Remark on the Swarm data residual distribution

Andrey Khokhlov (✉ fbmotion@gmail.com)

Express Letter

Keywords: Probability distributions, Magnetic field variations through time, Satellite magnetics

Posted Date: April 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-23276/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

¹ Remark on the Swarm data residual distribution

² A.Khokhlov^{1,2}

¹IEPT RAS, 84/32, Profsoyuznaya 117997 Moscow, Russia, fbmotion@gmail.com

²Geophysical Center RAS, 3, Molodezhnaya st., 119296 Moscow, Russia

Abstract

The behavior of data in space and time is usually based on a preliminary model of variability. Deviations observed from the model characterize both the model itself and possible data errors. Gaussian error theory is usually used to characterize the distribution of deviations, but very often non-gaussian nature of these deviations is visually obvious. We may see this on the semi-logarithmic chart of the corresponding histogram. Therefore, the percentage of large deviations may be significantly larger than that predicted with the help of the Gaussian error theory: a different method of processing the residuals is necessary for real data. If we assume that the variability itself depends on space and time (in the language of probability theory this behavior is characterized as a mixture of Gaussian distributions), then the non-gaussian shapes of histograms can be explained. For practical purposes, an explicit characterization of the mixing law is of interest, since the share of significant deviations is determined by this very law. The article substantiates the explicit description of the mixing law and shows how to calculate its parameters from the data using the example of magnetic data. Such an approach is also applicable in more general situations.

Keywords

Probability distributions, Magnetic field variations through time; Satellite magnetics

Introduction

Recall that typical appropriate statistical assumptions are being made with respect to the distribution of the uncertainties σ_i affecting the data used. The statistical properties of these uncertainties, however, are not always well characterized. In such circumstances, assuming that uncertainties follow a Gaussian distribution would *a priori* make sense, since such a distribution often arises naturally as a consequence of the central limit theorem when errors act in an additive manner (see, e.g., Feller 1971). Relying on this assumption, and provided that s_i is an adequate measure of the error affecting the datum γ_i , standard statistical estimations then used to infer the model. The normalized residuals $\left\{ \frac{\gamma_i - \tilde{\gamma}_i}{s_i} \right\}$ (here $\tilde{\gamma}_i$ being the datum value predicted by the model) are expected to follow a standard normal distribution. Yet, residuals often display a sharper distribution, sometimes much closer to that of so-called Laplace

distribution (e.g., Jackson et al. 2000; Walker & Jackson 2000; Panovska et al. 2012, 2015). We showed in Khokhlov & Hulot (2017) that residuals may be incorrectly normalized and therefore their common statistical distribution is a mixture of gaussian distributions (Barndorff-Nielsen et al. 1982) — this is, generally speaking, not at all new. In particular we demonstrated in Khokhlov & Hulot (2017) several examples of the variability in σ determination, that indeed leads to the non-gaussian shape of the histogram. Thus we assume that observable residuals θ is the mixture of individual gaussian random variables with zero expectations and random variances β^2 . The artificial intelligence approach provides the algorithm that, in principle, can recover the distribution of variances. However, this computational algorithm (EM-algorithm) is not perfect and too sensitive to the errors in the data. In the present note we argue that the distribution of random variable β can be well approximated by the lognormal distribution with pdf

$$f_{\beta}(t) = \frac{1}{ts\sqrt{2\pi}} \exp \left[-\frac{\ln^2 t}{2s^2} \right]$$

We also provide the method that recovers the value of s in the real data case.

Mixture model

The unformal interpretation

The mixture model is appropriate for the situation when the data is inhomogenous, for instance it comes from a several locations such that each region perturbe slightly the assumed data distribution law, i.e. the corresponding distribution formulae differs slightly in their parameters. In practice we often face with even simpler situation: each regional data is gaussian with mean zero but the corresponding σ -values depend on the region. However we rarely can select the region with absolutely homogenous data in it, therefore we better simulate this situations by means of sequential small perturbations of the initially homogenous gaussian population. Whether the limit distribution can be described providing the very small intermediate perturbations?

