. $\mathcal{N}(0, 1)$ entries. (d) Estimated parameter $\hat{\mathbf{B}}$ from RGrad based on a 70% training set.

4.2 Conformal Prediction Validation

To validate the efficacy of the proposed conformalized tensor completion (CTC) algorithm, we consider the simulation setting with $d \in \{40, 60, 80, 100\}$, $r \in \{3, 5, 7, 9\}$, $g(x, y) \in \{0, xy/15\}$. The noise tensor $\mathbf{\mathcal{E}}$ is simulated based on two different uncertainty regimes: 1) constant noise: $[\mathbf{\mathcal{E}}]_s \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$; 2) adversarial noise: $[\mathbf{\mathcal{E}}]_s$ follows independent Gaussian

distribution $\mathcal{N}(0, \sigma_s^2)$, with $\sigma_s = [2 \exp(\mathcal{B}_s^*) / [1 + \exp(\mathcal{B}_s^*)]]^{-1}$. The adversarial noise simulates cases where the missing entries have higher uncertainty than the observed entries.

For each simulation scenario, we apply the correctly specified CTC algorithm with P-AIC selected rank and call it **RGrad**. As a benchmark, we also consider two other versions of conformal inference: 1) **unweighted**: the unweighted conformal prediction; 2) **oracle**: the weighted conformal prediction with the true tensor parameter \mathcal{B}^* . We conduct simulation over $n = 30$ repetitions, and for each repetition, we randomly split the observed entries into a training and a calibration set with $q = 0.7$ and evaluate the constructed conformal intervals on the missing entries, denoted as \mathbb{S}_{miss} . For the tensor completion algorithm, we choose low Tucker rank tensor completion coupled with Riemannian gradient descent (Cai, Li & Xia 2022a). We use the absolute residual $\mathcal{S}(y, \hat{y}) = |y - \hat{y}|$ as the non-conformity score. To evaluate the conformal intervals, we define the average mis-coverage metric as:

$$\text{Average Mis-coverage \%} = \frac{100}{|\mathbb{Q}|} \sum_{\tau \in \mathbb{Q}} \left| \tau - \frac{1}{|\mathbb{S}_{miss}|} \sum_{s \in \mathbb{S}_{miss}} \mathbb{1}_{\{\mathbf{x}_s \in \widehat{C}_{\tau,s}(\hat{\mathbf{x}})\}} \right|, \quad (23)$$

with $\mathbb{Q} = \{0.80, 0.81, \dots, 0.98, 0.99\}$. We plot the average mis-coverage with $r = 3$ in Figure 2. We also plot the results with $r = 9$ in Section C.3 of the supplemental material.

According to the results, we find that with constant entry-wise uncertainty, even the unweighted conformal intervals perform decently, but still have more mis-coverage than the oracle case. Using our CTC algorithm significantly shrinks the mis-coverage and matches the performance of the oracle case. Under the adversarial noise regime, we observed significant mis-coverage ($> 10\%$) of the unweighted conformal prediction, and using the CTC algorithm provides conformal intervals with $< 1\%$ of mis-coverage, indicating that our method helps in constructing well-calibrated confidence intervals.

The mis-coverage is even worse for the unweighted conformal prediction when missingness is locally dependent based on the Ising model and the CTC algorithm still provides

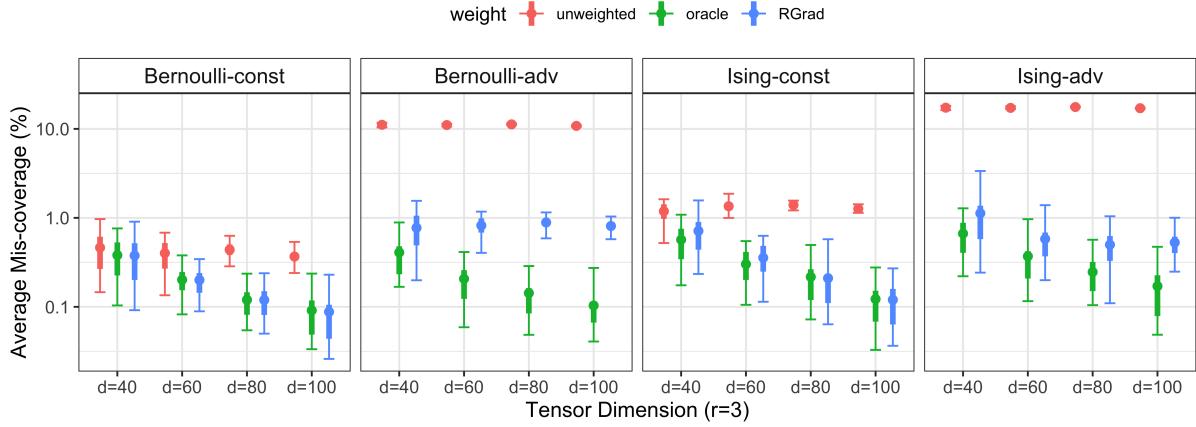


Figure 2: The average mis-coverage of three conformal prediction methods with $d \in \{40, 60, 80, 100\}$, $r = 3$ under the Bernoulli and Ising model. Two uncertainty regimes: constant noise (const) and adversarial noise (adv) are considered. Results are based on 30 repetitions, error bars show the 2.5%, 97.5% quantiles, and the thicker lines show the range of 25% to 75% quantile. The y-axis is plotted in log10-scale.

conformal intervals at the target coverage. In Figure 7 of Section C.3 of the supplemental material, we further show that the mis-coverage of the unweighted conformal prediction is mainly under-coverage as it cannot account for the increase of uncertainty in the testing set under adversarial noise.

To provide a full landscape on how the conformal intervals based on our CTC algorithm perform under different tensor rank r and tensor dimension d of the underlying parameter \mathcal{B}^* , we visualize in Figure 3 the empirical coverage of 90% and 95% conformal intervals under different missingness and uncertainty regimes by r/d , i.e. the rank-over-dimension of the tensor \mathcal{B}^* , based on our RGrad method. Generally speaking, the higher r/d is, the more difficult it is to estimate the missing propensity of the tensor data and thus the worse the coverage of the conformal intervals, which echoes our theoretical result in Section A.3 of the supplemental material. Therefore, we conclude that our proposed method would provide well-calibrated conformal intervals when the underlying missingness model has a low tensor rank relative to the tensor size (i.e. $r \ll d$).

In Section C.3 of the supplemental material, we also compare our RGrad approach with other binary tensor decomposition approaches such as CP and Tucker decomposition

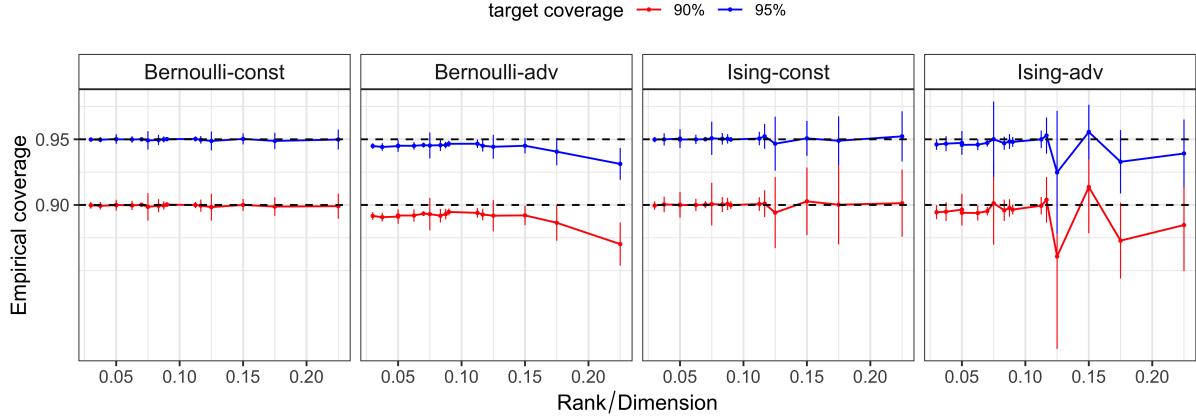


Figure 3: RGrad empirical coverage of the 90% and 95% conformal intervals under the Bernoulli and Ising model with two noise regimes. x-axis is the r/d of the tensor parameter \mathcal{B}^* . Results are based on $n = 30$ repetitions and error bars are ± 1.96 standard deviations.

for estimating the missing propensity and conducting conformal prediction. We find our method to perform consistently well under all kinds of dependency and uncertainty regimes.

