Statistical analysis of isocratic chromatographic data using Bayesian hierarchical modeling

Agnieszka Kamedulska, Łukasz Kubik, Paweł Wiczling

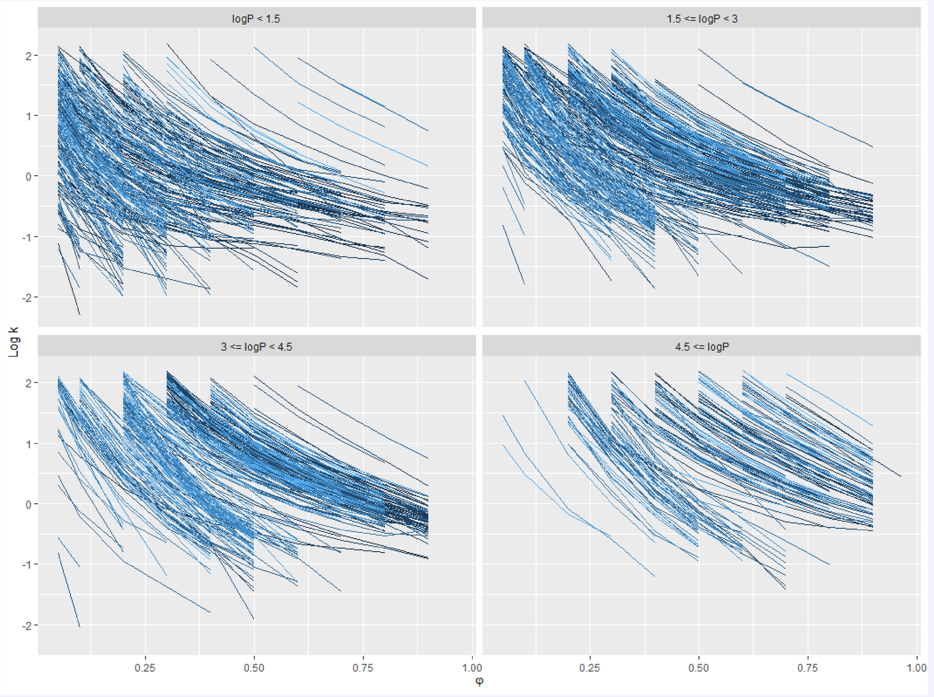
Department of Biopharmacy and Pharmacokinetics, Medical University of Gdańsk, M. Skłodowskiej-Curie 3a Street, 80-210 Gdańsk, Poland

Introduction

The chromatographic data is usually modeled considering one analyte at a time. It has certain limitations as no information is shared between analytes and consequently the model predictions poorly generalize to out-of-sample analytes. The methodology of full Bayesian inference with Markov Chain Monte Carlo sampling allows i) to incorporate prior knowledge about the likely values of model parameters, ii) to consider the between analyte variability and correlation between model parameters, iii) to explain the between analyte variability by available predictors, and iv) to share information across analytes. The latter is especially valuable when there is limited information in the data about certain model parameters. The results are obtained in the form of posterior probability distribution, that quantifies uncertainty about the model parameters and predictions. The posterior probability is also directly relevant for decision making.

Methodology

Raw Data



The dataset consists of isocratic reversed-phase high-performance liquid chromatography measurements of 1024 analytes using Agilent Eclipse Plus C18 stationary phase with 3.5 μm particles.
Data is publicly available, http://www.retentionprediction.org/hplc/database/ .

The statistical analysis was carried out in the Stan program coupled with R. Both programs are open-source.

Nonlinear Neue's model [1]

$$\boxed{\log k = \log k_w - \frac{S_1 \cdot \varphi}{1 + S_2 \cdot \varphi}}$$

$$S_1 = (\log k_w - \log k_a) \cdot \left(1 + 10^{\log S_{2A}}\right), \quad S_2 = 10^{\log S_{2A}}$$

Hierarchical model

$$\log k_{Obs} = f(R_i, \varphi_{i,j}) + \sigma$$
$$R_i = h(\theta, \log P_i) + \eta_{R,i}$$

R_i - individual (analytes-specific) parameters
σ - intra-analyte (residua) variability
θ - individual typical values
$\eta_{R,i}$ - inter-analyte variability

