STAT 576 Bayesian Analysis

Lecture 4: Asymptotic Properties of Bayesian Inference

Chencheng Cai

Washington State University

Normal Approximation to the Posterior Distribution

- Let $\hat{\boldsymbol{\theta}}$ be the maximize-a-posteriori (MAP) estimator, that is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\ p(\boldsymbol{\theta} \mid y)$$

- Consider a Taylor expansion of the $\log p(\boldsymbol{\theta} \mid y)$ at its mode $\hat{\boldsymbol{\theta}}$:

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

- The linear term is omitted because

$$\left[\frac{d}{d\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid y)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$$

Normal Approximation to the Posterior Distribution

- With the second approximation of the log-density around the mode:

$$\log p(\boldsymbol{\theta} \mid y) \approx \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

- we have the normal approximation of the posterior by

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y)$$

is the **observed information matrix**.

Normal Approximation to the Posterior Distribution

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

- The normal approximation works for any distribution of θ (with mode $\hat{\theta}$) when
  - $\hat{\boldsymbol{\theta}}$ is an inner point of Θ.
  - $\log p(\boldsymbol{\theta} \mid y)$ is second-order differentiable at $\hat{\boldsymbol{\theta}}$.
  - $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is positive-definite / non-singular.
- Using Bayes' rule, we have

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y) = \underbrace{-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y \mid \boldsymbol{\theta})}_{\text{info. from observations}} \quad \underbrace{-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta})}_{\text{info. from prior}}$$

Information Matrix

- Suppose we have i.i.d. observations $y = (y_1, \ldots, y_n)$ from a distribution $F_{\boldsymbol{\theta}}$ from a parametric family $\{F_{\boldsymbol{\theta}_0} : \boldsymbol{\theta} \in \Theta\}$ with true parameter $\boldsymbol{\theta}_0$.
- Then the observed information matrix is

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) - \frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta})$$

- With Law of Large Numbers, we know

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta}) \xrightarrow{F_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2} \log p(y_i \mid \boldsymbol{\theta})\right]$$

- Note: This is **NOT** the Fisher's information matrix because the expectation is taken under the true parameter $\boldsymbol{\theta}_0$.

Normal Approximation to the Posterior Distribution

- With the approximation from previous slide, we can revise the Taylor expansion of $\log p(\boldsymbol{\theta} \mid y)$ to

$$\log p(\boldsymbol{\theta} \mid y) = \log p(\hat{\boldsymbol{\theta}} \mid y) + \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_P(n\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$$

- Then the posterior can be approximated by

$$p(\boldsymbol{\theta} \mid y) \approx \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \hat{\boldsymbol{\theta}}, \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}_0}^{-1}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

- Or the rescaled version:

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathbb{E}_{\boldsymbol{\theta}_0}^{-1}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

where $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ is called the **local parameter** to $\hat{\boldsymbol{\theta}}$.

Normal Approximation to the Posterior Distribution

The approximation

$$p(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \;\middle|\; \mathbf{0}, \mathbb{E}_{\boldsymbol{\theta}_0}^{-1}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)$$

is still not satisfying as an asymptotic result because

- It is a finite sample approximation.
- The variance depends on the true parameter $\boldsymbol{\theta}_0$ and thus infeasible.
- The variance is random by involving $\hat{\boldsymbol{\theta}}$ in the formula.

Therefore, we need first to investigate the asymptotic behavior of $\hat{\boldsymbol{\theta}}$ itself.

Asymptotic Equivalence of MAP and MLE

- Maximize-a-posteriori estimator:

$$\hat{\boldsymbol{\theta}}_n^{(map)} = \arg\max \ \log p(\boldsymbol{\theta} \mid y) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta}) + \frac{1}{n}\log p(\boldsymbol{\theta})}_{f_n(\boldsymbol{\theta})}$$

- Maximum Likelihood Estimator:

$$\hat{\boldsymbol{\theta}}_n^{(mle)} = \arg\max \ \log p(y \mid \boldsymbol{\theta}) = \arg\max \ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{\theta})}_{g_n(\boldsymbol{\theta})}$$

- The difference $f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})$ does not uniformly converge to zero.
- But since $p(\hat{\boldsymbol{\theta}}_n^{(map)}) \geq p(\hat{\boldsymbol{\theta}}_n^{(mle)})$, as long as $\hat{\boldsymbol{\theta}}_n^{(mle)} \in \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta}) > 0\}$, we only need to consider the subset with positive prior density.
- A sufficient condition is (1) $\hat{\boldsymbol{\theta}}_n^{(mle)}$ is consistent for $\boldsymbol{\theta}_0$, and (2) $p(\boldsymbol{\theta})$ is strictly positive in a neighbor of $\boldsymbol{\theta}_0$.

