Density-on-scalar, scalar-on-density and density-on-density functional regression

30 April 2021

Karel Hron

Department of Mathematical Analysis and Applications of Mathematics

Faculty of Science – Palacký University, Olomouc, Czech Republic

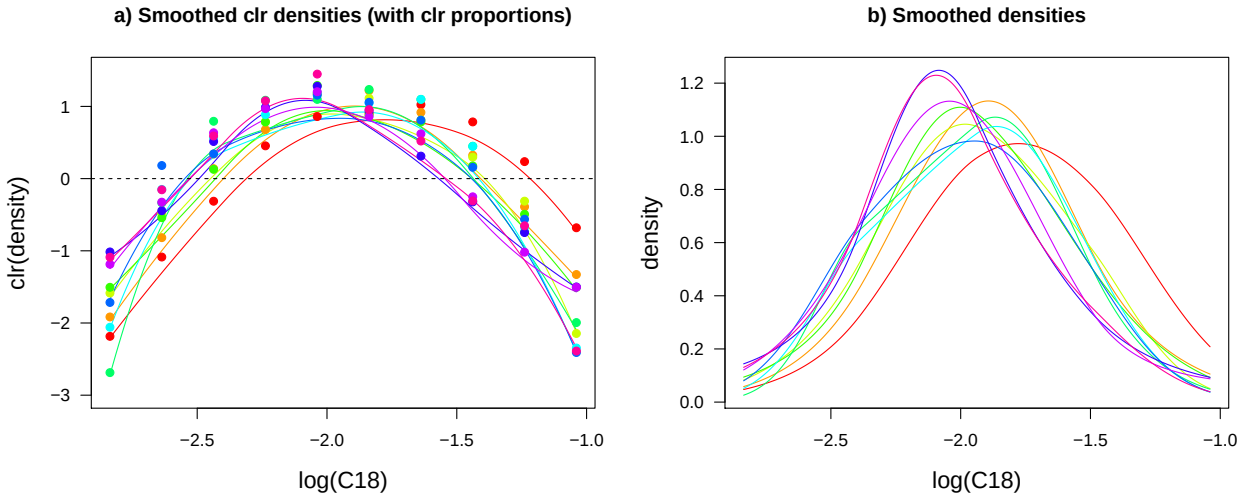
Outline

1. Density-on-scalar regression
2. Scalar-on-density regression
3. Density-on-density regression (outlook)

Modeling metabolite distributions in newborns

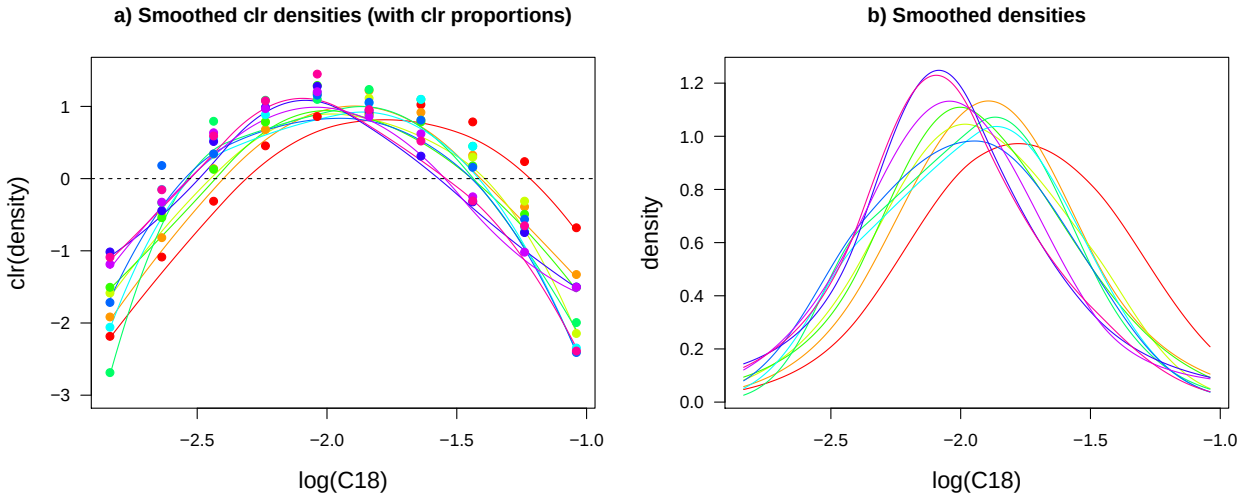
- Faculty Hospital in Olomouc: the weight and gender of every newborn are observed, together with 48 metabolic parameters measured from dried blood spots of each newborn
- The dataset we consider collects the data about 10,000 newborns with standard weights

