Rethinking Calibration as A Statistical Estimation Problem to Improve Measurement Accuracy – Supporting Materials

Sabrina Jaffe[1†], Duane Gossiaux[2], Reagan M. Errera[2], Emanuela Gionfriddo[3], Hongjun Wang[4], Curtis J. Richardson[4], Nipunika H. Godage[3] and Song S. Qian[1*†]

[1]Department of Environmental Sciences, The University of Toledo, 2802 Bancroft Street, Toledo, 43606, Ohio, USA.
[2]Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, 4840 South State Road, Ann Arbor, 48108, Michigan, USA.
[3]Department of Chemistry, The University of Toledo, 2802 Bancroft Street, Toledo, 43606, Ohio, USA.
[4]Nicholas School of the Environment, Duke University, 9 Circuit Drive, Durham, 27710, North Carolina, USA.

*Corresponding author(s). E-mail(s): song.qian@utoledo.edu;
Contributing authors: sabrina.jaffe@rockets.utoledo.edu; duane.gossiaux@noaa.gov; reagan.errera@noaa.gov; emanuela.giofriddo@utoledo.edu; hjwang78@duke.edu; curtr@duke.edu; nhirimu@rocketes.utoledo.edu;
†These authors contributed equally to this work.

In this document, we provide additional examples of using the Bayesian hierarchical modeling (BHM) approach for improving the measurement

accuracy of calibration methods. The BHM approach allows effective sharing of information within a single calibration test and across multiple tests and has been shown to reduce estimation and prediction uncertainty in many statistical estimation problems. In this document, we use three additional examples to illustrate (1) the computational details of applying BHM to calibration problems, (2) how sample size used to develop the calibration curve affects estimation accuracy, and (3) a comparison of estimation accuracy of three different calibration problems.

1 Additional Examples

The three examples used in this document include linear and nonlinear calibrations as well as a problem of estimating matrix effect.

1.1 Toledo Water Crisis

The Toledo water crisis example illustrates a nonlinear calibration curve problem similar to the data example in the main article. The data used here are cyanobacterial toxin microcystin (MC) concentration measurements from City of Toledo's Drinking Water Plant during the "Toledo Water Crisis" of 1-3 August 2014. During the crisis a "Do-Not-Drink" order was issued due to one MC measurement (2.7 μg/L) exceeding the Ohio drinking water quality standard (1 μg/L) on 1 August 2014. The enzyme-linked immunosorbent assay (ELISA) kit from Eurofins-Abraxis was used. It uses a sigmoid function (the four parameter logistic function) to describe the relationship between the instrumental response (a measure of color change) and MC concentrations (equation (1) in the paper). This example has data from six ELISA tests so that we can illustrate the computational details and make direct comparisons between the BHM estimates and the classical calibration estimates.

Following the US EPA's standard protocol, the standard commercial kit from Abraxis uses six standard solutions with known MC values ranging from 0 to 5.55 μg/L with two replicates each. The resulting 12 data points are either (1) directly used for fitting the calibration curve ($n = 12$) or (2) transformed by averaging the response replicate observations for each non-zero standard solution and dividing them by the average of the zero solution replicates to generate 5 "relative" responses for fitting the calibration curve ($n = 5$).

We use this example to illustrate the details of the results, including the effect of sample size on estimation accuracy and the benefit of pooling information using the Bayesian hierarchical model within a test (multiple water samples with unknown concentrations) and across multiple tests (multiple standard curves).

1.2 Ortho phosphate in the Stream and Wetland Assessment Management Park (SWAMP) at Duke University

Measuring phosphorus in water is an important and routine task in nearly all environmental monitoring and public health fields. The standard method used for analyzing phosphorus concentrations is based on the reaction of ammonium molybdate and potassium antimonyl tartrate in acid medium with orthophosphate (PO_4). The resulting phosphomolybdic acid is then reduced to intensely colored molybdenum blue by ascorbic acid (Ascorbic Acid Method, SM 4500-PE). The intensity of the color is proportional to the PO_4 concentration in the sample. Six standard solutions with known PO_4 concentrations and two blank samples are used to develop the linear calibration curve (i.e., $y = \beta_0 + \beta_1 x + \epsilon$).