Version of the general formula

If ζ is an arbitrary random variable with density f_{ζ} , then for fixed $y_0 > 0$ the ratio ζ/y_0 has density $f_{\zeta}(xy_0)y_0$, see also (see, e.g., Feller 1971). Let now denominator is not fixed but a positive random variable with density g_{η} , then we get the pdf for this ratio

$$h(x) = \int_0^{+\infty} f_{\zeta}(xy)y g_{\eta}(y) dy$$

We may now compare the mixture of unbiased Gaussian distributions (i.e. with pdf $f_\alpha = \mathcal{N}(0, \sigma^2)$) by randomizing their standard deviations using a random variable $\beta > 0$:

$$\begin{aligned} f_\theta(x) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{1}{2} \frac{x^2}{t^2}\right) t^{-1} f_\beta(t) dt \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y^2 x^2\right) y g_\eta(y) dy \end{aligned}$$

Obviously f_θ can be interpreted as the pdf of the ratio.

For instance, recall the following example of Khokhlov & Hulot (2017): the uniform mixture of unbiased Gaussian distributions with standard deviations varying between 0 and 1, i.e. mixing pdf is

$$f_\beta(t) = \begin{cases} 1 & 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

we may treat that mixture as the ratio of standard gaussian α divided by η — the inverse of uniform distribution:

$$g_\eta(y) = \begin{cases} \frac{1}{y^2} & y > 1 \\ 0 & y \leq 1 \end{cases}$$

Sequential small mixtures

The *small* multiplicative randomization is described in terms of a random $\beta > 0$ with pdf $f_\beta \sim 0$ out of $[1 - \varepsilon, 1 + \varepsilon]$ for some small ε , we may assume $\beta = e^\delta$ where expectancy $E(\delta) \sim 0$ and variance $D(\delta) \sim \varepsilon^2$.

For the sequential small mixtures (with independent β_i) then we get the ratio

$$\frac{\alpha}{\eta_1 \cdot \eta_2 \dots \eta_m} = \frac{\alpha}{e^{-\sum_i \delta_i}} \quad (1)$$

But under the mild conditions the distribution of $\sum_i^m \delta_i$ rapidly converges to gaussian distribution $\mathcal{N}(a, s^2)$ with $a \sim 0$ and $s \sim \sqrt{\sum_i^m \varepsilon_i^2}$, thus the limit pdf for the sequential *arbitrary, but small* mixtures can be approximated by

$$f_\theta(x) = \int_0^\infty \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2t^2\sigma^2}\right) \frac{1}{ts\sqrt{2\pi}} \exp\left[-\frac{\ln^2 t}{2s^2}\right] dt \quad (2)$$

for $a = 0$ and some suitable parameters s and σ .

Real data application

Here we use the same data as in Khokhlov & Hulot (2017): we consider the absolute scalar data acquired by two of the Swarm satellites (Satellites Alpha and Bravo) at quasi-latitudes ranging between $+55^\circ$ and

-55° , and computed residuals with respect to the so-called VFM model of Vigneron et al. (2015): for the the array ST_1 of one-day std of residuals, take a look at Fig.1 borrowed from our article. The satellite scalar data (Vigneron et al. 2015) cover a little less than a year (between November 29, 2013 and September 25, 2014) and were further selected following a number of criteria, among which magnetically quiet and night time conditions, to ensure that as little as possible non-modeled external signal is included in the data. This resulted in 42 160 data for the Alpha satellite and 42 175 for the Bravo satellite. These data can be expected to reflect the signal of the field of internal origin the model aims at modeling, any other source of signal being treated as a source of noise acting on top the very low instrumental and satellite noise (less than 0.3 nT, see Léger et al. 2015; Olsen et al. 2015; Fratter et al. 2016). The datasets used and analysed during the current study are available from the author on reasonable request.

Method and Numerical results

Rescale this array ST_1 as $r \mapsto \frac{\sigma_1}{r} = y$ where $\sigma_1 = \text{mean}(ST_1)$ and $r \in ST_1$; in virtue of the eq.2 the array $\{y_i\}$ is expected to obey the lognormal distribution with parameter s_1 (see eq.2), let's directly calculate σ_1 and s_1 .

