A simple mixture model for probability density estimation based on a quasi divergence

Osamu Komori[1,2] and Shinto Eguchi[2] [1]University of Fukui, [2]The Institute of Statistical Mathematics, Japan

1 Introduction

Let $\mathcal{F}$ be the space of all probability density functions with respect to a carrier measure λ, and let ν be a one-to-one mapping from $\mathcal{F}$ to $\mathcal{F}$.

Definition 1 *Let a functional D on $\mathcal{F}\times\mathcal{F}$ be a quasi-divergence if*

$$D(g,f)\geq 0 \tag{1}$$

with equality if and only if $f=\nu(g)$. Further D is called a divergence if ν is an identity mapping.

Let us take a way to introduce a class of quasi divergences. For this fix a function ϕ that is a strictly increasing and concave function. Thus we define a ϕ cross entropy as

$$C_\phi(g,f)=-\int\phi(f(x))g(x)d\lambda(x). \tag{2}$$

The corresponding loss for a data set $\{x_1,\ldots,x_n\}$ is given as

$$L_\phi(\theta)=-\frac{1}{n}\sum_{i=1}^{n}\phi(f_\theta(x_i)). \tag{3}$$

An argument from a variational calculus leads to the following inequality:

Proposition 1 *It holds for any f and g of $\mathcal{F}$ that*

$$C_\phi(g,f)\geq C_\phi(g,\xi(g)), \tag{4}$$

where $\xi(g)=\phi'^{-1}(c/g)$. Here $c>0$ is a normalizing factor satisfying

$$\int\phi'^{-1}\Big(\frac{c}{g(x)}\Big)d\lambda(x)=1. \tag{5}$$

As a result, we can define

$$D_\phi(g,f)=\int\{\phi(\xi(g(x)))-\phi(f(x))\}g(x)d\lambda(x), \tag{6}$$

where D_ϕ is called a quasi-divergence with the mapping ξ.

Remark 1 *There are two ways to make a divergence from the quasi-divergence. First, if we deform D_ϕ as*

$$D^*_\phi(g,f)=D_\phi(\xi^{-1}(g),f), \tag{7}$$

then $D^*_\phi(g,f)$ is a divergence given as in Definition 1. In effect,

$$D^*_\phi(g,f)=\int\frac{\phi(g(x))-\phi(f(x))}{\phi'(g(x))}d\lambda(x)\Big/\int\frac{1}{\phi'(g(x))}d\lambda(x), \tag{8}$$

which is nothing but the generalized KL divergence [1]

$$\mathbb{E}^{(\phi)}_g\{\phi(g(X))-\phi(f(X))\}, \tag{9}$$

where $\mathbb{E}^{(\phi)}_g$ denotes the generalized expectation with respect to g. For example, if we take a specific function as $\phi(f)=(f^\beta-1)/\beta$, then

$$D^*_\phi(g,f)=\frac{1}{\beta}\Big(1-\int g^{1-\beta}f^\beta d\lambda\Big)\Big(\int g^{1-\beta}d\lambda\Big)^{-1}, \tag{10}$$

which is proportional to the α-divergence with a relation to $\alpha=2\beta-1$. Second, we deform D_ϕ as

$$\begin{aligned}D^{**}_\phi(g,f)&=D_\phi(g,\xi(f))\\&=-\int\{\phi(\xi(f(x)))-\phi(\xi(g(x)))\}g(x)d\lambda(x).\end{aligned} \tag{11}$$

If we take as $\phi(f)=(f^\beta-1)/\beta$,

$$D^{**}_\phi(g,f)=-\frac{1}{\beta}\Bigg[\frac{\int f^{\frac{\beta}{1-\beta}}gd\lambda}{\big(\int f^{\frac{1}{1-\beta}}d\lambda\big)^\beta}-\Big(\int g^{\frac{1}{1-\beta}}d\lambda\Big)^{1-\beta}\Bigg], \tag{12}$$

which is nothing but the γ-power divergence when $\gamma=\beta/(1-\beta)$ [2].

2 Probability density estimation

We consider a simple mixture model

$$f_\theta(x)=\theta^\top f(x), \tag{13}$$

where $f(x)=(f_1(x),\cdots,f_J)$ and $\theta=(\theta_1,\cdots,\theta_J)$. Then we consider

$$\mathcal{L}_\phi(\theta,\omega)=\mathcal{L}_\phi(\theta)+\omega\sum_{j=1}^{J}|\theta_j|, \tag{14}$$

where $\mathcal{L}_\phi(\theta)=1/n\sum_{i=1}^n\phi\Big(\sum_{j=0}^J\theta_jf_j(x_i)\Big)$, $\theta_0=1-\sum_{j=1}^J\theta_j$ and $\theta_j\geq 0$ for $j=1,\ldots,J$. An non-informative probability density function $f_0(x)$ is introduced to enable us to do selection of density functions.

