Supplementary Appendix

A1. 2D Shape Example with Varying Size and Strength

We report here the snapshots of estimates with varying signal strength at sample size $n = 500$ (Figure S.1), and the line plot of the average root mean squared error (RMSE) for estimation of $\boldsymbol{B}$ with both varying sample size and signal strength (Figure S.2). The findings were summarized in Section 6.1.

A2. 2D Shape Example with Regularization

We have run a numerical experiment to illustrate regularized tensor regression estimation. The setup is the same as that in Figure 1 except that the sample size is reduced to 500, which is only barely larger than the number of parameters $380 = 5+3\times(64+64)-9$ of a rank-3 tensor model. Figure S.3 shows the outcome of applying the lasso penalty to $\boldsymbol{B}_d$ in the rank-3 tensor regression model. Recovered signals at three different values of $\lambda = 0, 100, 1000$ are displayed. Without regularization ($\lambda = 0$), the rank-3 tensor regression is difficult to recover some signals such as triangle, disk and butterfly, mainly due to a very small sample size. On the other hand, excessive penalization compromises the quality of recovered signals too, as evidently in those shapes at $\lambda = 1000$. Regularized estimation with an appropriate amount of shrinkage improves estimation quality, as seen in triangle and disk at $\lambda = 100$ and in butterfly at $\lambda = 1000$. In practice the tuning parameter is chosen by certain model selection criterion such as BIC or cross validation. Moreover, we have experimented with the bridge and SCAD penalties for the same data and obtained similar results. The desirable unbiased (or nearly unbiased) estimates from these concave penalties are reflected by the improved contrast in the recovered signal. For the sake of space, we do not show those figures here.

A3. Comparison with Classical Lasso

We compare our regularized fixed rank tensor estimate with a classical regularized model, the lasso applied to vectorized image covariates. Our purpose is to investigate which method could provide a better estimate to the complicated true array signal with a

