Nonstandard conditionally specified models for nonignorable missing data

Alexander M. Franks[a], Edoardo M. Airoldi[b,1], and Donald B. Rubin[b,c,1]

[a]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106; [b]Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, PA 19122; and [c]Qiu Chengtong Center for Mathematical Sciences, Tsinghua University, Beijing, 100084, China

Contributed by Donald B. Rubin, June 15, 2020 (sent for review April 18, 2019; reviewed by David Hoaglin and Hal Stern)

Data analyses typically rely upon assumptions about the missingness mechanisms that lead to observed versus missing data, assumptions that are typically unassessable. We explore an approach where the joint distribution of observed data and missing data are specified in a nonstandard way. In this formulation, which traces back to a representation of the joint distribution of the data and missingness mechanism, apparently first proposed by J. W. Tukey, the modeling assumptions about the distributions are either assessable or are designed to allow relatively easy incorporation of substantive knowledge about the problem at hand, thereby offering a possibly realistic portrayal of the data, both observed and missing. We develop Tukey's representation for exponential-family models, propose a computationally tractable approach to inference in this class of models, and offer some general theoretical comments. We then illustrate the utility of this approach with an example in systems biology.

missing not at random | nonignorable missingness mechanism | Tukey's representation | Bayesian analysis | exponential tilting

Missing data are ubiquitous in the social and biomedical sciences, and the credibility of any data analysis is dependent on the assumed mechanism that leads to the missing data, as well as on the mode of inference (1). Here, we work within a framework in which the estimand involves both observed and missing data (2). An important concept is that of ignorable missing data under which there is no need to specify a model for the missingness indicators to achieve valid Bayesian or likelihood-based inference (3–5).

There are two basic approaches to specify the joint distribution of the complete data (observed and missing) and missingness indicators. The first approach, called selection modeling, is to posit a standard model for the complete data and then specify a model that selects observed data from the complete data, referred to as the missingness mechanism (6). The second approach, called pattern-mixture modeling, is to specify separate distributions for each pattern of observed and missing data, thus eschewing explicit assumptions about the missingness mechanism (7, 8). The fundamental challenge with these two basic approaches is that assumptions about the missingness mechanism, whether explicit or implicit, are rarely testable from the observed data. As a result, literature on inference in the presence of missing data includes strategies for assessing sensitivity to model specifications (1, 7, 9, 10), often using a model assuming ignorability as a baseline.

Contributions

Most statistical analyses involving nonignorable missingness mechanisms use one of two approaches: pattern-mixture models or selection models. Here, we develop an alternative approach, evidently originally proposed by J. W. Tukey in a discussion of ref. 11, and described by ref. 12, which has thus far remained recondite. The key insight is to represent the joint distribution of the complete data and missingness indicators in a way such that assumptions are either assessable or, typically, allow the incorporation of substantive knowledge about the problem at hand, thereby offering a path to elicit a realistic portrayal of the data. This work is related to previous work on, so called, exponential-tilt pattern-mixture models for nonignorable missing data (13–16). We make the connection between exponential tilting and Tukey's representation, but we focus explicitly on describing the utility of Tukey's representation for Bayesian inference with nonignorable missing data (16). Also, we introduce a class of flexible and widely applicable models based on logistic missingness mechanisms and exponential-family models for the observed data.

Basic Models for Missing Data. Discussion of models for missing data can be found in a variety of places, including (11, 17). We introduce ideas in the simple case where the data are exchangeable scalar random variables.

Let $Y = (Y_1, Y_2, \ldots, Y_N)'$ be the complete data and $R = (R_1, R_2, \ldots, R_N)'$ represent the response indicators for Y; Y_i is "missing" when $R_i = 0$ and "observed" when $R_i = 1$. Assuming independence between observations (Y_i, R_i), the joint distribution for (Y_i, R_i) can be written in independent and identically distributed (i.i.d.) form:

$$P(Y, R \mid \theta) = \prod_i f(Y_i, R_i \mid \theta),$$

where θ is a parameter vector with a prior distribution. For notational simplicity, we focus on the case without covariates

Significance

We consider data-analysis settings where data are missing not at random. In these cases, the two basic modeling approaches are 1) pattern-mixture models, with separate distributions for missing data and observed data, and 2) selection models, with a distribution for the data preobservation and a missing-data mechanism that selects which data are observed. These two modeling approaches lead to distinct factorizations of the joint distribution of the observed-data and missing-data indicators. In this paper, we explore a third approach, apparently originally proposed by J. W. Tukey as a remark in a discussion between Rubin and Hartigan, and reported by Holland in a two-page note, which has been so far neglected.

Author contributions: A.M.F., E.A., and D.B.R. designed research; A.M.F., E.A., and D.B.R. performed research; A.M.F. and E.A. analyzed data; and A.M.F., E.A., and D.B.R. wrote the paper.

Reviewers: D.H., University of Massachusetts Medical School; and H.S., University of California, Irvine.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. airoldi@temple.edu or dbrubin@temple.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1815563117/-/DCSupplemental.

First published July 28, 2020.

although the approach can easily be extended to include covariates (*Theory*).

The Selection Factorization. The selection approach (6) factors the joint distribution of (Y_i, R_i) as

$$f(Y_i|\theta_Y)f(R_i|Y_i,\theta_{R|Y}), \quad [1]$$

using the distribution of the complete data, $P(Y_i|\theta_Y)$, and the missingness mechanism, i.e., $P(R_i|Y_i,\theta_{R|Y})$, which controls which complete-data values are observed and which are missing, where the parameters θ_Y and $\theta_{R|Y}$ are for the complete-data and missingness-mechanism parameters, respectively. Typical selection modeling has θ_Y distinct from $\theta_{R|Y}$, where "distinct" (3) means in disjoint parameter spaces and a priori independent (if distributions are specified). Common models for $f(Y_i|\theta_Y)$ include the normal, with mean μ and variance σ^2, i.e., $f(Y_i|\theta_Y) \sim N(\mu,\sigma^2)$ with $\theta_Y = (\mu,\sigma^2)$, or the Bernoulli with unknown probability of success, i.e., $f(Y_i|\theta_Y) \sim \text{Bern}(p)$ with $\theta_Y = p$. Common missingness mechanisms used in practice include the logistic and probit models (18), for the former $f(R_i|Y_i,\theta_{R|Y}) \sim \text{logit}^{-1}(\alpha+\beta Y_i)$ with $\theta_{R|Y} = (\alpha,\beta)$.

The Pattern-Mixture Factorization. The pattern-mixture approach (7, 8) is the alternative basic factorization. Here, the complete-data distribution is specified as a mixture of observed data and missing-data components,

$$f(Y_i, R_i|\theta) = f(Y_i|R_i,\theta_{Y|R})f(R_i|\theta_R), \quad [2]$$

where $\theta^{\text{obs}}_{Y|R} = (\theta^{\text{obs}}_{Y|R}, \theta^{\text{obs}}_{Y|R})$ and thus $\theta = (\theta^{\text{obs}}_{Y|R}, \theta^{\text{mis}}_{Y|R}, \theta_R)$ are the observed-data parameters, missing-data parameters, and the population fraction of observed data, respectively. Here $f(R_i|\theta_R)$, a Bernoulli distribution, is easily estimated by the fraction of indicators equal to one. The model for $f(Y_i|R_i=1,\theta^{\text{obs}}_{Y|R})$ is typically chosen to fit the observed data well. There is no information in the observed data or indicators, R_i, about $\theta^{\text{mis}}_{Y|R}$. This factorization omits an explicit specification for the missingness mechanism, about which there may be substantive knowledge. Tukey's representation offers another choice.

