Performance of the NCI method of dietary intakes in small sample sizes

Ljubomir Miljacic

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2019

Committee:
Kenneth Rice
Adam Szpiro

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Performance of the NCI method of dietary intakes
in small sample sizes

Ljubomir Miljacic

Chair of the Supervisory Committee:

Kenneth Rice

Department of Biostatistics

Modeling the population distribution of usual intake of episodically-consumed foods, primarily using 24-hour recall, is a challenging problem. The NCI method is the most recent major step in the evolution of specialized techniques that address it. This thesis is a contribution to the effort of spreading the usage of the method to a wider research community. We implement the basic method in the R environment, increasing the number of researchers and practitioners in the position to use the method. We observe the performance of the method in small samples using two different datasets, and in three potentially-useful scenarios. From these observations we draw recommendations for the minimal sample composition and size.

Contents

1 Introduction and motivation

For several decades there has been a sustained effort among researchers to construct statistical models that would accurately estimate the distribution of the dietary intake of foods in whole populations, for the purpose of informing public policy decisions, evaluating compliance with dietary recommendations, and assessing health risks [1, 2].

Food intake recommendations that aim to promote nutrient adequacy and public health are often expressed in terms of daily values [3]. As nutrients can be stored in the body and dietary intake varies on a daily basis, these targets are not expected to be achieved every day. Thus a key concept in assessing adherence to such recommendations is *usual intake*, defined formally as long-run average intake [2].

Two widely-used dietary data collection instruments are food frequency questionnaires (FFQs) and 24-hour recalls. FFQs are designed to measure long-term behaviour and are relatively inexpensive to collect, however they are hampered by the inability of individuals to accurately estimate their average food intake retrospectively, particularly over a long period of time. In fact using FFQs alone introduces substantial biases into usual intake estimates [2, 4–6]. In contrast, 24-hour recalls focus on a single day and thus greatly reduce systematic biases present in FFQs, while still providing rich detail about the types and amounts of foods consumed. However, they too have their drawbacks. Individual diets vary greatly from one day to another. Additionally, they still retain some measurement errors, bias, and often have to rely on standardized recipes and food composition databases and so cannot easily replicate between-person heterogeneity in nutrient intake. All these factors contribute to a considerable inaccuracy, making the measured intake on a single day a poor estimate of long-term mean individual intake.

Early attempts to compensate for this inaccuracy by averaging several (2–7) 24-hour recalls per individual were deemed unsatisfactory due to high respondent burden, and did not deliver high quality results [2]. Averages over small number of days inadequately represent individual usual intakes, thus an empirical distribution of these averages would be an inaccurate estimate of the population distribution.

This is illustrated in Figure 1, which compares several estimations of the distribution of usual intake of all fruits and vegetables in the US population

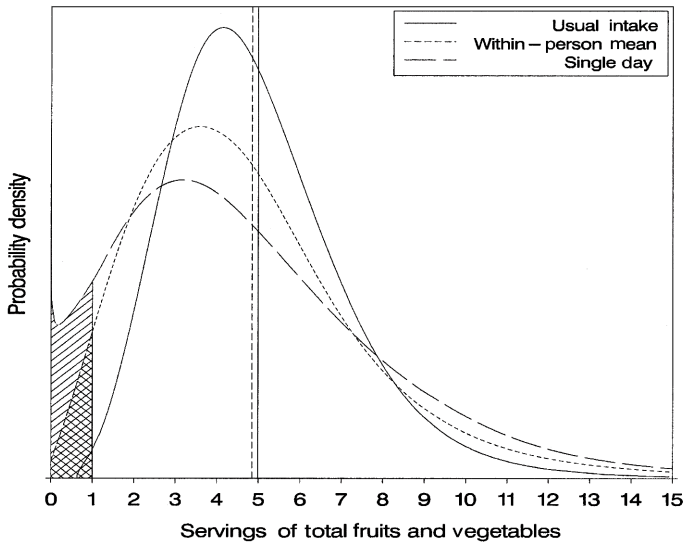


Figure 1: Comparing several estimations of the distribution of usual intake of all fruits and vegetables in the US population for the years 1994-1996. The broken line is based on a single 24-hour recall per person, the dashed line estimates within-person intake as average of two recalls, and the solid line uses a statistical model based on two recalls per person. The dashed vertical line marks the mean of the dashed line distribution, while the solid vertical line marks the mean of the other two distributions. Taken from Dodd *et al* (2006) [2].

for the years 1994-1996, where all the different approaches are based on 24-hour individual recalls. The Figure is taken from [2]. The broken line is based on a single recall per person, the dashed line estimates within-person intake as average of two recalls, and the solid line uses a more accurate statistical model based on two recalls per person. The dashed vertical line marks the mean of the dashed line distribution, while the solid vertical line marks the mean of the other two distributions. Although the means of all three distributions closely coincide (and two actually overlap), the tails are very different. For example, the percent of the population with usual intake of less then one serving per day (shaded region) is estimated as: 9.3%, 3.6%, and 0.4%. [2] We can see that, when the goal is to estimate various percentiles of the intake distribution, usual intake can not be represented by an average of any operationally feasible number of individual recalls.

2 Review of earlier methodology

2.1 The common modelling framework

All the statistical models based on 24-hour recall share a common basic framework, upon which they build their specific approaches and additional assumptions [2]. The framework can be summarized in three steps:

(a) Assume a relationship between individual 24-hour recalls and usual intake

(b) Partition the total variation in the measured data into within- and between-person components

(c) Estimate population distribution of usual intake accounting for the within-person variation

We describe each of these in more detail, below.

(a) Assume a relationship between individual 24-hour recalls and usual intake

The common assumption is that a 24-hour recall intake is an unbiased estimator of the usual intake, that is, of true single-day intake. Using measured

recalls, the usual intake distribution can be estimated empirically, but doing this reliably typically requires impractically-large samples. Hence it is desirable to take a parametric approach instead, and in particular to assume Normality of some form for both within- and between-person deviations. However, intake distribution (both in individuals and populations) is typically skewed to the left, often strongly. This is due at least in part to intake of course being non-negative, with a minimal value of zero occurring frequently for many nutrients, for occasional "feasting" days also occurring, with daily intake much higher than mean intake.

A common resolution is to assume that a Normal distribution approximates the distribution of non-linearly transformed observed data, instead of the observed data themselves. After the transformed data are fit with a statistical model, the results are back-transformed to the original scale. An issue that arises is that the 24-hour recall intake can be assumed to be unbiased estimator on the original scale, or on the transformed scale, but not both. If unbiasedness is on the transformed scale, then back-transformation of individual usual intakes is just the inverse of the original transformation. However, if unbiasedness is assumed on original scale, than back-transformation of means involves an additional adjustment, due to Jensen inequality.

Some evidence exists that neither of these two assumptions holds with great accuracy. Also, except for the small number of cases where unbiased biomarkers can be used as a check, it is impossible to know which of the two is more appropriate [2]. One such case is illustrated in Figure 2, taken from dodd06. This figure shows distributions of usual energy intake for female participants in the Observing Protein and Energy Nutrition study [7], estimated by different methods and unbiasedness assumptions, compared to the gold standard of biomarker measurements of the true intake (solid line). The long-dashed line shows estimates based on FFQs, the short-dashed line is the estimate based on two 24-hour recalls per person assuming unbiasedness on transformed scale (marked with "A" in the legend), and the dot-dashed line is the same except assuming unbiasedness on the original scale (marked with "B" in the legend). It is concluded that 24-hour recall exhibits a tendency toward underestimation, which is partially offset in analyses by assuming unbiasedness on the original scale.

(b) Partition the total variation in the measured data into within- and between-person components

Individual usual intakes (denoted μ_i for subjects $i = 1, 2, \ldots, n$) are ex-

pressed as the sum of the whole sample's mean (denoted μ) and person-specific deviations from it (denoted ϵ_i), representing the between-person variation. Hence we have

$$\mu_i = \mu + \varepsilon_i. \tag{1}$$

Within each individual, each particular 24-hour recall is expressed as the sum of the individual usual intake, and measurement-specific deviation from it (denoted ϵ_{ij}). These deviations represent the within-person variation, and we obtain

$$\mu_{ij} = \mu + \varepsilon_i + \varepsilon_{ij}. \tag{2}$$

The parameters of such a model will be estimated using standard maximum likelihood methods. To estimate the within-person variance (i.e. to obtain some estimate of the spread of the ϵ_{ij}) at least some individuals must have at least two 24-hour recalls.

(c) Estimate population distribution of usual intake accounting for the within-person variation

One rudimentary way to achieve this would be to shrink the individual observed means toward the overall 24-hour recall mean, then construct an empirical distribution of those shrunken values. If the within-person variance is σ^2_ω, and between-person variance is σ^2_b, and the empirical distribution of within-person means has variance $\sigma^2_b + \sigma^2_\omega/n$, a set of shrunk intake values with the desired variance σ^2_b is given by shrinking each individual mean to

$$\mu_i^{shrunk} = (1 - \omega) \cdot \mu + \omega \cdot \mu_i, \tag{3}$$

where

$$\omega = \sqrt{\frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\omega/n}}. \tag{4}$$

2.2 The Institute of Medicine Method (IOM)

Researchers representing the Institute of Medicine developed a detailed method [8] that builds on the common framework, and includes a power or log transformation of 24-hour recall data. Components of variance are estimated on the transformed scale, and individual means are shrunk toward the overall mean of the transformed means, then back-transformed to the

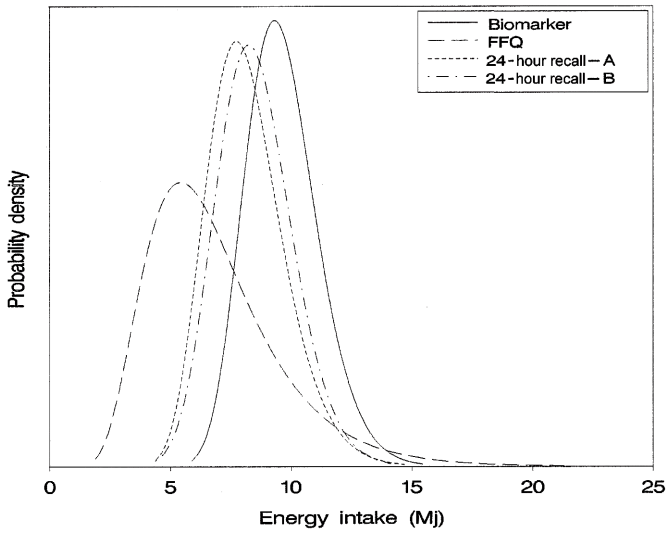


Figure 2: Distributions of usual energy intake for female participants in the Observing Protein and Energy Nutrition study [7], estimated by different methods and unbiasedness assumptions, and compared against the golden standard of doubly labeled water biomarker measurements of the true intake (solid line). Long-dashed line shows estimate based on FFQs. Short-dashed line is the estimate based on two 24-hour recalls per person assuming unbiasedness on transformed scale (marked with "A"), while dot-dashed line is the same except assuming unbiasedness on the original scale (marked with "B"). Taken from Dodd *et al* (2006) [2].

original scale. The method is therefore consistent with the assumption of unbiasedness on the transformed scale.

2.3 The Iowa State University Method (ISU)

Another method for modeling usual intake was developed at Iowa State University [9, 10]. In contrast to IOM, here unbiasedness is assumed on the original scale, and the method can be extended to 24-hour recalls from complex surveys.

The method is based on a complex model that uses two-stage transformation to obtain almost exact Normality of the transformed 24-hour recalls, but the transformation itself requires a large amount of data, with at least 50 subjects having double 24-hour recalls [2]. A unique feature of this approach is that it allows the within-person variance to vary across individuals, modelling the fact that some individuals have more daily variation in their diet then others. Unbiasedness of 24-hour intakes is assumed to hold on the original scale, for which back-transformation procedure is adjusted.

2.4 The Best-Power Method (BP)

Some developers of the ISU method proposed a simplified version, known as the Best-Power method. It uses only a single stage power or log transformation which is easy to back-transform with an adjustment for unbiasedness of recalls on the original scale. This version does not allow within-person variance to vary across individuals, and like ISU can be extended to recalls from complex surveys. Importantly, a simulation study comparing ISU and BP methods found the differences to be very small in practical terms [9]. This suggests that the frequently made assumption of within-person variance being constant for all individuals is quite useful, especially when sample size is modest. It adds robustness to the statistical model by removing a degree of freedom that contributes little to model's output.

2.5 The Iowa State University Food Method (ISUF)

The methods described thus far were developed to model usual intake where the distribution of single 24-hour recalls can be transformed to being approx-

imately Normal. This, it appears in practice, is the case for many foods and nutrients that many subjects consume daily, or almost daily. However, for episodically-consumed foods and nutrients (for example broccoli, fish, whole grains, etc) there will be many days of observed zero consumption, whether a person consumes that item only occasionally or strictly never. This leads to a distribution of intake that has a spike of zero observations in the left tail, and transforming it to Normality is not feasible. For this reason, the ISU method was extended [11] to explicitly deal with the spike at zero. Under this approach, the zero observations are treated separately form nonzero ones, this being motivated by seeing the n-day within person total intake as the product of probability of eating on any given day, and intake on a consumption day:

$$\frac{\text{total intake}}{n} = \frac{k}{n} \cdot \frac{\text{total intake}}{k}, \tag{5}$$

where k is the number of consumption days. The first step in ISUF is to estimate the distribution of single-day consumption probabilities in the population. This distribution is modeled as a multinomial where each subject's consumption probability can be either $0, 0.02, ..., 0.98$ or 1.0, i.e. one of 51 values. We estimate the proportion of individuals in the population having each of these values by the observed proportions of individuals who consumed $0, 1, ..., n$ out of n possible 24-hour recalls.

The distribution of usual intake *on a consumption day* is estimated by applying the ISU method on nonzero 24-hour recalls. Finally, the two distributions of consumption probability and usual consumption day intake are combined, while assuming that the two are independent, in other words that usual consumption-day intake is unrelated to consumption probability. However, Dodd *et al* [12] demonstrated that these two variables can be positively correlated in practice, as individuals who consume a food more when they do eat it also consume it more frequently. Neglecting this effect may produce additional bias, overestimating the amount consumed by people who consume more sporadically and underestimating the amount in those who consume more frequently.

An important limitation common to all these models is with subgroup analyses. Differences in usual intake between sexes or among age groups require separate estimations for each subgroup, in effect forcing a stratified approach between subgroups. None of the methods (IOM, ISU, BP, ISUF) can include adjustment for covariates in the context of unified model.

3 The NCI method

3.1 Introduction

Although 24-hour recalls provide rich detail on dietary intake of a given day and a reasonably small recall bias, collecting more than two recalls per individual is impractical, not only in large surveys such as NHANES [1, 13], but also often in small surveys, when operations on the population of interest can be challenging [14, 15]. For reasons of simplicity as well as practicality, the NCI method is presented here assuming that there are at most two 24-hour recalls per person. This is also the standard setup that NCI method is set to address, and generalization to a larger number of recalls is straightforward.

Like its antecedents described in Section 2, the NCI method [13, 16, 17] starts with the common modeling framework of Section 2.1, and uses insights gained from it to address the most general problem of estimating the population distribution of usual intake of episodically consumed foods. It addresses the following challenges and issues:

(a) Distinguishing and accounting for both within-person and between-person intake variability;

(b) Allowing for the highly skewed, non-negative intake data, with extreme values in the upper tail;

(c) Accounting for the number of days (potentially many days) without consumption;

(d) Allowing for correlation between probability of consumption and the consumption-day intake amount;

(e) Allowing for covariate treatment and covariate adjustment in the context of an unified model;

(f) Sacrificing the heterogeneity of within-person variance for a gain in model robustness, which is helpful for dealing with samples of modest size.

The most elaborate among the methods reviewed in Section 2 is the ISUF method, that meets all the challenges above except incorporating covariate

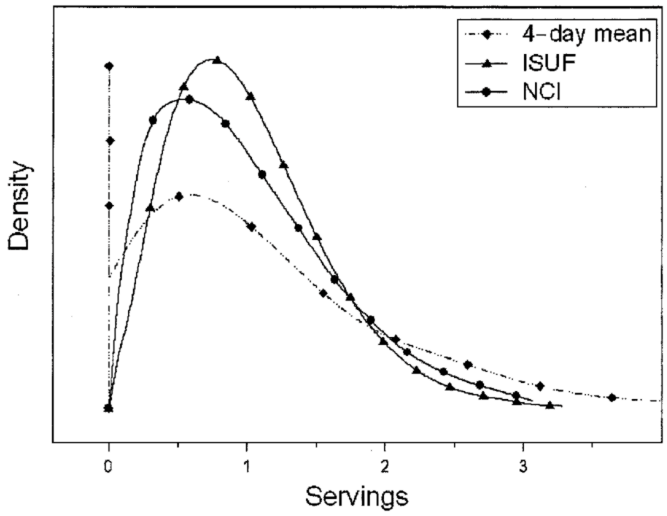


Figure 3: Estimated distributions of usual intake of whole grains for women in the EATS study, using different methods. Taken from Tooze *et al* (2006) [13].

information and allowing for the correlation between intake frequency and amount. That this correlation can be a relevant effect is illustrated in Figure 3, taken from [13]. It shows estimated distributions of usual intake of whole grains for women in the EATS study [18], using different methods. The distribution with the spike at zero is based on fitting usual intake values as just the mean of four 24-hour recalls per person, and it shows that whole grains can be considered being episodically consumed. Two other curves that are capable of addressing this phenomenon use ISUF (solid line connecting triangles) and NCI (solid line connecting dots) methods. As subsequent work showed [13], the difference between the two curves is due to ISUF's inability to correlate intake probabilities and amounts. When the correlation is constrained to be zero in NCI approach, the two curves nearly coincide. Since the ISUF method disregards the relationship between intake frequency and amount, it artificially increases the middle section of its curve at the expense of the tails.

An additional problem with the ISUF method is specific to the constraint of having only two 24-hour intakes per person. The observed probabilities can take only three values: $\{0, \frac{1}{2}, 1\}$. Estimating a smooth distribution of consumption probabilities for such discrete data can be difficult [13].

3.2 Assumptions and overview

The assumptions of the NCI method are [13]:

(a) That a 24-hour recall is an unbiased estimate of usual intake of episodically-consumed foods on the original scale of the intake amounts. In other words, repeated over many days, the average of observed amounts would approach the person's true usual intake;

(b) That 24-hour recalls do not misclassify the reported amounts: if the food was not consumed on a particular day if will be reported as zero, and if it was consumed it will be reported as non-zero;

(c) That non-zero recalls can be transformed to approximate Normality;

(d) That the sample contains at least a few individuals having at least two non-zero measured 24-hour recalls.

Applying the method consists of two steps. First, the statistical model is fit by a standard non-linear likelihood maximization numerical procedure. The model has two interlinked parts, one dealing with intake frequency and the other with intake amount, and it describes the relationship between usual intake and population covariates. Fitting it gives estimates of both within- and between-person variability of intakes. As part of the model, the non-zero intake amounts are transformed, flexibly, using the Box-Cox [19] transformation.

In the second step, additional and application-dependent statistical procedures are performed on the fitted model, to produce the final result of the analysis. These procedures typically consist of back-transforming the data to original scale, then calculating an empirical CDF and its various percentile point estimates. Variability of these percentiles can additionally be estimated using repeated sampling.

3.3 Formal model underpinning the NCI method

For an assumed sample of n individuals, each with two 24-hour recalls measured, the construction of the model starts with distinguishing consumption occurrence from consumption amount. Let random variable Y_{ij} denote the

24-hour recall measured amount for individual i at time $j = 1, 2$. Then denote R_{ij} as the indicator of occurrence, i.e.

$$R_{ij} = \begin{cases} 0\,, & Y_{ij} = 0 \\ 1\,, & Y_{ij} > 0\,. \end{cases} \tag{6}$$

The non-zero amounts of consumption are transformed to a scale where they will be assume to be Normally distributed:

$$S_{ij} \equiv [g(Y_{ij})|R_{ij} = 1]\,, \tag{7}$$

where

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}\,, & \lambda > 0 \\ \log(y)\,, & \lambda = 0 \end{cases}, \tag{8}$$

i.e. the Box-Cox transformation with $\lambda \geq 0$ as its single free parameter.

The intake frequency part is modeled as mixed-effects logistic model, with fixed effects given by covariates $\vec{\beta}_1$ and a random-intercept effect u_{1i}. Specifically

$$R_{ij} \,|\, \vec{\beta}_1, u_{1i} \;\sim\; \text{Bernoulli}\left(p_{ij}|\vec{\beta}_1, u_{1i}\right), \tag{9}$$

$$\text{and logit}(p_{ij} \,|\, \vec{\beta}_1, u_{1i}) \;=\; X_{1,ij}^T \cdot \vec{\beta}_1 + u_{1i}\,, \tag{10}$$

where X_1 is the design matrix of covariates for for intake frequency. The intake amount is modeled, on the transformed scale, using a mixed-effects non-linear model with fixed effects given by covariates $\vec{\beta}_2$ and a random-intercept effect u_{2i}. Specifically,

$$S_{ij} \,|\, \vec{\beta}_2, u_{2i} \sim N\big(X_{2,ij}^T \cdot \vec{\beta}_2 + u_{2i}\,, \sigma_e^2\big)\,, \tag{11}$$

where X_2 is the design matrix of covariates for intake amount, and σ_e^2 thus represents the between-person variance of intake on a consumption day.

The sets of covariates included in X_1 and X_2 can in principle be identical, overlapping, or altogether different, and their coefficients can similarly be related, or not.

Importantly, the random effects for frequency and amount are allowed to be correlated. This is achieved by assuming they have a bivariate Normal joint distribution:

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right). \tag{12}$$

The marginal variances σ_1^2 and σ_2^2 represent within-person variances of, respectively, frequency and non-zero intake. These are assumed to be the same for each individual, as suggested for the BP method of Section 2.4. Correlation coefficient ρ is the correlation between the individual frequency and consumption day amount.

The probability distribution function of Y_{ij} is a mixture:

$$f(y_{ij}) = p_{ij}^{r_{ij}} f_S(s_{ij}) + (1 - p_{ij})^{1-r_{ij}} \,, \tag{13}$$

and the total likelihood for n individuals is

$$\begin{aligned} L(\vec{\beta}_1, \vec{\beta}_2, \sigma_1, \sigma_2, \sigma_e, \rho, \lambda) \;\; &= \;\; \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \iint \prod_{j=1}^{2} \left[1 - p_{ij}(\vec{\beta}_1, u_{1i})\right]^{1-r_{ij}} \cdot p_{ij}(\vec{\beta}_1, u_{1i})^{r_{ij}} \cdot \\ &\quad \cdot f_S(s_{ij}|\vec{\beta}_2, u_{2i}, \sigma_e) \cdot f_{12}(u_{1i}, u_{2i}|\vec{\beta}_2, \sigma_1, \sigma_2, \rho)\, du_{1i}\, du_{2i} \,, \end{aligned} \tag{14}$$

where we denote

$$f_S(s_{ij}) = \begin{cases} 1 \,, & r_{ij} = 0 \\ f_S(g(y_{ij})) \cdot y_{ij}^{\lambda-1} \,, & r_{ij} = 1 \,. \end{cases} \tag{15}$$

The total likelihood (14) can be maximized using any non-linear optimization procedure in the space of its parameters, for example by quasi-Newton optimization with the likelihood approximated by adaptive Gaussian quadrature [20], starting from some reasonable guess for parameters values. However, this involves solving double integrals which may make the procedure too slow for practical purposes, at least in this raw form.

It is of interest to note that in the first publication of the NCI method [16], the amount part S_{ij} was not transformed to another scale, but instead assumed to be log-normal, i.e.

$$\log(S_{ij} \,|\, \vec{\beta}_2, u_{2i}) \sim N\big(X_{2,ij}^T \cdot \vec{\beta}_2 + u_{2i} \,, \sigma_e^2\big) \,. \tag{16}$$

While no reason was stated for the change—which was implemented soon after the first publication—it may be due to the double-integration, which can not be avoided under log-Normal assumptions. However, if S_{ij} is assumed Normal on some transformed scale, one of the integrals in (14) can be solved analytically, leading to a significant speedup in execution.