5 Data Application to TEC Reconstruction

Our proposed method can account for the locally-dependent data missingness, which is a common data missing pattern for spatial data, therefore we apply our method to a spatio-temporal tensor completion problem in this section as an application. Specifically, we consider the total electron content (TEC) reconstruction problem over the territory of the USA and Canada. The TEC data has severe missing data problems since they can be measured only if the corresponding spatial location has a ground-based receiver. An accurate prediction of the TEC can foretell the impact of space weather on the positioning, navigation, and timing (PNT) service (Wang et al. 2021, Younas et al. 2022). Existing literature (Pan et al. 2021, Sun et al. 2022, Wang et al. 2023) focuses on imputation and prediction of the global and regional TEC and lacks data-driven approaches for quantifying the uncertainty of the imputation and we aim at filling in this gap.

In Figure 4(a), we plot the TEC distribution over the USA and Canada from the VISTA

TEC database (Sun, Chen, Zou, Ren, Chang, Wang & Coster 2023). The VISTA TEC is a pre-imputed version of the Madrigal TEC (MIT Haystack Observatory 2012), which has $> 80\%$ of the data missing globally. We use the VISTA TEC as the ground truth and the Madrigal TEC data missingness to mask out entries in the VISTA TEC to simulate data missingness close to what scientists observe in practice.

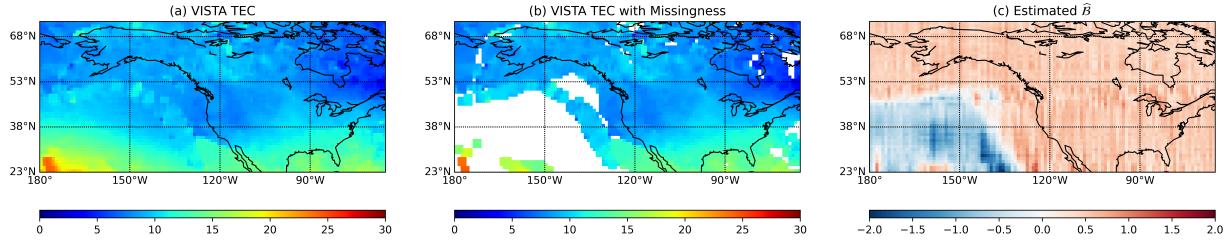


Figure 4: (a) The VISTA TEC at 00:02:30 UT, September 1, 2017. (b) The VISTA TEC in (a) with data missingness from the Madrigal TEC. (c) Fitted $\hat{\beta}$ based on the Ising model.

To set up the experiment, we use the first 20 days of data in September 2017, and each day consists of a tensor of size $50 \times 115 \times 96$. We use the first 5 days as a validation set to search for the best $g(\cdot, \cdot)$ function for the Ising model. For each day, we fit the CTC algorithm with a simple tensor completion algorithm based on (1) with a Tucker rank at $(3, 3, 3)$ and pick the tensor-train rank $\mathbf{r} = (r, r)$ by P-AIC. Based on Figure 3, we know that the Ising model exhibits under-coverage as r/d increases over 0.15, therefore, we select the rank r from $2 \leq r \leq 7$ only. For each day, we consider the Ising model with $g(x, y) = 5xy/4, h(x) = x/2$, the Bernoulli model with $g(x, y) = 0, h(x) = x/2$ and the unweighted conformal prediction for comparison. In Table 1, we report the results on the average mis-coverage % and the empirical coverage of 90% and 95% CI.

| method | mis-coverage % | 90% CI coverage % | 95% CI coverage % |
|------------|----------------|-------------------|-------------------|
| unweighted | 42.1(6.49) | 46.3(6.58) | 52.3(7.23) |
| Bernoulli | 23.1(5.34) | 64.6(5.97) | 76.8(5.03) |
| Ising | 6.01(2.45) | 90.0(6.06) | 94.2(3.74) |

Table 1: Mis-coverage % and empirical coverage of CI at 90% and 95% level for the unweighted conformal prediction and weighted conformal prediction with Bernoulli and Ising model for data during Sept 6 to Sept 20, 2017.

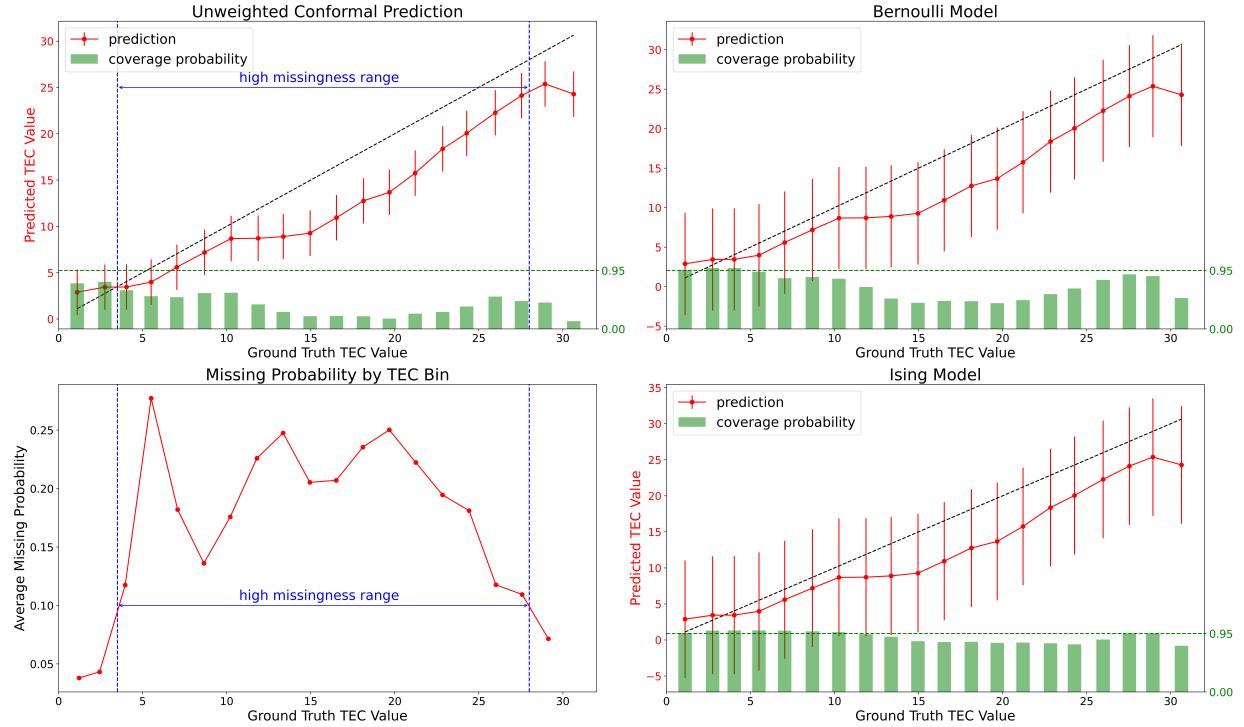


Figure 5: All except the lower-left panels show the average 95% conformal intervals and the empirical coverage for 20 different bins of TEC values on Sept 6, 2017. Each bin spans 1.5 TEC units. The lower-left panel shows the missing probability of different bins. A bin is termed “high missingness” if > 10% of the data is missing.

In Figure 5, we visualize the average 95% CI and its empirical coverage for 20 different bins of TEC values on Sept 6, 2017. It is shown that the data missingness is not uniform across different bins of TEC values and different bins have different distributions of the imputation errors (see how the prediction deviates from the truth), making the unweighted conformal prediction less favorable especially when data missingness is high. These empirical results reveal that by accounting for the heterogeneity and the spatial dependency of data missingness, one can construct well-calibrated confidence intervals using our method.

6 Conclusion

In this paper, we propose a data-driven approach for quantifying the uncertainty of tensor completion. Our method consists of two major steps. We first estimate the missing propensity of each tensor entry using a parameterized Ising model and then plug in the missing

propensity estimator to weight each tensor entry and then construct the confidence region with split conformal prediction. We implement the estimation of missing propensity with a computationally efficient Riemannian gradient descent algorithm and validate the resulting conformal intervals with extensive simulation studies and an application to regional TEC reconstruction. There are two limitations of our method. Firstly, we do not have a systematic approach to determine the best specification of the Ising model. Secondly, our Ising model can only account for locally-dependent missingness but not arbitrary missingness. We leave these topics to future research.

