Statistics and Its Interface Volume 9 (2016) 355–363

A choice model with a diverging choice set for POI data analysis

Xiaoling Lu[*,†], Junlong Zhao[‡], Yu Chen[§], and Hansheng Wang[§]

A point of interest (POI) is a geographical location, that might carry interest for the public. A POI provides a convenient way to register people's locations through mobile devices, which leads to POI data. POI data contain accurate location information and are extremely valuable for location based services (LBS). Accordingly, principled statistical methods, which can be used for regression and/or prediction are required. To partially fulfill this theoretical gap, we propose a conditional logit approach for POI choice analysis. This new model is a natural extension of the classical choice model (McFadden, 1974, 1978) but with two key characteristics. First, POIs located far away from the current position are less likely to be selected as the next POI choice. As a result, the distance (or its appropriate transformation) between the current position and the next POI candidate is an important predictor and should be included in the model. Second, the classical choice model considers a finite choice set. By contrast, the new model studies a diverging choice set, mainly because the total number of POI locations in practice is typically large. The diverging choice set produces an expensive computation of the maximum likelihood estimation (MLE). To alleviate computational costs, we further propose a constrained maximum likelihood estimation (CMLE) method. Compared with MLE, CMLE utilizes only those POIs located within a reasonable distance. This prioritization leads to a significant reduction in computation at a reasonable efficiency loss. To demonstrate the finite sample performance of the method, numerical studies based on both simulated and real datasets are presented.

Keywords and phrases: Choice model, Constrained maximum likelihood estimation, Diverging choice set, Location based service, Point of interest.

1. INTRODUCTION

With the rapid development of wireless communication and positioning technology, massive amounts of human movement data have been collected. Mining and understanding such data has gained substantial attention recently. Gonzalez et al. (2008), Song et al. (2010) and Yan et al. (2013) studied the basic laws of human mobility patterns. The results are important for urban planning and traffic forecasting. Zheng et al. (2009) proposed a tree-based hierarchical graph model to mine interesting locations and travel sequences from GPS trajectories. Li et al. (2010) proposed a two-stage algorithm to address the problem of mining periodic behaviors of moving objects. Yuan et al. (2012) used human mobility and points of interest (POI) data to discover functional regions in a city. Under a discrete choice formulation, Kumar et al. (2015) studied the dynamics of geographic choice. They applied the model to study restaurant choices in map search logs and showed that a four-parameter model based on combinations of lognormals displayed an excellent performance.

In this paper, we focus on analyzing POI data. A point of interest is a geographical location, that might carry interest for the public. Typical examples include universities, hospitals, gas stations, and airports. A POI provides a convenient way to register location of people. This registry is typically performed through mobile devices including GPS devices and smart phones. An example is given in Figure 1. The text message contained in the top rectangle box was posted by the last author of this work on Sina Weibo, which is the largest Twitter-type social media service in China. This message was posted using a smart phone. At the time the message was posted, the geographical location of the author was detected using the GPS system in the smart phone and then recorded in terms of longitude and latitude values. With the permission of the author, this information can be publicly shared on the social network. However, directly sharing this information produces an unsatisfactory user experience. Most people, including the author are not aware of the relationship between the actual location and the longitude-latitude values. To address this problem, Sina Weibo registered the location of the author to the nearest POI, which was "Peking University", given in the circled box in Figure 1. The author's followers know immediately that this message was posted when the author was near "Peking University". Although this location is not fully accurate, the reporting of this location produces a much improved user experience, because "Peking University" (as a POI) displays excellent public awareness.

The above illustration briefly explains the POI technique. This technique has been extensively implemented in mobile

*Corresponding author.
†Center for Applied Statistics, Data Mining Center, School of Statistics, Renmin University of China.
‡School of Mathematics and Systems Science, Beihang University.
§Guanghua School of Management, Peking University.

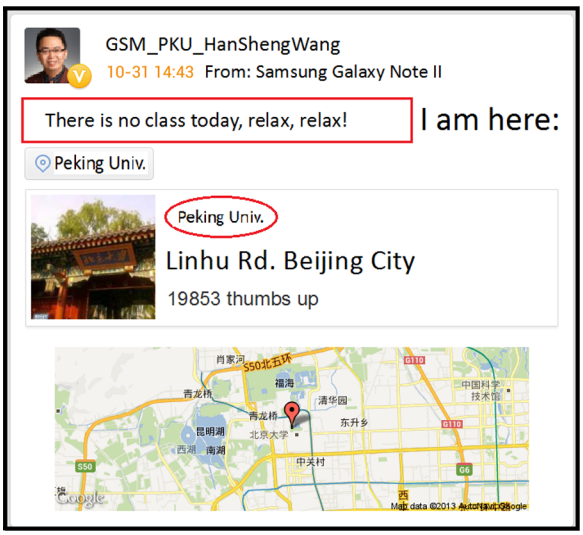

Figure 1. The posted text message on Sina Weibo.

internet related apps, with a particular focus on location based services (LBS). These apps include FourSquare, Facebook, Twitter, QQ, and WeChat. The widespread use of LBS apps increases the availability of POI data. With the help of POI data, one can easily infer people's general locations for shopping, dining, traveling, and etc. Thereafter, a new product/service can be designed and appropriate marketing measures can be taken. Thus, the value produced by POI data should be significant. Unfortunately, this value is not achieved in practice. One primary hurdle is the lack of a principled statistical model, that can be used for regression and/or prediction. This motivated us to develop this work to partially fulfill this important theoretical gap.

We present a study of one particular problem: POI choice. In other words, given a set of predictors and the current position of a user, we develop a process to predict the next POI choice. Let $\mathcal{P} = \{1, \cdots, K\}$ be a set collecting all possible POI candidates. Given the current position of a user and a set of predictive variables, the next POI is then selected from $\mathcal{P}$. Thus, the classical choice model (McFadden, 1974, 1978) can be considered subject to some challenges. The key challenge faced here is that the size of the POI choice set (denoted by $|\mathcal{P}|$) in practice is large. In many cases, the choice set could be comparable or even larger than the sample size. Therefore, directly applying the classical choice model is computationally challenging. However, POI data provides valuable location information. Accordingly, the distance (or its appropriate transformation) between the current position and the next POI candidate can be calculated. Intuitively, those POIs located far away from the current position are less likely to be selected. Thus, the actual choice set of a user likely contains only those POIs within a reasonable distance. This constraint produces a smaller choice set, that dynamically changes according to the current position of the user. Additionally, the constrained choice set could be much smaller than $|\mathcal{P}|$ in size and the computation should therefore be much easier.

Motivated by the above observations, we propose a conditional logit model for the POI choice analysis. This new model is a natural extension of the classical choice model (McFadden, 1974, 1978) but with two key features. First, as we discussed before, POIs located far away from the current position are less likely to be selected as the next POI. As a result, the distance (or its appropriate transformation) between the current position and the next POI candidate is an important predictor and should be included in the choice model. Second, the classical choice models consider finite choice sets. By contrast, the new model considers a diverging choice set. This consideration is mainly because the total number of POI locations (i.e., $|\mathcal{P}|$) in practice is large. The diverging choice set causes the computation of the maximum likelihood estimation (MLE) to be expensive. To alleviate the computational cost, we further propose a constrained maximum likelihood estimation (CMLE) method. Compared with the MLE, the CMLE utilizes only those POIs located within a reasonable distance, developing a significant reduction in computation.

