The compositional meaning of a detection limit

K. Gerald van den Boogaart[1] and Raimon Tolosana-Delgado[2] and Matevz Bren[3]

[1]Institut für Stochastik - TU Bergakademie Freiberg, Germany boogaart@math.tu-freiberg.de
[2]Laboratori d'Enginyeria Marítima, - Universitat Politècnica de Catalunya, Barcelona, Spain
[3]Faculty of Criminal justice and Security, University of Maribor, Slovenia and
Institute of Mathematics, Physics and Mechanic, Ljubljana, Slovenia

1 What is a detection limit

In compositional data analysis a value below detection limit (BDL) is typically modeled as the definitive information that the actual value is below some fixed value - the detection limit (see e.g. Palarea-Albaladejo et. al (2007, 2008)). Analytical chemistry (Heinrichs and Herrmann (1990); Fletcher (1981); Kellner et al. (2004)) however has a different view on measured concentrations. The measured concentration C_m is not the true concentration c_m of the measurant but a quantity computed from an observable quantity O_m through a calibration equation. E.g.:

$$C_m = \beta O_m - \alpha$$

when the observable is assumed to follow a linear model

$$O_m = a + bc_m + \epsilon_o$$

where ϵ_o is a random measurement error typically modeled by a normal distribution of mean 0 and variance σ_o^2. The variance of ϵ_o might or might not depend on the actual concentration c_m. The parameters σ_o^2, a and b are a-priorly unknown, but are typically estimated during a calibration of the measurement procedure taking place prior to the actual geochemical study we are interested in, i.e. they are assumed as approximately know. Setting $\alpha = \frac{a}{b}$ and $\beta = \frac{1}{b}$ and assuming the estimation errors negligible results in:

$$C_m = \beta(a + bc_m + \epsilon_o) - \alpha = \frac{1}{b}(a + bc_m + \epsilon_o) - \frac{a}{b} = c_m + \frac{\epsilon_o}{b}$$

with $\epsilon_m = \frac{\epsilon_o}{b} \sim N(0, \frac{\sigma_0^2}{b^2}) = N(0, \sigma_m^2)$ we get

$$C_m = c_m + \epsilon_m \tag{1}$$

The variances σ_o^2 and σ_m^2 respectively might depend on the actual concentration c_m. Three different scaling laws have a simple physical interpretation:

- Constant variance: The error might be the effect of sort of additive background noise, e.g. scattered material in a mass spectrometer. This sort of background also produces the intercept a in the observation equation. Some noise is always detected even if the measurant is simply not present at all.

- Variance proportional to the true value: If we count events proportional to the concentration of the measurant, e.g. in a mass spectrometer, we get Poisson counting errors. The variance of a Poisson distribution is proportional to its mean.

- Standard deviation proportional to the true value: Some observation methods have a limited range of proportionality to the concentration of the measurant. Thus it could be part of the measurement procedure to dilute the measurant into range or to measure a different total. The relative standard deviation of the measurement error is than rescaled with the dilution ratio (or total) when the original concentration is recomputed by dividing the result by the dilution ratio.

