Mixture of exponential distribution: Properties of the MLE

David Pham, Nicolas Ruckstuhl

2017-04-10

Slides availability

The slides are available on

github.com/davidpham87/sfs_seminar_2017S

Mixture of Exponential Distributions by Nicholas Jewell

- Paper in *The Annals of Statistics*, published in 1982.
- Author is statistician from United Kingdom who obtained his PhD in 1976.
- Currently researches in Berkeley in biostatistics, epidemiological data analysis, genomics and *survival analysis*.
- The paper was aimed to study survival analysis originally.

Exponential Distribution

1. How can we find absolutely continuous distributions, or how can we remember them?
2. How can we extend a family of distributions?

Exponential Distribution

1. How can we find absolutely continuous distributions, or how can we remember them?
2. How can we extend a family of distributions?

Take any positive function g such that $\int g(u)du < \infty$ and define the density as

$$f(x) = \frac{g(x)}{\int g(u)du}.$$

For example with $g(x) = \exp(-x)\, 1(x \geq 0)$,

$$\int_0^\infty g(x)dx = 1 < \infty.$$

Suppose X has g as density function. Then $Y = X/\lambda$ has density function

$$\partial_y\{P(Y \leq y)\} = \partial_y\{P(X \leq \lambda y)\} = \lambda g(\lambda y) = \lambda \exp(-\lambda y).$$

Exponential Distribution

The exponential distribution $\mathrm{Exp}(\lambda)$ with parameter $\lambda \in \mathbb{R}_+^*$ has a cumulative, resp. density, function defined as

$$F(x) = 1(x \geq 0)\ (1 - \exp\{-\lambda x\}),$$
$$f(x) = 1(x \geq 0)\ \lambda \exp(-\lambda x).$$

Application of the Exponential Distribution

- Describes lengths of interval of arrival times in Poisson process.
- Interpreted as the continous counterpart of geometric distribution (aka. waiting time). Application in particle decays, telephone calls, time until defaults.
- Memoryless property, suited for hazard rate portion and failure rates.

$$P(X > s + t | X > s) = P(X > t), \quad t, s \geq 0.$$

- A good approximate for extreme values in hydrology.

Properties and Relationship with other distributions

If $X, X_1, \dots, X_n$ are i.i.d and follows $\mathrm{Exp}(\lambda)$, then

- $X/\kappa \sim \mathrm{Exp}(\kappa\lambda)$.
- When $\lambda = 1/2$, then $X \sim \chi^2_2$.
- If $Y \sim \mathrm{Exp}(\nu)$ then $\min(X, Y) \sim \mathrm{Exp}(\lambda + \nu)$.
- $\sum_i X_i \sim \mathrm{Gamma}(n, \lambda)$, when $f^{\Gamma}_{\alpha,\beta}(x) \propto \beta^\alpha x^{\alpha-1} \exp(-\beta x)$.
- If $U_i \sim U(0,1)$, $i = 1, \dots, n$ and U_i are independent, then $n \min_i U_i \to \mathrm{Exp}(1)$.

Intuition for Mixture Distribution

Mixture models are best represented as hierarchical models. These are best understood through their sampling schemes in the discrete case.

Intuition for Mixture Distribution

Mixture models are best represented as hierarchical models. These are best understood through their sampling schemes in the discrete case.

1. First, select n different models and label them with $1, \ldots, n$ (e.g. $\mathcal{N}(\mu_i, \sigma)$, $i = 1, \ldots, n$).
2. Select a probability discrete distribution M over $1, \ldots, n$, considered as the weights of the previous models.
3. Then to sample an observation from the mixture models, sample one observation from M, then sample from the corresponding models.

Intuition for Mixture Distribution

Mixture models are best represented as hierarchical models. These are best understood through their sampling schemes in the discrete case.

1. First, select n different models and label them with $1, \dots, n$ (e.g. $\mathcal{N}(\mu_i, \sigma)$, $i = 1, \dots, n$).
2. Select a probability discrete distribution M over $1, \dots, n$, considered as the weights of the previous models.
3. Then to sample an observation from the mixture models, sample one observation from M, then sample from the corresponding models.

Visually, imagine a dice with n faces with the definition of a model on each face. You roll the dice and select the model on the resulting face. The mixture model is the distribution of this hierarchical model.