Modeling metabolite distributions in newborns



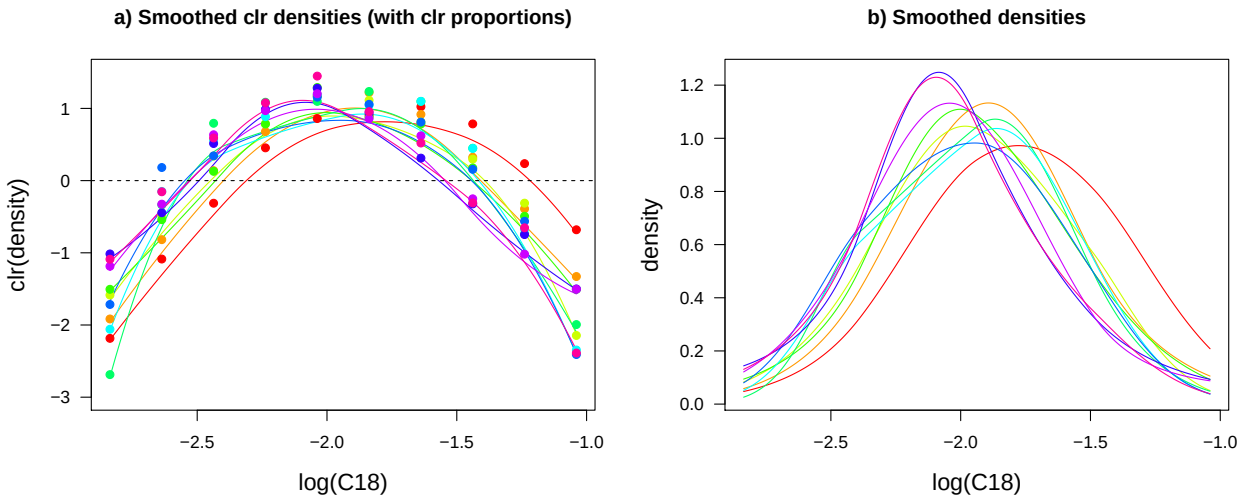
- Metabolite C18 is considered, which is presumed to be closely connected with the weight of newborns

Modeling metabolite distributions in newborns



- Metabolite C18 is considered, which is presumed to be closely connected with the weight of newborns
- The values of the logarithm of C18 were divided into 10 groups of equal size according to the logarithm of weight

Modeling metabolite distributions in newborns



- Metabolite C18 is considered, which is presumed to be closely connected with the weight of newborns
- The values of the logarithm of C18 were divided into 10 groups of equal size according to the logarithm of weight
- **Aim**: To model the dependence of C18 distribution on weight

Recall: Spline representation of PDFs

$$
\begin{array}{ccc}
f(t) \in \mathcal{B}^2(I) & \xrightarrow{\text{clr}} & \operatorname{clr}(f)(t) \in L_0^2(I) \\
 & & \Big\downarrow \text{spline representation} \\
\xi_k(t) \in \mathcal{C}_k^{\Delta\lambda}(I) & \xleftarrow[\text{clr}^{-1}]{} & s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}(I)
\end{array}
$$

Spline representation of PDFs: ZB-splines

- B-spline basis system having zero-integral on $I = [a, b]$ is needed for centred logratio (clr) transformed densities $\rightarrow$ ZB-splines
- given sequence of knots $\Delta\lambda$ and $k \geq 0$, $Z_i^{k+1}(t)$ are defined as

$$Z_i^{k+1}(t) := \frac{\mathrm{d}}{\mathrm{d}t} B_i^{k+2}(t)$$

$\rightarrow$ every spline $s_k(t) \in \mathcal{Z}_k^{\Delta\lambda}[a, b]$ in $L_0^2(I)$ has a unique representation ($g \ldots$ number of inner knots)

$$s_k(t) = \sum_{i=-k}^{g-1} b_i^z Z_i^{k+1}(t)$$

Spline representation of PDFs: CB-splines

- back-transformation of ZB-splines $Z_i^{k+1}(t)$ to $\mathcal{B}^2(I) \rightarrow$ CB-splines:

$$\zeta_i^{k+1}(t) = \exp[Z_i^{k+1}(t)]$$

Spline representation of PDFs: CB-splines

- back-transformation of ZB-splines $Z_i^{k+1}(t)$ to $\mathcal{B}^2(I) \rightarrow$ CB-splines:

$$\zeta_i^{k+1}(t) = \exp[Z_i^{k+1}(t)]$$

$\rightarrow$ every spline $\xi_k(t) \in \mathcal{C}_k^{\Delta\lambda}[a,b]$ in $\mathcal{B}^2(I)$ has a unique representation

$$\xi_k(t) = \bigoplus_{i=-k}^{g-1} b_i^z \odot \zeta_i^{k+1}(t)$$

$\xi_k(t) \ldots$ *compositional spline*

Function-on-scalar regression in Bayes spaces

- Model with functional response $y(t)$ from $\mathcal{B}^2(I)$ and scalar regressors, $x_0, \ldots, x_r$,

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} [x_{ij} \odot \beta_j](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, N$$

Function-on-scalar regression in Bayes spaces

- Model with functional response $y(t)$ from $\mathcal{B}^2(I)$ and scalar regressors, $x_0, \ldots, x_r$,

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} \left[x_{ij} \odot \beta_j\right](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, N$$

- We aim to minimize the residual sum of squares

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|_B^2 = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} \left[x_{ij} \odot \beta_j\right] \ominus y_i \right\|_B^2 .$$

Function-on-scalar regression in Bayes spaces

- Model with functional response $y(t)$ from $\mathcal{B}^2(I)$ and scalar regressors, $x_0, \ldots, x_r$,

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} [x_{ij} \odot \beta_j](t) \oplus \varepsilon_i(t), \quad i = 1, \ldots, N$$

- We aim to minimize the residual sum of squares

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\varepsilon_i\|_B^2 = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} [x_{ij} \odot \beta_j] \ominus y_i \right\|_B^2 .$$

- This minimization problem can be mapped into L_0^2 space and approach based on ZB-spline representation of clr transforms – $y_i(t)$ and $\beta_j(t)$ – is applied.