After the model is fitted, estimating the CDF of intake values from the parameter estimates is done by a Monte Carlo procedure, since the total

likelihood itself has no analytical solution. To numerically create a set of virtual intake values that would simulate the whole population intake distribution while reflecting the covariate structure of the individuals in the sample, $X_{1i}^T\,\vec{\beta}_1$ and $X_{2i}^T\,\vec{\beta}_2$ are calculated for every individual. Next, a set of virtual individuals ($k = 1 \ldots K$) is created by sampling a number of correlated bivariate Normal pairs $(u_{1i}^{(k)}, u_{2i}^{(k)})$, using the fitted values $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho}$. With this, each virtual individual (i, k) is represented by their fitted consumption probability $\hat{p}_i^{(k)}$ and mean consumption-day amount $\hat{\mu}_i^{(k)}$:

$$\hat{p}_i^{(k)} = X_{1i}^T\hat{\beta}_1 + u_{1i}^{(k)}\,, \tag{17}$$

$$\hat{\mu}_i^{(k)} = X_{2i}^T\hat{\beta}_2 + u_{2i}^{(k)}\,. \tag{18}$$

The model assumes that a virtual individuals' consumption-day intakes are Normally distributed on the transformed scale, having the within-person variance σ_e^2:

$$s_i^{(k)} \sim N(\mu_i^{(k)}, \sigma_e^2)\,. \tag{19}$$

When $s_i^{(k)}$ is back-transformed to the original scale to find a proper representation of the usual consumption-day intake of that virtual individual, it is not enough to use a simple Box-Cox inverse of the transformed scale mean $g^{-1}(\mu_i^{(k)})$. Due to Jensen's inequality and the non-linearity of $g(\cdot)$, we must instead calculate the mean of the back-transform of the whole Normal distribution for $s_i^{(k)}$.

One way to achieve this is to employ the nine-point approximation used previously by the ISU method [17, 21]. In this approximation, sets of nine points and nine weights are constructed in such a way that the first five moments are the same of the normal distribution $N(\mu_i^{(k)}, \sigma_e^2)$ and the discrete distribution with mass concentrated at the nine points, in proportion to the weights. Then, these nine discrete points of mass are individually back transformed via g^{-1}, with their mean then representing the virtual individual's usual consumption day intake $y_i^{(k)}$.

Finally, the usual intake of each virtual individual is given as the product $p_i^{(k)} \times y_i^{(k)}$. A very large sample of these usual intakes can be used to calculate the CDF empirically. From this CDF, various percentiles can be calculated, and corresponding 95% CIs can in principle be constructed by bootstrapping the original sample of surveyed individuals.

To estimate the variability of the estimated CDF percentiles when estimated from smaller samples, we can also repeatedly take small random samples

from a larger set of surveyed individuals, and examine the empirical variability in CDF percentiles estimated from each small sample.

3.4 Implementation in R

The NCI method, as outlined above, has been implemented in *SAS* [17]. The first task of this thesis is to efficiently implement the basic method in R, to make it more readily available to the research community. Subsequent updates to the code are expected to add features beyond the basic ones, for example allowing for complex survey weighting, assuming more than two 24-recalls per person, allowing for missing data, and optimizing the execution speed by interfacing with routines in C++ and other lower-level languages.

3.4.1 Likelihood maximization

To fit the NCI model, we start with the total likelihood (14) as the product of individual terms L_i, and distinguish four different cases based on possible (R_{i1}, R_{i2}) values. Throughout, we will make use of the following useful substitutions:

$$p_{ij} = \frac{\exp\big[X_{1,ij}^T \vec{\beta}_1 + u_{1i}\big]}{1 + \exp\big[X_{1,ij}^T \vec{\beta}_1 + u_{1i}\big]} \,, \tag{20}$$

$$\begin{aligned} f_{12}(u_{1i}, u_{2i}) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\big[-\frac{1}{2(1-\rho^2)}(\frac{u_{1i}^2}{\sigma_1^2} + \frac{u_{2i}^2}{\sigma_2^2} - 2\rho\frac{u_{1i}}{\sigma_1}\frac{u_{2i}}{\sigma_2})\big] \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\big(-\frac{u_{1i}^2}{2\sigma_1^2}\big) \cdot \left\{ \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\big[-\frac{(u_{2i} - \rho\frac{\sigma_2}{\sigma_1}u_{1i})^2}{2\sigma_2^2(1-\rho^2)}\big] \right\}. \end{aligned} \tag{21}$$

We also use the fact that the part in largest brackets in (21) integrates out: $\int_{-\infty}^{\infty} du_{2i} \big\{ \dots \big\} = 1$.

3.4.2 Likelihood maximization: Case $(0, 0)$

With

$$R_{i1} = 0 \,,\; R_{i2} = 0 \,,\; f_S(s_{i1}) = f_S(s_{i2}) = 1$$

then

$$L_i^{00} = \iint \big[1-p_{i1}\big]\cdot\big[1-p_{i2}\big]\cdot 1\cdot 1\cdot f_{12}\; du_{1i}\, du_{2i}$$
$$= \int_{-\infty}^{\infty} \big[1+\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1+u_1}\big]^{-1}\big[1+\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1+u_1}\big]^{-1}\cdot\frac{1}{\sqrt{2\pi}\sigma_1}\,\mathrm{e}^{-\frac{u_1^2}{2\sigma_1^2}}\,du_1\,, \quad (22)$$

which it is convenient, for coding purposes, to rewite as

$$L_i^{00} = D_{00}\cdot\int_{-\infty}^{\infty}\frac{1}{1+\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1+u_1}}\cdot\frac{1}{1+\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1+u_1}}\cdot\mathrm{e}^{\left(A_{00}+\frac{B_{00}^2}{C_{00}}\right)}\,du_1\,, \quad (23)$$

where

$$A_{00}\equiv -\frac{1}{2(1-\rho^2)}\cdot\frac{u_1^2}{\sigma_1^2}\,, B_{00}\equiv\frac{\rho}{1-\rho^2}\cdot\frac{u_1}{\sigma_1\sigma_2}\,, C_{00}\equiv\frac{1}{2(1-\rho^2)\sigma_2^2}\,, D_{00}\equiv\frac{1}{\sqrt{2\pi}\sigma_1}\,.$$

3.4.3 Likelihood maximization: other cases

The other cases follow as minor modifications to those in Section 3.4.2. With

$$R_{i1}=1\,,\; R_{i2}=0\,,\; f_S(s_{i2})=1\,.$$

then after some straightforward algebra analogous to (23) we find that

$$L_i^{10} = D_{10}\cdot\int_{-\infty}^{\infty}\frac{\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1+u_1}}{1+\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1+u_1}}\cdot\frac{1}{1+\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1+u_1}}\cdot\mathrm{e}^{\left(A_{10}+\frac{B_{10}^2}{C_{10}}\right)}\,du_1\,, \quad (24)$$

where

$$A_{10}\equiv A_{00}-\frac{1}{2\sigma_e^2}(X_{2,i1}^T\vec{\beta}_2-s_{i1})^2\,, B_{10}\equiv B_{00}-\frac{1}{\sigma_e^2}(X_{2,i1}^T\vec{\beta}_2-s_{i1})\,, C_{10}\equiv C_{00}+\frac{1}{2\sigma_e^2}\,,$$

$$D_{10}\equiv\frac{y_{i1}^{\lambda-1}}{2\sqrt{2}\pi\sigma_e\sigma_1\sigma_2\sqrt{1-\rho 2}}\cdot C_{10}^{-1/2}\,.$$

With

$$R_{i1}=0\,,\; R_{i2}=1\,,\; f_S(s_{i1})=1\,.$$

we similarly obtain

$$L_i^{01} = D_{01}\cdot\int_{-\infty}^{\infty}\frac{1}{1+\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1+u_1}}\cdot\frac{\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1+u_1}}{1+\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1+u_1}}\cdot\mathrm{e}^{\left(A_{01}+\frac{B_{01}^2}{C_{01}}\right)}\,du_1\,, \quad (25)$$

where

$$A_{01} \equiv A_{00} - \frac{1}{2\sigma_e^2}(X_{2,i2}^T\vec{\beta}_2 - s_{i2})^2\,, B_{01} \equiv B_{00} - \frac{1}{\sigma_e^2}(X_{2,i2}^T\vec{\beta}_2 - s_{i2})\,,$$

$$C_{01} \equiv C_{00} + \frac{1}{2\sigma_e^2} \equiv C_{10}\,, D_{01} \equiv \frac{y_{i2}^{\lambda-1}}{2\sqrt{2}\pi\sigma_e\sigma_1\sigma_2\sqrt{1-\rho 2}} \cdot C_{01}^{-1/2}\,.$$

Finally, for

$$R_{i1} = 1\,,\; R_{i2} = 1\,.$$

we obtain

$$L_i^{11} = D_{11} \cdot \int_{-\infty}^{\infty} \frac{\mathrm{e}^{X_{1,i1}^T\vec{\beta}_1 + u_1}}{1 + \mathrm{e}^{X_{1,i1}^T\vec{\beta}_1 + u_1}} \cdot \frac{\mathrm{e}^{X_{1,i2}^T\vec{\beta}_1 + u_1}}{1 + \mathrm{e}^{X_{1,i2}^T\vec{\beta}_1 + u_1}}\, \mathrm{e}^{\left(A_{11} + \frac{B_{11}^2}{C_{11}}\right)}\, du_1\,, \qquad (26)$$

where

$$A_{11} \equiv A_{00} - \frac{1}{2\sigma_e^2}\left[(X_{2,i1}^T\vec{\beta}_2 - s_{i1})^2 + (X_{2,i2}^T\vec{\beta}_2 - s_{i2})^2\right],$$

$$B_{11} \equiv B_{00} - \frac{1}{\sigma_e^2}\left[(X_{2,i1}^T + X_{2,i2}^T)\vec{\beta}_2 - (s_{i1} + s_{i2})\right],$$

$$C_{11} \equiv C_{00} + \frac{1}{\sigma_e^2}\,, D_{11} \equiv \frac{(y_{i1}y_{i2})^{\lambda-1}}{4\pi\sqrt{\pi}\sigma_e^2\sigma_1\sigma_2\sqrt{1-\rho 2}} \cdot C_{11}^{-1/2}\,.$$

3.4.4 Likelihood maximization: numeric optimization

We fit the model by maximizing log-likelihood, with likelihood being the product over individual terms $L_i^{(r_{i1},r_{i2})}$, as given in expressions above. Each term involves just one-dimensional integration, which is a straightforward task for numerical calculation as implemented in standard R routines, since the integration grid does not need to be adaptive but can be fixed, as confirmed in our tests. Additionally, many of the terms and factors in the expressions above are constants with respect to the parameters, and so can be calculated once, before the main loop over individual terms.

We maximize the log-likelihood with R's `nlm()` function, which is a part of its core distribution [22]. This routine minimizes a multi-variate function using a Newton-type steepest descent algorithm, evaluating the local Hessian numerically.

After the model is successfully fitted, we calculate the Hessian at the optimum using the `hessian()` function from the `numDeriv` package [23]. This Hessian is used when calculating the observed Fisher information and also for subsequent profiling steps.

The model parameters $(\sigma_1, \sigma_2, \sigma_e, \rho, \lambda)$, have constraints on their possible values, which is inconvenient for non-linear optimization as the search algorithm may try to explore beyond the allowed ranges and/or use step-halving or similar methods when the optimum is near the boundary. For this reason, we rescale these parameters to $(\sigma_1^*, \sigma_2^*, \sigma_e^*, \rho^*, \lambda^*)$, which all have the full $(-\infty, +\infty)$ range:

$$\sigma_1^* = \log(\sigma_1)\,,\ \sigma_2^* = \log(\sigma_2)\,,\ \sigma_e^* = \log(\sigma_e)\,,$$
$$\rho^* = \operatorname{artanh}(\rho)\,,\ \lambda^* = \operatorname{artanh}(2\lambda - 1)\,. \tag{27}$$

The optimizer 'sees' and operates solely on the rescaled parameters. After the model is fitted, the parameter estimates may be transformed back to the original scale, as needed, for later stages in the analysis.

3.4.5 Convergence issues

Our testing has showed that it is useful to limit the maximum allowed scaled step length (`stepmax`) to prevent `nlm()`'s algorithm leaving the area of interest in parameter space, and causing the numerical integration to fail—typically in its first few steps—to find the one-dimensional interval where integrand is non-zero. We therefore set the scaled step length to 3. Also, sometimes it takes many steps for the algorithm to converge, so we increase the maximum number of iterations before program is terminated (`interlim`) to 300.

To calculate the one-dimensional integrals numerically, we use R's `integrate()` function, part of its core distribution [22]. This uses a grid-based quadrature method, and since the grid is fixed and not adaptive, it can yields accurate results very quickly if the grid is set to have a good coverage of the domain where integrand is non-zero. The whole integration domain is $(-\infty, +\infty)$, and integrands are typically close to Gaussian and so notably non-zero only on a limited part of that domain. However, a pre-set grid may explore that part only partially or even miss it altogether, as illustrated in Figure 4.

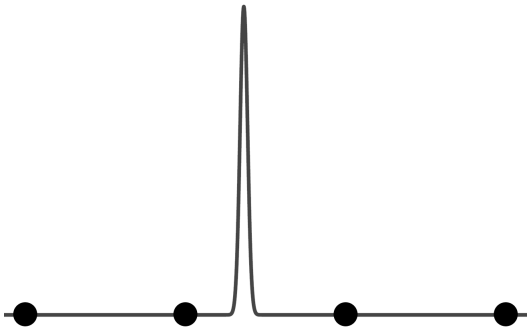


Figure 4: Failing to find the domain of non-zero integrand. The solid line represents one-dimensional Gaussian-like function to be integrated on $(-\infty, +\infty)$. Dots are fixed grid points that miss the non-zero part of the function. Using quadrature methods evaluating the function at these points would lead to underestimation of the integrand, as the domain of non-zero integrand is too far from the grid.

For this reason, we set the grid to span a wide area $(-10\,\sigma_1, +10\,\sigma_1)$, where σ_1 is current best estimate of that parameter. Also, we set the grid to be much denser than its default value, with `rel.tol` equal to 1.5×10^{-11} being an empirically good choice. Our tests showed that these settings mean only relatively small decrease in overall execution speed, relative to more common settings of $(-5\,\sigma_1, +5\,\sigma_1)$ for the grid span and the default value of `rel.tol`.

Using these steps, the algorithm proceeds to the optimum very reliably once it is not too far from the solution. However, the first few steps may be unstable, as the estimates still jump around in way that makes little use of the data, and may occasionally cause `integrate()` to abort due to floating point issues, being unable to perform when the parameter values happen get particularly awkward. To circumvent these issues, we employ a following strategy:

(a) Set a default value of zero for each integral

(b) Perform integration inside R's `try()` block, and if the procedure aborts, continue the code execution with the default zero value.

This approach substantially improves overall convergence rates. The rea-

soning behind it is that integrands are positive, and their integrals represent individual contributions to the likelihood. Setting an integral to zero acts as a penalty that reduces the likelihood and motivates the parameters away from such awkward areas in the parameter space. Alternative strategies are possible, for example reducing the maximal step length further. However, that might slow down the optimization by requiring more steps. After some experimenting, we moderately reduced `stepmax` and increased `interlim` relative to their default values, while applying the described strategy of dealing with unstable integration.

Convergence of the method for any given sample is not guaranteed, and it will occasionally fail. The `nlm()` function returns an integer `code` indicating why an optimization algorithm terminated:

1. Gradients were close to zero
2. Function value changed too little between steps
3. Failed to find a next step with a lower function value
4. Maximum number of iterations exceeded
5. Maximum step size exceeded

Although codes $1-3$ may indicate that a local maximum is found, only choice 1 can be accepted as choices 2 and 3 typically do not return a positive-definite Hessian. To increase the probability of an optimization ending with code 1, we employ a strategy inspired by simulated annealing approach:

(a) Perform optimization algorithm once; if it returns code=1 end here; if code≠1, continue;

(b) Randomly perturb each model parameter except λ from its converged value p_k, with:

$$p_k \longrightarrow p_k \cdot (1 + m \cdot \xi_k)\,, \quad \xi_k \sim \text{Uniform(-1.1)}\,. \tag{28}$$

These two steps alternate with the perturbation strength m, with each new pass, first incrementally increasing as $m = 0.01, 0.02, \ldots, 0.30$, then decreasing back as $m = 0.29, 0.28, \ldots, 0.01$. If after any single optimization run

code=1, the whole process ends with success. If it never happens, the case is recognized as not converged.

It is important to exclude the Box-Cox scaling parameter λ from being perturbed in this way, since it affects the data differently then other parameters do, in extremely non-linear and sensitive way. Furthermore, of all the parameters it is most informed by data, and perturbing it even by a small amount is unnecessary.

Without the last two strategies described (zero-integral penalty, simulated annealing), we find that typical convergence rate for small sized samples, up to n=200, is 60–80%. With these, the convergence rate typically reaches 95%, and increases further with system size.

3.4.6 Down-sampling and profiling

Maximizing the likelihood only provides point estimates of the model's parameters, which are in turn used to estimate percentiles of the intake distribution. To estimate variability of these quantities, we employ two approaches, using down-sampling—essentially a use of the "m out of n bootstrap" [24]—and profile likelihood.

We perform down-sampling in accordance with the correlation structure of the data, that is, by resampling (with replacement) entire individuals' clusters of two 24-hour recalls. Using a full dataset of n individuals, in this way we create many samples of size m, and the optimization algorithm is performed on each. The 2.5-97.5 percentile ranges are taken as measure of variability in samples of size m. We also use the mean of the point estimates from the samples of size m to indicate the magnitude and direction of small-sample biases in the point estimate.

This procedure is essentially the same as the "m out of n bootstrap" [24], with the difference that we do not aim to provide approximate confidence intervals based on the full dataset of size n, but instead to give a strongly data-based idea of how well an analysis would work (i.e. how variable and biased point estimates would be) in smaller studies from the same population. In this thesis, we are particularly interested in small samples and trends that develop with changing sample size. When m is small, our down-samples are small compared to the whole datasets they are drawn from, and very close to being random samples without replacement. We thus take these samples

to represent small random samples drawn from a population represented by the whole dataset.

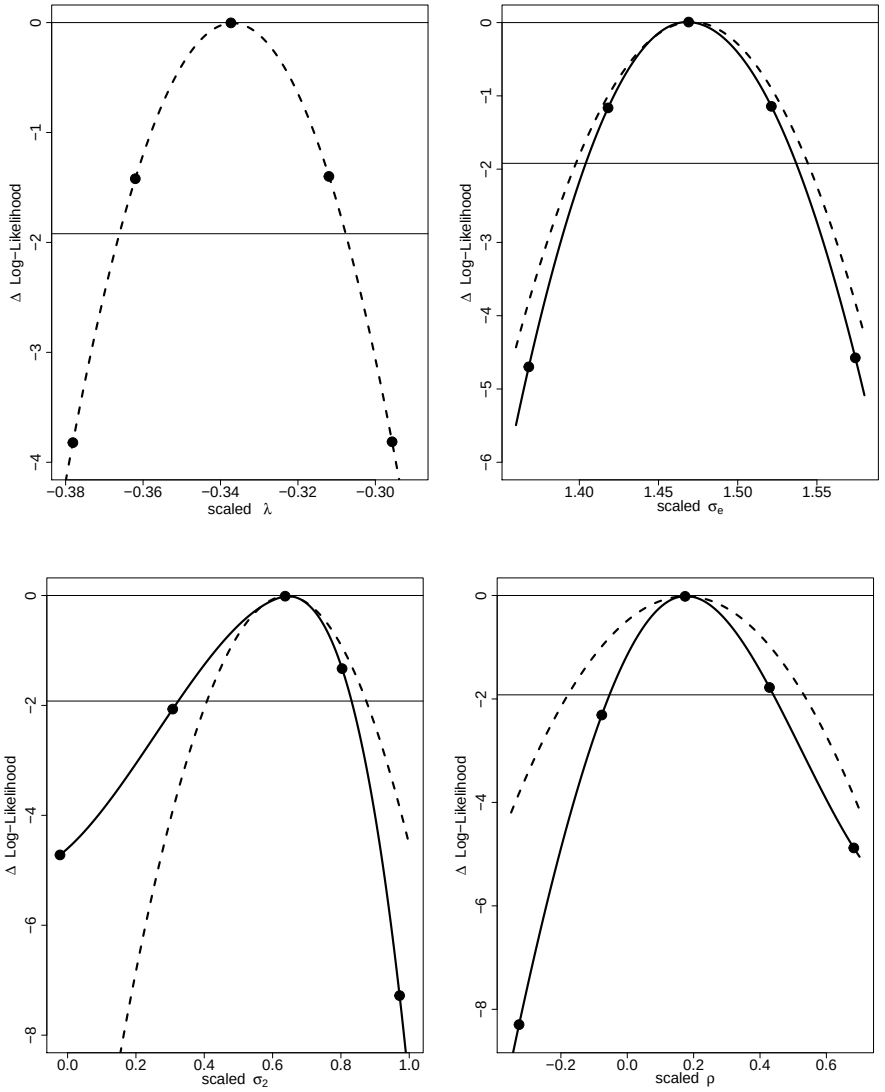
For a regular bootstrap procedure, a couple thousand of samples would typically suffice for inference [25]. However, since NCI likelihood optimization is slow, and we are interested in developing trends over a grid of sample size values and not precise calculations for specific datasets, we create many fewer samples at a particular sample size, typically only a hundred.

When down-sample size m is large, and we (as in practice) have a large original sample size n, our approach can be expected to work well but, unfortunately, the number of numeric integration steps required makes it impractically slow. In practice we find $m = 200$ to be the upper limit where repeated down-sampling is computationally tractable. For the behaviour of estimates at values of $m > 200$, we therefore instead consider a profiling procedure, and use the corresponding approximate confidence intervals over a handful of down-samples to indicate behavior of the point estimates. The process is repeated over a grid of down-sample sizes m.

For each down-sample of a specified size m, for a particular parameter p_k for which we want an approximate confidence interval, we select several values for p_k in the vicinity of the estimate p_k^0 obtained from the sample of size m. For each selected value, we fix that particular parameter and maximize the likelihood under this constraint. This gives the profile likelihood, a function of p_k. Under standard large-sample approximations and correct-model assumptions, Wilks' theorem states that twice the decrease in log-likelihood from the unconstrained to the constrained model is distributed as approximately χ^2_1. Thus, the p_k is the range where the likelihood decrease is smaller than $\frac{1}{2}\chi^2_{1,0.95}$ form an approximate 95% confidence interval for p_k.

Agreement between down-sampling and profiling in estimates of variability can act as an informal diagnostic that the asymptotic approximations that rely on quadratic likelihoods are not grossly violated, for a particular parameter. Figure 5 illustrates such several situations, all using the large NHANES dataset containing 16,776 individuals and two 24-hour recalls per person. On each panel, dots are constrained log-likelihood calculations; dashed lines are quadratic approximations to log-likelihood around a converged p_k^0 value, calculated using the appropriate eigenvalue of inverted Hessian; thick solid lines are added as eye-guides. The y-axis show the decrease in log-likelihood from the unconstrained value at the maximum, and the x-axis shows values of the selected parameter p_k, scaled as non-linear optimizer sees them. The

Figure 5: Profiling of some selected model parameters, with the full NHANES dataset. The ends of the approximate 95% confidence intervals are the points where the drop in the constrained log-likelihood exceeds $\frac{1}{2}\chi^2_{1,0.95}$. The better-informed we are about a parameter, the closer is its profiling curve to the quadratic approximation.



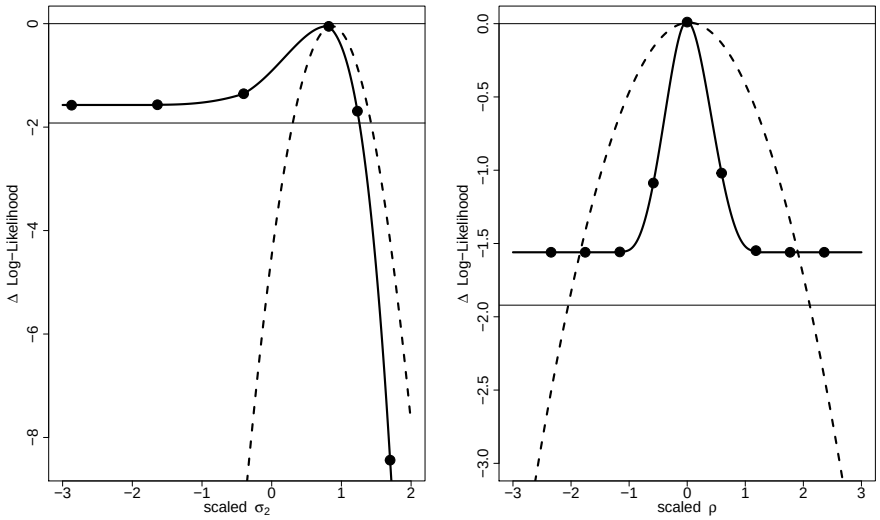


Figure 6: Profiling of some selected model parameters, with the NHANES dataset down-sampled to 850 randomly chosen individuals. As the overall information in the system decreases, the drop in log-likelihood in some parameters can be so reduced that it may not exceed the limiting value of $\frac{1}{2}\chi^2_{1,0.95}$.

thin horizontal lines denote zero and $-\frac{1}{2}\chi^2_{1,0.95}$, so 95% CIs are visualized as the points on x-axis where the solid eye-guide line crosses the lower thin horizontal line.