Tukey's Representation. John W. Tukey, recorded in ref. 12, suggested an alternative representation of the joint distribution for (Y_i, R_i), which he referred to as the "simplified selection model," with parameters $\theta^{\text{obs}}_{Y|R}$, $\theta_{R|Y}$, and θ_R:

$$f(Y_i, R_i \mid \theta) = f(Y_i \mid R_i=1,\theta^{\text{obs}}_{Y|R}) \times f(R_i=1 \mid \theta_R)\cdot \frac{f(R_i \mid Y_i,\theta_{R|Y})}{f(R_i=1 \mid Y_i,\theta_{R|Y})}, \quad [3]$$

where $\theta^{\text{obs}}_{Y|R}$ are the parameters of the observed data density. The missingness mechanism describes the probability that Y_i is observed or missing given its value. Here, we focus on models using Tukey's representation when the observed-data distribution is an exponential-family distribution and when the missingness mechanism$f(R_i=1 \mid Y_i,\theta_{R|Y})$ is the inverse-logit of some function of Y_i. Tukey's representation can be obtained from Bayes' rule or through an application of Brook's lemma (19, 20), commonly referenced in the theory of spatial autoregressive models (21). Brook's Lemma is only applicable when the so-called "positivity condition" (due to Hammersley and Clifford) is satisfied (20), which for Tukey's representation means that

If $P(R_i=r|\theta_R)>0$ and $P(Y_i=y|\theta_Y)>0$
Then $P(R_i=r,\, Y_i=y|\theta)>0$

for all pairs of values (r,y), where θ_Y are the parameters for the complete-data distribution of Y_i.* This condition enforces regularity in the way the supports of the marginal distributions relate to the support of the corresponding joint distribution and avoids pathological cases (e.g., the case in ref. 22). This condition is not trivially satisfied in missing-data problems. For instance, Tukey's representation cannot be applied to models where $P(R_i=1|Y_i<c,\theta_{R|Y})=0$, for some finite c, as when the complete data are normal but the observed data are truncated normal. Consequently, here we focus on problems where $P(R_i=1|Y_i,\theta_{R|Y})>0$, that is, where the support of the missing data is a nontrivial subset of the support of the observed data. Moreover, as we discuss later, the distributions specified in Eq. **3** must imply an integrable joint density (20). With Tukey's representation, the "integrability condition" constrains the rate at which the tails of the distribution for the observed data decrease relative to the rate at which the odds of a missing value increase. This condition is further discussed in *A Note on the Integrability Condition*.

Unlike related work with exponential-tilting models discussed in *Connections to Exponential Tilting*, Tukey's approach focuses on an explicit formulation of the missingness mechanism. We describe the utility of Bayesian inference with Tukey's representation, and we show that it is tractable when the missingness mechanism is logistic and the observed-data distribution is in the exponential family.

Advantages and Challenges of Tukey's Representation. A notion central to the arguments in this paper is that although joint distributions can be represented mathematically in several ways, a particular representation may involve components that are more easily elicited from investigators or more easily estimated from data. As Holland notes (12), a main advantage of Tukey's representation is that it involves the observed-data density, $f(Y_i \mid R_i=1,\theta^{\text{obs}}_{Y|R})$, and the marginal probability of a missing observation, both of which can be estimated directly, and the missingness mechanism, $f(R_i \mid Y_i,\theta_{R|Y})$, which is often natural to elicit in the context of a specific application.

Modeling and Inference Using Tukey's Representation. Let $f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R})$ denote the observed-data density parameterized by $\theta^{\text{obs}}_{Y|R}$. Using Eq. **3**, we can write the joint density for (y_i, r_i) as

$$f(y_i, r_i \mid \theta) \propto \begin{cases} f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R}) & \text{if } r_i = 1 \\ \frac{f(r_i=0|y_i,\theta_{R|Y})}{f(r_i=1|y_i,\theta_{R|Y})} f^{\text{obs}}(y_i|\theta^{\text{obs}}_{Y|R}) & \text{if } r_i = 0, \end{cases}$$

or

$$f(r_i=1 \mid y_i,\theta_{R|Y})^{r_i-1} f(r_i=0 \mid y_i,\theta_{R|Y})^{1-r_i} f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R}), \quad [4]$$

with normalization constant

$$Q(\theta_{Y|R=1},\theta_{R|Y}) = \left(1+\int \frac{f(r_i=0 \mid y_i,\theta_{R|Y})}{f(r_i=1 \mid y_i,\theta_{R|Y})} f^{\text{obs}}(y_i|\theta^{obs}_{Y|R})\, dy_i\right)^{-1}. \quad [5]$$

Theorem 1 documents that the normalization constant Q is the population fraction of data that are observed. In the class of "exponential-tilt pattern-mixture" models (13), it is assumed that

*In Tukey's representation $\theta_Y = (\theta^{obs}_{Y|R}, \theta_R, \theta_{R|Y})$.

the missing-data density is $f^{\text{mis}}(y_i) = e^{m(y_i)} f^{\text{obs}}(y_i)$ for some function m. Theorem 2 shows that with Tukey's representation

$$f^{\text{mis}}(y_i) \propto \frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} f^{\text{obs}}(y_i).$$

As such, Tukey's representation can be expressed in the exponential-tilt framework where $m(y_i) = \log\left(\frac{f(r_i=0|y_i,\theta_{R|Y})}{f(r_i=1|y_i,\theta_{R|Y})}\right) +$ const. Tukey's representation focuses on parameterizing the meaningful missingness mechanism $f(r_i = 1 \mid y_i, \theta_{R|Y})$, rather than on a hard to interpret "exponential-tilt function."

With Bayesian inference with missing data, at each iteration of a Markov chain Monte Carlo (MCMC) procedure, missing data often are imputed, and so in such a setting, it is advantageous for the missing-data density to have a tractable form. Below, we introduce a class of models for which computation of the normalization constant $Q(\theta_{Y|R=1}, \theta_{R|Y})$ is tractable and implies simple distributional forms for both the missing-data and complete-data distributions.

Exponential-Family Models. Suppose the observed-data distribution belongs to an exponential family and that the logit of the missingness mechanism is linear in the sufficient statistics of that family. Formally, let f^{obs} be an exponential-family distribution with natural parameter $\theta^{\text{obs}}_{Y|R} = \eta$, that is,

$$f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R}) = h(y_i) g(\eta) e^{T(y_i)'\eta}, \qquad [6]$$

where $g(\eta)$ is the normalization function and $T(y_i)$ is the natural sufficient statistic. A logistic missingness mechanism in $T(y_i)$, with $\theta_{R|Y} = (\alpha, \beta)$,

$$f(r_i = 1 \mid y_i, \theta_{R|Y}) = \text{logit}^{-1}\left(\alpha + T(y_i)'\beta\right) \qquad [7]$$

$$= \frac{1}{1 + e^{-\alpha - T(y_i)'\beta}}, \qquad [8]$$

implies that

$$\frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} = e^{-\alpha - T(y_i)'\beta}.$$

Then, as shown in Theorem 3, the normalization function Q in Eq. **5** can be written as a simple function of the normalization constant $g(\cdot)$ in the exponential-family formulation of f^{obs},

$$Q(\theta^{\text{obs}}_{Y|R}, \theta_{R|Y}) = \frac{g(\eta + \beta)}{g(\eta + \beta) + e^{\alpha} g(\eta)}. \qquad [9]$$

For the class of exponential family (EF)-logistic models defined by Eqs. **6** and **7**, the missing-data distribution, as specified in Eq. **10**, is from the same exponential family as the observed data with natural parameter $\theta^{mis}_{Y|R}$ (Theorem 3).† Here, we have $\theta^{\text{mis}}_{Y|R} = \eta + \beta$, and

$$f^{mis}(y|\eta, \beta) = h(y) g(\eta + \beta) e^{T(y)'(\eta+\beta)}. \qquad [10]$$

This statement is formalized in Theorem 2. Missing-data imputation in this model class is straightforward. The EF-logistic model corresponds to an exponential-tilt pattern-mixture model where $m(y_i)$ is parameterized as a linear function of the sufficient statistics of the observed-data distribution. Since a large source of uncertainty about any inferential target is due to the missingness mechanism, specifying a scientifically justifiable prior distribution for $\theta_{R|Y}$ is important.