Definition for exponential distribution

Let W be a random variable with mixture distribution $M(\lambda)$ with support $\mathbb{R}_+^*$ and density $m(\lambda)$. The quantity of interest is

$$\begin{aligned} F(t) &= \int_0^\infty \{1 - \exp(-\lambda t)\}\, dM(\lambda) \\ &= \int_0^\infty \{1 - \exp(-\lambda t)\}\, m(\lambda)\, d\lambda \\ &= \sum_{\{\lambda : m(\lambda) > 0\}} \{1 - \exp(-\lambda t)\}\, m(\lambda), \end{aligned}$$

where the last line only make sense when W is discrete. If M is the mixture distribution, then we denote $f_M(t)$ the associated density

$$f_M(t) = \int_0^\infty \lambda \exp(-\lambda t) dM(\lambda) = \int_0^\infty \{\lambda \exp(-\lambda t)\}\, m(\lambda)\, d\lambda.$$

Theorem

Theorem

Let $F(t) = \int_0^\infty \{1 - \exp(-\lambda t)\}\, dM(\lambda)$ *be the distribution function of a mixture of exponentially distributed random variables with parameter* λ *for the unknown mixture measure* M*. Given* n *independent observations,* $t_1, \ldots, t_n$ *from* F*, the MLE of* M *exists and is unique.*

Existence of the MLE

Let $\mathcal{F}$ be the set of measures on $[0, \infty)$ with a total mass of at most 1 and $\mathcal{F}_0$ the measures in $\mathcal{F}$ with no mass at 0.

Existence of the MLE

Let $\mathcal{F}$ be the set of measures on $[0, \infty)$ with a total mass of at most 1 and $\mathcal{F}_0$ the measures in $\mathcal{F}$ with no mass at 0.

Define the function $\phi(M) = (\phi_1(M), \ldots, \phi_n(M))$ on $\mathcal{F}$ with

$$\phi_j(M) = \int_0^\infty \lambda \exp(-\lambda t_j) dM(\lambda), \quad j = 1, \ldots, n.$$

Existence of the MLE

Let $\mathcal{F}$ be the set of measures on $[0, \infty)$ with a total mass of at most 1 and $\mathcal{F}_0$ the measures in $\mathcal{F}$ with no mass at 0.

Define the function $\phi(M) = (\phi_1(M), \ldots, \phi_n(M))$ on $\mathcal{F}$ with

$$\phi_j(M) = \int_0^\infty \lambda \exp(-\lambda t_j) dM(\lambda), \quad j = 1, \ldots, n.$$

Thus, the log-likehood function can be expressed as

$$\begin{aligned}\Phi &= \sum_{j=1}^{n} \log\left\{ \int_0^\infty \lambda \exp(-\lambda t_j)\; dM(\lambda) \right\} = \sum_{j=1}^{n} \log\{\phi_j(M)\} \\ &= \Phi\{\phi(M)\}.\end{aligned}$$

Existence of the MLE (Cont'd)

The set $\phi(\mathcal{F}) = \{\phi(M) : M \in \mathcal{F}\}$ is a subspace of $\mathbb{R}^n$ which is compact and convex. Observe that Φ is strictly concave on $\phi(\mathcal{F})$.

Thus Φ attains its maximum at a unique point of $\phi(\mathcal{F})$. This proves the existence of the MLE.

Uniqueness of the MLE

Denote the maximum of Φ over $\phi(\mathcal{F})$ by $\hat{\beta} \in \mathbb{R}^n$ and let $\hat{M} \in \mathcal{F}$ be such that

$$\phi(\hat{M}) = \hat{\beta}.$$

It can be shown that $\hat{M}$ has no mass at 0, that is $\hat{M} \in \mathcal{F}_0$, and that it has finite support.

Uniqueness of the MLE

Denote the maximum of Φ over $\phi(\mathcal{F})$ by $\hat{\beta} \in \mathbb{R}^n$ and let $\hat{M} \in \mathcal{F}$ be such that

$$\phi(\hat{M}) = \hat{\beta}.$$

It can be shown that $\hat{M}$ has no mass at 0, that is $\hat{M} \in \mathcal{F}_0$, and that it has finite support.

Hence $\hat{\beta}$ uniquely determines all different points $\{\lambda_1, \ldots, \lambda_r\}$ in the support of $\hat{M}$ and the number r.