We contribute the following to the literature. First, to the best of our knowledge, this investigation is one of the first studies to formally recognize the fundamental importance of POI choice data to both theory and practice. In terms of the theory, POI data develop a choice model with a diverging choice set. In terms of the application, POI data are likely one of the most important data types for the LBS related industry. Second, a choice model is developed by allowing the size of the choice set to be large. Furthermore, to alleviate the computational cost, a novel CMLE method is proposed. This CMLE differentiates our approach from the existing methods (McCullagh and Nelder, 1989).

The reminder of the article is organized as follows. The next section introduces the model and notations. Both the MLE and CMLE are also introduced. The corresponding asymptotic theories are developed. Numerical studies are presented in Section 3, including a simulation study and a real dataset analysis of Sina Weibo. This article is concluded with a brief discussion in Section 4.

2. METHODOLOGY

2.1 Maximum likelihood estimation

Let $1 \le i \le n$ denote a total of n different subjects, $Y_{0i} \in \mathcal{P} = \{1, \cdots, K\}$ be the current position, and $Y_i \in \mathcal{P}$ be the next POI choice. Because the total number of POI locations is large, we assume that $K \to \infty$ as $n \to \infty$. Let $X_{ik} = (X_{ik1}, \cdots, X_{ikp})^\top \in \mathbb{R}^p$ be the predictor associated with the ith subject and the kth POI candidate, and $\mathbb{X}_i = \{X_{ik} : 1 \le k \le K\}$ collect all the covariate information associated with i. In practice, X_{ik} is typically constructed by interacting the subject-specific (i.e., i-specific) information with that of the location (i.e., k-specific). Let d_{ik} be the distance between the current position and the kth POI candidate for the ith

subject. Depending on the real application, d_{ik} might be the original distance or its appropriate transformation (e.g., log-transformation). We then follow the approach by McFadden (1974, 1978) and assume the following choice model

$$\begin{aligned} P(Y_i = k|\mathbb{X}_i, Y_{0i}) &= p_{ik} \\ &= \exp(\alpha d_{ik} + \beta^\top X_{ik})\Big[\sum_{k'=1}^{K} \exp(\alpha d_{ik'} + \beta^\top X_{ik'})\Big]^{-1} \\ &= v_{ik}(\theta)\Big[\sum_{k'=1}^{K} v_{ik'}(\theta)\Big]^{-1}, \end{aligned} \tag{1}$$

where $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ is the unknown regression coefficient with a true value of $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})^\top \in \mathbb{R}^p$, $\alpha \in \mathbb{R}^1$ is the unknown scalar with a true value of α_0, $\theta = (\alpha, \beta^\top)^\top \in \mathbb{R}^{p+1}$ with a true value of $\theta_0 = (\alpha_0, \beta_0^\top)^\top \in \mathbb{R}^{p+1}$, and $v_{ik}(\theta) = \exp(\alpha d_{ik} + \beta^\top X_{ik})$. Because the POIs with larger distances are less likely to be selected next, we should expect $\alpha_0 < 0$. However, such a constraint is theoretically not required.

Define $Z_i = (Z_{i1}, \cdots, Z_{iK})^\top \in \mathbb{R}^K$ with $Z_{ik} = I(Y_i = k)$. The likelihood of Y_i is then given by

$$\left\{\prod_{k=1}^{K} v_{ik}(\theta)^{Z_{ik}}\right\}\left[\sum_{k=1}^{K} v_{ik}(\theta)\right]^{-1},$$

leading to the following log likelihood function

$$\ell_i(\theta) = \sum_{k=1}^{K} Z_{ik} \log v_{ik}(\theta) - \log \sum_{k=1}^{K} v_{ik}(\theta). \tag{2}$$

Accordingly, the full log likelihood function is the following:

$$\begin{aligned} \ell(\theta) &= \sum_i \ell_i(\theta) \\ &= \sum_i \left\{ \sum_k Z_{ik}\theta^\top \bar{X}_{ik} + \log \sum_k \exp(\theta^\top \bar{X}_{ik}) \right\}, \end{aligned}$$

where $\bar{X}_{ik} = (d_{ik}, X_{ik}^\top)^\top \in \mathbb{R}^{p+1}$. This function leads to the maximum likelihood estimator: $\hat{\theta} = \text{argmax}\ell(\theta)$.

Let $\dot{\ell}(\theta) \in \mathbb{R}^{p+1}$ and $\ddot{\ell}(\theta) \in \mathbb{R}^{(p+1)\times(p+1)}$ be the first and second order derivatives of $\ell(\theta)$, respectively. The following can be written:

$$\begin{aligned} \dot{\ell}(\theta) &= \sum_i \dot{\ell}_i(\theta) = \sum_i \sum_k (Z_{ik} - p_{ik})\bar{X}_{ik}, \\ \ddot{\ell}(\theta) &= \sum_i \ddot{\ell}_i(\theta) = \sum_i \hat{\Sigma}_{iK}, \quad \text{and} \\ \hat{\Sigma}_{iK} &= \sum_k p_{ik}\bar{X}_{ik}\bar{X}_{ik}^\top - \Big(\sum_k p_{ik}\bar{X}_{ik}\Big)\Big(\sum_k p_{ik}\bar{X}_{ik}\Big)^\top. \end{aligned}$$

Therefore, $E\{\dot{\ell}(\theta_0)\} = 0$. Define $n^{-1}E\{\ddot{\ell}(\theta_0)\} = E(\hat{\Sigma}_{ik}) = \Sigma_K$. Because $K \to \infty$ as $n \to \infty$, we assume further that $\Sigma_K \to \Sigma$ for some positive definite matrix $\Sigma \in \mathbb{R}^{(p+1)\times(p+1)}$.

Theoretically, we assume that different POI positions $\{\xi_1, \cdots, \xi_K\}$ are independently generated according to a probability distribution. Therefore, for two arbitrary different POI positions (e.g., ξ_{k_1} and ξ_{k_2} with $k_1 \neq k_2$), their mutual distance $\tilde{d}_{k_1k_2} = \|\xi_{k_1} - \xi_{k_2}\|$ follows a random variable with mean $\mu = E(\tilde{d}_{k_1k_2})$ and standard deviation $\sigma = (\text{var}(\tilde{d}_{k_1k_2}))^{1/2}$. Without loss of generality we assume that both $|\mu|$ and σ are bounded above.

Because $K \to \infty$, the MLE is more complicated than a typical one. To investigate the property of the MLE, we then have the following technical conditions.

(C1) Asymptotic Framework. $n^{-1}(\log K)^2 = o(1)$.

(C2) Tail Probability. There exists constants $c_0 > 0$ and $K_0 > 0$ such that $P(\|X_{ik} - E(X_{ik})\| > t) \leq c_0 \exp(-t^2/K_0)$ for every $1 \leq k \leq K$ and that $P\left(\sigma^{-1}|\tilde{d}_{k_1k_2} - \mu| > t\right) \leq c_0 \exp(-t/K_0)$, as $t > C \log K$, for some constant C being large.