In *Modeling Assumptions* and in *SI Appendix*, we describe extensions of this basic logistic-EF model.

Estimation and Inference. The primary estimands of interest are typically functions of the parameters specifying the complete-data distribution, θ_Y. Because the observed- and missing-data densities for EF-logistic models are both exponential families (Eqs. **6** and **10**), the complete-data distribution is a mixture of exponential families. In this paper, we focus on Bayesian inference for θ_Y via the likelihood, which is obtained as a product of terms (4) for each observation. In the EF-logistic model, analytic expressions for the normalization constant Q and the likelihood are available, and thus standard MCMC methods are applicable (23).

Note that an alternative strategy for estimation is to estimate $f^{\text{obs}}(y_i)$ nonparametrically using the empirical distribution of the observed data. When the complete-data estimands are simple univariate summaries of the complete-data distribution, this strategy is straightforward because it is easy to integrate a function with respect to the empirical distribution. However, this approach ignores the uncertainty about the true observed data density and may also suffer because, as a consequence of the positivity condition of Brook's lemma, Tukey's representation is appropriate only when the missing-data density is absolutely continuous with respect to the observed-data distribution. Using the empirical distribution as if it were the true observed-data distribution implies that the complete data, missing data, and observed data must all have the same discrete support as the finite observed data.

We take a simple approach to computation via MCMC, which is computationally less demanding than alternative methods that characterize the geometry of the solution space. Consider a simple normal-logistic model for illustration, with $T(y_i)' = (y_i, y_i^2)$ and $\theta_{R|Y} = (\alpha, \beta_1, \beta_2)$ with α corresponding to the intercept and $\beta = (\beta_1, \beta_2)'$ the rate at which the odds of selection change in y_i and y_i^2. Here, we assume the observed data follows a standard normal distribution, for convenience. Later, we consider more general cases, including a normal distribution with arbitrary mean μ and SD σ (*Modeling Assumptions*). Concretely,

$$f(r_i = 1 \mid y_i, \alpha, \beta) = \text{logit}^{-1}\left(\alpha + \beta_1 y_i + \beta_2 y_i^2\right) \qquad [11]$$

$$f^{\text{obs}}(y_i) = \text{Normal}\,(0, 1),$$

where the standard normal distribution has fixed and known natural parameters $\theta^{\text{obs}}_{Y|R} = (\eta_1, \eta_2) = (0, -1/2)$. Rather than specifying a prior distribution on $(\alpha, \beta_1, \beta_2)$, we specify a prior distribution on Q and β, but not on α. We then solve Eq. **9** for α to obtain

$$\alpha(Q, \eta, \beta) = \log\left(\frac{g(\eta_1 + \beta_1, \eta_2 + \beta_2)}{g(\eta_1, \eta_2)} \frac{(1 - Q)}{Q}\right), \qquad [12]$$

for general (η_1, η_2). For the normal distribution, $g(\eta_1, \eta_2) = \sqrt{-\eta_2} e^{\frac{\eta_1^2}{4\eta_2}}$.‡ For simplicity, assume $\beta_2 = 0$, that is, the log odds of missingness is linear in y_i only, and recall that the natural parameters for the standard normal are $\theta^{\text{obs}}_{Y|R} = (\eta_1, \eta_2) = (0, -1/2)$. Then, Eq. **12** simplifies to

$$\alpha = \frac{-\beta_1^2}{2} + \log\left(\frac{1 - Q}{Q}\right).$$

†In multiparameter exponential families, η and β are assumed to be column vectors of the same length. See the normal-logistic example in *Estimation and Inference*.

‡Note that, by convention, all multiplicative constants are part of the base measure, $h(y)$.

We use this strategy for inference to demonstrate the utility of Tukey's representation.

Illustration on Transcriptomic and Proteomic Data. We demonstrate the utility of Tukey's representation by revisiting a recent analysis of biological data aimed at quantifying the coordination between transcription and translation (24).

Scientific Question and Data. In experiments involving measurements of transcriptomic and proteomic data, messenger RNA (mRNA) transcripts and proteins that occur at low levels are less likely to be observed (25, 26). This makes it challenging to infer normalization constants for absolute protein levels (27), cluster genes into functionally related sets (28), infer the degree of coordination between transcription and translation (24), and determine the ratio of dynamic range inflation from transcript to protein levels (29). Here, we demonstrate how data analysis with Tukey's representation can be used to investigate some of these issues by assessing the sensitivity of estimands to different assumptions about the missingness mechanism.

In this analysis, we explore imputation of missing values in a dataset of mRNA and protein abundances in yeast *Saccharoymyces cerevisiae* in exponential-growth phase. We model transcript measurements (mRNA) from ref. 30 and protein measurements from ref. 31 on 5,854 genes. About 14% of the transcript measurements have missing values ($n_{\text{obs}} = 5034$), while about 36% missing of the protein measurements have missing values ($n_{\text{obs}} = 3747$). These data were gathered in experiments that were designed in part to understand the degree of coordination between transcription and translation, as well as to identify the relative dynamic ranges of transcript and protein abundances. We treat the complete-data mean and variance as estimands of interest in the analysis. We also consider the ratio of SDs from the two datasets as the quantity that describes the relative inflation of dynamic range between mRNA and protein. Note that in this application, we focus on estimands that are functions of marginal quantities only and ignore the dependence between mRNA and protein levels. For more complex estimands, one approach to modeling the multivariate structure would be to incorporate a copula, in addition to the marginal models, which describes the dependence between observations.

Modeling Assumptions. It is standard to assume that both complete-data mRNA and protein levels are log-normally distributed (32, 33), although this assumption may not be justified (34, 35) and is also not testable. Here, we use Tukey's factorization; we model the observed data as a mixture of normal distributions and specify a prior distribution for the parameters of the logistic missingness mechanism, instead of modeling the complete data directly. Together, these assumptions imply a more flexible distribution over complete-data densities. We found that a mixture with $K = 3$ components gave a reasonable approximation to the observed-data density.

In experiments measuring transcript and protein abundances, molecules that occur at lower abundances are typically much harder to measure. Thus, we expect a nonignorable missing-data mechanism in which the probability of observation decreases monotonically with decreasing abundance. Evidence suggests that a logistic missingness mechanism with a strictly positive slope, β, is plausible (36). However, as noted in ref. 37, missing values can occur for multiple reasons, at different stages of the data-collection process. Thus, we generalize the EF-logistic model to allow the selection mechanism to have a logistic form that asymptotes at some value less than one. The observed-data distribution and missingness mechanism together define the joint distribution for a single observation i:

$$f^{\text{obs}}(y_i | r_i = 1, \theta^{\text{obs}}_{Y|R}) \sim \sum_{k=1}^{K} w_k N(y_i; \mu_k, \sigma_k^2), \qquad [13]$$

$$f(r_i = 1 | y_i, \theta_{R|Y}) = \frac{\kappa e^{\beta y_i + \alpha}}{1 + e^{\beta y_i + \alpha}}, \qquad [14]$$

with $\theta^{\text{obs}}_{Y|R} = (\mu, \sigma, w)$ and $\theta_{R|Y} = (\alpha, \beta, \kappa)$. Here, $0 < (1 - \kappa) < 1$ corresponds to the fraction of data that is missing completely at random, and α and β describe the odds of a missing value, with β parameterizing the rate at which the odds of a missing value change in y_i. Under this model, the implied missing-data distribution is

$$\begin{aligned} f^{\text{miss}}(y_i | r_i = 0, \beta, \kappa, w_k, \mu_k, \sigma_k) \\ = (1 - \kappa^*) f^{\text{obs}}(y_i | r_i = 1, \mu_k, \sigma_k) \\ + \kappa^* \left(\sum_{k=1}^{K} w_k^* N(y_i; \mu_k + \beta \sigma_k^2, \sigma_k^2) \right). \end{aligned} \qquad [15]$$

The full derivation of the mixture weights w_k^* and κ^* (functions of w_k and κ) is given in *SI Appendix*.

