ACTIVE MATRIX FACTORIZATION FOR SURVEYS

BY CHELSEA ZHANG[1,*], SEAN J. TAYLOR[2], CURTISS COBB[3] AND JASJEET SEKHON[1,†]

[1]*Department of Statistics, University of California, Berkeley,* *cyz@berkeley.edu; †sekhon@berkeley.edu*

[2]*Lyft, Inc., seanjtaylor@gmail.com*

[3]*Facebook, Inc., ccobb@fb.com*

Amid historically low response rates, survey researchers seek ways to reduce respondent burden while measuring desired concepts with precision. We propose to ask fewer questions of respondents and impute missing responses via probabilistic matrix factorization. A variance-minimizing active learning criterion chooses the most informative questions per respondent. In simulations of our matrix sampling procedure on real-world surveys as well as a Facebook survey experiment, we find active question selection achieves efficiency gains over baselines. The reduction in imputation error is heterogeneous across questions and depends on the latent concepts they capture. Modeling responses with the ordered logit likelihood improves imputations and yields an adaptive question order. We find for the Facebook survey that potential biases from order effects are likely to be small. With our method, survey researchers obtain principled suggestions of questions to retain and, if desired, can automate the design of shorter instruments.

1. Introduction.

1.1. *Reducing response burden in surveys.* Modern surveys suffer from declining response rates which inflate administration costs and cast doubt on the validity of inferences. Research has long suspected that survey length plays a role. An inverse association between length and response rate appears in several meta-analyses of mail surveys (Edwards et al. (2002), Heberlein and Baumgartner (1978), Yammarino, Skinner and Childers (1991)). An experiment evaluating redesigns of the U.S. Census found that shortening the questionnaire increased response rate (Dillman, Sinclair and Clark (1993)). Another experiment showed a sizable negative effect of completion in relation to length in web surveys (Marcus et al. (2007)). There is disagreement about the direction and size of effect (Munger and Loyd (1988), Sheehan (2001)), although variation in reported effect sizes may be due to the disparity of survey modes and measures of length (Fan and Yan (2010)).

In addition to nonresponse, a longer instrument may be more susceptible to measurement error. Respondents may avoid the cognitive burden of surveys by taking mental shortcuts, such as selecting "don't know" or arbitrary responses, a behavior called satisficing (Krosnick (1991)). Satisficing in the form of straight-line responding—giving identical answers to consecutive items—occurs more frequently on a long instrument than a short one (Herzog and Bachman (1981)). There is also evidence that both interviewers and interviewees deliberately shortcut interviews to reduce cognitive burden, such as answering initial questions negatively in order to avoid follow-up questions (Tourangeau, Kreuter and Eckman (2015)).

To combat these issues, survey practitioners have suggested reducing respondent burden by asking fewer questions (Kreuter (2013)). This idea arose in an earlier era of survey research, when norms shifted from in-person to phone surveys; it became easier for contacts to prematurely end a survey by hanging up. Researchers adapted by administering shorter

phone surveys (Groves (2011)). Assigning a subset of questions to each respondent, otherwise known as matrix sampling, may increase response rate and reduce the bias of nonresponse (Munger and Loyd (1988)). Matrix sampling for the Consumer Expenditure Interview Survey adaptively assigns respondents to subquestionnaires to achieve minimum-variance estimates across expenditure types (Gonzalez and Eltinge (2008)). Early, Mankoff and Fienberg (2017) propose to choose questions sequentially in order to maximize the gain of information insofar as gain is traded off with dropout probability.

1.2. *The case for imputation.* These matrix sampling procedures create missingness in the response matrix which consists of the responses to all potential questions. Estimating marginal quantities, such as per-question population averages, is possible with the existing responses and optional weighting adjustment. However, in order to use complete-data methods without discarding data, missing responses must be imputed. Multiple imputation is a common approach (Reiter and Raghunathan (2007), Rubin (2004), Thomas et al. (2006)). It has been argued that surveys should quantify information content by means of a certain metric for imputation uncertainty rather than nonresponse rate (Wagner (2010)).

We list a few survey applications where imputation is useful:

1. Political scientists, campaigns and the media are interested in voter opinion on a variety of issues. One such research question concerns itself with voters holding their representatives accountable—whether agreement between constituent policy preferences and legislator votes affects constituent support for legislators (Ansolabehere and Jones (2010)). This type of analysis requires a complete set of individual responses to issue questions. A complete set of predictions is also useful for decision making. For instance, a political campaign attempting to persuade voters may want to be able to predict individual voter opinions from a minimum of questions. Armed with predictions on all issues, a canvasser could then customize her text to highlight areas of alignment between the voter and candidate.

2. Survey experiments measure the effects of treatment on survey outcomes. Whenever responses from a baseline survey are available, they can serve as pretreatment covariates for the estimation of heterogeneous treatment effects (Broockman, Kalla and Sekhon (2017)). Künzel et al. (2019) estimate treatment effects that are conditional upon baseline attitudes by way of metalearners that are built upon arbitrary supervised learning methods, such as random forests, neural networks and lasso. Although some base learners handle missingness natively, others require complete data. A sufficient sample size is crucial for many base learners which makes imputation more attractive than dropping incomplete cases.

3. Panel data collected from repeated surveys is prone to attrition. Chen, Felt and Huynh (2017) use top-up refreshment samples in order to account for nonignorable attrition in the yearly Canadian Financial Monitor which has an attrition rate of 50%. When a complete panel dataset is required for subsequent analysis, dropping incomplete cases may create survivorship bias, and imputation may be preferable (Productivity Commission (2018)). Popular methods for causal inference from panel data, such as the synthetic control method (Abadie, Diamond and Hainmueller (2010), Ben-Michael, Feller and Rothstein (2018)), take complete data as input. In fact, missingness is intrinsic to causal inference from panel data: once units undergo treatment, their control outcomes are missing. Athey et al. (2018) use matrix completion to impute these counterfactual outcomes.

4. Designers of large administrative surveys face important tradeoffs regarding which questions are worth asking and how to evolve surveys over time. Choosing from a large set of potential survey questions is particularly challenging when survey data is intended for general research usage without advance knowledge of estimands of interest. An adaptive sampling approach combined with an imputation model provides survey practitioners with a new design choice, a compromise between a short, incomplete survey and a long, expensive

and burdensome one. Though there may be challenges in adapting analyses to work with imputed data, the flexibility of a high-dimensional and easy-to-modify survey instrument could be worth the cost in many cases.

We focus on the first use case, measuring voter opinion across a wide spectrum of political questions. What we are envisioning is a less burdensome survey that asks a strategic subset of issue questions and imputes responses to the rest. The response matrix is likely to contain low-dimensional latent structure; we would expect many questions to correlate with partisanship, for instance. Udell and Townsend (2019) argue that this low-rank assumption is plausible for a general class of latent variable models.

With a sparse response matrix that is approximately low rank, a natural imputation strategy is matrix completion. Widely used in recommendation systems in order to predict user-item preferences across thousands of items, matrix completion enjoys theoretical guarantees and efficient algorithms. Multiple authors have suggested applying matrix completion to survey imputation (Candès and Recht (2009), Davenport et al. (2014), Josse, Husson et al. (2016), Klopp et al. (2015)), but published applications to real-world surveys are rare. One exception is concurrent work by Sengupta, Srebro and Evans (2018) that examines the predictive ability of matrix completion on survey responses collected by different elicitation strategies.

1.3. *The design opportunity.* With matrix factorization as the response model, we next consider how best to prioritize survey items. The field of active learning is concerned with selecting informative training points when label acquisition is expensive. A common but myopic baseline is uncertainty sampling which iteratively chooses the point with greatest predictive uncertainty. Other strategies choose the point that produces most disagreement within an ensemble or the point that would minimize expected predictive variance in the updated model (Settles (2009)).

Active learning for matrix factorization seeks the most informative entries in the response matrix. In the usual setting of a movie recommendation system, one baseline strategy simply prompts for the most popular items since users are more likely to recognize them and remain attentive (Elahi, Ricci and Rubens (2016)). Uncertainty sampling can be used with various models of unobserved matrix entries (Chakraborty et al. (2013), Sutherland, Póczos and Schneider (2013)). Influence-based strategies find the item that would produce the greatest change in predictions (Rubens and Sugiyama (2007)) or user parameters (Karimi et al. (2011a)). Silva and Carin (2012) maximize mutual information between selected and unobserved instances. Under distributional assumptions about ratings, such as stationarity, one could find the item order that directly minimizes prediction error (Golbandi, Koren and Lempel (2010), Karimi et al. (2011b)).

Although theoretical results show active learning has lower sample complexity than passive sampling in certain settings (Settles (2009)), there are several challenges to the adoption of active learning in practice. These include choosing a base learner and item selection strategy within the label budget as well as poor discovery of rare classes (Attenberg and Provost (2011)). In addition, datasets collected via active learning may lose their advantage over random sampling when transferring across models (Lowell, Lipton and Wallace (2019)).

1.4. *Optimal design in latent space.* Active learning for matrix factorization can be recast as placing respondents on latent scales with maximal precision. For our voter survey we seek questions that are most informative for locating a respondent along latent dimensions of opinion. The literature on item response theory has long pursued this goal. Computerized adaptive testing (CAT) algorithms choose questions online to precisely estimate an individual's latent ability within a fixed question budget. Montgomery and Cutler (2013) advocate for applying CAT methods to public opinion surveys. Using an item selection strategy that

minimizes expected posterior variance of a one-dimensional ability parameter, they shorten a political knowledge battery by 40% while retaining measurement accuracy.

In the general case of multiple latent dimensions, optimal item selection can take different forms. Optimal design approaches minimize the inverse Fisher information of the multidimensional ability or opinion parameter. This matrix is the asymptotic variance of the maximum likelihood estimate as well as the posterior variance under a noninformative prior. The D-optimal criterion, used by Segall (2009), minimizes the determinant of inverse information which equivalently minimizes the size of the posterior credibility region. Mulder and van der Linden (2009) argue that the A-optimal criterion, which minimizes the trace of inverse information, should act similarly to D-optimality since the trace includes the determinant as a factor. Their simulations show A- and D-optimality outperform random selection, while the E-optimal criterion of minimizing the maximum eigenvalue is worse than random.