(C3) Let Y_K, T_K be the population version of Y_i and $(X_{i1}, \cdots, X_{iK})$ for fixed K. There exists an open set Θ_0 of Θ which contains the true parameter θ_0. The conditional density of Y_K given T_K admits all three derivatives and $|\partial^3 f(Y_K; \theta, T_K))/\partial\theta_{j_1}\partial\theta_{j_2}\partial\theta_{j_3}| < M_K(Y_K, T_K)$ for some $M_K(Y_K, T_K)$ with $\sup_K E[M_K(Y_K, T_K)] < \infty$ for all $1 \leq j_1, j_2, j_3 \leq p$.

We can verify that (C2) holds when X_{ik} and the POIs are i.i.d. and follows a multivariate normal. Moreover, the assumption on $\tilde{d}_{k_1k_2}$ in (C2) is obvious when the POIs are located in a bounded region and K is large. (C3) is adapted from the assumption on the third derivative of density commonly used in the literature (Shao, 1997; Fan and Li, 2001). We then have the following theorem.

Theorem 1. *Assume $K \to \infty$ as $n \to \infty$, $\Sigma_K \to \Sigma$ for some positive definite matrix Σ as $K \to \infty$. Under assumptions (C1)–(C3), we then have $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, \Sigma^{-1})$ as $n \to \infty$.*

The proof is given in Appendix B. By Theorem 1, we know that, even with $K \to \infty$, the MLE $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal.

2.2 Constraint maximum likelihood estimation

Although $\hat{\theta}$ as a MLE is theoretically attractive, its practical computation is not easy. This complexity is mainly because the size of the choice set $\mathcal{P}$ (i.e., $|\mathcal{P}| = K$) is large, resulting in expensive computations. However, for those POIs with large d_{ik} values, their corresponding $\exp(\alpha d_{ik} + \beta^\top X_{ik})$ values are close to 0. Therefore, their contribution to the likelihood function is limited. Then, we might consider constraining our efforts to those POIs within a reasonable distance. To this end, define an index set $\mathcal{S}_c = \{i : d_{iY_i} < c\}$ for some constant $c > 0$ to collect these qualified subjects. Given

a subject $i \in \mathcal{S}_c$, the next POI choice Y_i must be selected from the following POI candidates $\mathcal{P}_{ic} = \{k : d_{ik} < c\}$. For an arbitrary $k \in \mathcal{P}_{ic}$, the corresponding likelihood is given by the following:

$$P(Y_i = k | i \in \mathcal{S}_c, \mathcal{P}_{ic}) = q_{ik} = v_{ik}(\theta) \Big[\sum_{k' \in \mathcal{P}_{ic}} v_{ik'}(\theta) \Big]^{-1}.$$

This function leads to the following constrained log likelihood:

$$\ell_c(\theta) = \sum_{i \in \mathcal{S}_c} \ell_{ci}(\theta) = \sum_{1 \le i \le n} \ell_{ci}(\theta) I(i \in \mathcal{S}_c),$$

$$\text{and } \ell_{ci}(\theta) = \sum_{k \in \mathcal{P}_{ic}} Z_{ik} \log v_{ik}(\theta) - \log \sum_{k \in \mathcal{P}_{ic}} v_{ik}(\theta). \tag{3}$$

Accordingly, the constrained maximum likelihood estimator (CMLE) can be computed as $\hat{\theta}_c = \text{argmax}_\theta \ell_c(\theta)$. Compare $\ell_c(\theta)$ against the genuine log likelihood. Two key constraints are noted. The first constraint is imposed by $\mathcal{S}_c$, which selects only those subjects whose next POI choice is sufficiently close to the current position (i.e., $i \in \mathcal{S}_c$). This constraint leads to a reduced sample size $\sum_i I(d_{iY_i} < c)$. The second constraint is imposed on POI choices (i.e., $k \in \mathcal{P}_{ic}$). This leads to reduced number of POI candidates (i.e., $|\mathcal{P}_{ic}|$) for each subject i. Both constraints reduce the computational cost at a cost of efficiency loss.

Let $\dot{\ell}_c(\theta) \in \mathbb{R}^{p+1}$ and $\ddot{\ell}_c(\theta) \in \mathbb{R}^{(p+1)\times(p+1)}$ be the first and second order derivatives of $\ell_c(\theta)$, respectively. Then, the following can be directly verified:

$$\dot{\ell}_c(\theta) = \sum_i \dot{\ell}_{ci}(\theta) I(i \in \mathcal{S}_c) = \sum_{i \in \mathcal{S}_c} \sum_{k \in \mathcal{P}_{ic}} (Z_{ik} - q_{ik}) \bar{X}_{ik},$$

$$-\ddot{\ell}_c(\theta) = -\sum_i \ddot{\ell}_{ci}(\theta) I(i \in \mathcal{S}_c) = \sum_{i \in \mathcal{S}_c} \hat{\Sigma}_{c,iK}(\theta),$$

$$\hat{\Sigma}_{c,iK}(\theta) = \sum_k q_{ik} \bar{X}_{ik} \bar{X}_{ik}^\top - \Big(\sum_k q_{ik} \bar{X}_{ik} \Big) \Big(\sum_k q_{ik} \bar{X}_{ik} \Big)^\top.$$

Therefore, $E\{\dot{\ell}_c(\theta_0)\} = 0$. Define $E(\hat{\Sigma}_{c,iK}(\theta_0)) = \Sigma_{cK}$ and $\pi_K = P(d_{iY_i} < c)$. Then $E\{-n^{-1}\ddot{\ell}_c(\theta_0)\} = \pi_K \Sigma_{cK}$. Because $K \to \infty$ as $n \to \infty$, we assume further that $\pi_K \to \pi_0$ and $\Sigma_{cK} \to \Sigma_c$ for some positive definite matrix $\Sigma_c \in \mathbb{R}^{(p+1)\times(p+1)}$. We then have the following theorem.

Theorem 2. *Assume $K \to \infty$ as $n \to \infty$, $\Sigma_{cK} \to \Sigma_c$ for some positive definite matrix Σ_c as $K \to \infty$. Under (C1)–(C3), we then have $\sqrt{n}(\hat{\theta}_c - \theta) \to_d N(0, \Sigma_c^{-1})$ as $n \to \infty$.*

The proof is is given in Appendix A. By Theorem 2, the CMLE $\hat{\theta}_c$ is also $\sqrt{n}$-consistent and asymptotically normal. However, compared with the MLE $\hat{\theta}$, its asymptotic efficiency is different. $\Sigma - \Sigma_c$ is a semipositive definite matrix. This implies that the CMLE is statistically less efficient than the MLE. However, its computation is simpler.

