Bayesian Conditional Tensor Factorizations for High-Dimensional Classification

Supplementary Appendix A: More properties of the conditional tensor factorization model

Matrix form

We format the conditional probability $P(y|x_1,\ldots,x_p)$ as a $d_1\times\cdots\times d_p$ vector

$$Vec\{P(y|-)\} = \big\{P(y|1,\ldots,1,1), P(y|1,\ldots,1,2),\ldots,P(y|1,\ldots,1,d_p),\ldots,\\ P(y|1,\ldots,d_{p-1},d_p),\ldots,P(y|d_1,\ldots,d_{p-1},d_p)\big\}'$$

and $\lambda_{h_1,\ldots,h_p}(y)$ as a $k_1\times\cdots\times k_p$ vector

$$Vec\{\Lambda(y)\} = \big\{\lambda_{1,\ldots,1,1}(y), \lambda_{1,\ldots,1,2}(y),\ldots,\\ \lambda_{1,\ldots,1,k_p}(y),\ldots,\lambda_{1,\ldots,k_{p-1},k_p}(y),\ldots,\lambda_{k_1,\ldots,k_p}(y)\big\}'.$$

Let $\pi^{(j)}$ be a $d_j\times k_j$ matrix with $\pi_v^{(j)}(u)$ as the (u,v)th element. It is a stochastic matrix, so rows sum to one, by constraint (3). Then representation (2) can be written in vector form:

$$Vec\{P(y|-)\} = \big(\pi^{(1)}\otimes\pi^{(2)}\otimes\cdots\otimes\pi^{(p)}\big)Vec\{\Lambda(y)\},\ \text{for } y=1,\ldots,d_0, \tag{A.1}$$

where $\otimes$ denotes the Kronecker product. Furthermore, if we let $Mat(P)$ and $Mat(\Lambda)$ be two stochastic matrices with the yth column $Vec\{P(y|-)\}$ and $Vec\{\Lambda(y)\}$ respectively for $y = 1, \ldots, d_0$, then we can write the above d_0 identities together as:

$$Mat(P) = \left(\pi^{(1)} \otimes \pi^{(2)} \otimes \cdots \otimes \pi^{(p)}\right) Mat(\Lambda).$$

Bias-variance trade-off

In tensor factorization model (2), the multirank k controls the sparsity, characterizing the impact of each predictor X_j through the "effective category count" k_j. For example, if the level of X_1, say $1, 2, 3$, can be divided into 2 classes $\{1\}$ and $\{2, 3\}$ such that $P(Y = y|X_1 = 2, \ldots, X_p = x_p) \equiv P(Y = y|X_1 = 3, \ldots, X_p = x_p)$, then k_1 is equal to 2. The following illustration suggests that to select k, we can use a hard clustering approximation by setting $\pi^{(j)}_{h_j}(x_j)$ to be either zero or one (section 4.2).

We initially provide a heuristic argument to demonstrate the tendency of our model to produce low mean squared error (MSE), which is defined as:

$$\begin{aligned}
\mathrm{MSE}(\tilde{P}) =& \int \sum_{y=1}^{d_0} E\big(\tilde{P}(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big)^2 G(dx_1, \ldots, dx_p) \\
=& \int \sum_{y=1}^{d_0} \big(E\tilde{P}(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big)^2 G(dx_1, \ldots, dx_p) \\
&+ \int \sum_{y=1}^{d_0} Var\tilde{P}(y|x_1, \ldots, x_p) G(dx_1, \ldots, dx_p) \\
\triangleq& \mathrm{Bias}^2(\tilde{P}) + \mathrm{Var}(\tilde{P}), \qquad \text{(A.2)}
\end{aligned}$$

where $\tilde{P}$ is an estimator of the truth P_0, G is the joint marginal distribution of the covariates X and the expectation is taken with respect to the joint distribution of (X, Y). Our focus is on obtaining accurate estimates of the conditional probability $P(Y|X)$; accurate estimates will lead to accurate classification while containing information on classification uncertainty, of critical importance in medical decision making among other areas.