Bayesian inference

$$P\left(\theta | \log k_{i,j}, \varphi_{i,j}\right) \propto P\left(\log k_{i,j} | \theta, \varphi_{i,j}\right) \cdot P(\theta)$$

Stan code of Neue's model

```
functions{
  real hplcmodel(real fi, real logkw, real logka,
  real logSA){
    real logk;
    real S1;
    S1 = (logkw - logka)*(1+10^logS2A);
    logk = logkw - S1 * fi / (1 + 10^logS2A * fi);
    return logk;
  }
}
```

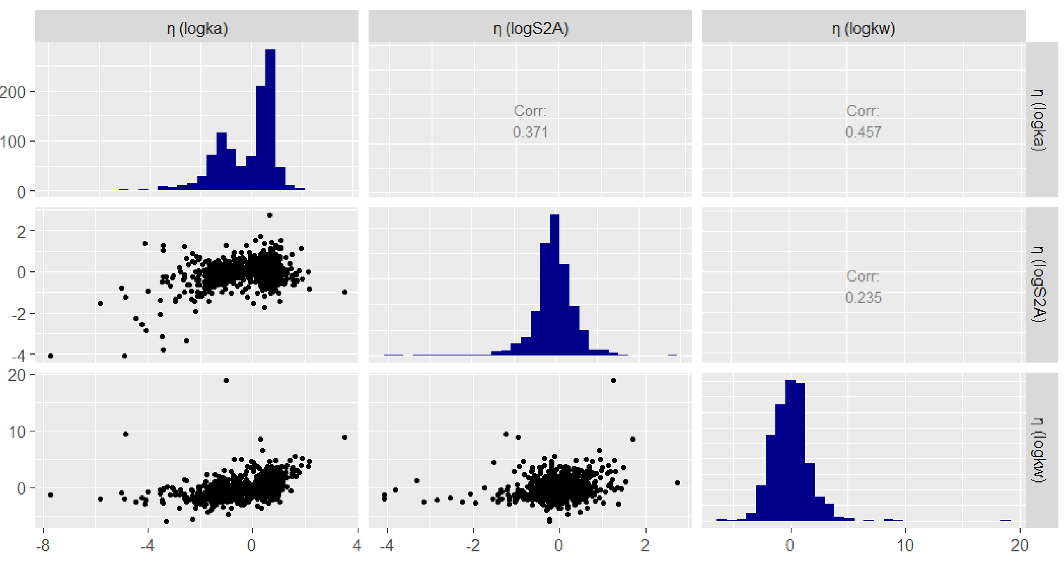
Initial model

$$\log k_{Obs} \sim \mathcal{N}\left(logk_{i,j}, \sigma\right)$$
$$\log k_{i,j} = \text{hplcmodel}(\varphi_{i,j}, \log k_{w,i}, \log k_{a,i}, \log S_{2A,i})$$

$$\begin{bmatrix} \log k_{w,i} \\ \log k_{a,i} \\ \log S_{2A,i} \end{bmatrix} = MST \begin{bmatrix} \nu, & \begin{matrix} \theta_{logk_{w,i}} + \beta_1 \cdot \log P_i, \\ \theta_{logk_{a,i}} + \beta_2 \cdot \log P_i, \\ \theta_{logS_{2A,i}}, \end{matrix} & \Omega \end{bmatrix}$$

Result

1. Our initial analysis indicated that the analytes form two clusters with different retention characteristics.



2. To describe this phenomenon was used a mixture model that assumes two data generating processes, each with their own set of parameters.

Improved model

$$\log k_{Obs} \sim \mathcal{N}\left(logk_{i,j}, \sigma\right)$$
$$\log k_{i,j} = \text{hplcmodel}(\varphi_{i,j}, \log k_{w,i}, \log k_{a,i}, \log S_{2A,i})$$

