On using nonparametric approaches for precipitation estimation

Mariusz Grządziel

1Wrocław University of Environmental and Life Sciences, Department of Mathematics, Grunwaldzka 53, 50357 Wrocław, Poland

Abstract. Nonparametric density estimation methods have been used for precipitation estimation for decades. The new approach for estimating the density proposed recently by Geenens and Wang appears to offer advantages over them as far as their behavior in the tail is concerned. Real data analyses presented in this paper confirm the usefulness of this new approach for precipitation modelling.

1 Introduction

Kernel density estimation has many applications in environmental sciences; compare e.g. [1–3]. However, in many practical situations, e.g. when we want to estimate precipitation distribution, we have to estimate density of positive variables which may be challenging. It is known that the kernel estimators have heavy bias at and near the boundary; compare [4, page 594]. Some remedies for this defect and other drawbacks of kernel estimators that arise in the case when the random variable of interest is positive were proposed in the literature. However, these remedies are not quite satisfactory and the problem of density estimation of positive random variables is still an area of active research.

We will present some new methods addressing this problem in Section 2; among them the method proposed by Geenens and Wang [5] deserves in our opinion particular interest. Its usefulness will be illustrated on real datasets in Section 3. Conclusions will be given in the final section.

2 Density estimation of positive variables

For a given a sample \(X_1, ..., X_n\) consisting of i.i.d. copies of a random variable \(X\) with unknown distribution \(F_X\) admitting a density \(f_X\), the kernel of estimator \(\hat{f}_X\) takes the form

\[
\hat{f}_X(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - X_k}{h}\right),
\]

where the kernel \(K\) is a non-negative function satisfying the conditions

\[
\int_{-\infty}^{\infty} K(x) \, dx = 1,
\]

and \(h > 0\) is known as the bandwidth. \(K\) is usually assumed to be symmetric about 0.

The statistical properties of the estimator (1) have been presented e.g. in [6]. It has been shown that it reliably estimates \(f_X\) when it is supported on \(\mathbb{R}\). If the support of \(f_X\) is bounded, the kernel estimator \(\hat{f}_X\) may have some drawbacks.

In this section we consider the case when \(F_X\) is the distribution of a positive continuous random variable \(X\) having density \(f_X\) supported on \(\mathbb{R}^+ = (0, +\infty)\). Such variables naturally arise in various areas of environmental sciences.

As it already has been mentioned, the kernel estimator (1) usually fails to correctly estimate the behaviour of \(f_X\) near 0; also its tail behaviour often is not satisfactory. In the neighbourhood of 0 the estimator may place positive probability mass on the negative half-line which results in the boundary bias. In the tail region the estimator may produce artificial local maxima at each observation, sometimes referred to as spurious bumps; compare [7]. Several remedies to overcome the boundary bias, such as the reflection method [8], were proposed in 1970-ies and 1980-ies.

In 2000 Chen [9] addressed both mentioned problems arising in the case when \(F_X\) is supported on \(\mathbb{R}^+ = (0, +\infty)\) by introducing the asymmetric kernel density estimator of \(f_X\)

\[
\hat{f}_X(x) = \frac{1}{n} \sum_{k=1}^{n} L_{x,h}(X_k),
\]

where \(L_{x,h}\) is an asymmetric \(\mathbb{R}^+\)-supported density whose parameters are functions of \(x\) and \(h\). He proposed two ‘asymmetric kernels’:

Corresponding author: mariusz.grzadziel@upwr.edu.pl

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
\[ L_{X|Y}^{a,b}(x) = \frac{f_X(x)}{f_Y(y)} \quad \text{for } x > 0 \]

(8)

where \( f_{X|Y} \) is the density of the Gamma distribution with the shape parameter \( \alpha \) and the scale parameter \( \beta \) while

\[ \rho(x,y) = \frac{c}{h(x-h)^2} + 1 \quad \text{if } x \geq 2h; \]

\[ \rho(x,y) = 0 \quad \text{if } y < 0 \]

(7)

Several types of asymmetric kernels were investigated in the subsequent literature [10–13]. However, it seems that the estimators using these kernels do not show significant advantage over the Chen estimators.

Another approach to estimating \( f_X \) supported on \( \mathbb{R}^+ = (0, +\infty) \) consists in estimating the density of \( Y = \log X \) and obtaining an estimator of the density of \( X \) using well-known results concerning functions of random variables. The fact that

\[ f_X(x) = \frac{f_Y(\log x)}{x} \quad \text{for } x > 0. \]

(9)

Here \( f_Y \) is a suitably chosen estimator of the density \( f_X \) of the random variable \( Y \) supported on \( \mathbb{R} \). We will refer to the estimator \( (9) \) as the log-transform estimator of the density function \( f_X \); compare [14].