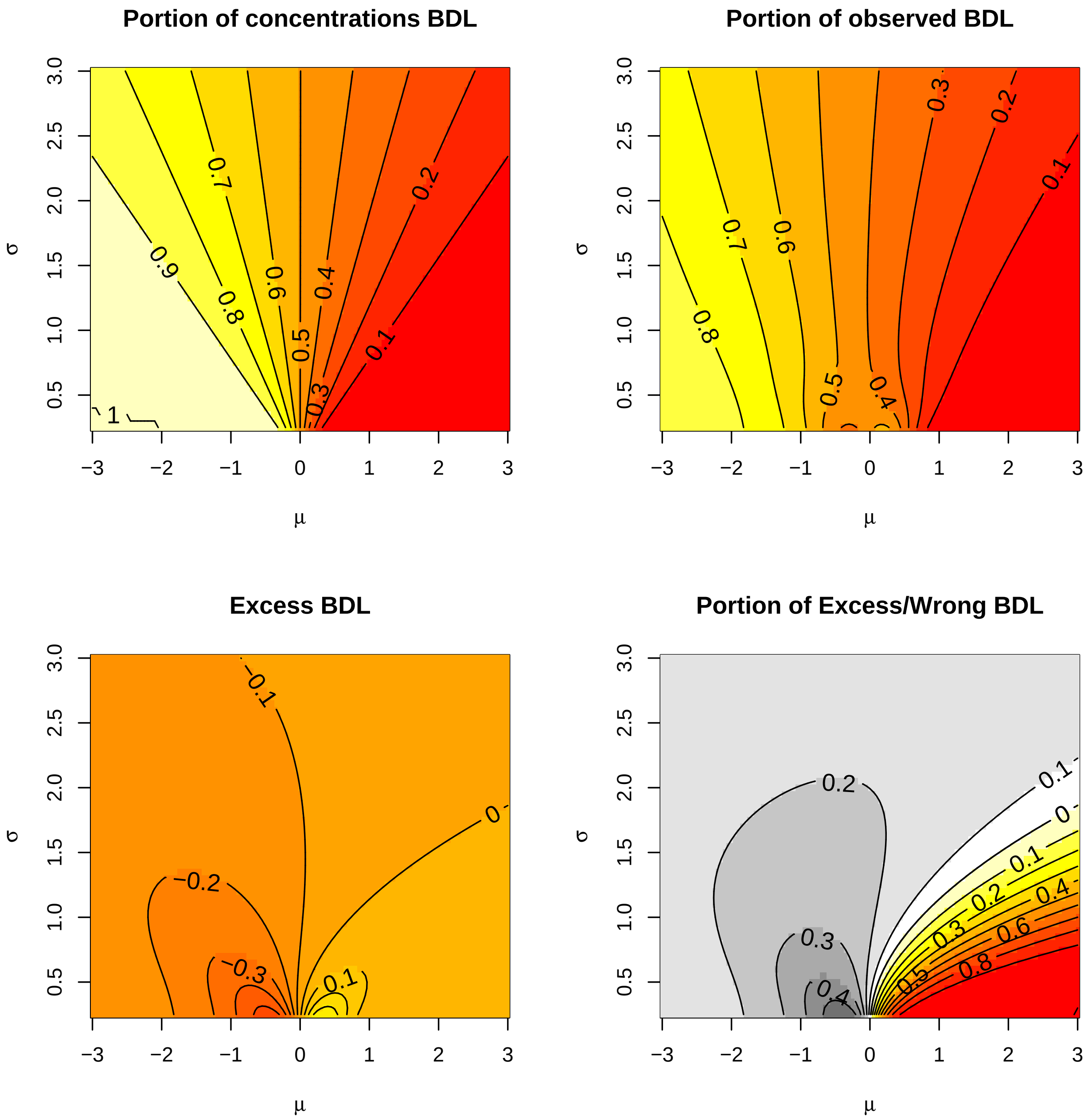

Figure 1: Numerically computed effect of additive normal errors on the observed portion below detection limit values of lognormal concentration distributions.