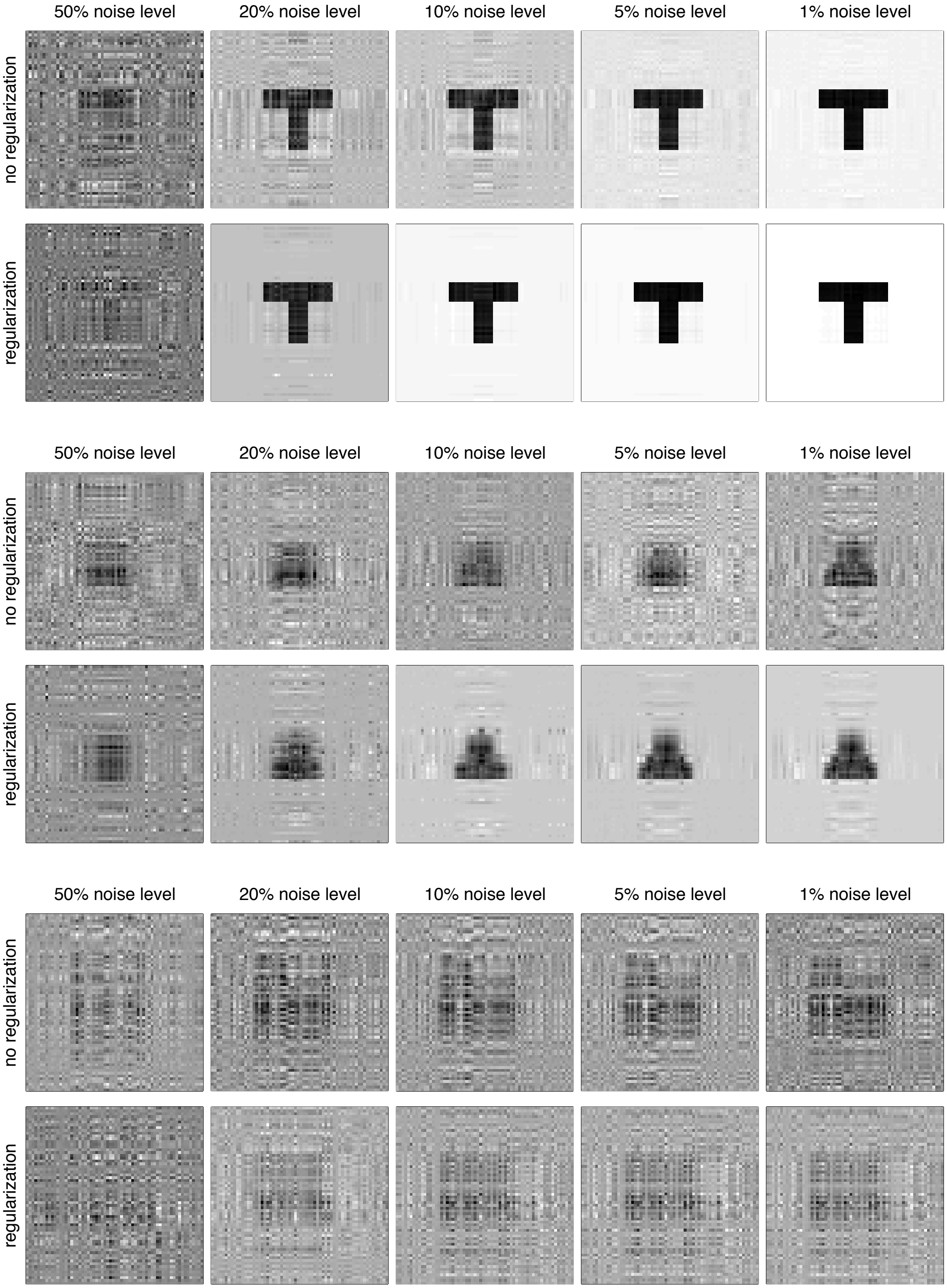

Figure S.1: Snapshots of tensor estimation with varying noise level. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black).

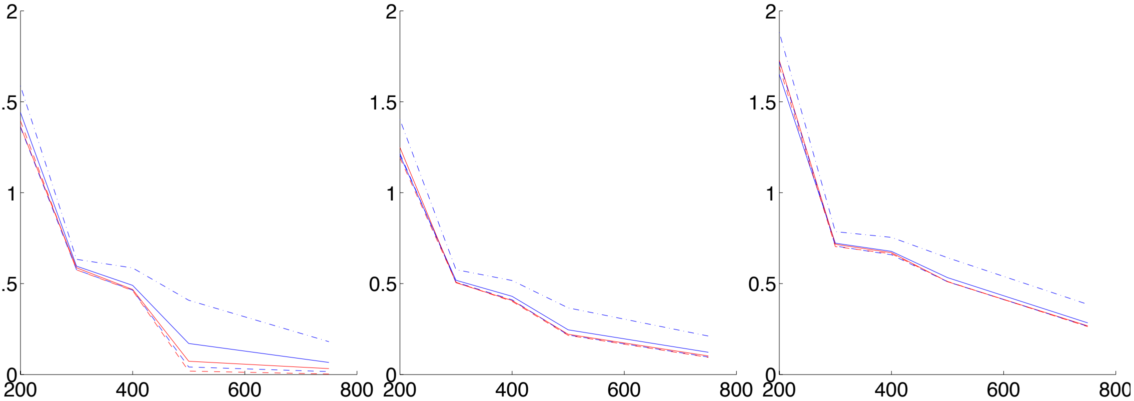

Figure S.2: Line plot of the average root mean squared error (RMSE, y-axis) for estimation of $\boldsymbol{B}$ with varying sample size (x-axis). Five lines denote the average RMSE when the noise level is 50% (blue dash-dot), 20% (blue solid), 10% (red solid), 5% (blue dash) and 1% (red dash) of the signal.

limited sample size. We reproduce the regularized tensor estimates of the "disk", "triangle", and "butterfly" signals in Figure S.3 with a rank 3 model. In addition, we also display the regular lasso estimates (i.e., lasso penalty applied to the vectorized matrix covariates) at the same sample size $n = 500$. The tuning parameter is chosen according to BIC. The results are shown in Figure S.4. It is clearly seen that the vector version of lasso estimates are far off from the truth whereas our tensor version estimates are much better.

A4. Algorithm Stability and Computing Time

We have carried out a numerical experiment to study the algorithm stability and the computing time. We report the results in Figure S.5. We adopt the setting of the illustrative example, using a "triangle" signal. Only one data instance was simulated with a fixed sample. Then the algorithm was initialized from 100 random starting points for tensor regression models at rank $r = 1, 2, 3, 4$. Box-plots of the final model deviances and wall clock run times are displayed in Figure S.5. All run times were recorded on a standard laptop computer with a 2.6 GHz Intel i7 CPU. As expected, higher rank models fit the data better, yielding smaller deviance, since the true signal is of a high rank. On the other hand, higher rank models are more vulnerable to local modes, as indicated