Geenens and Wang [5] pointed out that choosing a kernel estimator for estimating \( f_Y \) may result in some drawbacks. This is due to the fact that the kernel estimators are known to poorly estimate tails of densities, compare [4, page 594]. One way to overcome this shortcoming is to use the local likelihood density estimator of \( f_Y \) (instead of the kernel estimator). The estimator constructed using this approach has many favourable properties; compare [5]. A simulation study in [5, Section 7] indicates that when the degree of the polynomial that approximates the logarithm of the density of \( Y \) is 2 (compare [5, Section 2.3]), this estimator has small Mean Integrated Absolute Relative Error (MIARE); for an estimator \( f_X \) of the density \( f_X \) of a positive random variable \( X \) MIARE is defined as

\[ E \left( \int_0^x |f_X(x) - f_X(x)| dx \right). \]

(10)

This shows that this local log-quadratic transformation kernel density estimator based on the log-transformation (in short: LL-LT estimator) perform well in the tail.

This property was confirmed by the simulation study presented in [15], the version of the preprint [5] that was accepted for publication; in this simulation study other criterion than MIARE, the mean integrated squared error, was taken into account.

### 3 Applications to precipitation estimation

The applications of nonparametric methods to precipitation estimation were considered in [16]. The authors suggest using the log-transform density estimator \( (9) \) in which \( f_Y \) was computed using the kernel method with bandwidth proposed by Sheather and Jones [17]. The log-transform kernel density estimator was also used in [18] for estimating precipitation amounts for various time intervals.

Using nonparametric methods for estimating precipitation extremes appears to be particularly challenging. The estimator \( (9) \) in which \( f_Y \) is estimated using spline methods seems to be particularly promising for this aim [19].

#### 3.1 Real data analyses

We have performed analyses of three datasets that were labelled according to the places they were collected from:

- The dataset ‘Łódź’ contains the measurements of monthly precipitation in June (in mm) in the years 1954–1992 from the meteorological station Łódź-Lublinek.
- The dataset ‘Wichita’ contains the measurements of monthly precipitation in August (in mm) in Wichita, Kansas, in the years 1980–2011. This dataset is a part of the R package SPEI [20].
- The dataset ‘Kraków’ contains the annual (from May to September) maximal precipitation intensity measurements (mm/h) from Jagiellonian University Astronomical Observatory in Kraków in the years 1961–1985. The intensities are computed for the time interval of 24 hours. The dataset is taken from the monograph [21, page 354].

For the dataset ‘Łódź’ the Gamma distribution and the Lognormal distribution were fitted using the procedure `fitdist` from the package `fitdistplus` [22]. The LL-LT estimator was computed using the R script accompanying the paper [5] (available at the Journal of the Computationl and Graphical Statistics website).

### Table 1. Precipitation (mm) in June in Łódź (1954–1992): quantiles computed for the obtained distributions (estimates of the density function): Lognormal, Gamma, LL-LT, Gamma kernel, Gaussian kernel (with Sheather-Jones bandwidth).

| Distribution       | Quantile       | 0.01 | 0.99 | 3.0 |
|--------------------|----------------|------|------|-----|
| Lognormal          |                | 13.04| 18.49| 291.11|
| Gamma              |                | 8.22 | 14.51| 221.93|
| LL-LT              |                | 10.49| 19.85| 264.47|
| Gamma kernel       |                | 0.75 | 5.22 | 226.52|
| Gaussian Kernel(SJ)|                | -3.31| 7.36 | 236.42|
Gamma kernel estimate did well in the right tail, but its boundary behaviour seems to be less satisfactory; the values of the density function near the boundary seem to be overestimated using this approach.

Table 2. Precipitation (mm) in August in Wichita (1980-2011): quantiles computed for the estimates of the density.

| Distribution       | Quantile | 0.01  | 0.99  | 0.995 |
|--------------------|----------|-------|-------|-------|
| Lognormal          | $\Phi(3)$| 5.68  | 10.03 | 903.34|
| Gamma              |          | 2.41  | 6.84  | 421.50|
| Gamma ker.         |          | 3.21  | 7.24  | 413.48|
| Gaussian kernel (SJ)|        | 0.28  | 2.10  | 408.31|
|                    |          | -52.79| -27.52| 353.86|

We see that the LL-LT estimate did approximately as well as the Log-normal estimate. Let us note that the dataset ‘Łódź’ was also analyzed in [25]: the authors have demonstrated that the fit obtained for the Log-normal distribution was better than for the competing distributions considered in [25].