Such optimal design criteria will guide us in shortening our survey of voter opinion. The specific form of matrix factorization will determine the adaptivity of the survey. When we model responses as Gaussian, we obtain an interpretable but deterministic question order. When we model responses with a more realistic ordered logit likelihood, we obtain an adaptive question order, but we resort to approximate inference to handle nonconjugacy. We demonstrate the gains of active question selection for imputing left-out questions in extensive survey simulations. Finally, we present experimental evidence that active question selection improves efficiency.

2. Active matrix completion.

2.1. *Matrix completion methods.* Given n users and k questions, let R denote the $n \times k$ response matrix. Matrix factorization finds a low-rank decomposition of R, a set of user factors $U = [u_1, \ldots, u_n]^T$ and question factors $V^T = [v_1, \ldots, v_k]$ such that $R \approx UV^T$. Let r be the dimensionality of latent space, typically small. Then, $u_i, v_j \in \mathbb{R}^r$ for all i and j.

When R is partially observed, matrix completion adapts matrix factorization to approximately reconstruct observed entries while predicting missing entries. Let I be an indicator matrix for whether the corresponding responses in R exist. $I_{ij} = 1$ implies user i responded to question j with value R_{ij}. Matrix completion finds U and V that minimize the reconstruction error of $u_i^T v_j$ for R_{ij} on the set $\{(i, j) : I_{ij} = 1\}$.

The formulation of matrix completion that enforces a hard rank constraint is nonconvex and generally intractable (Fithian and Mazumder (2013)). It is common to work with a convex relaxation that instead regularizes the nuclear norm or sum of singular values (Srebro, Rennie and Jaakkola (2005)). This optimization problem seeks a matrix Z, in place of UV^T, that minimizes reconstruction error. It encourages a low-rank solution by favoring sparsity in the singular values according to regularization parameter λ.

$$(1) \qquad \min_Z \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} I_{ij} (R_{ij} - Z_{ij})^2 + \lambda \|Z\|_*.$$

The nuclear norm regularized problem (1) enjoys theoretical guarantees: recovery of the complete matrix occurs with high probability when $O(n \operatorname{polylog}(n))$ entries are observed at random, with or without noise (Negahban and Wainwright (2012), Recht (2011)). Moreover, (1) has an efficient solution in the SoftImpute algorithm by Mazumder, Hastie and Tibshirani (2010). SoftImpute iteratively computes the SVD of Z, soft-thresholds the singular values and updates the entries where $I_{ij} = 0$ with the prediction from the soft-thresholded SVD, until convergence. Hence, the solution can be expressed as $Z = UDV^T$ for some matrices $U \in \mathbb{R}^{n \times r}, D \in \mathbb{R}^{r \times r}, V \in \mathbb{R}^{k \times r}$.

An alternate formulation of matrix completion, introduced by Rennie and Srebro (2005), penalizes the Frobenius norm of U and V,

(2) $$\min_{U,V} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} I_{ij} (R_{ij} - u_i^T v_j)^2 + \frac{\lambda_U}{2} \|U\|_F^2 + \frac{\lambda_V}{2} \|V\|_F^2.$$

Problem (2) is nonconvex in U and V; it is solved via gradient descent or alternating least squares (Hastie et al. (2015)). This formulation is useful for large-scale problems with low rank, since it is less expensive to operate on U and V than Z. The solutions to (1) and (2) coincide if $\lambda_U = \lambda_V$ and the solution to (1) has rank at most r, due to an identity relating the nuclear norm and sum of Frobenius norms (Fithian and Mazumder (2013)).

The user and question factors resulting from either optimization are point estimates, as are the imputed survey responses. We seek a strategy for actively selecting the next survey question based on the uncertainty reduction achieved. To quantify uncertainty over imputed responses, we turn to probabilistic matrix factorization methods.

2.2. *Probabilistic matrix factorization.* Probabilistic matrix factorization (PMF) by Salakhutdinov and Mnih (2008) models user and question factors as independently normally distributed. Responses add zero-mean, constant-variance Gaussian noise to the inner product of user and question factors. Following the original notation,

$$u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_U, \Lambda_U^{-1}),$$

$$v_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_V, \Lambda_V^{-1}),$$

$$R_{ij} \mid U, V \overset{\text{ind.}}{\sim} \mathcal{N}(u_i^T v_j, \alpha^{-1}).$$

With zero-mean, isotropic priors, MAP estimation for PMF corresponds to solving the Frobenius norm regularized problem (2). Specifically, for $\mu_U = \mu_V = 0$, $\Lambda_U = \alpha_U I$ and $\Lambda_V = \alpha_V I$, the MAP estimate of U and V conditional on R is the solution to (2) with regularization parameters $\lambda_U = \alpha_U / \alpha$ and $\lambda_V = \alpha_V / \alpha$.

Bayesian probabilistic matrix factorization (BPMF) places additional normal-Wishart priors on the hyperparameters $\mu_U, \Lambda_U, \mu_V, \Lambda_V$ (Salakhutdinov and Mnih (2008)). Posterior inference is performed by Gibbs sampling. They derive the complete conditional for u_i as follows:

$$P(u_i \mid R, V, \mu_U, \Lambda_U, \alpha) \propto P(u_i, R_{i\cdot} \mid V, \mu_U, \Lambda_U, \alpha)$$

$$= \prod_{j=1}^{k} [P(R_{ij} \mid u_i^T v_j, \alpha^{-1})]^{I_{ij}} P(u_i \mid \mu_U, \Lambda_U^{-1}).$$

The complete conditional is conjugate normal with mean μ_i^* and precision Λ_i^*,

$$u_i \mid R, V, \mu_U, \Lambda_U, \alpha \sim \mathcal{N}(\mu_i^*, [\Lambda_i^*]^{-1}),$$

$$\Lambda_i^* = \Lambda_U + \alpha \sum_{j=1}^{k} I_{ij} v_j v_j^T,$$

$$\mu_i^* = [\Lambda_i^*]^{-1} \left(\alpha \sum_{j=1}^{k} I_{ij} R_{ij} v_j + \Lambda_U \mu_U \right).$$

This can be recognized as the posterior for Bayesian linear regression with a Gaussian prior and Gaussian noise: u_i are the coefficients, observed rows of V form the design matrix and

α^{-1} is the noise variance. The expression for μ_i^* also arises in MAP estimation for PMF with zero-mean, isotropic priors; it is the coordinate ascent update for u_i. For details, see Section A.1 of the Supplementary Material (Zhang et al. (2020)). The complete conditional for v_j involves analogous expressions.

2.3. *Active learning formulation.* Assume for simplicity we have a fixed question bank with known factors $v_1, \ldots, v_k$, learned from abundant existing data. A new user i enters the survey pool. We want to select questions optimally for learning u_i.

The PMF model admits a convenient online formulation for updating our knowledge about u_i given responses from this user. Suppose, after t responses, u_i is Gaussian with mean $\mu_i^{(t)}$ and variance $[\Lambda_i^{(t)}]^{-1}$. Next the user answers question j. The posterior for u_i is

$$u_i^{(t+1)} \mid R^{(t+1)}, V, \mu_U, \Lambda_U, \alpha \sim \mathcal{N}(\mu_i^{(t+1)}, [\Lambda_i^{(t+1)}]^{-1}),$$

$$(3) \qquad \Lambda_i^{(t+1)} = \Lambda_i^{(t)} + \alpha v_j v_j^T,$$

$$(4) \qquad \mu_i^{(t+1)} = [\Lambda_i^{(t+1)}]^{-1}(\alpha R_{ij} v_j + \Lambda_i^{(t)} \mu_i^{(t)}).$$

We consider how to choose question j optimally. Inspired by approaches in active learning and item response theory, we maximize a measure of posterior information or minimize a measure of posterior variance. We focus on the trace, the sum of posterior variance along latent directions. To select the $(t+1)$th question, we solve

$$(5) \qquad \min_j \mathrm{tr}[\Lambda_i^{(t+1)}]^{-1}.$$

Our criterion is Bayes A-optimal which corresponds to minimizing the quadratic loss of a point estimate of user factors (Chaloner and Verdinelli (1995)). Our criterion is also related to minimizing predictive variance. For simplicity, let $\mu(j)$ and $\Sigma(j)$ denote the posterior mean and variance of u_i after asking question j. Suppose we draw a new question $\tilde{v}$ from the uniform distribution on the unit sphere, independently of u_i. Our prediction of the response, $u_i^T \tilde{v}$, has variance at least $r^{-1} \mathrm{tr}\, \Sigma(j)$. This is shown in Section A.2 of the Supplementary Material (Zhang et al. (2020)). Thus, A-optimality minimizes a lower bound on predictive variance along uniform latent directions.

For more intuition, we rewrite our optimization problem, letting $\lambda_\ell(\cdot)$ denote the ℓth eigenvalue of a matrix. (5) is equivalent to

$$\min_j \sum_{\ell=1}^r \lambda_\ell([\Lambda_i^{(t+1)}]^{-1}) = \min_j \sum_{\ell=1}^r [\lambda_\ell(\Lambda_i^{(t+1)})]^{-1}.$$

This variance criterion penalizes small eigenvalues of the user precision, corresponding to latent directions with least information. Information is acquired by sampling questions whose factors lie in those directions. The optimal sampling strategy chooses questions based on their informativeness and their contribution to less explored directions. For a spherical user prior, provided questions are well distributed across latent directions, the strategy prefers new questions roughly orthogonal to previous questions. See Section A.3 of the Supplementary Material (Zhang et al. (2020)) for a more formal treatment.

Algorithm 1 summarizes our active strategy. Note some limitations of this initial version. First, the algorithm is greedy, selecting only one question at a time. Optimal queries for a multistep search horizon can be found with a branch-and-bound algorithm; Garnett et al. (2012) show unbounded gains over the greedy strategy in theory but marginal gains empirically. Second, the optimal sequence of questions can be computed offline, as the objective in (5) does not depend on response values. There is one active question order across all respondents. This unrealistic property is a consequence of Gaussian modeling of responses; we will dispense with it in Section 5.

