Local moment matching with Erlang mixture under automatic roughness penalization

IWSM 2023 @ Dortmund

Oskar Laverny[1]
Philippe Lambert[1,2]

July 17, 2023

[1] Institut de statistiques, biostatistiques et actuariat, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[2] Institut de Mathématique, Université de Liège, Liège, Belgium.

Table of content

Local moment matching problem

Local moment matching problem

Definition (Local moment matching problem on $\mathbb{R}_+$)

Setup: Let $X_1, \ldots X_n$ be a n-sample of the positive random variable X, and $B_1, \ldots B_J$ a finite partition of $\mathbb{R}_+$, $\boldsymbol{k} \in \mathbb{N}^J$ a vector of integers.

Observations: We only observe the following local moments:

$$\hat{\boldsymbol{\pi}} = \left\{ \hat{\pi}_j = \frac{1}{N} \sum_{i=1}^{N} 1_{X_i \in B_j} : j \in 1, \ldots, J \right\} \text{ and}$$

$$\hat{\boldsymbol{\mu}} = \left\{ \hat{\mu}_{j,k} = \frac{1}{N} \sum_{i=1}^{N} X_i^k 1_{X_i \in B_j} : j \in 1, \ldots J, k \in 1, \ldots, k_j \right\}.$$

Goal: Estimate the distribution of X from $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}})$.

This kind of summarized information may be due to confidentiality issues (e.g. GDPR). See Lambert[1] for a similar problem on a bounded support.

[1] Philippe Lambert. "Nonparametric density estimation and risk quantification from tabulated sample moments". In: *Insurance: Mathematics and Economics* 108 (2023), pp. 177–189.

Example (LogNormal simulated example)

This local moment problem is drawn from $N = 750$ observations of a $\texttt{LogNormal}(0, 0.5)$. In this example, the three first lines give estimated boxed moments. The fourth line is a little more involved: it prescribes a Value-At-Risk b_{J-1} and a Tail-Value-at-Risk $\hat{\mu}_{J-1,1}$ at the quantile level $1 - \hat{\pi}_4$. This data structure is classical when modeling insurance losses, which are usually divided into *attritional* and *large* losses.

j	$[b_{j-1}, b_j[$	n_j	k_j	π_j	$\hat{\mu}_{j,1}$	$\hat{\mu}_{j,2}$	$\hat{\mu}_{j,3}$	$\hat{\mu}_{j,4}$
1	$[0.000, 0.969[$	375	4	0.500	0.341	0.249	0.191	0.151
2	$[0.969, 1.877[$	300	4	0.400	0.535	0.742	1.065	1.581
3	$[1.877, 3.058[$	67	4	0.089	0.197	0.442	1.002	2.305
4	$[3.058, \infty[$	8	1	0.011	0.038			

Data loglikelihood

Definition (Theoretical moments)

From the (unknown) distribution of the r.v. X, we can construct :

$$\boldsymbol{\pi} = \{\pi_j = \mathbb{P}(X \in B_j) : j \in 1, ..., J\} \text{ and}$$
$$\boldsymbol{\mu} = \left\{\mu_{j,k} = \mathbb{E}\left(X^k \mathbb{1}_{X \in B_j}\right) : j \in 1, ...J, k \in 1, ..., k_j\right\}$$
$$\boldsymbol{\Sigma} = \left\{\Sigma_{(j,k),(i,m)} = \mu_{j,k+m}\mathbb{1}_{j=i} - \mu_{j,k}\mu_{i,m}\right\}$$

Therefore:

(i) $N\hat{\boldsymbol{\pi}} \sim \texttt{Multinomial}(\boldsymbol{\pi}, N)$

(ii) Conditionally on $\hat{\boldsymbol{\pi}}$, due to CLT, $\hat{\boldsymbol{\mu}} \sim \texttt{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/N)$ when $N \to \infty$.