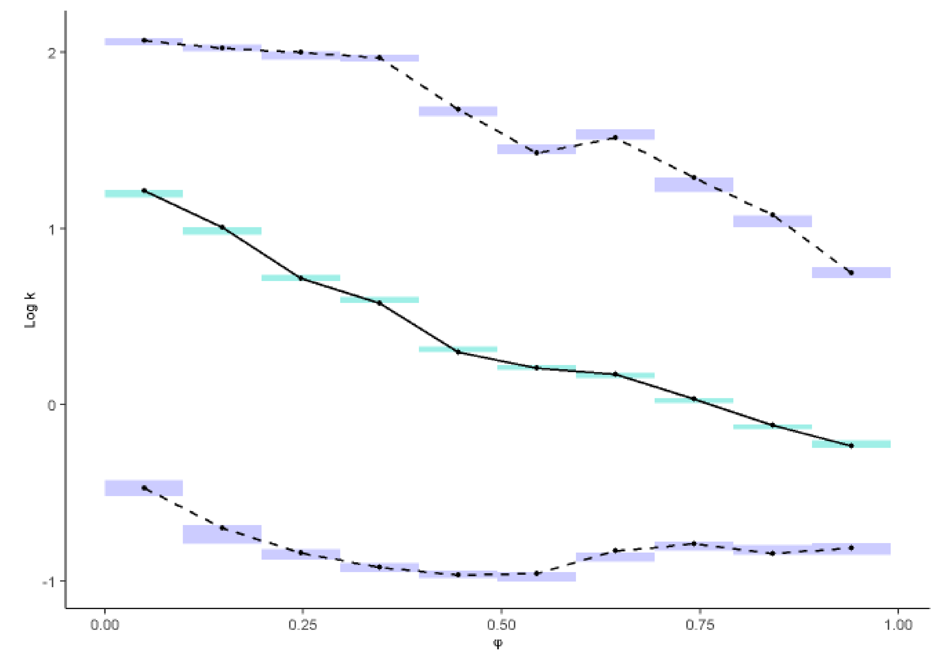
$$\begin{bmatrix} \log k_{w,i} \\ \log k_{a,i} \\ \log S_{2A,i} \end{bmatrix} = \text{Mixture}\left(\lambda, MVN\left(\begin{bmatrix} \log k_{w_1,i} \\ \log k_{a_1,i} \\ \log S_{2A,i} \end{bmatrix} \middle| \begin{bmatrix} \theta_{logk_{w_1,i}} + \beta_1 \cdot \log P_i, & \\ \theta_{logk_{a_1,i}} + \beta_3 \cdot \log P_i, & \Omega_1 \\ \theta_{logS_{2A,i}} + \beta_5 \cdot \log P_i, & \end{bmatrix}\right),\right.$$
$$\left. MVN\left(\begin{bmatrix} \log k_{w_2,i} \\ \log k_{a_2,i} \\ \log S_{2A,i} \end{bmatrix} \middle| \begin{bmatrix} \theta_{logk_{w_2,i}} + \beta_2 \cdot \log P_i, & \\ \theta_{logk_{a_2,i}} + \beta_4 \cdot \log P_i, & \Omega_2 \\ \theta_{logS_{2A,i}} + \beta_5 \cdot \log P_i, & \end{bmatrix}\right)\right).$$