To complete the specifications for the analysis, we propose prior distributions for the parameters Q, β and κ. As in the normal-logistic example in *Estimation and Inference*, rather than specifying a prior distribution for α, we specify a prior distribution on Q, which is well identified. Under this specification, α is a deterministic function of Q, β and κ. Computing the value of α is unnecessary as it does not appear in the missing-data density. In this application, our prior specification is

$$\begin{aligned} \beta &\sim \text{Beta}(1, 3) \\ Q &\sim \text{Uniform}(0, 1) \\ \kappa \mid Q &\sim 1 - (1 - Q)\text{Beta}(1, 2), \quad \kappa \geq Q. \end{aligned} \qquad [16]$$

We chose a uniform prior on Q, the population fraction of observed data. The results are not sensitive to this choice because the population fraction of observed data is well identified. On the contrary, κ and β are not estimable from data, and thus the results are very sensitive to the prior specifications. Ideally, these specifications should incorporate as much expertize and knowledge about the measurement technology or observation mechanism as possible. In the application we consider, Karpievitch et al. (37) and other authors reported that Missing Completely at Random (MCAR) censoring is expected to affect a relatively small proportion of the proteins (e.g., $< 20\%$). The parameter κ captures the fraction of missing data that is abundance-specific, not MCAR. Under our prior specification, the fraction of the missing data that is MCAR follows a $\beta(1, 2)$ distribution. This prior has high variance (reflecting our uncertainty) but implies on average that one-third of the missing data is MCAR. Note that κ must be greater than Q because the selection probabilities cannot be less than the population fraction of observed data.

The prior on β specifies beliefs about the rate at which log odds of missingness change. This depends on the measurement technology and experimental design and thus is expected to be quite variable across datasets. From Eq. **15**, we see that the mean of each missing-data component corresponds to a location shift of $\beta\sigma_k^2$ of the observed-data mean. This intuition can help us specify plausible priors on β. If we set $\beta = 0$, then all missingness is MCAR. Alternatively, one may want to calibrate the prior for β using ancillary data about the sensitivity of the measurement technology.

Draws from the full prior distribution are shown in Fig. 1, *Top Left* in gray, around the prior mean in black.

Data Analysis. Fig. 1, *Bottom Left* shows the fit to the protein measurements (31), when β is set to its median posterior

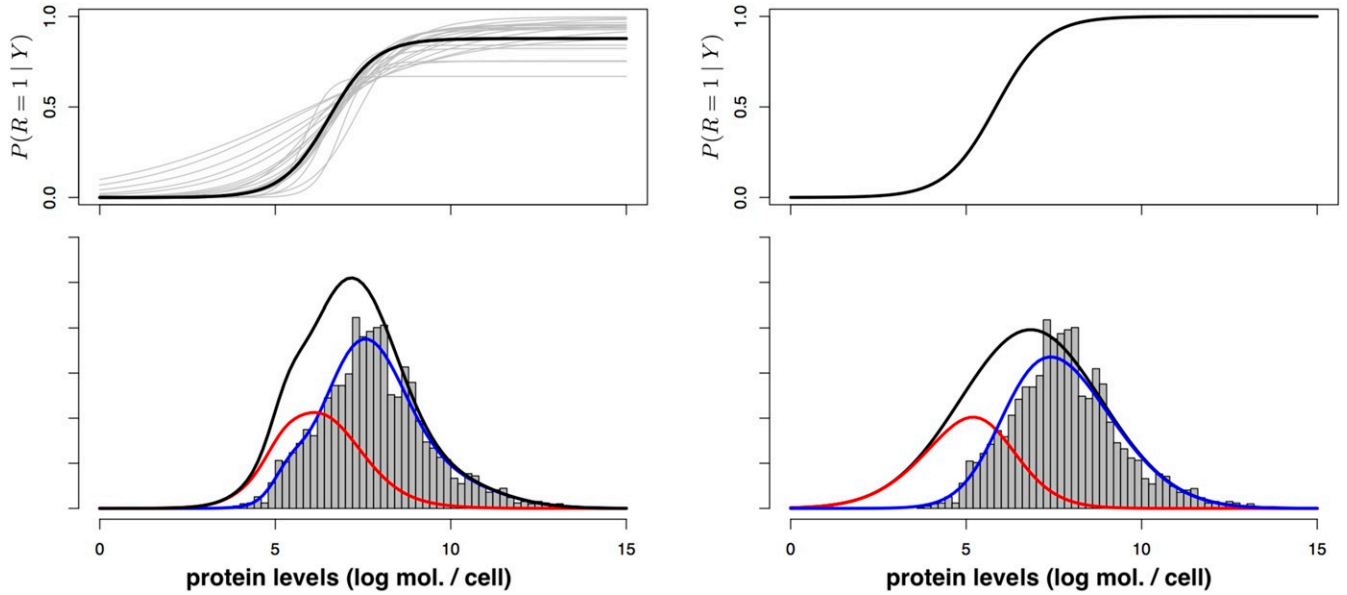

Fig. 1. Model fit to proteomic abundance data (log molecules per cell [log mol./cell]) from ref. 31 data using two approaches: Tukey's representation (*Left*) and the selection factorization (*Right*). The gray lines in *Top Left* represent draws of the selection mechanism from the prior distribution provided in Eq. **16**. The black, red, and blue lines in *Bottom Left* and *Bottom Right* correspond to the estimated densities of complete data, missing data, and observed data, respectively.

value. For comparison, Fig. 1, *Bottom Right* shows the fit of the selection-factorization model in refs. 24 and 29, which assumes the complete data are distributed according to a lognormal and the missingness mechanism is logistic, with the mean linear in y_i. The black, red, and blue lines, in both Fig. 1, *Bottom Left* and *Bottom Right*, correspond to the estimated densities of complete data, missing data, and observed data, respectively. Note that under the selection model, the estimated observed-data density (blue) is a poor fit to empirical distribution of the observed data, especially near the mode. The corresponding results for the transcript measurements (30) are provided in *SI Appendix*.

In Fig. 2, we compare the estimated complete-data means and SDs for the protein measurements using Tukey's representation model implied by Eqs. **13** and **16** to the estimates from the selection-factorization model in ref. 24. In the selection-model parameterization, not all parameters can be estimated from data (38). We found empirically that the likelihood of the selection factorization has two modes: one where $\kappa = 1$ (corresponding to the usual selection model with logistic asymptote of 1) and one where $\kappa = \frac{N_{\text{obs}}}{N}$ (corresponding to a fully MCAR model). For both sets of measurements, the estimates obtained with the MCAR model and with the selection-factorization model bracket the estimates obtained with Tukey's representation. Under the selection-factorization model, the complete-data SD is large and the mean is small relative to the estimates from Tukey's representation. The results suggest that the parametric assumptions associated with the selection-factorization models overly constrain the fit to the observed data.

Table 1 reports exact numerical estimates of the two estimands of interest.

Recent published analyses of data using the selection factorization found that translational regulation widens the dynamic range of protein expression (24, 29). One way to quantify the relative dynamic ranges of mRNA and protein is by computing the ratio of the SDs between log-mRNA and log-protein levels. A value of this ratio less than 1 suggests that the dynamic range of protein levels is smaller than that of mRNA and is taken as evidence of a suppressive role of translational regulation. A value greater than 1 is taken as evidence of amplification.

We used posterior estimates of the complete-data SDs, obtained from the three competing models fit to both protein and mRNA measurements (30, 31) to estimate the distribution of the dynamic range ratios, displayed in Fig. 3. The results obtained with Tukey's representation are consistent with those reported by ref. 29, suggesting that translational regulation reflects amplification of protein levels.

