POLYNOMIAL EQUIVALENCE OF THE KULLBACK INFORMATION FOR MIXTURE MODELS

Shaowei Lin
(with Carlos Améndola and Mathias Drton)
Singapore University of Technology and Design

IVC Vilnius 20180705

PART 1

Nonanalyticity of Kullback Information in Mixtures

Mild Analyticity Assumption

Kullback Divergence $K(q \| p) = \int q(x) \log \frac{q(x)}{p(x)} dx$

Kullback Information $K(\omega) = K(p_{\omega^*} \| p_\omega)$

Fisher Information $F(\omega^*) = \nabla^2 K(\omega^*)$

Analyticity of Kullback information often assumed in asymptotic theory.

Examples

1. *Fisher information is positive definite*
   (asymptotic normality of MLE, …)
2. *Log likelihood ratio* $\log q(x)/p(x)$ *is analytic*
   (asymptotics of marginal likelihood integral
   in singular learning theory)

S.Watanabe: Algebraic Geometry and Statistical Learning Theory, Cambridge Monographs on Applied and Computational Mathematics 25 (2009).

Nonanalyticity in Mixture Models

Exponential family $p_\omega(x) \propto \exp\{\omega x - A(\omega)\}$

Mixture model $\alpha p_{\vec{\omega}+\omega^*}(x) + (1-\alpha) p_{\omega^*}(x)$

Log-likelihood ratio

$$f(x|\alpha) = -\log \frac{\alpha p_{\vec{\omega}+\omega^*}(x) + (1-\alpha) p_{\omega^*}(x)}{p_{\omega^*}(x)}$$

$$= -\log\left\{1 + \alpha\left(e^{\vec{\omega}x - A(\vec{\omega}+\omega^*) + A(\omega^*)} - 1\right)\right\}$$

Power series coefficients grow quickly if x unbounded.

$$\frac{(-1)^{\mathrm{k}}}{k!}\frac{\partial^k f}{\partial \alpha^k}(x|0) = \frac{1}{k}\left\{e^{\vec{\omega}x - A(\vec{\omega}+\omega^*) + A(\omega^*)} - 1\right\}^k$$

$$\geq \frac{1}{k} e^{k\vec{\omega}x/2} \qquad \text{for all } x \in \mathcal{X}$$

Here, $\mathcal{X}$ is the set of all x where $\vec{\omega}\, x$ is sufficiently large.

Nonanalyticity in Mixture Models

Kullback information $K(\alpha) = \int p_{\omega^*}(x) f(x|\alpha) dx$

Power series coefficients $\frac{1}{k!}\frac{\partial^k K}{\partial \alpha^k}(0) = \int p_{\omega^*}(x) \frac{1}{k!}\frac{\partial^k f}{\partial \alpha^k}(x|0) dx$

As $k \to \infty$, the size of the coefficient is dominated by

$$\int_{\mathcal{X}} p_{\omega^*}(x) \frac{(-1)^{\mathrm{k}}}{k!} \frac{\partial^k f}{\partial \alpha^k}(x|0) dx \geq \int_{\mathcal{X}} p_{\omega^*}(x) \frac{1}{k} e^{\frac{k\vec{\omega}x}{2}} dx$$

$$= \frac{1}{k} \int_{\mathcal{X}} e^{\{\omega^* + k\vec{\omega}/2\}x - A(\omega^*)} dx$$

$$= \frac{1}{k} e^{A(\omega^* + k\vec{\omega}/2) - A(\omega^*)} C_k$$

where $C_k = \int_{\mathcal{X}} p_{\omega^* + k\vec{\omega}/2}(x) dx \to 1$.

Thus, if $A(\omega^* + k\vec{\omega}/2)/k \to \infty$, then radius of convergence is 0 and $K(\alpha)$ is nonanalytic. Examples: Gaussian, Poisson, gamma.

Gaussian Mixtures

Example. One-dimensional Gaussian.

$$\omega_1 = \frac{\mu}{\sigma^2},\ x_1 = t$$

$$\omega_2 = \frac{1}{\sigma^2},\ x_2 = -\frac{1}{2}t^2$$

$$A(\omega) = \frac{1}{2}\left(\frac{\omega_1^2}{\omega_2} + \log\frac{2\pi}{\omega_2}\right)$$

Now, if $\vec{\omega} = (\vec{\omega}_1, 0)$, then $\frac{1}{k}A\left(\omega^* + \frac{k\vec{\omega}}{2}\right) \approx \frac{\vec{\omega}_1^2}{2\omega_2}k \to \infty$ as $k \to \infty$.

