Supporting Information -- Statistical Background

With the limited space of a research paper, we are unable to fully explore the statistical basis of our work. In this supporting document, we detail our motivation in conducting the research and explain the statistics behind our thought process. We first explain why we stated that calibration is a statistical estimation problem. We then explain some relevant regression analysis results (specifically, the fitted versus predictive uncertainty) and why numerical approximation (Monte Carlo simulation) was used in our study for quantifying regression prediction uncertainty. Finally, we refer readers with several important references regarding a class of biased estimators that can improve calibration estimation accuracy.

Calibration is a statistical estimation problem

Although regression is the basic statistical procedure in a calibration problem, the current applications often treat the fitted standard curve as a known relationship between the instrumental response and the analyte concentration. As a result, the standard curve is typically fit with only the data from the standard solutions, and the unknown analyte concentration is "back-calculated" using the inverse function of the fitted standard curve.

However, when fitting the standard curve using linear or nonlinear regression, we assume that the instrumental response were measured with error and the concentration values are measured accurately. Likewise, we assume regression model parameters are constants but unknown. Fitting the regression model produces an estimator of these parameters and the estimator is a random variable because the instrumental responses are random. I now use a linear regression for simplicity. For a linear regression problem (i.e., $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, with a parameter vector of $\beta = \{\beta_0, \beta_1\}$ and $i = 1, \cdots, n$ are the n standard solutions), the estimator is a bivariate random variable follows a bi-variate normal distribution with the mean vector of $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the variance-covariance matrix of $\sigma^2 \left(X^T X\right)^{-1}$, where X is the design matrix. From this sampling distribution we derive the marginal distributions of the intercept and the slope (which are both normal distributions). If we replace σ^2 with the estimated residual variance, the marginal distributions are student-t distributions.

In a calibration problem, we have a number of test samples with unknown analyte concentrations. We measure the instrumental responses of the standard, as well as the test samples together. Conceptually, concentrations of test samples are constants, not random. Suppose that we have only one test sample with an instrumental response of y_0 and an unknown (but constant) concentration x_0. The problem is not just derive an estimator for the regression parameter vector, but also an estimator for the unknown concentration x_0. In other

words, we now treat x_0 as an unknown parameter of the problem. When using the maximum likelihood (ML) approach, the likelihood function of the combined data has $n+1$ terms (equation (2) of the manuscript), a function of both β and x_0. In statistics, this is the missing data problem (Little and Rubin 2019).

The point estimates (MLE) of β is the same as fitting the regression without x_0, and the MLE of x_0 is the inverse formula $\hat{x}_0 = (y_0 - \hat{\beta}_0)/\hat{\beta}_1$. The sampling distribution of x_0 is, however, not the normal distribution. Because the estimator is a ratio of two normal random variables, the sampling distribution of $\hat{x}_0$ is a Cauchy distribution, which doesn't have a moment-generating function, meaning that its mean and variance is analytically undefined. Likewise, the variances of $\hat{\beta}$ also changed because of the inclusion of y_0. For this reason, we declared that "quantifying the uncertainty of the estimated concentration is contentious and almost never reported in practice" (quoting Osborne, 1991). We interpret the quote as that "*properly estimated* uncertainty is never reported."

In our manuscript, we used two Monte Carlo simulation approaches to circumvent the analytical difficulties in quantifying the estimation uncertainty.

Monte Carlo simulation for quantifying fitted and predictive uncertainty

Understanding regression models using simulation

Monte Carlo simulation was first used to evaluate the performance of the inverse estimation approach, commonly used in the literature and in practice. Because the conventional practice is to fit the standard curve before calculating the unknown concentration, the unknown test sample concentration and its instrumental response were not included in the standard curve fitting. The relevant model uncertainty is the predictive uncertainty -- the uncertainty of predicting a future data point not used in model fitting. For a linear standard curve, the fitted standard error (when predicting the average of $\hat{y}_0$ given x_0) is

$$sefit(\hat{y}_0|x_0) = \hat{\sigma}\left[\frac{1}{n} + \frac{x_0 - \bar{x}}{SXX}\right]^{1/2},$$

which is the standard error of $\hat{y}_0$ given the known concentration of x_0, the estimation uncertainty of the mean of instrumental response. When predicting an individual observation of y_0, the prediction uncertainty is

$$sepred(\hat{y}_0|x_0) = \hat{\sigma}\left[1 + \frac{1}{n} + \frac{x_0 - \bar{x}}{SXX}\right]^{1/2}.$$

These two equations (Section 5.5, Qian, 2016) are obviously irrelevant to a calibration problem as we are interested in the unknown x_0. The point here is the difference between the fitted uncertainty and the prediction uncertainty, and the prediction uncertainty is the relevant one in a calibration problem. What we routinely used from a linear regression output (e.g., using R function `lm()`) are fitted uncertainty, not the prediction uncertainty. The prediction uncertainty, for example in R, is calculated in the function `predict()`. The prediction uncertainty $sepred$ for a nonlinear regression problem is case-specific. No general purpose formula is available.

