Nonparametric Tensor Estimation and Completion via Hypergraphon Learning

Miaoyan Wang
University of Wisconsin – Madison
miaoyan.wang@wisc.edu

Abstract

We consider the problem of tensor estimation from noisy observations with possibly missing entries. A nonparametric approach to tensor estimation is developed based on K-uniform hypergraphons. The hypergraphon model captures the key features of conditional independence arising in the generative process of multiway tensor data. The nonparametric hypergraphon representation of tensors encompasses many existing tensor models—such as CP models, Tucker models, shape constrained tensor models—as special examples. We develop a rate-optimal nonparametric tensor estimator using the stochastic blockmodel approximation to the underlying hypergraphon. The integrated risk bound for the hypergraphon estimation is established for specially structured functional classes. The result uncovers the joint contribution of the statistical bias-variance error and the agnostic discretization error. Numerical results demonstrate the robustness of our proposal over previous tensor methods and the attractive performance as the tensor order increases.

1 Introduction

2 Nonparametric Tensor Model via Hypergraphons

Let $\mathcal{Y} = [\![\mathcal{Y}(\boldsymbol{i})]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be an order-K $(d_1, \ldots, d_K)$-dimensional data tensor, where $\boldsymbol{i} = (i_1, \ldots, i_K)$ is the K-way index. We propose a two-stage generative process for the tensor observation. First, we draw a collection of i.i.d. random variables, $x_k(i) \in \mathcal{X}_k$, from some probability measure $(\mathcal{X}_k, \mu_{\mathcal{X}_k})$, for all $i \in [d_k]$, $k \in [K]$. The Cartesian product of the random variables, denoted $(x_1(i_1), \ldots, x_K(i_K))$, represents the latent features at position $\boldsymbol{i} = (i_1, \ldots, i_K)$ of the tensor. Second, conditional on the latent features, the tensor entries are drawn independently with mean $f((x_1(i_1), \ldots, x_K(i_K))$.

- Nonparametric mean: there exists a unknown function $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_K \mapsto [0,1]$ such that

$$\mathbb{E}\mathcal{Y}(\boldsymbol{i}) = f(x_1(i_1), \ldots, x_K(i_K)), \quad \text{for all } \boldsymbol{i} = (i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]. \tag{1}$$

  Here $x_k(i_k) \in \mathcal{X}_k$ denotes the latent feature associated with the i_k-th entry along the k-th mode of the tensor, where $k \in [K]$.
- Latent features: for each mode k, the latent features $\{x_k(i_k)\colon i_k \in [d_k]\}$ are sampled independently and identically from the probability measure $(\mathcal{X}_k, \mu_{\mathcal{X}_k})$.
- Conditionally independence: conditional on the latent features $x_k(i_k)$, the tensor entries $\mathcal{Y}(\boldsymbol{i})$ are independent, sub-Gaussian random variables; i.e. $\mathbb{E}\exp[t(\mathcal{Y}(\boldsymbol{i}) - \mathbb{E}(\mathcal{Y}(\boldsymbol{i}))] \le \exp(t^2\sigma^2/2)$ for all $\boldsymbol{i} \in [d_1] \times \cdots \times [d_K]$ and $t \in \mathbb{R}$.
- For simplicity, we set $\mathcal{X}_k = [0,1]$ and $\mu_{\mathcal{X}_k}$ the Lebesgue measure over $[0,1]$ for all $k \in [K]$. Furthermore, we assume the latent features are mutually independent across K modes.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

- Regularity conditions on f. Two possible options: 1. stepwise functions, 2. Holder smooth functions.

We call the model (1) the nonparametric tensor model because the function f is unknown and to be estimated.

In the case of binary tensor observations, our nonparametric model is closely connected to the hypergraphon model in the graphical literature. Specifically, let $\mathcal{G} = (V, E)$ be a K-uniform hypergraph, where $V = [d]$ is the node set and $E \subset V^{\otimes K}$ is the hyperedge set with each hyperedge connecting precisely K nodes, $K \leq d$. The hypergraphon model assumes that the hyperedges are generated through a symmetric, measurable function $f\colon [0,1]^K \to [0,1]$,

$$\mathbb{1}\{\boldsymbol{i} \in E\} \sim \text{Bernoulli}(f(x(i_1), \ldots, x(i_K)), \qquad \text{for all } \boldsymbol{i} \in [d]^K,$$

where $\{x(i)\colon i \in [d]\}$ is an i.i.d. random sample from $U[0,1]$, and the events $\mathbb{1}\{\boldsymbol{i} \in E\}$ are mutually independent conditional on $\{x(i)\}$. The function f is referred to as the K-uniform hypergraphon. Our nonparametric tensor model (1) generalizes the hypergraphon model by allowing more flexible observations with mode-specific latent features?? and asymmetric latent function f. For this reason, we adopt the terminology and call the function f a hypergraphon.

