Project Week 1: Snow Particles

Emanuele Sorgente

2024-03-01

Contents

Introduction

The report delves into an examination of snowflake diameters sourced from the Laboratory of Cryospheric Sciences at EPFL. This dataset, furnished by a PhD student, is integral for understanding avalanche dynamics, particularly in the context of fast-moving powder-snow avalanches. The sedimentation velocity, heavily influenced by snowflake size, holds paramount importance in modeling turbulent dilute suspension avalanches accurately. The report emphasizes the significance of precise size estimation, noting that errors can lead to substantial inaccuracies in velocity estimation.

Initially, the report conducts exploratory data analysis to glean insights into the dataset, crucially focusing on the distribution of particles across various size categories. Furthermore, the report highlights the methodology employed, particularly the utilization of binned data representing diameter ranges of snowflakes. Expert knowledge advocates for a bi-log-normal distribution as a suitable model for this data. Section 2 verifies this assumption, while Section 3 employs the EM algorithm and optimization techniques to estimate the five parameters characterizing this distribution. The report concludes with a validation step using parametric bootstrap to ascertain whether the snowflake diameters indeed adhere to the bi-log-normal distribution model.

Data Exploration

In this section, our primary objective is to validate the assumption that the bi-log-normal distribution accurately represents the data. While plotting a histogram seems straightforward, the data is provided in bins of varying sizes, requiring a more nuanced approach. Therefore, before constructing the histogram, we'll introduce jitter to the data using a uniform distribution within each bin. Additionally, we'll adjust the endpoint of the last bin to 3 to ensure comparability with the others, despite lacking data points in this last bin.

The outcome of interest is the variable `retained`, defining the proportion of particles belonging to every bin's diameter. The other variables are:

- The `startpoint` of the bin
- The `endpoint` of the bin
- The number of particles `detected`

Our visualization strategy involves plotting the distribution of snowflake diameters alongside observation frequencies. This step is crucial for accurately simulating snowflake diameters to investigate snow's aeolian transport dynamics.

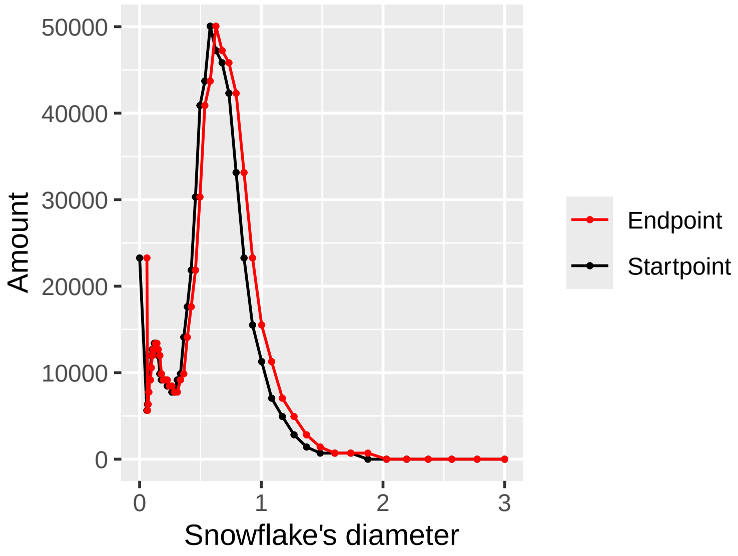

Figure 1: Lines defining the amount of snowflakes in each diameter's bin

Figure 1 provides a clear depiction of the original data behavior. We observe two distinct lines representing the startpoint and endpoint. Notably, over 20,000 observations fall within the bin between 0 and 0.06. Additionally, there are no observations for diameters greater than 2.

Observing in Figure 2 two slightly right-skewed peaks in the distribution suggests that the assumption of a bi-log-normal distribution holds true. This alignment corresponds with our expectations for a mixture of two log-normal distributions.

Model Fitting

In this section, our objective is to estimate the parameters for the jittered data assuming it conforms to a bi-log-normal distribution. Having observed the possibility of the data being distributed as a mixture of two log-normal distributions, we proceed to evaluate the parameters for the density function:

$$f(x) = (1-\tau)\frac{1}{x\sigma_1\sqrt{2\pi}}e^{-\frac{(\log(x)-\mu_1)^2}{2\sigma_1^2}} + \tau\frac{1}{x\sigma_2\sqrt{2\pi}}e^{-\frac{(\log(x)-\mu_2)^2}{2\sigma_2^2}} \tag{1}$$