Figure 5 shows results of profiling for a few selected model parameters with the full dataset of 16,776 individuals included. The upper left panel shows λ profiling, which we have found is the best-informed parameter at any sample size. We can see that the calculated points agree with the quadratic approximation, suggesting the CLT is acceptably accurate for this parameter. The next three panels show three more parameters where the accuracy of the CLT is brought into question with increasing concern, judging by discrepancies between the calculated points and their quadratic approximations. For all these parameters, the solid lines do cross the lower thin line, so the usual form of 95% confidence interval can be estimated, albeit with growing skepticism about the accuracy of its coverage.

To illustrate more concerning behavior, Figure 6 shows results of the profiling procedure for a few selected model parameters with a random draw of 850 samples from the NHANES dataset (i.e. a down-sample). We see that reducing the sample size prevents some solid lines from crossing the limiting solid lines, meaning that one or both ends of the usual approximate 95% confidence interval would be unbounded. A bootstrap procedure, in contrast, would provide a set of finite estimates, and intervals constructed from this approach would be bounded. This discrepancy suggests that there is insufficient information for standard large-sample approximations to be accurate enough. Moreover, the small drop in constrained log-likelihood across the whole parameter range indicates that the data is never particularly more or less likely under any pair of parameter values, i.e. that the data provide little information by which to identify and set of parameter values as better explanations for the observed data.

3.4.7 Finding approximate CI by profiling

Since each execution of the optimization algorithm takes considerable execution time, we seek to find the ends of the profile-likelihood approximate 95% confidence intervals in the smallest possible number of steps. Using base R's `uniroot()` [22] function to perform the optimization is impractical, as it employs unnecessarily-many steps. We instead exploit the fact—inherent in the asymptotic results being used for the inference—that the profile like-

lihood looks like a distorted quadratic curve. Examples of this can be seen in Figures 5 and 6. From the unconstrained solution, we perform two independent searches, to the left and to the right. In each search we fit a local parabola through the last three points calculated, with the next search point set where the parabola crosses the limiting $-\frac{1}{2}\chi^2_{1,0.95}$ line. This typically produces an accurate solution with only 2-4 calculations per one side, even when the profiling curve is quite distorted from the quadratic form yet still crosses the limiting line. This performance should be compared to 7-10 steps that `uniroot()` typically requires to reach convergence.

3.4.8 Multimodality of the likelihood

Despite all the effort put into estimates converging to the global maximum of the likelihood, a fundamental limitation remains that the likelihood may be multimodal, with no guarantee available the global maximum can be found with finite computing resources. Multimodality is common in even the simplest mixed models, as shown by Liu *et al* (2003) [26]. Addressing the issue of multimodality for the complex model in the NCI method goes beyond the goals of this thesis, and is made yet more challenging but the considerable processor time needed just to evaluate the likelihood. The most we can do here is be aware of the issue, and illustrate it by example.

Figure 7 illustrates the problem of multimodality. While profiling one of the parameters and with a small set of 330 individuals down-sampled from the NHANES dataset, we noticed our search algorithm underpinning the construction of confidence intervals was unexpectedly failing to find a solution. On closer inspection it was seen that for various values of p_k, the constrained log-likelihood oscillates between two distinct solutions, that are both local maxima. A more detailed search discovered two distinct curves of local maxima, as shown on Figure 7. The open circle is the unconstrained solution that the algorithm first found at p_k^0, and it lies on a solid line connecting local maxima as p_k is moved around. Then, another (dashed) line of solutions exists above it, corresponding to a better (i.e. higher-likelihood) estimate that would otherwise remained undetected. It is very hard to estimate how frequently such a situation occurs, as multimodal likelihood may not raise any warning flags. Even when multimodality is identified it can be considerable work to verify it.

The issue of multimodality can affect the results of the profiling procedure.

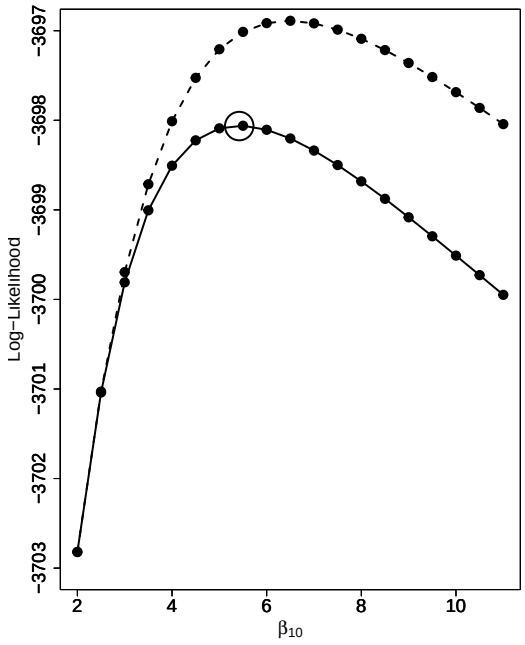


Figure 7: Illustration of the fundamental limitation of applying the NCI method, as only a local maximum is found directly (open circle, solid line), and searching for possibly existing better maxima (dashed line) is very costly. The small circles are data calculated in a profiling procedure for parameter β_{10}. The open circle shows the subpar solution to which the NCI method originally converged.

In searching for the ends of CI of a particular model parameter, several constrained optimizations are performed. If in just one of these runs the algorithm "lands" on a different line of solutions (for example in Figure 7, solution constrained to some value of β_{10} ends on the solid line, but the solution constrained to a slightly different value $\beta_{10} + d\beta_{10}$ ends on the dashed line reporting a sudden jump in log-likelihood), the search algorithm will fit an local parabola in the way that the limiting value of $\frac{1}{2}\chi^2_{1,0.95}$ will never be exceeded, and the CI will be reported as infinite. This can not be distinguished from CI being genuinely infinite due to small sample size. Thus, some CIs calculated by profiling will be incorrectly reported as infinite.

3.5 CDF and percentile estimation

After estimating model parameters from the (global) maximum of the likelihood, we now proceed to estimate the corresponding CDF and percentiles of the intake distribution. We do this in the straightforward manner described in Section 3.3, and using the base R function `ecdf()`. As samples we apply the method on differ in size $(30 - 1700)$, for each CDF estimation we choose the number of virtual individuals per a real individual so that we create a population of 100,000 virtual individuals (up to a round-off error). Some ambiguity remains present in implementing the 9-weights approximation, since we could not find a fully reproducible description of the method in the literature we surveyed [17, 21]. We thus implement and test our own version in the following manner.

Our goal is to estimate the mean of the back-transformed distribution of virtual-individual's mean consumption-day intake, given in Equation (19). On the transformed scale the intakes are normally distributed with mean $\mu_i^{(k)}$ and standard deviation σ_e. Estimating them naïvely, as $g^{-1}(\mu_i^{(k)})$, would introduce bias via Jensen's inequality.

Using the pair $(\mu_i^{(k)}, \sigma_e)$ as input, we set the positions of nine point masses $\{c_j\}$ at nine spread-out percentiles of the distribution of $s_i^{(k)}$, which is assumed to be Normal. The percentiles are

$$\frac{1}{100}, \frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, \frac{99}{100}. \tag{29}$$

The nine weights, $\sum_j \omega_j = 1$, are then calculated so that the first five moments of the discrete distribution $\{c_j, \omega_j\}$ match the first five exact moments

m_l of the fitted distribution of the $s_i^{(k)}$:

$$\sum_{j=1}^{9} \omega_j c_j^l = m_l \,, \;\; (l = 1 \ldots 5) \,. \tag{30}$$

Each weight ω_j is seen to be a linear combination of polynomials of the fifth order. It is convenient to use the base R function `poly()` that produces orthogonal polynomials (O_{jl}) over the specified set of nine positions $\{c_j\}$: $\omega_j = \sum_{l=0}^{5} O_{jl} S_l$, with S_l being the coefficients calculated from conditions in Equation (30).

Using matrix definitions, specifically defining

$$m = \begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_5 \end{bmatrix} , W = \begin{bmatrix} 1 \\ \omega_1 \\ \vdots \\ \vdots \\ \omega_9 \end{bmatrix} , S = \begin{bmatrix} S_0 \\ S_1 \\ \vdots \\ S_5 \end{bmatrix} , C = \begin{bmatrix} 1 & \cdots & 1 \\ c_1^1 & \cdots & c_9^1 \\ \vdots & & \vdots \\ c_1^5 & \cdots & c_9^5 \end{bmatrix} , O = \begin{bmatrix} O_{01} & \cdots & O_{09} \\ O_{11} & \cdots & O_{19} \\ \vdots & & \vdots \\ O_{51} & \cdots & O_{59} \end{bmatrix} ,$$

we therefore obtain nine weights by solving

$$\begin{aligned} m &= C \cdot W = C \cdot O \cdot S \quad , \\ \Longrightarrow \quad W &= O \cdot S = O \cdot (C \cdot O)^{-1} \cdot m \; . \end{aligned} \tag{31}$$

We illustrate this approach in Figure 8. Here we estimate the back-transform of the consumption-day usual intake of a virtual individual having $\mu_i^{(k)} = 15$, $\sigma_e = 3.0$, with Box-Cox parameter set to $\lambda = 0.33$. The left panels shows the inverse function g^{-1} (thick solid line), the Normal distribution of $s_i^{(k)}$ (thin solid Gaussian density), the naïve transformation $g^{-1}(\mu_i^{(k)})$ (thin horizontal line), and the 9-weights approximation result (dashed line). We can see that the 9-weights result is higher, which is expected by Jensen's inequality since g^{-1} is convex.

The right panel illustrates the accuracy of our version of the 9-weights approximation. Here, we represent the Normal distribution of $s_i^{(k)}$ with N sampled points, then calculate the first five moments of a discrete distribution, and set the weights $\{\omega_j\}$ to match these moments. As $N \to \infty$, these approximation approach the exact moments of the Normal distribution. However, for each sample, we can have an exact back-transformation

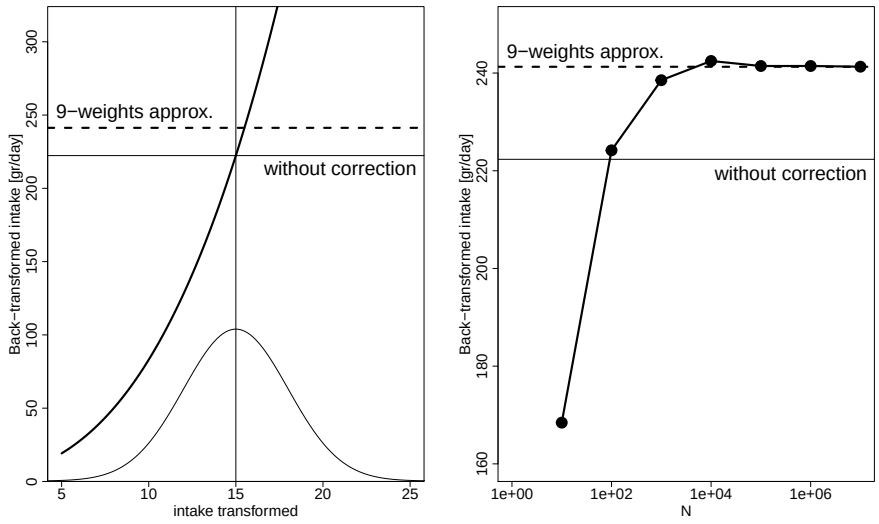


Figure 8: Illustration and test of accuracy of the implemented version of 9-weights approximation, see text.

result of its discrete distribution, by back-transforming each individual observation, then calculating their mean. These exact means are shown in the Figure as solid circles. We can see that as $N \to \infty$, the 9-weights result matches the result of the infinitely large sample of discrete points. Thus in this case the 9-weight approximation stands, in practice, not as an approximation but as an exact calculation.

It is interesting to note that in the original NCI literature [17], besides the 9-weights approximation, another method was concurrently used to address the estimation problem, based on second-order Taylor expansion of g^{-1}. This approach is faster but usable only when the usual intake on the original scale is not excessively skewed. The motivation for this seems to be primarily computational, but since the NCI method was published computing power has made calculating the transformed mean straightforward, we do not pursue this approach here.

4 Performance of the method

The second task of this thesis is to evaluate the performance of the NCI method when sample size N is small, for the purpose of informing researchers on whether to use the model at all. While the usual performance criteria of the method would be of interest, like CI coverage, power for related tests, etc, a more basic step is to work out at what level it is reasonable to start using this model, i.e. when is the fitting possible and there is enough information to say something about the uncertainty in the estimates, even if the inference is not calibrated perfectly.

We apply the method to two different datasets and for three testing scenarios, using down-sampling to illustrate the behavior at small N. In particular, we aim to evaluate:

- The probability that the estimates converge, when fitting the model.
- The behaviour of parameter estimates and corresponding confidence intervals, how much each parameter is informed by the available data, how accurately CLT approximations work, and the extent of any finite-sample bias in estimates relative to their larger-sample values.
- The behaviour of the estimated CDF of usual intake, via its various percentiles and their confidence intervals.

In order to explore the behaviour of the fixed-effects part of the model, we add age as a single covariate in both the intake probability and the intake amount parts, in both datasets used. Together with the five parameters intrinsic to the model, we have the total of nine parameters over which to maximize the likelihood:

$$\{\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \sigma_1, \sigma_2, \sigma_e, \rho, \lambda\}\,, \tag{32}$$

where (β_{10}, β_{11}) are the intercept and the age-slope of the intake probability part, and (β_{20}, β_{21}) are the intercept and the age-slope of the intake amount part of the model.

In the original publication of the NCI method, a simulation study was performed with a small sample size of n=100 individuals, each with seven 24-hour intakes [16]. Model parameters values were chosen, virtual intakes were

generated from the model, and parameters estimated by fitting the model on these datasets. This was repeated 100 times, for each of two sets of parameter values. Mean parameter estimates were, in all cases, close to specified values, and so it was concluded that the NCI method is approximately unbiased.

Similar simulations could be performed to investigate the coverage of the approximate 95% confidence intervals output by the NCI method; computing intervals for each parameter from each simulated dataset and seeing how frequently they cover the known true value of those parameters. As noted above, the intervals could be constructed by bootstrapping, or by use of profile likelihood. At small n the profile likelihood is tractable but (due to lack of information in the data) not likely to work well. In contrast the bootstrap may work better for the regression parameters, but has too high a computational burden for this form of calculation. For these reasons, a check on coverage is not included in this thesis.

4.1 NHANES dataset

The first dataset we analyze is from the publicly-available NHANES study [1], from which we use the data on total fish consumption of a large set of $N = 16,776$ individuals. There are $N0 = 12,570$ individuals with zero consumption on both 24-hour recalls ("zero-hits"), $N1 = 3,587$ individuals with one non-zero consumption recall ("single-hits"), and $N2 = 619$ individuals with two non-zero consumption recalls ("double-hits"). Double-hits are essential for estimation of the within-person intake variance. We fitted the model on this whole dataset, then used those parameter estimates as the starting point for any subsequent fits on subsets of the data.

To examine the assumption that consumption-day intakes are transformed to approximate Normality, we Box-Cox transformed the non-zero recalls using the converged value $\lambda = 0.337$, and made histograms of the untransformed and transformed recalls, and a Normal Q-Q plot of the transformed recalls. These are given in Figure 9. Strong skewness of the original data is apparent, and transforming to Normality seems to hold fairly well, except in tails.

We note a small mode in the left tail of transformed recalls, which somewhat reduces the effectiveness of the Box-Cox transformation in achieving Normality of residuals. This small mode does have practical meaning, however:

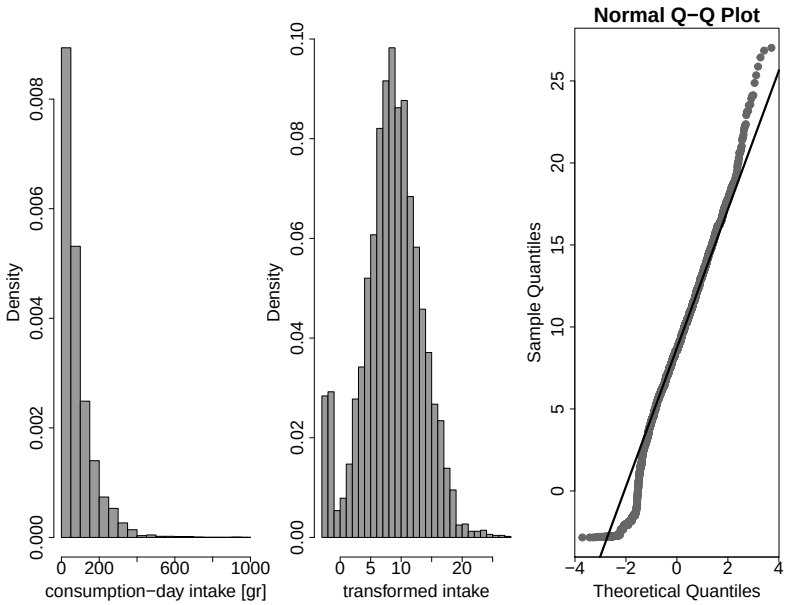


Figure 9: Informal comparison of the transformed NHANES dataset outcomes to Normality.

it consists of a very specific type of intakes of less than 1[gr/day] with fish being present in traces via supplements, pills, etc. Since this is not the part of typical fish consumption for food, in some studies it may be reasonable to introduce a lower cutoff where such minuscule intakes would be treated as zero. However, that would affect the intake probability part of the model, so it should be contemplated with care. As discussed previously in Section 3.2, a basic assumption of the model is that the reported amounts are not misclassified. We proceed by including all the data as given.

4.2 Hits-selection scenario

In the first scenario, we separate zero-, single-, and double-hit individuals into distinct groups. We explicitly set the counts $(N0, N1, N2)$ in each group, then construct a sample to fit the model on a subset of the data, randomly sampled with replacement from the three groups that constitute the original data.

In this way we explore the 3-dimensional $(N0, N1, N2)$ space in the following

manner: we explore one dimension NX ($X = 0, 1, 2$) at a time while keeping the other two constant. The NX values are explored in two ranges, in each case with values increasing in steps of 10:

Range 1 ($NX = 10 - 200$); at each value of NX we fit 100 samples; we take means of model parameters and CDF percentiles as their point-estimates, and their central 95% inter-percentile ranges as estimates of variation; we estimate convergence as fractions of converged runs at each NX;

Range 2 ($NX = 200 - 1500$); at each value of NX we fit four samples, and present model parameters with their various point-estimates, each having a confidence interval calculated by profiling.

In addition, in both ranges we present the observed Fisher information from all individual sample runs; we also plot estimates of variability: widths of confidence intervals in range 2 or inter-percentile ranges in range 1.

As a specific NX dimension is being explored, we first keep the other two groups size to be 10 (notation: "10-10"). We then repeat the whole procedure with their size set to be 50 (notation: "50-50"). Finally, we repeat again with their size set to be 100 (notation: "100-100").

The results are given in several figures. First we show the convergence rates in Figure 10, as line connected dots and with s horizontal line marking perfect 100% convergence serving as eyeguide. We show these results of all the repeats: 10-10, 50-50, and 100-100 in a single plot. NX explores only the first range up to 200, as only here we ran enough repetitions per NX value (100) to have a reliable estimate.

Next in Figure 11 we present model estimates (the upper nine panels) and their CI widths/inter-percentile ranges (the lower nine panels), with ($N0 = NX, N1 = 10, N2 = 10$), as noted in the caption. The two ranges of NX exploration are separated at 200 by a thin vertical line. In the upper nine panels, parameter point-estimates are black dots and their CIs/inter-percentile ranges are in gray. In the lower nine panels, CI widths/inter-percentile (IP) ranges are given as grey dots, while black lines are loess curves added as eye-guides and fitted separately in the two rages of NX. At the separation line of $NX = 200$, we change the way estimates of variability are calculated (as inter-percentile ranges versus by profiling), so if for a particular parameter CLT has not been established at that sample

size, the two loess curves may exhibit a large discontinuity. For a parameter that is well-informed by the data, the CLT does not appear the be grossly violated, and this discontinuity largely disappears. For some parameters, some ends of their CIs can attain their suprema/infima values e.g. -1, 0, or 1. Finally, we need to observe how estimates of variability based on inter-percentile ranges behave for larger samples too. For this purpose, we chosen three particular values $NX = 500, 1000, 1500$ where we fit 100 samples each, and present their inter-percentile ranges with a three large open dots with a black center.

In Figure 12, we present observed information (the upper nine panels) and CDF percentiles (the lower nine panels), with ($N0 = NX, N1 = 10, N2 = 10$), as noted in the caption. In the upper nine panels, observed information is given as grey dots; black lines are loess curves added as eye-guides and fitted separately in the two rages of NX that are separated at 200 by a thin vertical line. In the lower nine panels, percentiles are given as point-estimates (black dots) and inter-percentile ranges (grey lines), and only cover the first range of $NX \leq 200$.

Figures 13-28 continue this pattern, covering all the combinations of repetitions $(10-10), (50-50), (100-100)$, and cases where $NX = N0/N1/N2$.

Finally, it is of interest to note that in making Figures 11-28, some more model fits were excluded. Although all the runs finished with the `code`=1, signifying that a local minimum is found, occasionally a reported parameter value or its calculated estimate of variability would present a far outlier, being more than 8σ away from the other points at the same sample size. We opted to remove such cases when they are obviously standing out from the similar data, and their frequency was roughly estimated between $0.1-1\%$. It is not quite clear why these, relatively rare, events occur.

Similar outliers happen more frequently with calculated observed information. Estimating the Hessian accurately is known to be a numerically challenging task [27], and here the outliers may span tens of orders of magnitude in value so including these in a figure would mask the relevant details of the bulk of the useful data. When a parameter is well determined at a particular sample size, the frequency of outliers of this sort is roughly $1-2\%$. However, in a situation where a model parameter is poorly determined by the data, as much as 5% of calculated Hessian values may scatter uncontrollably over many orders of magnitude, while the rest of calculated points stay close together in value. We have opted to scale the Y-axis of our plots to exclude

these points, but also still retain the useful information present in the bulk of the datapoints.

A few selected random checks suggested that calculated Hessians are mostly positive definite. However a few cases were found where this was not so, and observed information for some parameter was reported negative. These cases were, together with the outliers, removed from all figures. In this respect, `hessian()` function from the `numDeriv` package performed better then `nlm()` function, by returning negative information less frequently. However, neither of them performed flawlessly. For this reason we avoided using calculated Hessians beyond observing their general trends with sample size, and as a starting point in profiling algorithm. For example, we did not consider Wald CIs based on the Hessian.

4.3 Frequency-selection scenario

In the second scenario of Section 4.1, we explicitly set the frequency of consumption and the total sample size. For the reason of simplicity, we start by assuming that each instance of consumption in a population has a fixed probability p. We again separate zero-, single-, and double-hit individuals into distinctive groups. We then sample a set of N i.i.d. random variables $X_i \sim \text{Binom}(2, p)$, and set the numbers $(N0, N1, N2)$ equal to the numbers of variables in the sample $\{X_i\}$ having $X_i = 0, 1, 2$. We now construct a sample from the NHANES population by randomly choosing with replacements $(N0, N1, N2)$ individuals from the groups of zero-, single-, and double-hitters, respectively.

This down-sampling method has the disadvantage of making the intake probability between-person variance σ_1 indistinguishable from zero, as this parameter is informed by the relations between the numbers of hitters $(N0, N1, N2)$, and here they are explicitly chosen by assuming that each person has the same p. However, other parameters in the model are still allowed to be informed by various aspects of the data, and since $(N0, N1, N2)$ vary uniformly with the size N, it would be expected that model parameters and CDF percentiles do not change much in point-estimates by N, however their estimates of variability will, which is the main expected benefit of this scenario.