For simplicity of exposition, assume the response Y to be binary. Denote by $\mathcal{T}$ the set of all conditional probability tensors parameterized by (2). Let $\mathcal{T}_0$ be a subset of $\mathcal{T}$ consisting of models with $\pi_{h_j}^{(j)}(x_j)$ being either zero or one. Then given k and π, $\pi^{(j)}$ uniquely determines a hard clustering of X_j: $X_j = x_j$ belongs to the $h_j(x_j)$th cluster, where $h_j(x_j)$ is the unique h_j such that $\pi_{h_j}^{(j)}(x_j) = 1$. Consider approximating P_0 by this subset $\mathcal{T}_0$. Intuitively, the best MSE attained within $\mathcal{T}_0$ gives an upper bound on the optimal MSE achievable by the whole model class $\mathcal{T}$. To demonstrate the bias-variance trade-off in terms of the selection of the multirank k, we compare the MSE of the maximum likelihood estimators (MLE) in model space $\mathcal{T}_0$ under different k and the clustering scheme determined by π. Define

$$\epsilon_M = \inf_{P \in \mathcal{T}_0 : |k(P)| \leq M} ||P - P_0||,$$

where $|k(P)|$ denotes the size of the multirank of the conditional probability tensor P and

$$||P - P_0|| = \left\{ \int \sum_{y=1}^{2} |P(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)|^2 G(dx_1, \ldots, dx_p) \right\}^{1/2}. \tag{A.3}$$

ϵ_M can be interpreted as the smallest error or bias caused by approximating P_0 using $P \in \mathcal{T}_0$ with size $|k(P)| \leq M$, related to compressibility of P_0.

Under degeneracy of the π's, $P(y|x_1, \ldots, x_p) = \lambda_{h_1(x_1)\ldots h_p(x_p)}(y)$, where $h_j(x_j)$ is defined previously as the unique h_j such that $\pi_{h_j}^{(j)}(x_j) = 1$. Given k and π, the MLE of $\lambda_{h_1\ldots h_p}$ is the sample frequencies of $Y_i = y$ among all observations with covariates $X_i = (X_{i1}, \ldots, X_{ip})$ satisfying $h_j(X_{ij}) = h_j$ for each $j = 1, \ldots, p$:

$$\hat{\lambda}_{h_1\ldots h_p}(i) = \frac{\sum_{(x_1,\ldots,x_p):h_j(x_j)=h_j} \sum_{i=1}^{n} I(X_{i1} = x_1, \ldots, X_{ip} = x_p, Y_i = i)}{\sum_{(x_1,\ldots,x_p):h_j(x_j)=h_j} \sum_{i=1}^{n} I(X_{i1} = x_1, \ldots, X_{ip} = x_p)}, \; i = 1, 2,$$

where 0/0 is defined to be 0 for simplicity. Although given k and π an unbiased estimator does not exist due to model misspecification, the following lemma shows that this MLE is still optimal in terms of minimizing the bias. A proof is sketched in the appendix.

Lemma 1 *Given k and π, among all estimators of λ's, the MLE defined above minimizes the $Bias^2(\tilde{P})$ in* (A.2).

This lemma indicates that the ϵ_M has another characterization as

$$\epsilon_M = \min_{(k,\pi):|k|\leq M,\pi \text{ degenerate}} \text{Bias}\big(\hat{P}(k,\pi)\big),$$

where $\hat{P}(k,\pi)$ is the MLE of P given (k,π).

Intuitively, under the degeneracy of π, n samples are separated into $|k|$ clusters to estimate the corresponding λ's, and the variance term in (A.2) should be of order $|k|/n$. The following lemma formalizes this and a proof is sketched in the appendix.

Lemma 2 *Given k and π, the $Var(\tilde{P})$ as defined in* (A.2) *for the MLE $\hat{P}$ satisfies*

$$Var\big(\hat{P}(k,\pi)\big) = C|k|/n + O(|k|/n^2), \tag{A.4}$$

where the constant $C \in [a,b]$, where $a,b>0$ only depends on P_0 and G.

Combining Lemma 1 and 2, given k and π, the MSE of MLE $\hat{P}$ satisfies:

$$\text{MSE}\big(\hat{P}(k,\pi)\big) \geq \epsilon_{|k|}^2 + C\frac{|k|}{n} + O(|k|/n^2).$$

This reflects the so-called bias-variance trade-off for our model: as $|k|$ increases, the model becomes more complex and thus the bias term decreases; however, the variance term increases as more parameters are introduced. Therefore, there exists an optimal model size $|k|$ that solves $|k| = n\epsilon_{|k|}^2$ minimizing the MSE. This typical trade-off also appears in the Assumption B in section 3.2 where the posterior convergence rate is studied.