The data we used were from the Stream and Wetland Assessment Management Park (SWAMP) located in the upper section of the Sandy Creek stream and floodplain in Duke Forest in Durham, North Carolina, USA. Duke University Wetland Center has been monitoring the water quality monthly in the SWAMP since mid-1990s. In this study, we focus on PO_4, as all other forms of phosphorus are measured by the ascorbic acid method after additional processes. A small subset of calibration tests for PO_4 in a 5-year period (2017-2021) was selected for this comparison study.

The PO_4 measurement protocol does not require replicate samples. As a result, measured responses from calibration samples contribute no additional information about the quality of the fitted calibration curve (residual variance) [1].

1.3 Measuring Xenobiotics in Human Blood

Concentrations of xenobiotics in human blood are often used to measure human exposure to organic pollutants, such as pesticides and pharmaceuticals. Measuring xenobiotic concentrations can be expensive and complicated, and ultimately, the concentration in a blood sample is estimated using a calibration method. A recent study explored the use of biocompatible solid phase microextraction (SPME) coupled with liquid chromatography mass spectrometry (LCMS) for xenobiotic analysis in plasma [2].

This study emphasized the feasibility of using non-human plasma-based standard solutions to reduce costs. Due to differential binding between target analytes and different plasmas, calibration curve coefficients are known to vary from plasma to plasma, an effect known as the matrix effect. When the matrix effect between human plasma and non-human plasma (e.g., bovine plasma) is stable, bovine plasma-based standard solutions can be used as

substitutes for human plasma standard solutions, which can greatly reduce measurement costs. Using calibration data for acetochlor (a widely used herbicide and known human carcinogen and thyroid disruptor) as an example, we illustrate the use of BHM for improving estimation accuracy (evaluated using quality control samples) and establishing matrix effects.

The example includes (1) the determination of the calibration curve format (a log-log linear model), (2) a comparison of the estimation accuracy using three sets of quality control samples, and (3) the estimated matrix effects (differences in fitted calibration curve coefficients, intercept and slope, among five mediums: human, bovine, rabbit, and rat plasma, as well as phosphate-buffered saline (PBS) buffer).

2 The Role of Sample Size

In a regression problem, we assume that the response variable is a normal random variable with the mean defined as a function of the predictor variable (the mean function) and a constant residual variance. The model-fitting process quantifies both the mean function coefficients and the residual variance based on available data. The estimation uncertainty is largely determined by the amount of information in the data, measured by the degrees of freedom (typically, the sample size minus the number of model coefficients to be estimated). The regression model coefficient uncertainty is a function of the model's residual variance. When a regression model's degrees of freedom is four or fewer, the uncertainty of the estimated residual variance is undefined, meaning that we cannot simultaneously quantify the model coefficients and the residual variance with confidence. In other words, a regression model with a degree of freedom of 4 or fewer cannot be reliably used for prediction, although a regression model fit with a small sample size

is more likely to have an impressive goodness-of-fit statistics. In the extreme case of fitting a linear model with two data points, we always obtain a perfect R^2 value of 1. However, with a degrees of freedom of 0, we have no information about the model's predictive accuracy (i.e., the resulting model is unreliable for prediction).

We illustrate the role of sample size by fitting the nonlinear four-parameter logistic function (equation (1) in the paper) using data from the first ELISA test from Toledo Water Crisis. We fit the regression model using a Bayesian regression method (with non-informative priors) and illustrate the strong correlation between model coefficients and the residual variance. We first use $n = 5$ (using the relative response) and then use $n = 12$.