Acknowledgement

We thank Shasha Zou for helpful discussions on the TEC data. YC is supported by NSF DMS 2113397, NSF PHY 2027555, NASA 22-SWXC22_2-0005 and 22-SWXC22_2-0015.

References

- Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in ‘Proc. 2nd International Symposium of Information Theory’, Akademiai Kiado, pp. 267–281.
- Barber, R. F. & Drton, M. (2015), ‘High-dimensional Ising Model Selection with Bayesian Information Criteria’, *Electronic Journal of Statistics* **9**, 567–607.
- Bhattacharya, B. B. & Mukherjee, S. (2018), ‘Inference in Ising Models’, *Bernoulli* **24**(1), 493–525.
- Bi, X., Qu, A. & Shen, X. (2018), ‘Multilayer Tensor Factorization with Applications to Recommender Systems’, *The Annals of Statistics* **46**(6B), 3308 – 3333.
URL: <https://doi.org/10.1214/17-AOS1659>
- Cai, C., Li, G., Poor, H. V. & Chen, Y. (2022), ‘Nonconvex Low-rank Tensor Completion from Noisy Data’, *Operations Research* **70**(2), 1219–1237.
- Cai, C., Poor, H. V. & Chen, Y. (2022), ‘Uncertainty Quantification for Nonconvex Tensor Completion: Confidence intervals, Heteroscedasticity and Optimality’, *IEEE Transactions on Information Theory* **69**(1), 407–452.
- Cai, J.-F., Li, J. & Xia, D. (2022a), ‘Generalized Low-rank plus Sparse Tensor Estimation by Fast Riemannian Optimization’, *Journal of the American Statistical Association* pp. 1–17.
- Cai, J.-F., Li, J. & Xia, D. (2022b), ‘Provable Tensor-Train Format Tensor Completion by Riemannian Optimization’, *Journal of Machine Learning Research* **23**(1), 5365–5441.
- Chen, H., Raskutti, G. & Yuan, M. (2019), ‘Non-convex Projected Gradient Descent for

Generalized Low-rank Tensor Regression', *The Journal of Machine Learning Research* **20**(1), 172–208.

Chen, Y., Fan, J., Ma, C. & Yan, Y. (2019), ‘Inference and Uncertainty Quantification for Noisy Matrix Completion’, *Proceedings of the National Academy of Sciences* **116**(46), 22931–22937.

Cipra, B. A. (1987), ‘An Introduction to the Ising Model’, *The American Mathematical Monthly* **94**(10), 937–959.

Csiszár, I. & Talata, Z. (2006), ‘Consistent Estimation of the Basic Neighborhood of Markov Random Fields’, *The Annals of Statistics* **34**(1), 123–145.

Guan, L. (2023), ‘Localized Conformal Prediction: A Generalized Inference Framework for Conformal Prediction’, *Biometrika* **110**(1), 33–50.

Gui, Y., Barber, R. & Ma, C. (2023), ‘Conformalized Matrix Completion’, *Advances in Neural Information Processing Systems* **36**, 4820–4844.

Holtz, S., Rohwedder, T. & Schneider, R. (2012), ‘On Manifolds of Tensors of Fixed TT-rank’, *Numerische Mathematik* **120**(4), 701–731.

Hong, D., Kolda, T. G. & Duersch, J. A. (2020), ‘Generalized Canonical Polyadic Tensor Decomposition’, *SIAM Review* **62**(1), 133–163.

Ji, C. & Seymour, L. (1996), ‘A Consistent Model Selection Procedure for Markov Random Fields based on Penalized Pseudolikelihood’, *The Annals of Applied Probability* **6**(2), 423–443.

Ke, Z. T., Shi, F. & Xia, D. (2019), ‘Community Detection for Hypergraph Networks via Regularized Tensor Power Iteration’, *arXiv preprint arXiv:1909.06503* .

Kolda, T. G. & Bader, B. W. (2009), ‘Tensor Decompositions and Applications’, *SIAM review* **51**(3), 455–500.

Kressner, D., Steinlechner, M. & Vandeheycken, B. (2014), ‘Low-rank Tensor Completion by Riemannian Optimization’, *BIT Numerical Mathematics* **54**, 447–468.

Lee, C. & Wang, M. (2020), Tensor Denoising and Completion based on Ordinal Observations, in ‘International conference on machine learning’, PMLR, pp. 5778–5788.

Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. (2018), ‘Distribution-free Predictive Inference for Regression’, *Journal of the American Statistical Association* **113**(523), 1094–1111.

Li, X., Xu, D., Zhou, H. & Li, L. (2018), ‘Tucker Tensor Regression and Neuroimaging Analysis’, *Statistics in Biosciences* **10**, 520–545.

Li, Z. & Xiao, H. (2021), ‘Multi-Linear Tensor Autoregressive Models’, *arXiv preprint arXiv:2110.00928*.

Luo, Y. & Zhang, A. R. (2022), ‘Tensor-on-tensor Regression: Riemannian Optimization, Over-parameterization, Statistical-computational gap, and their Interplay’, *arXiv preprint arXiv:2206.08756*.

Mao, H., Martin, R. & Reich, B. J. (2022), ‘Valid Model-Free Spatial Prediction’, *Journal of the American Statistical Association* pp. 1–11.

Matsuda, T., Uehara, M. & Hyvarinen, A. (2021), ‘Information Criteria for Non-normalized Models’, *Journal of Machine Learning Research* **22**(158), 1–33.

MIT Haystack Observatory (2012), ‘Madrigal database’. <http://millstonehill.haystack.mit.edu/>.

Oseledets, I. V. (2011), ‘Tensor-Train Decomposition’, *SIAM Journal on Scientific Computing* **33**(5), 2295–2317.

Pan, Y., Jin, M., Zhang, S. & Deng, Y. (2021), ‘Tec Map Completion through a Deep Learning Model: SNP-GAN’, *Space Weather* **19**(11), e2021SW002810.

Papadopoulos, H., Proedrou, K., Vovk, V. & Gammerman, A. (2002), Inductive Confidence Machines for Regression, in ‘Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13’, Springer, pp. 345–356.

Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. (2010), ‘High-dimensional Ising Model Selection using ℓ_1 -Regularized Logistic Regression’, *The Annals of Statistics* pp. 1287–1319.

Schwarz, G. (1978), ‘Estimating the Dimension of a Model’, *Annals of Statistics* **6**(2), 461–464.

Shafer, G. & Vovk, V. (2008), ‘A Tutorial on Conformal Prediction.’, *Journal of Machine Learning Research* **9**(3).

Steinlechner, M. (2016), ‘Riemannian Optimization for High-Dimensional Tensor Completion’, *SIAM Journal on Scientific Computing* **38**(5), S461–S484.

Sun, H., Chen, Y., Zou, S., Ren, J., Chang, Y., Wang, Z. & Coster, A. (2023), ‘Complete Global Total Electron Content Map Dataset based on a Video Imputation Algorithm VISTA’, *Scientific Data* **10**(1), 236.

Sun, H., Hua, Z., Ren, J., Zou, S., Sun, Y. & Chen, Y. (2022), ‘Matrix completion methods for the total electron content video reconstruction’, *The Annals of Applied Statistics* **16**(3), 1333–1358.

Sun, H., Manchester, W., Jin, M., Liu, Y. & Chen, Y. (2023), Tensor Gaussian Process with Contraction for Multi-Channel Imaging Analysis, *in* ‘Proceedings of the 40th International Conference on Machine Learning’, PMLR, pp. 32913–32935.

Tibshirani, R. J., Foygel Barber, R., Candes, E. & Ramdas, A. (2019), ‘Conformal Prediction under Covariate Shift’, *Advances in Neural Information Processing Systems (Neurips) 32*.

Tomioka, R. & Suzuki, T. (2014), ‘Spectral Norm of Random Tensors’, *arXiv preprint arXiv:1407.1870*.

Vovk, V., Gammerman, A. & Shafer, G. (2005), *Algorithmic Learning in a Random World*, Vol. 29, Springer.

Wang, M. & Li, L. (2020), ‘Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality’, *Journal of Machine Learning Research* **21**(1), 6146–6183.

Wang, M. & Zeng, Y. (2019), ‘Multiway Clustering via Tensor Block Models’, *Advances in neural information processing systems 32*.