The maximum likelihood estimator of p using the set $\{X_i\}$ is $\hat{p} = \frac{N1+2N2}{2N}$,

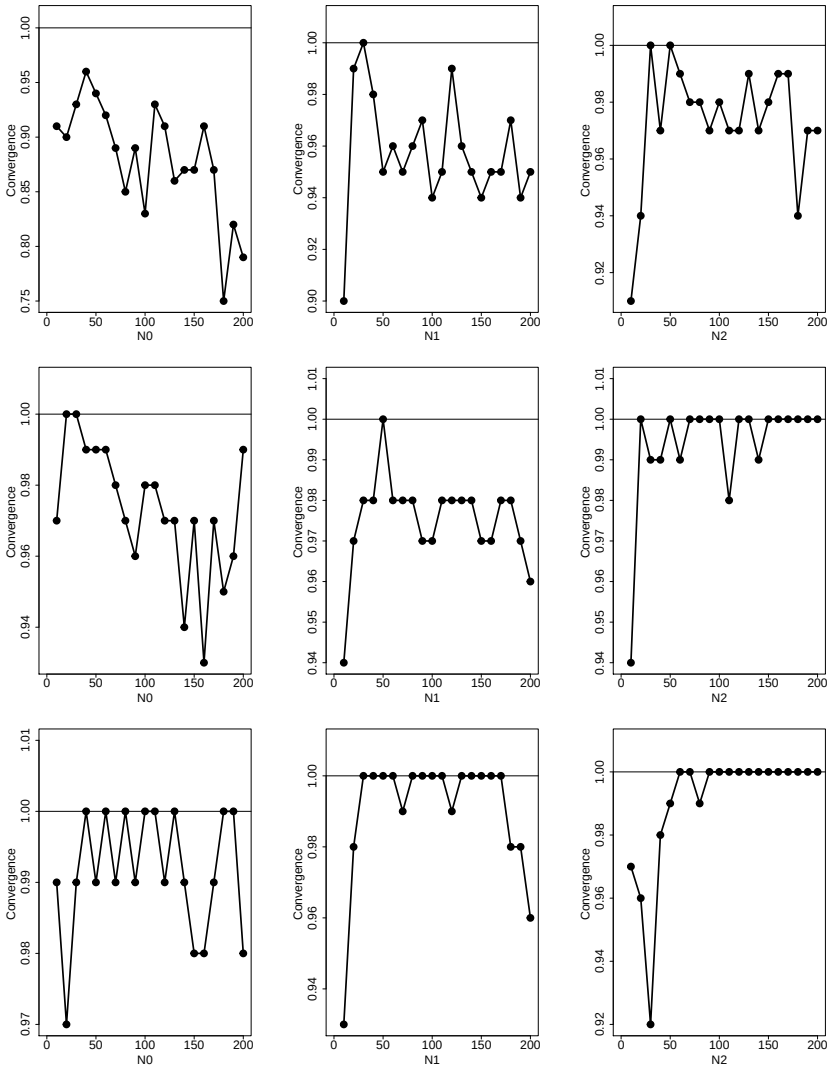


Figure 10: Convergence ratios: top row (10-10), middle row (50-50); bottom row (100-100).

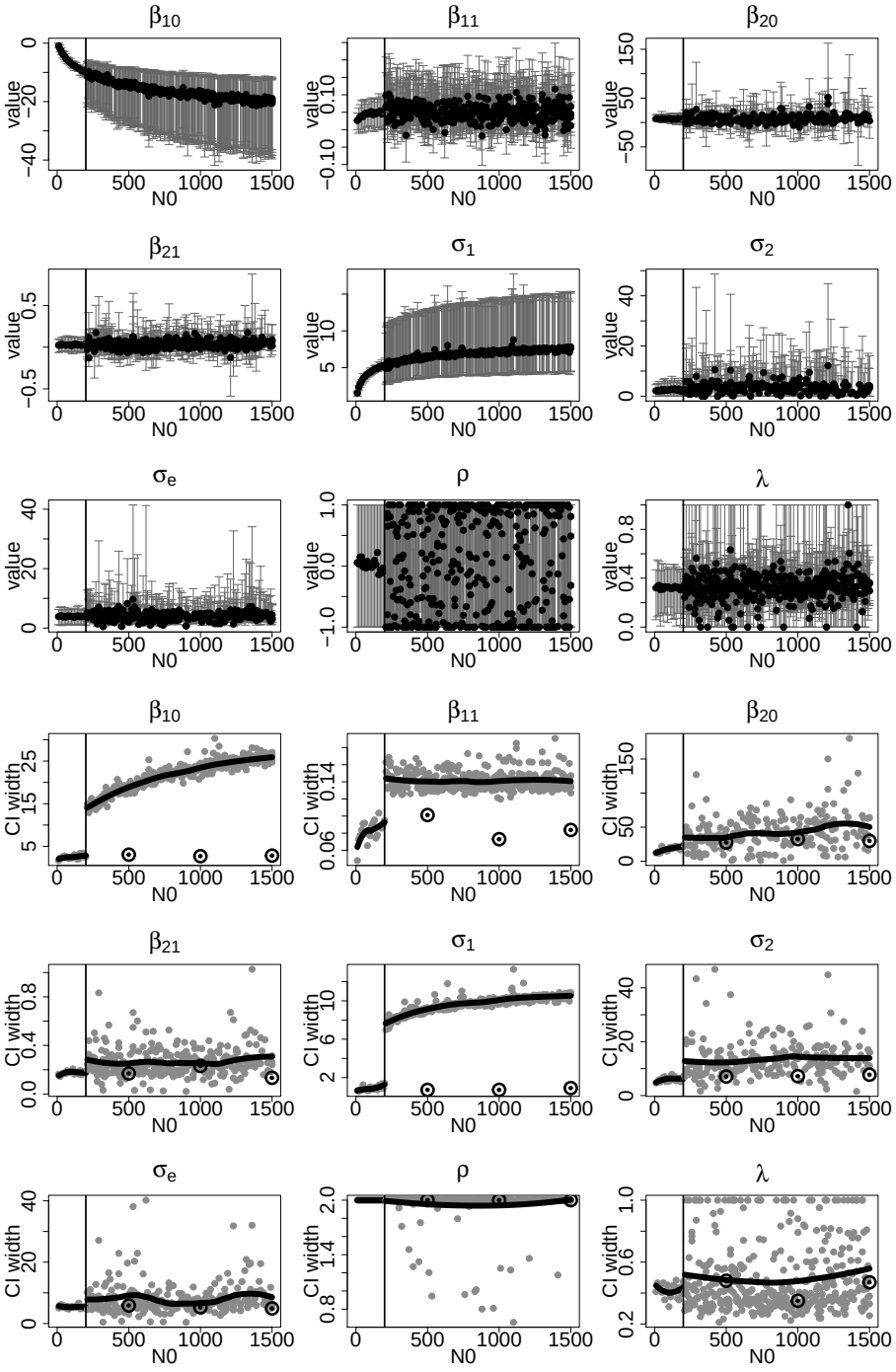


Figure 11: Parameter values & CI widths/IP ranges: NX=N0 (10-10)

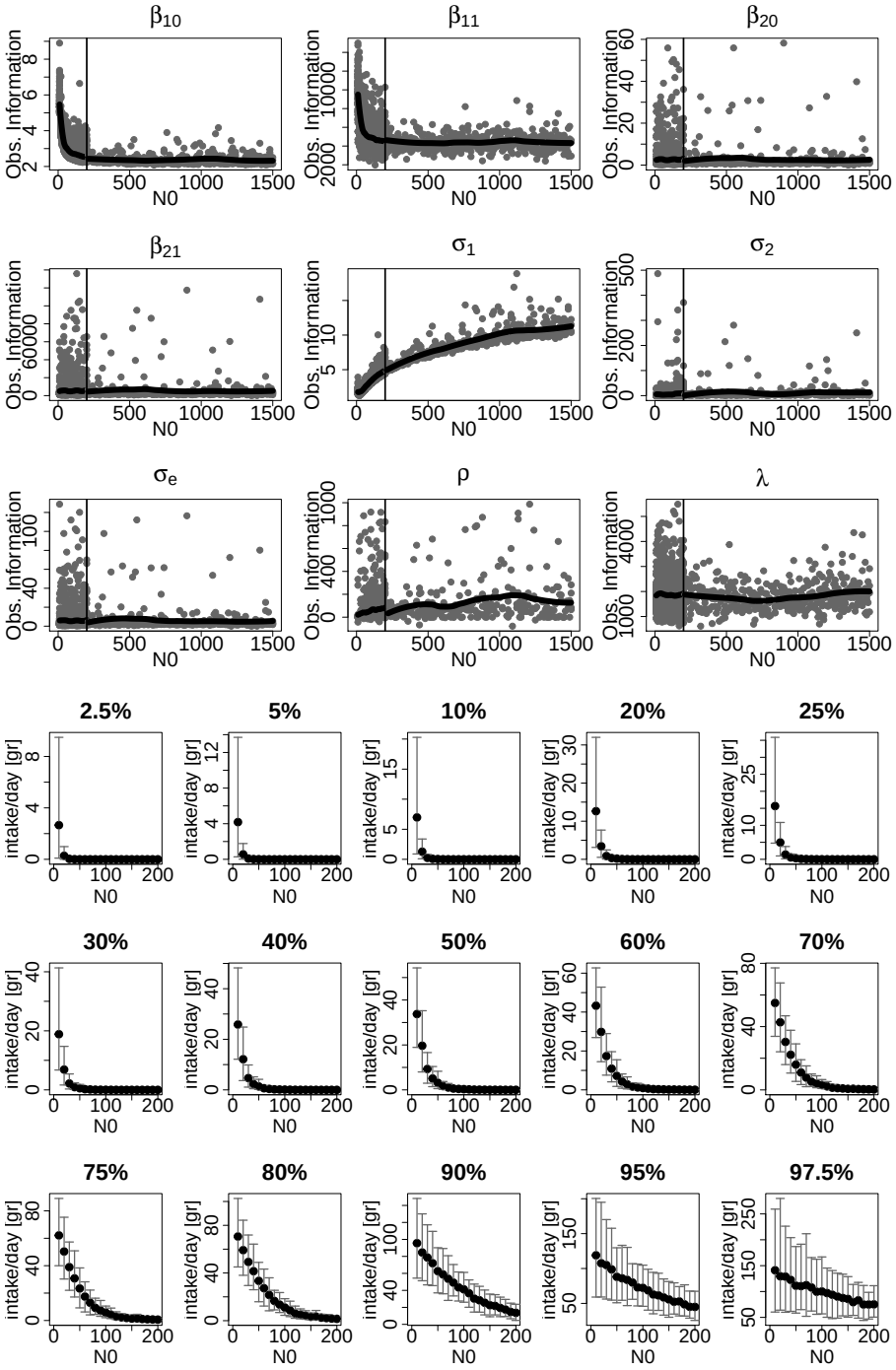


Figure 12: Observed information & CDF percentiles NX=N0 (10-10)

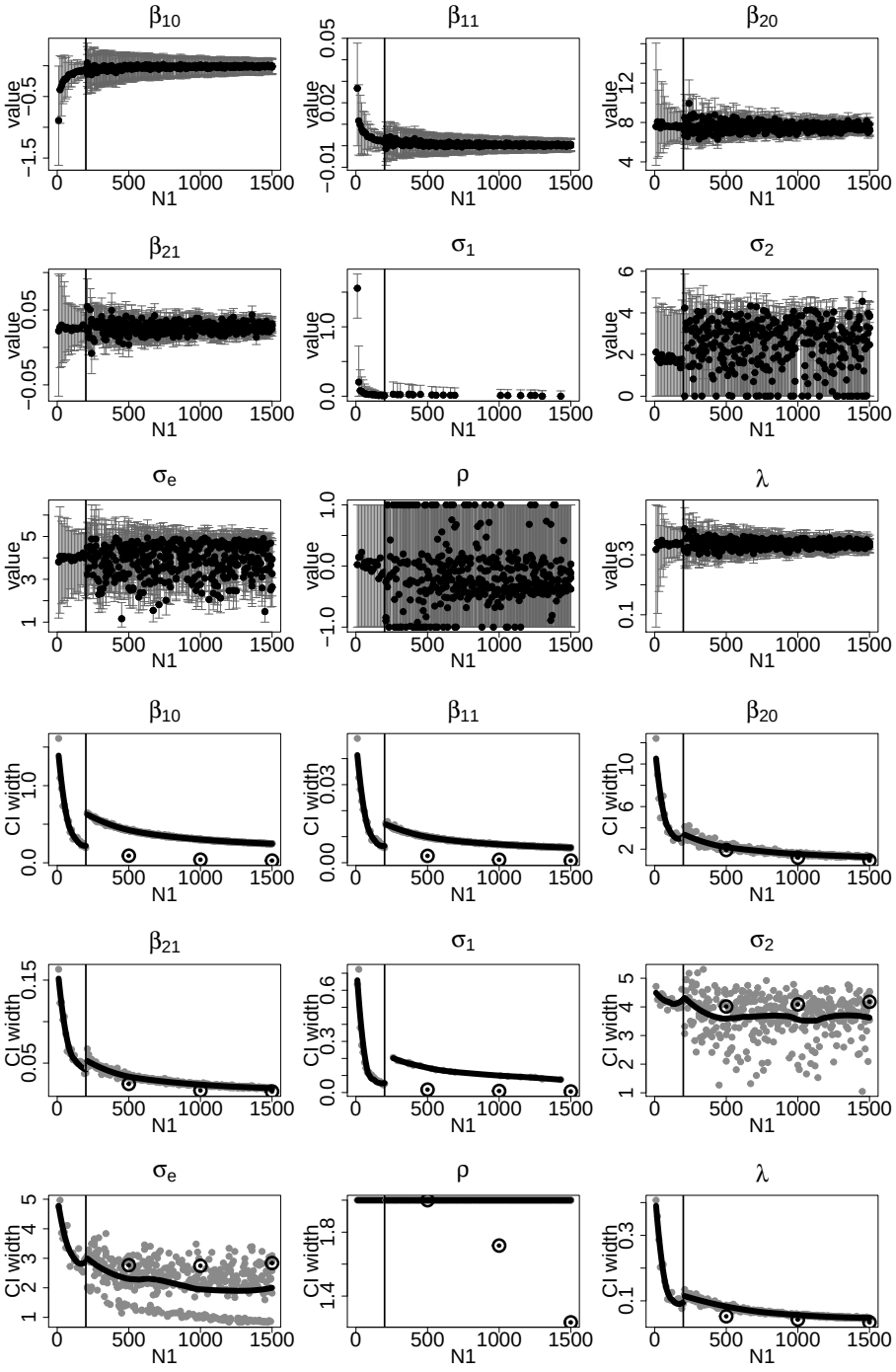


Figure 13: Parameter values & CI widths/IP ranges: NX=N1 (10-10)

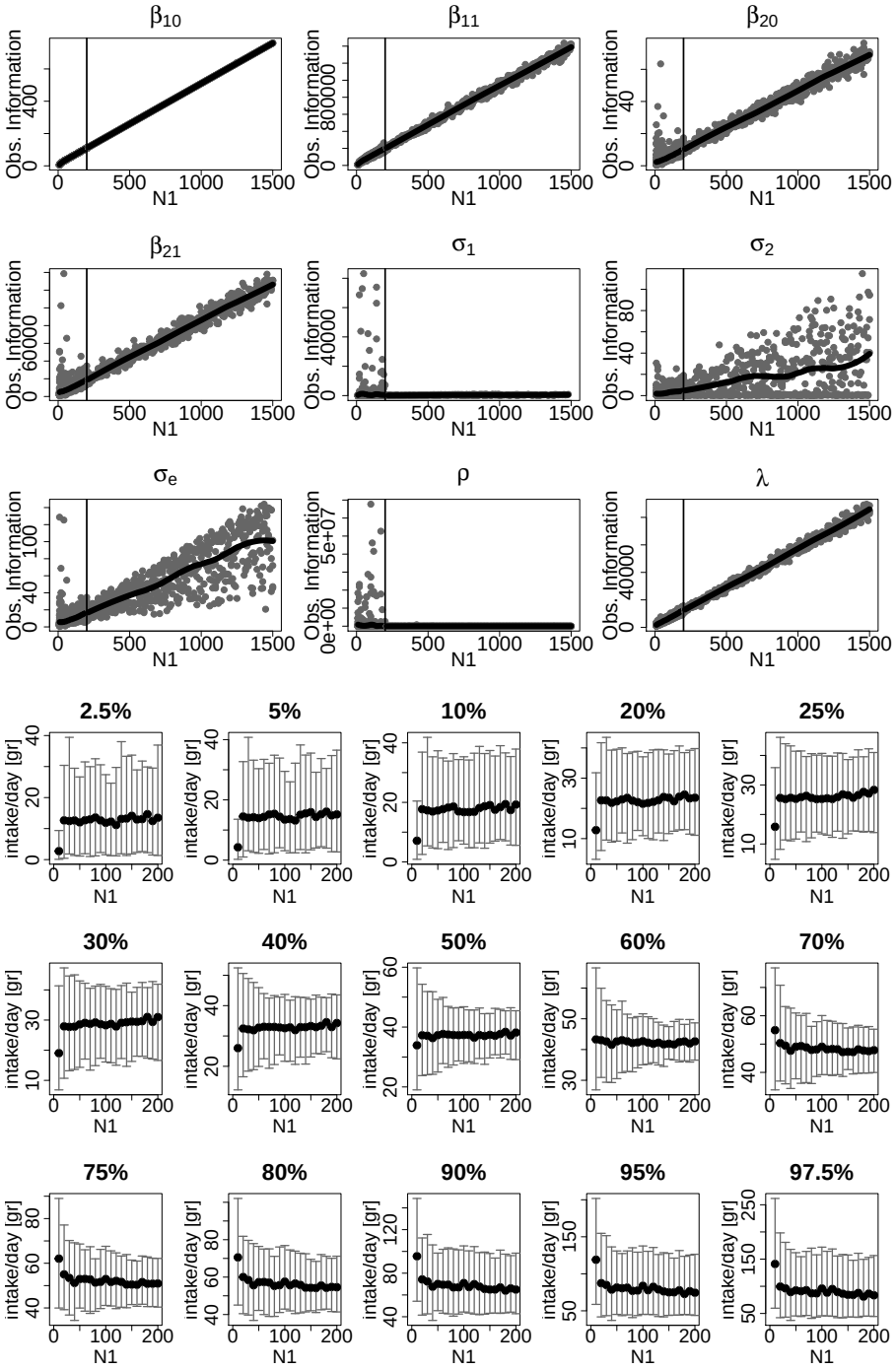


Figure 14: Observed information & CDF percentiles NX=N1 (10-10)

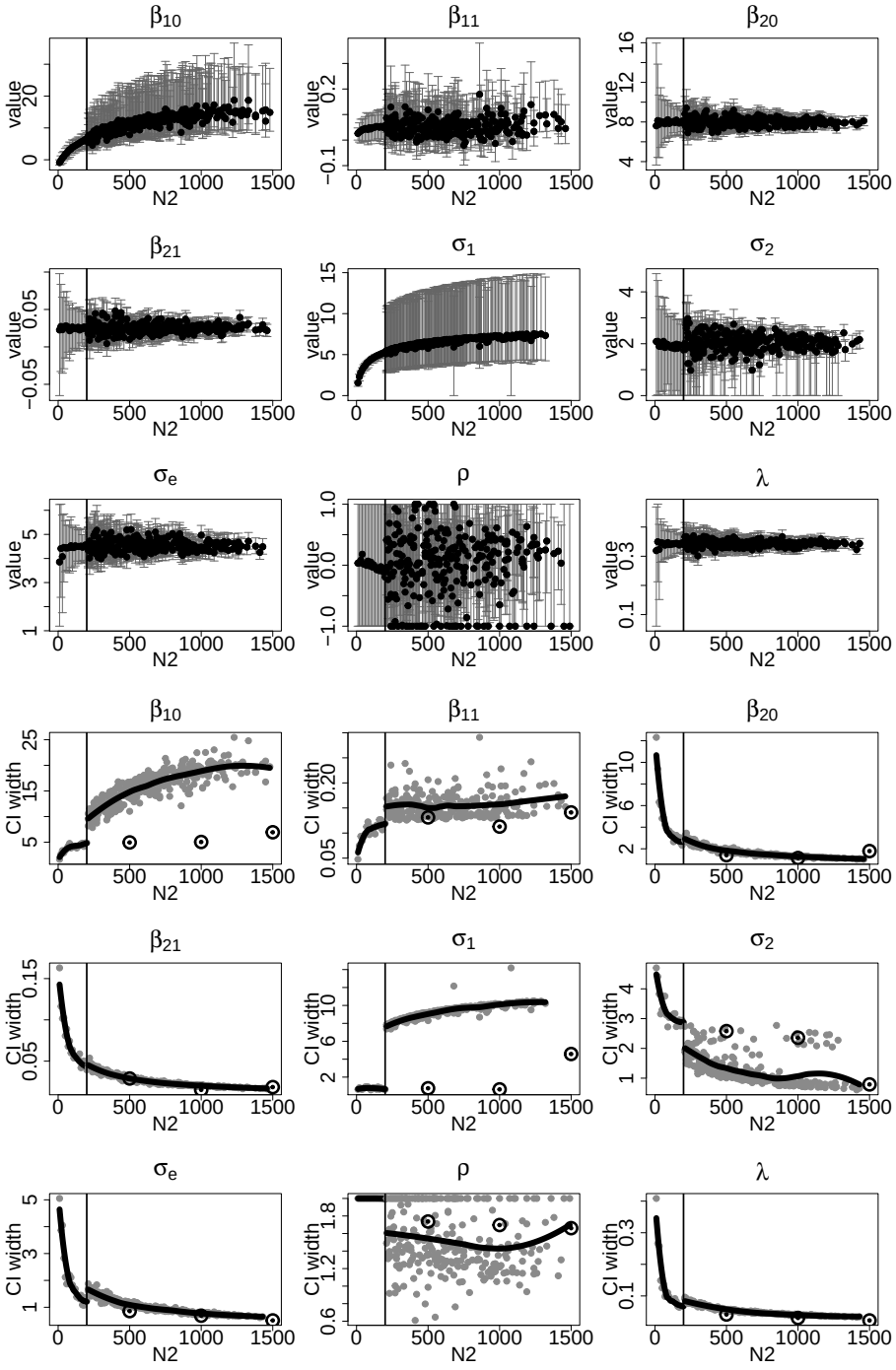


Figure 15: Parameter values & CI widths/IP ranges: NX=N2 (10-10)

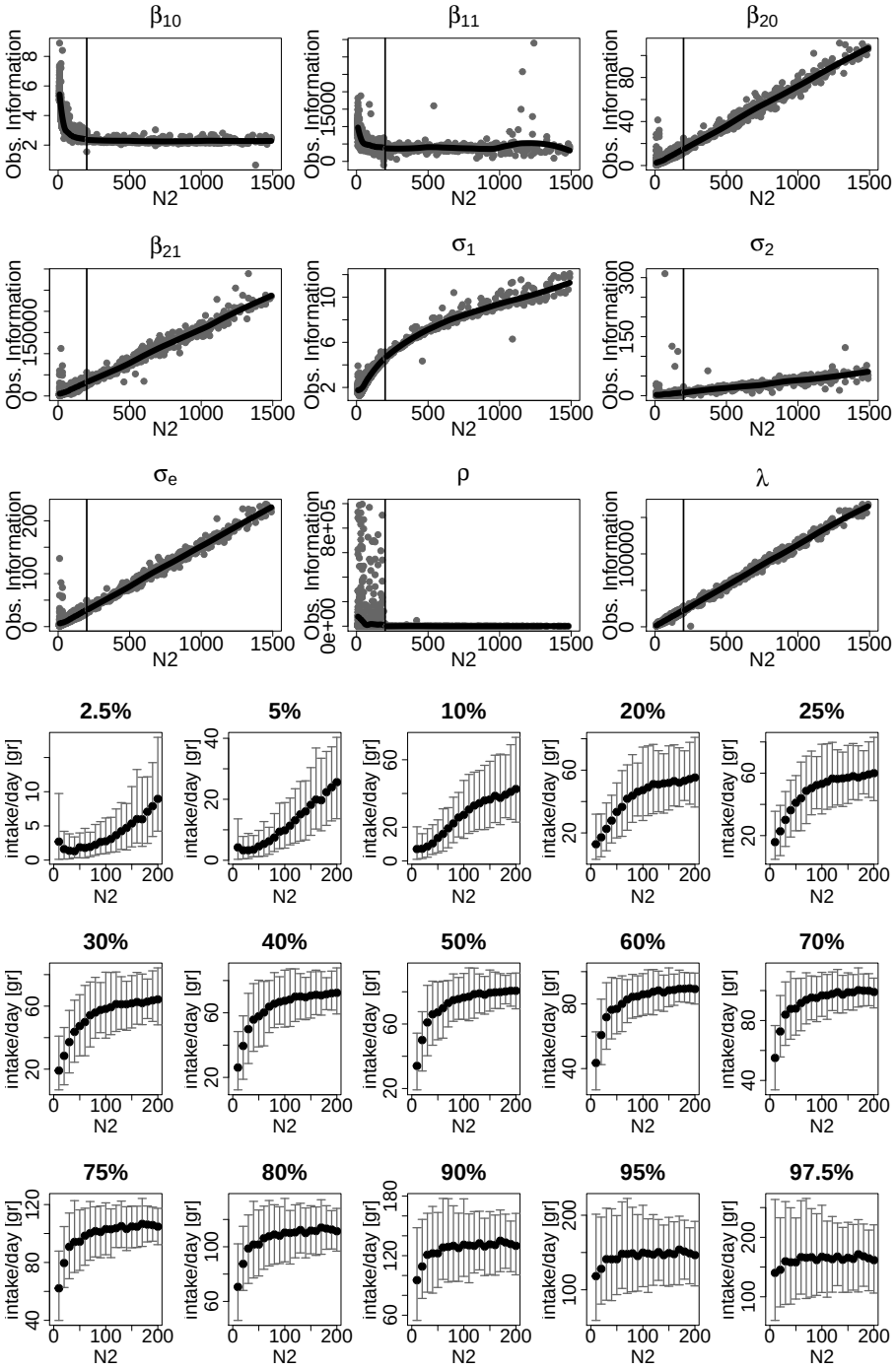


Figure 16: Observed information & CDF percentiles NX=N2 (10-10)