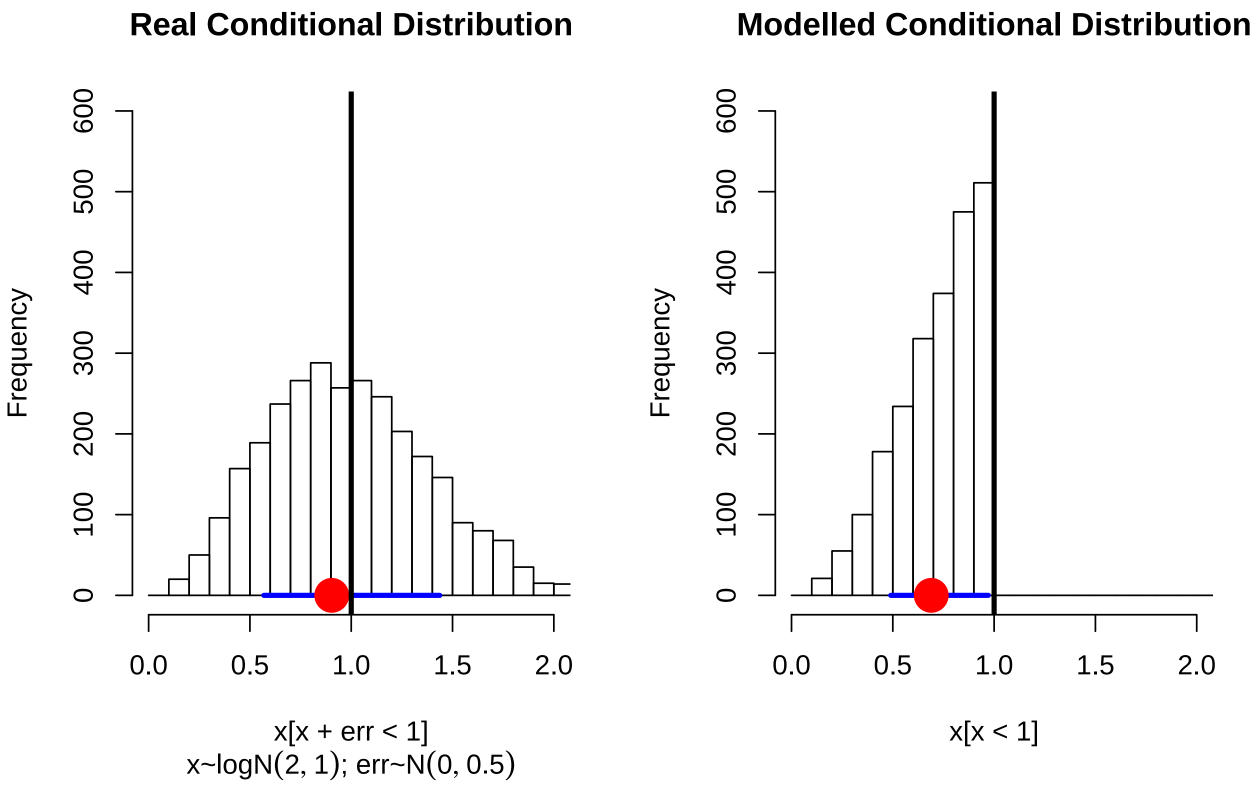

Figure 2: A simulation of the conditional distribution of values below detection limit for lognormal distribution with log-mean=2 and log-sd=1 and an additive measurement error of $N(0, 0.5)$ corresponding to a detection limit of 1. The left panel shows the true conditional distribution of true concentrations for values observed below detection limit. The right hand shows the conditional distribution assuming that wrong model that an observation below detection limit is corresponding to a value below the detection limit. The red dots show the conditional geometric mean and the blue line the conditional geometric standard deviation.

Additional sources of measurement errors or combinations of these different sources of errors can lead to a more complicated dependence of errors. For near 0 concentrations however typically the constant error will dominate, leading to the effect that zero or near zero concentrations c_m can lead to measurements observations O_m above or below a and thus to C_m below or above zero. To avoid such nonsense values analytical chemistry only reports measurements for which the observant O_m and/or respectively C_m is so high that from $c_m = 0$. Assuming a Gaussian distribution of the observation error this corresponds to a value $O_m \geq k\sigma_o + a$ or $C_m \geq k\sigma_m$, where k is some quantile of the normal distribution typically chosen as 2 (Fletcher, 1981, e.g.) or 3 (Kellner et al., 2004, e.g.) corresponding to probabilities of below 0.0023 or 0.00135 to report a nonzero value if actually no measurant is present. As a side effect a negative concentration is never reported. This value $k\sigma_m$ is reported as the detection limit. We will always use $k = 2$ in this paper.

The detection limit is thus the critical value of a Gauss test above which it is statistically proven that measurant is present (Fletcher (1981)). Analytical chemistry reports a value below detection limit, if the actual observation is so small that the statistical tests fails to prove that the measurant is present. This has several consequences for the statistical analysis of geochemical data in the context of lognormal and additive lognormal distributions, which we will discuss in detail. However in this his paper we will not discuss possible solutions specification of the goals of the statistical analysis.