Normal Approximation to the Posterior Distribution

- Under regularity conditions on the prevoius slide, we have

$$\hat{\theta}_n^{(map)} \xrightarrow{P} \boldsymbol{\theta}_0$$

- Therefore,

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \xrightarrow{P} \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{d^2}{d\boldsymbol{\theta}^2}\log p(y_i \mid \boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathcal{I}(\boldsymbol{\theta}_0)$$

- In this case, the approximation of the posterior is

$$p(\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \mid y) \approx \mathcal{N}\left(\boldsymbol{h} \,\middle|\, \mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)\right)$$

  with $\boldsymbol{h} = \sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ the **local parameter**.
- The unnormalized version is the distribution that is degenerate at $\boldsymbol{\theta}_0$.

$$p(\boldsymbol{\theta} \mid y) \approx \delta_{\boldsymbol{\theta}_0}$$

Bayes Estimator

Besides the MAP estimator, we can define a general Bayes estimator based on any loss function L.

- $L(\theta, \delta)$ is the **loss** in utitlity when the true parameter is θ while the estimator is δ.
  - Squared loss: $L(\theta, \delta) = (\theta - \delta)^2$
  - Misclassification loss: $L(y, \hat{y}) = \mathbb{I}\{y \neq haty\}$.
- The **risk** of an estimator δ is given by

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta)]$$

- The **Bayes risk** of an estimator δ is

$$R(\delta) = \mathbb{E}_{p(\theta)}[R(\theta, \delta)] = \mathbb{E}[L(\theta, \delta)]$$

- The **Bayes estimator** is the estimator $\hat{\theta}$ that minimizes the Bayes risk:

$$\hat{\theta}_n = \arg\min_{\delta \in \Theta} \; R(\delta)$$

Bayes Estimator

- Note that $R(\delta) = \mathbb{E}[\mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y]]$
- The Bayes estimator turns our to be the conditional optimizer:

$$\hat{\theta}_n(y) = \arg\min_{\delta \in \Theta} \ \mathbb{E}_{p(\theta|y)}[L(\theta, \delta) \mid y] = \arg\min_{\delta \in \Theta} \ \int L(\theta, \delta) p(\theta \mid y) d\mu$$

- Examples:
  - under squared loss: $\hat{\theta}_n$ is the posterior mean.
  - under absolute loss: $\hat{\theta}_n$ is the posterior median.
  - under cross entropy loss: $\hat{\theta}_n$ is the one with minimum Kullback-Leibler divergence.
- Do we still have the consistency result for Bayes estimators other than the MAP?
  Yes. Doob's Consistency Theorem.
- Do we still have the normal approximation for the posterior without utilizing the MAP?
  Yes. Berstein-Von Mises Theorem.

Counter-Examples

Before we move on to the Doob's Theorem and the Berstein-Von Mises Theorem. We first look at the a few counter-examples that are related to the key assumptions so far.

- Unidentifiable Models:
  Only observe the values of u for

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

- Non-fixed Number of Parameters:

$$y_i \sim \mathcal{N}(\theta_i, 1)$$

- Zero prior density at $\boldsymbol{\theta}_0$.
- Converge to the edge of the parameter space.

Notation

- Distribution family $\{P_\theta : \theta \in \Theta\}$
- For any measurable function $f : \mathcal{X} \to \mathbb{R}$,

  $$P_\theta f := \mathbb{E}_\theta[f(X)]$$

  is the expectation of f under probability measure P_θ.
- P_θ^n is the joint probability measure for n independent copies.
- $P_{\theta|y_1,y_2,\ldots,y_n}$ is the posterior probability measure given obervations $y_1, \ldots, y_n$.

Doob's Consistency Theorem

Definition (Consistency)

A sequence of posterior measures $P_{\theta|y_1,y_2,\ldots,y_n}$ is called consistent under θ_0 if under $P_{\theta_0}^{\infty}$-probability it converges in distribution to the measure δ_{θ_0} that is degenerate at θ_0, in probability. It is strongly consistent if this happens for almost every sequence $X_1, X_2, \ldots$.