3. NUMERICAL STUDIES

3.1 A simulation study

We devote this section to evaluate the finite sample performance of the CMLE and MLE methods. Specifically, the sample size is fixed at $n = 200$ and 500. The number of POI choices is fixed at $K = 100$, 200, and 500. For a given (n, K) combination and one particular simulation replication, the longitude and latitude of each POI are randomly generated from a standard normal distribution. Here the original Euclidean distance is used for d_{ik}. The predictor dimension is fixed at $p = 5$ and X_i is generated from a normal distribution with a mean of 0 and a covariance of $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$. Let $\theta_0 = (\alpha_0, \beta_0^\top)^\top = (-2, 2, 1, 0, 0, 0)^\top \in \mathbb{R}^6$. The current position is randomly selected from the K POI choices. The next location is randomly selected from the remaining POIs according to model (1). Subsequently, both the MLE and CMLE are computed. To select the cutoff value for the CMLE, we compute the distances between the current positions of the users and their next POI choices. Let C_τ be the corresponding τth quantile. We set $c = C_\tau$ with $\tau = 50\%$, 80%, and 100%, respectively. Accordingly, the percentage of the sample size used in the CMLE is aprroximately 50%, 80%, and 100% respectively. The CMLE with $c = C_{1.00}$ is different from the MLE. The difference is that the MLE makes use of all the samples (i.e., every $1 \le i \le n$) and also all the POIs (i.e., every $1 \le k \le K$). However, with $c = C_{1.00}$, the CMLE included all the samples, but not necessarily every POI.

The experiment is randomly replicated $M = 200$ times. Let $\hat{\theta}^{(m)} = (\hat{\theta}_j^{(m)} : 1 \le j \le p+1)$ be one particular estimator (e.g., MLE) obtained in the mth simulation replication. We then evaluate the estimation error by the root of the mean squared error as $\text{RMSE} = M^{-1} \sum_m \{\|\hat{\theta}^{(m)} - \theta_0\|^2/(p+1)\}^{1/2}$. Let $\widehat{\text{SE}}_j^{(m)^2}$ be the jth diagonal element of $\hat{\Sigma}^{(m)}$, which is the estimated asymptotic covariance for $\hat{\theta}_j^{(m)}$, according to either Theorem 1 or 2. Moreover, we compute a test statistic as $Z_j^{(m)} = \hat{\theta}_j^{(m)}/\widehat{\text{SE}}_j^{(m)}$ for each simulation replication m and each regression coefficient j. Fix the significance level to be $\alpha = 5\%$. We then reject the null hypothesis of $H_0 : \theta_j = 0$ vs. $H_1 : \theta_j \neq 0$, when $|Z_j^{(m)}| > z_{1-\alpha/2}$, where z_α stands for the αth quantile of a standard normal distribution. Define for each predictor an Empirical Rejection Probability (ERP) as $\text{ERP}_j = M^{-1} \sum_{m=1}^M I(|Z_j^{(m)}| > z_{1-\alpha/2})$. The ERP corresponds to the empirical size and power according to whether $\theta_j = 0$. Define $\mathcal{M}_0 = \{j : \theta_j = 0\}$ and $\mathcal{M}_1 = \{j : \theta_j \neq 0\}$. Summarize the average size and power as $\text{SIZE} = |\mathcal{M}_0|^{-1} \sum_{j \in \mathcal{M}_0} \text{ERP}_j$ and $\text{POWER} = |\mathcal{M}_1|^{-1} \sum_{j \in \mathcal{M}_1} \text{ERP}_j$, respectively. Here $|\mathcal{M}_0|$ and $|\mathcal{M}_1|$ stand for the size of $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. For a given cutoff value C_τ, the number of POIs involved for the ith subject is given by $|\mathcal{P}_{ic}|$, where $\mathcal{P}_{ic} = \{k : d_{ik} < C_\tau\}$. Accordingly, the Average Percentage of the POI (APP) involved by

the CMLE is given by APP $= |\mathcal{S}_c|^{-1} \sum_{i \in \mathcal{S}_c} |\mathcal{P}_{ic}|/K$. Larger APP values result in larger CPU times (CPU). The detailed results are given in Table 1 with standard deviation (sd) for the RMSE, APP and CPU in parentheses.

From Table 1, we draw the following conclusions: (1) The performance of the CMLE with $\tau = 50\%$ is not acceptable, because its RMSE value is substantially higher than that of the MLE. The standard deviation of the RMSE is also much higher than that of the MLE. Consider the case with $n = 500$ and $K = 100$, the RMSE of the CMLE with $\tau = 50\%$ is 0.25 with sd 0.15, which is much larger than 0.07 with a sd 0.03 of the MLE. (2) By contrast, the CMLE with $\tau = 80\%$ is much better. In fact, the performance of the CMLE with $\tau = 100\%$ is almost identical with that of the full scale MLE; see for example when $n = 200$ and $K = 500$, the RMSE of CMLE with $\tau = 100\%$ is 0.11, which is identical with that of the MLE up to two decimal digits. (3) However, the average number of POIs involved by the corresponding CMLE is typically much less than that of the MLE. For example, the APP value is 27.5% for the case with $n = 500$, $K = 200$, and $\tau = 80\%$. Therefore, the CPU time demanded by CMLE is substantially less.

3.2 A real example

We obtained a dataset from Sina Weibo (www.weibo.com), the largest Twitter-type social media service in China. The data contains 2,038 observations collected between January 24th, 2012 and October 16th, 2013. Each observation corresponds to one location transition occuring between two consecutive POIs. We require that the transition from one POI to another must occur within one hour. Otherwise, the dependence of the next POI choice on the current location could be extremely weak. In total, 1,154 unique POIs are involved, the density is given in Figure 2. A descriptive analysis reveals that the median and maximal values of the transitional distances (i.e., the distances between two consecutive POIs) are 1.04 and 61.06 kilometers, respectively. Therefore that $c_{0.50} = 1.04$ and $c_{1.00} = 61.06$. Similarly, we find that $c_{0.80} = 3.32$.

To explain the POI transitional behavior of a user, the following predictors are considered. The first predictor is the log-transformed inter-POI distance (d_{ik}), whose regression coefficient is expected to be negative. The other predictor X_{ik} is defined as the interaction of W_i and V_k. Here, W_i is set of subject-specific variables and V_k is location-specific. Specifically, W_i describes the ith user's gender (M or F) and residence (local resident or tourist). V_k classifies the POIs into the following seven categories: landmarks (LM, e.g., a very high building tower), dinning places (DP, e.g., restaurants, fast foods), shopping centers (SC, e.g., supermarket, shopping malls), transportation centers (TC, e.g., train stations, airports), tourist attractions (TA, e.g., national parks, museums), school areas (SA, e.g., middle schools, universities), and others (OT).