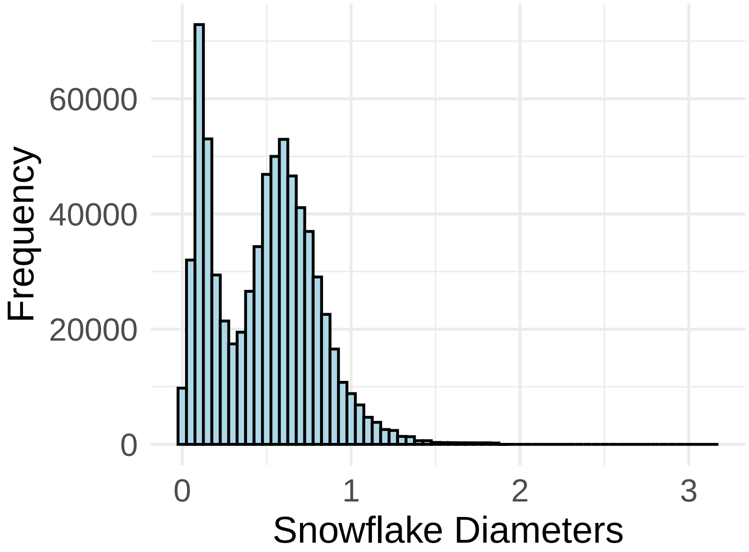

Figure 2: Histogram of the snowflake diameters after jittering the data, using in each bin a uniform distribution.

Our approach involves the following steps: First, utilizing the EM algorithm, we estimate the parameters μ_1, μ_2, σ_1, σ_2, and τ for the jittered data. Subsequently, we optimize the log-likelihood of the binned data (not the jittered data used for the EM algorithm), initiating with the values derived from the EM algorithm.

Following the initial estimation of parameters using the EM algorithm, I will delve into alternative methos to enhance the accuracy of estimation. This exploration will encompass techniques like local search, aimed at refining the parameter estimates for the model.

EM algorithm

To effectively apply the EM algorithm, we require initial values for each parameter: $\mu_1, \mu_2, \sigma_1, \sigma_2$ and τ. These values are determined by analyzing our histogram (refer to Figure 1). Considering the expectation and variance of a log-normal distribution with parameters μ, σ obtained from $\exp(\mu + \frac{\sigma^2}{2})$, $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$, we set our initial values as follows: $\mu_1 = -3, \mu_2 = -0.5, \sigma_1 = 0.3, \sigma_2 = 0.4$ and $\tau = 0.7$. While $\varepsilon = 0.001$ will constitute the threshold such that I stop the algorithm when the difference in the log-likelihoods is under ε.

We will leverage the outputs of the EM algorithm. We achieve this by generating a plot that compares the kernel density estimator (utilizing a Gaussian kernel) of our jittered data with the density of a bi-log-normal distribution, characterized by the results obtained from the EM algorithm.

We can see from Figure 3 two very similar curves, supporting our assumption of the bi-log-normal model. This plot is also useful to see that the EM algorithm converged to a reasonable solution.

Optimization

Although we applied the EM algorithm to the jittered data, we did not fully maximize the true log-likelihood of the binned data. The log-likelihood function for the binned data is given by:

$$f(d|\mu_1, \mu_2, \sigma_1, \sigma_2, \tau) \propto \prod_{j=1}^{n} \left[f(a_j) - f(a_{j-1})\right]^{b_j}, \tag{2}$$