Function-on-scalar regression in Bayes spaces

- Both the responses and the regression parameters are represented as ZB-splines:

$$\begin{aligned} \mathrm{clr}(y_i)(t) &= \sum_{m=-k}^{g-1} Y_{i,m+k+1} Z_m^{k+1}(t) \\ \mathrm{clr}(\beta_j)(t) &= \sum_{m=-k}^{g-1} b_{j,m+k+1} Z_m^{k+1}(t) \end{aligned}$$

$i = 1, \ldots, N, j = 0, \ldots, r$

$\Rightarrow$ The regression model is rewritten as

$$(\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_{g+k}) = \mathrm{X}(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{g+k}) + (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{g+k}).$$

Function-on-scalar regression in Bayes spaces

- Resulting estimates are obtained as

$$\operatorname{clr}(\hat{\beta}_j)(t) = \sum_{m=-k}^{g-1} \hat{b}_{j,m+k+1} Z_m^{k+1}(t), \quad j = 0, \ldots, r,$$

respectively,

$$\hat{\beta}_j(t) = \bigoplus_{m=-k}^{g-1} \hat{b}_{j,m+k+1} \odot \zeta_m^{k+1}(t), \quad j = 0, \ldots, r.$$

Function-on-scalar regression in Bayes spaces

- Resulting estimates are obtained as

$$\mathrm{clr}(\hat{\beta}_j)(t) = \sum_{m=-k}^{g-1} \hat{b}_{j,m+k+1} Z_m^{k+1}(t), \quad j = 0, \ldots, r,$$

respectively,

$$\hat{\beta}_j(t) = \bigoplus_{m=-k}^{g-1} \hat{b}_{j,m+k+1} \odot \zeta_m^{k+1}(t), \quad j = 0, \ldots, r.$$

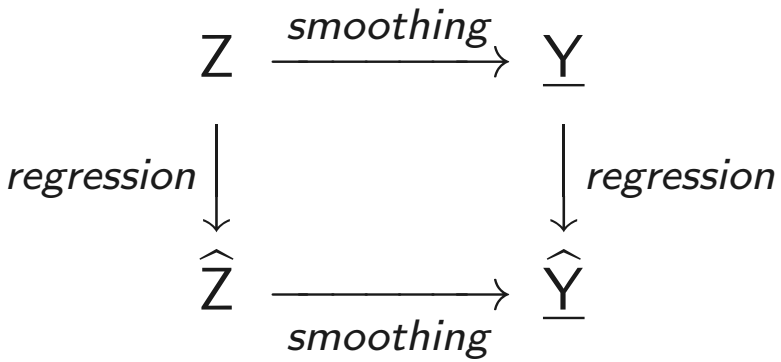
- Alternatively (Talská et al., 2018), for B-spline representation, $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$, $s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$, the zero integral condition $\sum_{i=-k}^{g} b_i (\lambda_{i+k+1} - \lambda_i) = 0$ is kept for $\mathrm{clr}(\hat{\beta}_j)(t)$

Two equivalent approaches to the problem

- One can show the coherence of the following approaches
  - Smoothing of the discretely clr transformed PDFs $\underline{Z} = (Z_1, Z_2, \ldots, Z_D)$, then regression with ZB-spline coefficients (we did)

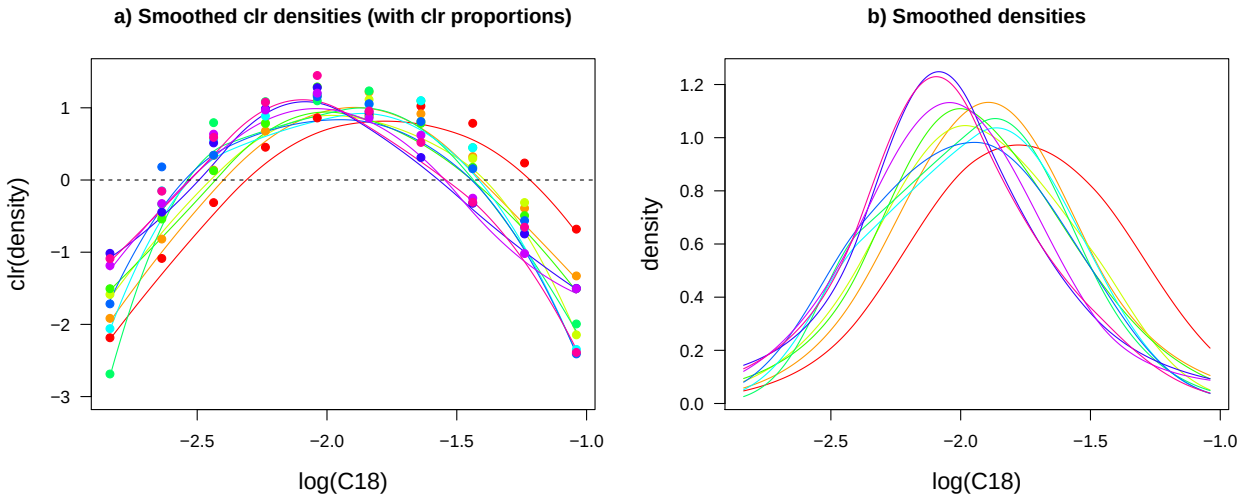
Two equivalent approaches to the problem

- One can show the coherence of the following approaches
  - Smoothing of the discretely clr transformed PDFs $\underline{Z} = (Z_1, Z_2, \ldots, Z_D)$, then regression with ZB-spline coefficients (we did)
  - Compositional regression (with $\underline{Z}$ as the response), then smoothing