Wang, Z., Zou, S., Liu, L., Ren, J. & Aa, E. (2021), ‘Hemispheric Asymmetries in the Mid-latitude Ionosphere During the September 7–8, 2017 Storm: Multi-instrument Observations’, *Journal of Geophysical Research: Space Physics* **126**, e2020JA028829.

Wang, Z., Zou, S., Sun, H. & Chen, Y. (2023), ‘Forecast Global Ionospheric TEC: Apply Modified U-Net on VISTA TEC Data Set’, *Space Weather* **21**(8), e2023SW003494.

Wei, B., Peng, L., Guo, Y., Manatunga, A. & Stevens, J. (2023), ‘Tensor Response Quantile Regression with Neuroimaging Data’, *Biometrics* **79**(3), 1947–1958.

Xia, D., Yuan, M. & Zhang, C.-H. (2021), ‘Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries’, *The Annals of Statistics* **49**(1).

Younas, W., Khan, M., Amory-Mazaudier, C., Amaechi, P. O. & Fleury, R. (2022), ‘Middle and Low Latitudes Hemispheric Asymmetries in $\Sigma O/N2$ and TEC during intense magnetic storms of solar cycle 24’, *Advances in Space Research* **69**, 220–235.

Yuan, M. & Zhang, C.-H. (2016), ‘On Tensor Completion via Nuclear Norm Minimization’, *Foundations of Computational Mathematics* **16**(4), 1031–1068.

SUPPLEMENTARY MATERIAL

This supplemental material contains three sections. Section A contains the proofs for proposition 2.1 and additional theoretical results for the estimation error and coverage guarantee for the Bernoulli model. Section B describes the technical lemmas used in Section A. Section C contains additional details and results of the simulation experiments.

A Proofs of Theorems and Propositions

Throughout this section, for any tensor $\mathcal{B} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, we use \bar{d}, d^* to denote $\sum_k d_k$ and $\prod_k d_k$, respectively. For any tensor-train rank $\mathbf{r} = (r_1, \dots, r_{K-1})$, we use r^* to denote $\prod_k r_k$. We use $c, c', C, C_0, C_1, \dots$ to denote positive absolute constants and $c_K, c'_K, C_K, C_{K,0}, C_{K,1}, \dots$ to denote positive constants that only relate to K . For two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we use $a_n \asymp b_n$ to represent $\lim_{n \rightarrow \infty} a_n/b_n = C > 0$, with C being finite.

A.1 Proof of Proposition 2.1

Proof. Given any testing entry $s^* \in \mathbb{S}_{miss}$, we relabel all elements in $\mathbb{S}_{cal} \cup \{s^*\}$ as $\{s_1, \dots, s_{n+1}\}$.

Now recall the definition of \mathcal{E}_0 as:

$$\mathcal{E}_0 = \left\{ \widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_{tr} \cup \mathbb{S}_{cal}, \mathbb{S}_{cal} \cup \{s^*\} = \{s_1, \dots, s_{n+1}\} \text{ and } \widetilde{\mathcal{W}}_s = -1 \text{ o.w.} \right\},$$

namely one observes data only at \mathbb{S}_{tr} and n out of $n+1$ entries from $\{s_1, \dots, s_{n+1}\}$.

Let V denote the non-conformity score of the testing entry, then the weighted exchangeability framework in Tibshirani et al. (2019) states that one can treat V as a weighted draw from $\{\mathcal{S}(\mathcal{X}_{s_1}, \widehat{\mathcal{X}}_{s_1}), \dots, \mathcal{S}(\mathcal{X}_{s_{n+1}}, \widehat{\mathcal{X}}_{s_{n+1}})\}$, with weight being:

$$P \left[V = \mathcal{S}(\mathcal{X}_{s_k}, \widehat{\mathcal{X}}_{s_k}) \middle| \mathcal{E}_0 \right] = \frac{P \left[\widetilde{\mathcal{W}}_{s_k} = -1, \widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_k \middle| \mathcal{E}_0 \right]}{\sum_{l=1}^{n+1} P \left[\widetilde{\mathcal{W}}_{s_l} = -1, \widetilde{\mathcal{W}}_s = 1 \text{ for } s \in \mathbb{S}_l \middle| \mathcal{E}_0 \right]},$$

where $\mathbb{S}_k = \{s_1, \dots, s_{n+1}\} \setminus \{s_k\}$, for $k = 1, \dots, n+1$. Multiplying both the numerator and the denominator by $P(\mathcal{E}_0)$ leads to the weight in the form of $p_k / \sum_{l=1}^{n+1} p_l$, with p_k defined as (4). The coverage guarantee in (6) is then a direct result of theorem 2 of Tibshirani et al. (2019). \square

A.2 Bernoulli Model Estimation Error Bound

In this subsection, we derive the error bound of the MPLE estimator $\hat{\mathcal{B}}$ under the assumption that $g(x, y) = 0$, i.e. all entries of $\mathcal{W}_{\mathbb{S}_{tr}}$ are observed independently with probability $q(\exp[-2h(\mathcal{B}_s^*)] + 1)^{-1}$. Evidently, under this assumption, the MPLE is identical to MLE since the pseudo-likelihood is also the true Bernoulli likelihood. Our main result is in Theorem A.5. To establish the theoretical result, we make several additional assumptions:

Assumption A.1. $h(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is a non-decreasing, non-constant twice continuously differentiable function with $h''(\cdot) \geq 0$.

Assumption A.2. The MPLE estimator $\hat{\mathcal{B}}$ and the true tensor parameter \mathcal{B}^* have bounded max-norm: $\|\hat{\mathcal{B}}\|_\infty, \|\mathcal{B}^*\|_\infty \leq \xi$.

We define $f(x) = \exp[2h(x)]/(1 + \exp[2h(x)])$ and the following two constants:

$$\alpha_\xi = \sup_{|x| \leq \xi} |2h'(x)|, \quad \gamma_\xi = \inf_{|x| \leq \xi} \min \left\{ \left[\frac{f'(x)}{f(x)} \right]^2 - \frac{f''(x)}{f(x)}, \frac{qf''(x)}{1 - qf(x)} + \left[\frac{qf'(x)}{1 - qf(x)} \right]^2 \right\}.$$

To see what these two constants represent, recall that the negative log-likelihood for $\mathcal{W}_{\mathbb{S}_{tr}}$ given \mathcal{B} can be written as the sum of each entry's negative log-likelihood $\ell_i([\mathcal{W}_{\mathbb{S}_{tr}}]_i | \mathcal{B})$, which is defined as:

$$\ell_i([\mathcal{W}_{\mathbb{S}_{tr}}]_i | \mathcal{B}) = - \left[\left(\frac{[\mathcal{W}_{\mathbb{S}_{tr}}]_i + 1}{2} \right) \log qf(\mathcal{B}_i) + \left(\frac{1 - [\mathcal{W}_{\mathbb{S}_{tr}}]_i}{2} \right) \log(1 - qf(\mathcal{B}_i)) \right].$$

It is not difficult to verify that α_ξ upper bounds $|\partial \ell_i(\cdot | \mathcal{B}) / \partial \mathcal{B}_i|$ and γ_ξ lower bounds $\partial^2 \ell_i(\cdot | \mathcal{B}) / \partial \mathcal{B}_i^2$ for all i as long as $\max_s |\mathcal{B}_s| \leq \xi$. By excluding the trivial case where $h(\cdot)$ is a constant function, α_ξ is strictly positive. If for all $|x| \leq \xi$, we have $1 - (1 - q)f(x) - f^2(x) > 0$,

then we can verify that $\gamma_\xi > 0$ for common choices of $h(\cdot)$, such as the logit model $h(x) = x/2$ or the probit model $h(x) = 2^{-1} \log[\Phi(x)/(1 - \Phi(x))]$. For the remainder of the appendix, we will assume generally that $\gamma_\xi > 0$, which is simply saying that the function $\ell_i(\cdot | \mathcal{B})$ is γ_ξ -strongly convex.

Finally, it is useful to define the tensor spectral norm and the tensor nuclear norm here:

Definition A.3. For a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, its spectral norm, denoted as $\|\mathcal{A}\|_\sigma$, is defined as:

$$\|\mathcal{A}\|_\sigma = \sup_{\mathbf{u}_1, \dots, \mathbf{u}_K} \langle \mathcal{A}, \mathbf{u}_1 \circ \dots \circ \mathbf{u}_K \rangle, \quad \mathbf{u}_k \in \mathbb{S}^{d_k-1}, \forall k,$$

where \circ denotes vector outer product and \mathbb{S}^{d_k-1} is a unit sphere in \mathbb{R}^{d_k} .