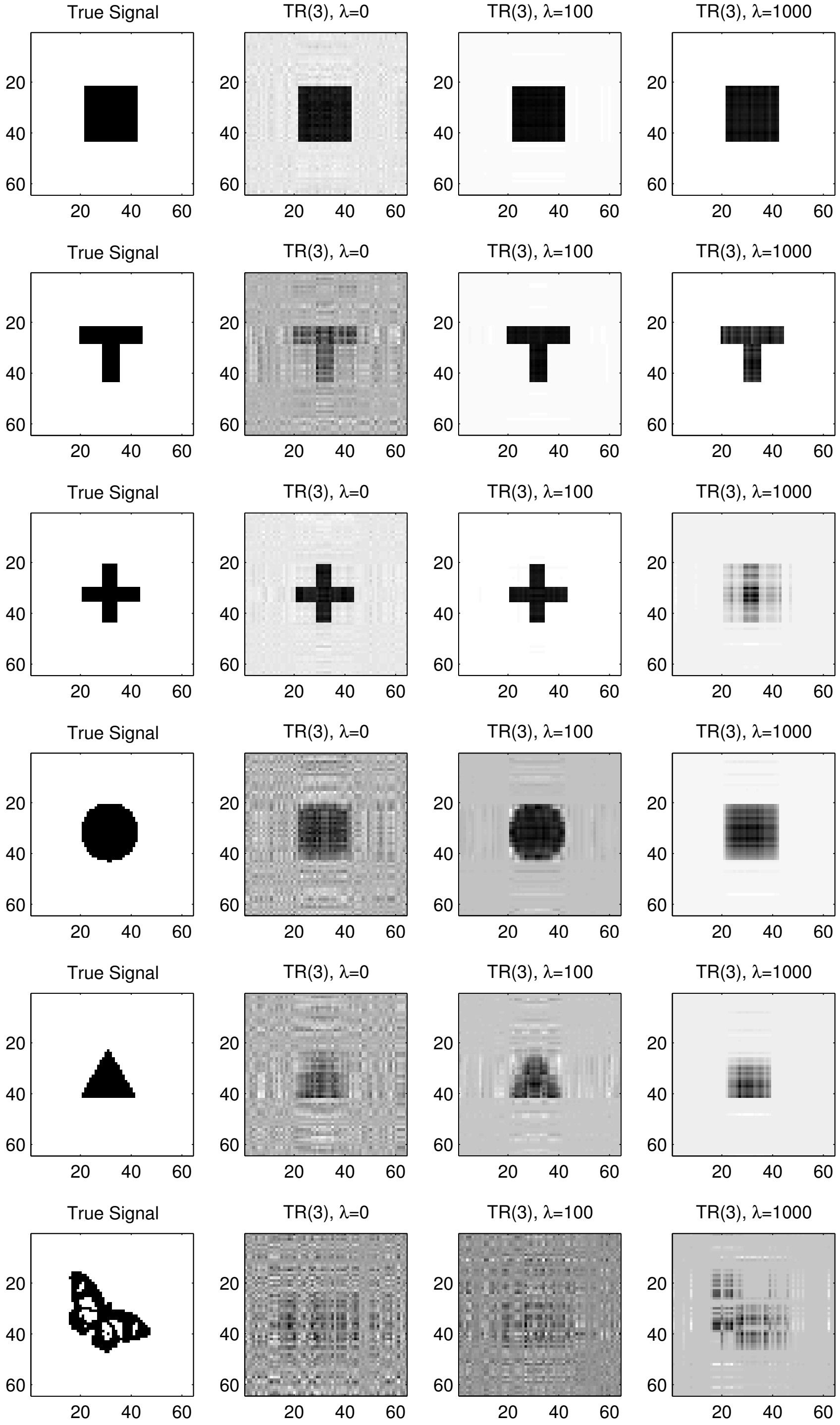

Figure S.3: Demonstration of lasso regularization. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black). The sample size is 500.

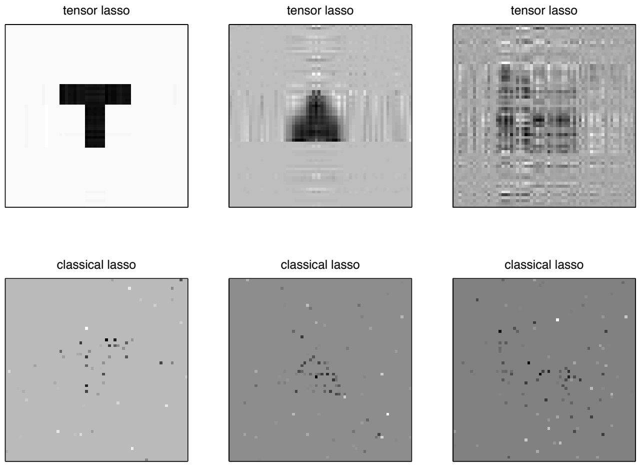

Figure S.4: Tensor lasso estimate (top) vs classical lasso estimate (bottom; applied on the vectorized matrix covariates) for T-shape, triangle and butterfly at sample size $n = 500$.