Hence, the (approximate) loglikelihood of the model is given by:

$$\ell_0(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left\{\hat{\boldsymbol{\pi}}' \log(\boldsymbol{\pi})\right\} + \left\{-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_{\Sigma}^2\right\}, \quad (1)$$

Regularized Erlang Mixtures

Definition (Mixtures of Gammas)

A positive real random variable X is said to be $\texttt{MixedGamma}(\nu, \theta)$ distributed, with mixing probability measure $\nu \in \mathcal{M}_+(\mathbb{R})$ and scale $s \in \mathbb{R}_+$, iff it has density

$$f(x) = \int_0^\infty \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha} \nu(d\alpha).$$

When $\mathrm{Supp}(\nu) \subseteq \mathbb{N}$, that is e.g. $\nu = \sum_{i \in \mathbb{N}} \omega_i \delta_i$, we say that $X \sim \texttt{MixedErlang}(\omega, \theta)$.

Theorem (Tijms[2])

The set of Erlang Mixtures is dense in the set of probability distributions over $\mathbb{R}_+$.

Note: There are multivariate extensions of the result, see Theorem 2.1 in Lee & Lin[3].

Remark

For $X \sim \texttt{MixedGamma}(\nu, \theta)$, $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are easily computed. Indeed,

$$\mathbb{E}\left(X^k \mathbb{1}_{X \in [a,b[}\right) = \theta^k \int \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} \left(\gamma\left(\frac{b}{\theta}, \alpha + k\right) - \gamma\left(\frac{a}{\theta}, \alpha + k\right)\right) \nu(d\alpha), \tag{2}$$

with these implicit definitions, we setup $\ell(\boldsymbol{\omega}, \theta) = \ell_0(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

[2] Henk C Tijms. *Stochastic models: an algorithmic approach*. Vol. 303. John Wiley & Sons Incorporated, 1994.

[3] Simon Lee and X Sheldon Lin. "Modeling Dependent Risks with Multivariate Erlang Mixtures". In: *Astin Bulletin* 42.1 (2012), pp. 153–180.

Erlang Mixtures models often require regularisation

Erlang Mixtures have been used as modeling tools for insurence purposes for a long time, see (among others) pioneering work from Lee & Lin[4]. However, they have a spiky behavior, which drives us toward roughness penalisation.

Definition (Roughness penalty from Gui, Huang & Lin[5])

$$\widetilde{\mathrm{Pen}}_r(\lambda, f) = \frac{\lambda}{2}\int f^{(r)}(x)^2 dx = \frac{\lambda}{2}\theta^{-(2r+1)}\boldsymbol{\omega}'\tilde{\boldsymbol{P}}_r\boldsymbol{\omega},$$

where $\tilde{\boldsymbol{P}}_r$ is a fixed positive semidefinite dense matrix with elements

$$\tilde{P}_{r,i,j} = \sum_{k=0}^{r}\sum_{\ell=0}^{r} c_{r,k}c_{r,\ell}\frac{\Gamma(i+j-k-\ell-1)}{\Gamma(i-k)\Gamma(j-\ell)}2^{-(i+j-k-\ell-1)}\mathbb{1}_{i-k>0,j-l>0},$$

where $c_{r,k}$ are finite difference coefficients of order r: $c_{0,0} = c_{1,0} = 1$ & $c_{r,k} = (c_{r-1,k} - c_{r-1,k-1})\mathbb{1}_{k\in\{0,\dots r\}}$.

Issue A: The matrix $\tilde{\boldsymbol{P}}_r$ is dense which is numerically cumbersome.

Issue B: We cannot calibrate λ without cross-validation, which is impossible in our settings.

[4] Simon C. K. Lee and X. Sheldon Lin. "Modeling and Evaluating Insurance Losses Via Mixtures of Erlang Distributions". In: *North American Actuarial Journal* 14.1 (Jan. 2010), pp. 107–130. ISSN: 1092-0277, 2325-0453. DOI: 10.1080/10920277.2010.10597580.