Algorithm 1: Active question selection for single user

Given: question factors $v_1, \ldots, v_k$
        prior parameters μ_U, Λ_U
        noise variance α^{-1}
        desired survey length T

Set unasked questions: $\mathcal{U} \leftarrow \{1, \ldots, k\}$
Set user prior: $\mu_0 \leftarrow \mu_U, \Lambda_0 \leftarrow \Lambda_U$
for $t \leftarrow 1$ **to** T **do**
    Choose next question, $j \leftarrow \arg\min_{\ell \in \mathcal{U}} \mathrm{tr}[(\Lambda_{t-1} + \alpha v_\ell v_\ell^T)^{-1}]$
    Collect response R_{ij}

    Update user posterior:
    $\Lambda_t \leftarrow \Lambda_{t-1} + \alpha v_j v_j^T$
    $\mu_t \leftarrow \Lambda_t^{-1}(\alpha R_{ij} v_j + \Lambda_{t-1}\mu_{t-1})$
    Mark question asked: $\mathcal{U} \leftarrow \mathcal{U} \setminus \{j\}$

---

3. Data and evaluation methods.

3.1. *Datasets*. We simulate active question selection on multiple datasets, summarized in Table 1. The Facebook survey is a survey of Facebook users, administered on the app or web interface, with a variety of questions about their experiences with the product and the company. The Facebook on-platform survey was administered in random order.

The Cooperative Congressional Election Survey (CCES) is a national Internet survey of U.S. adults that measures opinions about prevailing political issues and elected officials, before and after an election (Ansolabehere and Schaffner (2010)). Respondents are selected by matching an opt-in respondent pool to a stratified random sample from the American Community Survey. Our main results use the pre-election survey from 2016, though we also obtain results for the 2012 and 2018 surveys. We limit consideration to common content questions that ask respondents to evaluate national issues or entities on a binary or ordinal scale. For robustness checks, we expand the question set to include respondent demographics, party identification and other characteristics; we refer to this as the "full" CCES dataset. We exclude questions about actions in the past year and opinions of state or local representatives. We also exclude questions that are missing a majority of responses.

For each survey question, allowable responses are rescaled to $[-1, 1]$. Some responses will be missing, either because they were not present in the original dataset or because we dropped response values that violated the ordinal assumption. CCES has low overall missingness rates:

TABLE 1
Dataset characteristics. CCES 2016 *(full) refers to CCES* 2016 *with extra covariates*

Dataset	Number of respondents	Number of questions
Facebook on-platform survey	11,793	53
CCES 2012	54,535	29
CCES 2016	64,600	38
CCES 2016 (full)	64,600	61
CCES 2018	60,000	42
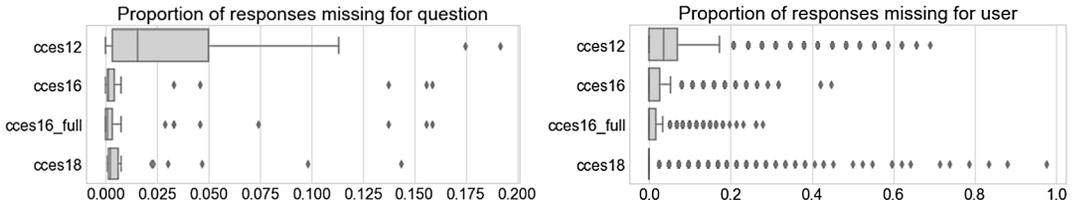
FIG. 1. *Missingness in the CCES by question and by user. There is less nonresponse in later years.*

3.7% in 2012, 1.5% in 2016 and 1.2% in 2018. The missingness distributions by question and by user are shown in Figure 1.

3.2. *Simulating the active strategy.* To simulate active question selection, we begin by randomly splitting the respondent set into a training half and a simulation half. We perform matrix factorization on the training responses to estimate question factors. We then hold out some responses in the simulation half and simulate a hypothetical survey on the remaining responses. In the hypothetical survey, we actively select the next question per respondent, reveal available responses to that question and update each user posterior. We predict held-out responses using the estimated question factors and the posterior means of user factors. We repeat this process until all questions have been asked. Since the active strategy is greedy, we can truncate the process at any point to obtain the actively chosen questions for a given survey length.

We compare the active strategy to two baselines: existing question order (available for CCES) and random question order (generated for each respondent). We focus on mean absolute error (MAE) and bias of predictions. We also compute mean squared error and the proportion of predictions with the wrong sign.

We determine the holdout set in two ways. The first method reserves a random 20% of each user's responses, effectively punching holes in the response matrix. We call this the "sparse" holdout set. It allows us to evaluate error averaged over questions and make summary comparisons of question selection strategies. The sparse holdout set has drawbacks: the artifice that the simulation procedure treats these responses as missing; and higher variability in per-question evaluation error due to using one-fifth of responses. Thus, to evaluate prediction error for individual questions, we use a second holdout method: leave-one-question-out (LOOCV) cross-validation. For each question, we simulate the survey on the response matrix with that column removed. LOOCV uses all available responses for a question to evaluate its imputation error but repeats the survey once per question. As LOOCV is more computationally demanding, we rely on the sparse holdout set to evaluate variations on the simulation procedure quickly.

3.3. *Simulation settings.* To estimate question factors from the training half, we use the SoftImpute algorithm which solves the nuclear norm regularized problem (1). Empirically, we have found that SoftImpute is efficient and produces stable estimates of V across simulations. We rely on the SoftImpute implementation in the `fancyimpute` package (Rubinsteyn and Feldman (2016)). The regularization parameter λ is selected by grid search with warm starts as recommended by Mazumder, Hastie and Tibshirani (2010), based on mean absolute error on a 20% validation set within the training half.

One variation is to estimate V by solving the Frobenius norm regularized problem (2), as it is equivalent to MAP estimation for PMF with simple priors. However, this nonconvex optimization yields highly variable question factors and orderings.

Our main results use a rank-4 matrix decomposition; SoftImpute solves (1) subject to this hard rank constraint. For any rank r, we search for λ as above. Setting $r = 4$ results in lower

prediction error than $r = 2$, while keeping the dimensionality of latent space manageable. A higher-rank decomposition ($r = 8$) does not reduce prediction error further. Greater r requires greater λ to avoid overfitting; this may overly shrink the highest-variance components.

In the active strategy, we set the prior mean μ_U and prior precision Λ_U using empirical Bayes. Specifically, we set μ_U and Λ_U^{-1} to the sample mean and covariance of the rows of UD, the implied user factors from SoftImpute. It remains to set the noise variance α^{-1}. By Popoviciu's inequality and the prior rescaling of responses to $[-1, 1]$, we know $\alpha^{-1} \leq 1$. Our main results use the upper bound ($\alpha^{-1} = 1$), though we tried smaller values[1].

We also allowed the active strategy to minimize measures of posterior variance other than the trace, like the determinant and maximum eigenvalue. These optimal design criteria result in similar overall predictive performance and active orderings.

Replication code is available in the Supplementary Material (Zhang et al. (2020)). Our main results follow. Results for alternate simulation settings appear in Appendix E of the Supplementary Material.

4. Results.

4.1. *The most informative issue questions.* The active strategy ranks questions by precision gained in the latent representation of a user. For more intuition, see Appendix B which visualizes users and questions in two-dimensional latent space. Below we showcase the questions from the 2016 CCES that the active strategy deems most informative. The Supplementary Material includes results for the 2018 CCES (Appendix F), the 2012 CCES (Appendix G) and the Facebook survey (Appendix H).

Active item selection produces a stable question order (Figure 2). The foremost question is whether to repeal the Affordable Care Act (ACA). Questions about immigration, abortion and environmental policies are prioritized: a 10-question active survey includes multiple questions on each topic.

We check the robustness of the active ordering to our question inclusion criteria for the CCES. We progressively add questions about respondent political affiliation, demographics, education and other characteristics. Questions with categorical responses, like race, are converted into indicators. Note this one-hot encoding artificially creates a separate survey question per response value; multinomial logit modeling would be more appropriate in practice. Appendix C contains active orderings with these additional questions.

The active ordering with augmented questions does not diverge much from the active ordering in Figure 2. Though questions about gender, party identification, parenthood and home ownership slot into the first 20 positions, questions about the environment, abortion and the ACA remain prominent. Gender and Obama approval displace immigration questions from the top 10. The active strategy postpones questions about race and education, possibly because these one-hot-encoded covariates are not well captured by a low-rank matrix decomposition.

4.2. *The latent space of issues.* We interpret the latent concepts for which the active strategy gathers information. Figure D.5 displays question factors as loadings on these latent concepts. The first direction broadly indicates partisanship: Democratic and Republican policies tend to have loadings with opposite sign. It makes sense that this principal component contains most prior variance. Partisanship is highly correlated with opinions about the environment, abortion and immigration, which explains their prominence in the active ordering.

---

[1]Preliminary probabilistic modeling of the 2016 and 2018 CCES with a rank-4 decomposition suggests that α^{-1} is likely between 0.7 and 0.75.
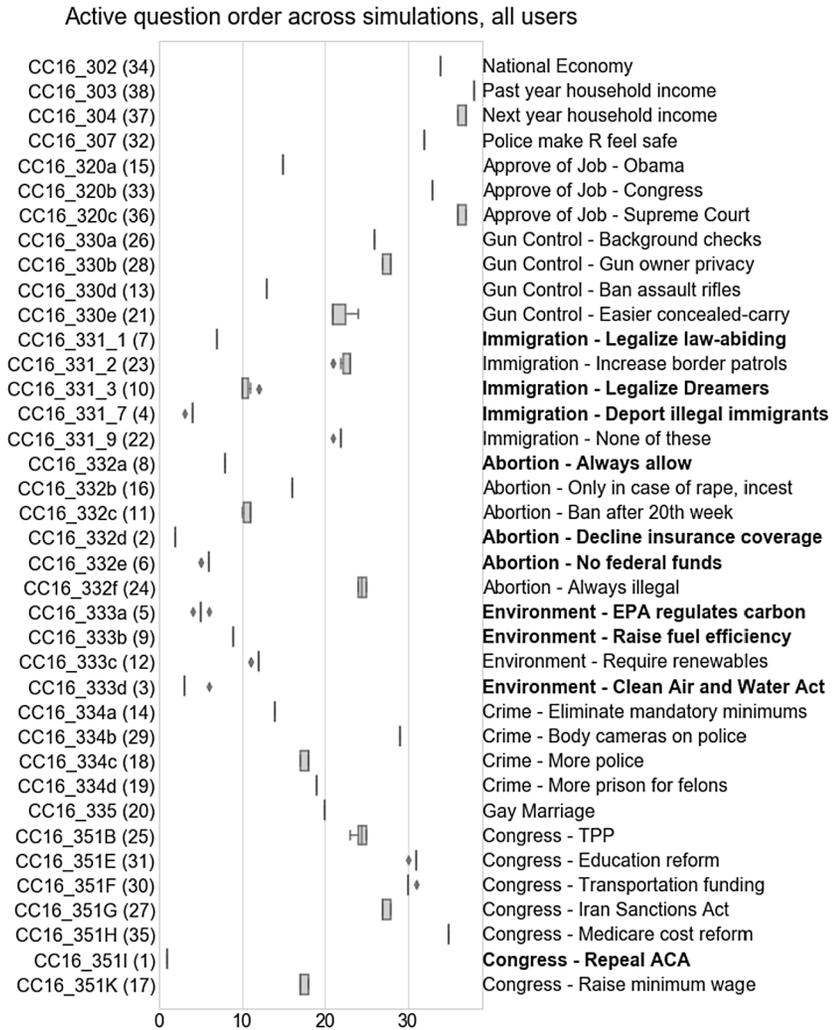
FIG. 2. *Active ordering for* 2016 *CCES using a rank-4 matrix decomposition with no questions held out. The box plot shows the rank of each question across* 10 *training/simulation splits. Position in the active ordering, according to median rank, is in parentheses. The top* 10 *questions appear in bold. Across simulations, the active survey leads with a question about Obamacare repeal and then cycles through questions about abortion, the environment and immigration.*

