Discussion of a Link between Brown and Jordan

Andrew R. Barron

Yale University, Department of Statistics & Data Science

Lawrence D Brown Memorial Workshop

December 1, 2018

The score function analysis of Larry Brown's estimation theory is linked to the log-likelihood ratio analysis of Michael Jordan.

Estimation Error and Score

- Larry Brown pioneered role of score function in analysis of the mean square of parameter estimation error

$$E[\mu|Y] - \mu$$

- Pointwise identity for exponential families with natural parameter μ, e.g. normal location,

$$\nabla \log p(Y) - \nabla \log p(Y|\mu) \;=\; E[\mu|Y] - \mu$$

- Implication: Mean squared error equals expected squared distance between scores

$$J(p,q) = E_p \|\nabla \log p(Y) - \nabla \log q(Y)\|^2$$

- Brown (1971): Which priors produce admissible estimators

- Brown (1982): Central Limit Theorem from decrease of $J(p_n, \phi)$

  p_n is the density for standardized sums of n i.i.d. random variables with added normals.

  ϕ is the corresponding normal density.

Log Likelihood Ratio

- Michael Jordan, with Ahmed El Alaoui, multivariate normal $Y_k = \mu_k + W_k$, rectangularly arrayed, $k = (i, j)$ with

$$\mu_{i,j} = \bar{\beta} u_i v_j$$

- Prior: u_i and v_j i.i.d. with $N \times M$ observations, $M/N = \alpha$.
- Investigate asymptotics of the log-likelihood ratio

$$\log L(Y; \beta) = \log p_\beta(Y)/p_0(Y)$$

- Mean is the Kullback divergence

$$E \log L(Y; \beta) = K(p_\beta, p_0)$$

- Address contiguity question of when there is a finite limit.
- One tool therein: Gaussian Poincaré inequality

$$E(\log L - E \log L)^2 \leq E \|\nabla \log L\|^2 = J(p_\beta, p_0)$$

- Additional tool available:

$$K(p_\beta, p) = (1/2) \int_0^\beta J(p_t, p_0)\, dt$$

Score Projection

- Setting: Continuous-valued Y, hidden variable or param X
- Relationship between marginal and conditional density

$$p(y) = \int p(y|x)P(dx)$$

$$\nabla p(y) = \int \nabla p(y|x)P(dx)$$

- Relationship between marginal and conditional score

$$\nabla \log p(y) = \frac{\nabla p(y)}{p(y)} = \int \left[\frac{\nabla p(y|x)}{}\right] \frac{P(dx)}{p(y)}$$

$$= \int [\nabla \log p(y|x)] P(dx|y)$$

- The score of Y is the best estimator of the score of $Y|X$.

Score Projection

- Setting: Continuous-valued Y, hidden variable or param X
- Relationship between marginal and conditional density

$$p(y) = \int p(y|x)P(dx)$$

$$\nabla p(y) = \int \nabla p(y|x)P(dx)$$

- Relationship between marginal and conditional score

$$\nabla \log p(y) = \frac{\nabla p(y)}{p(y)} = \int \left[\frac{\nabla p(y|x)}{p(y|x)}\right] \frac{p(y|x)P(dx)}{p(y)}$$

$$= \int [\nabla \log p(y|x)] P(dx|y)$$

- The score of Y is the best estimator of the score of $Y|X$.

Normal Location Example

$$Y = \mu + Z$$

- Linear Score

$$\nabla \log p(Y|\mu) = \mu - Y$$

- Score from Bayes factor $p(Y) = \int p(Y|\mu) P(d\mu)$ is the posterior mean of the conditional score

$$\nabla \log p(Y) = E[\mu|Y] - Y$$

- Error

$$\nabla \log p(Y) - \nabla \log p(Y|\mu) = E[\mu|Y] - \mu$$

- Expected square is the risk

$$J(p, p_\mu)$$

Exponential Families

- Naturally parameterized

$$p(y|\mu) = e^{\mu \cdot Y - \psi(\mu)} p_0(y)$$

- Conditional Score

$$\nabla \log p(Y|\mu) \;=\; \mu \;+\; \nabla \log p_0(Y)$$

- Minus score $-\nabla \log p_0(Y)$ is best unbiased estimator of μ
- Unconditional Score