2 There are too many BDLs

In cases, where the true concentration is slightly above the detection limit, the probability of actually observing it, as a BDL is nearly 50%. Likewise actual positive values below the detection limit have

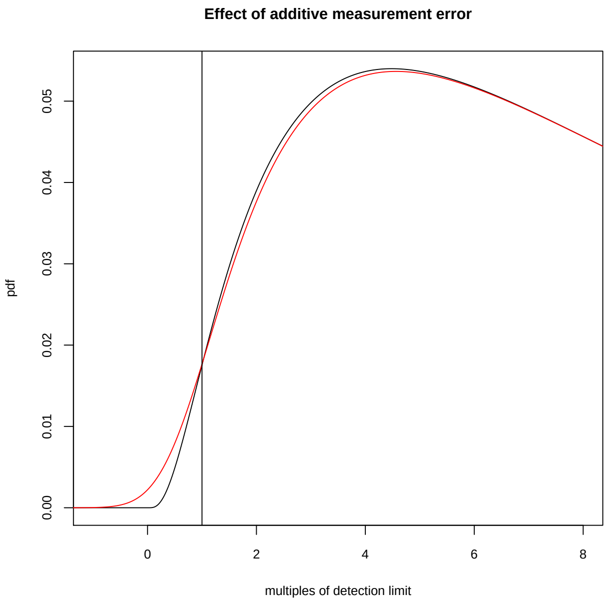

Figure 3: Density of a lognormal concentration distribution (black line) and the density (red line) of the distribution of the values that would be observed with this distribution under the model of an additive normal error. The vertical line marks the detection limit $2\sigma_m$

a similar probability of being observed as above detection limit. However typically the probability density of the true concentrations drops towards zero. Thus there will be more observations wrongly below the detection limit than wrongly above the detection limit. This is illustrated in figure 3, where we plot the density of a log-normal distribution as an example for a concentration distribution and the resulting density of values observed in case of an additive normal measurement error. Below the detection limit this density is clearly higher than the true density, while above the detection limit the density is nearly unchanged by the convolution. I.e. while the distribution of the actually observed measurements is not so different from the true concentration above detection limit, there are less quantitative measurements and more values classified below detection limit than true concentrations are below the detection limit. This is shown in figure 1. Each panel of the figure has a parameter log-median and log-standard deviation of the lognormal distribution measured in multiples of detection limit. The panels show the true portion of concentrations below detection limit, portion of observations below the detection limit, the difference (i.e. the additional portion), and the portion of excess and wrong BDLs, respectively. The panel of excess and wrong BDLs shows portion of the wrong BDLs on all BDLs in color and the portion of missing BDLs in gray scale. Thus the number of observed BDLs differs substantially from the the number of concentrations below the detection limit. A value of 1 in the color area means e.g. that nearly all observed BDL are originating from concentrations above detection limit. A value of 0.5 in the grey area would mean that 50% of the concentration below detection limit are not reported as such.

3 Imputation based Estimation is biased

In state of the art imputation algorithms (Palarea-Albaladejo et. al, 2007, 2008, e.g.) for compositional data the values are imputed according to the conditional expectation or conditional distribution given being below detection limit in an additive lognormal model. Further estimates are based on these imputed values. The portion of actually observed BDLs is different – often higher – than the portion of BDLs in the distribution. Using the conditional distribution of additive lognormal distribution will still lead to a wrong portion of the low values. The estimates of mean compositions based on an imputed datasets in the state of the art model are thus biased towards lower values in this component.