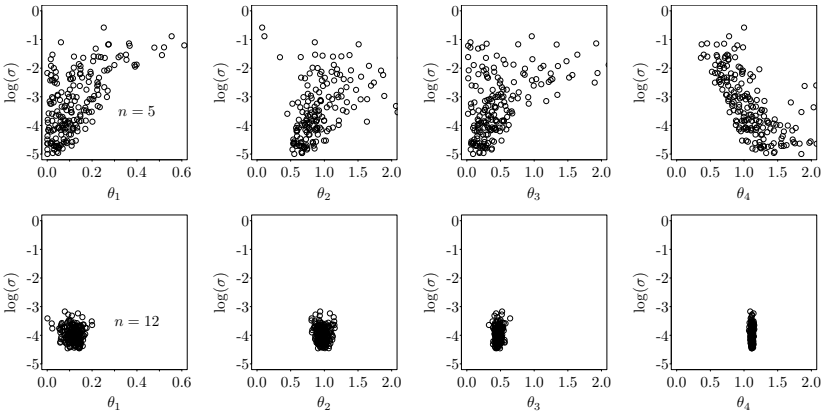

Fig. 1 Uncertainty in the fitted standard curve (four-parameter logistic function) coefficients (θ_1–θ_4) and residual variance are represented by their joint posterior distribution, expressed using bivariate scatter plots of the log standard deviation against coefficients using random samples from the joint distribution. The standard curve fit with mean relative response ($n = 5$) has a much higher uncertainty (top row) than the curve fit with untransformed raw data ($n = 12$, bottom row).

When fitting the model using relative responses ($n = 5$), the marginal posterior distributions of θ_i and σ^2 are funnel-shaped (Figure 1, top row): the range of θ_i is highly dependent on the value of σ^2. With $n = 5$ (or 1 degree of freedom), we have no information to adequately quantify σ^2. As a result, we

have no information about the model coefficients. When fitting the same model with $n = 12$ (8 degrees of freedom), we are able to better quantify σ^2 (Figure 1, bottom row). As a result, the estimated θ_i are more stable. In all six tests, the apparent R^2-values for the models fit with $n = 5$ are larger than the same for models fit with $n = 12$. The effects of sample size on the estimated concentration values are almost never mentioned in practice. This is because the estimation uncertainty is usually evaluated using the sampling distribution of an estimator. For the calibration problem, the sampling distribution of the inverse-function estimator is only available for the linear calibration problem. However, the variance of the sampling distribution, a Cauchy distribution, is undefined. Using the Bayesian estimation method, we can evaluate the estimation uncertainty using the posterior distribution of the unknown analyte concentration. The variances of the estimated MC concentrations (expressed as the standard deviations of posterior distributions) using the inverse-function method is orders of magnitude higher when using the standard curve fitted with $n = 5$ than the same using $n = 12$ (Figure 2). In other words, the R^2-value should not be used as the sole criterion of a regression model. Additionally, fitting a calibration curve using average responses of standard solution replicates is inadvisable.

3 Hierarchical Modeling Within and Across Multiple Calibration Curves

While adequate sample size is important in fitting a regression model, the calibration problem is about the estimation of the "missing" concentration data. As such we should focus on the estimator of the analyte concentrations. When we use the conventional inverse-function-based method or the Bayesian method with flat priors, we emphasize the unbiased nature of the

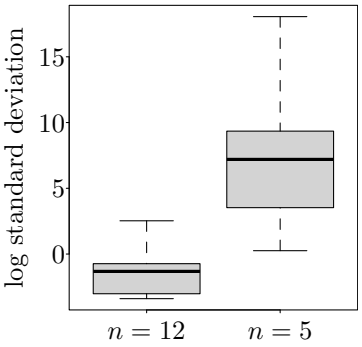

Fig. 2 Estimation uncertainty (expressed as standard deviation of the estimated MC concentrations) using the inverse-function method is a function of the sample sized used for fitting the standard curve. The boxplots represent the 80 estimation log standard deviations from the 40 water samples (each with a replicate) in one of the 6 ELISA tests conducted during the Toledo Water Crisis.

estimator. Because we almost always measure multiple analyte concentrations at the same time, we can drastically improve estimation accuracy by inducing biases in the estimator [3]. In modern statistics, we can achieve such improvement by using the Bayesian hierarchical modeling (BHM) approach [4]. We apply the BHM approach on two levels. One is within each test, by considering all unknown analyte concentrations as exchangeable (i.e., they are different but otherwise uncertain) and imposing a common prior distribution. The other is across all tests by considering the standard curve coefficients as exchangeable and imposing a common prior.