$$\nabla \log p(Y) \;=\; E[\mu|Y] \;+ \nabla \log p_0(Y)$$

- Score difference equals error of Bayes estimator

$$\nabla \log p(Y) - \nabla \log p(Y|\mu) \;=\; E[\mu|Y] \;-\; \mu$$

- Expected square is the risk

$$J(p, p_\mu)$$

Likelihood Link

- Role of score in analysing $K(p,q) = E[\log p(Y)/q(Y)]$
- Intergral relationship

$$K(p,q) \;=\; (1/2)\int_{t>0} J(p_t,q_t)\,dt$$

- Arises in three groups of scholarship
  - B. 84,86, based on Stam 59, where p_t, q_t are densities for Y_t when X has density p, q

$$Y_t \;=\; e^{-t}X \;+\; \sqrt{1-e^{-2t}}\,W$$

  - Bakry & Emery 85, where p_t, q_t are densities for Brownian diffusion Y_t, with log-concave limit density f, when Y_0 has density p, q

$$dY_t \;=\; -(1/2)\nabla \log f(Y_t)\,dt \;+\; dW_t$$

  - Guo, Shamai, Verdú 05. Shannon Info as integral of MSE.

Back to Jordan

- Multivariate normal $Y_k = \mu_k + W_k$, rectangularly arrayed, $k = (i, j)$ with

$$\mu_{i,j} = \bar{\beta} u_i v_j$$

- Prior: u_i and v_j i.i.d. with $N \times M$ observations, $M/N = \alpha$.
- Investigate asymtotics of log-likelihood ratio mean
- Mean is the Kullback divergence

$$E_\beta \log L(Y; \beta) = K(p_\beta, p_0)$$

- Also stated to be of interest to study

$$-E_0[\log L(T; \beta)] = K(p_0, p_\beta)$$

- Integrals of J_t are available for the study of each.

An Information Trick

- Jordan's -$E_0[\log L(T;\beta)]$ is a special case of

$$K(p_0,p) = -E_0 \log \int [p(Y|\mu)/p(Y|0)]P(d\mu)$$

$$= -E_0 \log \int \exp\{\log p(Y|\mu)/p(Y|\mu_0)\}P(d\mu)$$

- Generalized Holder inequality (provable from non-negativity of Kullback divergence), shows it it not more than

$$\leq -\log \int \exp\{E_0[\log p(Y|\mu)/p(Y|\mu_0)]\}P(d\mu)$$

$$= -\log \int \exp\left\{-\bar{\beta}^2(\sum_{i,j} u_i v_j)^2/2\right\} P(du)P(dv).$$

These tools look promising as additional methods by which to show a finite limit for the expected log likelihood ratio in Michael Jordon's problem.

The Central Limit Problem

For independent identically distributed random variables $X_1, X_2, \ldots, X_n$, with $E[X] = 0$ and $VAR[X] = \sigma^2 = 1$, consider the standardized sum

$$\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}.$$

Let its density function be f_n and its distribution function F_n.

Let the standard normal density be ϕ and its distribution function Φ.

Natural questions:

- In what sense do we have convergence to the normal?
- Do we come closer to the normal with each step?
- Can we give clean bounds on the "distance" from the normal and a corresponding rate of convergence?

Convergence

- **In distribution:** $F_n(x) \to \Phi(x)$

  Classic: Fourier method or expansion of expectation of smooth functions.

  Linnik 59, Brown 82 via info measures applied to smoothed distributions.

- **In density:** $f_n(x) \to \phi(x)$

  Prohorov 52 showed $||f_n - \phi||_1 \to 0$ iff f_n exists eventually.

  Kolmogorov & Gnedenko 54 $||f_n - \phi||_\infty \to 0$ iff f_n bounded eventually.

- **In Shannon Information:** $H(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i) \to H(Z)$

  B. 84, 86 shows $D(f_n||\phi) \to 0$ iff it is eventually finite.

- **In Fisher Information:** $I(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i) \to 1/\sigma^2$

  Johnson & B. 04 shows $J(f_n||\phi) \to 0$ iff it is eventually finite.