Borrowing of information

The previous section discussed the bias-variance trade-off for a subclass of models specified by (2), where π's are degenerate at zero and one. In this section, we illustrate another desirable property by allowing π's to be continuous on $[0, 1]$: borrowing of information across cells corresponding to each combination of $X_1, \ldots, X_p$. Letting $w_{h_1,\ldots,h_p}(x_1, \ldots, x_p) = \prod_j \pi_{h_j}^{(j)}(x_j)$, model (2) is equivalent to

$$P(Y = y|X_1 = x_1, \ldots, X_p = x_p) = \sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1, \ldots, x_p)\lambda_{h_1\ldots h_p}(y), \tag{A.5}$$

and constraints (3) imply $\sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1, \ldots, x_p) = 1$. In the special case when π is degenerate, $\lambda_{h_1\ldots h_p}(y)$ is just the conditional probability of $Y = y$ given the observations in cluster $h_1(X_1) = h_1, \ldots, h_p(X_p) = h_p$ (for details, refer to the descriptions in the paragraph before (A.3)). If π's are allowed to be continuous, then our model essentially uses a kernel estimate that allows borrowing of information across clusters via a weighted average of the cluster frequencies.

To illustrate the strength of this, consider a simplified example involving one covariate X with m categories and a binary response Y. In fact, each category of X can correspond to a cluster as in the preceding paragraph and the implications can be extended to our model by changing the notations. Let $P_j = P(Y = 1|X = j)$ for $j = 1, \ldots, m$. Then the MLE for $(P_1, \ldots, P_m)$ is sample frequencies $(s_1/n_1, \ldots, s_m/n_m)$, denoted by $(\hat{P}_1, \ldots, \hat{P}_m)$, where $s_j = \sharp\{i : y_i = 1 \text{ and } x_i = j\}$ and $n_j = \sharp\{i : x_i = j\}$. Instead, kernel estimates (A.5) are

$$\tilde{P}_k = \left\{1 - \sum_{j \neq k} w_{jk}\right\}\hat{P}_k + \sum_{j \neq k} w_{jk}\hat{P}_j, \ k = 1, \ldots, m,$$

where w_{jk} could be considered as the weight of the contribution to cluster k by cluster j. MLE corresponds to a special case when $w_{jk} = 0$ for all $j \neq k$. We use squared loss to compare these two estimators. After some calculations,

$$E\{L(\hat{P}, P)\} = \sum_{j=1}^{m} E(\hat{P}_j - P_j)^2 = \sum_{j=1}^{m} \frac{P_j(1-P_j)}{n_j},$$

and $E\{L(\tilde{P}, P)\} = \sum_{j=1}^{m} E(\tilde{P}_j - P_j)^2$ is a function of w_{jk}'s, whose partial derivative with respect to $w_{jk}(j \neq k)$ at zero is

$$\left.\frac{\partial E\{L(\tilde{P}, P)\}}{\partial w_{jk}}\right|_{w_{st}=0, \forall s \neq t} = -2\frac{P_k(1-P_k)}{n_k}.$$

This implies that $E\{L(\tilde{P}, P)\}$ will be reduced by $2\frac{P_k(1-P_k)}{n_k}$ for every unit increasing of w_{jk} near zero. Particularly when n_k is small, borrowing information from other cluster $j(\neq k)$ will considerably reduce $E\{L(\tilde{P}, P)\}$ compare to MLE. In the special case when all w_{jk} are equal, $E\{L(\tilde{P}, P)\}$ can attain a minimum

$$\begin{aligned} E\{L(\hat{P}, P)\}\left[1 - \left(1 - \frac{1}{m}\right)\frac{E\{L(\hat{P}, P)\}}{E\{L(\hat{P}, P)\} + \frac{1}{m-1}\sum_{i<j}(P_i - P_j)^2}\right] \\ \in \left(\frac{1}{m}E\{L(\hat{P}, P)\}, E\{L(\hat{P}, P)\}\right). \end{aligned}$$

This suggests that when P_j's are similar, the estimate $\tilde{P}$ can reduce the risk up to only $1/m$ the risk of estimating $\hat{P}$ separately. If P_j's are not similar, $\tilde{P}$ can still reduce the risk considerably when the cell counts $\{n_j\}$ are small.