In the six Toledo Water Crisis ELISA tests, we have 11 to 40 water samples for each test. Among these water samples is a quality assurance (QA) sample with known MC concentration of 0.75 μg/L. We compare the estimated MC concentrations of the QA sample from using the inverse-function method, the Bayesian method with flat prior, and the BHM approach (Figure 3). The estimation uncertainty of the inverse-function based method is estimated based on a Monte Carlo simulation where random samples of the standard

curve coefficients were used to represent the model uncertainty. Some random samples of the coefficients result in non-real value solutions (log of negative values) and must be excluded. As a result, uncertainty of the inverse-function method in Figure 3 is underestimated. We note that the inverse-function method estimated QA sample concentrations vary more from test to test compared to the estimates from the BHM estimates. The large among test variation is represented in the absolute difference between the estimated and the true concentrations (absolute error, Figure 4).

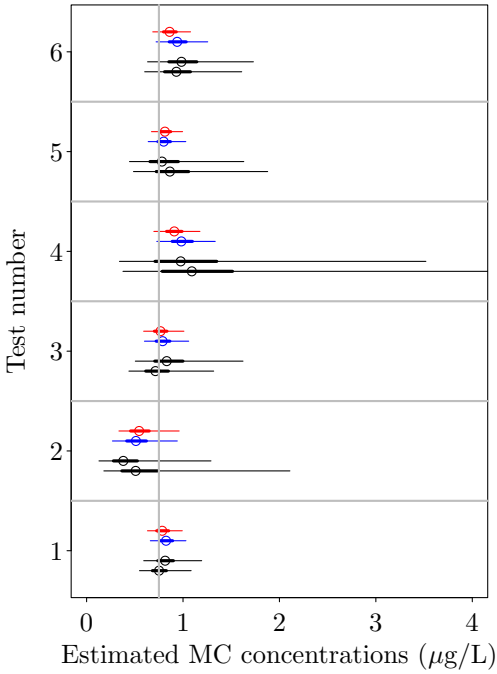

Fig. 3 The estimated QA sample concentrations using inverse-function method (black), BHM within test (blue), and BHM within and across test (red) are compared. The open circles are the estimated means, the thick and thin lines are the 50% and 95% credible intervals, respectively. The true QA sample concentration (0.75 μg/L) is shown by the gray vertical line.

For the two linear calibration examples from SWAMP monitoring data, we see the same overall pattern of reduced bias when using BHM (Figures 4).

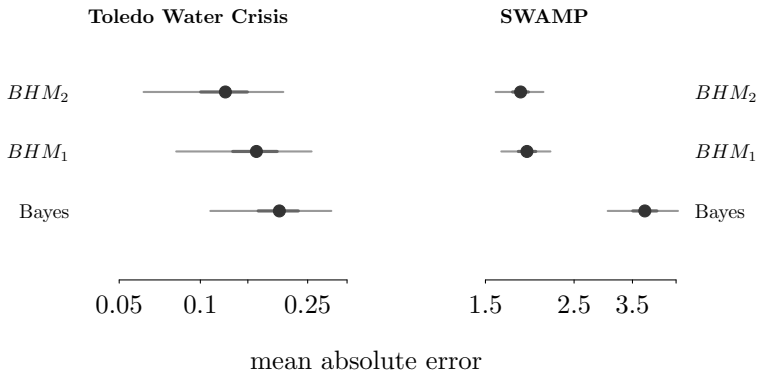

Fig. 4 ELISA estimation biases (mean absolute errors) of the inverse-function method (MLE_{12}), Bayesian method with flat prior (Bayes), BHM within test (BHM_1), and BHM within and across tests (BHM_2) are compared (to the known control sample concentrations: 0.75 μg/L [microcystins] for the Toledo water crisis data and 50 μg/L [total phosphorus])