We then apply our method to the dataset. As suggested by the simulation study, the CMLE with $\tau = 80\%$ is considered because of its good balance between computational cost and statistical efficiencies. For a comparison purpose,

Table 1. Detailed Simulation Results based on 200 Replications

		$n = 200$			
K		50%	80%	100%	MLE
100	RMSE	0.47	0.19	0.12	0.12
		(0.31)	(0.08)	(0.04)	0.04)
	APP (%)	22.00	29.60	59.60	100
		(2.90)	(3.40)	(8.50)	(-)
	CPU (second)	0.41	0.66	1.31	2.62
		(0.09)	(0.12)	(0.30)	(0.15)
	SIZE (%)	3.50	5.00	2.80	4.80
	POWER (%)	85.30	100	100	100
200	RMSE	0.40	0.19	0.12	0.11
		(0.25)	(0.09)	(0.04)	0.04)
	APP (%)	21.00	27.90	57.60	100
		(2.70)	(2.90)	(8.10)	(-)
	CPU (second)	0.86	1.15	1.96	3.77 9
		(0.19)	(0.21)	(0.45)	(0.34)
	SIZE (%)	3.00	6.20	7.30	3.70
	POWER (%)	84.70	100	100	100
500	RMSE	0.43	0.17	0.11	0.11
		(0.27)	(0.07)	(0.04)	0.03)
	APP (%)	19.40	25.90	55.60	100
		(2.00)	(2.20)	(7.80)	(-)
	CPU (second)	1.78	2.45	5.39	10.41
		(0.25)	(0.48)	(1.29)	(1.08)
	SIZE (%)	5.00	3.70	4.80	5.50
	POWER (%)	82.30	100	100	100
		$n = 500$			
K		50%	80%	100%	MLE
100	RMSE	0.25	0.12	0.08	0.07
		(0.15)	(0.05)	(0.02)	(0.03)
	APP (%)	22.50	29.50	65.70	100
		(2.40)	(2.80)	(7.90)	(-)
	CPU (second)	1.85	2.68	5.8	9.72
		(0.29)	(0.40)	(1.02)	(0.81)
	SIZE (%)	4.70	6.20	5.30	4.50
	POWER (%)	97.00	100	100	100
200	RMSE	0.26	0.11	0.07	0.07
		(0.16)	(0.05)	(0.03)	(0.02)
	APP (%)	21.00	27.50	62.50	100
		(2.00)	(2.10)	(6.80)	(-)
	CPU (second)	2.80	4.32	10.54	19.89
		(0.34)	(0.50)	(1.57)	(0.66)
	SIZE (%)	5.80	5.00	5.20	4.30
	POWER (%)	97.50	100	100	100
500	RMSE	0.25	0.11	0.07	0.07
		(0.16)	(0.05)	(0.02)	(0.02)
	APP (%)	19.30	25.90	60.00	100
		(1.40)	(1.60)	(6.30)	(-)
	CPU (second)	5.83	9.27	60.72	116.49
		(0.63)	(1.19)	(8.37)	(3.64)
	SIZE (%)	5.20	4.00	5.30	4.70
	POWER (%)	97.50	100	100	100

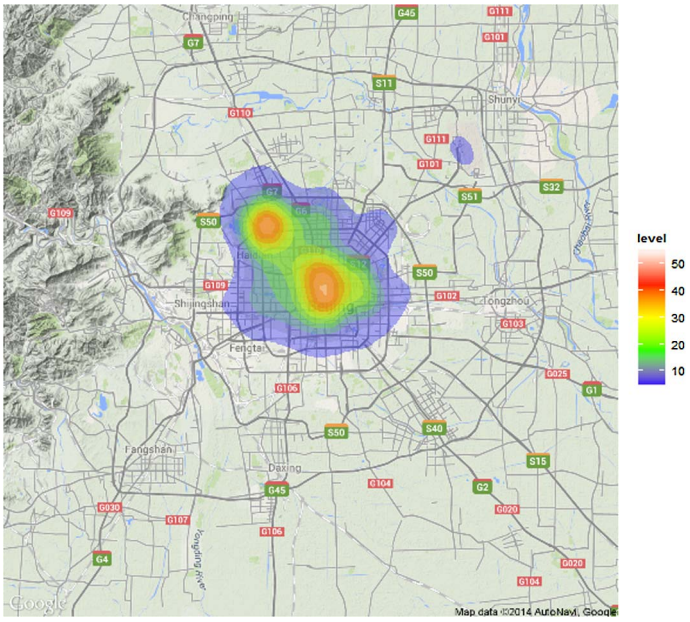

Figure 2. The heat map of POIs in Beijing.

Table 2. The Estimates and Associated SEs for the Real Data Analysis

Effects	80%	ML
M → DP	10.74 (0.56)	13.82 (0.43)
F → DP	12.19 (0.12)	15.29 (0.11)
M → LM	11.98 (0.25)	14.94 (0.18)
F → LM	12.54 (0.09)	15.45 (0.07)
M → SC	11.06 (0.70)	14.38 (0.58)
F → SC	12.08 (0.20)	15.42 (0.17)
M → TC	13.28 (0.38)	14.69 (0.45)
F → TC	12.82 (0.21)	15.39 (0.16)
M → TA	13.15 (0.17)	15.84 (0.14)
F → TA	13.76 (0.06)	16.64 (0.05)
M → SA	11.59 (0.34)	14.85 (0.32)
F → SA	12.47 (0.11)	15.93 (0.09)
M → OT	11.36 (0.35)	13.88 (0.30)
d_{ik}	-2.53 (0.08)	-1.65 (0.04)

we also compute the MLE. The interaction between the residence and POI category is not significant at the 5% level and is thus excluded. The model is refitted and the detailed results are given in Table 2. Here, the category of Female and Others, denoted by F → OT, is treated as the baseline. The analysis results obtained by CMLE with $\tau = 80\%$ and MLE are qualitatively similar. However, in terms of computation time demanded, their difference is significant. The average percentage of POI involved by CMLE with $\tau = 80\%$ is only 11.36% (131.05 out of 1154). The CPU time demanded by CMLE with $\tau = 80\%$ is 1.87 hours on a workstation with Intel(R) Xeon(R) CPU E5_2603 1.80GHz. However, the CPU time of MLE is 22.38 hours.

From Table 2, all of the estimates are significant at the 1% level. The estimate of α is negative, confirming our expectation that people tend to select places within a reasonable distance. To interpret the estimation result, consider the estimates for F → SC. The estimates are 12.08 and 15.42 for CMLE and MLE, respectively. Both results suggested that females are more likely to choose shopping centers as their next POI choice than others (i.e., OT). Other estimates can be interpreted similarly.

To further validate the prediction accuracy of this model, we randomly divide the sample into two equal sized sets, which are denoted as $\mathcal{D}_0$ and $\mathcal{D}_1$. $\mathcal{D}_0$ is used for training and $\mathcal{D}_1$ is used for testing. Subsequently, the CMLE with $\tau = 80\%$ is estimated based on $\mathcal{D}_0$, producing the CMLE estimator. With the CMLE estimator, the next POI choice probabilities can be computed for every $i \in \mathcal{D}_1$ and $1 \le k \le K$. For each testing sample $i \in \mathcal{D}_1$, the POI with the largest choice probability (denoted by $\hat{Y}_i$) is predicted to be the next choice. The resulting forecasting accuracy is computed as $\text{FA} = |\mathcal{D}_1|^{-1} \sum_{i \in \mathcal{D}_1} I(Y_i = \hat{Y}_i)$. For a reliable estimation, the experiment is randomly replicated 20 times and the averaged forecasting accuracy (AFA) is computed. The resulting AFA values is 10.5%. Given that the number of candidate POI choices is more than one thousand, this forecasting accuracy is encouraging.

In real practice, one can typically place at least five advertisements on a mobile device (e.g., a smart phone). Therefore, we can develop at least five different POI predictions for each subject. The resulting prediction accuracy should be further improved to an extent. With the CMLE, we can compute the next POI choice probability for each subject i and each location k. For a given subject i, we can rank different locations according to their choice probabilities. We next use $\hat{Y}_i^{(k)}$ to indicate the location with the kth largest choice probability. Therefore $\hat{Y}_i^{(1)} = \hat{Y}_i$. We then collect the top five POIs with the largest choice probability by $\mathcal{C}_i = \{Y_i^{(k)} : 1 \le k \le 5\}$. We then re-define the forecasting accuracy as $\text{FA} = |\mathcal{D}_1|^{-1} \sum_{i \in \mathcal{D}_1} I(Y_i \in \mathcal{C}_i)$. The resulting AFA = 29.4%, which is satisfactory for industry applications.