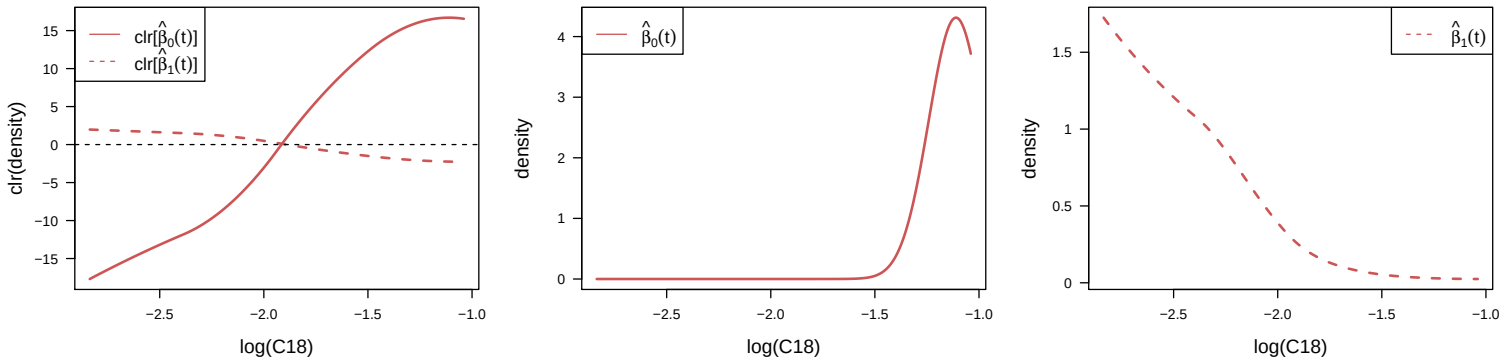
Modeling metabolite distributions in newborns

To remind:



Aim: To model the dependence of C18 distribution on weight of newborns

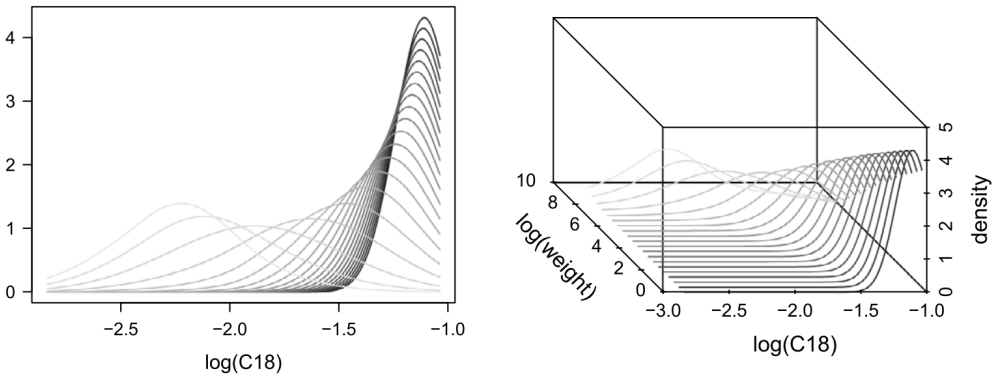
Results



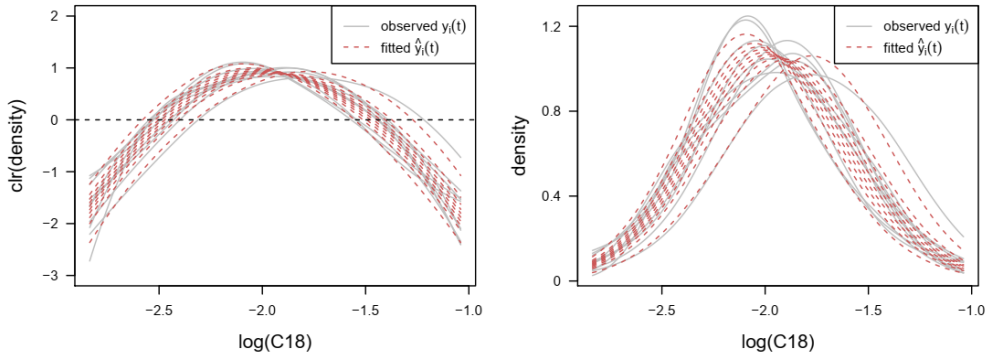
Estimates of regression functions (in the case of girls)

- Intercept: baseline PDF
- Slope: if the weight increases, PDFs tend to be more concentrated in the left part of the domain, and associated with higher variabilities (see bootstrap results)

Goodness of fit

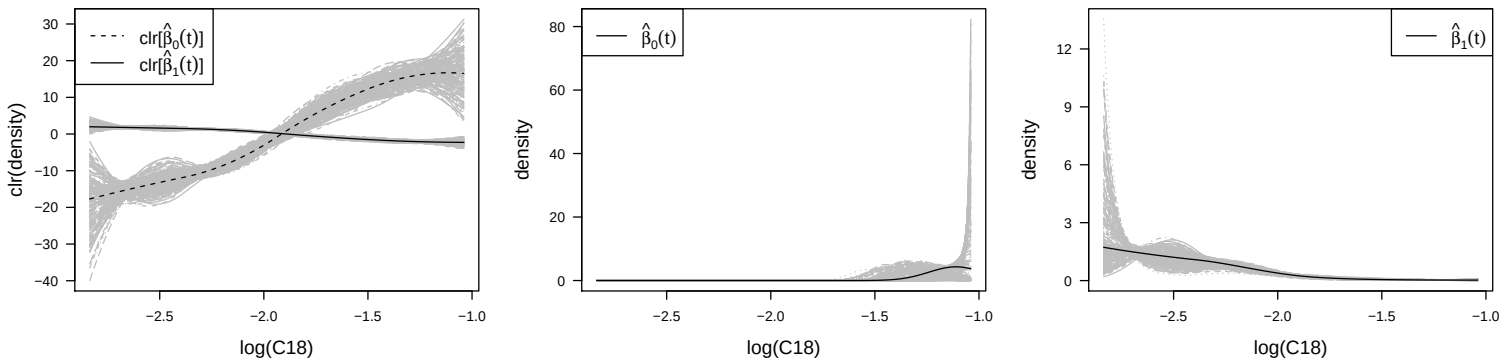


2D and 3D graphs of predicted distributions for increasing sequence of 20 values of log weights (girls)



Comparison of observed y (grey) and fitted $\hat{y}$ (colored curves) distributions in L_0^2 and $\mathcal{B}^2$ space; $R^2 = 0.728$.