We use $(\mathcal{X}_k, \mu_{\mathcal{X}_k}, f)$ to denote the sampling scheme for the latent features and the hypergraphon associated with our nonparametric tensor model (1). By specializing the latent features in $\mathcal{X}_k$ and the function f, the conditional mean model (1) incorporates several common previously-studied tensor models as special cases.

Low-rank model. Let $\mathcal{X}_k \subset \mathbb{R}^{r_k}$ be a bounded close set, and $\mu_{\mathcal{X}_k}$ a probability measure over $\mathcal{X}_k$. Consider a multilinear hypergraphon

$$\begin{aligned} f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_K &\to \mathbb{R} \\ (x_1, \ldots, x_K) &\mapsto \mathcal{C} \times_1 x_1^T \times_2 \cdots \times_K x_K^T, \end{aligned} \tag{2}$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is a fixed coefficient tensor. Let $x_k(i_k) \in \mathcal{X}_k$ be the realization of the mode-k latent feature at index $i_k \in [d_k]$, and $\boldsymbol{X}_k = [x_k(1)|\ldots|x_k(d_k)] \in \mathbb{R}^{r_k \times d_k}$ the corresponding feature matrix. Then, model (1) induces a rank-$(r_1, \ldots, r_K)$ Tucker model:

$$\mathbb{E}\left(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\right) = \mathcal{C} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T. \tag{3}$$

Similarly, our model incorporates the CP tensor model by setting $r_1 = \cdots = r_K = r$ and a super-diagonal core tensor $\mathcal{C}$.

Nonlinear single-index model. Consider the same setting as in Example 1. Let $f' = g \circ f$, where f is defined as in (2), $g\colon \mathbb{R} \to \mathbb{R}$ is a nonlinear monotonic function, and $\circ$ denotes the function composition. Then the model (1) induces a nonlinear single-index model:

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) = g(\mathcal{C} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T).$$

Here the function g could be either parametric such as a logistic function as in Bradly-Terry model, or nonparametric such as a monotonic, Lipschitz function as in [1]. Note that, with the nonlinear transformation, the data tensor is likely to have full rank in expectation.

Stochastic transitivity model. Let $\mathcal{X}_k \subset \mathbb{R}$ be a bounded close set, and $\mu_{\mathcal{X}_k}$ a probability measure over $\mathcal{X}_k$. Consider a monotonic hypergraphon $f\colon \mathbb{R}^K \to \mathbb{R}$ in that

$$f(x_1, \ldots, x_K) \leq f(x_1', \ldots, x_K')$$

whenever $x_k \leq x_k'$ for all $k \in [K]$. Then, model (1) reduces to the strong stochastic transitivity model; i.e., there exist a set of permutations $\sigma_k\colon [d_k] \to [d_k]$ such that the entries are monotonically increasing along the permuted indices:

$$\mathbb{E}\mathcal{Y}(\sigma_1(i_1), \ldots, \sigma_K(i_K)) \leq \mathbb{E}\mathcal{Y}(\sigma_1(i_1'), \ldots, \sigma_K(i_K')), \tag{4}$$

whenever $\sigma_k(i_k) \leq \sigma_k(i_k')$ for all $k \in [K]$. The strong stochastic transitivity (4) is also known as rank-1 permutation model; it was initially proposed for the matrix case $K = 2$. Our formulation extends the model to higher-order cases. More generally, by setting f as a mixture of shape-constrained functions over multivariate latent features $\mathcal{X}_k = \mathbb{R}^r$, our model encompasses the more general low permutation-rank models and statistical seriation models.