This brief case study demonstrates the relative ease of applied data analysis with Tukey's representation models and the increased flexibility of models specified using this full conditional specification. By directly modeling the observed data, we avoid the need for Monte Carlo integration of the missing data and do not require parametric specifications for the complete-data density as is typical for selection models. By modeling the selection function directly, we are also able to express uncertainty about the missing-data density beyond the simple location and scale changes typical in pattern-mixture model sensitivity analyses.

Discussion

Tukey's representation provides a powerful alternative for specifying missing-data models. It allows analysts to eschew some difficult questions about identifiability in models for nonignorable missing data (38) by factoring the joint distribution of the complete data, Y, and missing-data indicators, R, in such a way that the missingness mechanism is the only component that must rely on assumptions unassessable using observed data.

Theory. Thus far, we largely worked with exponential-family models. Here, we make formal statements about exponential-family models, as well as give results that hold more generally.

Theorem 1. *The normalization constant $Q(\theta^{\text{obs}}_{Y|R}, \theta_{R|Y})$, given in Eq. **5**, is equal to the population fraction of observed data:*

$$\left(1 + \int \frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R})\, dy_i\right)^{-1} = \mathbb{E}[r_i \mid \theta^{\text{obs}}_{Y|R}, \theta_{R|Y}].$$

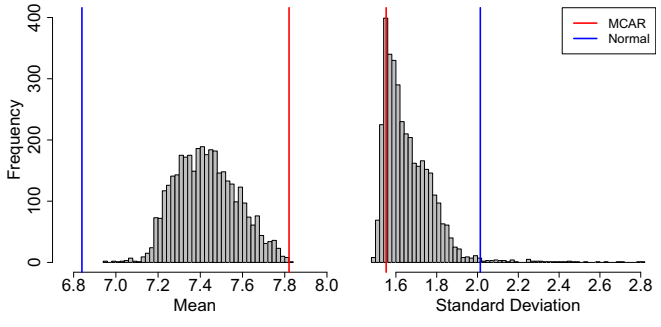

Fig. 2. Posterior distributions of the complete-data mean (*Left*) and complete-data SD (*Right*) for protein data (31). The MCAR estimates (red) and an estimate assuming normality of the complete data (blue) are shown as vertical lines for comparison. Under the prior distribution in Eq. **16**, estimates using the MCAR and the selection-factorization models are at opposite ends of these posterior distributions.

Proof (SI Appendix): A consequence of Theorem 1 is that the missing-data density can be expressed as a function of the observed-data density.

Theorem 2. *If the positivity condition is satisfied, i.e., f^{mis} is absolutely continuous with respect to f^{obs}, then f^{mis} can be expressed as a function of the observed-data density and the selection function*

$$f^{\text{mis}}(y_i \mid \theta_{R|Y}, \theta_{Y|R}) = \frac{Q(\theta_{Y|R}, \theta_{R|Y})}{1 - Q(\theta_{Y|R}, \theta_{R|Y})} \frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} f^{\text{obs}}(y_i \mid \theta_{Y|R}). \quad [17]$$

Proof (SI Appendix): This result is a consequence of setting $r_i = 0$ in the complete-data likelihood (Eq. **4**). Eq. **17** can help assess the plausibility of various missingness mechanisms—not at random, completely at random, and at random (5)—by viewing them as functions of the odds of a missing value versus an observed value, $\frac{f(r_i=0|y_i,\theta_{R|Y})}{f(r_i=1|y_i,\theta_{R|Y})}$. For instance, when the odds have low variance, it may be more reasonable to assume the missing-data mechanism is completely at random, or at random.

Eq. **17** also leads to a general understanding of the main result regarding exponential families, which can be summarized in the following statement.

Theorem 3. *Assume the observed-data distribution, $f^{\text{obs}}(y \mid \theta^{\text{obs}}_{Y|R})$, belongs to an exponential family, with natural parameter $\theta^{\text{obs}}_{Y|R} = \eta$ and natural sufficient statistic $T(y)$, and that the selection function, $f(r = 1 \mid y, \theta_{R|Y}) = \text{logit}^{-1}(\alpha + T(y)'\beta)$ with $\theta_{R|Y} = (\alpha, \beta)$. Then the implied missing-data distribution, $f^{\text{mis}}(y \mid \theta^{\text{mis}}_{Y|R})$, is in the same exponential family as the observed-data distribution, with natural parameter $\theta^{\text{mis}}_{Y|R} = \eta + \beta$. When $T(y)$ is P-dimensional, η and β are also P-dimensional vectors. The normalization constant of the complete-data distribution has the form*

$$\left(1 + \int \frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} f^{\text{obs}}(y_i \mid \theta^{\text{obs}}_{Y|R})\, dy_i\right)^{-1} = \frac{g(\eta + \beta)}{g(\eta + \beta) + e^{\alpha} g(\eta)}, \quad [18]$$

where $g(\eta)$ is the expression for the normalization constant in the exponential family.

Proof (SI Appendix): For Corollary 1, assume that the observed-data distribution is a K-component mixture of distributions in a common exponential family, with natural parameters $\theta^{\text{obs}}_{Y|R} = (\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(K)})$, natural sufficient statistic $T(y)$, and mixture weights w_k,

$$f^{\text{obs}}(y \mid \theta^{\text{obs}}_{Y|R}) = \sum_{k=1}^{K} w_k f^{\text{obs}}(y \mid \eta^{(k)}),$$

and that the selection function is logistic in $T(y)$, $f(r = 1 \mid y, \theta_{R|Y}) = \text{logit}^{-1}(\alpha + T(y)'\beta)$. Then the implied missing-data distribution, $f^{\text{mis}}(y|\theta^{\text{mis}}_{Y|R})$, is a K-component mixture of distributions in the same exponential family as the observed-data components, with natural parameters, $\theta^{\text{mis}}_{Y|R} = (\eta^{(1)} + \beta, \eta^{(2)} + \beta, \ldots, \eta^{(K)} + \beta)$, and weights

$$w_k^* = w_k \frac{g_k\left(\eta^{(k)}\right)}{g_k\left(\eta^{(k)} + \beta\right)} \Bigg/ \left(\sum_{k=1}^{K} w_k \frac{g_k\left(\eta^{(k)}\right)}{g_k\left(\eta^{(k)} + \beta\right)}\right),$$

where $g(\eta)$ is the expression for the normalization constant in the common exponential family.

Tukey's representation can be extended to model incomplete data accounting for observed covariates, for instance, by simply conditioning on x,

$$f(y_i, r_i \mid x_i, \theta) \propto \prod_{i=1}^{N} \left[f(y_i \mid r_i = 1, x_i, \theta^{\text{obs}}_{Y|R}) \cdot \frac{f(r_i \mid y_i, x_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, x_i, \theta_{R|Y})}\right]. \quad [19]$$

The factor $f(y_i \mid r_i = 1, x_i, \theta^{\text{obs}}_{Y|R})$ is estimable from the observed values. A potential challenge in applications that include covariates is the need to specify the selection probabilities, $f(r_i \mid y_i, x, \theta_{R|Y})$, for all values of x.