4. CONCLUDING REMARKS

The wide spread use of mobile devices increases the availability of location-related data. POI data are likely one of the most important types of data and are critically important for the LBS-related industry. We consider this work to be an initial attempt to address problems related to POI choice analysis. Our work serves as a call to the statistical community to investigate mobile-internet related data analysis. Further research along this direction is required.

ACKNOWLEDGEMENTS

Lu's research was supported in part by the National Natural Science Foundation of China (NSFC 61472475). Zhao's research was supported by the National Natural Science

Foundation of China (NSFC, 11471030, 11101022) and the Fundamental Research Funds for the Central Universities. Wang's research was supported in part by the National Natural Science Foundation of China (NSFC, 11131002, 11271032), Fox Ying Tong Education Foundation, the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University.

APPENDIX A. THE PROOF OF THEOREM 2

One can verify easily that the log likelihood function $\ell(\theta)$ is a strictly convex function. As a result, it can have only one maximizer. Therefore, the theorem conclusion follows, if we can show the existence of at least one local maximizer that satisfies the theorem conclusion.

Recall the notations $\mathcal{P}_{ic} = \{k : d_{ik} < c\}$ and $E\{-\ddot{\ell}_{ic}(\theta_0)\} = \pi_K \Sigma_{cK}$, with $\Sigma_{cK} \to \Sigma_c$. Let $n_c = |\mathcal{S}_c|$. For simplicity of notations, we denote $\hat{\Sigma}_{c,iK}(\theta_0)$ as Q_{Ki} and $\tilde{Q}_{Ki} = I(i \in S_c)\hat{\Sigma}_{c,iK}(\theta_0)$. Furthermore, by definition we have

$$
\begin{aligned}
-\frac{1}{n_c}\frac{\partial^2 \ell_c(\theta_0)}{\partial\theta\partial\theta^\top} &= \frac{n}{n_c}\cdot\frac{1}{n}\sum_{1\le i\le n} I(i\in S_c)\hat{\Sigma}_{c,iK}(\theta_0) \\
&= \frac{n}{n_c}\cdot\frac{1}{n}\sum_{1\le i\le n} I(i\in S_c)\cdot \\
&\quad \Big\{\sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik}\bar{X}_{ik}^\top q_{ik} - \sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik}q_{ik}\sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik}^\top q_{ik}\Big\} \\
&= \frac{n}{n_c}\cdot\frac{1}{n}\sum_{1\le i\le n}\tilde{Q}_{Ki}.
\end{aligned} \tag{4}
$$

It is clear that both $\{Q_{Ki}, i \in S_c, K = 1, 2, \cdots,\}$ and $\{\tilde{Q}_{Ki}, K = 1, 2, \cdots,\}$ are the triangular array. Write $Q_{Ki} = (Q_{Ki,st} : 1 \le s, t \le (p+2)) \in \mathbb{R}^{(p+2)\times(p+2)}$. Subsequently, we are going to establish the theorem conclusion according to the following four steps.

In step 1, we show that $n_c/n \to_p \pi_0$. Note that $E(n_c)/n = \pi_K \to \pi_0$, it is sufficient to show $n_c/n \to_p \pi_K$, where $\pi_K = P(i \in S_c)$. In the 2nd step, we establish the following important conclusion.

$$
n^{-1}\max_{1\le s,t\le p+2} E(\tilde{Q}^2_{Ki,st}) \to 0. \tag{5}
$$

This conclusion is used subsequently to establish the following two conclusions in the 3rd and 4th steps, respectively.

$$
\Big\| -\frac{1}{n_c}\frac{\partial^2 \ell_c(\theta_0)}{\partial\theta\partial\theta^\top} - \Sigma_c\Big\| \to_p 0, \tag{6}
$$

$$
n_c^{-1/2}\frac{\partial \ell_c(\theta_0)}{\partial\theta} \xrightarrow{d} N(0, \Sigma_c), \tag{7}
$$

where $\|A\|^2 = tr(A^\top A)$ for an arbitrary matrix A. Lastly, the asymptotic normality of the MLE $\hat{\theta}$ is established based on (6) and (7).

Step 1. Recall that $\pi_K > 0$. This conclusion is obvious base on the definition of n_c.

Step 2. To prove (5), we write $\bar{X}_{ik} = \{\bar{X}_{ik,j} : 1 \le j \le (p+2)\}^\top \in \mathbb{R}^{p+2}$. Then, $\bar{X}_{ik,1} = d_{ik}$, $\bar{X}_{ik,2} = 1$, and $\bar{X}_{ik,j} = X_{ik,j-2}$ for $j \ge 2$. Next, by definition we have

$$
\begin{aligned}
\tilde{Q}^2_{Ki,st} &= I(i\in S_c)\Big\{\sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik,s}\bar{X}_{ik,t}q_{ik} - \\
&\quad \Big(\sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik,s}q_{ik}\Big)\Big(\sum_{k\in\mathcal{P}_{ic}} \bar{X}_{ik,t}q_{ik}\Big)\Big\}^2 \\
&= I(i\in S_c)\Big[\sum_{k\in\mathcal{P}_{ic}}\Big\{\bar{X}_{ik,s} - \Big(\sum_{k'\in\mathcal{P}_{ic}} \bar{X}_{ik',s}q_{ik}\Big)\Big\} \\
&\quad \Big\{\bar{X}_{ik,s} - \Big(\sum_{k'\in\mathcal{P}_{ic}} \bar{X}_{ik',s}q_{ik}\Big)\Big\}q_{ik}\Big]^2 \\
&\le \tilde{Q}_{Ki,ss}\tilde{Q}_{Ki,tt}.
\end{aligned}
$$

Next note that $0 \le \tilde{Q}_{Ki,ss} \le \sum_{k\in\mathcal{P}_{ic}} \bar{X}^2_{ik,s}q_{ik} \le (\max_{1\le k\le K}|\bar{X}_{ik,s}|)^2 \le (G+\mu_0)^2$, where $\mu_0 = \|E(X_{ik})\| + \mu < \infty$ and

$$
G = \max_{1\le k\le K}\|X_{ik} - E(X_{ik})\| + \max_{1\le k\le K}\|d_{ik} - \mu\| := G_1 + G_2.
$$

Consider G_1 first. We then have

$$
\begin{aligned}
EG_1^4 &\le 4\int_0^\infty t^3 P(G_1 > t)dt \\
&= 4\left(\int_0^{\sqrt{K_0\log K}} + \int_{\sqrt{K_0\log K}}^\infty\right) t^3 P(G_1 > t)dt \\
&\le K_0^2(\log K)^2 + 4\int_{\sqrt{K_0\log K}}^\infty t^3 P(G_1 > t)dt.
\end{aligned} \tag{8}
$$