Stochastic block model. Let $\mathcal{X}_k = [0,1]$ and $\mu_{\mathcal{X}_k}$ the Lebesque measure over $[0,1)$. For each k, write $\mathcal{X}_k = [0,1/r_k) \cup [2/r_k, 3/r_k) \cup \cdots \cup [(r_k-1)/r_k, 1)$ as a disjoint union of r_k equal-sized intervals. Define a piecewise constant hypergraphon

$$f\colon [0,1]^K \to \mathbb{R}$$
$$(x_1,\ldots,x_K) \mapsto \sum_{j_1,\ldots,j_K} c_{j_1,\ldots,j_K} \mathbb{1}\left\{x_k \in \left[\frac{j_k-1}{r_k}, \frac{j_k}{r_k}\right), \text{ for all } k \in [K]\right\},$$

where $\mathcal{C} = [\![c_{j_1,\ldots,j_K}]\!] \in \mathbb{R}^{r_1 \times \cdots \times r_k}$ is a fixed tensor specifying the block means. Let $x_k(i_k) \in \mathcal{X}_k$ be the realization of the mode-k latent feature at index $i_k \in [d_k]$. Then model (1) reduces to a stochastic block model,

$$\mathbb{E}\mathcal{Y}(\boldsymbol{i})|\{x_k(i_k)\} = c_{j_1,\ldots,j_K}, \tag{5}$$

where $j_k \in [r_k]$ is the mode-k block index for which $(j_k-1)/r_k \le x_k(i_k) \le j_k/r_k$, $k \in [K]$.

3 Identifiability and problem statements

We examine the model identifiability in this section. The nonparametric model (1) is determined by the random latent features and the hypergraphon, denoted $(\mathcal{X}_k, \mu_{\mathcal{X}_k}, f)$. Our goal is to estimate the function f from a noisy tensor observation $\mathcal{Y}$, without knowing the latent features. The function f is nonidentifiable due to two complications, as we describe below.

The first complication comes from the indeterminacy of the feature space $\mathcal{X}_k$ and the domain of function f. In particular, $(\mathcal{X}_k, \mu_{\mathcal{X}_k}, f)$ and $(g_k(\mathcal{X}_k), \mu_{g_k(\mathcal{X}_k)}, f \circ g)$ induce the same conditional model, where g_k is a measurable function defined on $\mathcal{X}_k$ and $f \circ g$ denotes the function $(x_1,\ldots,x_K) \mapsto f(g_1(x_1),\ldots,g_K(x_K))$. As an example, we show a different latent variable representation to construct the model in Example 4.

Example 1 (A different latent feature sampling for stochastic block model)**.** Let $\mathcal{X}_k = \{0,1\}^{r_k}$ and $\mu_{\mathcal{X}_k}$ the multinomial distribution with equal probability over r_k categories. The latent feature is encoded as an indicator vector $x_k = (0,\ldots,0,1,0,\ldots,0)^T \in \mathcal{X}_k$, where the position of the entry 1 follows from the distribution $\mu_{\mathcal{X}_k}$. Now consider a multilinear hypergraphon $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_K \to \mathbb{R}$ defined in (2). The induced conditional mean model (3) has the following form:

$$\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_K) = \mathcal{C} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T, \tag{6}$$

where $\boldsymbol{X}_k \in \{0,1\}^{r_k \times d_k}$ is a membership matrix that collects the sampled latent features along mode k. Note that model (6) is exactly the same as model (5) in Example 4.

In view of the above discussion, we fix $\mathcal{X}_k = [0,1]$ and set $\mu_{\mathcal{X}_k}$ the Lebesgue measure over $[0,1]$ in the sequel. The hypergraphon of interest is defined on $[0,1]^K$. The following result ensures that our assumption is general enough to incorporate separable Hilbert latent spaces such as the Examples 1–4 mentioned in Section (2).

Proposition 1 (Reduction to 1-d uniform sampling)**.** *Let $\mathcal{X}$ be a separable Hilbert space equipped with a probability measure $\mu_{\mathcal{X}}$. There exists a transform sampling function $g\colon [0,1] \to \mathcal{X}$, such that*

$$s \sim \mathrm{Unif}(0,1) \Rightarrow g(s) \sim \mu_{\mathcal{X}}.$$

The transformation relates the sampling from general latent feature space $(\mathcal{X}, \mu_{\mathcal{X}})$ to the Lebesque measure on $[0,1]$. If $\mathcal{X}$ is one-dimensional, then a well known example of g is the inverse cumulative distribution function for $\mu_{\mathcal{X}}$. For general feature space such as r-dimensional Euclidean space, the function $g(\cdot)$ is given in Appendix.

However, this f may not satisfy the continuity condition. holds under additional assumption $\mathcal{X}$ has bounded density..