Hence, the Kullback information of Gaussian mixtures is nonanalytic.

See work of Watanabe, Yamazaki and Aoyagi (2004) where they also proved that $K(\omega)$ is equivalent to a polynomial. We extend their results to polynomial families using algebraic geometry.

S. Watanabe, K. Yamazaki, and M. Aoyagi: Kullback information of normal mixture is not an analytic function, Technical Report of IEICE, NC2004-50 (2004).

PART 2

Equivalence of Kullback Information to a Polynomial

Equivalence is Enough

Definition. Loss functions $f, g: \Omega \to \mathbb{R}_{\geq 0}$ are *equivalent* if there exist constants $c_1, c_2 > 0$ such that $c_1 g(\omega) \leq f(\omega) \leq c_2 g(\omega)$ for all $\omega \in \Omega$. Easy to check for reflexivity, symmetry and transitivity.

Equivalent functions produce the same asymptotic properties.

Example. For large N, the log Laplace integral is asymptotically

$$\log Z_f(N) = \log \int_\Omega e^{-Nf(\omega)} d\omega$$
$$\approx -\lambda_f \log N + \left(\theta_f - 1\right) \log\log N + C$$

if $f(\omega)$ vanishes in Ω. $\left(\lambda_f, \theta_f\right)$ is the real log canonical threshold of f. If f is equivalent to g, then their RLCTs are the same.

Question. Is Kullback information equivalent to an analytic function?

Milder Upper-Bound Assumption

Fix ω^* and rewrite the Kullback information as

$$K(\omega) = K(p_{\omega^*} \| p_\omega) = \int \left| \frac{p_\omega(x)}{p_{\omega^*}(x)} - 1 \right|^2 S\left(\frac{p_\omega(x)}{p_{\omega^*}(x)}\right) p_{\omega^*}(x) dx$$

where real-analytic $S(t)$ satisfies $-\log \mathrm{t} = -(t-1) + (t-1)^2 S(t)$.

Assumption 1. Parameter space Ω is compact and semi-analytic.

Assumption 2. There exists real-analytic $\bar{S}(x)$ such that

$p_{\omega^*}(x) \le \bar{S}(x)$ and $S\left(\frac{p_\omega(x)}{p_{\omega^*}(x)}\right) \le \bar{S}(x)$ for all $\omega \in \Omega$;

$\bar{K}(\omega) = \int \left| \frac{p_\omega(x)}{p_{\omega^*}(x)} - 1 \right|^2 \bar{S}(x) p_{\omega^*}(x) dx$ is finite, real-analytic.

Polynomial Families

Definition. A family $\{p_\omega\}$ of distributions is *polynomial* if

1. Each moment $m_\omega(\gamma) = \mathbb{E}\left[X_1^{\gamma_1} \cdots X_m^{\gamma_m}\right]$ exists and is polynomial in ω;
2. Each p_ω is defined uniquely by its moments.

See work of Belkin and Sinha (2010).

Example. Gaussian, Poisson, gamma, binomial distributions are polynomial, but Weibull, Cauchy distributions are not.

Proposition. Mixtures of polynomial families are polynomial.

M. Belkin, and K. Sinha: Polynomial learning of distribution families, 51st Annual IEEE Symposium on Foundations of Computer Science (2010), 103-112.

Equivalence to Sum of Squares

Despite being nonanalytic, the Kullback information is equivalent to a polynomial, so asymptotic laws of the mixture model may be derived.

Main Theorem. Under Assumptions 1 & 2, if $\{p_\omega\}$ is a polynomial family, then $K(\omega)$ is equivalent to the polynomial

$$M(\omega) = \sum_{1\leq|\gamma|\leq\ell} \left(m_\omega(\gamma) - m_{\omega^*}(\gamma)\right)^2 .$$

Corollary. The RLCT of $K(\omega)$ equals the RLCT of the ideal

$$\langle m_\omega(\gamma) - m_{\omega^*}(\gamma) \colon 1 \leq |\gamma| \leq \ell \,\rangle \;.$$

We may thus use ideal-theoretic techniques to compute the RLCT.