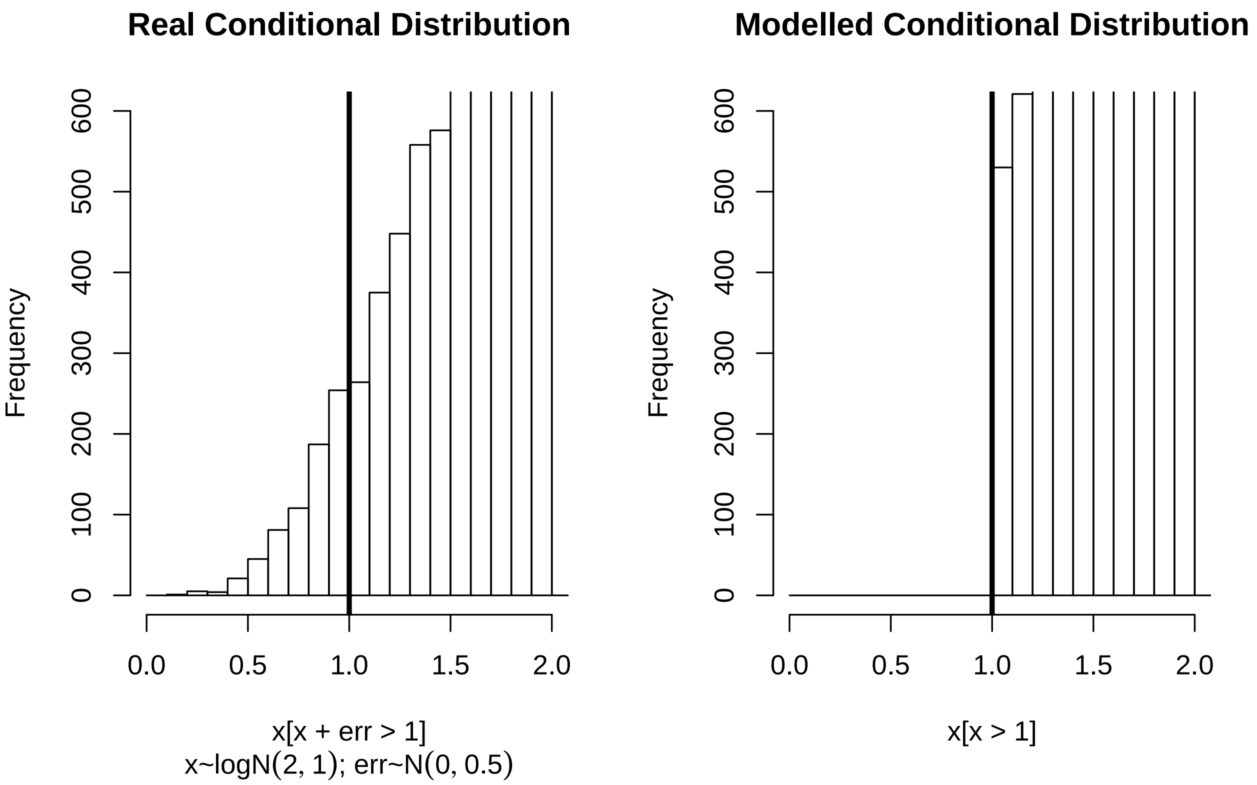

Figure 4: A simulation of the conditional distribution of values above detection limit for lognormal distribution with log-mean=2 and log-sd=1 and an additive measurement error of $N(0, 0.5)$ corresponding to a detection limit of 1. The left panel shows the true conditional distribution of true concentrations for values observed above detection limit. The right hand shows the conditional distribution assuming that wrong model that an observation below detection limit is corresponding to a value above the detection limit.

Figure 2 shows the difference between the two conditional distributions for an example situations. The left panel shows a simulation of the true conditional distribution of concentrations given that the observed value is BDL and the right panel shows the conditional distribution which is used in the state of the art modeling principle.

However simply replacing the conditional distributions does not solve the problem. Figure 4 shows the opposite problem that the actually observed values do not represent an unbiased sample of the true concentrations. Unlike the classical model it is thus not possible to do a separate imputation with the conditional distribution, because even the values above detection limit are biased due to the censoring with the detection limit. I.e. a replaced imputation would lead to an overestimation of mean concentration in this component.

An naive idea would be to simply model the observable rather than the true concentrations. However as visible in figure 3 the distribution of the observable values is not limited to positive values and thus not compatible with the basic principles of compositional data analysis. A “correct” imputation of BDLs is thus impossible.

4 High relative errors for low values

The detection limit is the limit of observations that with high probability should not be observed if we have an actual concentration of zero. However the typical assumption of compositional data analysis is anyway that zero is an impossible value occurring with probability zero. It is much more likely to have values between zero and the detection limit or around the detection limit. Assuming the symmetry of the additive error we would have a reasonable probability for true concentrations at the detection limit to be observed as 0 or as twice the detection limit. Small values are thus observed with a substantial