Similar results were obtained for the dataset ‘Wichita’. The LL-LT and the Gamma estimates appear to perform best; the $\Phi(3)$ quantile for the Log-normal distribution in the Table 2 seems to be unreasonably large.

Table 3. Maximum rainfall intensity (mm/h) in Kraków (1961–1985) computed for the time interval of 24 hours: quantiles computed for the density estimates.

| Distribution       | Quantile | 0.99  | 0.995 |
|--------------------|----------|-------|-------|
| Lognormal          |          | 3.39  | 3.72  |
| Gamma              |          | 3.23  | 3.47  |
| GEV                |          | 3.02  | 6.30  |
| LL-LT              |          | 4.21  | 4.83  |
| Gamma kernel       |          | 3.65  | 3.77  |
| Gaussian kernel (SJ)|        | 3.57  | 3.93  |
for this aim. There is a need of an extensive simulation study comparing this method with other new promising nonparametric approaches that may be used for precipitation estimation.

References

1. K.D. Kim, J.H. Heo, J. Hydrology 260, 176–193 (2002)
2. R.A. Ferreyra, G. P. Podestá, C. D. Messina, D. Letson, J. Dardanelly, E. Guevara, S. Meira, Agric. For. Meteorol. 107, 177–192 (2001)
3. A. Michalski, Meteorol. Hydrol. Water Manage, 4, 41–46 (2016)
4. S.J. Sheather, Stat. Sci. 19, 588–597 (2004)
5. G. Geenens, C. Wang, J. Comput. Graph. Stat. (to be published)
6. M.P. Wand, M.C. Jones, Kernel Smoothing (Chapman and Hall, Boca Raton, 1995)
7. P. Hall, C. Minnotte, C. Zhang, Ann. Statist. 32, 2124–2141 (2004)
8. E. Schuster, Comm. Statist. Theory Methods 14, 1123–1136 (1985)
9. S.X. Chen, Ann. Inst. Statist. Math. 52, 471–480 (2000)
10. Jin, X. and Kawczak, J. Ann. Econ. Finance 4, 103–124 (2003)
11. O. Scaillet, J. Nonparametr. Stat. 16, 217–226 (2004)
12. M. Hirukawa, M. Sakudo, J. Nonparametr. Stat. 27, 41–63 (2015)
13. G. Igarashi, Commun. Stat. Theory Methods 45, 6670–6687 (2016)
14. A. Charpentier, E. Flachaire, L’Actualité Économique 91, 141–159 (2015)
15. G. Geenens, C. Wang, Local-likelihood transformation kernel density estimation for positive random variables (arXiv preprint, arXiv:1602.04862 v1 [stat.ME], 2016)
16. B. Rajagopalan, U. Lall, D. Tarboton, Stoch. Hydrol. Hydraul. 11, 523–547 (1997)
17. S.J. Sheather, M.C. Jones, J.R. Stat. Soc. Ser. B Stat. Methodol. 53, 683–690 (1991)
18. T. Mosthaf, A. Bardossy, Hydrol. Earth Syst. Sci. 21, 2643–2481 (2017)
19. W. Huang, D. Nychka, H. Zhang, Modeling Precipitation Extremes using Log-Histospline (arXiv preprint, arXiv:1802.09387v1, 2018)
20. S. Begueria, S. Vicente-Serrano, SPEI: Calculation of the Standardised Precipitation- Evapotranspiration Index (R package, version 1.7, http://CRAN.R-project.org/package=SPEI, 2017)
21. S. Węglarczyk, Statistics in environmental engineering (in Polish) (Technical University in Kraków Press, Kraków, 2010)
22. M.L. Delignette-Muller, C. Dutang, J. Stat. Softw. 64, 1–34 (2015)

23. J. Moss, M. Tveten, *kdensity: Kernel Density Estimation with Parametric Starts and Asymmetric Kernels* (R package, version 1.0, https://CRAN.R-project.org/package=kdensity, 2018)

24. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, 2015)

25. E. Gąsiorek, E. Musiał, J. Ecol. Eng. 16, 44–53 (2015)

26. A. Stephenson, R News 2, 31–32 (2002)