We can also apply Tukey's factorization with multivariate data in situations where we have a monotone missing-data mechanism (39). A missing-data pattern is called "monotone" if $\vec{y}_i$ is a K-dimensional multivariate random variable that can be ordered such that if y_i^j is missing, then all variables y_i^k, $k > j$ are also missing. In this case, the complete-data distribution can be written using Tukey's representation as

$$f(\vec{y}, \vec{r} \mid \theta) \propto \prod_{i=1}^{N} \prod_{k=1}^{K} \left[f\left(y_i^k \mid r_i^k = 1, r_i^{k-1} = \ldots = r_i^1 = 1, y_i^{k-1}, \ldots, y_i^1, \theta^{\text{obs}}_{Y|R}\right) \times \frac{f(r_i^k \mid y_i^k, r_i^1 = \ldots = r_i^{k-1} = 1, y_i^1, \ldots, y_i^{k-1}, \theta_{R|Y})}{f(r_i^k = 1 \mid y_i^k, r_i^1 = \ldots = r_i^{k-1} = 1, y_i^1, \ldots, y_i^{k-1}, \theta_{R|Y})}\right], \quad [20]$$

Table 1. Estimates for the quantities of interest obtained with three models, from protein and mRNA data

Estimand	Tukey's Rep.	Selection	MCAR	Dataset
Mean	7.42 (7.18, 7.73)	6.84	7.82	Prot.
SD	1.66 (1.52, 1.94)	2.01	1.55	Prot.
Mean	0.51 (0.44, 0.59)	0.35	0.60	mRNA
SD	1.13 (1.07, 1.23)	1.23	1.08	mRNA
Ratios	1.48 (1.28, 1.73)	1.62	1.43	Both

The dynamic range ratios are computed using both datasets. We report maximum-likelihood point estimates for both the MCAR and selection models. We report posterior medians and 95% posterior intervals (in parentheses) for Tukey's representation. Units are in log molecules per cell. Tukey's Rep., Tukey's representation.

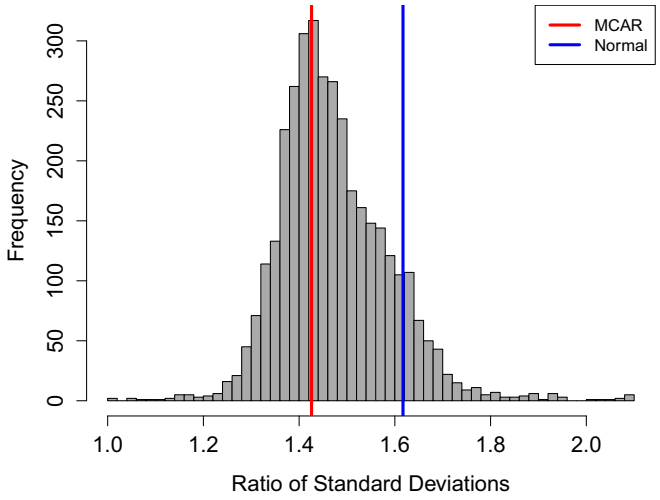

Fig. 3. Posterior distribution of dynamic range ratios obtained using Tukey's representation (histogram), the maximum-likelihood estimation dynamic range on the normal selection-factorization model (blue) (29) and the MCAR model (red).

where, as in the univariate setting, $f(y_i^k \mid r_i^k = 1, r_i^{k-1} = \ldots = r_i^1 = 1, y_i^{k-1}, \ldots, y_i^1, \theta_{Y|R}^{\text{obs}})$ is observed but $f(r_i^k = 1 \mid y_i^k, r_i^1 = 1, \ldots, r_i^{k-1} = 1, y_i^1, \ldots, y_i^{k-1}, \theta_{R|Y})$ is not.

Connections to Exponential Tilting. Tukey's representation has an interesting connection to exponential-tilting methods (13–15). Both approaches model the missing-data distribution by modifying the observed-data distribution by a multiplicative factor. However, the strategies to obtain such factors, and their interpretations, differ.

In exponential tilting, we write $f^{\text{mis}}(y)$ directly, using a tilting function $q(y)$, as

$$f^{\text{mis}}(y) = \frac{e^{-q(y)}}{\int e^{-q(y)} f^{\text{obs}}(y)\, dy} f^{\text{obs}}(y),$$

whereas Tukey's representation induces $f^{\text{mis}}(y)$, indirectly, as

$$f^{\text{mis}}(y) = \frac{Q}{1-Q} \frac{1-\pi(y)}{\pi(y)} f^{\text{obs}}(y).$$

The multiplication factor in the exponential-tilting formulation is a direct consequence of the choice of $q(y)$. The multiplication factor in Tukey's representation is a function of the fraction of observed data Q and of the odds of missingness.

The relation between exponential tilting and Tukey's representation is similar in spirit to the relation between a frequentist penalized likelihood and a posterior distribution, in which the penalty is implied by the choice of a prior distribution. Tukey's representation offers a principled way to derive specific forms of $q(y)$ from assumptions made on f^{mis}.

Importantly, however, exponential tilting requires specifying the function $q(y)$ somewhat arbitrarily, and because of that, the data analysis may often require a serious sensitivity analysis, for the results to be defensible. In contrast, Tukey's representation requires the specification of a fully generative model, the pieces of which are arguably easier to defend in scientific applications.

To explore this equivalence in more detail, we write f^{mis} in exponential-tilting form by exponentiating the log of the multiplicative factor from Theorem 2,

$$\begin{aligned} &f^{\text{mis}}(y_i \mid \theta_{R|Y}, \theta_{Y|R}) \\ &\quad = \exp\left\{ \log\left(\frac{Q(\theta_{Y|R}, \theta_{R|Y})}{1 - Q(\theta_{Y|R}, \theta_{R|Y})} \frac{f(r_i = 0 \mid y_i, \theta_{R|Y})}{f(r_i = 1 \mid y_i, \theta_{R|Y})} \right) \right\} f^{\text{obs}} \\ &\qquad \times (y_i \mid \theta_{Y|R}). \end{aligned}$$

The right hand side of this equation can be seen as $\exp\{-q(y_i, \theta_{\text{tilt}})\} f^{\text{obs}}(y_i \mid \theta_{Y|R})$, where $q(y, \theta_{\text{tilt}})$ is a function-valued sensitivity parameter specified in exponential-tilting models. In Tukey's representation, when using the logistic selection function, the exponential factor has interpretable components: the log odds of missingness and the log odds of selection. More generally, the parametric form for q is often a complicated function of the selection parameters, even when the equivalent selection function in Tukey's representation is easily interpretable.

Tukey's representation opens the door to more transparent analyses in problems that involve missing data. For example, in the model in *Illustration on Transcriptomic and Proteomic Data*, we can derive the implied tilting function (*SI Appendix*) given the missingness mechanism $f(r_i \mid y_i, \theta_{R|Y})$ and the normalization constant $Q(\theta_{Y|R}, \theta_{R|Y})$. Here, we have the following exponential-tilt function:

$$\begin{aligned} -q(y, \theta_{\text{tilt}}) = &\log\left(e^{-\alpha} \left(\sum_k^K \frac{w_k}{\kappa} \frac{g(\eta)}{g(\eta^*)} \right) + \frac{1-\kappa}{\kappa} \right) \\ &+ \log\left(\frac{1}{\kappa} e^{-\beta y - \alpha} + \frac{1-\kappa}{\kappa} \right), \end{aligned} \qquad [21]$$

where $\eta^* = \eta + \beta$. The complex nature of this function highlights the importance of prior specification for the selection function, as opposed to the tilting function. As another example, in this paper, we focus on cases where we specify a prior distribution for β, the rate at which the odds of selection change in $T(y_i) = y_i$. We show that for fixed β, α is identified in the two-parameter logistic model. However, in some applications, we may have more prior knowledge about α, the log odds of missingness given $y_i = 0$. In this case, if we specify a prior distribution on α, then β is identified. This further illustrates how, in many contexts, it is easier to elicit priors and justify parameterizations for the fully generative specification of the selection function.

Even in light of specific equivalences for specific choices of the various factors, working with the $q(y)$ function implied by Tukey's generative approach may be more easily defensible, say, in medical, biological, economics, or legal contexts.