by larger variations, and takes longer to converge. The overall run time, however, is remarkably fast. For instance, the median run time of fitting a rank 3 model in this example is about 5 seconds. Fitting a rank-3 logistic model to the 3D ADHD data in Section 6.2.3 took about 285 seconds for 10 runs from 10 random starting points, averaging < 30 seconds per run.

A5. Proofs

Proof of Lemma 1

For the first identity it is enough to check that the mode-d matricization of $\boldsymbol{b}_1 \circ \cdots \circ \boldsymbol{b}_D$ is $\boldsymbol{b}_d(\boldsymbol{b}_D \otimes \cdots \otimes \boldsymbol{b}_{d+1} \otimes \boldsymbol{b}_{d-1} \otimes \cdots \otimes \boldsymbol{b}_1)^{\mathsf{T}}$, which is easily seen to hold elementwise. The scalar product $\prod_{d' \neq d} b_{d' i_{d'}}$ appears as the j-th element of the row vector $(\boldsymbol{b}_D \otimes \cdots \otimes \boldsymbol{b}_{d+1} \otimes \boldsymbol{b}_{d-1} \otimes \cdots \otimes \boldsymbol{b}_1)^{\mathsf{T}}$ where $j = 1 + \sum_{d' \neq d} (i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$. The matricization of a sum of arrays equals sum of their matricizations. Therefore the first identity holds.

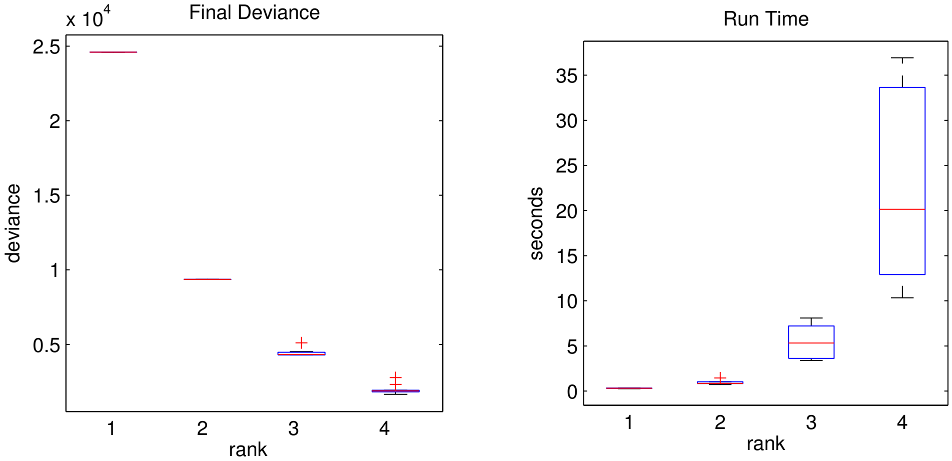

Figure S.5: Algorithm stability and run time. The same algorithm is initialized from 100 random starting points to fit tensor regression models at rank $r = 1, 2, 3, 4$ respectively. The final model deviances and wall clock timings are reported.

For the second identity,

$$
\begin{aligned}
\operatorname{vec} \boldsymbol{B} &= \operatorname{vec}\Big(\sum_{r=1}^{R} \boldsymbol{b}_1^{(r)} \circ \cdots \circ \boldsymbol{b}_D^{(r)}\Big) = \sum_{r=1}^{R} \operatorname{vec}(\boldsymbol{b}_1^{(r)} \circ \cdots \circ \boldsymbol{b}_D^{(r)}) \\
&= \sum_{r=1}^{R} \boldsymbol{b}_D^{(r)} \otimes \cdots \otimes \boldsymbol{b}_1^{(r)} = (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1)\mathbf{1}_R.
\end{aligned}
$$