Then we take the gradient descent approach similar to that of [3], where the gradient $\mathbf{g}(\theta)=(\mathbf{g}_1(\theta),\mathbf{g}_2(\theta),\ldots,\mathbf{g}_J(\theta))^\top$ is defined by

$$\mathbf{g}_j(\theta)=\begin{cases}\frac{\partial}{\partial\theta_j}\mathcal{L}_\phi(\theta)+\omega\theta_j & \text{if } \theta_j\neq 0\\ \frac{\partial}{\partial\theta_j}\mathcal{L}_\phi(\theta)-\omega\mathrm{sign}(\frac{\partial}{\partial\theta_j}\mathcal{L}_\phi(\theta)) & \text{if } \theta_j=0 \text{ and } |\frac{\partial}{\partial\theta_j}\mathcal{L}_\phi(\theta)|>\omega\\ 0 & \text{otherwise,}\end{cases} \tag{15}$$

for $j=1,\ldots,J$. The range of optimization for a scalar ρ is given as

$$\rho_{\mathrm{edge}}(\theta)=\min_{j=1,\ldots,J}\Big\{\frac{\theta_j}{\mathbf{g}_j(\theta)}\Big|\ \mathrm{sign}(\theta_j)=-\mathrm{sign}(\mathbf{g}_j(\theta))\neq 0\Big\}. \tag{16}$$

---

1. Set $\theta_0^{(1)}=1$ and $\theta_j^{(1)}=0$ for $j=1,\ldots,J$.
2. For $t=2,\ldots,T$,
   (a) Update $\theta_j^{(t)}=\max\big(0,\theta_j^{(t-1)}+\rho_{\mathrm{opt}}\,\mathbf{g}_j(\theta^{(t-1)})\big)$ for $j=1,\ldots,J$ where
   $$\rho_{\mathrm{opt}}=\underset{0\leq\rho\leq\rho_{\mathrm{edge}}(\theta^{(t-1)})}{\mathrm{argmax}}\ \mathcal{L}_\phi\Big(\theta^{(t-1)}+\rho\,\mathbf{g}(\theta^{(t-1)}),\omega\Big) \tag{17}$$
   (b) Update $\theta_0^{(t)}=\max\big(0,1-\sum_{j=1}^J\theta_j^{(t)}\big)$.
3. Apply the EM algorithm to $f_j(x)$ $(j\neq 0)$ in the active set $\mathcal{A}$ with the initial value $\theta_j^{(T)}$ to obtain $\hat\theta_j$. And set $\hat\theta_j=0$ for $f_j(x)$ in $\mathcal{A}^c$.
4. Output
   $$\hat f_\phi(x)=\frac{1}{\phi'\Big(\sum_{j=1}^J\hat\theta_jf_j(x)\Big)}\Bigg/\int\frac{1}{\phi'\Big(\sum_{j=1}^J\hat\theta_jf_j(x)\Big)}d\lambda. \tag{18}$$