Definition A.4. For a tensor $\mathcal{C} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, its nuclear norm $\|\mathcal{C}\|_*$ is defined as:

$$\|\mathcal{C}\|_* = \inf \left\{ \sum_r \lambda_r \middle| \mathcal{C} = \sum_r \lambda_r \mathbf{u}_1 \circ \dots \circ \mathbf{u}_K, \mathbf{u}_k \in \mathbb{S}^{d_k-1}, \forall k \right\}.$$

With the aforementioned assumptions and notations, we have the following non-asymptotic bound on $\|\widehat{\mathcal{B}} - \mathcal{B}^*\|_{\text{F}}$:

Theorem A.5. Assume that $g(x, y) = 0$ and assumption A.1 and A.2 hold, and further assume that $\widehat{\mathcal{B}}$ reaches the global minimum of the negative log-likelihood $\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B})$ and the entry-wise negative log-likelihood is γ_ξ -strongly convex with $\gamma_\xi > 0$, then:

$$P \left(\frac{1}{\sqrt{d^*}} \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_{\text{F}} \leq 2C_{K,1} \frac{\alpha_\xi}{\gamma_\xi} \sqrt{\frac{r^* \bar{d}}{d^*}} \right) \geq 1 - \exp(-C_1 \bar{d} \log K), \quad (24)$$

where $C_1, C_{K,1}$ are some positive constants.

Proof. Using Taylor expansion upon $\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \widehat{\mathcal{B}})$ at $\mathcal{B} = \mathcal{B}^*$ yields:

$$\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \widehat{\mathcal{B}}) = \ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B}^*) + \left\langle \nabla \ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B}^*), \widehat{\mathcal{B}} - \mathcal{B}^* \right\rangle + \frac{1}{2} \text{vec}(\widehat{\mathcal{B}} - \mathcal{B}^*)^\top \mathbf{H}(\check{\mathcal{B}}) \text{vec}(\widehat{\mathcal{B}} - \mathcal{B}^*), \quad (25)$$

where $\check{\mathcal{B}}$ is a convex combination of $\widehat{\mathcal{B}}$ and \mathcal{B}^* . Since, by assumption, $\widehat{\mathcal{B}}$ reaches the global minimum of $\ell(\mathcal{W}_{\mathbb{S}_{tr}} | \mathcal{B})$, or $\ell(\mathcal{B})$ in short, we have $\ell(\widehat{\mathcal{B}}) \leq \ell(\mathcal{B}^*)$, and thus the sum of the

last two terms in (25) are no greater than zero.

For the first term, let $\mathcal{G}^* = \nabla \ell(\mathbf{B}^*)$ and \mathcal{G}^* satisfies:

$$[\mathcal{G}^*]_s = -[1 - f(\mathbf{B}_s^*)] \cdot 2h'(x) \cdot \mathbb{1}_{\{[\mathbf{W}_{\mathbb{S}_{tr}}]_s=1\}} + \frac{qf(\mathbf{B}_s^*)[1 - f(\mathbf{B}_s^*)]}{1 - qf(\mathbf{B}_s^*)} \cdot 2h'(x) \cdot \mathbb{1}_{\{[\mathbf{W}_{\mathbb{S}_{tr}}]_s=-1\}}, \quad (26)$$

and it is easy to verify that $E[[\mathcal{G}^*]_s] = 0$ and $\|\mathcal{G}^*\|_\infty \leq \alpha_\xi$. By Lemma B.1, we can lower bound the first term as:

$$\langle \nabla \ell(\mathbf{W}_{\mathbb{S}_{tr}} | \mathbf{B}^*), \widehat{\mathbf{B}} - \mathbf{B}^* \rangle \geq -\|\mathcal{G}^*\|_\sigma \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_*. \quad (27)$$

By Lemma B.2, we have $\text{rank}^{\text{tt}}(\widehat{\mathbf{B}} - \mathbf{B}^*) \leq 2r$, and then by Lemma B.3, we have $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_* \leq \sqrt{(2r_1) \cdots (2r_{K-1})} \cdot \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F$. Therefore, to lower bound the RHS of (27), we only need to upper bound the spectral norm of \mathcal{G}^* . Since entry-wisely, \mathcal{G}^* is mean-zero and bounded by α_ξ (therefore the sub-Gaussian norm is α_ξ), we can apply Lemma B.4 and get:

$$P\left(\|\mathcal{G}^*\|_\sigma \leq \sqrt{8\alpha_\xi^2 \left[\bar{d} \log 5K + \log \frac{2}{\delta}\right]}\right) \geq 1 - \delta. \quad (28)$$

By setting $\delta = \exp(-C_1 \bar{d} \log K)$, with C_1 be some absolute constant, we can simplify (28) as:

$$P\left(\|\mathcal{G}^*\|_\sigma \leq C_K \alpha_\xi \sqrt{\bar{d}}\right) \geq 1 - \exp(-C_1 \bar{d} \log K), \quad (29)$$

with $C_K = \sqrt{8(\log 5K + C_1 \log K + 1)}$.

Combining these results, we can lower bound the RHS of (27) by:

$$-\|\mathcal{G}^*\|_\sigma \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_* \geq -C_{K,1} \alpha_\xi \sqrt{\bar{d}} \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F, \quad (30)$$

with probability at least $1 - \exp(-C_1 \bar{d} \log K)$, where $C_{K,1} = 2^{(K-1)/2} C_K$.

For the quadratic form in (25), we have:

$$\frac{1}{2} \text{vec}(\widehat{\mathbf{B}} - \mathbf{B}^*)^\top \mathbf{H}(\check{\mathbf{B}}) \text{vec}(\widehat{\mathbf{B}} - \mathbf{B}^*) \geq \frac{\gamma_\xi}{2} \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F^2 > 0. \quad (31)$$

Combining (30) and (31), we obtain:

$$\mathrm{P} \left(\frac{1}{\sqrt{d^*}} \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_{\mathrm{F}} \leq 2C_{K,1} \frac{\alpha_\xi}{\gamma_\xi} \sqrt{\frac{r^* \bar{d}}{d^*}} \right) \geq 1 - \exp(-C_1 \bar{d} \log K),$$

which completes the proof. \square

Remark A.6. Under the scenario where $d_1 \asymp \dots \asymp d_K \asymp O(d)$ and $r_1 \asymp \dots \asymp r_{K-1} \asymp O(r)$, the result in (24) can be reduced to:

$$\mathrm{P} \left(\frac{1}{\sqrt{d^*}} \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_{\mathrm{F}} \leq 2C_K \frac{\alpha_\xi}{\gamma_\xi} \sqrt{\left(\frac{r}{d}\right)^{K-1}} \right) \geq 1 - \exp(-C_1 \bar{d} \log K).$$

So the estimating error can scale with r/d , where lower r/d generally poses an easier binary tensor decomposition problem with lower rooted mean-squared error.