Proof of Proposition 1

Proof of global convergence follows from standard arguments for algorithms that monotonically increase objective function (de Leeuw, 1994; Lange, 2004, 2010). Under the blockwise strict concavity condition (2), the block update is well-defined and differentiable. Then algorithmic map M is a composition of $D+1$ differentiable maps and, by implicit function theorem, continuous. Let $\boldsymbol{\theta}^{(t)}$ be the sequence generated by M and $\boldsymbol{\theta}$ be any accumulation point of $\boldsymbol{\theta}^{(t)}$. Since the algorithm always increase objective value, $\ell(M(\boldsymbol{\theta}^{(t)})) \ge \ell(\boldsymbol{\theta}^{(t)})$. Taking limit gives $\ell(M(\boldsymbol{\theta})) = \ell(\boldsymbol{\theta})$ by continuity of M and ℓ. Thus any accumulation point of algorithmic sequence is a stationary point of ℓ. The set of accumulation points is contained in $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \ge \ell(\boldsymbol{\theta}^{(0)})\}$ and thus compact by condition (1). Compactness implies that this set of accumulation points is also connected (Lange, 2010, Propitions 8.2.1 and 15.4.2). Discreteness of the stationary points of ℓ implies that the number of stationary points is finite. Otherwise there is a sequence of stationary points whose limit is not isolated. Finally the set of accumulation points is a connected subset of these finite number of stationary points, thus is a single point. In other words the algorithmic sequence $\boldsymbol{\theta}^{(t)}$ converges to a stationary point of ℓ.

Proof of local convergence relies on the Ostrowski's theorem (Ostrowski, 1960), which states that the sequence $\boldsymbol{\theta}^{t+1} = M(\boldsymbol{\theta}^t)$ is locally attracted to $\boldsymbol{\theta}^\infty$ if the spectral radius of the differential of the algorithmic map $\rho[dM(\boldsymbol{\theta}^\infty)]$ is strictly less than 1. We partition the Hessian of the objective function ℓ at $\boldsymbol{\theta}^\infty$ as

$$d^2\ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \begin{pmatrix} d^2_{00}\ell & & \mathbf{0} & \\ & d^2_{11}\ell & \cdots & d^2_{1D}\ell \\ \mathbf{0} & \vdots & \ddots & \vdots \\ & d^2_{D1}\ell & \cdots & d^2_{DD}\ell \end{pmatrix} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}^\intercal,$$

where $\boldsymbol{L}$ is the strictly block lower triangular part and $\boldsymbol{D}$ is the block diagonal part. Then it can be shown that the differential of map M is

$$dM(\boldsymbol{\theta}^\infty) = -(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}^\intercal.$$

Note $\boldsymbol{\theta}^\infty$ being a strict local maximum implies that $d^2\ell(\boldsymbol{\theta}^\infty)$ is strictly negative definite and thus the diagonal blocks $d^2_{dd}\ell$, $d = 0, \ldots, D$ are strictly negative definite too. Therefore the block lower triangular matrix $(\boldsymbol{L} + \boldsymbol{D})$ is invertible as it shares the same eigenvalues as its diagonal blocks. The spectral radius of $-(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}^\intercal$ is strictly less than one. Therefore the iterates are locally attracted to $\boldsymbol{\theta}^\infty$.

Proof of Lemma 2

By Lemma 1,

$$\boldsymbol{B}_{(d)} = \boldsymbol{B}_d(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1)^\intercal.$$

Using the well-known fact that $\text{vec}(\boldsymbol{XYZ}) = (\boldsymbol{Z}^\intercal \otimes \boldsymbol{X})\text{vec}(\boldsymbol{Y})$,

$$\text{vec}\boldsymbol{B}_{(d)} = [(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_d}]\text{vec}(\boldsymbol{B}_d).$$

Thus we have

$$\begin{aligned}\boldsymbol{J}_d &= D\boldsymbol{B}(\boldsymbol{B}_d)\\ &= D\boldsymbol{B}(\boldsymbol{B}_{(d)}) \cdot D\boldsymbol{B}_{(d)}(\boldsymbol{B}_d)\\ &= \boldsymbol{\Pi}_d \frac{\partial \text{vec}\boldsymbol{B}_{(d)}}{\partial(\text{vec}\boldsymbol{B}_d)^{\mathsf{T}}}\\ &= \boldsymbol{\Pi}_d[(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_n}].\end{aligned}$$

Combining gives

$$\begin{aligned}&D\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= D\eta(\boldsymbol{B}) \cdot D\boldsymbol{B}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= (\text{vec}\boldsymbol{X})^{\mathsf{T}}[\boldsymbol{J}_1 \ \boldsymbol{J}_2 \ \cdots \ \boldsymbol{J}_D].\end{aligned}$$