As with the ELISA data, BHM within a test resulted in a substantial reduction in bias. The addition of cross-test BHM produced marginal improvement. For the mass-spectromertry data, the reduction of bias is a combined result of within test BHM and the BHM model across calibration curves fit to data from 6 different mediums (Figure 6).

4 Matrix Effect

In the xenobiotics measurement example, [5] first explored whether the linear calibration curve should be fit in the original scale or the log-log scale. Graphical display (scatter plots) of the measured response and known concentration values indicate that the calibration curve is better defined in the log-log scale (Figure 5). This result can be justified by the Beer-Lambert law which states that the response would increase proportionally as the concentration increases. A log-log linear relationship of

$\log(y) = \beta_0 + \beta_1 \log(x) + \epsilon$ reflects this assumption well. That is, the regression indicates that for every 1% increase in x, y would increase by β_1% [6].

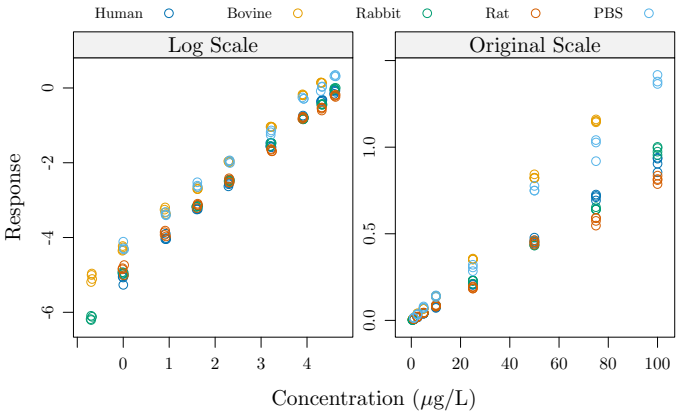

Fig. 5 Scatter plot of the SPME-LCMS data for acetochlor (the response, y-axis, is the area of the analyte). In the log-log scale, the difference among different mediums lies in the intercept and the slopes are nearly the same. The response is analyte area and the concentration is in μg/L.

We use data for acetochlor, a chloroacetanilide type herbicide, to illustrate how the use of BHM improves estimation accuracy.

In the experiment reported by Godage et al [5], there are three groups of quality assessment samples spanning across the concentration spectrum of the target analyte concentration to represent low, median, and high levels. As in the other two examples, we analyzed the data using both the conventional inverse-estimation method and the BHM method. In both methods, we used predesignated standard solutions and treat quality control sample concentration values as unknown to evaluate estimation accuracy. There are six mediums (human, bovine, rabbit, and rate plasma, and PBS solution) for each xenobiotic.

For the inverse-estimation method, we fit each of the six calibration curves independently and calculate the "unknown" concentrations using the

respective resulting linear functions. The BHM method combines data from all six test mediums to estimate the quality control sample concentrations and the six calibration curve coefficients together.

The estimation bias, in this case, is estimated using the absolute difference between the estimated log concentration and the true log concentration because the calibration curve is fit at the log concentration scale, indicating that the bias is likely proportional to the true concentration. The estimated log bias is converted to percent. Similarly, we conclude that using BHM is beneficial (Figure 6).