The Projection Tool

For each X_i,

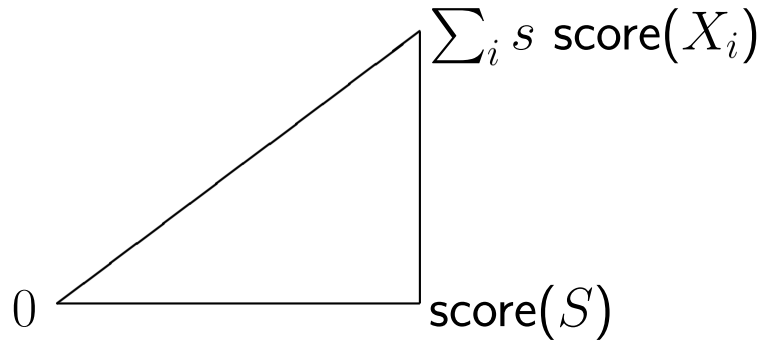
$$\text{score}(S) = E\big[\mathsf{S}(X_I) \,\big|\, S\big]$$

Hence, for weights w_i that sum to 1,

$$\text{score}(S) = E\left[\sum_i w_i \text{ score}(X_i) \,\middle|\, S\right]$$

Pythagorean inequality

The Fisher info. of the sum is the mean squared length of the projection



$$I(S) \le E\left[\sum_i w_i \text{ score}(X_i)\right]^2$$

The Heart of the Matter

- Pythagorean inequality:

$$I(\mathsf{sum}_{\mathsf{total}}) \leq E\left[\sum_{s\in\mathcal{S}} w_s \ \mathsf{score}(\mathsf{sum}_s)\right]^2$$

- Apply a variance drop lemma to get

$$I(\mathsf{sum}_{\mathsf{total}}) \leq r(\mathcal{S}) \sum_{s\in\mathcal{S}} w_s^2 I(\mathsf{sum}_s)$$

- Central limit analysis: quantify the Fisher gap by Pythagorean identity:

  gap = Expected norm square between $\sum w_s$ score(sum_s) and score$(\mathsf{sum}_{\mathsf{total}})$

  The only additive functions of a sum are the linear functions

  Score close to linear means the log density is close to the normal

Comment on CLT rate bounds

For iid X_i let $J_n = J(f_n||\phi)$ and $D_n = D(f_n||\phi)$.

- Supposing the distribution of the X_i to have a finite Poincaré constant R, Johnson & Barron '04: Pythagorean identity for score projection yields

$$J_n \leq \frac{2R}{n + (2R - 1)} J_1$$

$$D_n \leq \frac{2R}{n + (2R - 1)} D_1$$

- The finite Poincaré assumption implies finite moments of all orders.

- Osipov, Petrov 1967: Local limit thm, Edgeworth expan, $E|X|^m < \infty$,

$$[1 + |x|^m] \, (f_n(x) - \phi_m(x)) \, = \, o(n^{-(m-2)/2})$$

- Bobkov, Chistyakov, Götze 2013: Fourth moment and D_1 finite

$$D_n \, = \, 0(1/n)$$

Other remarkable implications and extensions

- Stochastic Diffusion: $dX_t = (1/2)\nabla \log q(X_t)dt + dB_t$ where B_t is a standard Brownian motion. Let p_t and $q_t = q$ be the density of X_t when at $t = 0$ the density is respectively $p_0 = p$ or $q_0 = q$.
- Bakry, Emery 85: Differential identity $d_t\, D(p_t\|q) = -(1/2)J(p_t\|q)$ from Kolmogorov Fokker-Planck equation (or the process generator) yielding

$$D(p||q) = \frac{1}{2}\int_0^\infty J(p_t\|q)dt$$

- Bakry, Emery 85: Log Sobolev inequalities: Suppose q is strictly log concave, that is, there is a $c > 0$ such that $Hessian(-\log q(X)) \geq c\, I$, then

$$d_t\, J(p_t\|q) \leq -c\, J(p_t\|q).$$

Hence $J(p_t\|q) \leq e^{-c\,t} J(p\|q)$. Integrating yields Log Sobolev inequality

$$D(p\|q) \leq \frac{1}{2c} J(p\|q)$$

Thus is obtained rapid mixing of the process

$$D(p_t\|q) \leq \frac{1}{2c} J(p_t\|q) \leq \frac{1}{2c} e^{-c\,t} J(p\|q)$$

Outline

- Entropy and the Central Limit Problem[1]
- Entropy Power Inequality (EPI)[2]
- Monotonicity of Entropy and subset sum EPI[3]
- Projection and Fisher Information
- Rates of Convergence in the CLT[4]
- Stam-Gross Log-Sobolev Inequality[2,5]
- Convergence to Invariance

[1] Andrew Barron, Annals of Probability 1986

[2] A. J. Stam, Inform & Control 1959

[3] Mokshay Madiman and Andrew Barron, ISIT 2006 and IEEE IT 2007

[4] Oliver Johnson and Andrew Barron, PTRF 2004

[5] L. Gross, Amer. J. Math. 1975

[6] D. Bakry, M. Emery, Springer 1985