A.3 Bernoulli Model Conformal Inference Coverage Guarantee

In this subsection, we utilize the theoretical result in Theorem A.5 and derive the coverage probability lower bound of the CTC algorithm under the Bernoulli model. The result will reveal how the estimating error of \mathcal{B}^* propagates into the mis-coverage rate. To begin with, we state an essential lemma, which is a trivial extension of Theorem 3.2 of Gui et al. (2023) under the conformalized matrix completion context:

Lemma A.7 (Theorem 3.2 of Gui et al. (2023)). *Let $\widehat{\mathcal{X}}$ be the output of any tensor completion algorithm, and $\widehat{\mathcal{B}}$ be the output of the RGrad algorithm and both $\widehat{\mathcal{X}}, \widehat{\mathcal{B}}$ are based on \mathbb{S}_{tr} only, then given that $g(x, y) = 0$, we have:*

$$\mathrm{E} \left[\frac{1}{|\mathbb{S}_{miss}|} \sum_{s \in \mathbb{S}_{miss}} \mathbb{1}_{\{\mathcal{X}_s \in \widehat{C}_{1-\alpha,s}(\widehat{\mathcal{X}})\}} \right] \geq 1 - \alpha - \mathrm{E}[\Delta], \quad (32)$$

where $\widehat{C}_{1-\alpha,s}(\widehat{\mathcal{X}})$ is the conformal interval for testing entry s at $(1 - \alpha)$ level by the CTC algorithm and Δ is defined as:

$$\Delta = \frac{1}{2} \sum_{s \in \mathbb{S}_{cal} \cup \{s^*\}} \left| \frac{\exp[-2h(\widehat{\mathcal{B}}_s)]}{\sum_{s \in \mathbb{S}_{cal} \cup \{s^*\}} \exp[-2h(\widehat{\mathcal{B}}_s)]} - \frac{\exp[-2h(\mathcal{B}_s^*)]}{\sum_{s \in \mathbb{S}_{cal} \cup \{s^*\}} \exp[-2h(\mathcal{B}_s^*)]} \right|. \quad (33)$$

We neglect the proof here since the generalization from matrix to tensor setting is trivial as one can matricize the tensor into a matrix and the result holds automatically. By Lemma A.1 in Gui et al. (2023), one can further upper bound Δ by:

$$\Delta \leq \frac{\|\exp[-2h(\widehat{\mathcal{B}})] - \exp[-2h(\mathcal{B}^*)]\|_1}{\sum_{s \in \mathbb{S}_{cal}} \exp[-2h(\widehat{\mathcal{B}}_s)]}, \quad (34)$$

where the $h(\cdot)$ is applied to tensors element-wisely and $\|\cdot\|_1$ is the element-wise tensor ℓ_1 norm. The quantity Δ is trivially bounded by 1 as it is the total-variation (TV) distance between two CDFs of discrete random variables. With this lemma, we now formally state our main result:

Theorem A.8. *Assume that the same assumptions hold as Theorem A.5 and further denote $l_\xi = \inf_{|x| \leq \xi} \exp[-2h(x)]$, $u_\xi = \sup_{|x| \leq \xi} \exp[-2h(x)]$. The $(1 - \alpha)$ -level conformal interval $\widehat{C}_{1-\alpha,s}(\widehat{\mathcal{X}})$ satisfies:*

$$\begin{aligned} \mathbb{E} \left[\frac{1}{|\mathbb{S}_{miss}|} \sum_{s \in \mathbb{S}_{miss}} \mathbb{1}_{\{\mathcal{X}_s \in \widehat{C}_{1-\alpha,s}(\widehat{\mathcal{X}})\}} \right] &\geq 1 - \alpha - \frac{2C_{K,1}c_\xi}{(1-c)(1-q)} \sqrt{\frac{r^* \bar{d}}{d^*}} \\ &\quad - \exp[-C_1 \bar{d} \log K] - \exp \left[-\frac{c^2(1-q)d^* l_\xi}{2} \right], \end{aligned} \quad (35)$$

for any $0 < c < 1$, where q is the train-calibration split probability in the CTC algorithm and $c_\xi = u_\xi \alpha_\xi^2 / (\gamma_\xi l_\xi^2)$.

Proof. Given Lemma A.7, the coverage guarantee can be derived if one can characterize an upper bound for $\mathbb{E}[\Delta]$. To upper bound Δ , we start from (34) and bound the numerator on the RHS of (34) as:

$$\begin{aligned} \|\exp[-2h(\widehat{\mathcal{B}})] - \exp[-2h(\mathcal{B}^*)]\|_1 &\leq \sup_{|x| \leq \xi} |\exp[-2h(x)] \cdot 2h'(x)| \cdot \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_1 \\ &\leq u_\xi \alpha_\xi \cdot \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_1 \leq u_\xi \alpha_\xi \cdot \sqrt{d^*} \|\widehat{\mathcal{B}} - \mathcal{B}^*\|_{\text{F}}. \end{aligned} \quad (36)$$

Then we can apply the result of Theorem A.5 to further bound (36) with high probability.

For the denominator of the RHS of (34), we can lower bound it first as $n_{cal} l_\xi$, and for

n_{cal} , since each tensor entry can become a calibration point independently with probability $\exp[-2h(\mathcal{B}_s)](1 - q)$, where $0 < q < 1$ is the train-calibration set split probability, we can then apply the Chernoff bound and obtain:

$$P(n_{cal} \leq (1 - c)(1 - q)\|\exp[-2h(\mathcal{B}^*)]\|_1) \leq \exp\left[-\frac{c^2(1 - q)\|\exp[-2h(\mathcal{B}^*)]\|_1}{2}\right], \quad (37)$$

for any $0 < c < 1$. By denoting the event $\{n_{cal} \geq (1 - c)(1 - q)\|\exp[-2h(\mathcal{B}^*)]\|_1\}$ as \mathcal{E}_0 and the event in (24) as \mathcal{E}_1 and noticing that $\|\exp[-2h(\mathcal{B}^*)]\|_1 \geq d^*l_\xi$, then we have:

$$P\left(\Delta \leq \frac{2C_{K,1}}{(1 - c)(1 - q)} \cdot \frac{u_\xi\alpha_\xi^2}{\gamma_\xi l_\xi^2} \cdot \sqrt{\frac{r^*\bar{d}}{d^*}}\right) \geq 1 - \exp[-C_1\bar{d}\log K] - \exp\left[-\frac{c^2(1 - q)d^*l_\xi}{2}\right],$$

where the probability is the lower bound of the probability of the event $\mathcal{E}_0 \cap \mathcal{E}_1$. With this tail bound on Δ , one can upper bound $E[\Delta]$ as:

$$E[\Delta] \leq \frac{2C_{K,1}}{(1 - c)(1 - q)} \cdot \frac{u_\xi\alpha_\xi^2}{\gamma_\xi l_\xi^2} \cdot \sqrt{\frac{r^*\bar{d}}{d^*}} + \exp[-C_1\bar{d}\log K] + \exp\left[-\frac{c^2(1 - q)d^*l_\xi}{2}\right], \quad (38)$$

and thereby completes the proof. \square

Remark A.9. Under the scenario where $d_1 \asymp \dots \asymp d_K \asymp O(d)$ and $r_1 \asymp \dots \asymp r_{K-1} \asymp O(r)$, the coverage shortfall in (35), i.e. the difference between the lower bound in (35) and $(1 - \alpha)$, can be simplified into:

$$\frac{c_{K,\xi}}{(1 - c)(1 - q)} \cdot \sqrt{\left(\frac{r}{d}\right)^{K-1}} + \exp[-c_K d] + \exp[-c'_{K,\xi}c^2(1 - q)d^K],$$

where $c_{K,\xi}, c'_{K,\xi}$ are positive constants that only relate to K and ξ . The first term is of polynomial order with respect to r/d while the other two terms are of exponential order with respect to d , therefore the first term is the dominating term and the under-coverage of the conformal intervals scale primarily with r/d .

B Technical Lemmas

All technical lemmas listed in this section are cited from existing works. Therefore, we omit the proof here and refer our readers to the corresponding papers cited.

Lemma B.1 (Lemma 1 of Wang & Li (2020)). *For two tensors $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, their inner product $\langle \mathbf{A}, \mathbf{B} \rangle$ can be bounded as:*

$$|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_\sigma \|\mathbf{B}\|_*,$$

where $\|\cdot\|_\sigma, \|\cdot\|_*$ are the tensor spectral norm and the tensor nuclear norm, respectively.

Lemma B.2 (Lemma 24 of Cai, Li & Xia (2022b)). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be two low tensor-train rank tensors with $\text{rank}^{tt}(\mathbf{A}) = \mathbf{r}_1, \text{rank}^{tt}(\mathbf{B}) \leq \mathbf{r}_2$, respectively. Then one has:*

$$\text{rank}^{tt}(\mathbf{A} + \mathbf{B}) \leq \mathbf{r}_1 + \mathbf{r}_2.$$

Lemma B.3 (Lemma 25 of Cai, Li & Xia (2022b)). *Let $\mathbf{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be a low tensor-train rank tensor with $\text{rank}^{tt}(\mathbf{A}) = \mathbf{r} = (r_1, \dots, r_{K-1})$ and has a left-orthogonal representation $\mathbf{A} = [\mathcal{T}_1, \dots, \mathcal{T}_K]$, then:*

$$\|\mathbf{A}\|_* \leq \sqrt{r_1 \cdots r_{K-1}} \cdot \|\mathbf{A}\|_{\text{F}}.$$

Lemma B.4 (Theorem 1 of Tomioka & Suzuki (2014)). *For a random tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ with mean-zero and independent sub-Gaussian entries with sub-Gaussian norm σ , its spectral norm satisfies:*

$$\|\mathbf{A}\|_\sigma \leq \sqrt{8\sigma^2 \left[\bar{d} \log 5K + \log \frac{2}{\delta} \right]},$$

with probability at least $1 - \delta$.