To evaluate the prediction uncertainty of a calibration problem (estimating the uncertainty of $\hat{x}_0$ given an observed y_0 and the fitted standard curve), we used Monte Carlo simulation as described in Section 9.2 of Qian (2016) (we use $\tilde{\theta}$ to denote a random sample of θ):

1. We first simulate the uncertainty in the estimated residual variance. The estimated residual variance has a sampling distribution described by

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p).$$

   From the fitted standard curve, we have $\hat{\sigma}^2$. We then draw a random number ($\tilde{c}_i$) from the $\chi^2(n-p)$ distribution and calculate $\tilde{\sigma}_i^2 = (n-p)\hat{\sigma}^2/\tilde{c}_i$, which yields one random sample of σ^2.

2. Substitute the random sample $\tilde{\sigma}_i^2$ to the sampling distribution of $\hat{\beta}$, we can draw a sample of standard curve parameters.

3. Draw a random number $\tilde{\epsilon}_i$ from $N(0, \tilde{\sigma}_i^2)$ and calculate $\tilde{x}_0^i = (y_0 - \tilde{\beta}_0^i + \tilde{\epsilon}_i)/\tilde{\beta}_1^i$

4. Repeat 1-3 many times ($i = 1, \cdots, k$), we obtain many random samples $\tilde{x}_0^i$ to quantify the uncertainty of $\hat{x}_0$.

This Monte Carlo simulation allows us to numerically calculate the estimation uncertainty of $\hat{x}_0$ using $\tilde{x}_0^i, i = 1, \cdots, k$. However, because the instrumental response values of test samples were not considered in fitting the standard curve, the resulting curve may be problematic for some testing samples, especially for the nonlinear standard curve in our manuscript. When the instrumental response value fell out of the range set by the estimated bounds (θ_1 and θ_4 in equation (7)), the estimate fails and dropped from the simulation process. Consequently, uncertainty measures are underestimated when using the remaining $\tilde{x}_0^i$'s. Therefore, the standard curve is better fitted including instrumental responses from both the standard solutions and test samples. This is the task of the second Monte Carlo simulation.

Simulation as a numeric algorithm in Bayesian computation

The second Monte Carlo simulation was used to estimate all unknown quantities (both regress parameters and unknown test sample concentrations) based on the likelihood function (equation (2) in the manuscript). We used the Markov chain Monte Carlo (MCMC) simulation method commonly used in modern Bayesian computation. The details of MCMC is not important for the current manuscript. We documented the computational details in Chapter 2 of Qian et al (2022).

By using MCMC to directly tackle the joint likelihood function, we leverage the instrumental responses of the test samples to avoid ill-fitted standard curve that could lead to non-existing concentration estimates for some test samples.

When using MCMC for one calibration test, we used non-informative priors for all unknown quantities and draw random samples directly from their joint posterior distribution. Using a Bayesian approach gives us the flexibility to incorporate some important statistical theories on optimal estimation.

Biased versus unbiased estimators

The least squares method, or more generally, the maximum likelihood estimator (MLE) is unbiased and most efficient (or least variable *among all unbiased* estimators. However, in the late 1950s and early 1960s, Stain's paradox shook the statistics world by proving that an estimator (the James-Stein estimator) outperforms MLE when there are multiple means to be estimated. The paradox suggests that when we have only one variable to estimate (e.g, one unknown concentration of a test sample), MLE is the best estimator; when there are more than three variables to estimate (e.g., more than three unknown concentrations from test samples), the James-Stein estimator would result in better (more accurate) estimates. After 10 years of research, Efron and Morris (1977) detailed Stein's paradox in a *Scientific American* paper. Since then, many statisticians connected the James-Stein estimator to empirical Bayes method and explained the improved accuracy as a result of leveraging information among multiple samples. Recently, Gelman et al (2022) showed that the Bayesian hierarchical model (BHM) is a generalization of the James-Stein estimator. They further recognized that the Bayesian hierarchical model as one of the most important statistical discoveries in the past 50 years. Although the effect of improving overall accuracy is well recognized in statistics, these statistical results, have not yet penetrated in many applied science fields, largely due to the increased computational burden of combining data from multiple sources. In our paper, we showed that BHM can be used to leverage two sources of information: multiple test samples with unknown concentrations and multiple calibration tests in a typical chemistry lab.

Our study is an early attempt of introducing BHM into quantitative chemistry. The current manuscript demonstrated the improvement of BHM. We also posted our sister paper in my GitHub repository to demonstrate the practical feasibility of implement BHM for routine application.