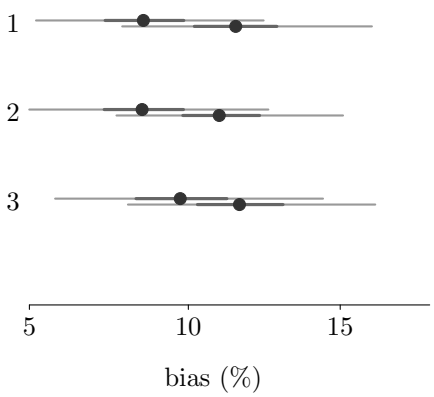

Fig. 6 Estimation accuracy, measured by the relative biases are compared between the inverse-estimation method (bottom line of each pair) and the BHM method (top line of each pair). The three pairs of lines represent (from top to bottom) quality control sample concentrations at the low (2.5-5 μg/L), median (10-25 μg/L), and high (50-75 μg/L) groups. A bias of 5%, for example, represents that, in the concentration scale, the estimated concentration is within 5% of the true concentration.

The advantages of BHM (or multilevel model) for estimating multiple (related) regression model coefficients is well-documented in the literature (e.g., [7, 8]). In our analysis we find that the estimated regression model

coefficients for the six mediums are independent of each other with low estimation uncertainty (Figure 7). As a result, using animal plasma (or PBS) as standard solution medium is feasible.

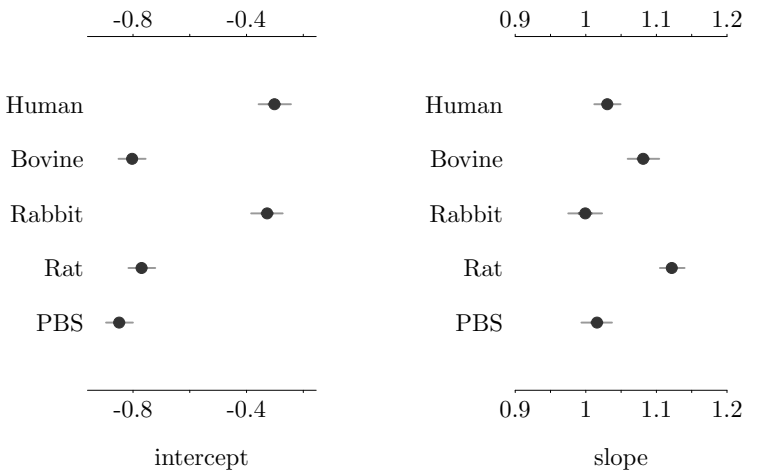

Fig. 7 BHM estimated calibration curve coefficients are of low estimation uncertainty, indicating that using non-human plasma as substitution for human plasma standard solution is feasible.

4.1 Summary

In all three examples, the BHM method consistently outperform the conventional inverse-estimation method in terms of measurement accuracy. We gauge the estimation accuracy by calculating the absolute difference between the estimated concentration values and the known (or true) concentration values (the bias). This method is preferred as bias serves as the pertinent measure for calibration problems.

References

[1] Hunter, W.G., Lamboy, W.F.: A Bayesian analysis of the linear calibration

problem. Technometrics **23**(4), 323–328 (1981)

[2] Godage, N.H., Qian, S.S., Gudjoe, E., Gionfriddo, E.: Enhancing quantitative analysis of xenobiotics in plasma through cross-matrix calibration and Bayesian hierarchical modeling. ACS Measurement Science Au **under review** (2023)

[3] Efron, B.: Biased versus unbiased estimation. Advances in Mathematics **16**, 259–277 (1975)

[4] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. CRC Press, Boca Raton, Florida (2014)

[5] Godage, N.H., QIan, S.S., Cudjoe, E., Gionfriddo, E.: Enhance quantitative analysis of Xenobiotics in blood plasma through cross-matrix calibration and Bayesian hierarchical modeling. ACS Measurement Science Au **4**(1), 127–135 (2024)

[6] Qian, S.S.: Environmental and Ecological Statistics with R, 2nd edn. Chapman and Hall/CRC Press, Boca Raton, Florida (2016)

[7] Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York (2007)

[8] Qian, S.S., Cuffney, T.F., Alameddine, I., McMahon, G., Reckhow, K.H.: On the application of multilevel modeling in environmental and ecological studies. Ecology **91**, 355–361 (2010)