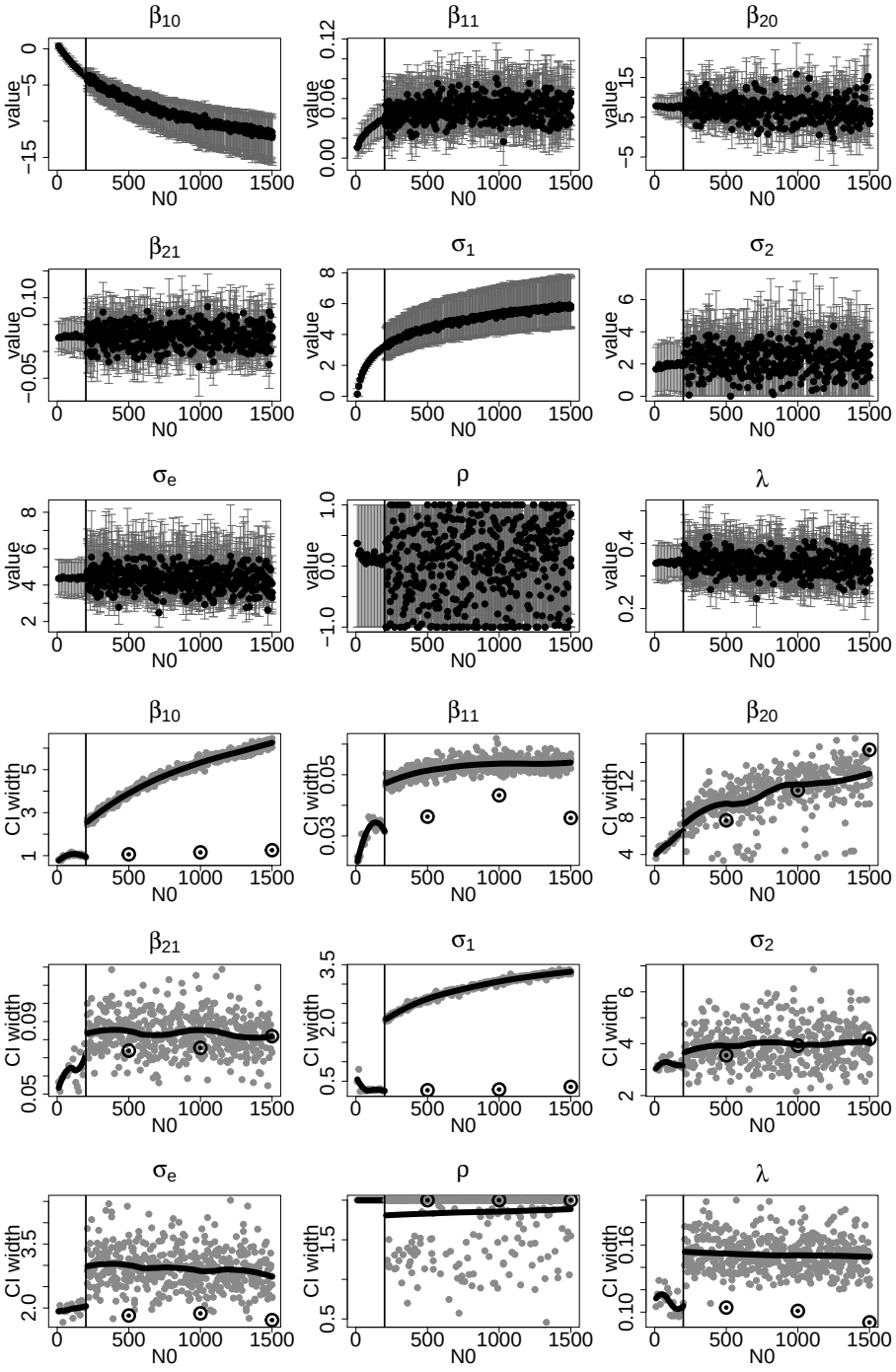


Figure 17: Parameter values & CI widths/IP ranges: NX=N0 (50-50)

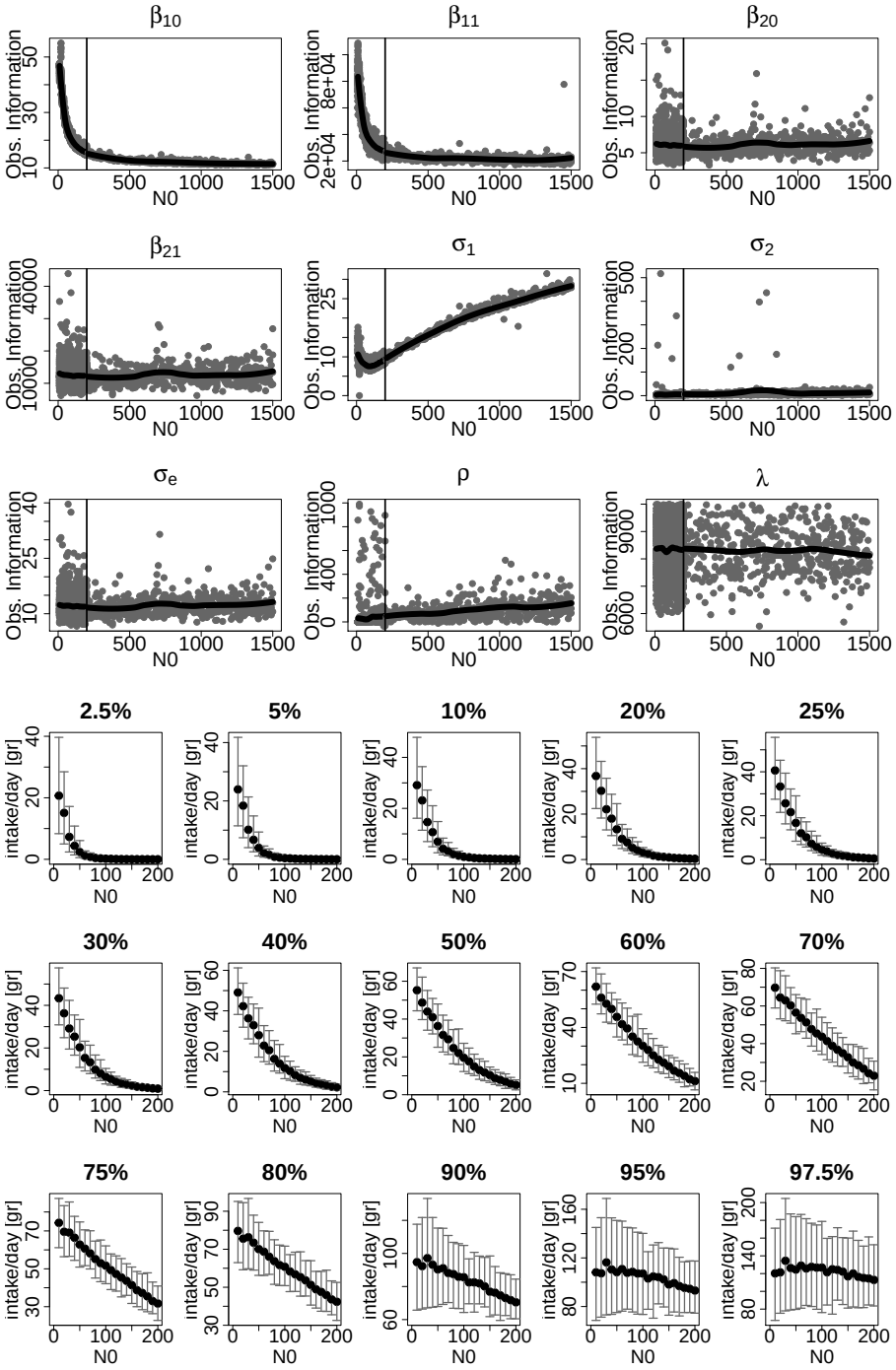


Figure 18: Observed information & CDF percentiles NX=N0 (50-50)

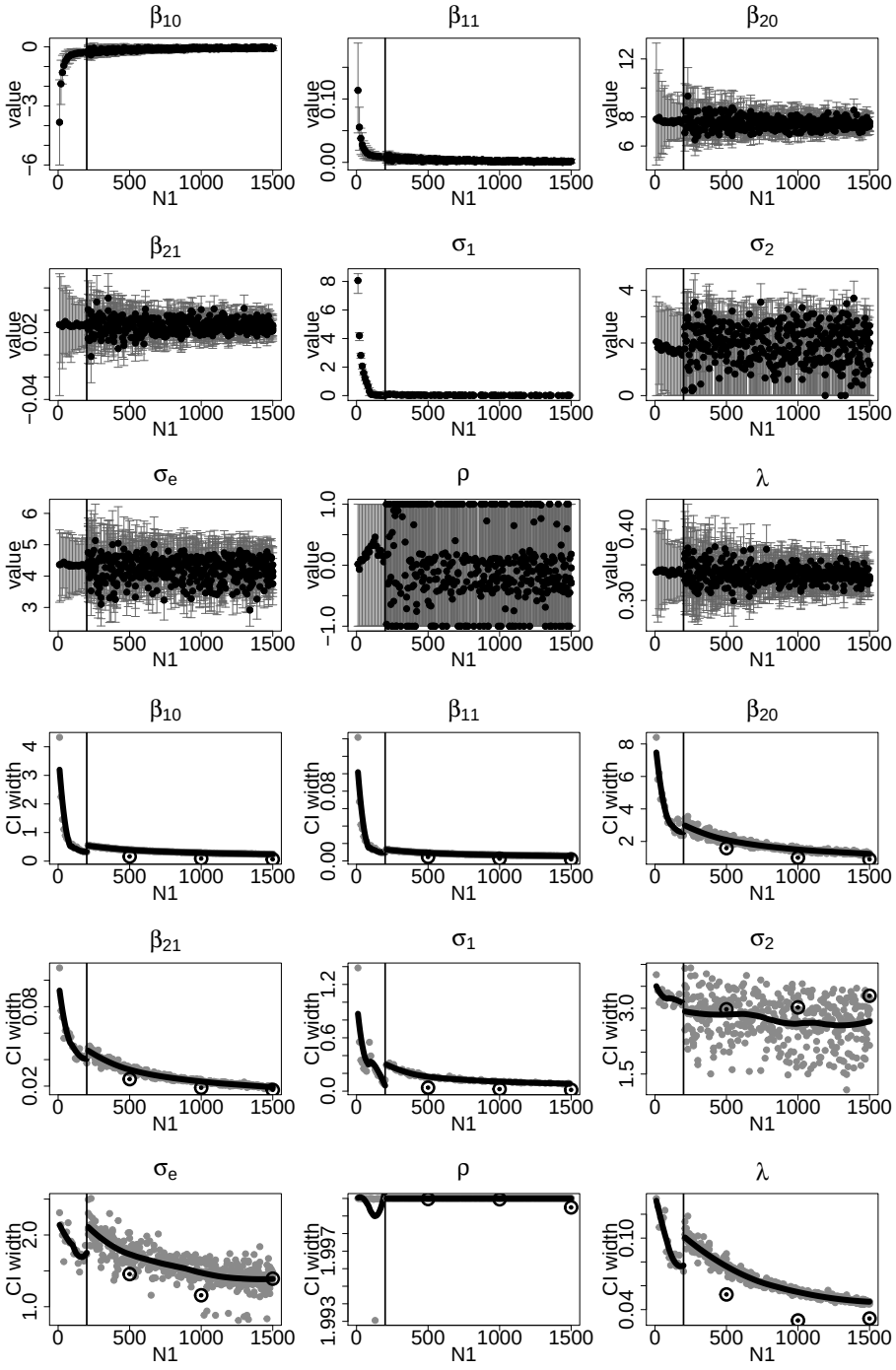


Figure 19: Parameter values & CI widths/IP ranges: NX=N1 (50-50)

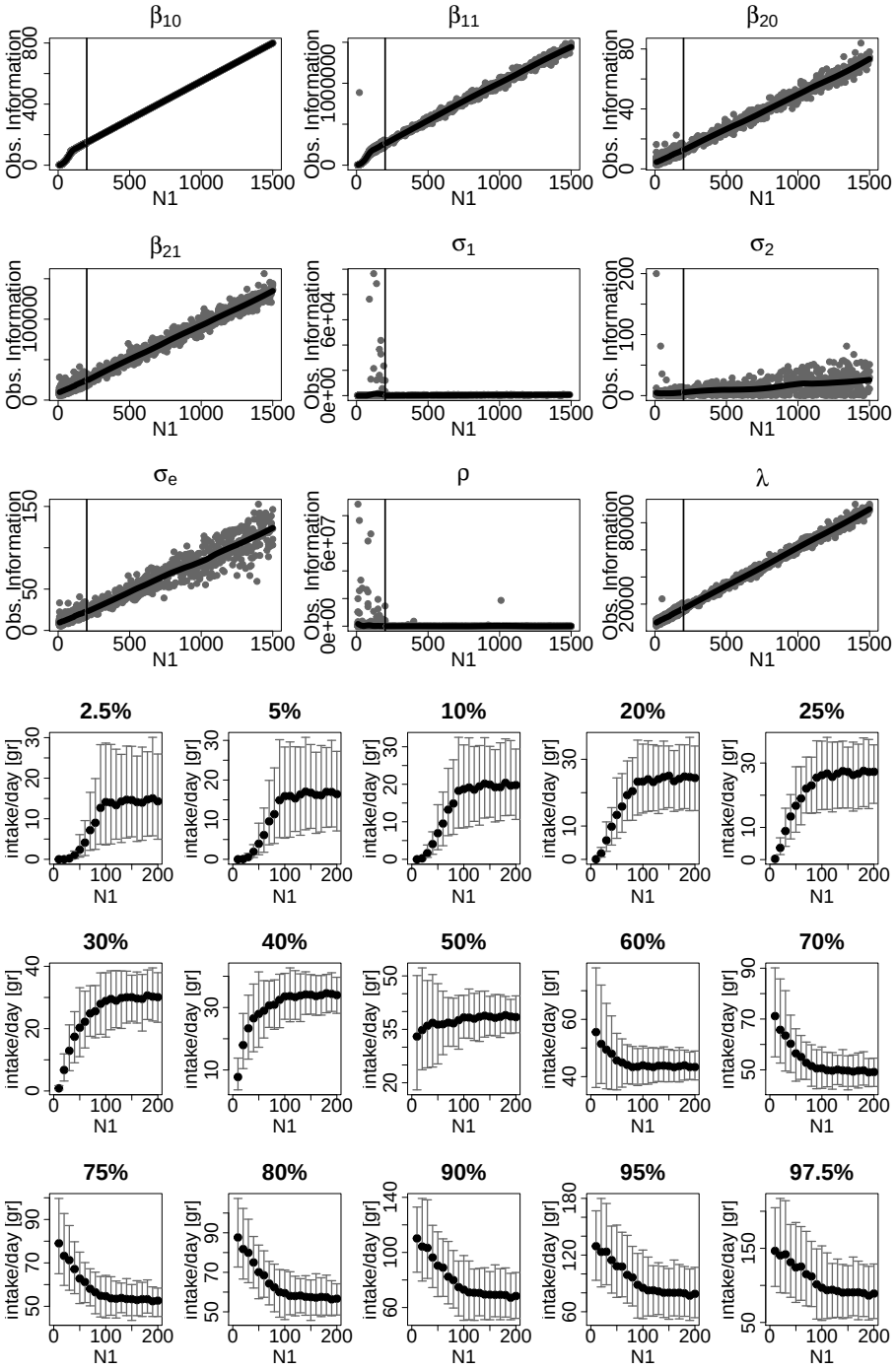


Figure 20: Observed information & CDF percentiles NX=N1 (50-50)

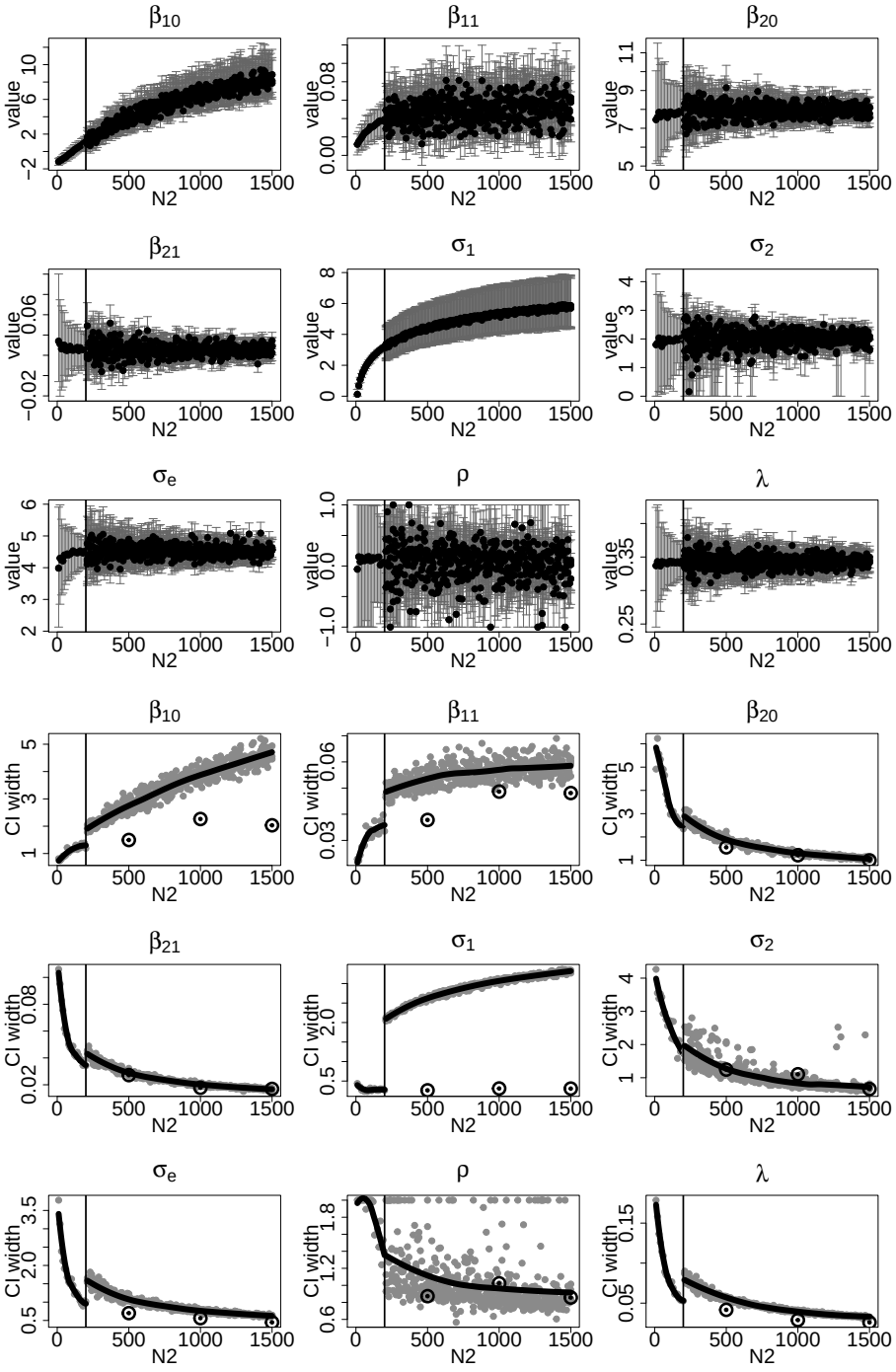


Figure 21: Parameter values & CI widths/IP ranges: NX=N2 (50-50)

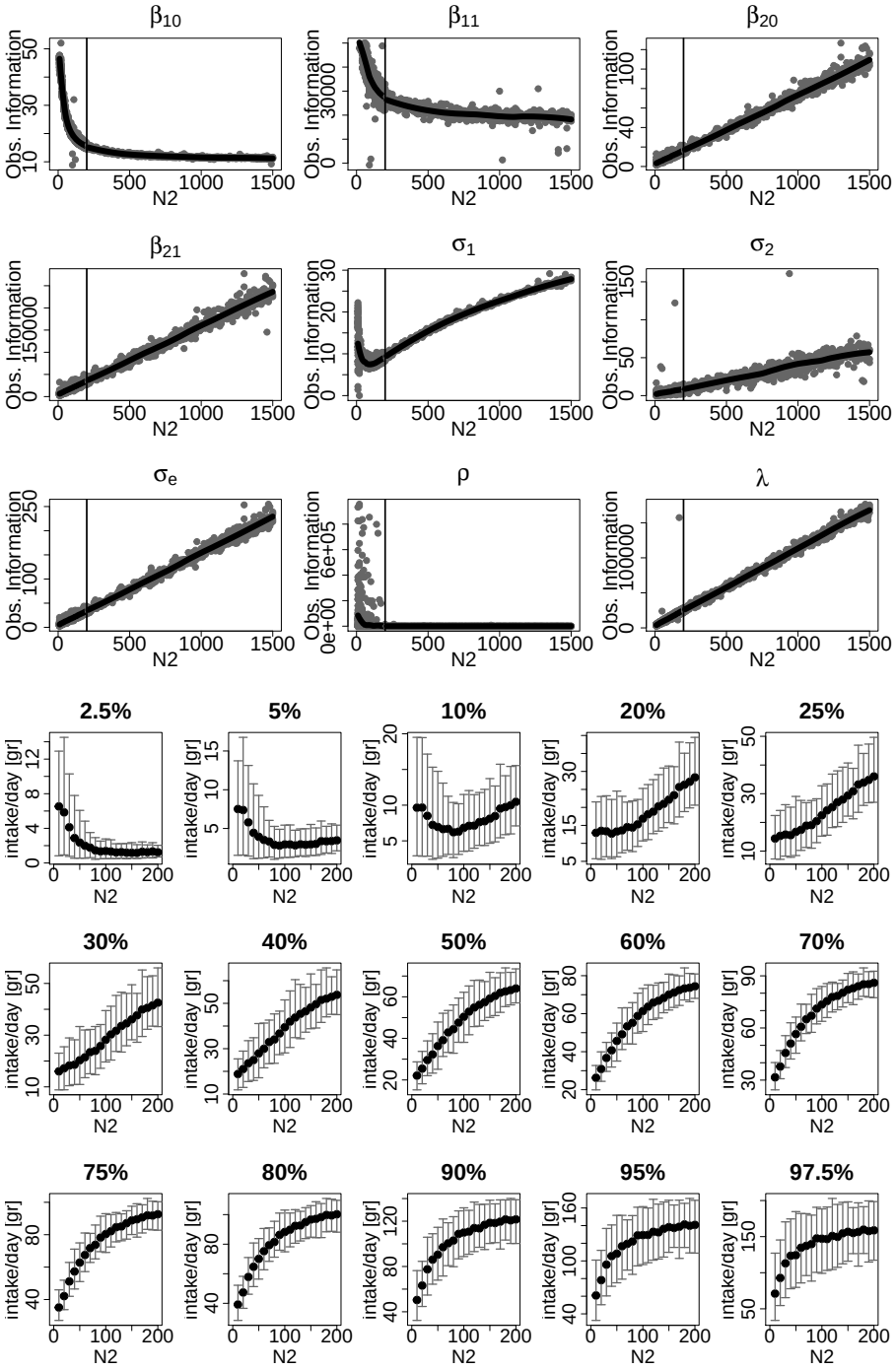


Figure 22: Observed information & CDF percentiles NX=N2 (50-50)