The second component seems to represent level of bipartisan support. For instance, increasing prison sentences for repeat felons, requiring police to wear body cameras and enacting background checks for all gun purchases are broadly popular policies supported by 84%, 87% and 90% of respondents, respectively. These load highly in the negative direction. Opposite these is a "none-of-the-above" question about immigration policies, which only 5% of respondents supported. Similar analyses for 2018 (Figure F.8) and 2012 (Figure G.8) also find that the two highest-variance latent directions capture partisanship and bipartisanship.

In 2016, the third principal component correlates support for greener environmental policies, support for abortion restrictions and opposition to same-sex marriage. This suggests a group of socially conservative respondents who are concerned about the environment. The fourth component correlates support for greener environmental policies and opposition to abortion restrictions with support for tougher crime and immigration policies. The active strategy doubles down on environmental, abortion and immigration questions in order to as-
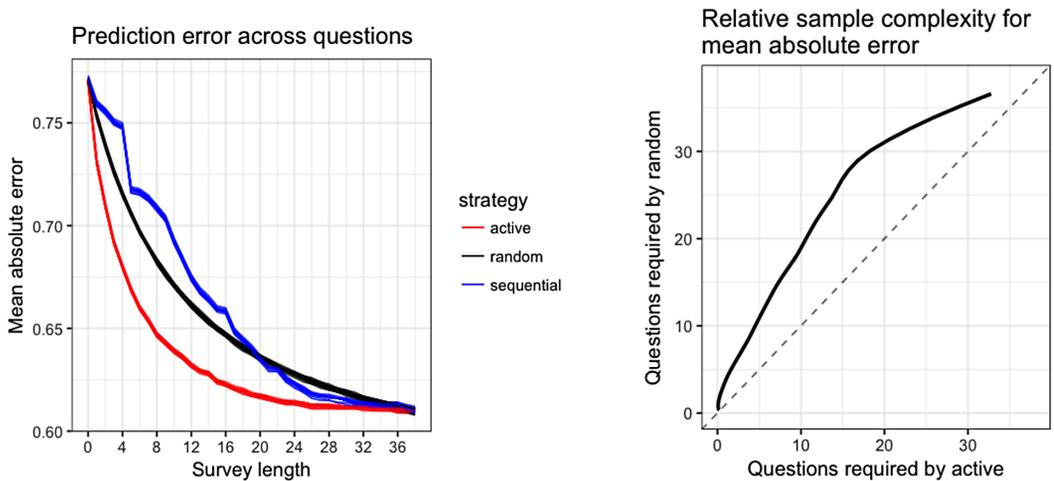
(a) Prediction error on the sparse holdout set, measured across 10 iterations of simulating the 2016 CCES. We show MAE averaged over all questions.

(b) Sample complexity of active ordering relative to random ordering for the 2016 CCES.

FIG. 3. *Summary measures of imputation error for simulated surveys using each question selection strategy. The training/simulation split is fixed; the source of randomness across iterations is the sparse holdout set. The solid line in panel* (b) *plots the number of questions required by each strategy to attain the same level of error in panel* (a). *We fit, for strategy* $s \in \{\text{random}, \text{active}\}$, *a loess smoother* $f_s(\epsilon)$ *to predict the number of questions required for error level* ϵ. *We plot* $(f_{\text{active}}(\epsilon), f_{\text{random}}(\epsilon))$ *for the range of* ϵ *attained by both strategies. For instance, the imputation error after five actively chosen questions is comparable to that after* 10 *randomly chosen questions. Since the curve lies above the dashed* $45°$ *line, active sampling outperforms random.*

certain membership in these groups. Going beyond two latent dimensions helps to identify parts of the electorate that do not behave according to conventional partisan wisdom.

4.3. *Efficiency gains under active selection.* We now examine the imputation ability of matrix factorization with responses revealed by active selection. On the sparse holdout set, predictions with actively chosen questions outperform predictions with randomly or sequentially chosen questions (Figure 3(a)). The active strategy attains lower imputation error, averaged over questions, for simulated surveys of short or medium length. Only when two-thirds of the questions have been asked do the strategies converge in overall performance; this error level is the minimum achievable by low-rank matrix factorization on this dataset. This result is consistent across error measures (Figure D.1) and optimal design criteria (Figure E.3).

Another measure of efficiency gain is sample complexity—the number of responses required by each strategy to reach a given error level. The active strategy almost always requires fewer responses (Figure 3(b)). Suppose we ask 20 random questions of each respondent; the active strategy reaches the same imputation quality with nearly half as many questions.

The active strategy minimizes the optimal design criterion, as Figure D.4 verifies. It is advisable to balance exploration and exploitation when optimizing under uncertainty. We introduce exploration with a simple ϵ-greedy modification to the active strategy: choose a random question with probability $\epsilon = 0.05$; otherwise choose the A-optimal question. ϵ-greedy question selection does not outperform active question selection in terms of prediction error (Figure D.6). Since our simulations treat question factors as fixed, exploration cannot reduce their estimation error. Henceforth, we focus on the active ($\epsilon = 0$) strategy, with the caveat that some form of exploration is preferable when question factors contain nontrivial uncertainty.
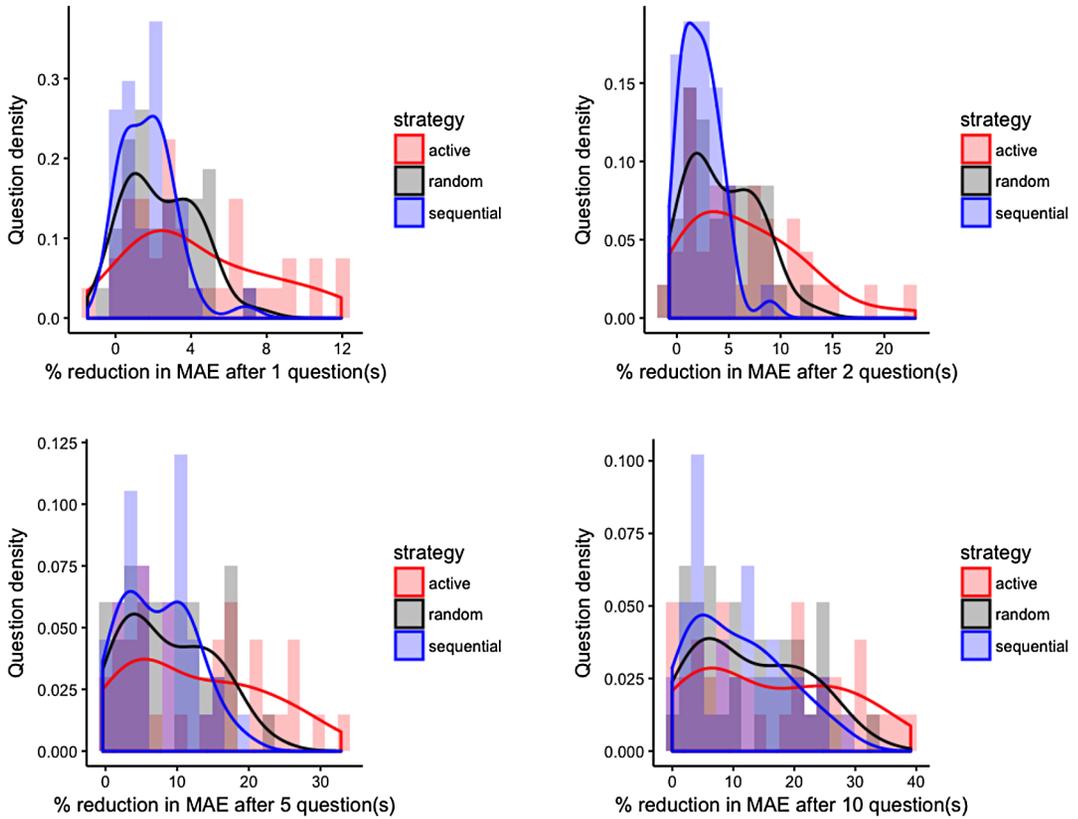
FIG. 4. *Reduction in imputation error from presurvey levels for the* 2016 *CCES. We show percent reduction in mean absolute error as a distribution over questions, smoothed by kernel density estimation. Error per question is evaluated by leave-one-out cross-validation. In short surveys a number of questions experience the greatest error reduction under the active strategy.*

4.4. *Uneven error reduction across questions.* Evaluation on the sparse holdout set masks considerable heterogeneity in which questions benefit from active selection. Figure 4 plots the reduction in LOOCV error per question for various survey lengths. More questions experience at least a 20% reduction in imputation error after five actively chosen questions, compared to five randomly or sequentially chosen questions. The advantage of the active strategy is apparent after a single question. As survey length increases, all strategies achieve greater error reduction, and this advantage narrows.

The error reduction of a single actively chosen question is more notable for the 2018 CCES: at least 12% on half of that year's question set (Figure F.3). Political preferences, at least those captured by the CCES, have become more predictable if we know what to ask.

Figure 5 locates the questions for which the imputation abilities of active and random selection diverge. Active selection delivers greater error reduction for questions concerning abortion, immigration and the environment—the very topics front-loaded by the active strategy. Active selection also benefits predictions of Obama approval and support for gun restrictions. The predictive advantage of active learning comes not from covering these topics exhaustively but from sampling informative items within correlated sets. Meanwhile, for questions about crime and the economy the predictive advantage of active learning is minimal.