Another interesting feature of our tensor model is the special structure of the weights w's in (A.5). Consider a class of continuous $\tilde{\pi}$'s indexed by a single parameter $c \in (0, 1)$

characterizing the strength of borrowing information,

$$\tilde{\pi}_{h_j}^{(j)}(x_j) = (1 - k_j c) I\big\{h_j = h_j(x_j)\big\} + cI\big\{h_j \neq h_j(x_j)\big\},$$

for $h_j \leq k_j$ and all possible x_j's. This $\tilde{\pi}$ still satisfies constraint (A.12) and the weight becomes

$$\tilde{w}_{h_1,\ldots,h_p}(x_1, \ldots, x_p) = \prod_{j=1}^{p} (1 - k_j c)^{I\{h_j = h_j(x_j)\}} c^{I\{h_j \neq h_j(x_j)\}}.$$

When c is small, given x, the weight of the contribution by the cluster indexed by $(h_1, \ldots, h_p)$ is approximately equal to c^s, where $s = \sum_{j=1}^{p} I\{h_j \neq h_j(x_j)\}$ is the number of latent classes not shared by $(h_1, \ldots, h_p)$ and $(h_1(x_1), \ldots, h_p(x_p))$, i.e. the Hamming distances between the latent class indices. This special structure in the weights suggests that similar clusters should share more information.

Supplementary Appendix B: Proof of Lemma 1

Given the degeneracy of π, the bias square term can be written as

$$\mathrm{Bias}^2 = \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1 \ldots h_p}} \big(E\tilde{\lambda}_{h_1 \ldots h_p}(y) - P_0(y|x_1, \ldots, x_p)\big)^2 G(dx_1, \ldots, dx_p),$$

where $A_{h_1 \ldots h_p} = \{(x_1, \ldots, x_p) : h_j(x_j) = h_j, j = 1, \ldots, p\}$ and $\tilde{\lambda}$'s are arbitrary estimators of λ's. It can be verified that the above expression is minimized if and only if:

$$E\tilde{\lambda}_{h_1 \ldots h_p}(y) = \frac{\int_{A_{h_1 \ldots h_p}} P_0(y|x_1, \ldots, x_p) G(dx_1, \ldots, dx_p)}{\int_{A_{h_1 \ldots h_p}} G(dx_1, \ldots, dx_p)} \tag{A.6}$$

holds for all possible $(h_1, \ldots, h_p)$. So we only need to check the the MLE $\hat{\lambda}$'s satisfy this condition.

Let $N_{x_1,\ldots,x_p} = \sum_{i=1}^n I(X_{i1} = x_1, \ldots, X_{ip} = x_p)$, $\bar{N}_{h_1,\ldots,h_p} = \sum_{A_{h_1\ldots h_p}} N_{x_1,\ldots,x_p}$, $X = \{X_1, \ldots, X_p\}$ and $Y = \{Y_1, \ldots, Y_p\}$. By the iterative expectation formula:

$$E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}(y) = \sum_{A_{h_1\ldots h_p}} E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} P_0(y|x_1, \ldots, x_p). \tag{A.7}$$

Note that

$$N_{x_1,\ldots,x_p}\Big|\bar{N}_{h_1,\ldots,h_p} \sim \mathrm{Bin}\left(\bar{N}_{h_1,\ldots,h_p}, \frac{G(x_1, \ldots, x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1, \ldots, dx_p)}\right). \tag{A.8}$$

Combining this and the iterative expectation formula:

$$E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} P_0(y|x_1, \ldots, x_p) = \frac{G(x_1, \ldots, x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1, \ldots, dx_p)} P_0(y|x_1, \ldots, x_p). \tag{A.9}$$

Combining (A.7) and (A.9) together, we can prove that (A.6) holds for the MLE $\hat{\lambda}$.

Supplementary Appendix C: Proof of Lemma 2

Under the same notation as in Lemma 1,

$$\begin{aligned}
\mathrm{Var} =& \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_{X,Y}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1, \ldots, dx_p) \\
=& \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_X Var_{Y|X}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{Y|X}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1, \ldots, dx_p) \\
&+ \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_X\big(E_{Y|X}\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1, \ldots, dx_p) \\
\triangleq& S_1 + S_2,
\end{aligned}$$

where $E_{Y|X}$ and $Var_{Y|X}$ stand for taking conditional expectation and variance given X, respectively.