Uniqueness of the MLE

Denote the maximum of Φ over $\phi(\mathcal{F})$ by $\hat{\beta} \in \mathbb{R}^n$ and let $\hat{M} \in \mathcal{F}$ be such that

$$\phi(\hat{M}) = \hat{\beta}.$$

It can be shown that $\hat{M}$ has no mass at 0, that is $\hat{M} \in \mathcal{F}_0$, and that it has finite support.

Hence $\hat{\beta}$ uniquely determines all different points $\{\lambda_1, \ldots, \lambda_r\}$ in the support of $\hat{M}$ and the number r.

Let p_m be the mass of $\hat{M}$ at λ_m, for $m = 1, \ldots, r$. Then, for $j = 1, \ldots, n$, we have that

$$\hat{\beta}_j = \phi_j(\hat{M}) = \int_0^\infty \lambda e^{-\lambda t_j} d\hat{M}(\lambda) = \sum_{m=1}^{r} p_m \lambda_m e^{-\lambda_m t_j}.$$

Uniqueness of the MLE (cont'd)

Since we have proven the existence of the MLE, there exist at least one solution $(p_1, \ldots, p_r)$ to these equations.

Suppose $(q_1, \ldots, q_r)$ is another solution. Then it holds that

$$\sum_{m=1}^{r} (p_m - q_m)\, \lambda_m e^{-\lambda_m t_j} = 0, \quad j = 1, \ldots, n.$$

Uniqueness of the MLE (cont'd)

Since we have proven the existence of the MLE, there exist at least one solution $(p_1, \ldots, p_r)$ to these equations.

Suppose $(q_1, \ldots, q_r)$ is another solution. Then it holds that

$$\sum_{m=1}^{r} (p_m - q_m)\, \lambda_m e^{-\lambda_m t_j} = 0, \quad j = 1, \ldots, n.$$

Polya and Szego (1925) showed that any non-identically vanishing exponential polynomial of this form has at most $(r-1)$ distinct zeros. But $r \leq n$, which is also shown using the same result. This contradiction shows that the MLE is unique.

Consistency

Theorem

The sequence M_n converges in distribution with probability one to the true mixing distribution M_0.

We will prove a slightly weaker theorem, where we assume some additional regularity condition (*i.e.* limits and integral can be switched).

For any G in $\mathcal{F}_0$, denote f_g its density function given by $f_G(t) = \int \lambda \exp(-\lambda t) g(\lambda) d\lambda$, (if G is absolutely continuous).

Proof

- From measure theory, we know there exists a subsequence converging weakly to a positive measure M on $\mathbb{R}_+^*$ with total mass of at most 1.
- Hence, the empirical distribution F_n associated with $t_1, \ldots, t_n$ converges weakly to the distribution function F of M as $n \to \infty$.

We now have to prove that $M = M_0$.

As always, suppose $M \neq M_0$ and we need to find a contradiction. We will show that, under the assumption that $M \neq M_0$, the quantity

$$E\{f_{M_0}(t)/f_M(t)\} - 1,$$

where the expectation is taken with respect to the probability measure induced by f_{M_0}, is non-positive and positive at the same time, which is impossible. Denote for simplicity

$$\psi(G) = E\Big[\log\{f_G(t)/f_{M_0}(t)\}\Big], \quad G \in \mathcal{F}_0.$$

Observe that $\psi(M) \leq 0$

The non-positiveness of $\psi(M)$ is demonstrated thanks to Jensen's inequality (as *log* is concave)

$$\psi(M) = E\Big\{ \log \big[f_M(t)/f_{M_0}(t)\big] \Big\} \leq \log \Big\{ E\big[f_M(t)/f_{M_0}(t)\big] \Big\}.$$

Remark that,

$$E\big[f_M(t)/f_{M_0}(t)\big] = \int_{\mathbb{R}} \frac{f_M(t)}{f_{M_0}(t)} f_{M_0}(t) dt = \int_{\mathbb{R}} f_M(t) dt = 1.$$

Hence $\psi(M) \leq 0$ with equality if and only if $M = M_0$.