C Appendix for Simulation

C.1 Details of Simulation Setup

We summarize the data-generating model of all essential tensors involved in the simulation experiment in Table 2.

| Tensor | Generating Model | Additional Details |
|-----------------|---|---|
| \mathcal{B}^* | $\mathcal{B}^* = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ $\mathcal{C} \in \mathbb{R}^{r \times r \times r}, \mathbf{U}_i \in \mathbb{R}^{d_i \times r_i}$ | $\mathcal{C} \stackrel{i.i.d.}{\sim} 0.5 \cdot \mathcal{N}(-1, 0.5) + 0.5 \cdot \mathcal{N}(1, 0.5)$ $\mathbf{U}_i = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{bmatrix}^\top$ and each row of \mathbf{U}_i has $\lceil d_i/r_i \rceil$ ones. |
| \mathcal{W} | $p(\mathcal{W}) \propto \exp[-\mathcal{H}(\mathcal{W} \mathcal{B}^*)]$ based on (7) and (8) | simulate by block-Gibbs MCMC, where in each proposal we first sample $\mathbb{I}_1 = \{(i_1, \dots, i_K) \sum_k i_k \text{ is odd}\}$ then \mathbb{I}_1^c . Each block is a Bernoulli model. |
| \mathcal{X} | $\mathcal{X} = \mathcal{X}^* + \mathcal{E}$ | \mathcal{X} is then masked by \mathcal{W} . |
| \mathcal{X}^* | $\mathcal{X}^* = \mathcal{C}^* \times_1 \mathbf{U}_1^* \times_2 \mathbf{U}_2^* \times_3 \mathbf{U}_3^*$ $\mathcal{C}^* \in \mathbb{R}^{3 \times 3 \times 3}, \mathbf{U}_i^* \in \mathbb{R}^{d_i \times 3}$ | $\mathcal{C}^*, \mathbf{U}_1^*, \mathbf{U}_2^*, \mathbf{U}_3^* \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ |
| \mathcal{E} | $[\mathcal{E}]_s \stackrel{\text{independent}}{\sim} \mathcal{N}(0, \sigma_s^2)$ | $\sigma_s = \begin{cases} 1 & \text{constant noise} \\ 0.5 [1 + \exp(-\mathcal{B}_s^*)] & \text{adversarial noise} \end{cases}$ |

Table 2: Details of the tensors generated in the simulation experiment.

C.2 Results on the Missing Propensity Estimation Error

We examine here the effectiveness of the RGrad algorithm for recovering the tensor parameter \mathcal{B}^* from a single observation \mathcal{W} . We consider $d \in \{40, 60, 80, 100\}$ and $r \in \{3, 5, 7, 9\}$ when simulating \mathcal{B}^* . For simulating \mathcal{W} using the Ising model, we fix $h(x) = x/2$ and consider either $g(x, y) \in \{0, xy/15\}$, where we term the case with $g = 0$ as the (independent) Bernoulli model and the case with $g(x, y) = xy/15$ as the (product) Ising model. We split the training and calibration set randomly based on a 70% – 30% ratio.

Under each combination of the choices of (d, r, g) , we generate $n = 30$ repetitions from a single chain of MCMC and fit RGrad to each repetition with the correctly specified $g(\cdot, \cdot)$

and a working rank $r' \in \{2, 3, \dots, 15\}$. In Table 3, we present the average rank selected by the P-AIC and P-BIC under the Bernoulli and Ising model with various (d, r) combinations.

| Bernoulli Model ($g(x, y) = 0$) | | | | | | | | |
|-----------------------------------|-----------------|-----------------|------------|------------|----------|------------|------------|------------|
| | P-AIC | | | | P-BIC | | | |
| rank | $d = 40$ | $d = 60$ | $d = 80$ | $d = 100$ | $d = 40$ | $d = 60$ | $d = 80$ | $d = 100$ |
| $r = 3$ | 3.0 | 3.0 | 3.0 | 3.0 | 2.0 | 2.0 | 3.0 | 3.0 |
| $r = 5$ | 5.0 | 5.0 | 5.0 | 5.0 | 2.0 | 2.1(0.3) | 4.0 | 5.0 |
| $r = 7$ | 6.2(0.4) | 7.0 | 7.0 | 7.0 | 2.0 | 2.0 | 2.0 | 2.3(0.4) |
| $r = 9$ | 6.0(0.8) | 8.8(0.4) | 9.0 | 9.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| Ising Model ($g(x, y) = xy/15$) | | | | | | | | |
| | P-AIC | | | | P-BIC | | | |
| rank | $d = 40$ | $d = 60$ | $d = 80$ | $d = 100$ | $d = 40$ | $d = 60$ | $d = 80$ | $d = 100$ |
| $r = 3$ | 3.4(2.0) | 3.0 | 3.0 | 3.0 | 2.0 | 3.0 | 3.0 | 3.0 |
| $r = 5$ | 7.7(4.1) | 5.0 | 5.0 | 5.0 | 2.0 | 4.0 | 5.0 | 5.0 |
| $r = 7$ | 13.9(0.2) | 7.0 | 7.0 | 7.0 | 2.0 | 2.1(0.2) | 4.0(0.2) | 4.7(0.4) |
| $r = 9$ | 13.9(0.2) | 9.0 | 9.0 | 9.0 | 2.0 | 2.0 | 2.0 | 3.9(0.3) |

Table 3: Model selection result of the Bernoulli model and Ising model. Each number is the mean rank selected by P-AIC/P-BIC with $n = 30$ repetitions followed by its standard deviations, if non-zero. Boldface are the cases where the true rank is within 1.96 standard deviations of the average rank.

Based on these numerical results, we find that the consistency of P-AIC and P-BIC depends on r/d , or the “low-rankness” \mathcal{B}^* . For tensors with high d and low r , both P-AIC and P-BIC are consistent, and the inconsistency emerges as r/d becomes larger. Generally speaking, P-AIC is more robust than P-BIC and is consistent across most of the simulation scenarios except for two cases with small tensor sizes. We therefore suggest using P-AIC for rank selection.

We then evaluate the fitted $\widehat{\mathcal{B}}$ with relative squared error (RSE) defined as: $\|\widehat{\mathcal{B}} - \mathcal{B}^*\|_F/\|\mathcal{B}^*\|_F$. The results, as plotted in Figure 6, exhibit a tendency that as r/d becomes larger, so does the RSE, which echoes the results of the model selection. Additionally, the estimation error is lower for the Ising model, as compared to the Bernoulli model, given the same r and d . We interpret this result as that the Ising model estimator can leverage the additional information from neighbors to infer the missing propensity of each tensor entry.

In Figure 1(d) of the main paper, we plot the estimator for \mathcal{B}^* shown in 1(a) by RGrad based on a randomly chosen 70% training set and it is clear that $\hat{\mathcal{B}}$ mimics \mathcal{B}^* very well.