Uncertainty assessment



Bootstrap results (based on resampling of the model-residuals): black curves indicate estimates of regression parameters, grey lines indicate the $R = 200$ bootstrap estimates for both the regression parameters.

Most of the uncertainty in β_0 is shown in the right part of domain, whereas for β_1 it is mostly present in the left part of domain.

Compositional function-on-scalar regression: R code

https://github.com/AMenafoglio/BayesSpaces-codes

(with special thanks to Ivana Pavlů, *Palacký University*)

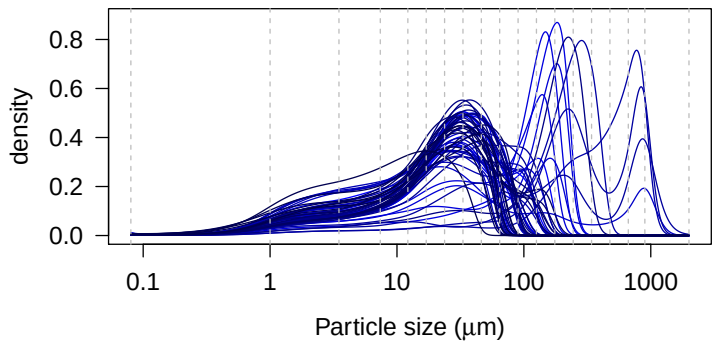
Particle size distributions

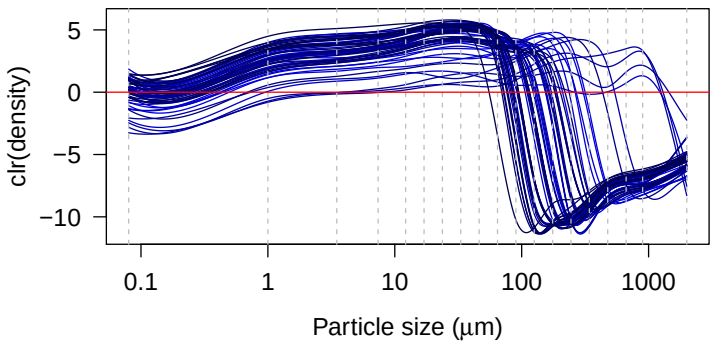
- Sediment samples were obtained in the Czech Republic in the Skalka Reservoir and in the Ohře River floodplain upstream of the reservoir

Particle size distributions

- Sediment samples were obtained in the Czech Republic in the Skalka Reservoir and in the Ohře River floodplain upstream of the reservoir
- **Aim**: To analyze how the geochemical composition is influenced by particle size distribution (PSD) of the samples

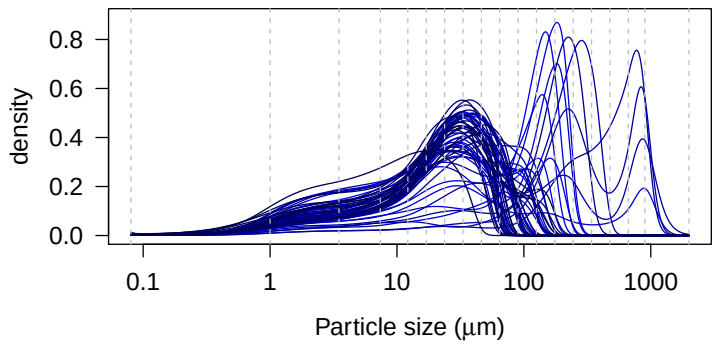
Particle size distributions

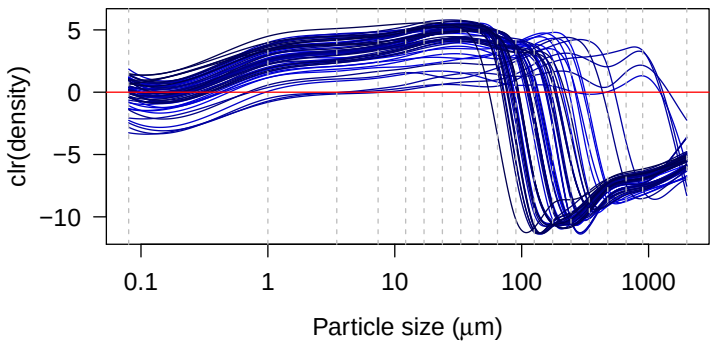




- $N = 59$ discretely sampled PSDs obtained by laser granulometry were turned into smoothed clr transformed densities

Particle size distributions





- $N = 59$ discretely sampled PSDs obtained by laser granulometry were turned into smoothed clr transformed densities
- Functional character of PSDs should be taken into account for regression modeling

Scalar-on-density regression model ($\mathcal{B}^2$ space)