Main result for the consistency of the posterior measure:

Theorem (Doob's Consistency Theorem)

Suppose that the sample space $(\mathcal{X}, \mathcal{A})$ is a subset of Euclidean space with its Borel σ-field. Suppose that $P_\theta \neq P_{\theta'}$ whenever $\theta \neq \theta'$. Then for every prior probability measure Π on Θ the sequence of posterior measures is consistent for Π-almost every θ.

Doob's Consistency Theorem — Proof

- The probability space we are working with: $\theta \sim \Pi$ and $y_1, y_2, \cdots \mid \theta \sim P_\theta$ i.i.d..
- Let Q be the joint probability measure on $\mathcal{X}^\infty \times \Theta$ such that the joint distribution $(y_1, \ldots, y_n, \theta)$ is a cylinder of the space.
- **Step 1:** Claim: there exists a measurable function $h : \mathcal{X}^\infty \to \Theta$ such that

$$h(x_1, x_2, \ldots) = \theta, \quad Q - a.s.$$

- **Step 2:** Then, for any bounded, measurable function $f : \Theta \to \mathbb{R}$, we construct a sequence $\eta_1, \eta_2, \ldots$ by

$$\eta_n = \mathbb{E}[f(\theta) \mid y_1, \ldots, y_n].$$

- η_n is a martingale. By Doob's martingale convergence theorem, we have

$$\eta_n \to \eta_\infty := \mathbb{E}[f(\theta) \mid y_1, y_2, \ldots] = f(h(y_1, y_2, \ldots)), \quad Q - a.s.$$

Theorem (Doob's Martingale Convergence Theorem)

Suppose X_n *is a super-martingale that satisfies* $\sup_n \mathbb{E}[|X_n|] < +\infty$*. Then* $X_\infty = \lim_n X_n$ *exists almost surely, and* $X_n \to X_\infty$ *a.s.*

Doob's Consistency Theorem — Proof

- Recall: for any bounded, measurable function f, we have

$$\mathbb{E}[f(\theta) \mid y_1, \ldots, y_n] \to f(h(y_1, y_2, \ldots)), \quad Q - a.s.$$

Lemma (Convergence-Determining Class)

There exists a countable set of continous functions $f : \mathbb{R}^k \to [0, 1]$ that $X_n \xrightarrow{\mathcal{D}} X$ if and only if $\mathbb{E}[X_n] \to \mathbb{E}[X]$ uniformly in $f \in \mathcal{F}$.

- With the countable convergence-determing class, we have

$$P_{\theta|y_1,\ldots,y_n} \xrightarrow{\mathcal{D}} \delta_{h(y_1,y_2,\ldots)}, \quad Q - a.s.$$

End of Step 2.

- Now we need to traslate the right-hand side to δ_{θ_0}.

Doob's Consistency Theorem — Proof

- **Step 3:** Let $C \subset \mathcal{X}^\infty \times \Theta$ be the subset that all current results hold,
- that is the intersection of all $Q - a.s.$ sets so far.
- By Fubini's Theorem, we have

$$1 = Q(C) = \iint \mathbb{I}\{(y, \theta) \in C\} dP_\theta^\infty(y) d\Pi(\theta) = \int P_\theta^\infty(C_\theta) d\Pi(\theta),$$

  where $C_\theta = \{y : (y, \theta) \in C\}$.
- We immediately have $P_\theta^\infty(C_\theta) = 1$ for Π-almost every θ.
- For those θ_0 that $P_\theta^\infty(C_\theta) = 1$, we have $(y, \theta_0) \in C$ for $P_{\theta_0}^\infty$-almost every sequence $y_1, y_2, \dots,$ then

$$P_{\theta|y_1,\dots,y_n} \xrightarrow{\mathcal{D}} \delta_{h(y_1,y_2,\dots)} = \delta_{\theta_0}$$

- **Now the theorem is proved**.

Doob's Consistency Theorem — Proof of Step 1

Claim: there exists a measurable function $h : \mathcal{X}^\infty \to \Theta$ such that

$$h(x_1, x_2, \dots) = \theta, \quad Q - a.s.$$

Definition (Accessibility)

A measurable function $f : \Theta \to \mathbb{R}$ is called accessible if there exists a sequence of measurable functions $h_n : \mathcal{X}^n \to \mathbb{R}$ such that

$$\int |h_n(y) - f(\theta)| \wedge 1 dQ(y, \theta) \to 0.$$

- The claim is eqivalent to say all $f(\theta) = \theta_0$ is accessible.
- We can show: every Borel measurable function is accessible.