4.5. *Where imputation fails.* The bounds in Figure 5 indicate what extent of error reduction is possible from knowing no responses ("presurvey") to knowing all available responses
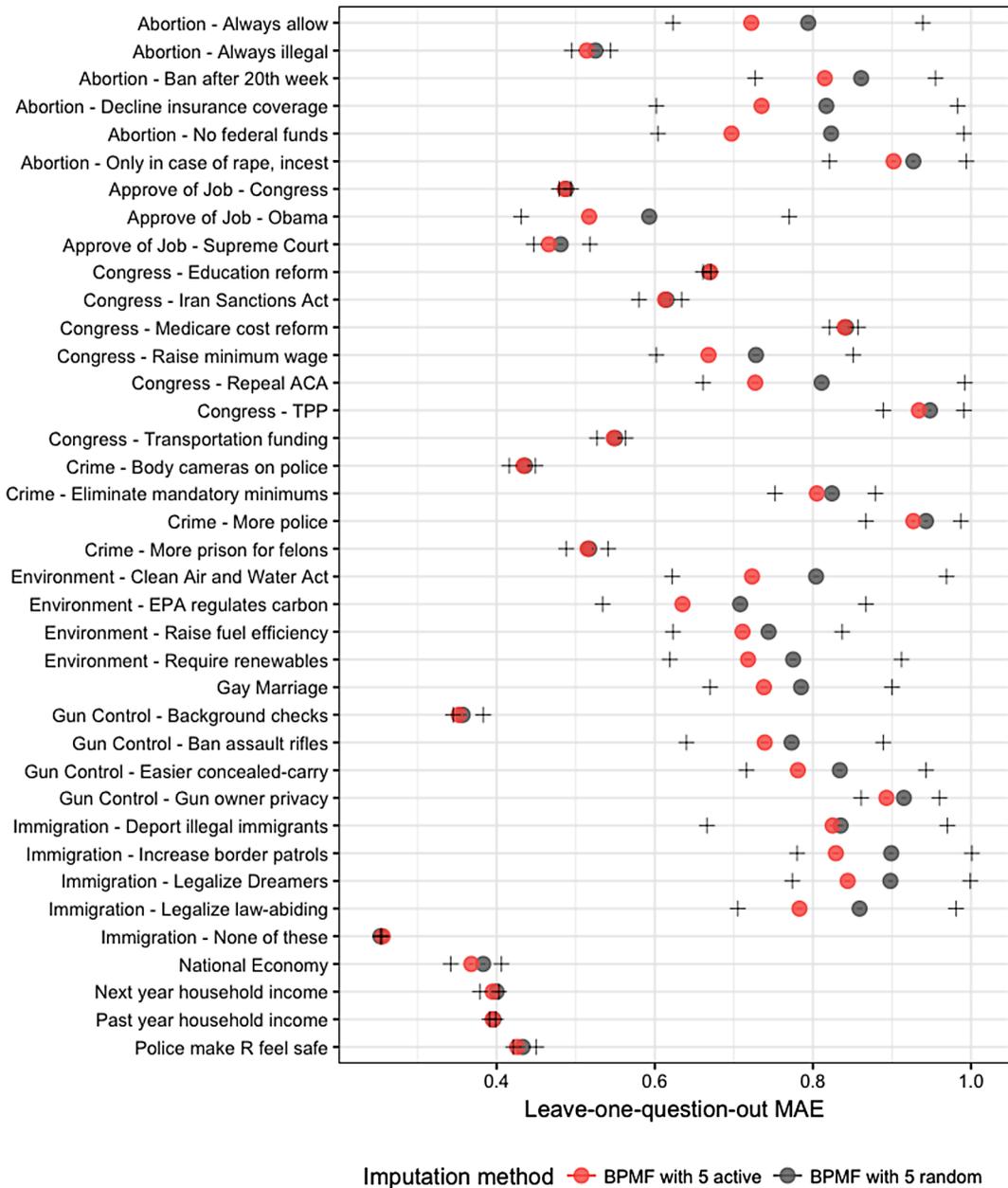
FIG. 5. *Mean absolute prediction error per question after a short survey using active or random question se-lection. We simulate the survey with one question held out, reveal responses to five actively or randomly chosen questions and predict responses to the held-out question. We repeat this for each question to produce leave-one-out cross-validation error. 2σ confidence intervals are not visible due to the number of respondents in the simulation half. The gray "+" bounds represent prediction error under the presurvey condition of no knowledge (right) and the oracle condition of knowing all other responses (left). On questions where presurvey and oracle error are well separated, active selection usually gives better imputations. On questions where little error reduction is possible, active selection does not help.*

("oracle"). The oracle bound quantifies the irreducible error of imputing each question with low-rank matrix factorization. Some questions are inherently harder to impute; their oracle MAE is close to 1. Other questions have low oracle error and low presurvey error. In both cases, whether question selection is active or random makes little difference. For intuition, on

a binary question with possible responses $\{-1, 1\}$, MAE of 1 is achievable by (i) randomly guessing -1 or 1 with equal probability or (ii) always predicting 0.

One component of irreducible error is bias. The presurvey and oracle bias per question are shown in Figure D.3. There is little difference between the two; bias is mostly determined once question factors have been estimated. Low-rank matrix factorization produces small bias relative to MAE: less than 0.1 in either direction for all questions, and less than 0.05 for all but two questions. There is no clear relationship between questions with high irreducible error and those with relatively high bias.

The relative irreducible errors are affected by our decision to scale all responses to $[-1, 1]$. Questions with evenly distributed binary responses will tend to have higher irreducible error than those with lopsided binary responses or ordinal responses. The alternative of zero-centered, unit-variance scaling would complicate interpretation, as the response values would depend on the response distribution. Our choice of scaling balances competing goals of interpretability and standardization.

With this caveat in mind, it may be ineffective to impute questions with high irreducible error. The active strategy has little room to help. The survey researcher is advised to include such questions in the eventual survey, regardless of their position in the active question order, if accurate measurement of these constructs is a priority.

5. Modeling ordinal responses. In this section we adjust the model to better capture binary and ordinal response values. We replace the Gaussian likelihood for responses with the ordered logit likelihood. The ordered logit model, also known as the proportional odds model, is prevalent in social science research (Fullerton and Xu (2012)). As we will see, this change breaks the determinism of the active order; question selection will now depend on the respondent's previous answers.

Quantized outputs in the response matrix have been modeled with a variety of link functions, including logistic, probit and multinomial logit (Cao and Xie (2015), Davenport et al. (2014), Klopp et al. (2015)). These works formulate matrix completion as maximum likelihood with nuclear norm regularization. We continue with a probabilistic matrix factorization approach. The ordered logit model is nonconjugate; so, for posterior inference we employ variational Bayes, also used for matrix completion in Lim and Teh (2007) and Seeger and Bouchard (2012). We forfeit the closed-form posterior update exploited by the active strategy for PMF, but the Laplace approximation offers a way forward.

5.1. *Probabilistic matrix factorization model.* Our ordered logit response model retains the normal priors for user and question factors. Responses are integer-valued starting at 1. We allow heterogeneity across questions: the number of response values can differ across questions, as can response frequencies. The model becomes

$$u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_U, \Lambda_U^{-1}),$$

$$v_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_V, \Lambda_V^{-1}),$$

$$R_{ij} \mid U, V \overset{\text{ind.}}{\sim} \text{OrderedLogit}(u_i^T v_j, \beta_j).$$

R_{ij} takes values in $\{1, 2, \ldots, M_j\}$, where M_j is the question-specific maximum response value. Let π_{ijm} denote the probability that $R_{ij} = m$. The probabilities $\{\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijM_j}\}$ are defined by the logistic link and a series of question-specific cutpoints $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,M_j-1})$. For simplicity of presentation, we drop the indexing for user i and question j.

Thus, R_{ij} takes values in $\{1, 2, \ldots, M\}$ with probabilities $\{\pi_1, \pi_2, \ldots, \pi_M\}$, parameterized by cutpoints $\beta = (\beta_1, \ldots, \beta_{M-1})$ as follows:

$$\text{logit}\left(\sum_{k=1}^{m} \pi_k\right) = u_i^T v_j + \beta_m \quad (m = 1, \ldots, M - 1),$$

$$\pi_M = 1 - \sum_{k=1}^{M-1} \pi_k.$$

5.2. *Inference.* We perform posterior inference on U and V, once when estimating user and question factors from the training half and again every time we expand the survey by one question for the simulation half. Using the updated posteriors, we compute prediction error on held-out survey responses at each iteration. The steps are detailed in Algorithm 2.

We obtain approximate posteriors for U and V given R in the above model using mean-field variational inference. We implement this in `edward` (Tran et al. (2016)). Our variational distributions are fully factorized Gaussian,

$$q(U) = \prod_{i=1}^{n} q(u_i) = \prod_{i=1}^{n} \prod_{j=1}^{r} \mathcal{N}(\mu_{ij}, \sigma_{ij}),$$

$$q(V) = \prod_{j=1}^{k} q(v_j) = \prod_{j=1}^{k} \prod_{i=1}^{r} \mathcal{N}(\nu_{ji}, \tau_{ji}).$$

Variational inference finds parameters $\{\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}\}_{i,j}$ that maximize the evidence lower bound or, equivalently, minimize the KL divergence between the variational distribution and the true posterior. We employ priors $\mu_U = \mu_V \equiv 0$ and $\Lambda_V = \Lambda_V \equiv \mathbf{I}_r$. To obtain

---

Algorithm 2: Active strategy simulation for ordered logit model

---

Given: training responses R^{train}
cutpoints β
desired survey length T

$\{\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}\}_{i,j} \leftarrow$ `approximate_U_V_posterior`$(R^{\text{train}}, \beta)$
Set $(v_1, \ldots, v_k)$ to variational means $(\nu_1, \ldots, \nu_k)$
Set Λ_U^{-1} to empirical covariance of $\{\mu_1, \ldots, \mu_n\}$
Initialize record of questions per user: $\mathcal{O}_i \leftarrow \varnothing, \mathcal{U}_i \leftarrow \{1, \ldots, k\} \, \forall i$
Initialize revealed responses R^{asked} to empty $n \times k$ matrix
for $t \leftarrow 1$ **to** T **do**
    Ask one question of all users:
    for $i \leftarrow 1$ **to** n **do**
        Choose next question, $j \leftarrow \arg\min_{j \in \mathcal{U}} \text{tr}[\Lambda_U + \sum_{\ell \in \mathcal{O}} \mathcal{J}^\ell(u_i; R_{i\ell}) + \mathcal{I}^j(u_i)]^{-1}$
        Collect response: $R_{ij}^{\text{asked}} \leftarrow R_{ij}$
        Mark question asked: $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup \{j\}, \mathcal{U}_i \leftarrow \mathcal{U}_i \setminus \{j\}$