- $f_1, \ldots, f_N$ ... functional measurements (PDFs) of the predictor f in $\mathcal{B}^2(I)$
- $y_1, \ldots, y_N$ ... a sample from the distribution of the real variable y such that $(y_i, f_i), i = 1, \ldots, N$
- The **functional linear model** for the ith observation y_i associated with the predictor f_i is expressed as

$$y_i = \beta_0 + \langle \beta_1(t), f_i(t) \rangle_{\mathcal{B}} + \varepsilon_i, \quad i = 1, \ldots, N, \quad t \in I$$

- $\beta_0 \in \mathbb{R}$, $\beta_1(t) \in \mathcal{B}^2(I)$ ... unknown regression parameters, ε is a random vector of i.i.d. random errors with zero-mean

Scalar-on-density regression model (L_0^2 space)

- Clr transformation of the sampled PDFs f_i:

$$\mathrm{clr}(f_i)(t) = f_i^c(t) = \ln f_i(t) - \frac{1}{\eta}\int_I \ln f_i(s)\,\mathrm{d}s$$

Scalar-on-density regression model (L_0^2 space)

- Clr transformation of the sampled PDFs f_i:

$$\mathrm{clr}(f_i)(t) = f_i^c(t) = \ln f_i(t) - \frac{1}{\eta}\int_I \ln f_i(s)\,\mathrm{d}s$$

$\Rightarrow$ The regression model can be equivalently restated as

$$\begin{aligned} y_i &= \beta_0 + \langle \mathrm{clr}(\beta_1)(t), \mathrm{clr}(f_i)(t)\rangle_2 + \varepsilon_i \\ &= \beta_0 + \int_I \mathrm{clr}(\beta_1)(t)\cdot \mathrm{clr}(f_i)(t)\,dt + \varepsilon_i \end{aligned}$$

- The regression parameters are estimated by minimizing

$$\mathrm{SSE}(\beta_0,\beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \langle \mathrm{clr}(\beta_1)(t), \mathrm{clr}(f_i)(t)\rangle_2)^2$$

Basis system representation

- For estimation of β_1 in the clr space, a proper basis representation should be used
- **Possible danger:** the total number of basis functions will exceed or at least approach the number of observations (N) ($\rightarrow$ LS estimation in the associated regression model will fail)

Basis system representation

- For estimation of β_1 in the clr space, a proper basis representation should be used
- **Possible danger:** the total number of basis functions will exceed or at least approach the number of observations (N) ($\rightarrow$ LS estimation in the associated regression model will fail)
- **Way out:** use any of regularization approaches (Ramsay and Silverman, 2005; Talská et al., 2021):

Basis system representation

- For estimation of β_1 in the clr space, a proper basis representation should be used
- **Possible danger:** the total number of basis functions will exceed or at least approach the number of observations (N) ($\rightarrow$ LS estimation in the associated regression model will fail)
- **Way out:** use any of regularization approaches (Ramsay and Silverman, 2005; Talská et al., 2021):
  - Keep dimensionality of the basis expansion low (low-dimensional regression)

Basis system representation

- For estimation of β_1 in the clr space, a proper basis representation should be used
- **Possible danger:** the total number of basis functions will exceed or at least approach the number of observations (N) ($\rightarrow$ LS estimation in the associated regression model will fail)
- **Way out:** use any of regularization approaches (Ramsay and Silverman, 2005; Talská et al., 2021):
  - Keep dimensionality of the basis expansion low (low-dimensional regression)
  - Estimation using a roughness penalty (penalized regression)

Basis system representation

- For estimation of β_1 in the clr space, a proper basis representation should be used
- **Possible danger:** the total number of basis functions will exceed or at least approach the number of observations (N) ($\rightarrow$ LS estimation in the associated regression model will fail)
- **Way out:** use any of regularization approaches (Ramsay and Silverman, 2005; Talská et al., 2021):
  - Keep dimensionality of the basis expansion low (low-dimensional regression)
  - Estimation using a roughness penalty (penalized regression)
  - Reduce dimensionality using SFPCA (SFPCA regression)

Basis system representation

- Basis (possibly ZB-spline) expansion of $f_i(t)$ and $\beta_1(t)$:

$$\mathrm{clr}(f_i)(t) = \sum_{k=1}^{K_f} c_{ik}\varphi_k(t), \;\; i = 1, \ldots, N, \quad \mathrm{clr}(\beta_1)(t) = \sum_{k=1}^{K_\beta} b_k\psi_k(t)$$

$\Rightarrow$ Reducing functional model into multivariate case:

$$\mathbf{y}_{N\times 1} = \mathbf{X}_{N\times(K_\beta+1)}\,\overline{\mathrm{b}} + \varepsilon_{N\times 1} \quad (\overline{\mathrm{b}} \text{ includes also the intercept})$$

Basis system representation

- Basis (possibly ZB-spline) expansion of $f_i(t)$ and $\beta_1(t)$:

$$\mathrm{clr}(f_i)(t) = \sum_{k=1}^{K_f} c_{ik}\varphi_k(t), \;\; i = 1, \ldots, N, \quad \mathrm{clr}(\beta_1)(t) = \sum_{k=1}^{K_\beta} b_k\psi_k(t)$$

$\Rightarrow$ Reducing functional model into multivariate case:

$$\mathbf{y}_{N\times 1} = \mathbf{X}_{N\times(K_\beta+1)}\,\overline{\mathrm{b}} + \varepsilon_{N\times 1} \quad (\overline{\mathrm{b}} \text{ includes also the intercept})$$

- Assuming that $K = K_f = K_\beta$ has been fixed, we can go for

$$\mathrm{clr}(\tilde{f}_i)(t) = \sum_{j=1}^{K} c_{ij}\xi_j(t), \quad \mathrm{clr}(\beta_1)(t) = \sum_{j=1}^{K} b_j\xi_j(t), \qquad (1)$$

$t \in I$, where $c_{ij} = \langle \mathrm{clr}(\tilde{f}_i), \xi_j\rangle_2$ and $b_j = \langle \mathrm{clr}(\beta_1), \xi_j\rangle_2$ are scores associated with the j-th SFPC $\xi_j, j = 1, \ldots, K$.