Now repeat all these computations for arrays $ST_{0.25}$, $ST_{0.5}$, $ST_{0.75}$ (i.e. corresponding to time intervals of 0.25 to 0.75 day), here are the results:

ST_1 : Satellite A $\sigma_1 = 2.41$ $s_1 = 0.33$, Satellite B $\sigma_1 = 2.40$ $s_1 = 0.36$

$ST_{0.75}$: Satellite A $\sigma_{0.75} = 2.34$ $s_{0.75} = 0.36$, Satellite B $\sigma_{0.75} = 2.35$ $s_{0.75} = 0.39$

$ST_{0.5}$: Satellite A $\sigma_{0.5} = 2.22$ $s_{0.5} = 0.39$, Satellite B $\sigma_{0.5} = 2.21$ $s_{0.5} = 0.46$

$ST_{0.25}$: Satellite A $\sigma_{0.25} = 1.99$ $s_{0.25} = 0.43$, Satellite B $\sigma_{0.25} = 1.96$ $s_{0.25} = 0.49$

As often happens, a limited amount of lognormal data cannot provide stable statistical estimates, so what are the "true values" of s and σ ? To answer this question let's use the following well-known method of

95 the statistical moments of θ , namely:

$$\begin{aligned}
E|\theta| &= \int_{-\infty}^{+\infty} |x| f_{\theta}(x) dx \\
&= \int_0^{\infty} \left[\int_{-\infty}^{+\infty} |x| \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{x^2}{t^2\sigma^2}\right) dx \right] \cdot \frac{\exp\left(-\frac{\ln^2 t}{2s^2}\right)}{ts\sqrt{2\pi}} dt \\
&= \int_0^{\infty} \sigma \sqrt{\frac{2}{\pi}} \cdot t \frac{1}{ts\sqrt{2\pi}} \exp\left(-\frac{\ln^2 t}{2s^2}\right) dt = \sigma \sqrt{\frac{2}{\pi}} E\beta \\
&= \sigma \sqrt{\frac{2}{\pi}} e^{s^2/2} \\
E\theta^2 &= \int_{-\infty}^{+\infty} x^2 f_{\theta}(x) dx = \int_0^{\infty} E\alpha^2 \cdot t^2 \frac{1}{ts\sqrt{2\pi}} \exp\left(-\frac{\ln^2 t}{2s^2}\right) dt \\
&= \sigma^2 E\beta^2 = \sigma^2 e^{2s^2}
\end{aligned}$$

96 Thus we get the explicit expressions of the unknown parameters

$$\begin{cases} s^2 = \ln \frac{2}{\pi} + \ln E\theta^2 - 2 \ln E|\theta| \\ \sigma^2 = \frac{\pi^2}{4} \cdot \frac{(E|\theta|)^4}{E\theta^2} \end{cases} \quad (3)$$

97 In practice we recover from real data the estimates of the moments $E\theta^2$, $E|\theta|$ and then get (the estimates
98 of) the unknown parameters.

99 Conclusions

100 Hereby we added the quantitative details of the data distribution to the qualitative analysis of it that was
101 published in Khokhlov & Hulot (2017): namely, using formula 3, we may now recover the estimates of the
102 parameters s and σ (the latter can be treated as an estimate for the "inner precision" of measurements);
103 Fig. 2 actually confirm the fact that this close-to-Laplacian distribution indeed can be represented as the
104 result of lognormal mixture according to formula 2.

105 Declarations

106 Availability of data

107 Here we use the same Swarm data as in Khokhlov & Hulot (2017). The data simulated and analysed
108 during the current study are available from the author on reasonable request.

109 Competing interests

110 There are no competing interests.

Funding

This work is partly supported by the Russian Science Foundation grant 17-77-20034.

Authors contributions

I am the single author of this short note and all results above are created by myself.

Acknowledgments

Author wish to thank Gauthier Hulot for discussions and suggestions.

