Some Theory of Likelihood

Friday 15$^{\text{th}}$ December, 2017

1 To be done

- Understand existing theory in exponential family. Some paper to be read: Portnoy (1988); Ghosal (2000).
- Give the theory of posterior Bayes factor under exponential family.
- Beyond exponential family. (**?**).
- Bartlett correction.
- General integral likelihood ratio test.
- Nonasymptotic. Read Spokoiny (2012)'s paper.
- Consider the sparse case as in Stadler and Mukherjee (2017).

2 Introduction

3 Results for exponential family

The content of this section is adapted from Ghosal (2000).

The following result, known as acute angle principle, is a key tool for the analysis.

Lemma 1 (J. M. Ortega (1987), Theorem 6.3.4.)**.** *Let C be an open, bounded set in $\mathbb{R}^n$ and assume that $F : \bar{C} \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ is countinuous and satisfies $(x - x_0)^T F(x) \geq 0$ for some $x_0 \in C$ and all $x \in \partial C$. Then $F(x) = 0$ has a solution in $\bar{C}$.*

We make the following assumptions.

Assumption 1. *The p dimensional independent random samples $X_1, \ldots, X_n$ are from a standard exponential family with density*

$$f(x; \theta_n) = \exp[x^T \theta_n - \psi_n(\theta_n)]$$

*with respect to μ_n. Where $\theta_n \in \Theta_n$, an **open** subset of $\mathbb{R}^n$. Sometimes we suppress the subscript n.*

The true parameter is denoted by θ_0. To prevent θ_0 approaching the boundary as $n \to \infty$, we assume that for a fixed $\epsilon_0 > 0$ independent of n, $B(\theta_0, \epsilon_0) \subset \Theta$.

It's well known that $\mathrm{E}\, X_1 = \psi'(\theta_0)$ and $\mathrm{Var}\, X_1 = \psi''(\theta_0)$. $\psi''(\theta_0)$ is also the Fisher information matrix. We assume that $\psi''(\theta_0)$ is positive definite.

Assumption 2. *$p \to \infty$ as $n \to \infty$.*

Let the positive definite matrix $\mathbf{J}$ be the square root of $\psi''(\theta_0)$, that is $\psi''(\theta_0) = \mathbf{J}^2$. The MLE $\hat{\theta}$ of θ is unique and satisfies $\psi'(\hat{\theta}) = \bar{X}$.

For a square matrix $\mathbf{A}$, $\|\mathbf{A}\|$ will stand for its operator norm.

The function $\psi(\theta)$ is in fact the cumulant generating function of X_1. Portnoy (1988) gave the following Taylor series expansions.

Proposition 1. *Suppose Assumption 1 holds. For any θ and θ_0 in Θ, the following expansions hold for some $\tilde{\theta}$ between θ and θ_0:*

$$\begin{aligned}
\psi(\theta) =&\psi(\theta_0) + (\theta-\theta_0)^T\psi'(\theta_0) + \frac{1}{2}(\theta-\theta_0)^T\psi''(\theta_0)(\theta-\theta_0)\\
&+ \frac{1}{6}\operatorname{E}_{\theta_0}\Big((\theta-\theta_0)^T(U-\operatorname{E}_{\theta_0}U)\Big)^3\\
&+ \frac{1}{24}\Big\{\operatorname{E}_{\tilde{\theta}}\Big((\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\Big)^4 - 3\big[\operatorname{E}_{\tilde{\theta}}\big((\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\big)^2\big]^2\Big\},
\end{aligned}$$

$$\begin{aligned}
\alpha^T\psi'(\theta) =&\alpha^T\psi'(\theta_0) + \alpha^T\psi''(\theta_0)(\theta-\theta_0)\\
&+ \frac{1}{2}\operatorname{E}_{\theta_0}\big((\theta-\theta_0)^T(U-\operatorname{E}_{\theta_0}U)\big)^2\alpha^T(U-\operatorname{E}_{\theta_0}U)\\
&+ \frac{1}{6}\operatorname{E}_{\tilde{\theta}}\big((\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\big)^3\alpha^T(U-\operatorname{E}_{\tilde{\theta}}U)\\
&- \frac{1}{2}\operatorname{E}_{\tilde{\theta}}\big((\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\big)^2\operatorname{E}_{\tilde{\theta}}(\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\alpha^T(U-\operatorname{E}_{\tilde{\theta}}U),
\end{aligned}$$

$$\alpha^T\psi'(\theta) =\alpha^T\psi'(\theta_0) + \alpha^T\psi''(\theta_0)(\theta-\theta_0) + \frac{1}{2}\operatorname{E}_{\tilde{\theta}}\big((\theta-\theta_0)^T(U-\operatorname{E}_{\tilde{\theta}}U)\big)^2\alpha^T(U-\operatorname{E}_{\tilde{\theta}}U),$$

where U is a random variable with density $f(x;\theta)$, α is a fixed p dimensional vector.