[5] Wenyong Gui, Rongtan Huang, and X. Sheldon Lin. "Fitting Multivariate Erlang Mixtures to Data: A Roughness Penalty Approach". In: *Journal of Computational and Applied Mathematics* 386 (Apr. 2021), p. 113216. ISSN: 03770427. DOI: 10.1016/j.cam.2020.113216.

A better regularisation through P-splines.

Idea: Penalise finite differences of ν instead of f.

Definition (Penalisation of the sequence of modes)

For $f \sim \texttt{MixedErlang}(\boldsymbol{\omega}, \theta)$, the sequence of modes of the Erlang densities is $\boldsymbol{y} = \left(y_i = \frac{i^i e^{-i}}{i!}\right)_{i \in \mathbb{N}}$. The corresponding difference matrix is denoted by $\boldsymbol{D}_r$, such that

$$D_{r,k,l} = c_{r,l-k} y_l \mathbb{1}_{l-k \leq r}.$$

Note: The matrix $\boldsymbol{D}_r$ is sparse!

Thus, $\boldsymbol{D}_r \boldsymbol{\omega}$ are r-order finite differences of the sequence of modes. The corresponding penalty writes

$$\frac{\lambda}{2} \|\boldsymbol{D}_r \boldsymbol{\omega}\|_2^2.$$

Bayesian P-splines interpretation: This is equivalent as setting a prior $\boldsymbol{D}_r \boldsymbol{\omega} | \lambda \sim \texttt{Normal}(0, \lambda \boldsymbol{I})$. Assuming furthermore that we assign a (uninformative, high variance) $\texttt{Gamma}(a_\lambda, b_\lambda)$ prior on λ, the final penalization is

$$\mathrm{Pen}_r(\lambda, \boldsymbol{\omega}) = \frac{1}{2} \left\{(n - r) \log(\lambda) + \lambda \|\boldsymbol{D}_r \boldsymbol{\omega}\|_2^2\right\} + \left\{(a_\lambda - 1) \log(\lambda) - \lambda b_\lambda^{-1}\right\},$$

Numerical scheme via Laplace approximations

The final complete llh is given by $\ell(\boldsymbol{\omega}, \theta, \lambda) = \ell(\boldsymbol{\omega}, \theta) - \mathrm{Pen}_r(\lambda, \boldsymbol{\omega})$.

The *a posteriori* marginal loglikelihood for λ can be easily expressed as $\ell(\lambda) = \ell(\boldsymbol{\omega}, \theta, \lambda) - \ell(\boldsymbol{\omega}, \theta|\lambda)$.

Definition (Hessian notations)

$$\boldsymbol{H}(\boldsymbol{\omega}, \theta) = -\frac{\partial^2}{\partial^2(\boldsymbol{\omega}, \theta)}\ell(\boldsymbol{\omega}, \theta)$$

$$\boldsymbol{P}_r = -\frac{\partial}{\partial\lambda}\frac{\partial^2}{\partial^2(\boldsymbol{\omega}, \theta)}\mathrm{Pen}_r(\lambda, \boldsymbol{\omega}) = \begin{pmatrix} -\boldsymbol{D}_r'\boldsymbol{D}_r & 0 \\ 0 & 0 \end{pmatrix}$$

$$\boldsymbol{H}(\boldsymbol{\omega}, \theta, \lambda) = -\frac{\partial^2}{\partial^2(\boldsymbol{\omega}, \theta)}\ell(\boldsymbol{\omega}, \theta, \lambda) = \boldsymbol{H}(\boldsymbol{\omega}, \theta) - \lambda\boldsymbol{P}_r.$$

Property (λ's estimating equation through Laplace approximation)