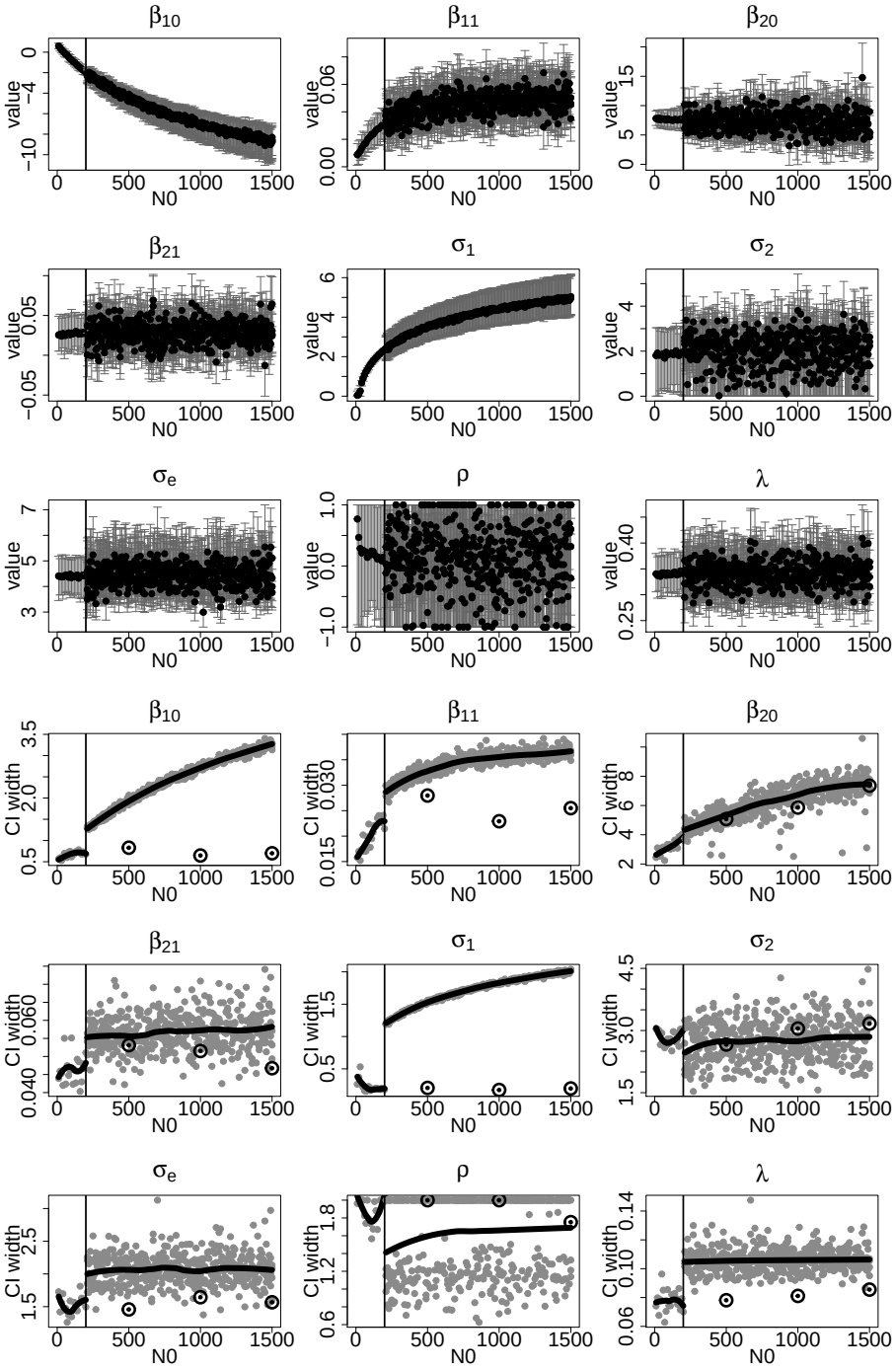


Figure 23: Parameter values & CI widths/IP ranges: NX=N0 (100-100)

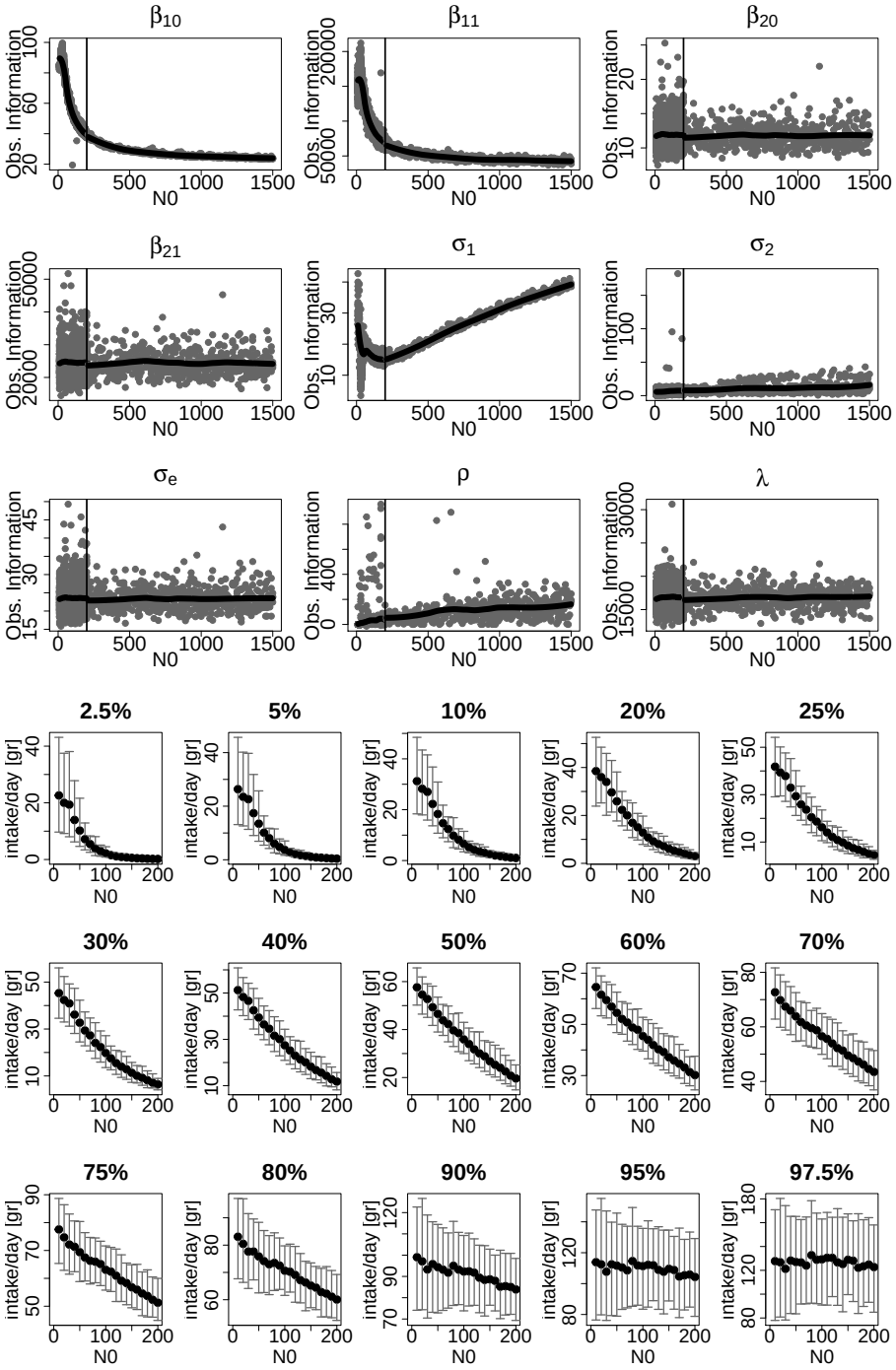


Figure 24: Observed information & CDF percentiles NX=N0 (100-100)

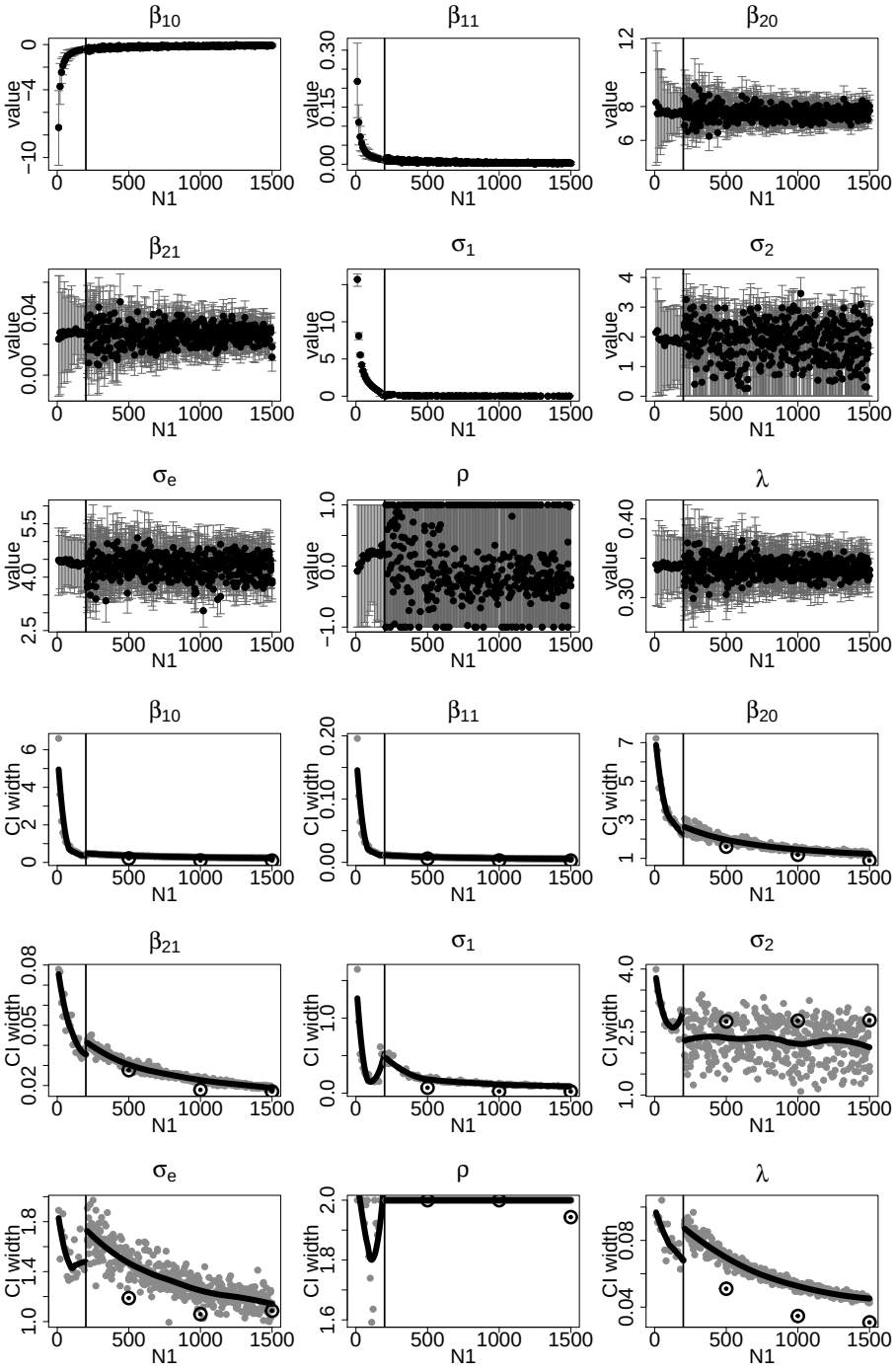


Figure 25: Parameter values & CI widths/IP ranges: NX=N1 (100-100)

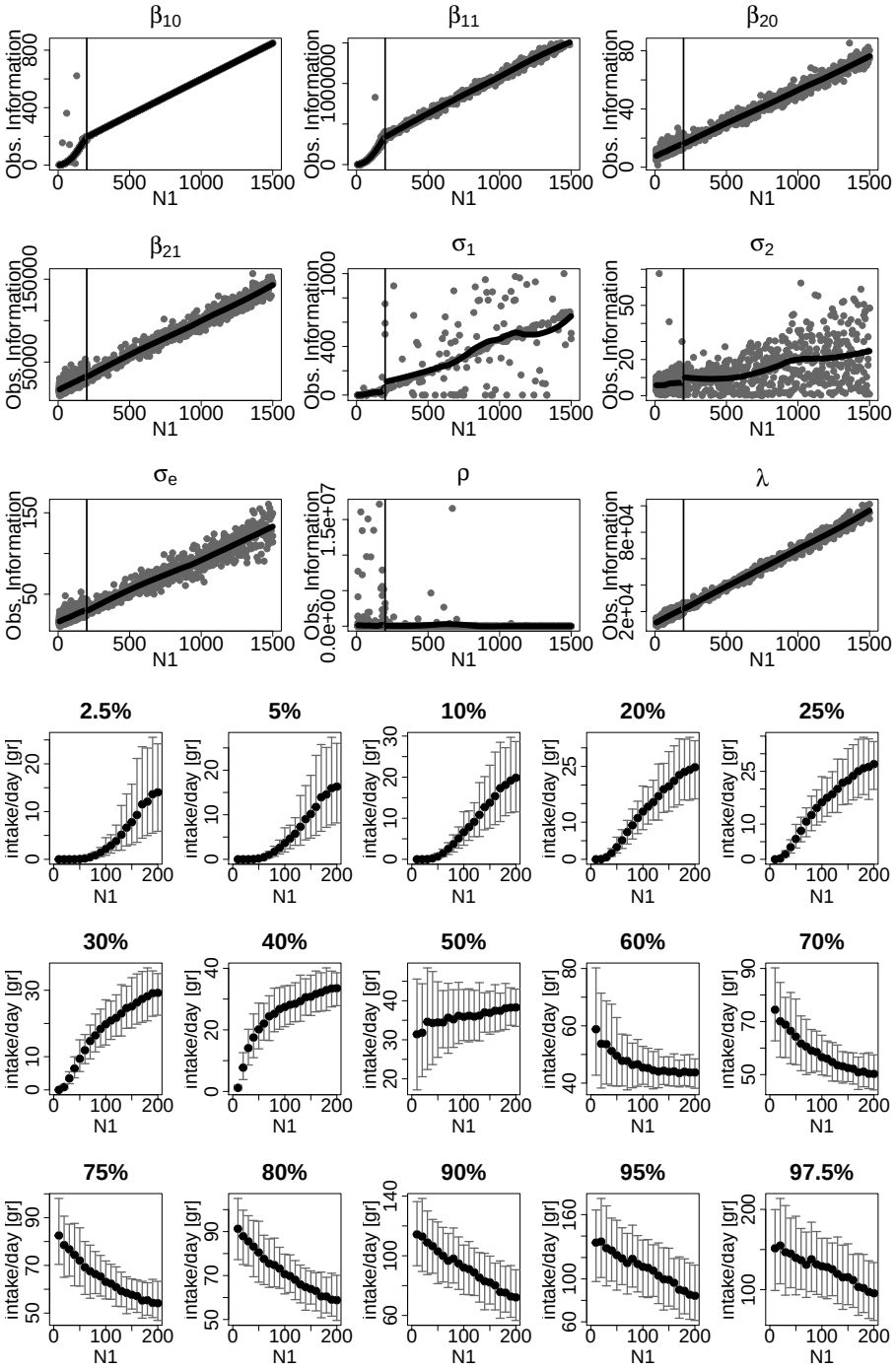


Figure 26: Observed information & CDF percentiles NX=N1 (100-100)

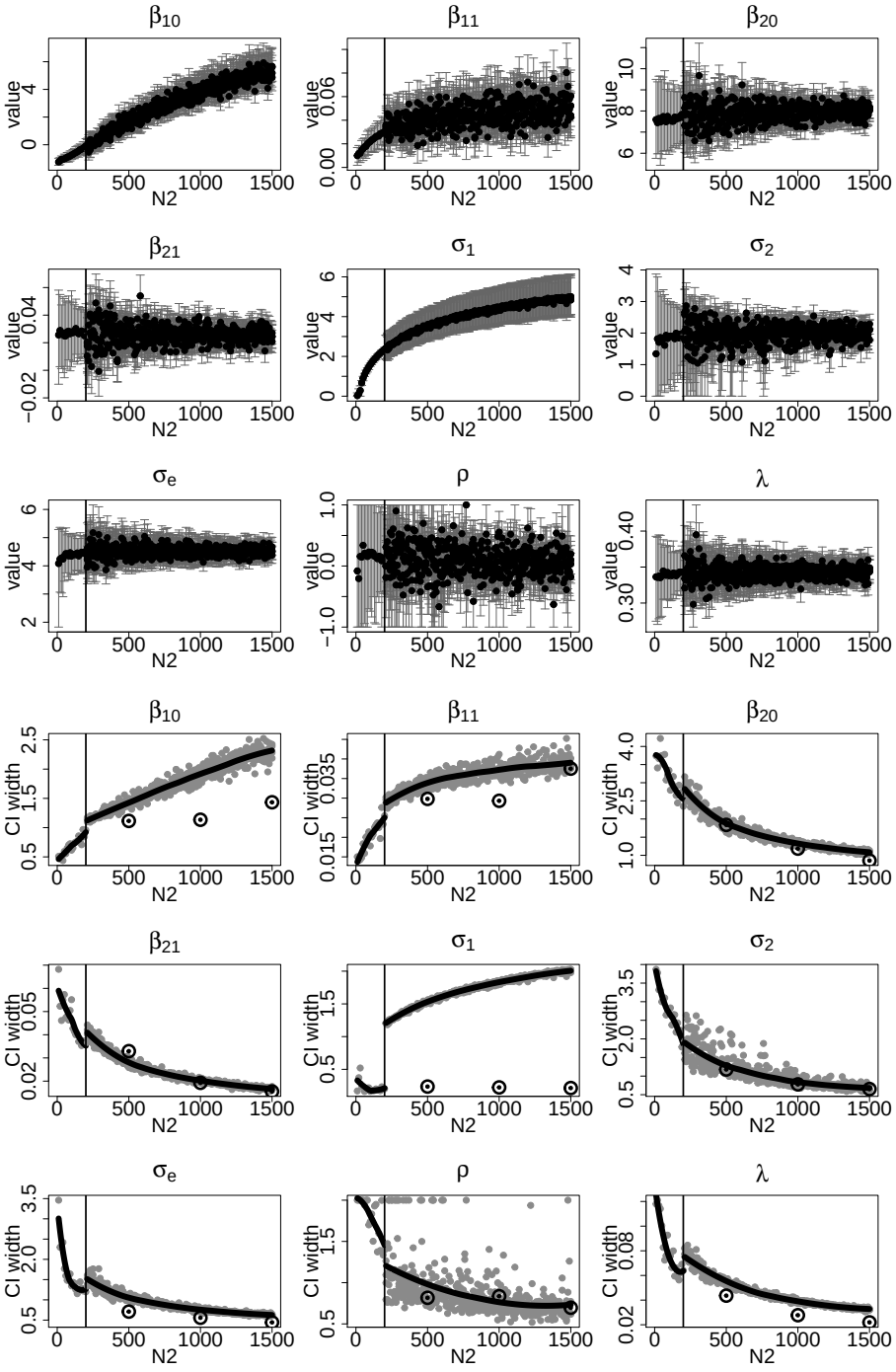


Figure 27: Parameter values & CI widths/IP ranges: NX=N2 (100-100)

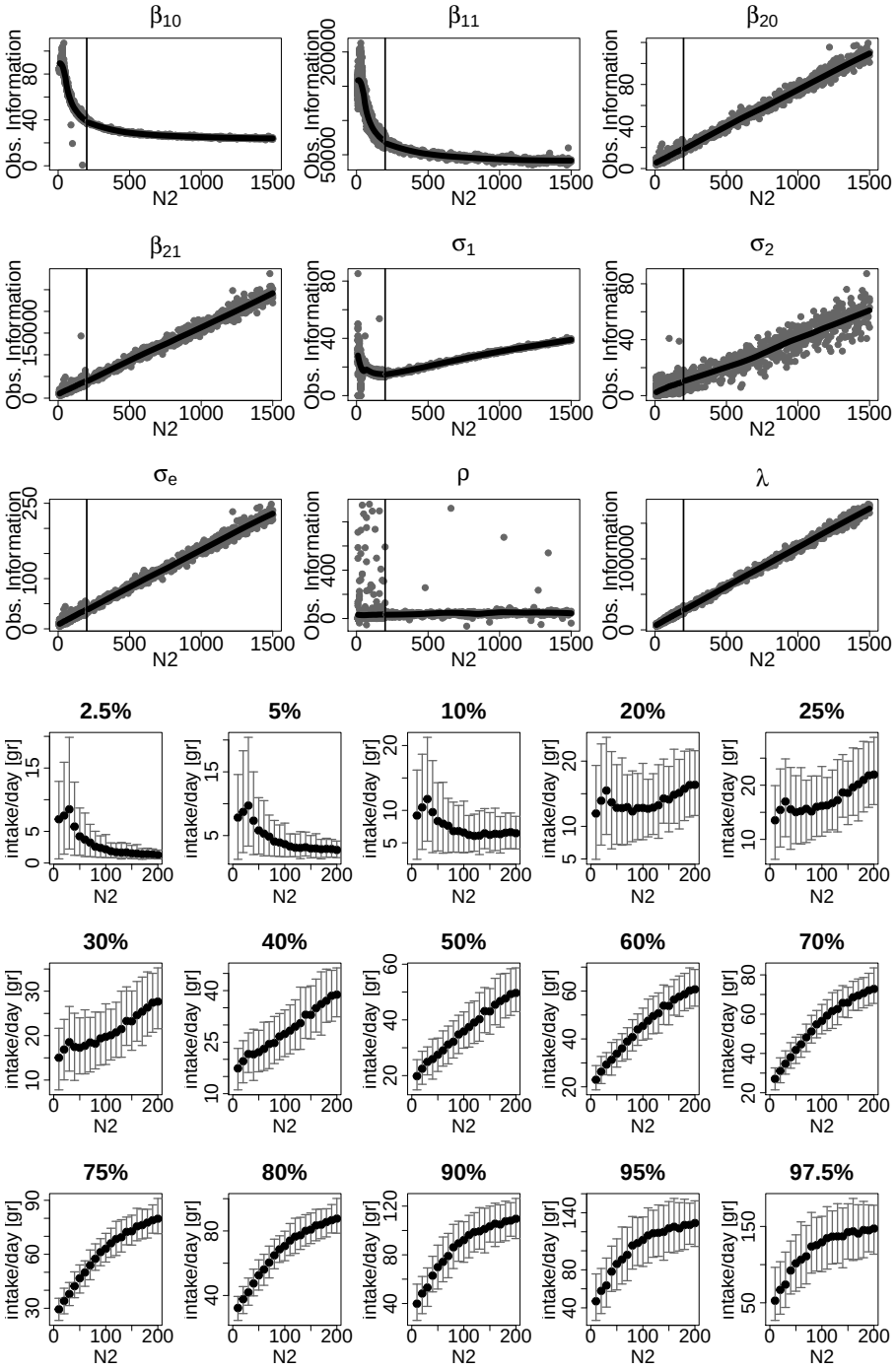


Figure 28: Observed information & CDF percentiles NX=N2 (100-100)

and from the known $(N0, N1, N2)$ values for the NHANES dataset, $\hat{p}_{NHANES} \approx 0.14$. Estimating the same value in the Nez Perce dataset, to be introduced in detail in Section 4.4, $\hat{P}_{Nez\,Perce} \approx 0.26$. Guided by these two examples of real consumption, we chose to observe this scenario at three values of consumption probability: $(0.10, 0.25, 0.50)$.

We explore the sample size N with the range $20 - 1500$. For the interval $20 - 200$, N is increasing in steps of 20, and for the interval $200 - 1500$, N is increasing steps of 50. At each value of N, we fit 100 samples, and take means of model parameters and CDF percentiles as their point-estimates, and their central 95% inter-percentile ranges as estimates of variability; we estimate convergence as fractions of converged runs at each N.

We present the results in several figures. First, in Figure 29, we show the convergence ratios for the three p values in, as line-connected dots and with horizontal lines marking perfect 100% convergence serving as visual guides.

Next, in Figure 30, we present parameter estimates (the upper nine panels) and their inter-percentile (IP) ranges (the lower nine panels) for $P = 0.10$, as noted in the caption. In the upper nine panels, parameter point-estimates are black dots and their inter-percentile ranges are in gray. In the lower nine panels, inter-percentile ranges are given as grey dots, while black lines are loess curves added as eye-guides. For some parameters, some ends of their inter-percentile ranges can attain their supremum/infimum values $(-1, 0, 1)$.

In Figure 31, we present observed information (the upper nine panels) and CDF percentiles (the lower nine panels) for $p = 0.10$, as noted in the caption. In the upper nine panels, observed information is given as grey dots; black lines are loess curves added as eye-guides. In the lower nine panels, percentiles are given as point-estimates (black dots) and inter-percentile ranges (grey lines).

Figures 32-35 continue this pattern, covering all three cases of p. The note from the previous section about excluding additional outliers as applicable also here.

4.4 Nez Perce Tribe dataset and size-selection scenario

The second dataset we employ is from a study of fish consumption by the Nez Perce Tribe [14]. This is data on combined finfish and shellfish consumption,

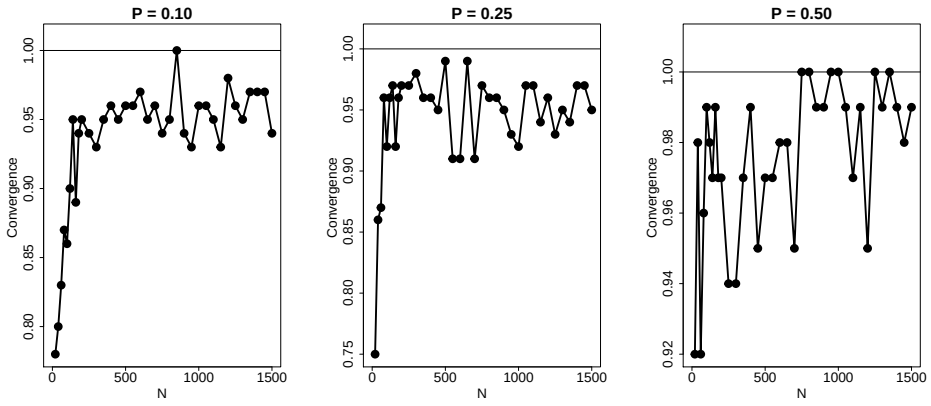


Figure 29: Convergence ratios in frequency-selection scenario.

from $N = 399$ individuals who had two 24-hour recalls measured. There are $N0 = 234$ zero-hits, $N1 = 122$ single-hits, and $N2 = 43$ double-hits. These numbers reflect a practical aspect of applying the NCI method; conducting two interviews per individual is difficult in a given situation, and so the accrued sample size is modest. Also, the original survey was a complex one, with a complex set of weights for individuals and consumption occasions. For the purpose of this thesis, as the code currently addresses only the basic NCI model, we resort to a simplification by assuming that the dataset is instead a simple random sample from the tribe population.