$E\{f_{M_0}(t)/f_M(t)\} - 1 \geq 0$

Observe that if $H = (1-\epsilon)M + \epsilon M_0$ for $0 \leq \epsilon \leq 1$, then

$$f_H(t) = (1-\epsilon)f_M(t) + \epsilon f_{M_0}(t).$$

Hence using the concavity of the log function, one gets

$$\begin{aligned}
\psi(H) &= E\Big\{ \log f_H(t) - \log f_{M_0}(t) \Big\} \\
&> E\Big[(1-\epsilon)\log f_M(t) + \epsilon \log f_{M_0}(t) - \log f_{M_0}(t) \Big] \\
&= E\Big[(1-\epsilon)\log f_M(t) - (1-\epsilon)\log f_{M_0}(t) \Big] = (1-\epsilon)\,\psi(M)
\end{aligned}$$

Hence

$$\frac{\psi(H) - \psi(M)}{\epsilon} > \frac{(1-\epsilon)\psi(M) - \psi(M)}{\epsilon} = -\psi(M) > 0,$$

as $\psi(M) < 0$, for $M \neq M_0$.

$E\{f_{M_0}(t)/f_M(t)\} - 1 \geq 0$ (cont'd)

Taking the limit yields

$$\lim_{\epsilon \to 0} \frac{\psi\{(1-\epsilon)M + \epsilon M_0\} - \psi(M)}{\epsilon} > 0.$$

The LHS of the equation can be developed, by using the Taylor approximation around 0 of $\log(1-x) = -x + O(x^2)$,

$$\begin{aligned}
&\epsilon^{-1}[\psi\{(1-\epsilon)M + \epsilon M_0\} - \psi(M)] \\
&= \epsilon^{-1}\, E\Big[\log\{(1-\epsilon)f_M(t) + \epsilon f_{M_0}(t)\} - \log f_M(t)\Big] \\
&= \epsilon^{-1}\, E\Big(\log[f_M(t) - \epsilon\{f_M(t) - f_{M_0}(t)\}] - \log f_M(t)\Big) \\
&= \epsilon^{-1}\, E\Big(\log[1 - \epsilon\{1 - f_{M_0}(t)/f_M(t)\}]\Big) \\
&= E\Big[\{f_{M_0}(t)/f_M(t) - 1\} + O(\epsilon C)\Big] \xrightarrow{\epsilon \to 0} E[f_{M_0}(t)/f_M(t)] - 1
\end{aligned}$$

Hence $E[f_{M_0}(t)/f_M(t)] - 1 > 0.$

$E\{f_{M_0}(t)/f_M(t)\} - 1 \le 0$

To show the non-positiveness, consider the log-likelihood function as a function of $\epsilon \in [0,1]$ of $H_\epsilon = (1-\epsilon)M_n + \epsilon M_0$. with respect of the actual data, that is the map

$$\epsilon \to \sum_{j=1}^{n} \log\left[f_{M_n}(t_j) + \epsilon\{f_{M_0}(t_j) - f_{M_n}(t_j)\}\right].$$

This function has a maximum at $\epsilon = 0$, because M_n is the maximum likelihood estimator. Then differentiating w.r.t ϵ, and evaluating at $\epsilon = 0$ yields

$$\sum_{j=1}^{n}\left(\frac{f_{M_0}(t_j)}{f_{M_n}(t_j)} - 1\right) \le 0 \Longrightarrow \frac{1}{n}\sum_{j=1}^{n}\frac{f_{M_0}(t_j)}{f_{M_n}(t)} \le 1.$$

$E\{f_{M_0}(t)/f_M(t)\} - 1 \leq 0$ (cont'd)

It follows that

$$1 \geq \frac{1}{n}\sum_{j=1}^{n}\frac{f_{M_0}(t_j)}{f_{M_n}(t_j)} = \int_0^\infty \frac{f_{M_0}(t)}{f_{M_n}(t)} dF_n(t) = E\left\{\frac{f_{M_0}(t)}{f_M(t)}\right\},$$

where the last equality is justified by our regularity conditions (when limit and integral can be switched for M_n and F_n).

Proof (end)

Hence under the assumption $M \neq M_0$, the quantity

$$E\{f_{M_0}(t)/f_M(t)\} - 1$$

is nonpositive and positive at the same time, which is impossible. Hence $M = M_0$, and thus any convergent subsequence of $\{M_n\}$ converges to M_0 with probability one.

Conclusion

- Introduction to the exponential distribution and mixture models.
- Presentation of the existence, uniqueness and consistency of the ML estimate for mixture of exponential distribution.

Question?

Thanks for your attention! Any questions?