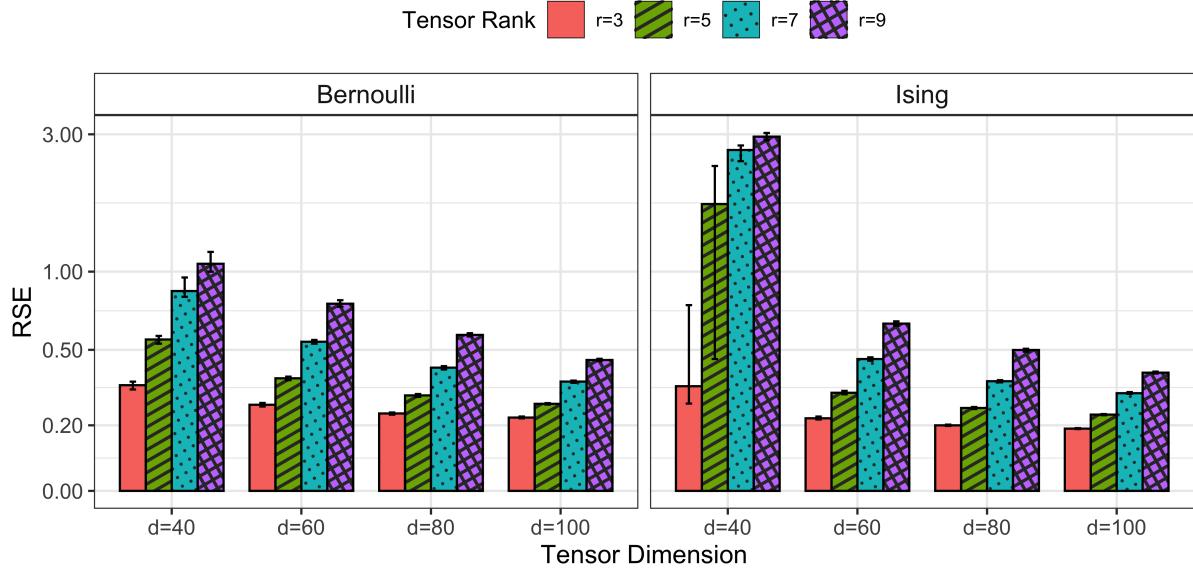


Figure 6: Relative square error of the MPLE $\hat{\mathcal{B}}$ under the Bernoulli (left) and Ising model (right). The results are based on $n = 30$ repetitions with the working rank of each sample determined by P-AIC and each model is fitted by a randomly chosen 70% training set. Error bars show the 2.5% and 97.5% quantiles.

C.3 Results on Conformal Prediction Validation

As a companion result of Figure 2, we plot the empirical coverage and half of the average confidence interval width of three conformal prediction methods under different simulation scenarios in Figure 7. The mis-coverage of the unweighted conformal prediction comes from under-coverage and is associated with shorter confidence intervals. The reason why unweighted conformal prediction has under-coverage is that under the adversarial noise setting, entries with higher missing propensity also have higher uncertainty, and using a uniform weight underestimates the uncertainty of a missing entry. As one can tell from Figure 7, our CTC algorithm matches the oracle case quite well and provides well-calibrated confidence intervals.

Apart from these results, we also compared other binary tensor decomposition methods

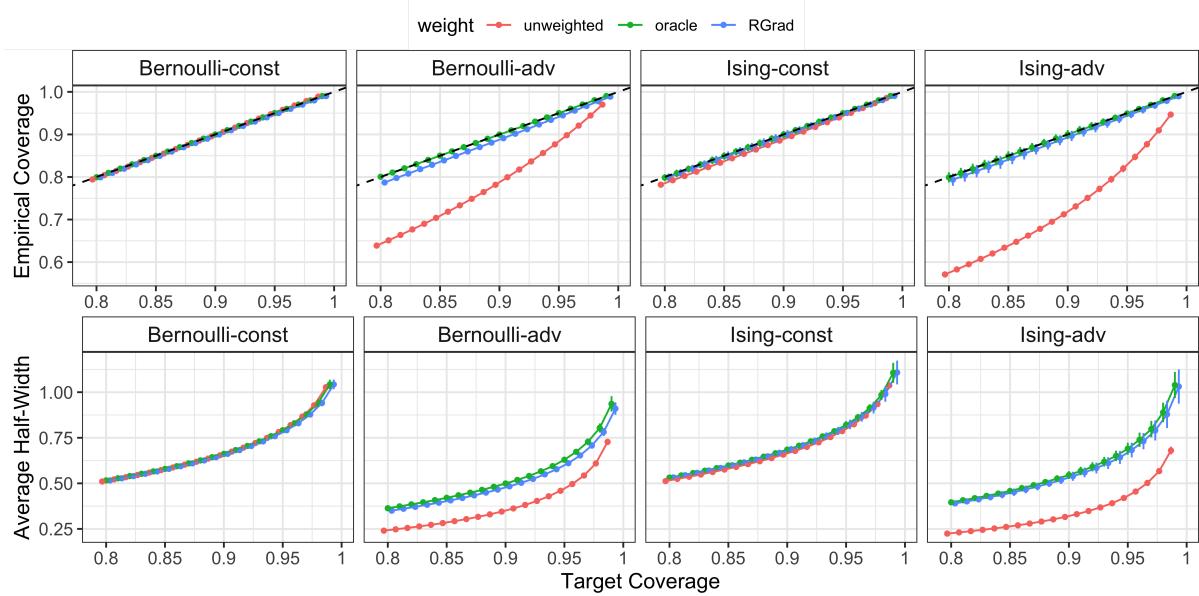


Figure 7: Empirical coverage and average confidence interval half-width of the three conformal prediction methods across the Bernoulli and Ising model with constant (const) or adversarial (adv) noise. Results are based on $n = 30$ repetitions and error bars are ± 1.96 standard deviations.

for estimating the missing propensity and conducting the weighted conformal prediction with our method. We mainly consider two competing methods other than the unweighted and oracle conformal prediction: 1) **GCP**: binary tensor decomposition with generalized CP-decomposition (Wang & Li 2020, Hong et al. 2020); 2) **Tucker**: binary tensor decomposition with generalized Tucker-decomposition (Lee & Wang 2020, Cai, Li & Xia 2022a). Different from our approach, these two methods assume independence among all the binary entries and thus they are misspecified under the Ising model. We conduct **GCP** with gradient descent following (Hong et al. 2020) and **Tucker** with Riemannian gradient descent following (Cai, Li & Xia 2022a) and select the corresponding ranks using the BIC criterion, as suggested by the literature. We consider $r = 3, d \in \{40, 60\}$ and list the average mis-coverage % under the constant and adversarial noise regimes as well as the RSE of the estimated $\widehat{\mathcal{B}}$ in Table 4.

Our finding from Table 4 is that our method consistently provides well-calibrated confidence intervals close to the oracle case and performs on average better than the GCP and

Tucker method. Our mis-coverage % is statistically significantly better (p -value < 0.005) than the Tucker method under the adversarial noise regimes across different tensor dimensions and missingness generating models. The GCP method, surprisingly, provides confidence intervals close to our method but has significantly larger RSE for the estimator $\widehat{\mathcal{B}}$. We found that CP-decomposition tends to underestimate the weights of the calibration data, therefore it has more testing data points with infinitely wide confidence intervals, making it less favorable.

| (d, Model) | Method | const. mis-coverage % | adv. mis-coverage % | RSE |
|---------------------|------------|-----------------------|---------------------|---------------------|
| (40, Bern) | unweighted | 0.463(0.244) | 11.1(0.389) | / |
| | oracle | 0.381(0.205) | 0.409(0.232) | / |
| | GCP | 0.373(0.183) | 1.66(0.965) | 0.522(0.069) |
| | Tucker | 0.380(0.219) | 0.841(0.431) | 0.295(0.008) |
| | RGrad | 0.377(0.235) | 0.773(0.404) | 0.345(0.010) |
| (60, Bern) | unweighted | 0.401(0.165) | 11.0(0.241) | / |
| | oracle | 0.202(0.082) | 0.207(0.105) | / |
| | GCP | 0.203(0.092) | 0.380(0.298) | 0.281(0.036) |
| | Tucker | 0.199(0.079) | 0.842(0.231) | 0.244(0.005) |
| | RGrad | 0.200(0.078) | 0.821(0.226) | 0.271(0.004) |
| (40, Ising) | unweighted | 1.19(0.298) | 17.3(0.528) | / |
| | oracle | 0.568(0.278) | 0.666(0.331) | / |
| | GCP | 0.870(0.597) | 1.24(0.840) | 1.81(0.621) |
| | Tucker | 0.504(0.241) | 1.80(0.653) | 0.444(0.010) |
| | RGrad | 0.713(0.377) | 1.13(1.21) | 0.341(0.304) |
| (60, Ising) | unweighted | 1.35(0.243) | 17.2(0.310) | / |
| | oracle | 0.302(0.136) | 0.370(0.242) | / |
| | GCP | 0.349(0.181) | 0.638(0.506) | 1.59(1.16) |
| | Tucker | 0.329(0.216) | 2.03(0.368) | 0.404(0.007) |
| | RGrad | 0.356(0.154) | 0.580(0.339) | 0.224(0.003) |

Table 4: Method comparisons of different conformal prediction methods with $r = 3$. The results include the average mis-coverage % defined in (23) under the constant (const.) and adversarial (adv.) noise regimes as well as the relatively squared error (RSE) of the estimator $\widehat{\mathcal{B}}$.