    $\{\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}\}_{i,j} \leftarrow$ `approximate_U_V_posterior`$(R^{\text{asked}} \cup R^{\text{train}}, \beta)$
    Set $(u_1, \ldots, u_n)$ to variational means $(\mu_1, \ldots, \mu_n)$
    Set R_{ij}^{pred} to mean of OrderedLogit$(\mu_i^T v_j, \beta)$ random variable $\forall i, j$
    Compute prediction error of R^{pred} for R

---

cutpoints β, we follow the inverse approach in the `rstanarm` package (Gabry and Goodrich (2016)). For each question we draw from the simplex probabilities $\pi = (\pi_1, \ldots, \pi_M)$ corresponding to the ordinal response values. Specifically, we draw $\pi \sim \text{Dirichlet}(c_1, \ldots, c_M)$, where the concentration parameters are prior counts of the response values. That is, we set c_m equal to the number of times a respondent answers m to this question in the training half. We then apply the logit transform to all but the last entry of $\text{cumsum}(\pi)$, obtaining

$$\beta_m = \text{logit}\left(\sum_{k=1}^{m} \pi_k\right) \quad (m = 1, \ldots, M-1).$$

To predict R_{ij}, we set u_i and v_j equal to their variational means μ_i and v_j and compute the mean of the resulting ordered logit random variable.

5.3. *Active learning formulation.* The ordered logit response model also entails changes to our item selection strategy. Again, we treat V as fixed; we approximate V with the variational means $\{v_j\}_{j=1}^{k}$. For now, we restrict attention to a single user in the model,

$$u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_U, \Lambda_U^{-1}),$$

$$R_{ij} \mid U, V \overset{\text{ind.}}{\sim} \text{OrderedLogit}(u_i^T v_j, \beta).$$

We seek the question j that maximizes information about u_i. Since we lack a closed-form update for the posterior of u_i, we do not minimize an exact measure of posterior variance. Instead, we work with the variance of the Laplace approximation, or the portion of this variance that we can control through item selection—the Fisher information. This approach follows the optimal design literature, notably Segall (2009), who applies it to the logistic likelihood for binary responses. Our approach is an ordered logit generalization of Segall (2009).

Fisher information is computed around a value of u_i. We estimate u_i with the mean of the user variational distribution, μ_i, from the latest round of probabilistic matrix factorization. Repurposing this provisional estimate of u_i is more computationally efficient than the alternative of computing a MAP estimate of u_i in the single-user model.

We denote the Fisher information gained from a response to question j as $\mathcal{I}^j(u_i)$ and the *observed* Fisher information from observing response m to question j as $\mathcal{J}^j(u_i; m)$. Then,

$$\mathcal{J}^j(u_i; m) = -\frac{\partial^2}{\partial u_i \, \partial u_i^T} \log \text{Pr}(R_{ij} = m \mid u_i, v_j, \beta)$$

and, again letting π_{ijm} denote $\text{Pr}(R_{ij} = m \mid u_i, v_j, \beta)$,

$$\mathcal{I}^j(u_i) = E[\mathcal{J}^j(u_i; R_{ij})] = \sum_{m=1}^{M} \pi_{ijm} \mathcal{J}^j(u_i; m).$$

In the ordered logit model, $\mathcal{J}^j(u_i; m)$ and $\mathcal{I}^j(u_i)$ have closed-form expressions. The Hessians are computed with autodifferentiation in `edward`.

Let $\mathcal{O}$ contain the indices of past questions and $\mathcal{U}$ the indices of unasked questions. We compute the sum of observed information over $\mathcal{O}$ and consider adding a Fisher information term for question $j \in \mathcal{U}$. We determine which question would contribute the most information to u_i in expectation. More formally, we find the question that minimizes the variance of the Laplace approximation A-optimally,

$$(6) \qquad \min_{j \in \mathcal{U}} \text{tr}\left[\Lambda_U + \sum_{\ell \in \mathcal{O}} \mathcal{J}^\ell(u_i; R_{i\ell}) + \mathcal{I}^j(u_i)\right]^{-1}.$$

When expanding the survey by one question, the per-user item selection problems can be solved in parallel. This subprocedure is placed in context in Algorithm 2.

5.4. *Results.* We make two modifications to our evaluation procedure. First, predictions by the ordered logit model are on the same scale as the original ordinal responses; questions with more allowable responses will tend to have higher error. Thus, we rescale prediction error per question to be comparable to that of PMF, which predicts responses rescaled to $[-1, 1]$. Second, inference and item selection are more computationally intensive with the ordered logit model. To avoid the additional computational burden of leave-one-question-out cross-validation, we resort to five-fold cross-validation on questions. Results appear in Appendix J of the Supplementary Material (Zhang et al. (2020)).

We find that imputation error under active question selection is reduced faster with ordered logit response modeling than with PMF. This is especially apparent after two questions. The difference after one question comes from the matrix factorization step rather than item selection, since the first actively chosen question is the same under both models. We have modeled responses more appropriately as ordinal and introduced additional parameters in the form of question-specific cutpoints. Item selection with the ordered logit likelihood may contribute to imputation gains starting with the second question; this warrants further investigation.

With five actively selected questions, the ordered logit predictions are nearly at oracle level. Not much room for improvement remains; it may be worth terminating the survey here. The presurvey and oracle error bounds are close to those of PMF. Note that oracle error may exceed mid-survey error in some cases, due to the possibility of increased bias with more responses. Indeed, greater bias is a shortcoming of the ordered logit procedure: unlike with PMF, the oracle bias can be considerably higher than the presurvey bias. This is unsurprising given our reliance on approximate inference.

With the ordered logit model there is more variability in the active question order within one simulation than across PMF simulations. This can be seen by plotting individual paths through survey questions (Figure J.7). The variability in question rank arises not from a few common question orderings but, rather, from diverse personalized paths dependent on responses to previous questions.

Across these individual paths, immigration, abortion and environmental policies are prioritized, and ACA repeal remains the top question across sampled users. These similarities to the PMF active ordering reinforce our earlier findings on question importance. Some questions that appear late in the PMF ordering have highly variable position in the adaptive ordering. Examples include perception of the U.S. economy over the past year and whether abortion should always be illegal. Approval of Obama is asked anywhere from second to 20th. This was the least stable question in our PMF robustness checks—it leaps into second place with the addition of covariate questions—so its variable position under the ordered logit model is natural.

6. Experimental comparison. Thus far, the imputation gains of active question selection have only been shown in simulations that reorder existing survey responses. To verify these gains experimentally, we conducted a second version of the Facebook survey using the deterministic active ordering from rank-4 PMF. Active order was compared to random order and an order generated by a survey methodologist, henceforth called expert order. The expert order was deterministic aside from certain random-order subsets of questions. To reduce burden, we trimmed the survey from 53 to 33 questions by removing some question groupings. Respondents were recruited from four countries and randomly assigned to conditions. The active, random and expert conditions had 4224, 4211 and 4177 respondents, respectively.

Our primary concern is the imputation performance of short surveys designed by each strategy. As before, our evaluation truncates the survey to a given length and computes the error of PMF predictions on remaining responses under each strategy. In addition, we reuse the notion of presurvey imputation error, the error of predictions using the user prior. Presurvey error on a given question can be slightly different across conditions (Figure I.2), due
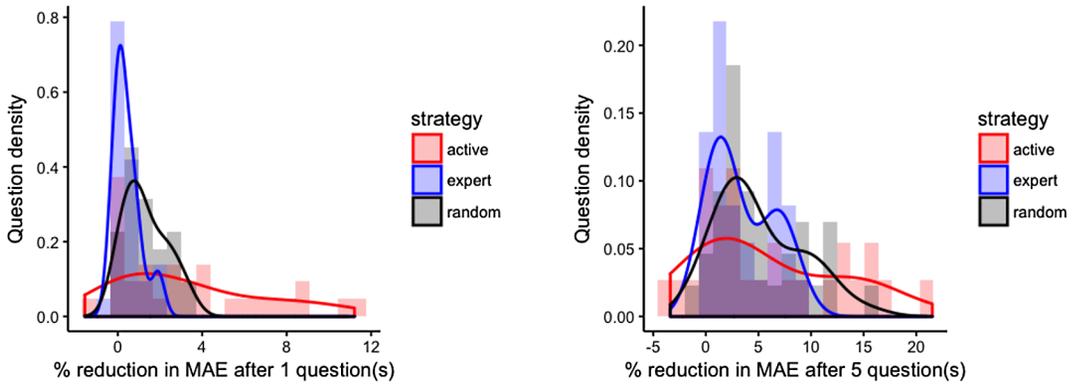
FIG. 6. *Reduction in imputation error from presurvey levels for the Facebook survey experiment. We show percent reduction in mean absolute error as a distribution over questions, smoothed by kernel density estimation. We omit questions with undefined imputation error after the truncated survey, namely, the first question in the active order (left) and the first five questions in the active order (right). The empirical distribution of error reduction under the active strategy contains more right-tail mass despite omitting the first five actively selected questions, two of which had high error reduction after a one-question survey.*

to biases discussed in Section 7. We control for these small biases by computing, for each condition, the reduction from its own pre-survey error.

The following analysis differs from the simulation-based comparisons in two ways. First, while the simulations held out each question in turn to compute LOOCV error, here we respect the order in which experimental responses are gathered. Suppose we are interested in imputation error after a five-question survey. For any question and condition, the computation of imputation error is restricted to respondents who answered this question after the first five. Hence, imputation error after a five-question active survey can only be computed for questions after the fifth in the (deterministic) active ordering. Second, our response predictions use question factors estimated from the first Facebook survey. As the surveys occurred a year apart, the latent structure may have changed. Imputation for the survey experiment may improve with updated question factors, and the active order may no longer be optimal.