Let, for $c \geq 0$,

$$\begin{aligned}
B_{1n}(c) &= \sup\Big\{\operatorname{E}_\theta|a^T\mathbf{J}^{-1}(U-\operatorname{E}_\theta U)|^3 : \|a\| = 1, \|\mathbf{J}(\theta-\theta_0)\| \leq c\Big\},\\
B_{2n}(c) &= \sup\Big\{\operatorname{E}_\theta|a^T\mathbf{J}^{-1}(U-\operatorname{E}_\theta U)|^4 : \|a\| = 1, \|\mathbf{J}(\theta-\theta_0)\| \leq c\Big\},
\end{aligned}$$

Consistency.

Theorem 1. *Suppose Assumption 1 holds. Assume that for all $M > 0$, $M\sqrt{p/n}B_{1n}(M\sqrt{p/n}) \to 0$. Assume that for every $M > 0$, we have $\{\theta : \|\mathbf{J}(\theta-\theta_0)\| \leq M\sqrt{p/n} \subset\}\Theta$ for large n. Then*

$$\|\mathbf{J}(\hat{\theta}-\theta_0)\| = O_P(\sqrt{p/n}).$$

Proof. The MLE $\hat{\theta}$ is unique and satisfies $\bar{X} - \psi'(\hat{\theta}) = 0$. By Lemma 1, the inequality

$$\sup_{\|\mathbf{J}(\theta-\theta_0)\|=c}(\theta-\theta_0)^T(\bar{X}-\psi'(\theta)) \leq 0$$

implies $\|\mathbf{J}(\hat{\theta}-\theta_0)\| \le c$. By proposition 1, for θ satisfying $\|\mathbf{J}(\theta-\theta_0)\| = c$, we have

$$\begin{aligned}
(\theta-\theta_0)^T(\bar{X}-\psi'(\theta)) &= (\theta-\theta_0)^T(\bar{X}-\psi'(\theta_0)) - (\theta-\theta_0)^T\psi''(\theta_0)(\theta-\theta_0) - \frac{1}{2}\,\mathrm{E}_{\tilde{\theta}}\left((\theta-\theta_0)^T(U-\mathrm{E}_{\tilde{\theta}}U)\right)^3 \\
&= (\theta-\theta_0)^T\mathbf{J}\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0)) - c^2 - \frac{1}{2}\,\mathrm{E}_{\tilde{\theta}}\left((\theta-\theta_0)^T(U-\mathrm{E}_{\tilde{\theta}}U)\right)^3 \\
&\le c\|\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0))\| - c^2 - \frac{1}{2}\,\mathrm{E}_{\tilde{\theta}}\left((\theta-\theta_0)^T\mathbf{J}\mathbf{J}^{-1}(U-\mathrm{E}_{\tilde{\theta}}U)\right)^3 \\
&\le c\|\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0))\| - c^2 + \frac{1}{2}c^3 B_{1n}(c).
\end{aligned}$$

Since $\mathrm{E}\,\|\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0))\|^2 = \mathrm{tr}\,\mathrm{Var}(\mathbf{J}^{-1}\bar{X}) = p/n$, we have $\|\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0))\| = O_P(\sqrt{p/n})$. Hence for every $\delta > 0$, there is an M such that $\|\mathbf{J}^{-1}(\bar{X}-\psi'(\theta_0))\| \le M\sqrt{p/n}$ with probability at least $1-\delta$. Taking $c = 2M\sqrt{p/n}$ yields that with probability at least $1-\delta$,

$$\sup_{\|\mathbf{J}(\theta-\theta_0)\| = M\sqrt{p/n}} (\theta-\theta_0)^T(\bar{X}-\psi'(\theta)) \le -2M^2\frac{p}{n} + 2M^2\frac{p}{n}\left(2M\sqrt{p/n}B_{1n}(2M\sqrt{p/n})\right),$$

which is less than 0 eventually. Hence for large n, with probability at least $1-\delta$, $\|\mathbf{J}(\theta-\theta_0)\| \le M\sqrt{p/n}$. This proves $\|\mathbf{J}(\theta-\theta_0)\| = O_P(\sqrt{p/n})$.

□

Appendices

Appendix A haha1

Appendix B haha2

References

References

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis*, 74(1):49–68.

J. M. Ortega, W. C. R. (1987). *Iterative Solution of Nonlinear Equations in Several Variables.* Classics in Applied Mathematics, 30. Society for Industrial Mathematics.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, 16(1):356–366.

Spokoiny, V. (2012). Parametric estimation. finite sample theory. *Annals of Statistics*, 40(6):2877–2909.

Stadler, N. and Mukherjee, S. (2017). Two-sample testing in high dimensions. *Journal of The Royal Statistical Society Series B-statistical Methodology.*