Parameter estimates

- The resulting LS estimate of the parameter β_1 in $L_0^2(I)$ space,

$$\mathrm{clr}(\hat{\beta}_1)(t) = \sum_{j=1}^{K} \hat{b}_j \xi_j(t),$$

can be mapped to the original $\mathcal{B}^2(I)$ space by using exponential:

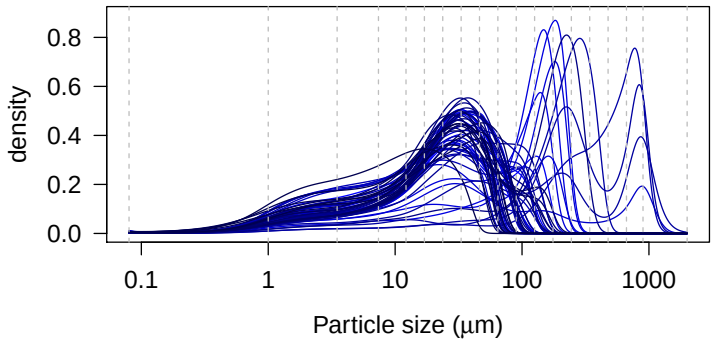
$$\hat{\beta}_1(t) = \bigoplus_{j=1}^{K} \hat{b}_j \odot \zeta_j(t),$$

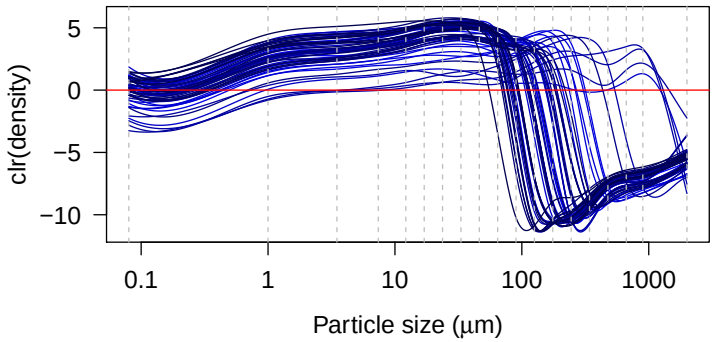
where $\zeta_j = \exp[\xi_j], j = 1, \ldots, K$

- For the purpose of interpretation, it is preferable to consider β_1 in $L_0^2(I)$ space

Interpretation of β_1

- Back to example with PSDs:

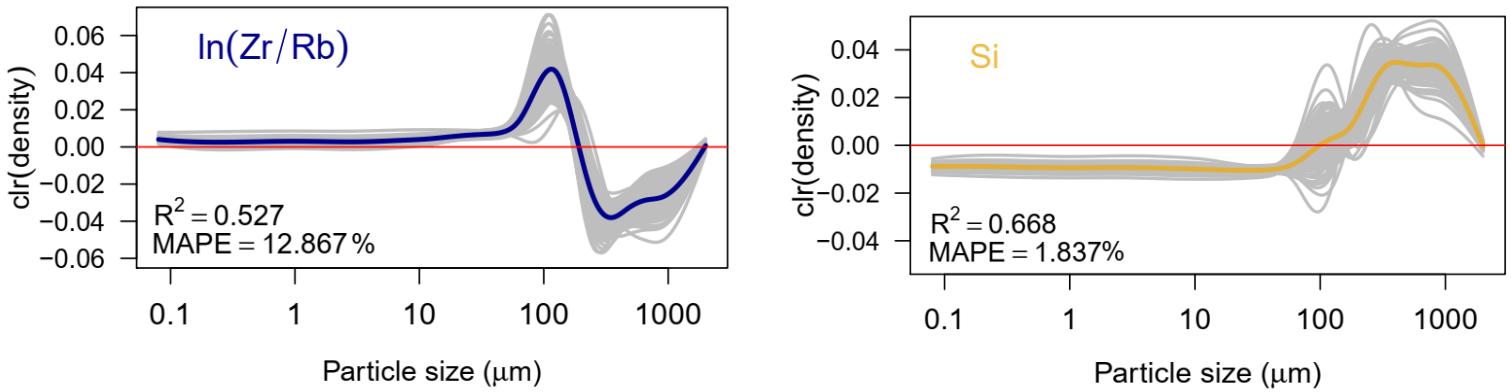




- Generally, *positive values of the regression parameter (in its L_0^2 representation) for a specific fraction induce higher values of the response, by considering the course (absolute values) of the sampled PSDs, and vice versa for negative values*