References

- Barndorff-Nielsen O , Kent J , Sørensen M (1982) Normal variance-mean mixtures and z distributions. International Statistical Review 50:145–159, doi: 10.2307/1402598
- Feller W (1971) An introduction to probability theory and its applications. Vol.2 (3rd edition) Wiley
- Fratter I, Léger JM, Bertrand F, Jager T, Hulot G, Brocco L, Vigneron P, (2016) Swarm Absolute Scalar Magnetometers first in-orbit results. Acta Astronautica **121**:76–87 doi: 10.1016/j.actaastro.2015.12.025
- Jackson A, Jonkers ART, Walker MR (2000) Four centuries of geomagnetic secular variation from historical records. Phil Trans R Soc Lond A 358:957–990 doi: 10.1098/rsta.2000.0569.
- Khokhlov A, Hulot G (2017) On the cause of the non-Gaussian distribution of residuals in geomagnetism. Geophys J Int 209:1036–1047 doi: 10.1093/gji/ggx071
- Léger JM, Jager T, Bertrand F, Hulot G, Brocco L, Vigneron P, Lalanne X, Chulliat A, Fratter I (2015) In-flight performances of the absolute scalar magnetometer vector mode on board the Swarm satellites. Earth Planets Space 67:57–68 doi: 10.1186/s40623-015-0231-1
- Olsen N, Hulot G, Lesur V, Finlay CC, Beggan C, Chulliat A, Sabaka TJ, Floberghagen R, Friis-Christensen E, Haagmans R, Kotsiaros S, Lühr H, Tøffner-Clausen L, Vigneron P (2015) The Swarm Initial Field Model for the 2014 geomagnetic field. Geophys Res Lett **42**:1092–1098 doi: 10.1002/2014GL062659
- Panovska S, Finlay CC, Donadini F, Hirt A, (2012) Spline analysis of Holocene sediment magnetic records: Uncertainty estimates for field modeling. JGR Solid Earth 117:B02101 doi:10.1029/2011JB008813
- Panovska S, Korte M, Finlay CC, Constable CG (2015) Limitations in paleomagnetic data and modelling techniques and their impact on Holocene geomagnetic field models. Geophys J Int, 202(1):402–418 doi:10.1093/gji/ggv137

Figure 1. Standard deviations (in nT) computed every day for the mid-latitude residuals of the Swarm scalar data used to compute the VFM model of Vigneron et al. (2015). Blue large dots: data from the Swarm Alpha satellite and red dots: data from the Swarm Bravo satellite. Days are counted in Julian days, with 2000 January 1 taken as the reference.

Vigneron P, Hulot G, Olsen N, Lger JM, Jager T, Brocco L, Sirol O, Coisson P, Lalanne X, Chulliat
A, Bertrand F, Boness A, Fratter I (2015) A 2015 International Geomagnetic Reference Field (IGRF)
Candidate Model Based on Swarms Experimental Absolute Magnetometer Vector Mode Data. Earth
Planets Space 67:95 doi:10.1186/s40623-015-0265-4

Walker MR, Jackson A (2000) Robust modelling of the Earth's magnetic field. Geophys J Int, 143(3):799–
808 doi:10.1046/j.1365-246X.2000.00274.x

Figure 2. Left: Histogram of the residuals (circles) of the Swarm Alpha scalar data used to compute the VFM model of Vigneron et al. (2015) together with histogram (triangles) of an identical number of simulated mixture of Gaussian distributions according to the parameters $s = 0.41$, $\sigma = 2.18$ recovered from the real data; right: the same plots but for the Swarm Bravo scalar data, parameters $s = 0.47$ and $\sigma = 2.10$.

Figures

Figure 1

Standard deviations (in nT) computed every day for the mid-latitude residuals of the Swarm scalar data used to compute the VFM model of Vigneron et al. (2015). Blue large dots: data from the Swarm Alpha satellite and red dots: data from the Swarm Bravo satellite. Days are counted in Julian days, with 2000 January 1 taken as the reference.

Figure 2

Left: Histogram of the residuals (circles) of the Swarm Alpha scalar data used to compute the VFM model of Vigneron et al. (2015) together with histogram (triangles) of an identical number of simulated mixture of Gaussian distributions according to the parameters $s = 0.41$, $\sigma = 2.18$ recovered from the real data; right: the same plots but for the Swarm Bravo scalar data, parameters $s = 0.47$ and $\sigma = 2.10$.