For the Hessian,

$$h_{(i_d,r),(i_{d'},r')} = \sum_{j_1,\ldots,j_D} x_{j_1,\ldots,j_D} \frac{\partial^2 b_{j_1,\ldots,j_D}}{\partial \beta_{i_d}^{(r)} \partial \beta_{i_{d'}}^{(r')}}.$$

The second derivative in the summand is nonzero only if $j_d = i_d$, $j_{d'} = i_{d'}$, $r = r'$, and $d \neq d'$. Therefore

$$h_{(i_d,r),(i_{d'},r')} = 1_{\{r=r',d\neq d'\}} \sum_{j_d=i_d,j_{d'}=i_{d'}} x_{j_1,\ldots,j_D} \prod_{d''\neq d,d'} \beta_{j_{d''}}^{(r)},$$

where the sum is over $\prod_{d''\neq d,d'} p_{d''}$ terms. It is easy to see that $h_{(i_d,r),(i_{d'},r')}$ are the entries of the matrix

$$\boldsymbol{X}_{(dd')}(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_{d'+1} \odot \boldsymbol{B}_{d'-1} \odot \cdots \odot \boldsymbol{B}_1).$$

Proof of Proposition 2

Since $\mu = b'(\theta)$, $d\mu/d\theta = b''(\theta) = \sigma^2/a(\phi)$ and

$$\begin{aligned}\nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) &= \frac{y - b'(\theta)}{a(\phi)} \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \nabla \eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= \frac{(y-\mu)\mu'(\eta)}{\sigma^2}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}(\text{vec}\boldsymbol{X})\end{aligned}$$

by Lemma 2. Further differentiating shows

$$\begin{aligned}
& d^2\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \\
= & -\frac{1}{\sigma^2}\nabla\mu(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) d\mu_i(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) + \frac{y-\mu}{\sigma^2} d^2\mu(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \\
= & -\frac{[\mu'(\eta)]^2}{\sigma^2}([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}\text{vec}\boldsymbol{X})([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}\text{vec}\boldsymbol{X})^{\mathsf{T}} \\
& + \frac{(y-\mu)\theta''(\eta)}{\sigma^2}([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}\text{vec}\boldsymbol{X})([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}\text{vec}\boldsymbol{X})^{\mathsf{T}} \\
& + \frac{(y-\mu)\theta'(\eta)}{\sigma^2} d^2\eta(\boldsymbol{B}).
\end{aligned}$$

It is easy to see that $\mathbf{E}[\nabla\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] = \mathbf{0}$ and $\mathbf{E}[-d^2\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] = \boldsymbol{I}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$, thus (9) follows.

Proof of Proposition 4

The following useful result relates local identifiability of parametric models to their Fisher information matrix.

Lemma 4. *(Rothenberg, 1971, Theorem 1) Let θ_0 be a regular point of the information matrix $I(\theta)$. Then θ_0 is locally identifiable if and only if $I(\theta_0)$ is nonsingular.*

The regularity assumptions for Lemma 4 are satisfied by tensor model: (1) the parameter space $\mathcal{B}$ is open, (2) the density $p(y, \boldsymbol{x}|\boldsymbol{B})$ is proper for all $\boldsymbol{B} \in \mathcal{B}$, (3) the support of the density $p(y, \boldsymbol{x}|\boldsymbol{B})$ is same for all $\boldsymbol{B} \in \mathcal{B}$, (4) the log density $\ell(\boldsymbol{B}|y, \boldsymbol{x}) = \ln p(y, \boldsymbol{x}|\boldsymbol{B})$ is continuously differentiable, and (5) the information matrix

$$\boldsymbol{I}(\boldsymbol{B}) = [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \left[\sum_{i=1}^{n} \frac{\mu'(\eta_i)^2}{\sigma_i^2}(\text{vec}\,\boldsymbol{x}_i)(\text{vec}\,\boldsymbol{x}_i)^{\mathsf{T}}\right] [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]$$

is continuous in $\boldsymbol{B}$ by Proposition 2. Then, by Lemma 4, $\boldsymbol{B}$ is locally identifiable if and only if $\boldsymbol{I}(\boldsymbol{B})$ is nonsingular.

Proof of Theorem 1

It suffices to show the consistency of the estimated factor matrix $\hat{\boldsymbol{B}}_{nd}$, $d = 1, \ldots, D$, which implies the consistency of the tensor estimate $\hat{\boldsymbol{B}}_n = [\![\hat{\boldsymbol{B}}_{n1}, \ldots, \hat{\boldsymbol{B}}_{nD}]\!]$ by continuous mapping theorem. The following well-known theorem is our major tool for establishing consistency.