Gaussian Mixtures

Example. Two-dimensional Gaussians with standard variance.

p_ω : (α_1, α_2)-mixture of Gaussians with means $(\mu_{11}, \mu_{12}), (\mu_{21}, \mu_{22})$

p_{ω^*} : unmixed Gaussian with mean (μ_1^*, μ_2^*)

The Kullback information $K(\omega)$ is equivalent to the polynomial

$$\begin{aligned} P(\omega) = &(\alpha_1\mu_{11} + \alpha_2\mu_{21} - \mu_1^*)^2 + \\ &(\alpha_1\mu_{12} + \alpha_2\mu_{22} - \mu_2^*)^2 + \\ &(\alpha_1\mu_{11}^2 + \alpha_2\mu_{21}^2 - \mu_1^{*2})^2 + \\ &(\alpha_1\mu_{11}\mu_{12} + \alpha_2\mu_{21}\mu_{22} - \mu_1^*\mu_2^*)^2 + \\ &(\alpha_1\mu_{12}^2 + \alpha_2\mu_{22}^2 - \mu_2^{*2})^2 + \\ &(\alpha_1\mu_{11}^3 + \alpha_2\mu_{21}^3 - \mu_1^{*3})^2 + \\ &(\alpha_1\mu_{11}^2\mu_{12} + \alpha_2\mu_{21}^2\mu_{22} - \mu_1^{*2}\mu_2^*)^2 + \\ &(\alpha_1\mu_{11}\mu_{12}^2 + \alpha_2\mu_{21}\mu_{22}^2 - \mu_1^*\mu_2^{*2})^2 + \\ &(\alpha_1\mu_{12}^3 + \alpha_2\mu_{22}^3 - \mu_2^{*3})^2 \end{aligned}$$

Hence, the maximum likelihood variety $\{\omega : K(\omega) = 0\}$
is a fiber over the secant map of Veronese embeddings.

PART 3

Proof of Equivalence

Comparing Distributions

Let $\phi_\omega(t) = \int e^{itx} p_\omega(x)dx$ be the characteristic function of $p_\omega(x)$.
If all the moments $m_\omega(\gamma)$ of p_ω exist, then

$$\phi_\omega(t) = \sum_\gamma \frac{i^{|\gamma|}}{|\gamma|!} \binom{|\gamma|}{\gamma} t^\gamma m_\omega(\gamma).$$

Kullback loss	$K(\omega) = K(p_{\omega^*} \| p_\omega)$
Density loss	$P(\omega) = \int \left(p_\omega(x) - p_{\omega^*}(x)\right)^2 dx$
Characteristic loss	$\Phi(\omega) = \int \left(\phi_\omega(t) - \phi_{\omega^*}(t)\right)^2 dt$
Moment loss	$M(\omega) = \sum_{1\le|\gamma|\le\ell} \left(m_\omega(\gamma) - m_{\omega^*}(\gamma)\right)^2$

Proof of Main Theorem

Step 1. Under Assumptions 1 & 2,
show that $K(\omega)$ is equivalent to $P(\omega)$.
Use resolution of singularities.

Step 2. Show that $P(\omega)$ is equal to $\Phi(\omega)$.
Use Fourier transform and Parseval's Theorem.

Step 3. Assuming $\{p_\omega\}$ is a polynomial family and Ω is compact,
show that $\Phi(\omega)$ is equivalent to $M(\omega)$.
Use functional analysis, Hilbert Basis Theorem,
and the Cauchy-Schwarz inequality.

Step 3. Ideal-Theoretic Approach.

Step 3. Assuming $\{p_\omega\}$ is a polynomial family and Ω is compact, show that $\Phi(\omega)$ is equivalent to $M(\omega)$.
Use Hilbert Basis Theorem, functional analysis, and the Cauchy-Schwarz inequality.

$$\Phi(\omega) = \int \left\{\sum_\gamma \frac{i^{|\gamma|}}{|\gamma|!}\binom{|\gamma|}{\gamma} t^\gamma \left(m_\omega(\gamma) - m_{\omega^*}(\gamma)\right)\right\}^2 dt$$

$$M(\omega) = \sum_{1\le|\gamma|\le\ell} \left(m_\omega(\gamma) - m_{\omega^*}(\gamma)\right)^2$$

Lemma. Let $F(\omega) = \int f_x(\omega)^2\, dx$ and $G(\omega) = \int g_t(\omega)^2\, dt$ be integrals/sums of squares with $f_x(\omega), g_t(\omega)$ real-analytic in ω.
If Ω is compact and $f_x(\omega) \in \langle g_t(\omega)\rangle$ for each x, then there exists a constant $c > 0$ such that $F(\omega) \le cG(\omega)$ for all $\omega \in \Omega$.
Proof. Cauchy-Schwarz inequality.