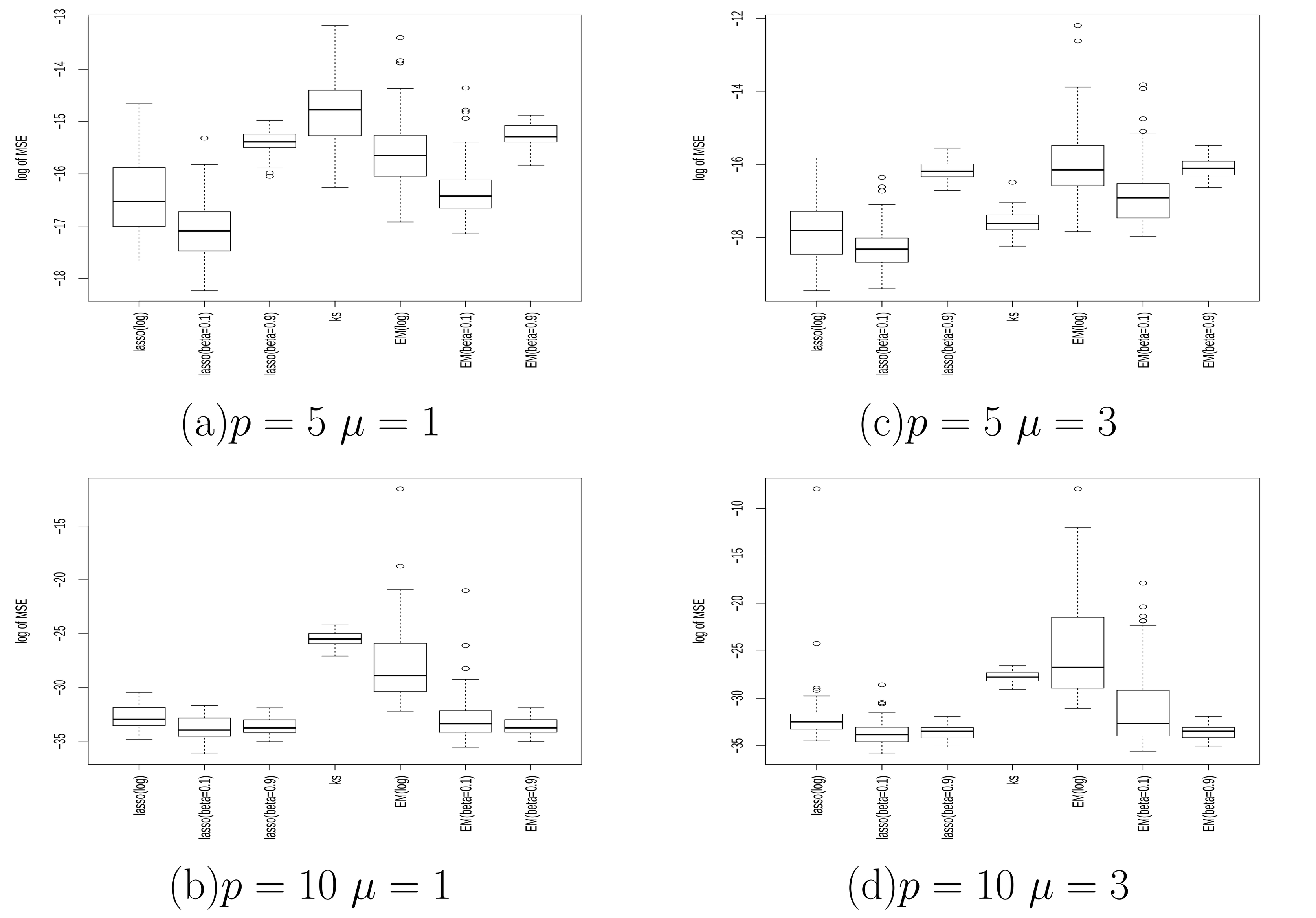
---

3 Simulation studies

We generate random variables from the normal mixture as

$$x_i\sim\pi_0N(0,I_p)+\pi_1N(\mu_1,I_p)+\pi_2N(\mu_2,I_p),\ i=1,\ldots,n \tag{19}$$

where $\pi_0=\pi_1=\pi_2=1/3$, $\mu_1=(\mu,\ldots,\mu)^\top$, $\mu_2=-\mu_1$ and $n=90$. And we consider $f_0(x)=f(x,0,1000\times I_p)$. We compare the performance of lasso algorithm based on $\phi(t)=\log(t)$ and $(t^\beta-1)/\beta$ with $\beta=0.1$ and $\beta=0.9$, and the kernel density estimation method by [4] using the `R` package `ks`.



(a)$p=5$ $\mu=1$ (c)$p=5$ $\mu=3$

(b)$p=10$ $\mu=1$ (d)$p=10$ $\mu=3$

Figure 1. Boxplots of log of MSE for lasso method (lasso(log), lasso(beta=0.1) and lasso(beta=0.9)) and ks and EM-like algorithm (EM(log) EM(beta=0.1), EM(beta=0.9)) based on 50 repetitions of simulations.

References

[1] Eguchi, S., Komori, O and Ohara, A. Information geometry associated with two generalized means (in preparation).

[2] Fujisawa, H. and Eguchi, S. (2008) Robust parameter estimation with a small bias against heavy contamination.*Journal of Multivariate Analysis* **99**, 2053–2081.

[3] Goeman, J. J. (2010) L1 penalized estimation in the Cox proportional hazards model.*Biometrical Journal* **20**, 3375–3387.

[4] Duong, T. (2010) ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software* **21**, 1–16.