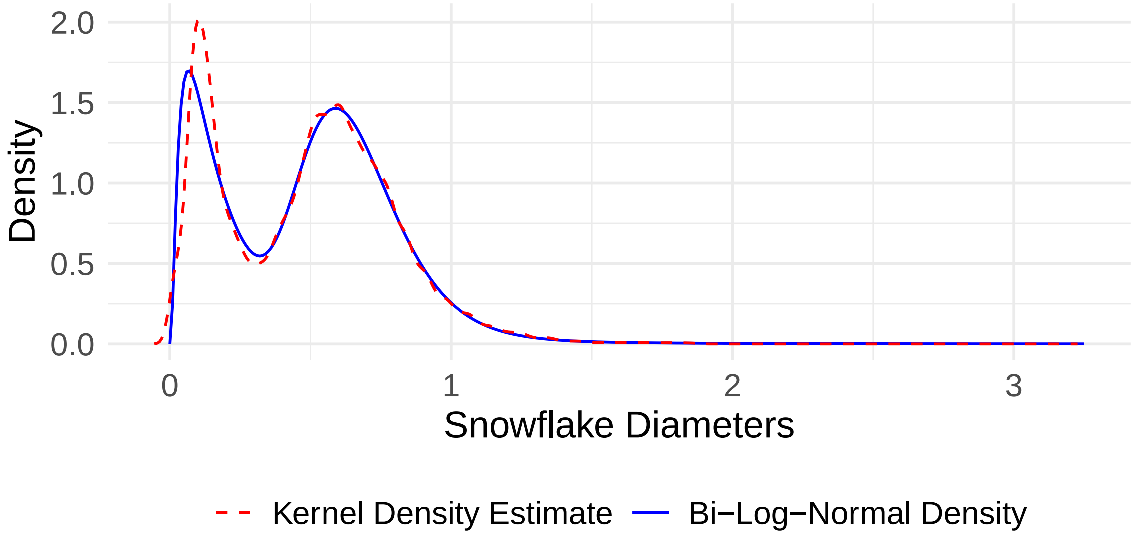

Figure 3: Kernel density estimate (using a Gaussian kernel) of the jittered data alongside the density obtained by the bi-log normal model with parameters derived from the EM algorithm.

where a_j represents the bin endpoints, b_j denotes the observed frequencies for each bin, and d signifies the full dataset and f was before defined in equation 1.

Nonetheless, the outcomes generated by the EM algorithm are expected to closely approximate the maximum of the log-likelihood. Therefore, we employ these results as initial values for optimization purposes, utilizing the Nelder-Mead optimization algorithm.

Our final optimized estimates of the parameters for the bi-log-normal distribution are as follows:

- $\mu_1 = -2.0219331$
- $\mu_2 = -0.4543924$
- $\sigma_1 = 0.6060781$
- $\sigma_2 = 0.3009943$
- $\tau = 0.6495080$

Figure 4 illustrates that the Bi-Log-Normal density, using the optimized parameters, fits the data significantly better than the density obtained after the EM algorithm.

Now that we have the optimized parameters for our parametric model of a bi-log-normal distribution, it is time to rigorously assess the validity of our assumption. While we have observed visual indications that the data may conform to a bi-log-normal distribution.

Parametric Bootstrap

In this section, we assess whether our data conforms to a mixture of log normal distributions using a parametric bootstrap algorithm. The aim is to determine if the observed distribution fits within the family of bi-log-normal distributions.

We frame our hypotheses as follows:

$$H_0 : F \in \mathcal{F} = \{F_\lambda | \lambda \in \Lambda\}$$

$$H_1 : F \notin \mathcal{F}$$

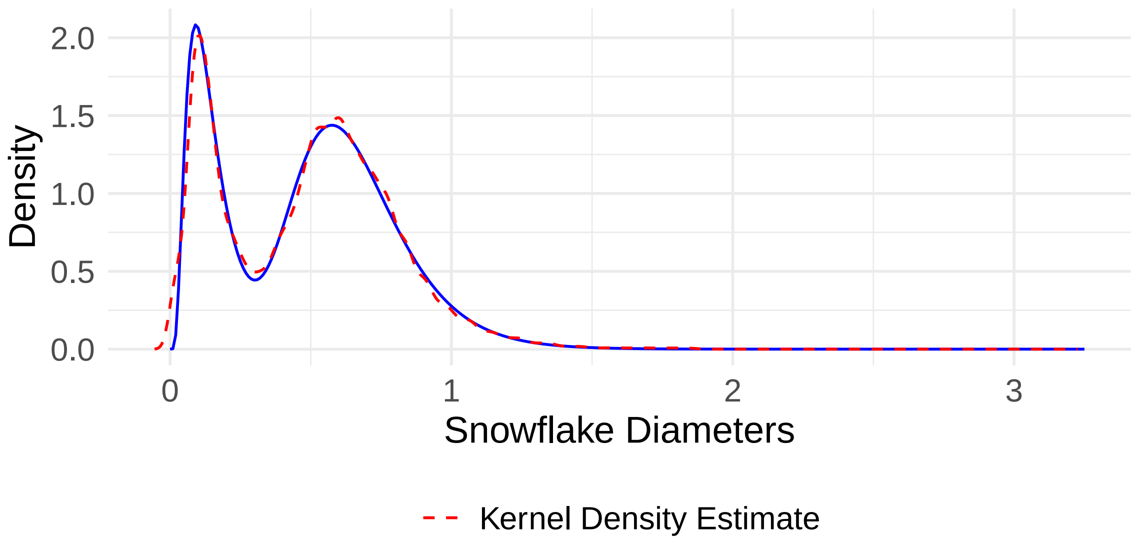

Figure 4: Kernel density estimate (using a Gaussian kernel) of the jittered data alongside the density obtained by the bi-log normal model with optimized parameters.

where $\mathcal{F}$ represents the set of bi-log-normal distributions.