When fitting the model on the subsets from this dataset, we start from the universal starting parameter values gained from fitting the whole NHANES dataset. (This is addressed later.) The age covariate in the dataset was available only in a discretized form, with the following grouping: $18{-}29, 30{-}39, 40{-}49, 50{-}59, 60{+}$. These were coded as integers $1{-}5$.

To examine how the Box-Cox transformation provides approximately Normal outcomes, we calculated the non-zero recall intakes using the fitted value of $\hat{\lambda} = 0.155$. In Figure 36 we give histograms of untransformed and transformed recall intakes, and a Normal Q-Q plot of the transformed intakes. The data is notably less skewed than the NHANES data, and the transformation appears to provide approximately Normal outcomes.

In this scenario we vary the total sample size. We construct a sample by randomly choosing with replacement a fixed number N of individuals present.

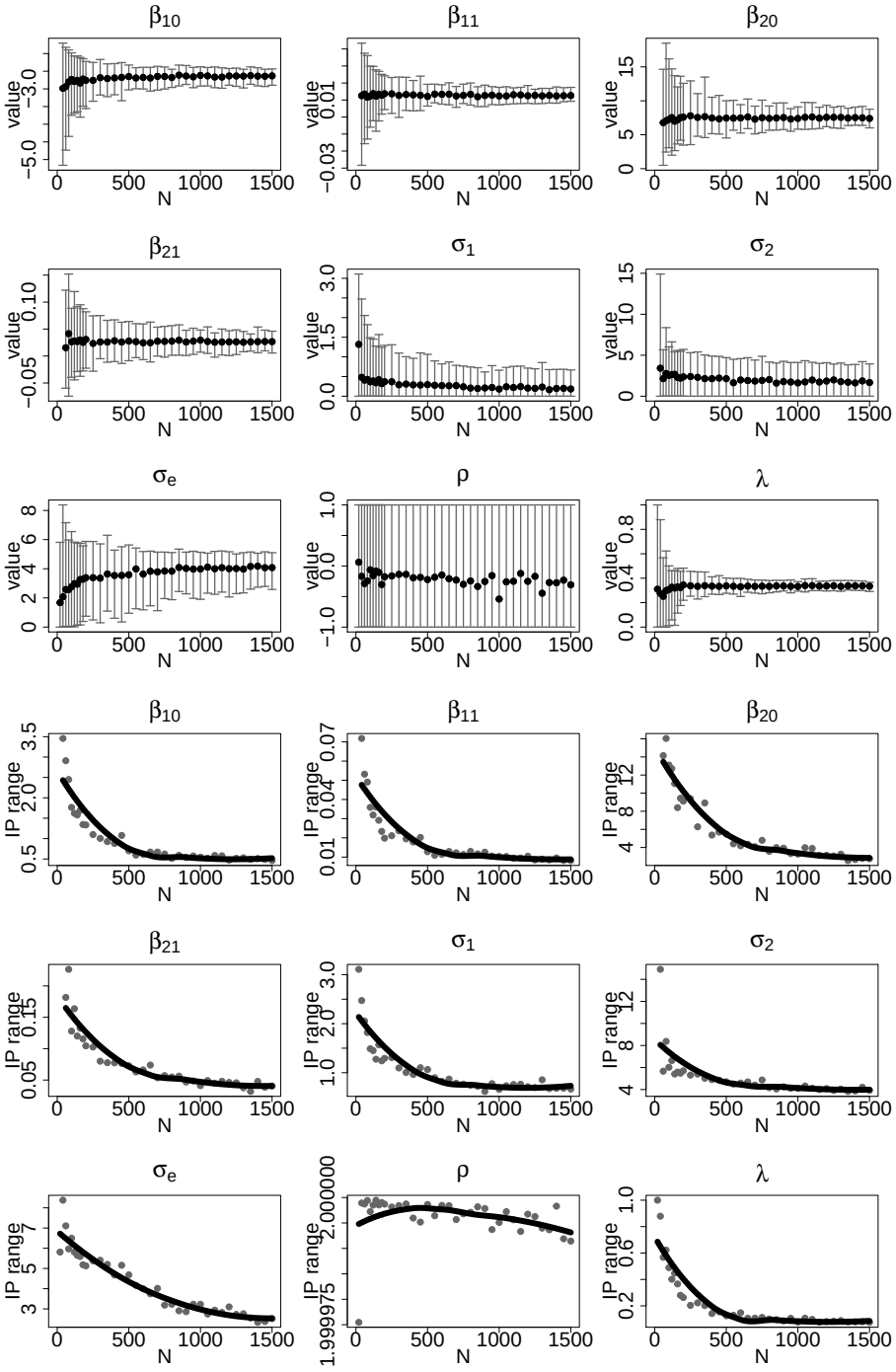


Figure 30: Parameter values & inter-percentile (IP) ranges; P = 0.10

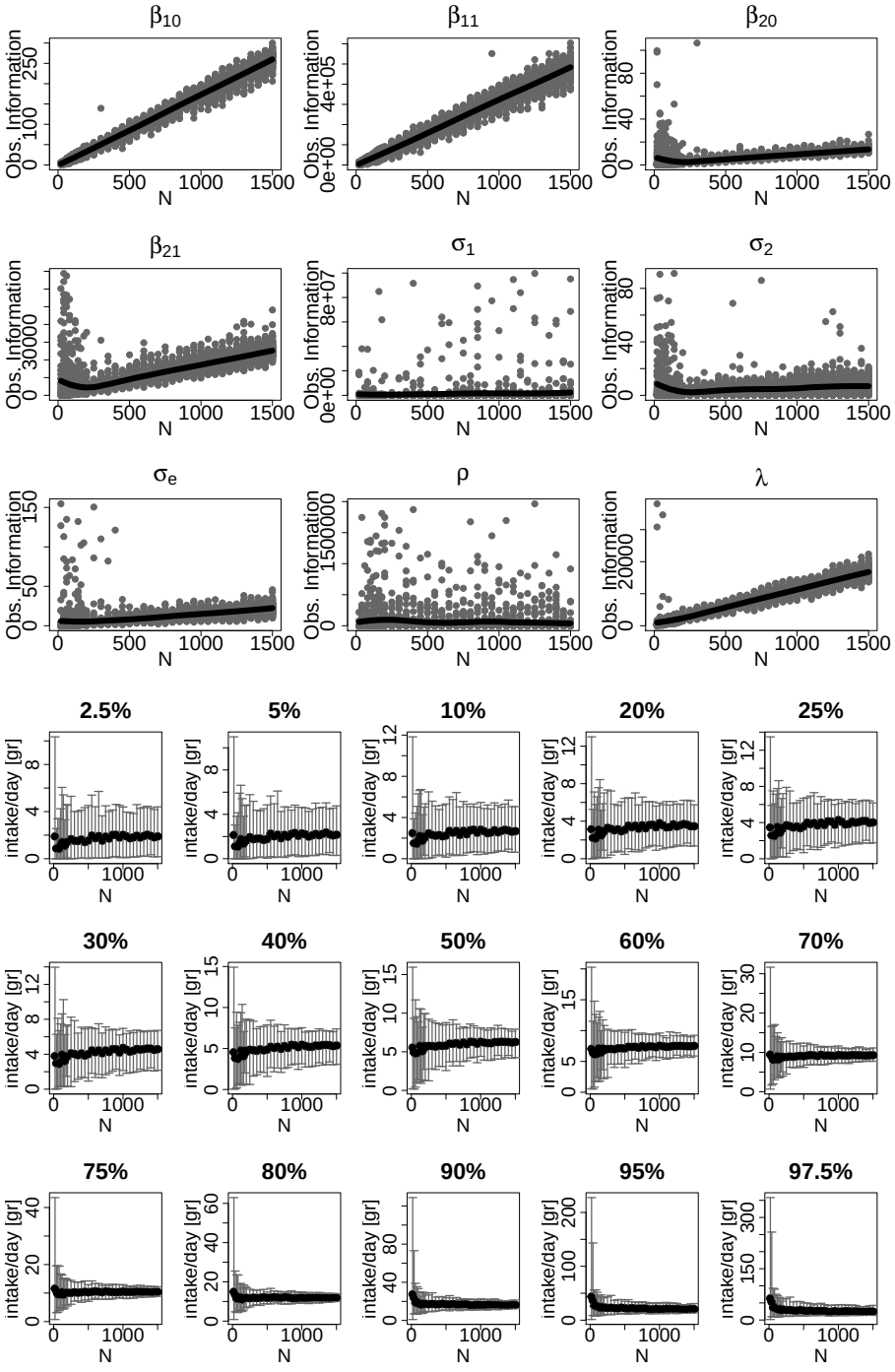


Figure 31: Observed information & CDF percentiles; P = 0.10

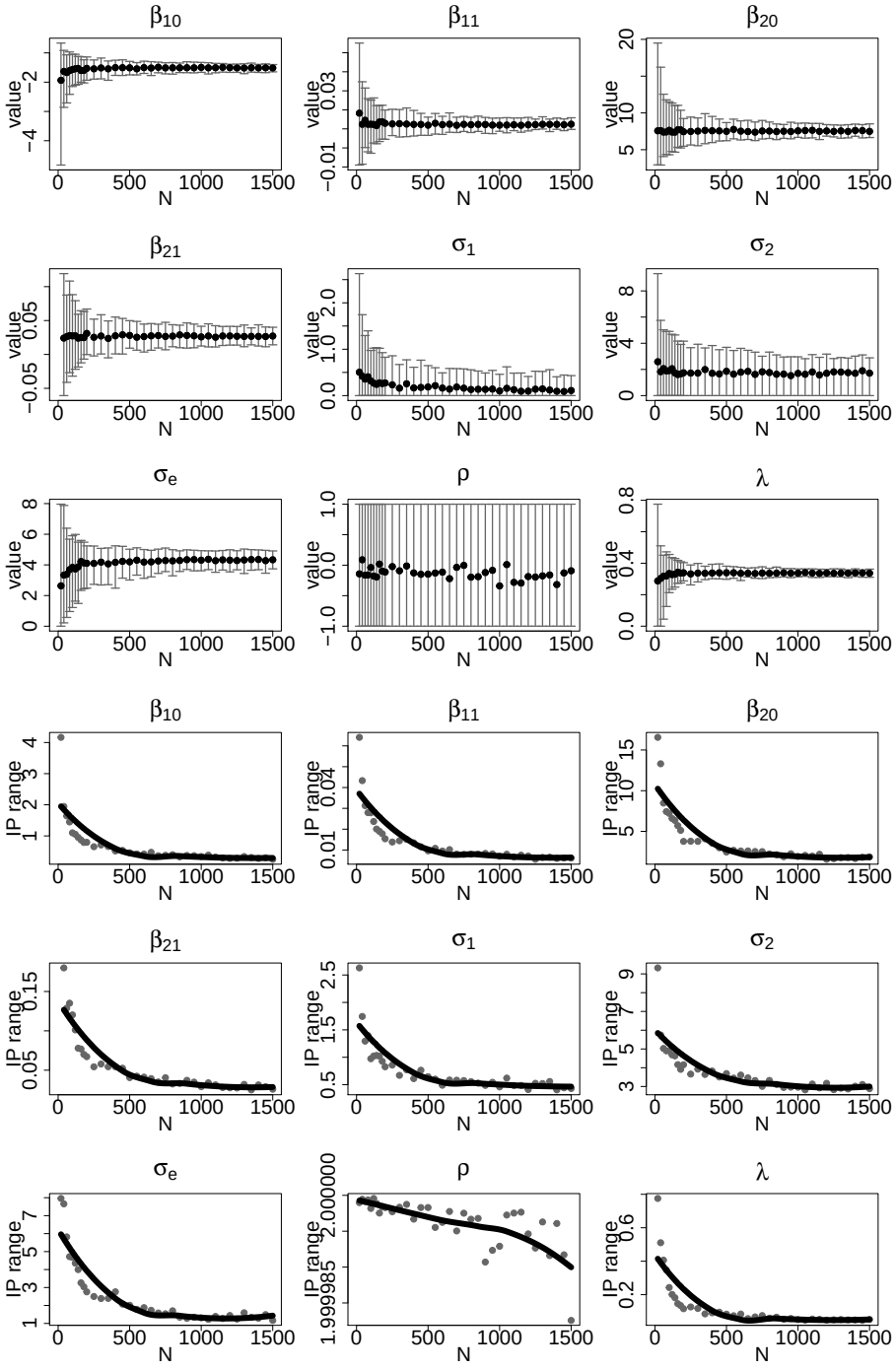


Figure 32: Parameter values & IP ranges; P = 0.25

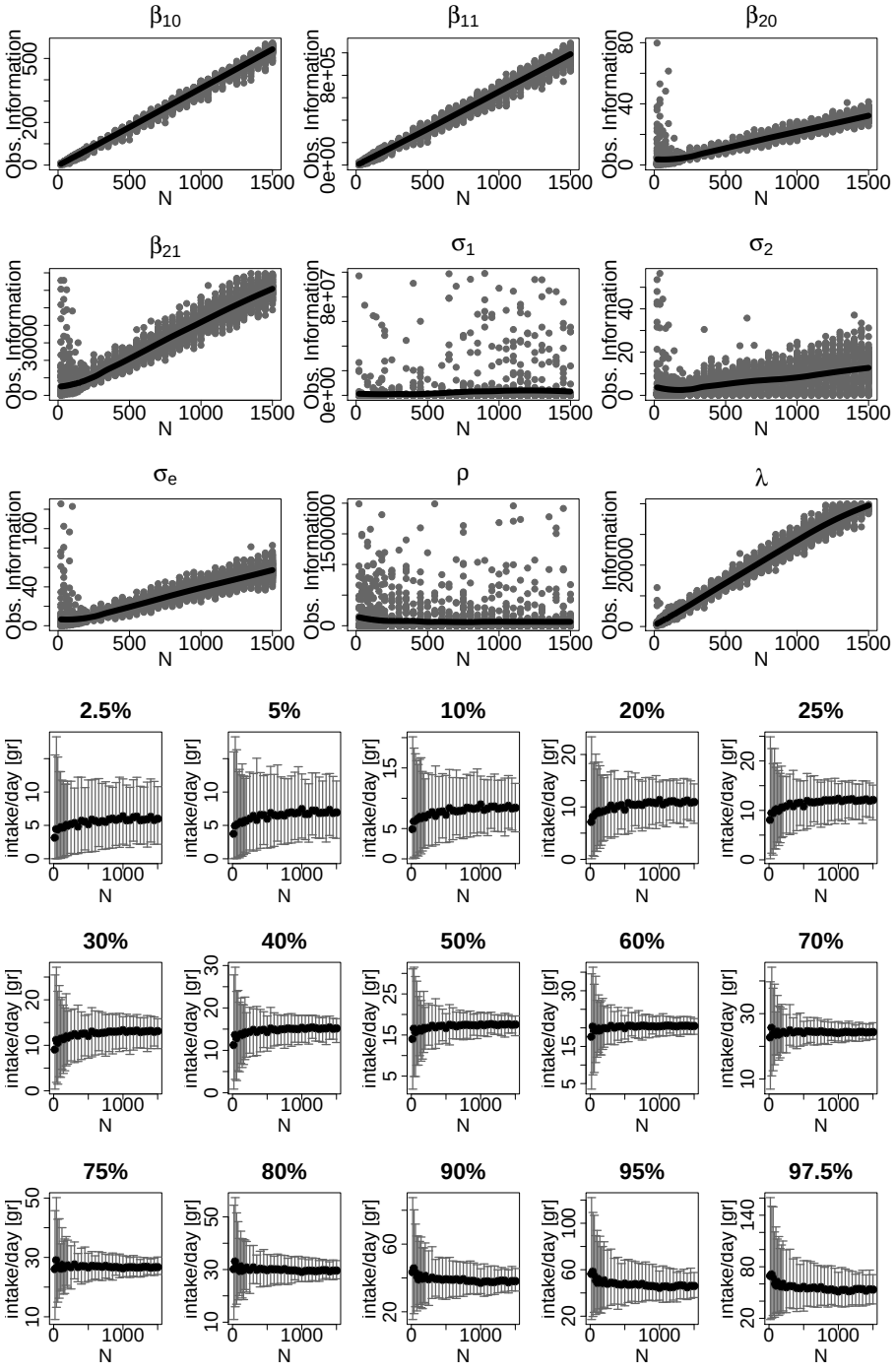


Figure 33: Observed information & CDF percentiles; P = 0.25

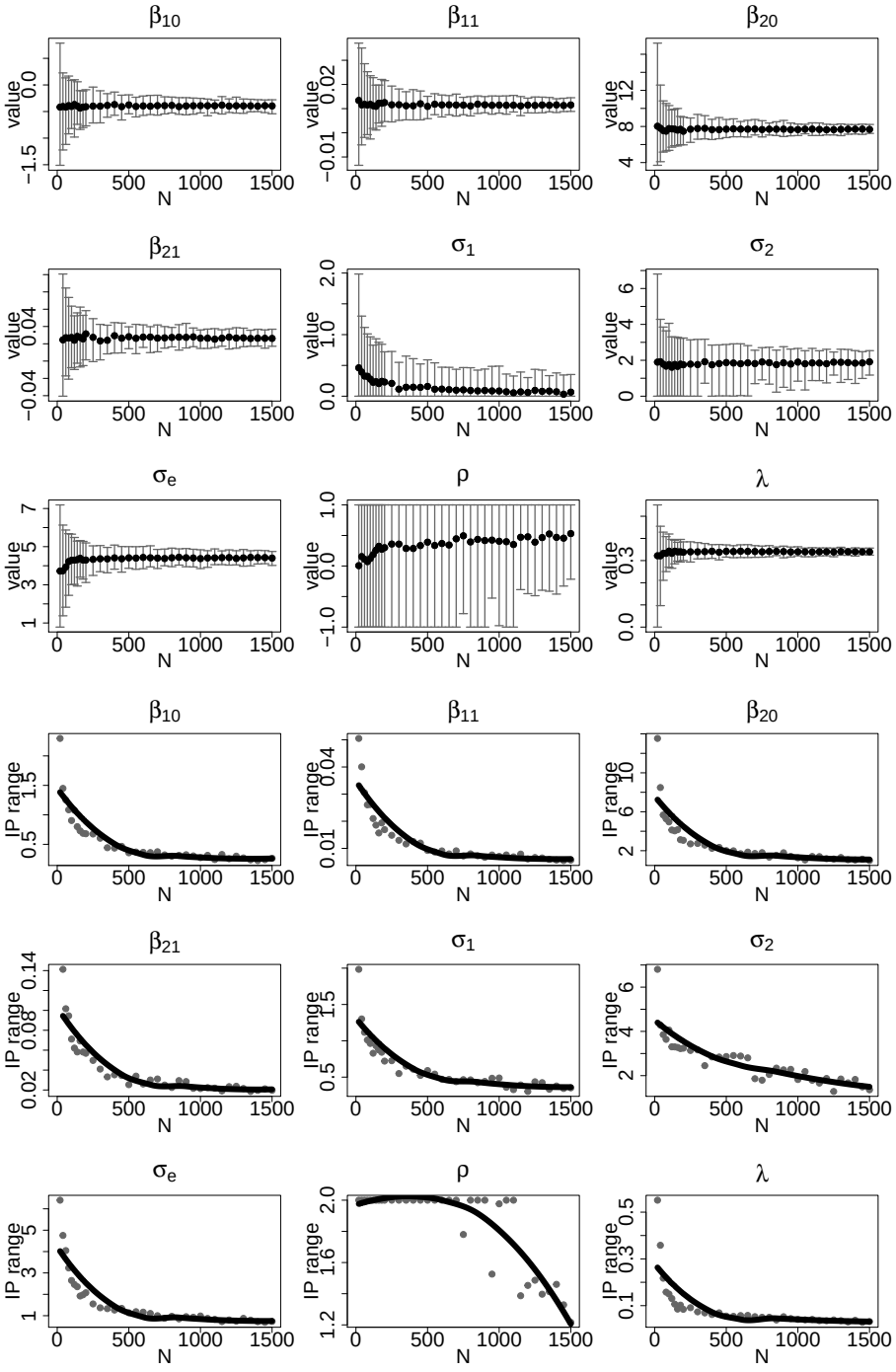


Figure 34: Parameter values & IP ranges; P = 0.50

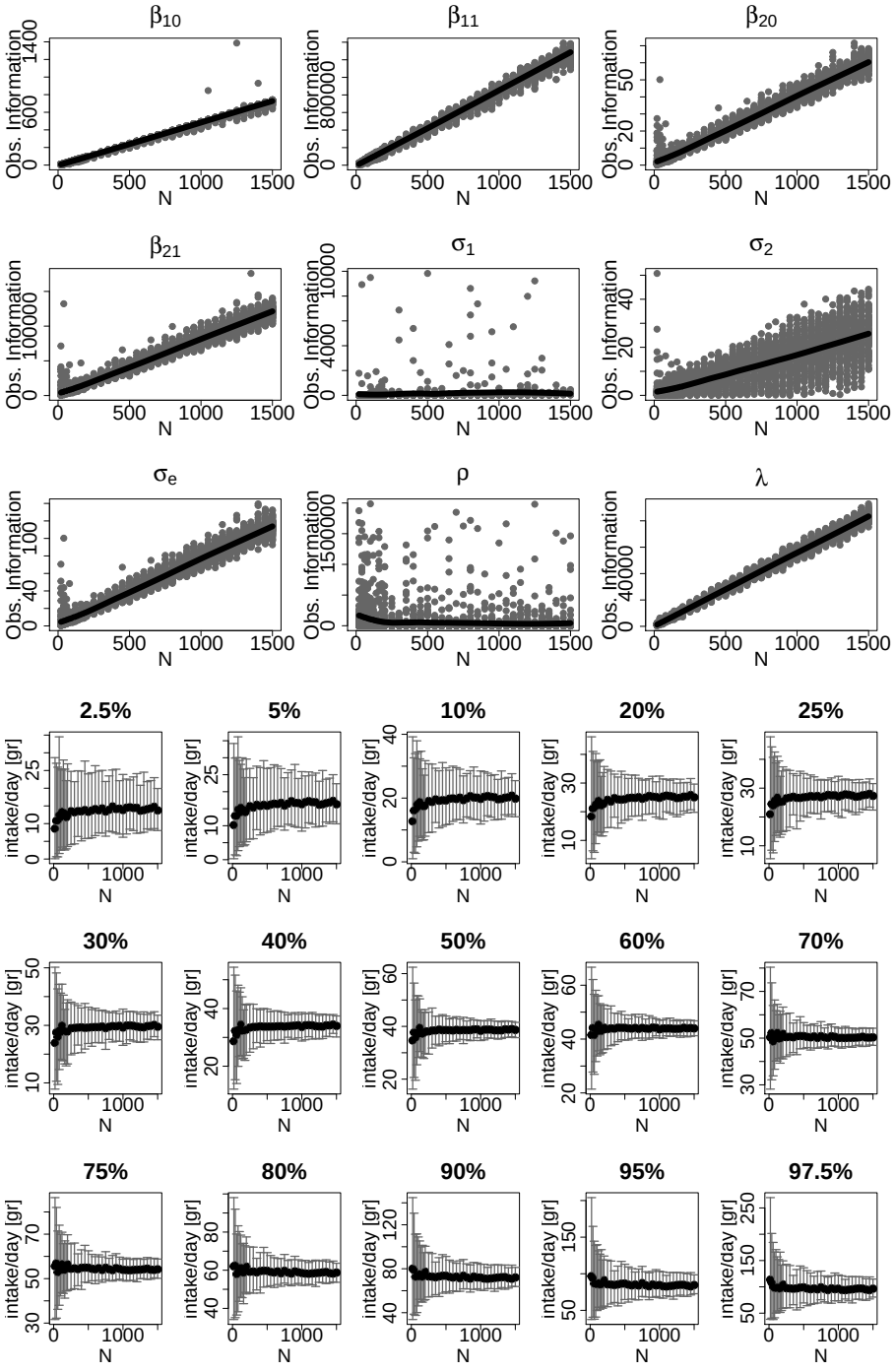


Figure 35: Observed information & CDF percentiles; P = 0.50

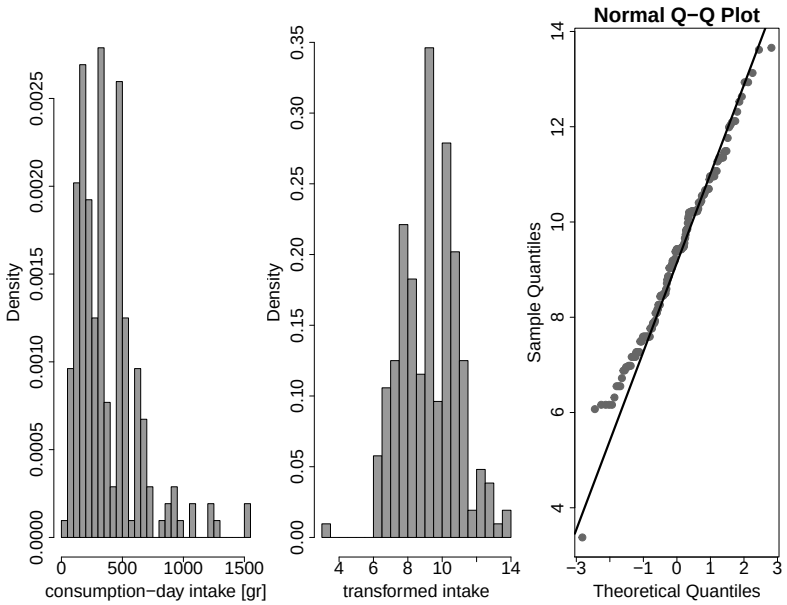


Figure 36: Transformation to normalcy check for the Nez Perce Tribe dataset.