Using a Laplace approximation given by $\ell(\boldsymbol{\omega}, \theta|\lambda) \approx \frac{1}{2}\log|\boldsymbol{H}(\boldsymbol{\omega}, \theta, \lambda)|$, *we have*

$$\frac{\partial}{\partial\lambda}\ell(\lambda) = \frac{1}{2}\left\{\sum_{i=1}^{n+1}\frac{\eta_i}{1-\lambda\eta_i} - \|\boldsymbol{D}_r'\boldsymbol{\omega}\|_2^2 - \frac{n-r-2a_\lambda+2}{\lambda} - 2b_\lambda^{-1}\right\} \tag{3}$$

where $\eta_1, \ldots, \eta_{n+1}$ *are eigenvalues of the matrix* $\boldsymbol{H}(\boldsymbol{\omega}, \theta)^{-1}\boldsymbol{P}_r$

Proposed estimation scheme

The algorithm jumps between the following two steps:

(i) Update $(\boldsymbol{\omega}, \theta) = \arg\min \ell(\boldsymbol{\omega}, \theta, \lambda)$ at the current value of λ,

(ii) Update λ by minimizing $\ell(\lambda)$ at the current value of $(\boldsymbol{\omega}, \theta)$.

until a given convergence criterion is met. *Our implementation runs until `Float64` precision is reached.*

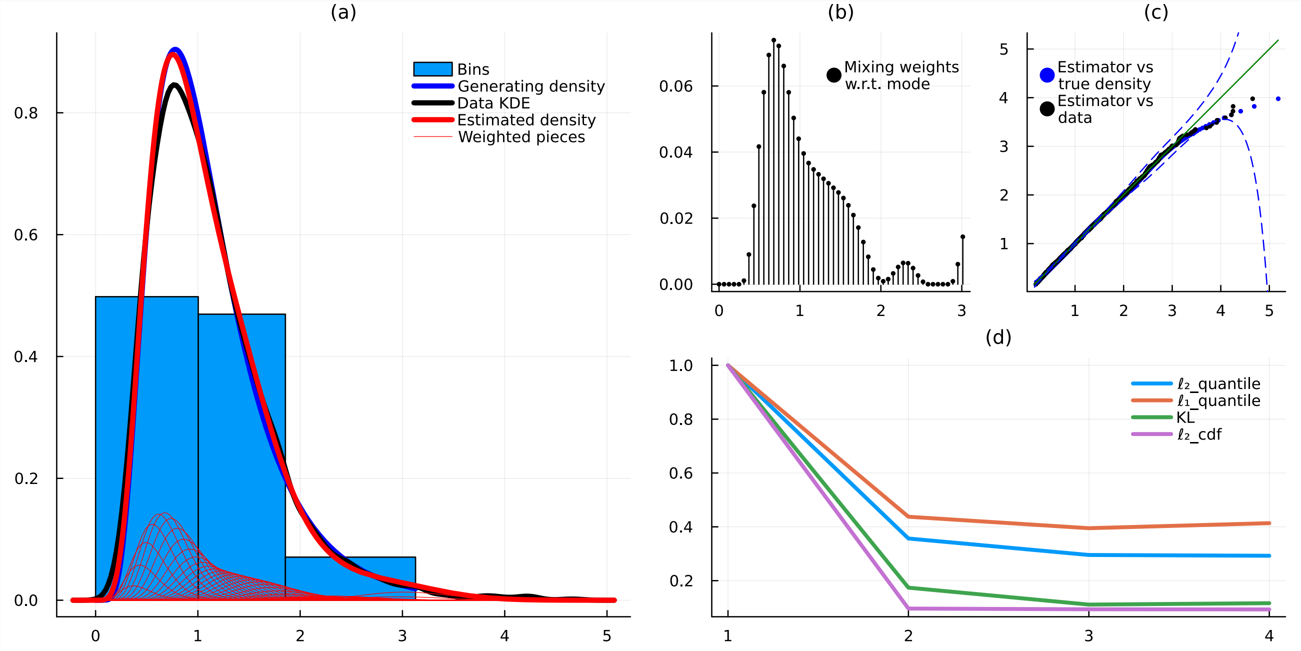
Remark: You may interpret the estimating equation for λ as the result of a mixed-effect regression analysis, see Eilers & Al[6,7].