Doob's Consistency Theorem — Proof of Step 1

Want to show: every Borel measurable function is accessible.

- **Step 1.1:** $f(\theta) = P_\theta(A)$ for any measurable set A is accessible.
- We can choose $h_n(y) = n^{-1} \sum_{i=1}^n \mathbb{I}\{y_i \in A\}$ and by LLN.
- **Step 1.2:** every function that is measurable in the σ-field generated by accessible functions is accessible.
- **Step 1.3:** Since $(\mathcal{X}, \mathcal{A})$ is Euclidean, there exits a countable measure determing subcollection $\mathcal{A}_0 \subset \mathcal{A}$.
- For A ranging over $\mathcal{A}_0$, the function $P_\theta(A)$ separates the points of Θ because of the identifiability. These functions generates the Borel σ-field on Θ.
- **Step 1.4:** Therefore all Borel measurable functions are accessible.

Doob's Consistency Theorem — Proof of Step 1.2

To show: every function that is measurable in the σ-field generated by accessible functions is accessible.

Lemma

Let $\mathcal{F}$ be a linear subspace of $\mathcal{L}^1(\Pi)$ with the properties:

1. *if $f, g \in \mathcal{F}$, then $f \wedge g \in \mathcal{F}$;*
2. *if $0 \leq f_1 \leq f_2 \leq \cdots \in \mathcal{F}$, and $f_n \uparrow f \in \mathcal{L}^1(\Pi)$, then $f \in \mathcal{F}$;*
3. *$1 \in \mathcal{F}$.*

Then $\mathcal{F}$ contains everty $\sigma(\mathcal{F})$-measurable function in $\mathcal{L}^1(\Pi)$.

Proof:

- Let $\mathcal{A}_0 = \{A : \mathbf{1}_A \in \mathcal{F}\}$
- $\mathcal{A}_0$ is a π-system and a λ-system. By Dynkin Theorem, $\mathcal{A}_0$ is a σ-field.
- For any $f \in \mathcal{F}$, the function $n(f - \alpha)_+ \wedge 1$ is in $\mathcal{F}$ and converges to $\mathbb{I}\{f > \alpha\}$. So $\{f > \alpha\} \in \mathcal{A}_0$.
- So $\sigma(\mathcal{F}) \subset \mathcal{A}_0$.

Doob's Consistency Theorem — Proof of Step 1.3

Lemma

Let $\mathcal{F}$ be a countable collection of measurable functions $f : \Theta \subset \mathbb{R}^k \to \mathbb{R}$ that separates the points of Θ. Then the Borel σ-field and the σ-field generated by $\mathcal{F}$ on Θ coincide.

Quadratic Mean Differentiability (QMD)

- Now we consider expand the likelihood function at the true parameter θ_0 with local parameter h.
- Taylor expansion:

$$\log \prod_{i=1}^{n} p(y_i \mid \theta_0 + h/\sqrt{n}) = \log \prod_{i=1}^{n} p(y_i \mid \theta_i) + \frac{h}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}(\theta_0; y_i) + \frac{h^2}{2n} \sum_{i=1}^{n} \ddot{\ell}(\theta_0; y_i) + o(h^2/n)$$

- By Law of Large Numbers, we have

$$\log \prod_{i=1}^{n} \frac{p(y_i \mid \theta_0 + h/\sqrt{n})}{p(y_i \mid \theta_0)} = h\Delta_{n,\theta_0} - \frac{1}{2} h^2 \mathcal{I}(\theta_0) + o_P(1),$$

where $\Delta_{n,\theta_0} = n^{-1/2} \sum_{i=1}^{n} \dot{\ell}(\theta_0; y_i)$ and $\mathcal{I}(\theta_0) = -P_{\theta_0} \ddot{\ell}(\theta_0; y_i)$.
- Do we require the second-order Differentiability of ℓ to have this result?