Still, it is apparent after one question that the active order produces greater error reduction than random and expert order on a subset of questions (Figure 6). Error reduction grows as more questions are asked; at five questions the random and expert order begin to catch up. For all strategies the distribution of error reduction appears bimodal, suggesting that one set of questions is easier to impute than the rest. Error reduction in the experimental data is consistent with reduction in LOOCV error on simulated data (Figure I.3). The questions we expect to benefit, actually benefit, and they benefit more from active order in short surveys.

7. Limitations of evaluation.

7.1. *Order effects.* Simulation-based estimates of error reduction may not generalize well to practice due to changes in respondent behavior under new survey designs. For instance, our simulations assume that responses do not depend on the order in which questions appear. In actuality, observed CCES responses are specific to the historical CCES question order; an individual's potential response to a question may be different under algorithmic question order. Our evaluation procedure assumes that the active-order and random-order interventions have zero effect on responses. In other words, there are no order effects.

In this section we estimate the magnitude of order effects in the initial randomly-ordered Facebook survey to gauge how much bias the active strategy could introduce. Position effects capture the effect of a question appearing at a given position in the instrument, such as early
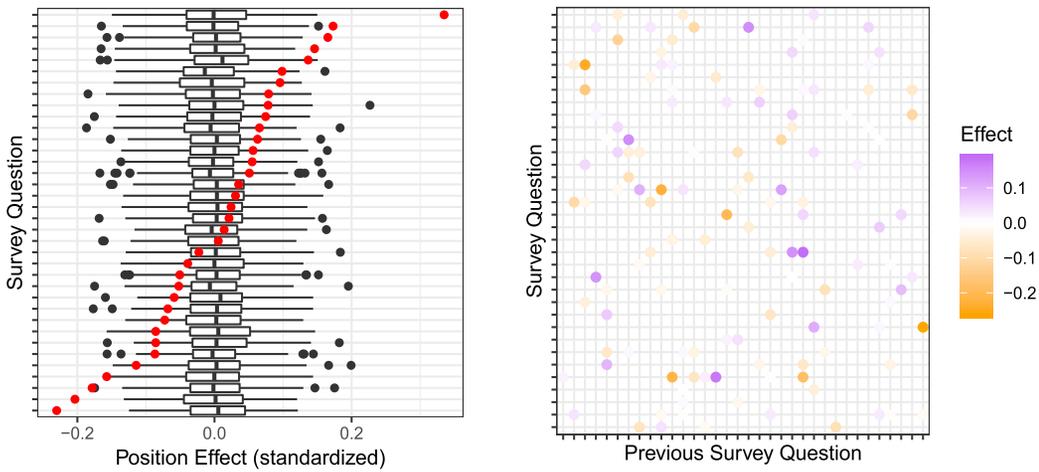
(a) Empirical (red) vs. null distribution (box-plots) of question position effects, obtained by permuting question order and fitting linear regressions of standardized survey responses.

(b) Effect of previous question on next question response for the Facebook survey. To detect large order effects, we estimate these via L1-penalized regression on all pairwise question interactions.

FIG. 7. *Estimated order effects in the Facebook survey. While we do detect position and interaction effects, these are not large.*

or late. Interaction effects capture whether the previous survey question that a user answered affects the response to the following question.

To estimate position effects, we fit one linear regression per survey question with the relative position of the question in the order as a predictor. In this model we use data from only completed surveys (about 30% of the surveys) in order to preclude attrition bias. We use this model to estimate the difference in the standardized response for each question appearing at the end of the survey compared to the beginning. As a null distribution, we randomly re-order the survey and fit the same model 200 times. The results are presented in Figure 7(a). A number of survey questions exhibit position effects, in which responses vary significantly depending on whether they were asked early or late in the survey. The worst-case bias appears to be about 0.3 standard deviations on the response scale.

We also estimate first-order interaction effects, making the Markov assumption that conditioning on the identity of the previous question is sufficient. We fit an L1-penalized regression with a parameter for all pairs of survey questions and previous questions, using 10-fold cross-validation to select the optimal penalty parameter. The results of this model are illustrated in Figure 7(b). About 10% of the possible question pairs exhibit a nonzero interaction effect. Some questions tend to be influential on the following question (columns with multiple points), while others tend to be affected by the prior question (rows with multiple points). Similarly to position effects, the interaction effects we observe are usually less than 0.2 standard deviations on the response scale. In regards to the choice of whether to include the pairs beginning at odd- or even-numbered question positions, these effects are not robust. In sum, we have not detected large, persistent interaction effects.

The subsequent experiment confirms that the order effects for the Facebook survey are small. We find that both active and expert orderings induce order effects relative to random order (Figure I.1). Where it exists, the bias in mean response is usually less than 0.2 standard deviations on the response scale, consistent with our estimates from the first Facebook survey.

The evaluation of the Facebook survey experiment has the advantage of using real, not potential, responses under each assigned ordering condition; we do not need to assume the absence of order effects. However, if order effects exist, our evaluation is agnostic to which

ordering is "correct," that is, to the context in which the responses should be interpreted. An alternative approach is to designate a standard ordering (e.g., expert) and attempt to adjust observed responses under the other orderings to potential responses under the standard ordering. Error reduction would be interpretable with respect to a single ordering but with inflated uncertainty.

7.2. *Attrition.* Although the experiment alleviates concerns over introducing bias by way of an altered survey design, it provokes a different issue, the potential for higher nonresponse under the active ordering. Compared to random and expert, the active order increases attrition on the initial question and maintains a lower proportion of respondents over the course of the survey (Figure I.4). A possible explanation is that the active order places more controversial questions upfront. One proxy for controversiality is how often a question is skipped; of the first four questions in the active order, three are frequently skipped by respondents in both active and expert conditions. The expert order experiences the least attrition over the first half of the survey, compensating in part for its slower reduction of imputation error.

If dropoff is indeed greater for the active ordering, our simulations may overstate the benefit of active learning. Future work could account for attrition by giving the active strategy a smaller question budget than other strategies. Otherwise, simulations could model probabilistic dropoff after each question with probabilities estimated from a pilot sample. A more pragmatic active learning objective would represent the tradeoff between information gain and dropoff. In addition, a wide pilot survey could be undertaken to discover informative questions with low rates of nonresponse.

7.3. *Measurement error.* On the other hand, the reduced burden of actively designed surveys may result in lower measurement error. Our simulations do not account for the possibility that a shorter survey may reduce satisficing and improve response quality. Roberts et al. (2019) review studies on satisficing across survey modes and domains, finding mixed evidence of a relationship between predictors of satisficing (task difficulty, respondent ability, respondent motivation) and degraded response quality. Even if satisficing is not a problem for the CCES, the world of surveys is broad, and respondent motivation could vary widely.

When satisficing occurs, it is not inconceivable that an imputed response represents the respondent's view more accurately than the collected response. In this study, evaluations compute error treating observed responses as ground truth. But observed responses are subject to measurement error. If we regard observed responses as noisy realizations of true opinion, as PMF does, we can attribute some of the irreducible error of imputation to measurement error. Our evaluation procedure may thus understate the potential of our method for quality imputations as well as quality responses.

8. Discussion and future work. If we commit to shortening a survey, we will impute uncollected responses with error. We identify questions that diminish imputation error via active learning that maximizes user precision in the latent space of concepts. As we explore the CCES questions preferred by the active strategy, we develop insights about the most informative set of questions for predicting political opinion. The active ordering offers a new notion of feature importance for domains with wide item sets and low-dimensional latent structure.

We have presented two variants of the active strategy. The first produces a deterministic question order which follows from Gaussian response modeling in PMF. This is convenient for survey applications that require a predictable design upfront. It is also computationally simple. However, the property that future questions do not depend on past responses is unrealistic, as evidenced by the improved imputation ability of the ordered logit response model.

With the ordered logit likelihood, the active ordering adapts to collected responses at the price of increased computation.

Incorporating side information about respondents should lead to further efficiency gains. Theoretical results have shown that sufficiently informative side information improves the sample complexity of matrix completion (Chiang, Hsieh and Dhillon (2015), Xu, Jin and Zhou (2013)). In Appendix K we explore two simple ways to leverage respondent covariates: by computing subgroup-specific user priors, and by treating covariates as always-revealed responses in matrix factorization. The latter method reduces imputation error early in the survey at the cost of introducing bias. Many avenues exist for more sophisticated modeling of side information. One could map covariates to a personalized user prior with multilevel regression or a Gaussian process (Adams, Dahl and Murray (2010), Zhou et al. (2012)). Another approach is to add terms for user and question covariates to the response specification (Athey et al. (2018), Porteous, Asuncion and Welling (2010)). Our simulations with free covariates imply an exchange rate between side information and survey responses. Trading off the information gain and acquisition cost of both in a user-specific way is a design opportunity.

Our methods thus far have assumed away uncertainty in estimated question factors. Future work should incorporate this uncertainty into the active learning objective, for instance with bandit algorithms that use an upper confidence bound for the optimality criterion or sample from its distribution. The active strategy could also account for temporal uncertainty if one believes that latent concepts are time-varying; explicitly modeling these dynamics could produce an active ordering that evolves automatically.

Matrix completion as an imputation procedure can suffer from both nonresponse and model misspecification bias. The standard matrix completion loss assumes that entries are missing completely at random. The tenuous plausibility of this assumption is exacerbated by the active strategy, which tailors questions to a user's inferred latent position. In an effort to de-bias the loss, Srebro and Salakhutdinov (2010) regularize a weighted nuclear norm, while others apply inverse propensity weights to reconstruction terms (Athey et al. (2018), Schnabel et al. (2016)). Alternatively, one could model the missing-data mechanism explicitly (Marlin and Zemel (2009)). Even with these corrections, a low-rank linear decomposition cannot capture all of the response variance; the remainder is reflected in oracle error. It could be worth exploring nonlinear matrix factorization in the form of Gaussian process latent variable models (Lawrence and Urtasun (2009)).

Aside from model improvements, future work should devote special attention to the logistics of survey administration. The difficulty of implementing an active ordering depends on the degree of adaptivity. The deterministic order from PMF is relatively straightforward to apply across survey modes; the adaptive order powered by the ordered logit model would require more infrastructure. Web and computer-assisted telephone surveys could lean on back-end software to suggest the next question. In-person field surveys would need a mechanism for inputting responses and quickly receiving the next question, such as a mobile app that calls a low-latency API for the active ordering. It would be productive to integrate with existing survey platforms to make the active strategy available. Additionally, suggesting questions in batch may be more practical than sequentially; this calls for optimal design with a multistep horizon. It is also an opportunity to design logical question groupings informed by domain knowledge of practitioners.