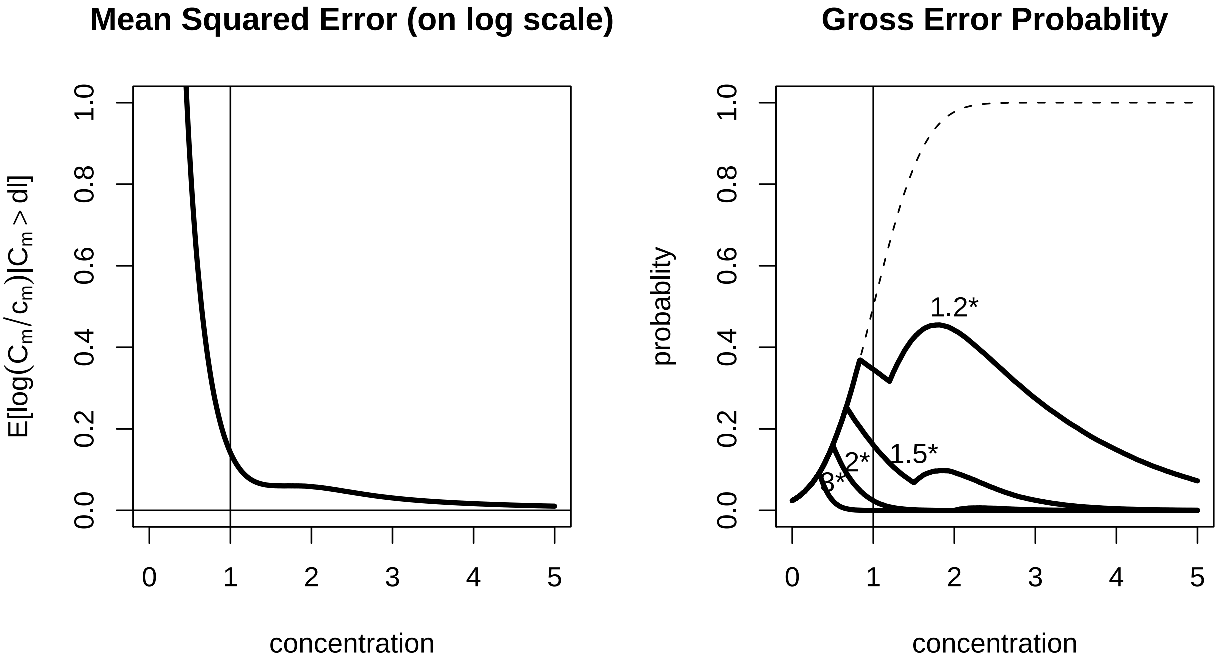

Figure 5: The errors as a function the true concentration given as multiples of the detection limit. The left panel shows the MSLE (Mean Squared Logarithmic error) and the right panel the probability of a gross overestimation by given factors. The dashed line is the probability of being observed at all.

relative error. This is illustrated in figure 5. The left panel shows the expected squared error of logs as a function of the true concentration, conditional to being above detection limit. The unconditional expected squared error of logs without a detection limit is anyway infinite. The right panel shows the probability to differ from the true value by at least a given factor (upwards or downwards). Both figures show that concentrations near to the detection have a substantially higher relative error and in cases of relatively small compositional variance are even prone to gross outliers. All probabilities include the fact that the concentration might be not observed at all because of the detection limit. For concentrations around the detection limit the gross error probability might be quite substantial in a relative analysis.

5 Subcompositional incoherence of geochemical measurement errors

The relative measurement error thus depends on the ratio of concentration and detection limit. A dataset with a different total will be measured with a different relative error and will thus have a different distribution. Similarly a third component can change the precision of the measurement of two other components. This is illustrated in figure 6 with a simulated additive lognormally distributed dataset with parameters $\mu = c(0.3, 0.3, 0.7)$ and

$$\Sigma_{clr} = \begin{pmatrix} 0.055 & 0.045 & -0.1 \\ 0.045 & 0.055 & -0.1 \\ -0.100 & -0.100 & 0.2 \end{pmatrix}$$

and detection limit of 0.04 in all components with the corresponding additive measurement error $N(0, 0.02^2)$. One can clearly see the dependence of the relative errors in the second ilr balance on the first ilr balance.