Next, for an arbitrary fixed constant u, define a constant $\tilde{u}_K = [K_0(u^2+1)\log K]^{1/2}$ with K_0 being the constant defined in (C2). We then have $P(G_1 > \tilde{u}_K) \le c_0 K\exp\{-(u^2+1)\log K\} = c_0\exp(-u^2\log K)$. Let $t = \tilde{u}_K$ and $\sigma = (2\log K)^{-1/2}$. Then apply this back to (8), we have

$$
\begin{aligned}
&\int_{\sqrt{K_0\log K}}^\infty t^3 P(G_1 > t)dt \\
&\le c_0(K_0\log K)^2\int_0^\infty (u^3+u)\exp(-u^2\log K)du \\
&= c_0\sqrt{\pi}K_0^2(\log K)^{\frac{3}{2}}\frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty (u^3+u^2)\exp(-\frac{u^2}{2\sigma^2})du \\
&= (\log K)^{\frac{3}{2}}O(1).
\end{aligned}
$$

Consequently, we have $EG_1^4 \le C_1(\log K)^2$, for some constant C_1. Note that $\mu < \sigma\log K$, as $K \to \infty$. Then, as $t > C\log K$ for some constant $C > 0$, we have by (C2) that

$$
\begin{aligned}
&P\left(\left|\frac{d_{ik}-\mu}{\sigma}\right|>t\right)\\
&=\sum_{k'=1}^{K}P\left(\left|\frac{d_{Y_{0i}k}-\mu}{\sigma}\right|>t|Y_{0i}=k'\right)P(Y_{0i}=k')\\
&=\sum_{k'=1}^{K}P\left(\left|\frac{d_{k'k}-\mu}{\sigma}\right|>t\right)P(Y_{0i}=k')\\
&\le c_0\exp(-t/K_0).
\end{aligned}
$$

Therefore, $\sigma^{-1}(d_{ik}-\mu)$ has the same tail probability as $X_{ik,s}-E(X_{ik,s})$. Let $\tilde{G}_2=\sigma^{-1}G_2$. Then as the argument for G_1, we have $E(\tilde{G}_2^4)\le C_2(\log K)^2$ for some constant C_2. Therefore, $E(G_2^4)\le C_2\sigma^4(\log K)^2$. Combined these argument together, we have $EG^4\le C_3\sigma^4(\log K)^2$ for some constant C_3.

Combined with the fact $\mu_0<\infty$, we have for some constant C_4, $\max_{1\le s,t\le p+2}E(\tilde{Q}^2_{Ki,st})\le\max_{1\le s\le p+2}E(\tilde{Q}^2_{Ki,ss})\le C_4[\sigma^4(\log K)^2+\mu^4]$. This combined with (C1) leads to the (5).

Step 3. We next consider (6). Note that $E(\tilde{Q}_{Ki})=\pi_K\Sigma_{cK}:=\tilde{\Sigma}_{cK}$. Recall that p is fixed. We first show that $n^{-1}\sum_{1\le i\le n}\tilde{Q}_{Ki,st}\to_p\tilde{\Sigma}_{cK,st}$, where $\tilde{\Sigma}_{cK,st}$ is the (s,t) element of $\tilde{\Sigma}_{cK}$. By (5), we have, for any $\epsilon>0$

$$
\begin{aligned}
(9)\qquad &P\left(\left|n^{-1}\sum_{1\le i\le n}\left\{\tilde{Q}_{Ki,st}-\tilde{\Sigma}_{cK,st}\right\}\right|>\epsilon\right)\\
&\le\epsilon^{-2}n^{-1}E(\tilde{Q}^2_{Ki,st})\to 0.
\end{aligned}
$$

Therefore, $n^{-1}\sum_{1\le i\le n}\tilde{Q}_{Ki,st}\to_p\tilde{\Sigma}_{cK,st}$. Recall the $\pi_K\to\pi_0,\Sigma_{cK}\to\Sigma_c$, we have $\tilde{\Sigma}_{cK}\to\pi_0\Sigma_c$. Combining with (4) and the fact that $n/n_c\to_p 1/\pi_0$, the conclusion (6) holds.

Step 4. Now, we consider the proof of (7). Let $R_{cK,i}=\partial\ell_{ci}(\theta_0)/\partial\theta\in\mathbb{R}^{p+2}$ with $i\in\mathcal{S}_c$. Then $\partial\ell_c(\theta_0)/\partial\theta=\sum_{1\le i\le n}R_{cK,i}I(i\in S_c)$. Accordingly,

$$
\begin{aligned}
&n_c^{-1/2}\Sigma_{cK}^{-1/2}\frac{\partial\ell_c(\theta_0)}{\partial\theta}\\
&\quad=(n_c/n)^{-1/2}\cdot n^{-1/2}\sum_{1\le i\le n}\Sigma_{cK}^{-1/2}R_{cK,i}I(i\in S_c)\\
&\quad:=(n_c/n)^{-1/2}\cdot n^{-1/2}\sum_{1\le i\le n}\tilde{R}_{cK,i},
\end{aligned}
$$

where $\tilde{R}_{cK,i}=\Sigma_{cK}^{-1/2}R_{cK,i}I(i\in S_c)$. It is easy to see that $E\left(R_{cK,i}I(i\in S_c)\right)=0$ and $\operatorname{cov}\left(R_{cK,i}I(i\in S_c)\right)=\pi_K\Sigma_{cK}=\tilde{\Sigma}_{cK}$. Consequently, $E(\tilde{R}_{cK,i})=0$ and $\operatorname{cov}\left(\tilde{R}_{cK,i}\right)=\pi_K I_{p+2}$, where I_{p+2} is the identity matrix and $\pi_K=P(i\in S_c)$. Note that $\{\tilde{R}_{cK,i},i\in S_c,K\ge 1\}$ is also triangular array with both $n_c\to\infty$ and $K\to\infty$. It is then sufficient to show that the following Linderberg condition

(10)
$$
\lim_{K\to\infty}n^{-1}\sum_{1\le i\le n}E\left(\|\tilde{R}_{cK,i}\|^2I(\|\tilde{R}_{cK,i}\|>\epsilon\ n^{1/2})\right)=0.
$$

Note that $\tilde{R}_{cK,i}$ are independent and identically distributed. We thus have

$$
\begin{aligned}
&n^{-1}\sum_{1\le i\le n}E\left(\|\tilde{R}_{cK,i}\|^2I(\|\tilde{R}_{cKi}\|>\epsilon n^{1/2})\right)\\
&\quad=E\left(\|\tilde{R}_{cK,1}\|^2I(\|\tilde{R}_{K,1}\|>\epsilon n^{1/2})\right).
\end{aligned}
$$

As a result, it suffices to show that $\lim_{K\to\infty}E\|\tilde{R}_{cK,1}\|^2<\infty$. For any K, we have $E\|\tilde{R}_{cK,1}\|^2=E\{\operatorname{trace}(\tilde{R}_{cK,1}\tilde{R}_{cK,1}^\top)\}=\pi_K(p+2)\le p+2<\infty$ and consequently, the Lindeberg condition (10) holds. Recall that $\pi_K\to\pi_0$, as $K\to\infty$. Therefore, as $K\to\infty$,

$$
n^{-1/2}\Sigma_{cK}^{-1/2}\frac{\partial\ell_c(\theta_0)}{\partial\theta}\xrightarrow{d}N(0,\pi_0I_{p+2}).
$$

Since $\Sigma_{cK}^{-1/2}\Sigma_c^{1/2}\to I_{p+2}$, as $K\to\infty$, we have $n^{-1/2}\frac{\partial\ell_c(\theta_0)}{\partial\theta}\xrightarrow{d}N(0,\pi_0\Sigma_c)$. Combined with $n_c/n\to_p\pi_0$. We have $n_c^{-1/2}\frac{\partial\ell_c(\theta_0)}{\partial\theta}\xrightarrow{d}N(0,\Sigma_c)$ This completes the proof of (7).