To execute the parametric bootstrap algorithm, we follow the procedure outlined in [2]. We generate 100 bootstrap datasets, each containing 705047 datapoints, as in the original data, by sampling from a bi-log-normal distribution using the optimized parameters obtained earlier. For each bootstrap dataset, we undertake the following steps:

1. Bin the data using the original dataset's bins.
2. Estimate the parameters of the bi-log-normal model using the EM algorithm.
3. Optimize the true log-likelihood of the binned, bootstrapped data.

With these steps completed, we proceed to estimate the p-value for testing H_0 against H_1. This involves comparing the empirical cumulative distribution function (CDF) of each bootstrapped dataset with the parametrized CDF using the parameters obtained by EM and optimization. We compute the Kolmogorov-Smirnov statistic

$$T_b^\star = \sup_x |\widehat{F}^\star_{N,b} - F_{\widehat{\lambda}_b^\star}|$$

for each bootstrapped dataset, where $\widehat{F}^\star_{N,b}$ represents the empirical CDF of the binned b-th bootstrapped dataset and $F_{\widehat{\lambda}_b^\star}$ is the CDF parametrized by the optimized parameters $(\lambda_b^\star)$. Additionally, we calculate the Kolmogorov-Smirnov statistic

$$T = \sup_x |\hat{F}_N(x) - F_{\hat{\lambda}}(x)|$$

of the empirical CDF of the original binned data and the CDF parametrized by the optimized parameters of the original data.

The p-value estimate is derived as:

$$\text{p-value} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}_{[T_b^\star \geq T]}\right)$$

where $B = 1000$ is the number of bootstrapped datasets.

After conducting these calculations, we obtain a p-value of 0.1338661. This relatively large p-value suggests that we do not have sufficient evidence to reject the null hypothesis, H_0, which supports the assumption of a bi-log-normal model for our data. Consequently, we can confidently utilize our optimized parameters to generate new data points for our snowflake experiment.

To conclude this section, we show in Figure 5 the empirical cumulative distribution function (CDF) of our data alongside the parametrized CDF using our optimized parameters. It emerges from visual inspection of the plot that the our bi-log normal distribution with our optimized parameters approximates with a very good fitting the binned data. This further supports our previous supposition from Figure 4 that the Bi-Log-Normal distribution is highly effective in modeling the data.

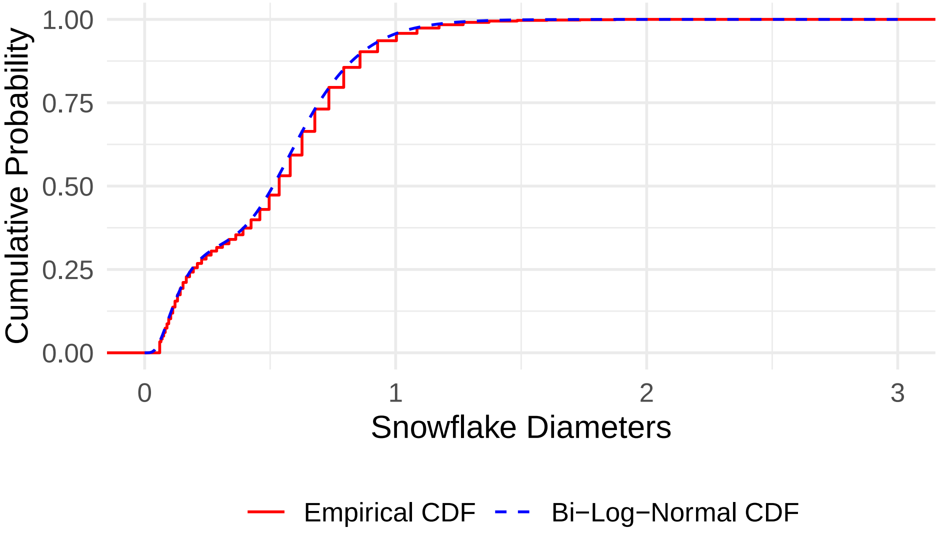

Figure 5: Bi-Log Normal CDF and Empirical CDF

Conclusion

In this report, we have presented statistical evidence indicating that the distribution of snowflake diameters follows a bi-log-normal distribution. Through the utilization of the EM algorithm and optimization techniques applied to the log-likelihood function. We also employed bootstrap validation, which corroborated our findings and supported the appropriateness of the bi-log-normal distribution for describing the data. We successfully determined a set of parameters that accurately describe our bi-log-normal model, effectively capturing the characteristics of the data.

Furthermore, our analysis demonstrates that even when working with binned data, it is feasible to identify a suitable statistical model for analysis. Nonetheless, it's important to note that utilizing non-binned data is preferable, as it offers a more comprehensive dataset and enhances the accuracy of our statistical models.