A Note on the Integrability Condition. Not all integrable specifications for $f^{\text{obs}}(y_i \mid \theta_{Y|R}^{\text{obs}})$ and $f(r_i \mid y_i, \theta_{R|Y})$ imply a proper distribution for $f^{\text{mis}}(y_i \mid \theta_{Y|R}^{\text{obs}}, \theta_{R|Y})$. The integrability condition requires the sum $\theta_{Y|R}^{\text{obs}} + \theta_{R|Y}$ to lie in the natural parameter space of the exponential family. In practice, analysts may want to consider missing-data mechanisms that involve a richer set of parameters, $\tilde{\theta}_{R|Y}$, such as including an intercept, as demonstrated in *Illustration on Transcriptomic and Proteomic Data*. In such cases, $\theta_{R|Y}$ is taken to denote the subset of parameters in $\tilde{\theta}_{R|Y}$ that multiply the sufficient statistics of f^{obs}. The derivations in *Modeling and Inference Using Tukey's Representation* can easily be extended to this situation, accordingly.

For example, assume that the natural parameter $\theta_{Y|R}^{\text{obs}} = \eta$, and that the missing-data mechanism is logistic with extended parameter vector $\tilde{\theta}_{R|Y} = (\alpha, \beta) = (\alpha, \theta_{R|Y})$ and $f(r_i = 1 \mid y_i, \alpha, \beta) = (1 + e^{-(\alpha + T(y_i)'\beta)})^{-1}$. Then, Eqs. **9** and **10** become

$$Q(\eta,\beta)=\frac{g(\eta+\beta)}{g(\eta+\beta)+g(\eta)e^{\alpha}}, \quad [22]$$

$$f^{\text{mis}}(y\mid\eta,\beta)=h(y)g(\eta+\beta)e^{T(y)'(\eta+\beta)}. \quad [23]$$

The class of EF-logistic models defined in Eqs. **6** and **7** can be further generalized in two useful ways, while maintaining its desirable properties. For instance, generalizing f^{obs} to be a mixture of exponential families is straightforward (Corollary 1) and does not increase computation substantially. Relaxing assumptions about the missingness mechanism can be more difficult. Still, it is possible to model $f(r_i\mid y_i,\theta_{R|Y})$ with a mixture of logistic functions, including a missingness mechanism where a fraction of the data is missing completely at random (as is shown in the applied example).

Inferential Strategies. Recall the simple normal-logistic model,

$$f(r_i=1\mid y_i,\alpha,\beta)=\text{logit}^{-1}(\alpha+\beta y_i)$$
$$f^{\text{obs}}(y_i)=\text{Normal}\,(0,1).$$

The inferential strategy proposed was to posit prior distributions on β and Q. Each iteration of the MCMC sampler yields a sample for β and Q and implied sample for α. A conceptually simpler approach to inference would be to place a prior distribution on all of the parameters of the missingness mechanism (i.e., α and $beta$) and solve for the implied Q at each iteration of the sampler.

In situations where the number of missing values is itself missing, as with truncated data, specifying a prior distribution for all of the parameters of the missingness mechanism would lead to an implied prior distribution for the unknown number of missing values, or equivalently, the population fraction of observed data Q; this approach was also used in the original expectation-maximization algorithm paper (40).

In situations where the number of missing values is known, however, as with censored data, and therefore Q can be estimated from observed data, the support of the likelihood is a constrained parameter space, and a number of choices for the prior distribution on β would lead to a posterior distribution that is challenging to explore using Monte Carlo methods. Specifically, because the population fraction of observed data are identifiable, Theorem 1 describes a moment constraint that restricts the region, where the parameters of the missingness mechanism have positive support, to a lower dimensional ridge. Fig. 4 illustrates this phenomenon for the simple normal-logistic model and increasing sample size.

Sequential Monte Carlo and other specialized Monte Carlo methods that exploit the geometry of the support of posterior distribution may provide solutions in this situation (41–44).

Concluding Remarks

In this paper, we used EF-logistic models to illustrate how Tukey's representation can be used to encode nonmonotonicity in the missingness mechanism and to model data with complex distributional forms. The EF-logistic models are widely applicable as they can be applied to data that are well approximated by mixtures of exponential families. Although not explored here, similar logic can be applied to facilitate inference for models with nonlogistic selection mechanisms that can be well approximated by mixtures of logistic functions. These EF-logistic models could also be used to facilitate tipping-point analyses (45) or to incorporate subjective model uncertainty via prior distributions on the missingness mechanism (2).

Tukey's representation is most useful when it is feasible to posit reasonable prior distributions on the selection mechanism. Translating expert knowledge into a functional form can be challenging, in general, and a logistic missingness mechanism is not always a good choice. In practice, Tukey's representation should be used in concert with strategies for expert prior elicitation (46–48). Nevertheless, prior elicitation for Tukey's representation is simpler than for other factorizations, because it involves only the set of parameters $\theta_{R|Y}$. In contrast, the selection factorization requires additional assumptions about the complete-data density.

In many settings, like the example presented in our applied analysis, we may be able to collect data that partially inform the specification for the selection mechanism. As such, when possible, we can design experiments to learn about the functional form of $f(r_i|y_i,\theta_{R|Y})$ as well as to further refine prior distributions for $\theta_{R|Y}$. Along these lines, Tukey's representation may be useful in the context of multiphase inference, which is intimately related to problems in missing data (49). In these problems, when preprocessing data, we often have strong knowledge (or control) of the missingness mechanism yet a weaker understanding of the underlying scientific model.

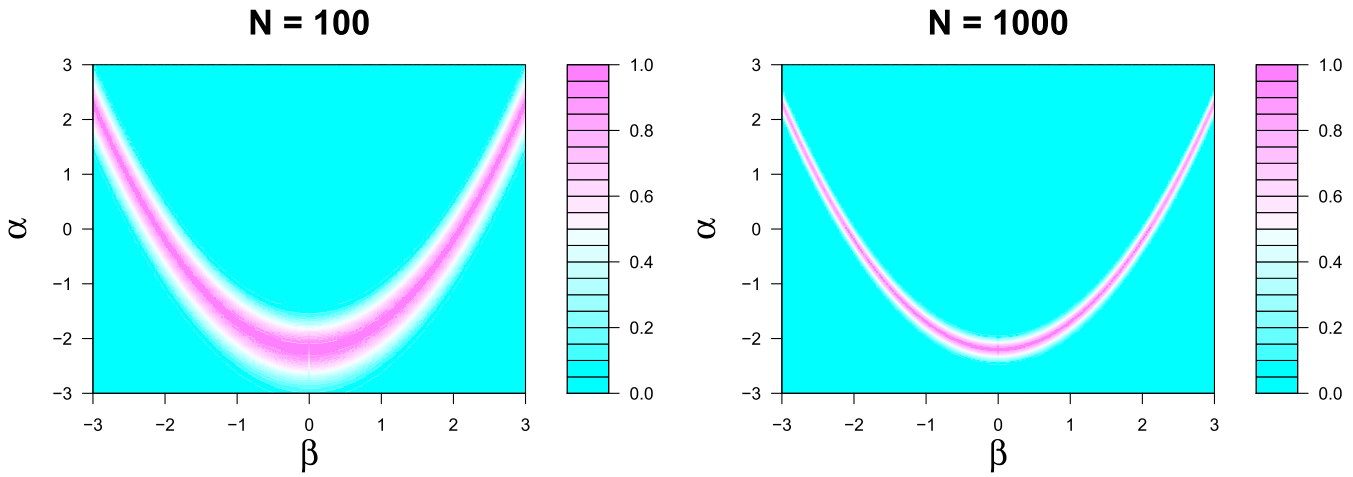

Fig. 4. The region of positive support for the likelihood, restricted to the parameters of the missingness mechanism, is increasingly constrained as the population fraction of observed data, *Q*, is estimated with increasingly high precision. This intuition is illustrated by the width of the ridge, which is a function of the amount of information about *Q*. We simulated data from a standard normal distribution and a logistic missingness mechanism. The parameters (α, β) were set to get 90% missing data. The sample size determines the amount of information: $N = 100$ (*Left*) and $N = 1{,}000$ (*Right*).