Estimation of S_1: First, we estimate the integrand in S_1 similar to (A.7):

$$\begin{aligned}
&E_X Var_{Y|X}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{Y|X}\hat{\lambda}_{h_1\ldots h_p}\big)^2 \\
=&\sum_{A_{h_1\ldots h_p}} E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}^2_{h_1,\ldots,h_p}} P_0(y|x_1,\ldots,x_p)\big(1-P_0(y|x_1,\ldots,x_p)\big) \\
=&\frac{\int_{A_{h_1\ldots h_p}} P_0(y|x_1,\ldots,x_p)\big(1-P_0(y|x_1,\ldots,x_p)\big)G(dx_1,\ldots,dx_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)} E_X \frac{I(\bar{N}_{h_1,\ldots,h_p}>0)}{\bar{N}_{h_1,\ldots,h_p}},
\end{aligned}$$

where the last step is by (A.8) and the iterative expectation formula. Since $\bar{N}_{h_1,\ldots,h_p} \sim \mathrm{Bin}\big(n, \int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)\big)$, by the asymptotic expansion for the expectation of reciprocal of Binomial random variables in Stephan (1945),

$$E_X \frac{I(\bar{N}_{h_1,\ldots,h_p}>0)}{\bar{N}_{h_1,\ldots,h_p}} = \frac{1}{n\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)} + O(n^{-2}), \tag{A.10}$$

we obtain

$$S_1 = C_1 \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} (1/n + o(n^{-2})) = 2C_1|k|/n + O(|k|/n^2),$$

where C_1 is some constant with lower and upper bounds independent of n.

Estimation of S_2: By (A.9), the integrand in S_2 is:

$$\begin{aligned}
&E_X\big(E_{Y|X}\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 \\
=&E_X\Bigg(\sum_{A_{h_1\ldots h_p}}\bigg(\frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} - \frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)}\bigg)P_0(y|x_1,\ldots,x_p)\Bigg)^2.
\end{aligned}$$

Similar to (A.8), the joint conditional distribution of $N_{x_1,\ldots,x_p}$ given $\bar{N}_{h_1,\ldots,h_p}$ follows a multinomial distribution:

$$\{N_{x_1,\ldots,x_p}:(x_1,\ldots,x_p)\in A_{h_1\ldots h_p}\}\big|\bar N_{h_1,\ldots,h_p}$$
$$\sim \mathrm{Multi}\bigg(\bar N_{h_1,\ldots,h_p},\bigg\{\frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1\ldots h_p}}G(dx_1,\ldots,dx_p)}:(x_1,\ldots,x_p)\in A_{h_1\ldots h_p}\bigg\}\bigg).$$

As a result, by the iterative expectation formula, $E_X\big(E_{Y|X}\hat\lambda_{h_1\ldots h_p}-E_{X,Y}\hat\lambda_{h_1\ldots h_p}\big)^2$ is also proportional to $E_X\frac{I(\bar N_{h_1,\ldots,h_p}>0)}{\bar N_{h_1,\ldots,h_p}}$. Therefore, by (A.10)

$$S_2=C_2\sum_{y=1}^{2}\sum_{h_1,\ldots,h_p}(1/n+o(n^{-2}))=2C_2|k|/n+O(|k|/n^2),$$

where C_2 is some constant with lower and upper bounds independent of n.

Combining the estimation of S_1 and S_2, we obtain the desired results with $C=2C_1+2C_2$.

Supplementary Appendix D: Proof of Theorem 2

The following two lemmas are needed to prove this theorem. The proof of lemma 1 can be found in Jiang (2006), and the proof of lemma 2 follows the line of Ghosal et al. (2000) and is given here.

Lemma 3 *Let $\mathcal{P}$ be a subset of all probability measures of X, $P_0\in\mathcal{P}$ and d be the total variance distance, then for each $\epsilon>0$ and $n>0$, there exists a test ϕ_n such that*

$$P_0^n\phi_n\le N\bigg(\frac{\epsilon}{4},\mathcal{P},d\bigg)\exp\bigg(-\frac{n}{8}\epsilon^2\bigg),$$
$$\sup_{P\in\mathcal{P}\cap\{P:d(P,P_0)\ge\epsilon\}}P^n(1-\phi_n)\le\exp\bigg(-\frac{n}{8}\epsilon^2\bigg),$$

where P^n is the $n-$fold of P.