3. Summary of marginal posterior distributions of model parameters

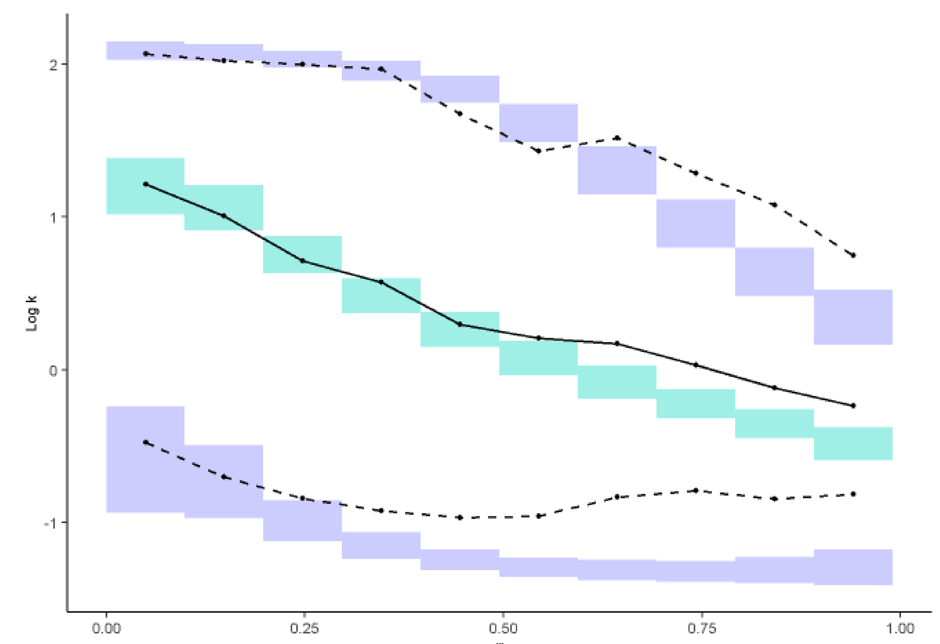
	mean	sd	2.5%	50%	97.5%	n_eff	Rhat
logkwHat[1]	2.90	0.14	2.62	2.90	3.19	3562	1
logkwHat[2]	4.45	0.18	4.09	4.45	4.79	3074	1
logkaHat[1]	-2.41	0.08	-2.56	-2.41	-2.26	4455	1
logkaHat[2]	-0.68	0.03	-0.74	-0.68	-0.62	3857	1
logS2AHat	0.45	0.01	0.42	0.45	0.47	4103	1
beta[1]	0.13	0.05	0.04	0.13	0.22	5410	1
beta[2]	0.34	0.06	0.21	0.34	0.46	4547	1
beta[3]	0.05	0.03	0.00	0.05	0.10	4977	1
beta[4]	0.06	0.01	0.04	0.06	0.08	5568	1
beta[5]	0.00	0.01	-0.01	0.00	0.01	4642	1
omega1[1]	1.83	0.07	1.69	1.83	1.98	3228	1
omega1[2]	1.14	0.04	1.05	1.13	1.23	3724	1
omega1[3]	0.31	0.01	0.29	0.31	0.34	2756	1
omega2[1]	2.09	0.10	1.90	2.09	2.31	2015	1
omega2[2]	0.38	0.02	0.35	0.38	0.41	4813	1
omega2[3]	0.20	0.01	0.18	0.20	0.23	2170	1
rho1[1,2]	0.39	0.05	0.29	0.39	0.47	3023	1
rho1[1,3]	0.07	0.06	-0.05	0.07	0.18	2910	1
rho1[2,3]	0.51	0.04	0.43	0.52	0.59	3760	1
rho2[1,2]	0.67	0.04	0.58	0.67	0.74	2023	1
rho2[1,3]	0.11	0.07	-0.04	0.11	0.25	1986	1
rho2[2,3]	-0.55	0.06	-0.66	-0.55	-0.42	1428	1
lambda	0.52	0.02	0.48	0.52	0.56	3579	1
sigma	0.04	0.00	0.04	0.04	0.04	1522	1

$$\boxed{\Omega = \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{bmatrix} \cdot \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & \rho_{2,2} & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & \rho_{3,3} \end{bmatrix} \cdot \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{bmatrix}}$$