[6] Paul H. C. Eilers and Brian D. Marx. "Flexible Smoothing with B-splines and Penalties". In: *Statistical Science* 11.2 (May 1996). ISSN: 0883-4237. DOI: 10.1214/ss/1038425655.

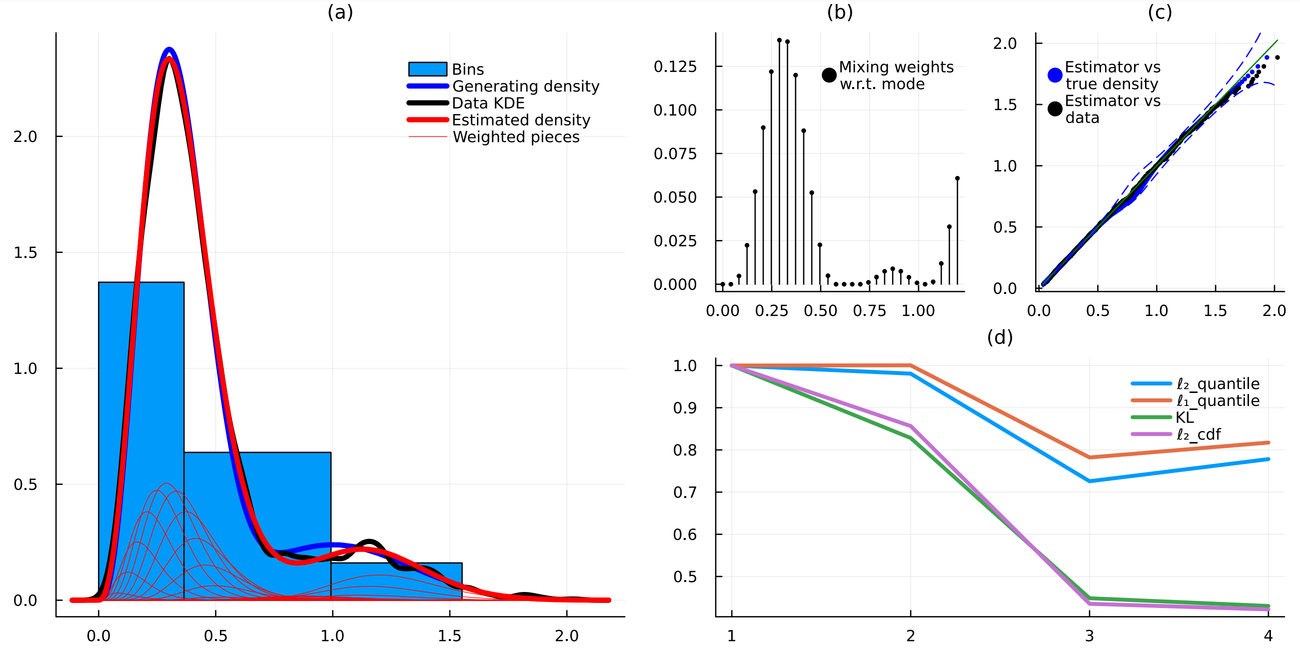
[7] Paul H. C. Eilers. "The Truth about the Effective Dimension: The Truth about the Effective Dimension". In: *Statistica Neerlandica* 72.3 (Aug. 2018), pp. 201–209. ISSN: 00390402. DOI: 10.1111/stan.12131.

Simulated examples

Results on Exemple 1



A Normal/Beta mixture



Conclusion

Conclusion

A few take-away points:

(i) The denseness of the class of Erlang mixtures makes it a good approximator for positive random variables, including multivariate random vectors, under different setups such as censure or truncation.

(ii) The Bayesian interpretation of finite differences roughness penalization, linked to the mixed effects models, allows for automatic and efficient selection of penalisation parameters without cross-validation.

(iii) Laplace approximation allows to derive confidence intervals without running full blown-up MCMC.

(iv) There is a potential for extension to censure & truncation, multivariate mixed Erlangs, and even other generic approximators (non-positively supported datasets).

Thanks !