This way of constructing the sample is very straightforward, as it is simply a down-sampled version of the original dataset. Here we do not have a freedom to explore anything besides the widening of the inter-percentile ranges as the sample gets smaller, and whether a bias in estimates develops. Yet we also do not introduce any artificial changes in the relations between different aspects of the model, so it is a reasonable test of the behavior of the NCI model in a realistic setting.

In Figure 37, we show convergence rate as functions of sample size (full dots), with loess curves added as eyeguides. The sample size N is explored in the range $30 - 400$, in steps of size 10. The proportion of successful (i.e. converged) model fits stays relatively high at $\approx 70\%$ at the smallest sample size of 30, where the number of double-hits is only 3.2 on average. To test the influence of starting parameter values, we introduced alternative set of small samples. There we first fitted the full Nez Perce dataset starting from values gained from fitting the full NHANES dataset. Then, we used those parameter values as the starting point for fitting the alternative sample set. This alternative set produced no meaningful changes in model parameters or CDF percentiles. The only noticeable difference was that convergence

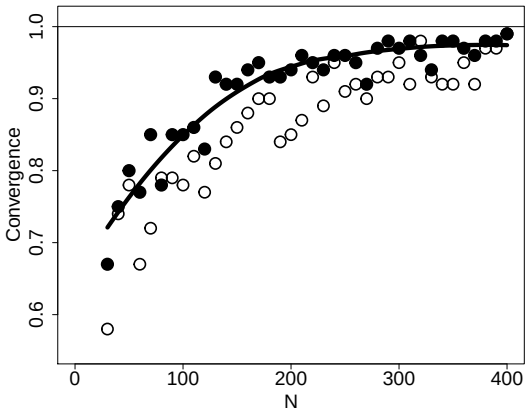


Figure 37: Convergence ratio as the function of sample size for two different choices of starting parameter values, Nez Perce Tribe dataset.

rates were slightly lower in the alternative sample set, shown in Figure 37 as open circles. We conclude that starting from NHANES fitted parameter values gives slightly better performance, and report the rest of our results using this as the starting point.

Figures 38 and 39 present model parameters and their inter-percentile ranges, the observed Fisher information, and CDF percentiles as functions of sample size, in full analogy with the similar figures from sections 4.2 and 4.3.

In addition, values of model parameters, observed information, and percentiles from the optimization of the full dataset of $N = 399$ individuals are added to their respective figures as open circles.

From Figure 39 we note that model parameters remain largely unbiased except at the smallest sample size. However, CDF percentiles seem to exhibit a drift with reducing sample size: small percentiles seem to drift upward while larger percentiles seem unchanged or drift downward. To examine this more closely, we ran down-sampled cases until accruing 1000 converged runs at each of four selected sample sizes: 30, 60, 90, and 120. For these sizes, in Table 1 we present a few selected percentiles and the mean of estimated intake distribution. For each we show estimated mean values, standard deviations, standard errors, and coefficients of variance. The value of percentiles and the mean, calculated by fitting the original dataset with 399 individuals, is given in brackets.

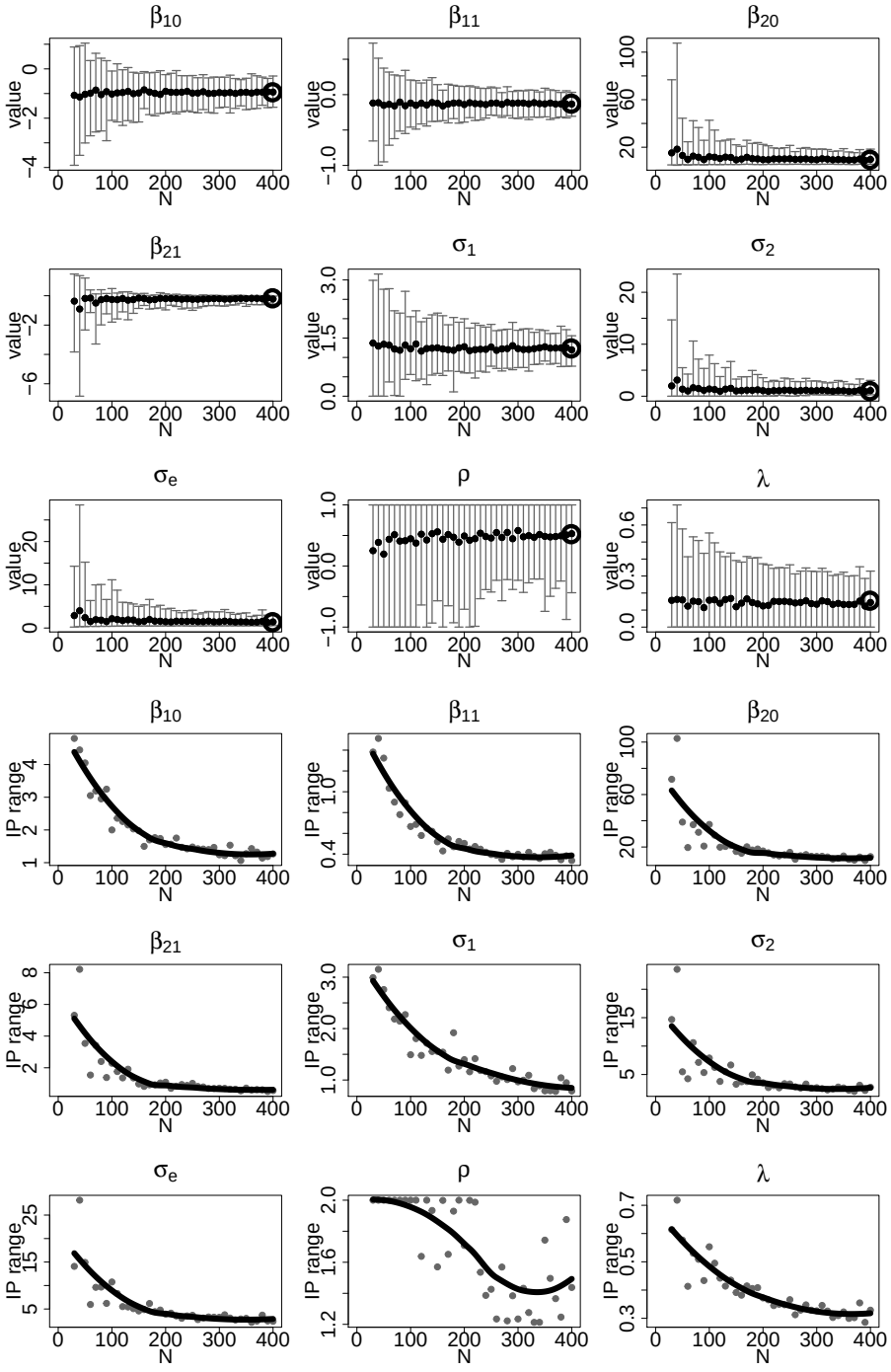


Figure 38: Parameter values & IP ranges, Nez Perce Tribe dataset.

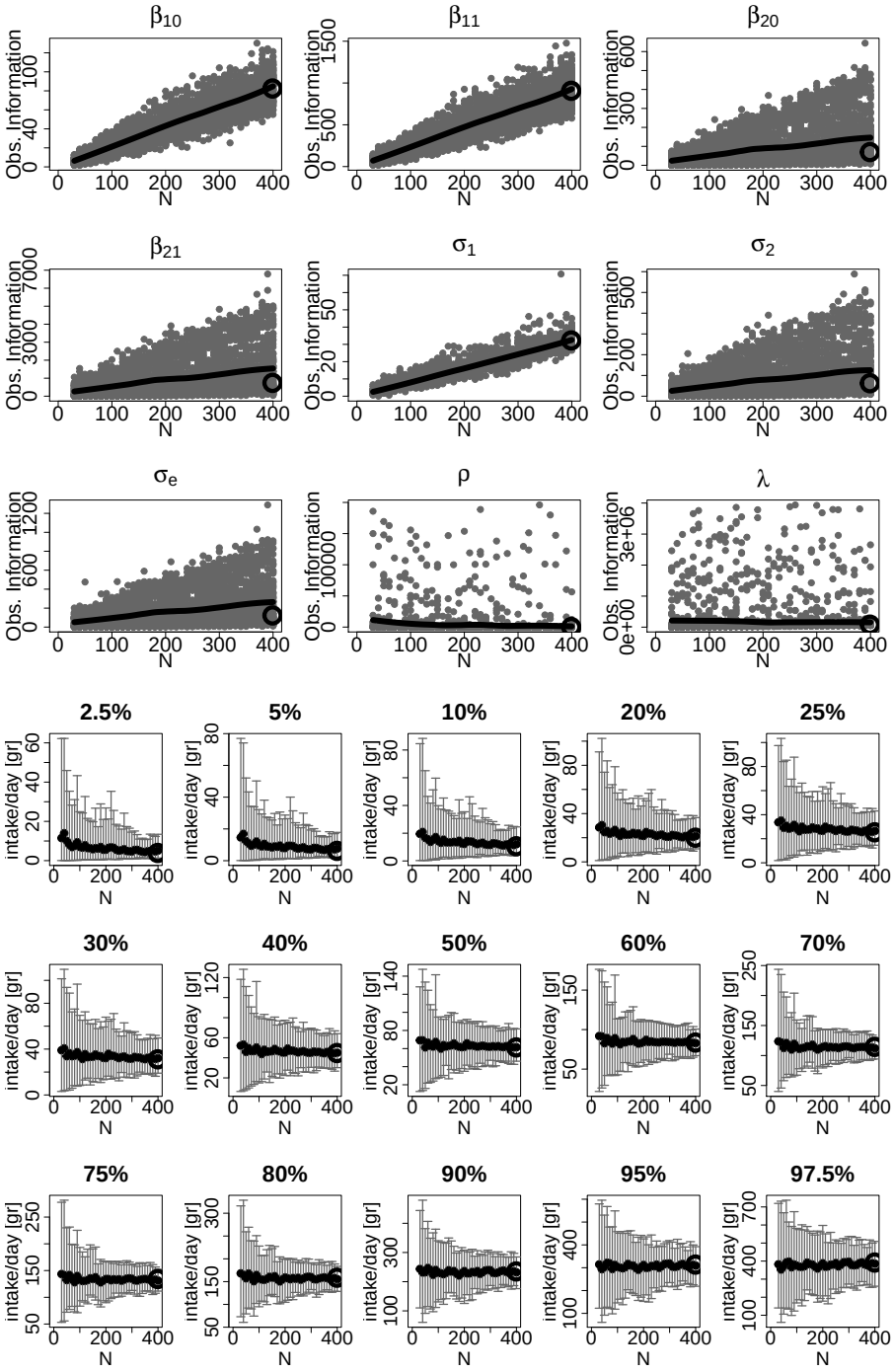


Figure 39: Observed information & CDF percentiles, Nez Perce Tribe dataset.

At these small sample sizes, the lower percentiles (2.5%, 25%, 50%) are larger than their full dataset values, while the higher percentiles (90%, 95%, 97.5%) are smaller. The mean and the 75% did not change much from their full dataset values. This suggests that the distribution as a whole does not shift upwards nor downwards at smallest sample sizes. However it becomes narrower, and this can be attributed to the reduced number of real individuals from whom a set of virtual individuals is constructed, and that are used to estimate the empirical CDF. As the number of real individuals is reduced, the range of their covariate values (here, age) also shrinks.

5 Discussion and concluding remarks

Modeling the population distribution of usual intake of episodically-consumed foods, primarily using 24-hour recall, is a challenging task. The NCI method stands as the most recent major step in the evolution of specialized techniques that address this task, as an integrated modeling framework [2] and it represents an advance in dietary assessment in general [13].

However, there currently exist two stumbling blocks for spreading and facilitating the usage the method to a more general community of researchers. First, it is implemented in *SAS*, which significantly limits its accessibility. Second, real surveys of usual intake often have only modest samples sizes, and the general performance of the model when sample sizes are small needs to be understood.

With this thesis we tried to address these two obstacles. First, we implemented the basic method in the R environment, making it widely available and free of charge, increasing the number of researchers and practitioners in the position to actually use the NCI method.

Second, we have presented the performance of the method for small sample sizes using two different datasets, and in three potentially-useful scenarios. The aspects of the model whose behaviour we observed are: the proportion of model fits that successfully converge, parameter and their variability estimate, Fisher information associated with individual model parameters, and CDF percentiles and their variability estimates.

In Section 4.2, we explored the behavior at different counts of zero-, single- and double-hitters. Here a researcher may get an idea of how would the NCI

percentile		*size*	*mean*	*sd*	*st.err.*	*CV*(%)
(3.90)	2.5%					
		30	12.05	18.06	0.58	149.88
		60	8.99	12.43	0.39	138.26
		90	8.39	12.15	0.37	144.82
		120	7.21	8.43	0.26	116.92
(25.35)	25%					
		30	33.54	27.40	0.88	81.69
		60	30.51	21.11	0.66	69.19
		90	30.50	19.19	0.58	62.92
		120	29.56	15.54	0.48	52.57
(61.26)	50%					
		30	66.89	33.88	1.09	50.65
		60	63.85	25.24	0.79	39.53
		90	65.01	23.07	0.7	35.49
		120	64.65	19.04	0.59	29.45
(134.12)	75%					
		30	137.51	50.60	1.63	36.80
		60	132.74	35.37	1.11	26.65
		90	134.40	34.82	1.06	25.91
		120	134.42	27.45	0.84	20.42
(235.07)	90%					
		30	234.53	92.51	2.98	39.44
		60	228.97	71.88	2.25	31.39
		90	230.55	66.96	2.04	29.04
		120	231.90	55.66	1.71	24.00

percentile		*size*	*mean*	*sd*	*st.err.*	$CV(\%)$
(313.68)	95%					
		30	303.67	134.89	4.34	44.42
		60	298.61	107.63	3.37	36.04
		90	300.66	96.70	2.94	32.16
		120	303.87	82.25	2.53	27.07
(391.87)	97.5%					
		30	370.37	183.07	5.90	49.43
		60	366.08	148.12	4.64	40.46
		90	368.54	129.20	3.93	35.06
		120	373.86	111.77	3.43	29.90
(98.15)	*mean*					
		30	102.58	31.73	1.02	30.93
		60	98.56	23.13	0.73	23.47
		90	99.29	23.42	0.71	23.59
		120	99.23	17.99	0.55	18.13

Table 1: Descriptive statistics of a few selected percentiles and the mean of estimated intake distribution, at four selected sample sizes. Each intake value is an average over 1000 converged runs, down-sampled from the full dataset. The values calculated by fitting the full dataset are shown in brackets. All intake values are given in $[gr/day]$. Nez Perce Tribe dataset.

method perform in a sample where these numbers are available to him or could be estimated, or even set *a priori* in the study's sampling design. In practice, with episodically-consumed food, a major concern is whether $N2$ is large enough (at least several tens is often expected) for the CDF percentiles to be reliable. For this we first concentrate on Figure 20. We notice that both values and variability of percentiles do experience some change for $N1 \leq 100$, and become very stable for larger $N1$. Also, in Figure 28, where $N1$ is set to 100, we see that smaller percentiles (2.5%-30%) do experience a sudden increase in variability for values $N2 \leq 50$, while larger percentiles (40%+) behave very smoothly at any $N2$. From these observations, we can make a recommendation that a sample should have at least $N1 = 100$ and $N2 = 50$ for its CDF estimate to be free from large decrease in precision due to its inadequate size.

In Section 4.3 we attempted to mimic situations with various realistic probabilities of consumption, so that a researcher can have an idea of what amount of robustness to expect from the model if he can come up with an estimate for the frequency of consumption of a particular food or nutrient. Interestingly, Figures 31, 33, 35 suggest that all probabilities tested ($P = 0.10, 0.25, 0.50$) exhibit the same behaviour of CDF percentiles, their variability abruptly increasing at sample sizes below approximately 200. Therefore, we recommend $N = 200$ to be a minimal value of the total sample size for NCI method to be used.

In Section 4.4 we performed a straightforward reduction of the system size down to near zero, and noticed widening of inter-percentile ranges, lack of change in model parameters values, and a shrinking of the intake distribution as a whole without a significant effect on its mean. In this sense, mean usual intake of NCI model may be tentatively considered unbiased in the sense that the average mean intake over many small samples, would approximately recover the mean intake of a very large sample. However, this is not the case with the tails of estimated CDF.

In the two real-life examples that we examined, NHANES and Nez Perce Tribe datasets of fish consumption, correlation between probability and amount of intake was rather small, and not statistically significant for the sample sizes explored, except in situations where double-hitters dominate the sample size, making consumption probability close to one. The least-informed parameter is intake probability between-person variance σ_1, and we never saw indications that the CLT was accurate for these estimates.

The most-informed parameter is the Box-Cox transformation parameter λ which nearly always reaches statistical significance in the settings we considered. The rest of parameter estimates fall somewhere in-between, and their behavior is situation-dependent. The convergence of the estimates is very good, typically occurring in 95% or more datasets, except at the smallest sizes. The lowest convergence rate was found around 70% for a system of total size 30. The rate of convergence was found to either increase or remain unchanged with increasing system size.

Our code is currently capable of solving the basic NCI method, and can be extended in several potential directions. Examples are: adding weights to address complex surveys and further speed optimization by transferring numerical load to routines in lower-level languages. Particularly interesting development would be adding a Bayesian framework, so that a relevant external information that originated beyond the sparsely available 24-hour recalls at hand could be imported via properly chosen set of priors.

References

[1] C. E. Woteki, "Integrated NHANES: Uses in National Policy," *The Journal of Nutrition*, vol. 133, no. 2, pp. 582S–584S, 02 2003. [Online]. Available: https://doi.org/10.1093/jn/133.2.582S

[2] K. W. Dodd, P. M. Guenther, L. S. Freedman, A. F. Subar, V. Kipnis, D. Midthune, J. A. Tooze, and S. M. Krebs-Smith, "Statistical methods for estimating usual intake of nutrients and foods: a review of the theory," *Journal of the american Dietetic association*, vol. 106, no. 10, pp. 1640–1650, 2006.

[3] US Food and Drug Administration. (2019) How to understand and use the nutrition facts label. [Online]. Available: https://www.fda.gov/food/nutrition-education-resources-materials/how-understand-and-use-nutrition-facts-label

[4] K. M. FLEGAL and F. A. Larkin, "Partitioning macronutrient intake estimates from a food frequency questionnaire," *American journal of epidemiology*, vol. 131, no. 6, pp. 1046–1058, 1990.

[5] A. F. Subar, V. Kipnis, R. P. Troiano, D. Midthune, D. A. Schoeller, S. Bingham, C. O. Sharbaugh, J. Trabulsi, S. Runswick, R. Ballard-Barbash *et al.*, "Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the open study," *American journal of epidemiology*, vol. 158, no. 1, pp. 1–13, 2003.

[6] V. Kipnis, A. F. Subar, D. Midthune, L. S. Freedman, R. Ballard-Barbash, R. P. Troiano, S. Bingham, D. A. Schoeller, A. Schatzkin, and R. J. Carroll, "Structure of dietary measurement error: results of the open biomarker study," *American journal of epidemiology*, vol. 158, no. 1, pp. 14–21, 2003.

[7] L. S. Freedman, D. Midthune, R. J. Carroll, S. Krebs-Smith, A. F. Subar, R. P. Troiano, K. Dodd, A. Schatzkin, P. Ferrari, and V. Kipnis, "Adjustments to improve the estimation of usual dietary intake distributions in the population," *The Journal of nutrition*, vol. 134, no. 7, pp. 1836–1843, 2004.

[8] Institute of Medicine, Food and Nutrition Board. (2003) Dietary reference intakes: Applications in dietary planning. [Online]. Available: https://nap.edu/books/0309088534/html

[9] S. M. Nusser, A. L. Carriquiry, K. W. Dodd, and W. A. Fuller, "A semiparametric transformation approach to estimating usual daily intake distributions," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1440–1449, 1996.

[10] P. M. Guenther, P. S. Kott, and A. L. Carriquiry, "Development of an approach for estimating usual nutrient intake distributions at the population level," *The Journal of nutrition*, vol. 127, no. 6, pp. 1106–1112, 1997.

[11] L. Nusser SM, Fuller WA, Guenther PM. Adjusting for measurement error and non-normality in 24-hour food intake data. In: Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, *Survey measurement and process quality*. Wiley New York, 1997.

[12] K. Dodd, A. Carriquiry, W. Fuller, and C. Chen, "On the estimation of usual intake distributions for foods," in *PROCEEDINGS-AMERICAN STATISTICAL ASSOCIATION BIOMETRICS SECTION*, 1997, pp. 100–105.

[13] J. A. Tooze, D. Midthune, K. W. Dodd, L. S. Freedman, S. M. Krebs-Smith, A. F. Subar, P. M. Guenther, R. J. Carroll, and V. Kipnis, "A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution," *Journal of the American Dietetic Association*, vol. 106, no. 10, pp. 1575–1587, 2006.

[14] Polissar NL, Salisbury A, Ridolfi C, Callahan K, Neradilek M, Hippe DS, Beckley WH. (2015) A Fish Consumption Survey of the Nez Perce Tribe. The Mountain-Whisper-Light Statistics, Pacific Market Research, Ridolfi, Inc. [Online]. Available: https://www.epa.gov/sites/production/files/2017-01/documents/fish-consumption-survey-nez-perce-dec2016.pdf

[15] Polissar NL, et. al. (2015) A Fish Consumption Survey of the Shoshone-Bannock Tribes. The Mountain-Whisper-Light Statistics, Pacific Market Research, Ridolfi, Inc. [Online]. Available: https://www.epa.gov/sites/production/files/2017-01/documents/fish-consumption-survey-shoshone-bannock-dec2016.pdf

[16] J. A. Tooze, G. K. Grunwald, and R. H. Jones, "Analysis of repeated measures data with clumping at zero," *Statistical methods in medical research*, vol. 11, no. 4, pp. 341–355, 2002.

[17] J. A. Tooze, V. Kipnis, D. W. Buckman, R. J. Carroll, L. S. Freedman, P. M. Guenther, S. M. Krebs-Smith, A. F. Subar, and K. W. Dodd, "A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the nci method," *Statistics in medicine*, vol. 29, no. 27, pp. 2857–2868, 2010.

[18] A. F. Subar, F. E. Thompson, V. Kipnis, D. Midthune, P. Hurwitz, S. McNutt, A. McIntosh, and S. Rosenfeld, "Comparative validation of the block, willett, and national cancer institute food frequency questionnaires: the eating at america's table study," *American journal of epidemiology*, vol. 154, no. 12, pp. 1089–1099, 2001.

[19] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[20] J. C. Pinheiro and D. M. Bates, "Approximations to the log-likelihood function in the nonlinear mixed-effects model," *Journal of computational and Graphical Statistics*, vol. 4, no. 1, pp. 12–35, 1995.

[21] S. M. Nusser, A. L. Carriquiry, K. W. Dodd, and W. A. Fuller, "A semiparametric transformation approach to estimating usual daily intake distributions," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1440–1449, 1996.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[23] P. Gilbert and R. Varadhan, *numDeriv: Accurate Numerical Derivatives*, 2019, r package version 2016.8-1.1. [Online]. Available: https://CRAN.R-project.org/package=numDeriv

[24] P. J. Bickel, F. Götze, and W. R. van Zwet, "Resampling fewer than n observations: gains, losses, and remedies for losses," in *Selected works of Willem van Zwet*. Springer, 2012, pp. 267–297.

[25] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[26] J. Liu and J. S. Hodges, "Posterior bimodality in the balanced one-way random-effects model," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 247–255, 2003.

[27] A. Szpiro, *BIOSTAT 571*. Lecture notes, UW 2019.