Quadratic Mean Differentiability (QMD)

Definition (Quadratic Mean Differentiability)

The probility family $\{P_\theta : \theta \in \Theta\}$ is called differentiable in quadratic mean at θ_0 if there exists a measurable vector function $\dot{\ell}(\theta)$ such that

$$\int \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h^T\dot{\ell}(\theta_0)\right]^2 d\mu = o(\|h\|^2), \quad h \to 0.$$

- QMD does not require the existence of $\dot{\ell}$ everywhere.
- Instead, it finds a proxy function that works as $\dot{\ell}$ as long as the **overall** error is controlled.

Quadratic Mean Differentiability (QMD)

Theorem

Suppose that Θ *is an open subset of* $\mathbb{R}^k$, *and the probability family* $\{P_\theta : \theta \in \Theta\}$ *is differentiable in quadratic mean at* θ_0. *Then* $P_{\theta_0}\dot{\ell}(\theta_0) = 0$ *and the Fisher information matrix* $\mathcal{I}(\theta_0) = P_{\theta_0}\dot{\ell}(\theta_0)\dot{\ell}(\theta_0)^T$ *exists. Furthermore, for every converging sequence* $h_n \to h$ *as* $n \to \infty$,

$$\log \prod_{i=1}^{n} \frac{p(y_i \mid \theta_0 + h/\sqrt{n})}{p(y_i \mid \theta_0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \dot{\ell}(\theta_0) - \frac{1}{2} h^T \mathcal{I}(\theta_0) h + o_P(1).$$

Bernstein-Von Mises Theorem

Theorem (Bernstein-Von Mises)

Suppose the probability family $\{P_\theta : \theta \in \Theta\}$ *is differentiable in quadratic mean at* θ_0 *with nonsingular Fisher information matrix* $\mathcal{I}(\theta_0)$*, and suppose that for any* $\epsilon > 0$ *there exists a sequence of tests* ϕ_n *such that*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\epsilon} P_\theta^n (1-\phi_n) \to 0.$$

Futhermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 *with a continuous density function at* θ_0*. Then the corresponding posterior distribution satisfy*

$$\left\| P_{\sqrt{n}(\theta-\theta_0)|y_1,,,,y_n} - \mathcal{N}\left(\Delta_{n,\theta_0}, \mathcal{I}(\theta_0)^{-1}\right) \right\|_{TV} \xrightarrow{P_{\theta_0}^n} 0$$

Bernstein-Von Mises Theorem

- The assumption on the distribution family is weak. (QMD)
- $\phi_n : \mathcal{X}^n \to \{0, 1\}$ is a test with $\phi_n(y_1, \dots, y_n) = 1$ meaning "reject".
- The assumption

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\epsilon} P^n_\theta(1-\phi_n) \to 0.$$

  means there exists a sequence of tests that distinguishes θ_0 from any other points.
- The **total variation** distance between two distributions F_1 and F_2 is defined as

$$\|F_1 - F_2\|_{TV} = \sup_{A \in \mathcal{F}} |F_1(A) - F_2(A)| = \frac{1}{2}\int |f_1(x) - f_2(x)| d\mu(x)$$

- The in probability convergence is w.r.t. $P^n_{\theta_0}$, because the randomness of the left-hand side is the observations $y_1, \dots, y_n$.

Bernstein-Von Mises Theorem — Proof

Step 0: Notations.

- We use the local parameter $h = \sqrt{n}(\theta - \theta_0)$.
- The prior on θ, Π, is translated to the prior on h, Π_n, by

$$\Pi_n(A) = \Pi(\theta_0 + A/\sqrt{n}) \quad \text{for any measurable set } A.$$

- For a given set C, let Π_n^C be the probability measure by restricting Π_n to C and then renormalizing.
- We write $P_{n,h}$ as the distribution of $y_1, \ldots, y_n \mid \theta_0 + h/\sqrt{n}$.
- Let $P_{n,C} = \int P_{n,h}\, d\Pi_n^C(h)$ be the average probability measure on C.
- The posterior distributions with priors Π_n and Π_n^C are $P_{h|y_1,\ldots,y_n}$ and $P^C_{h|y_1,\ldots,y_n}$.

Bernstein-Von Mises Theorem — Proof

Step 1: show $P_{\theta|y_1,\ldots,y_n}$ and $P^{C_n}_{\theta|y_1,\ldots,y_n}$ are close.