The second complication comes from the invariance of distribution with respect to the measure-preserving mapping. More precisely, two hypergraphons f and f' define the same distribution over $\mathcal{Y}$ if there exist a pair of measure-preserving maps $\phi, \phi'\colon [0,1]^K \to [0,1]^K$ such that

$$f(\phi(x_1,\ldots,x_K)) \overset{a.s.}{=} f'(\phi'(x_1,\ldots,x_K)). \tag{7}$$

When the equation (7) holds, we say f and f' are *weakly isomorphic*, denoted as $f \overset{w.i.}{=} f'$. The notion $\overset{w.i.}{=}$ defines an equivalence relation in the space of measurable functions from $[0,1]^K$ to $[0,1]$. As a consequence, one can only estimate the equivalence class of f, and we refer to hypergraphon estimation as the problem of estimating an equivalence class of f.

In the present paper, we consider the following two problems:

- Q1 [Signal tensor estimation]. Estimate the conditional mean tensor Θ from a single observation of data tensor $\mathcal{Y}$.
- Q2 [Hypergraphon estimation]. Estimate the hypergraphon $f\colon [0,1]^K \to [0,1]$ from a single observation of data tensor $\mathcal{Y}$.

We allow missing values in the data tensor; the framework therefore also extends to tensor completion. Specifically, we introduce a masking tensor $\mathcal{Z} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ with each entry drawn independently as Bernoulli(p) for some $p \in (0,1]$.

The function can be seen as a kernel function for random tensor models.

4 Single index model

Let $\boldsymbol{S}^r = \{\boldsymbol{x} \in \mathbb{R}^r\colon \|\boldsymbol{x}\|_2 \le 1\}$ be an r-dimensional unit ball. We also write $\boldsymbol{x} = (\boldsymbol{x}(1), \dots, \boldsymbol{x}(r))^T$, where $\boldsymbol{x}(i) \in \mathbb{R}$ denotes i-th element in the vector $\boldsymbol{x}$. Consider the following generative process:

- Consider an m-variate function

$$f\colon \boldsymbol{S}^r \times \cdots \times \boldsymbol{S}^r \to [0,1],$$
$$(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) \mapsto f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) = \sum_{i=1}^{r} \boldsymbol{x}_1(i) \cdots \boldsymbol{x}_m(i),$$

  By the multilinearlity and boundednss of $\boldsymbol{S}^r$, f is a 1-Lipschitz function.
- Draw i.i.d. samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d$ from $\boldsymbol{S}^r$.
- Define the signal tensor

$$\Theta(i_1, \dots, i_m) = f(\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_m}), \quad \text{for all } (i_1, \dots, i_m) \in [d]^m.$$

  By construction, Θ is a rank-r tensor.

Theorem 1. *Let $\Theta \in [0,1]^m$ be an order-m tensor generated from steps 1-3. Then, for every $k \in \mathbb{N}_+$, there exists a block-k tensor such that*

$$\|\Theta - \mathit{Block}(\Theta; K)\|_F^2 \le \frac{d^m}{K^{2/r}}.$$

Proof. Because $\boldsymbol{S}^r$ is a compact set, $\boldsymbol{S}^r$ can be covered by K Euclidean balls with radius $K^{-1/r}$. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_K$ be the center of these balls. Then, for every $i \in [d]$, there exists $k = k(i) \in \{1, \dots, K\}$ such that

$$\|\boldsymbol{x}_i - \boldsymbol{y}_{k(i)}\|_2 \le \frac{1}{K^{1/r}}.$$

Therefore,

$$|f(\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_m}) - f(\boldsymbol{y}_{k(i_1)}, \dots, \boldsymbol{y}_{k(i_m)})|^2 \le \frac{L^2}{K^{2/r}}.$$

Notice that there are in total K center points, so the

$$\bar{\Theta}(i_1, \dots, i_m) := f(\boldsymbol{y}_{k(i_1)}, \dots, \boldsymbol{y}_{k(i_m)})$$

defines a block tensor $\bar{\Theta}$ with K blocks on each model. We claim, for every $k \in \mathbb{N}$, there exists a block tensor

$$\|\Theta - \mathrm{Block}(\Theta; K)\|_F^2 \le \frac{d^m}{K^{2/r}}.$$

□

Acknowledgements

References

[1] Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881, 2015.

[2] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723, 2019.