Lemma 5. *(van der Vaart, 1998, Theorem 5.7) Let M_n be random functions and let M be a fixed function of θ such that*

$$\sum_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0)$$

for every $\epsilon > 0$ and

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to 0 \textit{ in probability.}$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}) \geq M_n(\theta_0) - o_P(1)$ converges in probability to θ_0.

To apply Lemma 5 in our setting, we take the nonrandom function M to be $\boldsymbol{B} \mapsto \mathbb{P}_{\boldsymbol{B}_0}[\ell(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{B})]$ (or its modifications) and the sequence of random functions to be $M_n : \boldsymbol{B} \mapsto \frac{1}{n}\sum_{i=1}^n \ell(y_i, \boldsymbol{x}_i|\boldsymbol{B}) = \mathbb{P}_n M$, where $\mathbb{P}_n$ denotes the empirical measure under $\boldsymbol{B}_0$. Then M_n converges to M a.s. by strong law of large number. The first condition requires that $\boldsymbol{B}_0$ is a well-separated maximum of M. This is guaranteed by the (global) identifiability of $\boldsymbol{B}_0$ and information inequality. The second uniform convergence condition is most convenient to be verified by the Glivenko-Cantelli theory (Pollard, 1984; van der Vaart, 1998; van der Vaart and Wellner, 2000).

For binary regression, the following proof is an expansion of (van der Vaart, 1998, Example 5.40) to tensor binary regression. The density is $p_{\boldsymbol{B}}(y|\boldsymbol{x}) = y\pi(\boldsymbol{B}, \boldsymbol{x}) + (1 - y)(1 - \pi(\boldsymbol{B}, \boldsymbol{x}))$, where $\pi(\boldsymbol{B}, \boldsymbol{x}) = g^{-1}(\langle \boldsymbol{B}, \boldsymbol{x} \rangle)$, where g is the link function. For instance, $\pi(\boldsymbol{B}, \boldsymbol{x}) = 1/(1 + e^{-\langle \boldsymbol{B}, \boldsymbol{x} \rangle})$ corresponds to the logit link and $\pi(\boldsymbol{B}, \boldsymbol{x}) = \Phi(\langle \boldsymbol{B}, \boldsymbol{x} \rangle)$ the probit link. Take $m_{\boldsymbol{B}} = \ln[(p_{\boldsymbol{B}} + p_{\boldsymbol{B}_0})/2]$. First we show that $\boldsymbol{B}_0$ is a well-separated maximum of the function $M(\boldsymbol{B}) := \mathbb{P}_{\boldsymbol{B}_0} m_{\boldsymbol{B}}$. The global identifiability of $\boldsymbol{B}_0$ and information inequality guarantee that $\boldsymbol{B}_0$ is the unique maximum of M. To show that it is a well-separated maximum, we need to verify that $M(\boldsymbol{B}_k) \to M(\boldsymbol{B}_0)$ implies $\boldsymbol{B}_k \to \boldsymbol{B}_0$. Suppose $M(\boldsymbol{B}_k) \to M(\boldsymbol{B}_0)$, then $\langle \boldsymbol{B}_k, \boldsymbol{X} \rangle \to \langle \boldsymbol{B}_0, \boldsymbol{X} \rangle$ in probability. If $\boldsymbol{B}_k$ are bounded, then $\mathbf{E}[\langle \boldsymbol{B}_k - \boldsymbol{B}_0, \boldsymbol{X} \rangle^2] \to 0$ and $\boldsymbol{B}_k \to \boldsymbol{B}_0$ by nonsingularity of $\mathbf{E}[(\text{vec}\boldsymbol{X})(\text{vec}\boldsymbol{X})^{\mathsf{T}}]$. On the other hand, $\boldsymbol{B}_k$ cannot escape to infinity. If they do, then $\langle \boldsymbol{B}_k, \boldsymbol{X} \rangle / \|\boldsymbol{B}_k\| \to 0$ in probability which in turn implies that $\boldsymbol{B}_k / \|\boldsymbol{B}_k\| \to \mathbf{0}$. For the uniform convergence, we see that the class of functions $\{\langle \boldsymbol{B}, \boldsymbol{X}\rangle, \, \boldsymbol{B} \in \mathcal{B}\}$ form a Vapnik-Červonenkis (VC) class. This is true because it is a collection of finite number of polynomials of degree D and then apply the VC vector space argument (van der Vaart and Wellner, 2000, 2.6.15). This implies that $\{\pi(\langle \boldsymbol{B}, \boldsymbol{X}\rangle), \boldsymbol{B} \in \mathcal{B}\}$ is a VC class since π is a monotone function (van der Vaart and Wellner, 2000, 2.16.18). Now $m_{\boldsymbol{B}}$ is Lipschitz in π and π_0 since