Step 1. Resolution of Singularities

Recall that

$$K(\omega) = \int \left| \frac{p_\omega(x)}{p_{\omega^*}(x)} - 1 \right|^2 S\left(\frac{p_\omega(x)}{p_{\omega^*}(x)}\right) p_{\omega^*}(x) dx,$$

$$P(\omega) = \int \left| \frac{p_\omega(x)}{p_{\omega^*}(x)} - 1 \right|^2 p_{\omega^*}(x)\, p_{\omega^*}(x) dx,$$

$p_{\omega^*}(x) \le \bar{S}(x)$ and $S\left(\frac{p_\omega(x)}{p_{\omega^*}(x)}\right) \le \bar{S}(x)$,

$$\bar{K}(\omega) = \int \left| \frac{p_\omega(x)}{p_{\omega^*}(x)} - 1 \right|^2 \bar{S}(x)\, p_{\omega^*}(x) dx.$$

Step 1. Resolution of Singularities

a. Since $\overline{K}(\omega)$ is analytic, there exists resolution of singularities $\pi: \mathcal{M} \to \Omega$ with $\mathcal{M}$ compact such that in each of chart of $\mathcal{M}$,

$$\overline{K}(\pi(\mu)) = \int \left|\frac{p_{\pi(\mu)}(x)}{p_{\omega^*}(x)} - 1\right|^2 \bar{S}(x) p_{\omega^*}(x) dx = \mu^{2\kappa}.$$

b. By comparing terms, there exists real-analytic $a(x,\mu)$ such that

$$\frac{p_{\pi(\mu)}(x)}{p_{\omega^*}(x)} - 1 = a(x,\mu)\mu^{\kappa}.$$

c. In each chart, $\mu^{2\kappa} \geq K\big(\pi(\mu)\big) = b_K(\mu)\mu^{2\kappa}$ where

$$b_K(\mu) = \int a(x,\mu)^2\, S\left(\frac{p_{\pi(\mu)}(x)}{p_{\omega^*}(x)}\right) p_{\omega^*}(x) dx.$$

The chart is compact, so $b_K(\mu)$ is bounded below.
Hence, $K\big(\pi(\mu)\big)$ is equivalent to $\mu^{2\kappa}$ in the chart.

Step 1. Resolution of Singularities

d. Similarly, $\mu^{2\kappa} \geq P\big(\pi(\mu)\big) = b_P(\mu)\mu^{2\kappa}$ where

$$b_P(\mu) = \int a(x,\mu)^2\, p_{\omega^*}(x)^2 dx.$$

The chart is compact, so $b_P(\mu)$ is bounded below.
Hence, $P\big(\pi(\mu)\big)$ is equivalent to $\mu^{2\kappa}$ in the chart.

e. Since $K(\pi(\mu))$ and $P\big(\pi(\mu)\big)$ are both equivalent to $\mu^{2\kappa}$ in every chart and there are finitely many charts in $\mathcal{M}$, they are equivalent over $\mathcal{M}$ and hence over Ω as well.

References

1. M. Belkin, and K. Sinha: Polynomial learning of distribution families, 51st Annual IEEE Symposium on Foundations of Computer Science (2010), 103-112.
2. S. Lin: Ideal-theoretic Strategies for Asymptotic Approximation of Marginal Likelihood Integrals, Journal of Algebraic Statistics 8:1 (2017).
3. S.Watanabe: Algebraic Geometry and Statistical Learning Theory, Cambridge Monographs on Applied and Computational Mathematics 25 (2009).
4. S. Watanabe, K. Yamazaki, and M. Aoyagi: Kullback information of normal mixture is not an analytic function, Technical Report of IEICE, NC2004-50 (2004) 41-46.

Thank you ☺