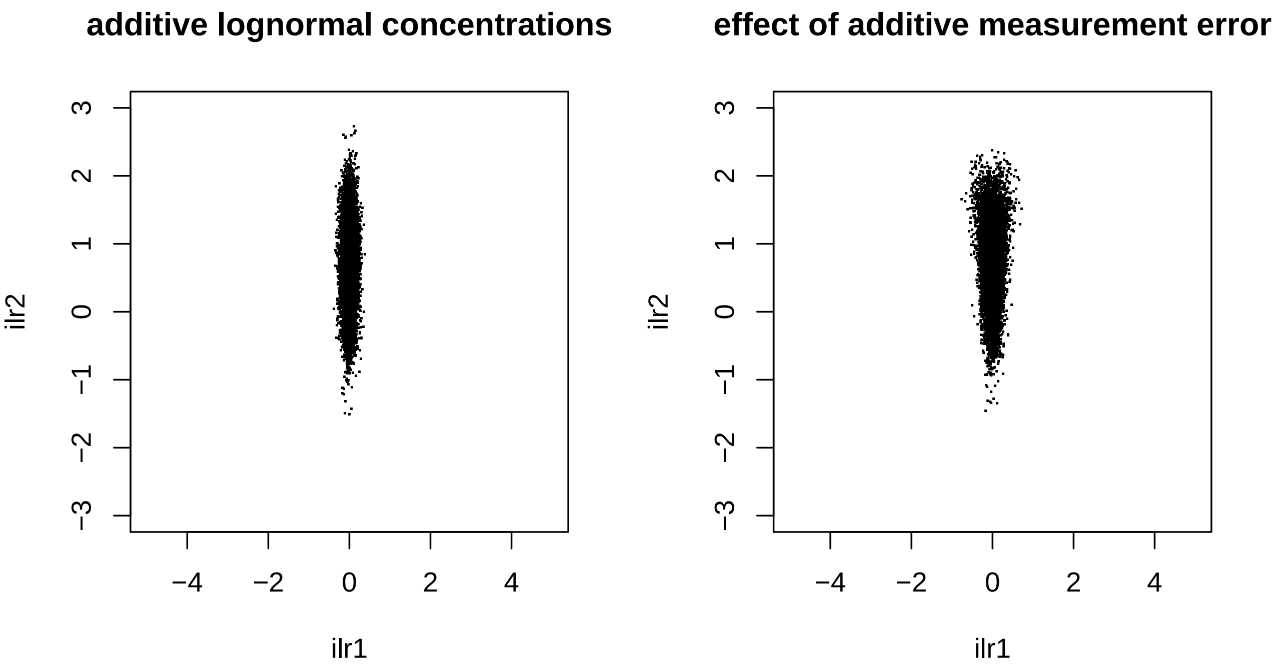

Figure 6: The left panel shows a additive lognormal in ilr coordinates. The right panel shows the same compositional dataset, where an additive error is applied and all missing values omitted.

6 Conclusions

The chemical interpretation of the detection limit and its stochastic model counterpart has thus different consequences for the statistical analysis than we would expect from the word by word interpretation of "below detection limit" as a concentration below some limit. The state of the art model on BDL compositional analysis is biased.

Some ideas are not directly applicable to the true effects of measurement errors near the detection limit. Even the basic principles like subcompositional coherence and the requirement of the independence of the analysis from the total are not fully valid near the detection limit.

References

Fletcher, W.K. (1981). *Handbook of Exploration Geochemistry, Volumne 1, Analytical Methods in Geochemical Processing* Elsevier Scientific publishing Company, Amsterdam, Oxford, New York. 255 p.

Heinrich, H., A.G. Herrmann (1990). *Praktikum der Analytischen Geochemie* Springer Lehrbuch, Springer Verlag, Berlin 669 p.

Kellner, R., J.M. Mermet, M. Otto, M. Valcárecel, H.M. Widmer (2004). *Analytical Chemistry, A Modern Approach to Analytical Science, Second edition* Whiley-VCH 1181 p.

Palarea-Albaladejo, J. and J.A. Martín-Fernández (2007) A modified EM alr-algorithm for replacing rounded zeros in compositional data sets *Mathematical Geology 39*, 625–645.

Palarea-Albaladejo, J. and J.A. Martín-Fernández (2008) A modified EM alr-algorithm for replacing rounded zeros in compositional data sets *Computers & Geosciences 34*(8), 902–917.