$$\frac{\partial m_{\boldsymbol{B}}}{\partial \pi} = \frac{\partial m_{\boldsymbol{B}}}{\partial \pi_0} = \frac{2y-1}{y\pi + (1-y)(1-\pi) + y\pi_0 + (1-y)(1-\pi_0)} \leq \frac{1}{\pi_0} + \frac{1}{1-\pi_0}.$$

A Lipschitz composition of a Donsker class is still a Donsker class (van der Vaart, 1998, 19.20). Therefore $\{\boldsymbol{B} \mapsto m_{\boldsymbol{B}}\}$ is a bounded Donsker class with the trivial envelope function 1. A Donsker class is certainly a Glivenko-Cantelli class. Finally the Glivenko-Cantelli theorem establishes the uniform convergence condition required by Lemma 5.

When the parameter is restricted to a compact set, $\mu = g^{-1}(\langle \boldsymbol{B}, \boldsymbol{x}\rangle)$ is confined in a bounded interval and the log-likelihood ℓ is Lipschitz on the finite interval. If follows that $\{\ell(\boldsymbol{B}) = \ell \circ g^{-1} \circ \langle \boldsymbol{B}, \boldsymbol{X}\rangle, \boldsymbol{B} \in \mathcal{B}\}$ is a Donsker class as composition with a monotone or Lipschitz function preserves the Donsker class. Therefore the Glivenko-Cantelli theorem establish the uniform convergence. Compactness of parameter space implies that $\boldsymbol{B}_0$ is a well-separated maximum if it is the unique maximizer of $M(\boldsymbol{B}) = \mathbb{P}_{\boldsymbol{B}_0} m_{\boldsymbol{B}}$ (van der Vaart, 1998, Exercise 5.27). Uniqueness is guaranteed by the information inequality whenever $\boldsymbol{B}_0$ is identifiable. This verifies the consistency for normal and Poisson regressions.

Proof of Lemma 3

By a well-known result (Lehmann and Romano, 2005, Theorem 12.2.2) or (van der Vaart, 1998, Lemma 7.6), it suffices to verify that the density is continuously differentiable in parameter for μ-almost all x and that the Fisher information matrix exists and is continuous. The derivative of density is

$$\nabla p(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \nabla e^{\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)} = p(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D),$$

which is well-defined and continuous by Proposition 2. The same proposition shows that the information matrix exists and is continuous. Therefore the tensor regression model is q.m.d.

Proof of Theorem 2

The following result relates asymptotic normality to the densities that satisfy q.m.d.

Lemma 6. *(van der Vaart, 1998, Theorem 5.39) Suppose that the model* $(P_\theta : \theta \in \Theta)$ *is q.m.d. at an inner point* θ_0 *of* $\Theta \subset \mathbb{R}^k$*. Furthermore, suppose that there exists a measurable function* $\dot{\ell}$ *with* $\mathbf{P}_{\theta_0}\dot{\ell}^2 < \infty$ *such that, for every* θ_1 *and* θ_2 *in a neighborhood of* θ_0*,*

$$|\ln p_{\theta_1}(x) - \ln p_{\theta_2}(x)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|.$$

If the Fisher information matrix I_{θ_0} *is nonsingular and* $\hat{\theta}_n$ *is consistent, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1). \tag{10}$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $I_{\theta_0}^{-1}$*.*

Lemma 3 shows that tensor regression model is q.m.d. By Proposition 2 and chain rule, the score function

$$\dot{\ell}_{\boldsymbol{B}}(y, \boldsymbol{x}) = d\ell(\boldsymbol{B}) = \frac{(y-\mu)\mu'(\eta)}{\sigma^2}(\operatorname{vec}\boldsymbol{x})^{\mathsf{T}}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]$$

is uniformly bounded in y, $\boldsymbol{x}$, and $\boldsymbol{B}$ ranging over compacta and continuous in $\boldsymbol{B}$ for every y and $\boldsymbol{x}$. For sufficiently small neighborhood U of $\boldsymbol{B}_0$, $\sup_U \|\dot{\ell}_{\boldsymbol{B}}\|$ is square-integrable. Thus the local Lipschitz condition is satisfied and Lemma 6 applies.