Lemma 4 *If $\Pi_n(P:||\log\frac{P}{P_0}||_\infty<\epsilon_n^2)>\exp(-Cn\epsilon_n^2)$, then for any test ϕ_n, the following* *inequality holds:*

$$E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n) \leq$$
$$P_0^n\phi_n + \exp((1+C)n\epsilon_n^2)\Pi_n(\mathcal{P}_n^c) + \exp((1+C)n\epsilon_n^2) \sup_{\mathcal{P}_n \cap \{P : d(P,P_0)\geq\epsilon_n\}} P^n(1-\phi_n).$$

Lemma 2 We can divide the l.h.s. into two pieces

$$\begin{aligned} &E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n) = \\ &\quad E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n)\phi_n \\ &\quad + E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n)(1-\phi_n). \end{aligned} \tag{A.11}$$

The first term satisfies

$$E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n)\phi_n \leq P_0^n\phi_n. \tag{A.12}$$

Next we will estimate the second term. By definition, we have

$$\begin{aligned} &E_{P_0}\Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1,\ldots,X_n)(1-\phi_n) = \\ &\quad E_{P_0}\frac{\int_{d(P,P_0)\geq\epsilon_n}\prod_{i=1}^n \frac{P}{P_0}(X_i)d\Pi_n(P)(1-\phi_n)}{\int\prod_{i=1}^n \frac{P}{P_0}(X_i)d\Pi_n(P)}. \end{aligned} \tag{A.13}$$

Let $K_n = \{P : ||\log\frac{P}{P_0}||_\infty < \epsilon_n^2\}$. Using the condition $\Pi_n(K_n) > \exp(-Cn\epsilon_n^2)$, we have

$$\begin{aligned} \int\prod_{i=1}^n \frac{P}{P_0}(X_i)d\Pi_n(P) &\geq \int_{K_n}\prod_{i=1}^n \frac{P}{P_0}(X_i)d\Pi_n(P) \\ &\geq \Pi_n(K_n)\exp(-n\epsilon_n^2) \geq \exp(-(1+C)n\epsilon_n^2)\ a.s.P_0^n. \end{aligned}$$

By Fubini's theorem and the fact $0 \leq \phi_n \leq 1$

$$\begin{aligned}
&E_{P_0} \int_{d(P,P_0)\geq\epsilon_n} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi_n(P)(1-\phi_n) \\
&\leq \Pi_n(\mathcal{P}_n^c) + \int_{\mathcal{P}_n \cap \{P: d(P,P_0)\geq\epsilon_n\}} P^n(1-\phi_n) d\Pi_n(P) \\
&\leq \Pi_n(\mathcal{P}_n^c) + \sup_{\mathcal{P}_n \cap \{P: d(P,P_0)\geq\epsilon_n\}} P^n(1-\phi_n).
\end{aligned}$$

Combining the above assertions and equation (A.13), we can see that

$$\begin{aligned}
&E_{P_0} \Pi_n(P : d(P,P_0) \geq \epsilon_n | X_1, \ldots, X_n)(1-\phi_n) \\
&\leq \exp((1+C)n\epsilon_n^2) E_{P_0} \int_{d(P,P_0)\geq\epsilon_n} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi_n(P)(1-\phi_n) \\
&\leq \exp((1+C)n\epsilon_n^2)\Pi_n(\mathcal{P}_n^c) + \exp((1+C)n\epsilon_n^2) \sup_{\mathcal{P}_n \cap \{P: d(P,P_0)\geq\epsilon_n\}} P^n(1-\phi_n).
\end{aligned} \tag{A.14}$$

Combining (A.11), (A.12) and (A.14) will lead to the conclusion.

Theorem 4 in the paper Let the test in the lemma 2 to be the test ϕ_n defined in lemma 1 with $\epsilon = M\epsilon_n$ and $M^2 > 16 + 8C$. Using the condition (a), (b) in the Theorem 4, we have

$$\begin{aligned}
&E_{P_0} \Pi_n(P : d(P,P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \leq \\
&\quad \exp(-n\epsilon_n^2) + \exp(-n\epsilon_n^2) + \exp(-n\epsilon_n^2) = 3\exp(-n\epsilon_n^2).
\end{aligned}$$

So

$$E_{P_0} \sum_n \Pi_n(P : d(P,P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \leq 3 \sum_n \exp(-n\epsilon_n^2) < \infty.$$

Thus we have

$$\sum_n \Pi_n(P : d(P, P_0) \geq M\epsilon_n | X_1, \ldots, X_n) < \infty \; a.s. P_0^n,$$

and

$$\Pi_n(P : d(P, P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \to 0 \; a.s. P_0^n.$$

References

Ghosal, H., J. K. Ghosh, and A. W. Van Der Vaart (2000). Convergence rates of posterior distributions. *Ann. Statist. 28*, 500–531.

Jiang, W. (2006). Bayesian variable selection for high dimensional generalized linear models. *Ann. Statist. 35*, 1487–1511.

Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. *Ann. Math. Statist. 16*, 50–61.