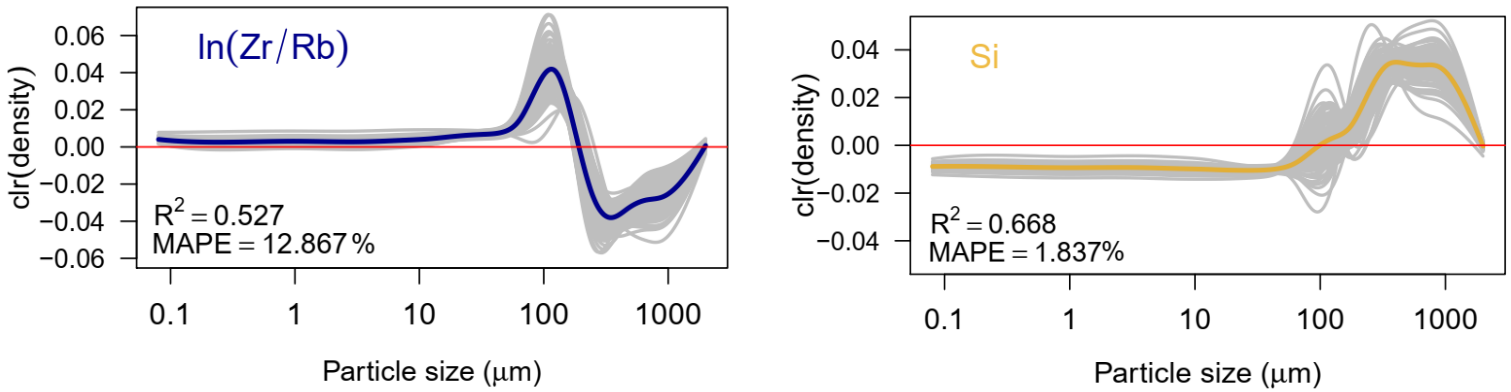
- Response: concentrations of chemical elements, expressed as pairwise log-ratios or clr coefficients/pivot coordinates (Filzmoser et al., 2018)

Results



Zr/Rb (log-)ratio is elevated at larger contribution of medium sized fraction, clr(Si) is most abundant in quartz of sand grain size (particles > 100 μm).

Results



Zr/Rb (log-)ratio is elevated at larger contribution of medium sized fraction, clr(Si) is most abundant in quartz of sand grain size (particles $> 100\ \mu m$).

- Incorporating an uncertainty in the estimation of functional regression parameters $\rightarrow$ bootstrap confidence bands
- They can be used to identify the parts of the domain I which are not related to the response: here the bands would contain zero values

Compositional scalar-on-function regression: R code

https://github.com/AMenafoglio/BayesSpaces-codes

(with special thanks to Ivana Pavlů, *Palacký University*)

Density-on-density regression model

- The regression model can be defined for response PDFs $g_i(t)$ and predictor PDFs $f_{ij}(t), j = 1, \ldots, r$ as

$$\mathrm{clr}(g_i)(t) = \mathrm{clr}(\beta_0)(t) + \sum_{j=1}^{r} \int_I \mathrm{clr}(\beta_j)(s,t), \mathrm{clr}(f_{ij})(s) ds + \mathrm{clr}(\varepsilon_i)(t)$$

[1] E.g., using the PLS approach: Beyaztas, U., Shang, H.L.: *On function-on-function regression: partial least squares approach*. Environmental and Ecological Statistics 27, 95–114, 2020

Density-on-density regression model

- The regression model can be defined for response PDFs $g_i(t)$ and predictor PDFs $f_{ij}(t), j = 1, \ldots, r$ as

$$\mathrm{clr}(g_i)(t) = \mathrm{clr}(\beta_0)(t) + \sum_{j=1}^{r} \int_I \mathrm{clr}(\beta_j)(s,t), \mathrm{clr}(f_{ij})(s) ds + \mathrm{clr}(\varepsilon_i)(t)$$

- For estimation of (functional) regression parameters the previous procedures for function-on-scalar and scalar-on-function regressions can be combined

[1] E.g., using the PLS approach: Beyaztas, U., Shang, H.L.: *On function-on-function regression: partial least squares approach*. Environmental and Ecological Statistics 27, 95–114, 2020

Density-on-density regression model

- The regression model can be defined for response PDFs $g_i(t)$ and predictor PDFs $f_{ij}(t), j = 1, \ldots, r$ as

$$\mathrm{clr}(g_i)(t) = \mathrm{clr}(\beta_0)(t) + \sum_{j=1}^{r} \int_I \mathrm{clr}(\beta_j)(s,t), \mathrm{clr}(f_{ij})(s) ds + \mathrm{clr}(\varepsilon_i)(t)$$

- For estimation of (functional) regression parameters the previous procedures for function-on-scalar and scalar-on-function regressions can be combined
- Adapt simply any of existing approaches[1] with the ZB-spline or any other relevant basis for representation of PDFs

[1] E.g., using the PLS approach: Beyaztas, U., Shang, H.L.: *On function-on-function regression: partial least squares approach*. Environmental and Ecological Statistics 27, 95–114, 2020

References

Filzmoser, P., Hron, K., Templ, M.: *Applied compositional data analysis*. Springer Series in Statistics. Springer, Cham, 2018.

Ramsay, J., Silverman, B.W.: *Functional data analysis*, 2nd ed. Springer, New York, 2005.

Talská, R., Menafoglio, A., Machalová, J., Hron, K., Fišerová, E.: Compositional regression with functional response. Computational Statistics and Data Analysis 123, 66–85, 2018.

Talská, R., Hron, K., Matys Grygar, T.: Compositional scalar-on-function regression with application to sediment particle size distributions. Mathematical Geosciences, DOI: 10.1007/s11004-021-09941-1, 2021.

Talská, R., Menafoglio, A., Hron, K., Egozcue, J.J., Palarea-Albaladejo, J.: *Weighting the domain of probability densities in functional data analysis*. Stat 9 (1), e283, 2020.

van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V.: *Bayes Hilbert spaces*. Australian & New Zealand Journal of Statistics 56, 171–194, 2014.