Finally, in this paper, we focus on the class of problems where the data are univariate and i.i.d. Extending this methodology to a broader class of multivariate missing-data problems is challenging. We show that Tukey's representation is easily extensible to montone missing data, where the observed-data models can easily be replaced by conditional models. For more general missing-data patterns, Tukey's representation is nontrivial. However, we believe that Tukey's representation can be a particularly useful tool for specifying joint multivariate distributions using only the full conditionals. Empirical work in this area has shown that imputation using a Gibbs sampler can be effective, even though the specified conditional densities can be incompatible (e.g., do not imply a proper joint distribution) (50, 51). In these so-called partially incompatible Gibbs samplers, each Gibbs would involve missing-data imputation of a single variable given the rest, through Tukey's representation. Such extensions are the subject of future research by us.

In conclusion, we argue that Tukey's representation, which represents a hybrid of the selection and pattern-mixture models is an underresearched yet promising alternative for modeling nonignorable missing data.

Data Availability. All raw input and processed output data are available in Dryad (DOI: 10.5061/dryad.rg367 and DOI: 10.5061/dryad.d644f).

ACKNOWLEDGMENTS. We are grateful to Dr. Shahab Jolani (Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University) and Dr. Stef Van Buuren (Netherlands Organization for Applied Scientific Research) for sharing preliminary work and analyses that contributed to the framing of this paper. We are also grateful to the two reviewers for detailed comments and suggestions that helped improve this paper.

1. R. J. Little, D. B. Rubin, *Statistical Analysis with Missing Data* (John Wiley & Sons, 2015).
2. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (Wiley Classic Library, 2004).
3. D. B. Rubin, Inference and missing data. *Biometrika* **63**, 581–592 (1976).
4. D. B. Rubin, Bayesian inference for causal effects: The role of randomization. *Ann. Stat.* **6**, 34–58 (1978).
5. F. Mealli, D. B. Rubin, Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102**, 995–1000 (2015).
6. D. B. Rubin, Characterizing the estimation of parameters in incomplete data problems. *J. Am. Stat. Assoc.* **69**, 467–474 (1974).
7. D. B. Rubin, Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.* **72**, 538–543 (1977).
8. R. J. Little, Pattern-mixture models for multivariate incomplete data. *J. Am. Stat. Assoc.* **88**, 125–134 (1993).
9. D. O. Scharfstein, A. Rotnitzky, J. M. Robins, Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.* **94**, 1096–1120 (1999).
10. R. Andrea, D. Scharfstein, T. L. Su, J. Robins, Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113 (2001).
11. R. J. Glynn, N. M. Laird, D. B. Rubin, "Selection modeling versus mixture modeling with nonignorable nonresponse" in *Drawing Inferences from Self-Selected Samples*, H. Wainer, Ed. (Springer, 1986), pp. 115–152.
12. P. Holland, "Discussion 4: Mixture modeling versus selection modeling with nonignorable nonresponse" in *Drawing Inferences from Self-Selected Samples*, H. Wainer, Ed. (Springer, 1986), pp. 143–149.
13. J. Birmingham, A. Rotnitzky, G. M. Fitzmaurice, Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *J. Roy. Stat. Soc. B Stat. Methodol.* **65**, 275–297 (2003).
14. S. Vansteelandt, A. Rotnitzky, J. Robins, Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841–860 (2007).
15. D. Scharfstein, A. McDermott, W. Olson, F. Weigand, Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Stat. Biopharm. Res.* **6**, 338–348 (2014).
16. A. R. Linero, M. J. Daniels, Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Stat. Sci.* **33**, 198–213 (2018).
17. G. Molenberghs, G. M. Fitzmaurice, M. G. Kenward, A. A. Tsiatis, G. Verbeke, Eds, *Handbook of Missing Data Methodology* (Chapman & Hall/CRC Press, 2015).
18. A. Gelman *et al.*, *Bayesian Data Analysis* (Chapman and Hall/CRC, 2013).
19. D. Brook, On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481–483 (1964).
20. J. Besag, Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. B* **36**, 192–236 (1974).
21. N. Cressie, *Statistics for Spatio-Temporal Data* (Wiley, Hoboken, NJ, 2011).
22. A. Gelman, X. Meng, A note on bivariate distributions that are conditionally normal. *Am. Statistician* **45**, 125–126 (1991).
23. C. P. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer-Verlag, ed. 2, 2004).
24. A. M. Franks, G. Csárdi, D. A. Drummond, E. M. Airoldi, Estimating a structured covariance matrix from multilab measurements in high-throughput biology. *J. Am. Stat. Assoc.* **110**, 27–44 (2015).
25. T. C. Walther, M. Mann, Mass spectrometry–based proteomics in cell biology. *J. Cell Biol.* **190**, 491–500 (2010).
26. W. W. Soon, M. Hariharan, M. P. Snyder, High-throughput sequencing for biology and medicine. *Mol. Syst. Biol.* **9**, 640 (2013).
27. Y. V. Karpievitch, A. R. Dabney, R. D. Smith, Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinf.* **13**, S5 (2012).
28. O. Troyanskaya *et al.*, Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
29. G. Csárdi, A. Franks, D. S. Choi, E. M. Airoldi, D. A. Drummond, Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* **11**, e1005206 (2015).
30. V. Pelechano, J. E. Pérez-Ortín, There is a steady-state transcriptome in exponentially growing yeast cells. *Yeast* **27**, 413–422 (2010).
31. S. Ghaemmaghami *et al.*, Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
32. M. Bengtsson, A. Ståhlberg, P. Rorsman, M. Kubista, Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* **15**, 1388–1392 (2005).
33. P. Lu, C. Vogel, R. Wang, X. Yao, E. M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).
34. C. Lu, R. D. King, An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. *Bioinformatics* **25**, 2020–2027 (2009).
35. N. F. Marko, R. J. Weil, Non-Gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PloS One* **7**, e46935 (2012).
36. C. Vogel, E. M. Marcotte, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
37. Y. Karpievitch *et al.*, A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **25**, 2028–2034 (2009).
38. W. Miao, P. Ding, Z. Geng, Identifiability of normal and normal mixture models with nonignorable missing data. *J. Am. Stat. Assoc.* **111**, 1673–1683 (2016).
39. G. Molenberghs, B. Michiels, M. G. Kenward, P. J. Diggle, Monotone missing data and pattern-mixture models. *Stat. Neerl.* **52**, 153–161 (1998).
40. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* **39**, 1–22 (1977).
41. A. Doucet, N. de Freitas, N. Gordon, Eds, *Sequential Monte Carlo Methods in Practice* (Springer, 2001).
42. J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, ed. 2, 2008).
43. M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. B Stat. Methodol.* **73**, 123–214 (2011).
44. E. Airoldi, B. Haas, "Polytope samplers for inference in ill-posed inverse problems" in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Journal of Machine Learning Research, Workshop & Conference Proceedings, 2011), pp. 110–118.
45. V. Liublinska, D. B. Rubin, Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat. Med.* **33**, 4170–4185 (2014).
46. A. O'Hagan *et al.*, *Uncertain Judgements: Eliciting Experts' Probabilities* (John Wiley & Sons, 2006).
47. M. Kynn, "Eliciting expert knowledge for Bayesian logistic regression in species habitat modelling," PhD thesis, Queensland University of Technology, Brisbane, Australia (2005).
48. S. M. Paddock, P. Ebener, Subjective prior distributions for modeling longitudinal continuous outcomes with non-ignorable dropout. *Stat. Med.* **28**, 659–678 (2009).
49. A. W. Blocker, X. L. Meng, The potential and perils of preprocessing: Building new foundations. *Bernoulli* **19**, 1176–1211 (2013).
50. S. Van Buuren, J. P. Brand, C. Groothuis-Oudshoorn, D. B. Rubin, Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simulat.* **76**, 1049–1064 (2006).
51. D. B. Rubin, Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57**, 3–18 (2003).