4. Visual predictive check



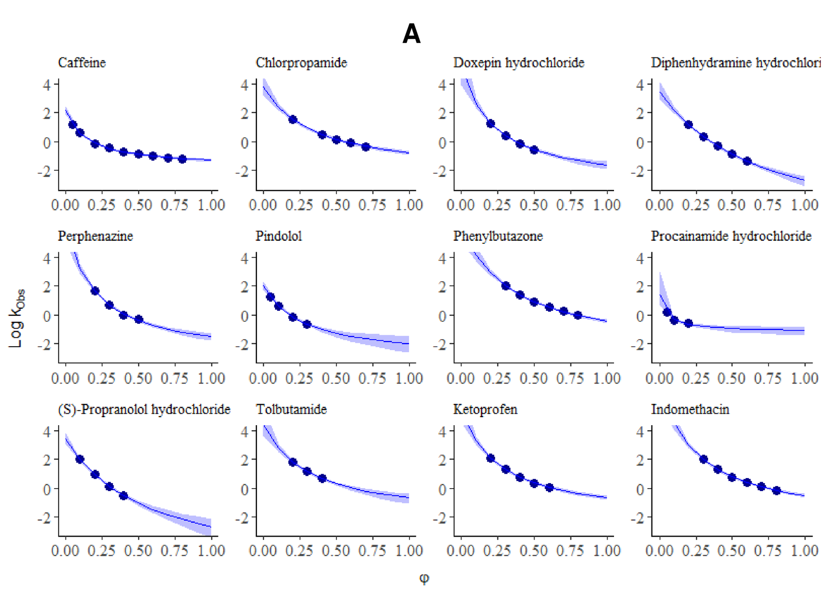
A

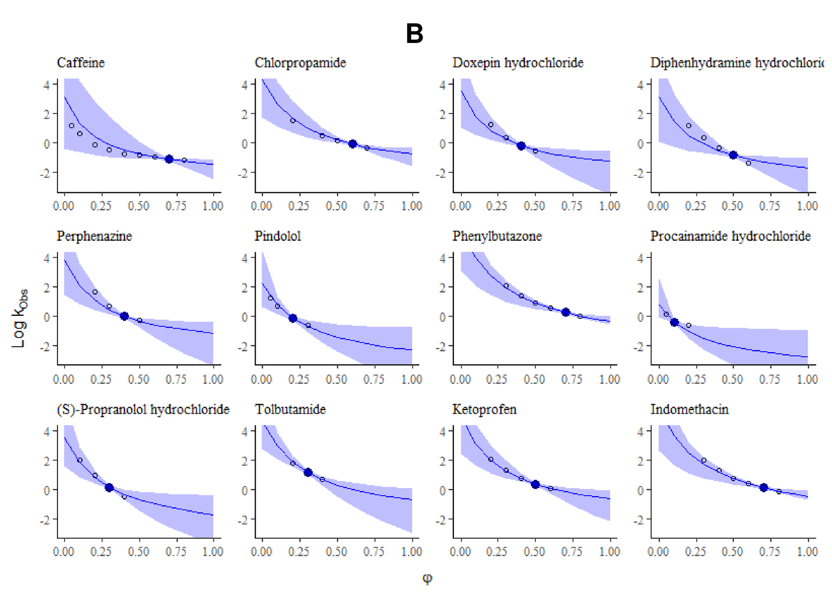


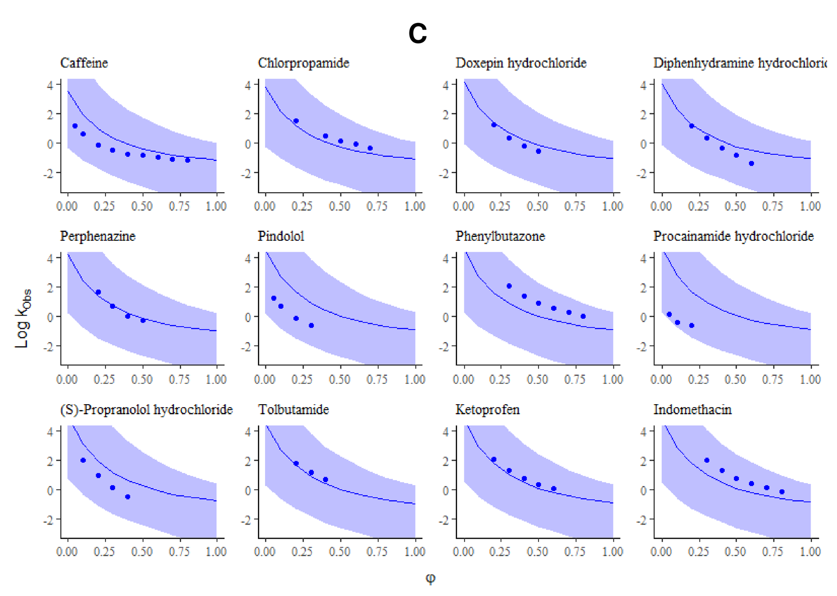
B

A. Predictions correspond to the future observations of the same analyte.
B. Predictions correspond to the future observations of a new analyte.

Conclusion







The proposed model gives insight into behavior of analytes in the chromatographic column and can be used to make predictions for structurally diverse set of analytes.

A. Prediction of future observations of the same analyte.
B. Prediction of future observations of the same analyte using reduced data.
C. Prediction of future observations of a new analyte.

References

[1] Neue, U. D., Phoebe, C. H., Tran, K., Cheng, Y.-F., Lu, Z. (2001). Dependence of reversed-phase retention of ionizable analytes on pH, concentration of organic solvent and silanol activity. Journal of Chromatography A, 925(1), 49–67.

[2] Kubik, Ł., Kaliszan, R., Wiczling, P. (2018). Analysis of Isocratic-Chromatographic-Retention Data using Bayesian Multilevel Modeling. Analytical Chemistry, 90(22), 13670–13679.