- Let C_n be the ball with radius M_n.
- For any measurable set B, (let $y = (y_1, \ldots, y_n)$)

$$\begin{aligned}
P_{h|y}(B) - P^{C_n}_{h|y}(B) &= P_{h|y}(B \cap C_n^c) + P_{h|y}(B \cap C_n) - P^{C_n}_{h|y}(B \cap C_n) - P^{C_n}_{h|y}(B \cap C_n^c) \\
&= P_{h|y}(B \cap C_n^c) + P_{h|y}(B \cap C_n) - P^{C_n}_{h|y}(B \cap C_n) \\
&= P_{h|y}(B \cap C_n^c) + P_{h|y}(C_n) P^{C_n}_{h|y}(B \cap C_n) - P^{C_n}_{h|y}(B \cap C_n) \\
&= P_{h|y}(B \cap C_n^c) - P_{h|y}(C_n^c) P^{C_n}_{h|y}(B \cap C_n) \\
&= P_{h|y}(B \cap C_n^c) - P_{h|y}(C_n^c) P^{C_n}_{h|y}(B) \\
&\leq 2P_{h|y}(C_n^c)
\end{aligned}$$

- Therefore,

$$\left\| P_{h|y} - P^{C_n}_{h|y} \right\|_{TV} \leq 2P_{h|y}(C_n^c)$$

Bernstein-Von Mises Theorem — Proof

- Let U be a ball around zero with fixed radius.
- Then

$$
\begin{aligned}
P_{n,U} P_{h|y}(C_n^c)(1-\phi_n) &= P_{n,U} \int_{C_n^c} \frac{p_{n,h}(y)(1-\phi_n)}{\int p_{n,\tilde{h}}(y) d\Pi_n(\tilde{h})} d\Pi_n(h) \\
&= \int_U \left[\int_{\mathcal{X}^n} p_{n,h'}(y) \int_{C_n^c} \frac{p_{n,h}(y)(1-\phi_n)}{\int p_{n,\tilde{h}}(y) d\Pi_n(\tilde{h})} d\Pi_n(h) dy \right] d\Pi_n^U(h') \\
&= \frac{1}{\Pi_n(U)} \int_U \int_{\mathcal{X}^n} \int_{C_n^c} \frac{p_{n,h}(y) p_{n,h'}(y)(1-\phi_n)}{\int p_{n,\tilde{h}}(y) d\Pi_n(\tilde{h})} d\Pi_n(h) dy d\Pi_n(h') \\
&= \frac{\Pi_n(C_n^c)}{\Pi_n(U)} P_{n,C_n^c} P_{h|y}(U)(1-\phi_n) \\
&\leq \frac{1}{\Pi_n(U)} \int_{C_n^c} P_{n,h}(1-\phi_n) d\Pi_n(h)
\end{aligned}
$$

- The integrand converges pointwise to 0. But that's not enough.

Bernstein-Von Mises Theorem — Proof

Lemma

There exists a sequence of tests ϕ_n and a constant c such that for every sufficiently large n and every $\|\theta - \theta_0\| \geq M_n/\sqrt{n}$,

$$P_{\theta_0}^n \phi_n \to 0, \quad P_\theta^n (1 - \phi_n) \leq \exp\left\{-cn(\|\theta - \theta_0\|^2 \wedge 1)\right\}$$

proof sketch:

- For $M_n/\sqrt{n} \leq \|\theta - \theta_0\| \leq \epsilon$, we set $\phi_n = \mathbb{I}\{(\mathbb{P}_n - P_{\theta_0})\dot{\ell}^L(\theta_0) \geq \sqrt{M_n/n}\}$
- For $\|\theta - \theta_0\| > \epsilon$, we first choose k such that $P_{\theta_0}^k \phi_k < 1/4$ and $P_\theta^k (1 - \phi_k) < 1/4$ as the assumption in the BVM theorem. For $n = mk$, let $\psi_1, \ldots, \psi_m$ be ϕ_k applied to $(y_1, \ldots, y_k), \ldots, (y_{(m-1)k+1}, \ldots, y_{mk})$. Let $\phi_n = \mathbb{I}\{\bar{\psi} \geq 1/2\}$.

Bernstein-Von Mises Theorem — Proof

Return to our Step 1 of the main proof.