Step 5. In the last step, we investigate the asymptotic behavior of the MLE $\hat{\theta}$. We start with investigating its existence. Following similar technique as in Fan and Li (2001), it suffices to show that

$$
\lim_{C\to\infty}P\left(\sup_{\|u\|=C}\ell(\theta_0+n_c^{-1/2}u)\le\ell(\theta_0)\right)=1.
$$

We introduce some notations. For any vector $\mathbf{a}=(a_1,\cdots,a_m)^T,\mathbf{b}=(b_1,\cdots,b_m)^T$, denote $|\mathbf{a}|=\sum_i|a_i|$, $D^{\mathbf{a}}f(x)=\partial^{|a|}f/\partial a_1\cdots\partial a_m$ and $\mathbf{b}^{\mathbf{a}}=\prod_{i=1}^m b_i^{a_i}$. Then by Taylor expansion for multivariate function with integral remainder, we have

$$
\begin{aligned}
(11)\qquad &\ell_c(\theta_0+n_c^{-1/2}u)-\ell_c(\theta_0)\\
&\quad=n_c^{-1/2}\frac{\partial\ell_c(\theta_0)}{\partial\theta^\top}u+n_c^{-1}u^\top\frac{\partial^2\ell_c(\theta_0)}{\partial\theta\partial\theta^\top}u\\
&\quad+n_c^{-3/2}\sum_{|a|=3}R_{\mathbf{a}}(\theta_0+n_c^{-1/2}u)u^{\mathbf{a}},
\end{aligned}
$$

where $R_{\mathbf{a}}(\theta_0+n_c^{-1/2}u)=2^{-1}\int_0^1(1-t)^2D^{\mathbf{a}}\ell_c(\theta_0+tn_c^{-1/2}u)dt$ with $|R_{\mathbf{a}}(\theta_0+n_c^{-1/2}u)|\le 2^{-1}\max_{|\mathbf{a}|=3}\max_{\theta\in\Theta}|D^{\mathbf{a}}\ell_c(\theta)|$. By the assumption (C3), we see that $n_c^{-1}|R_{\mathbf{a}}(\theta_0+n_c^{-1/2}u)|\le 2^{-1}M(Y_K,T_K)=O_p(1)$. Combined with the fact that $n_c\to\infty$ and $\|u\|$ is fixed, we have that the last term on the right side of (11) is of order $o_p(\|u\|^2)$.

From Step 2, it follows that $n_c^{1/2}\partial\ell_c(\theta_0)/\partial\theta^\top = O_p(1)$. So the first term is of the order $O_p(\|u\|)$. Also by (6) and (11), we have

$$\ell_c(\theta_0 + n_c^{-1/2}u) = \ell_c(\theta_0) + O_p(\|u\|) - u^\top\Sigma_c u + o_p(\|u\|^2).$$

As C being sufficiently large, the third term on the right hand side dominate the second term and consequently, there exists $\sqrt{n_c}$-consistency estimator, denoted as $\hat{\theta}$. Because $\hat{\theta}$ is $\sqrt{n_c}$-consistent, the standard Taylor's expansion type argument can be applied (Shao, 1997; Fan and Li, 2001), which leads to

$$\begin{aligned} 0 &= n_c^{-1/2}\frac{\partial\ell(\hat{\theta})}{\partial\theta} \\ &= n_c^{-1/2}\frac{\partial\ell(\theta_0)}{\partial\theta} + n_c^{-1/2}(\hat{\theta}-\theta)^\top\frac{\partial^2\ell_c(\theta_0)}{\partial\theta\partial\theta^\top} + o_p(\|\hat{\theta}-\theta\|^2). \end{aligned}$$

This conclusion, together with (7), (6), leads to $\sqrt{n_c}(\hat{\theta} - \theta_0) \to_d N(0, \Sigma_c^{-1})$. This completes the entire theorem proof.

APPENDIX B. PROOF OF THEOREM 1

The first conclusion of Theorem can be viewed as a special case of $c = \infty$. As $c = \infty$, we have $n_c = n, K_c = K$. Then conclusion Theorem 1 is derived based on the Theorem 2.

Received 28 September 2014

REFERENCES

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360. MR1946581

Gonzalez, M., Hidalgo, C., and Barabasi, A. (2012), "Understanding individual human mobility patterns," *Nature*, 453, 5 June, 779–782.

Kumar, R., Mahdian, M., Pang, B., Tomkins, A., and Vassilvitskii, S. (2015), "Driven by Food: Modeling Geographic Choice," *WSDM'15*, February 2–6, Shanghai, China, 213–222.

Li, Z., Ding, B., Han, J., Kays, R., and Nye, P. (2012), "Mining Periodic Behaviors for Moving Objects," *KDD'10*, July 25–28, 2012, Washington, DC, USA.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, New York. MR3223057

McFadden, D. (1974), "Conditional logit analysis of qualitative choice behavior," *Paul Zarembka, ed. Frontiers in Econometrics*, Academic Press, New York, 3–118.

McFadden, D. (1978), "Modelling the choice of residential location," *Anders Karlqvist et al. eds. Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam, 75–96.

Shao, J. (1997), *Mathematics Statistics*, Springer, New York.

Song, C., Qu, Z., Blumm, N., and Barabasi, A. (2012), "Limits of predictability in human mobility," *Science*, 327, 19 February, 1018–1021. MR2643139

Yan, X., Han, X., Wang, B., and Zhou, T. (2012), "Diversity of Individual Mobility Patterns and Emergence of Aggregated Scaling Laws," *Physics.soc-ph*, arXiv:1211.2874v2, 23 Sep 2013.

Yuan, J., Zheng, Y., and Xie, X. (2012), "Discovering Regions of Different Functions in a City Using Human Mobility and POIs," *KDD'12*, August 12–16, 2012, Beijing, China, 186–194.

Zheng, Y., Zhang, L., Xie, X., and Ma, W. (2009), "Mining Interesting Locations and Travel Sequences from GPS Trajectories," *IW3C2*, April 20–24, 2009, Madrid, Spain.

Xiaoling Lu
Center for Applied Statistics
Data Mining Center
School of Statistics
Renmin University of China
Beijing 100872
China
E-mail address: xiaolinglu@ruc.edu.cn

Junlong Zhao
School of Mathematics and Systems Science
Beihang University
Beijing 100191
China
E-mail address: zhaojunlong928@126.com

Yu Chen
Guanghua School of Management
Peking University
Beijing 100871
China
E-mail address: yu.chen@pku.edu.cn

Hansheng Wang
Guanghua School of Management
Peking University
Beijing 100871
China
E-mail address: hansheng@pku.edu.cn