- Let $D \leq 1$ be sufficiently small such that $\pi(\theta)$ is uniformly bounded on $\|\theta - \theta_0\| \leq D$.
- Then

$$
\begin{aligned}
P_{n,U} P_{h|y}(C_n^c)(1-\phi_n) &\leq \frac{1}{\Pi_n(U)} \int_{C_n^c} P_{n,h}(1-\phi_n) d\Pi_n(h) \\
&\leq \frac{1}{\Pi_n(U)} \int_{\|h\| \geq M_n} e^{-c(\|h\|^2 \wedge n)} d\Pi_n(h) \\
&= \frac{1}{\Pi_n(U)} \left(\int_{M_n \leq \|h\| \leq D\sqrt{n}} + \int_{\|h\| \geq D\sqrt{n}} \right) e^{-c(\|h\|^2 \wedge n)} d\Pi_n(h) \\
&\leq K \left(\int_{\|h\| \geq M_n} e^{-c\|h\|^2} dh + \sqrt{n^k} e^{-cD^2 n} \right) \to 0
\end{aligned}
$$

- Therefore $P_{h|y}(C_n^c) \xrightarrow{P_{\theta_0}^n} 0$.

Bernstein-Von Mises Theorem — Proof

Step 2: show that $\mathcal{N}(\Delta_{n,\theta_0}, \mathcal{I}(\theta_0)^{-1})$ and $P^{C_n}_{h|y}$ are close.

- Now let C be the ball with fixed radius M around 0. Let $\mathcal{N}^C(\mu, \Sigma)$ be the normal distribution restricted to C.
- Then

$$\|\mathcal{N}^C(\Delta_{n,\theta_0}, \mathcal{I}(\theta_0)^{-1}) - P^C_{h|y}\|_{TV}$$

$$= \int \left(1 - \frac{d\mathcal{N}^C}{dP^C_{h|y}}\right)_+ dP^C_{h|y} = \int \left(1 - \frac{d\mathcal{N}^C(h) \int_C p_{n,g}(y) d\Pi_n(g)}{\mathbb{I}\{h \in C\} p_{n,h}(y) d\Pi_n(h)}\right)_+ dP^C_{h|y}(h)$$

$$\leq \iint \left(1 - \frac{p_{n,g}(y) d\Pi_n(g) d\mathcal{N}^C(h)}{p_{n,h}(y) d\Pi_n(h) d\mathcal{N}^C(g)}\right)_+ d\mathcal{N}^C(g) dP^C_{h|y}(h)$$

Bernstein-Von Mises Theorem — Proof

- It suffices to show the integral converges to 0 in mean under $P_{n,C}$.
- Notice that

$$P_{n,C}(dy)P^C_{h|y}(dh)\mathcal{N}^C(dg) = \Pi^C_n(dh)P_{n,h}(dy)\mathcal{N}^C(dg)$$

- Since Π^C_n and $\mathcal{N}^C$ are bounded on C, they can be replaced by a multiple of uniform measure λ_C
- Therefore, it suffices to show the integrand converges to 0 in probability

$$\lambda_C(dh)P_{n,0}(dy)\lambda_C(dg)$$

- It follows from the expansion theorem of QMD family.

Summary

- Both Doob's Consistency Theorem and Bernstein-Von Mises Theorem requires QMD.
- Some sufficient condition for QMD:

Lemma

For every θ in an open subset of $\mathbb{R}^k$, let p_θ be a μ-probability density. Assume the map $\theta \mapsto \sqrt{p_\theta(x)}$ is continously differentiable for every x. If the elements of the matrix

$$\mathcal{I}(\theta) = \int \frac{\dot{p}_\theta}{p_\theta} \frac{\dot{p}_\theta^T}{p_\theta} p_\theta d\mu$$

are well defined and continuous in θ. Then the map $\theta \mapsto \sqrt{p_\theta(x)}$ is QMD with $\dot{\ell}(\theta) = \dot{p}_\theta / p_\theta$.

Summary

- Under regularity conditions, the Doob's consistency theorem gives

$$p(\theta \mid y) \xrightarrow{\mathcal{D}} \delta_{\theta_0}$$

- Under regularity conditions, the Bernstein-Von Mises Theorem gives

$$\left\| p(\sqrt{n}(\theta - \theta_0) \mid y) - \mathcal{N}(\Delta_{n,\theta_0}, \mathcal{I}(\theta_0)^{-1}) \right\|_{TV} \xrightarrow{P} 0$$

or the resclaed version

$$\left\| p(\theta \mid y) - \mathcal{N}\left(\theta_0 + \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}(\theta_0 \mid y_i), \frac{1}{n}\mathcal{I}(\theta_0)^{-1}\right) \right\|_{TV} \xrightarrow{P} 0$$