Ultimately, the degree to which active question selection influences survey design rests with the survey researcher. Our method can guide the researcher in creating shorter instruments by suggesting informative questions in a principled manner. At the other end of the spectrum, it can automate adaptive survey design.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/20-AOAS1322SUPPA; .pdf). We provide proofs, additional simulation and experimental results, robustness checks for the active ordering, a two-dimensional illustration of the active strategy, and exploratory work incorporating side information.

Replication code (DOI: 10.1214/20-AOAS1322SUPPB; .zip). This archive contains the CCES datasets, simulation code and analysis notebooks. It includes the implementation of the ordered logit model.

REFERENCES

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929 https://doi.org/10.1198/jasa.2009.ap08746

ADAMS, R. P., DAHL, G. E. and MURRAY, I. (2010). Incorporating side information in probabilistic matrix factorization with Gaussian processes. Preprint. Available at arXiv:1003.4944.

ANSOLABEHERE, S. and JONES, P. E. (2010). Constituents' responses to congressional roll-call voting. *Amer. J. Polit. Sci.* **54** 583–597.

ANSOLABEHERE, S. and SCHAFFNER, B. (2010). CCES common content, 2012. V3 [Version]. Available at http://hdl.handle.net/1902.1/17705 (accessed June 2, 2014).

ATHEY, S., BAYATI, M., DOUDCHENKO, N., IMBENS, G. and KHOSRAVI, K. (2018). Matrix completion methods for causal panel data models Technical report, National Bureau of Economic Research, Cambridge, MA.

ATTENBERG, J. and PROVOST, F. (2011). Inactive learning?: Difficulties employing active learning in practice. *ACM SIGKDD Explor. Newsl.* **12** 36–41.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2018). The augmented synthetic control method. Preprint. Available at arXiv:1811.04170.

BROOCKMAN, D. E., KALLA, J. L. and SEKHON, J. S. (2017). The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Polit. Anal.* **25** 435–464.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CAO, Y. and XIE, Y. (2015). Categorical matrix completion. In 2015 *IEEE* 6*th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* (*CAMSAP*) 369–372. IEEE, Piscataway, NJ.

CHAKRABORTY, S., ZHOU, J., BALASUBRAMANIAN, V., PANCHANATHAN, S., DAVIDSON, I. and YE, J. (2013). Active matrix completion. In 2013 *IEEE* 13*th International Conference on Data Mining* 81–90. IEEE, Piscataway, NJ.

CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304. MR1390519

CHEN, H., FELT, M.-H. and HUYNH, K. P. (2017). Retail payment innovations and cash usage: Accounting for attrition by using refreshment samples. *J. Roy. Statist. Soc. Ser. A* **180** 503–530. MR3601781 https://doi.org/10.1111/rssa.12208

CHIANG, K.-Y., HSIEH, C.-J. and DHILLON, I. S. (2015). Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems* 3447–3455.

PRODUCTIVITY COMMISSION (2018). Superannuation: Assessing efficiency and competitiveness. Productivity Commission draft report overview. Published April.

DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E. and WOOTTERS, M. (2014). 1-bit matrix completion. *Inf. Inference* **3** 189–223. MR3311452 https://doi.org/10.1093/imaiai/iau006

DILLMAN, D. A., SINCLAIR, M. D. and CLARK, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opin. Q.* **57** 289–304.

EARLY, K., MANKOFF, J. and FIENBERG, S. E. (2017). Dynamic question ordering in online surveys. *J. Off. Stat.* **33** 625–657.

EDWARDS, P., ROBERTS, I., CLARKE, M., DIGUISEPPI, C., PRATAP, S., WENTZ, R. and KWAN, I. (2002). Increasing response rates to postal questionnaires: Systematic review. *BMJ* **324** Art. ID 1183.

ELAHI, M., RICCI, F. and RUBENS, N. (2016). A survey of active learning in collaborative filtering recommender systems. *Comput. Sci. Rev.* **20** 29–50. MR3519039 https://doi.org/10.1016/j.cosrev.2016.05.002

FAN, W. and YAN, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Comput. Hum. Behav.* **26** 132–139.

FITHIAN, W. and MAZUMDER, R. (2013). Flexible low-rank statistical modeling with side information. Preprint. Available at arXiv:1308.4211.

FULLERTON, A. S. and XU, J. (2012). The proportional odds with partial proportionality constraints model for ordinal response variables. *Soc. Sci. Res.* **41** 182–198.

GABRY, J. and GOODRICH, B. (2016). rstanarm: Bayesian applied regression modeling via Stan. R Package Version 2.

GARNETT, R., KRISHNAMURTHY, Y., XIONG, X., SCHNEIDER, J. and MANN, R. (2012). Bayesian optimal active search and surveying. Preprint. Available at arXiv:1206.6406.

GOLBANDI, N., KOREN, Y. and LEMPEL, R. (2010). On bootstrapping recommender systems. In *Proceedings of the* 19*th ACM International Conference on Information and Knowledge Management* 1805–1808. ACM, New York.

GONZALEZ, J. M. and ELTINGE, J. L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *Proceedings of the Section on Survey Research Methods* 2081–2088. American Statistical Association, Alexandria, VA.

GROVES, R. M. (2011). Three eras of survey research. *Public Opin. Q.* **75** 861–871.

HASTIE, T., MAZUMDER, R., LEE, J. D. and ZADEH, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16** 3367–3402. MR3450542

HEBERLEIN, T. A. and BAUMGARTNER, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *Am. Sociol. Rev.* 447–462.

HERZOG, A. R. and BACHMAN, J. G. (1981). Effects of questionnaire length on response quality. *Public Opin. Q.* **45** 549–559.

JOSSE, J., HUSSON, F. et al. (2016). missMDA: A package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70** 1–31.

KARIMI, R., FREUDENTHALER, C., NANOPOULOS, A. and SCHMIDT-THIEME, L. (2011b). Towards optimal active learning for matrix factorization in recommender systems. In 2011 *IEEE* 23*rd International Conference on Tools with Artificial Intelligence* 1069–1076. IEEE, Piscataway, NJ.

KARIMI, R., FREUDENTHALER, C., NANOPOULOS, A. and SCHMIDT-THIEME, L. (2011a). Non-myopic active learning for recommender systems based on matrix factorization. In 2011 *IEEE International Conference on Information Reuse & Integration* 299–303. IEEE, Piscataway, NJ.

KLOPP, O., LAFOND, J., MOULINES, É. and SALMON, J. (2015). Adaptive multinomial matrix completion. *Electron. J. Stat.* **9** 2950–2975. MR3439190 https://doi.org/10.1214/15-EJS1093

KREUTER, F. (2013). Facing the nonresponse challenge. *Ann. Am. Acad. Polit. Soc. Sci.* **645** 23–35.

KROSNICK, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* **5** 213–236.

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci.* **116** 4156–4165.

LAWRENCE, N. D. and URTASUN, R. (2009). Non-linear matrix factorization with Gaussian processes. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 601–608. ACM, New York.

LIM, Y. J. and TEH, Y. W. (2007). Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop* **7** 15–21.

LOWELL, D., LIPTON, Z. C. and WALLACE, B. C. (2019). Practical obstacles to deploying active learning. In *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP–IJCNLP*) 21–30.

MARCUS, B., BOSNJAK, M., LINDNER, S., PILISCHENKO, S. and SCHÜTZ, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in web surveys. *Soc. Sci. Comput. Rev.* **25** 372–383.

MARLIN, B. M. and ZEMEL, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the Third ACM Conference on Recommender Systems* 5–12. ACM, New York.

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

MONTGOMERY, J. M. and CUTLER, J. (2013). Computerized adaptive testing for public opinion surveys. *Polit. Anal.* **21** 172–192.

MULDER, J. and VAN DER LINDEN, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika* **74** 273–296. MR2507381 https://doi.org/10.1007/s11336-008-9097-5

MUNGER, G. F. and LOYD, B. H. (1988). The use of multiple matrix sampling for survey research. *J. Exp. Educ.* **56** 187–191.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

PORTEOUS, I., ASUNCION, A. and WELLING, M. (2010). Bayesian matrix factorization with side information and Dirichlet process mixtures. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

RECHT, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12** 3413–3430. MR2877360

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542 https://doi.org/10.1198/016214507000000932

RENNIE, J. D. and SREBRO, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning* 713–719. ACM, New York.

ROBERTS, C., GILBERT, E., ALLUM, N. and EISNER, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opin. Q.* **83** 598–626.

RUBENS, N. and SUGIYAMA, M. (2007). Influence-based collaborative active learning. In *Proceedings of the 2007 ACM Conference on Recommender Systems* 145–148. ACM, New York.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. MR2117498

RUBINSTEYN, A. and FELDMAN, S. (2016). fancyimpute: Multivariate imputation and matrix completion algorithms implemented in Python.

SALAKHUTDINOV, R. and MNIH, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning* 880–887. ACM, New York.

SCHNABEL, T., SWAMINATHAN, A., SINGH, A., CHANDAK, N. and JOACHIMS, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. Preprint. Available at arXiv:1602.05352.

SEEGER, M. and BOUCHARD, G. (2012). Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Artificial Intelligence and Statistics* 1012–1018.

SEGALL, D. O. (2009). Principles of multidimensional adaptive testing. In *Elements of Adaptive Testing* 57–75. Springer, Berlin.

SENGUPTA, N., SREBRO, N. and EVANS, J. (2018). Simple surveys: Response retrieval inspired by recommendation systems. Preprint. Available at arXiv:1901.09659.

SETTLES, B. (2009). Active learning literature survey Technical report, Dept. Computer Sciences, Univ. Wisconsin-Madison.

SHEEHAN, K. B. (2001). E-mail survey response rates: A review. *J. Comput.-Mediat. Commun.* **6** JCMC621.

SILVA, J. and CARIN, L. (2012). Active learning for online Bayesian matrix factorization. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 325–333. ACM, New York.

SREBRO, N., RENNIE, J. and JAAKKOLA, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems* 1329–1336.

SREBRO, N. and SALAKHUTDINOV, R. R. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems* 2056–2064.

SUTHERLAND, D. J., PÓCZOS, B. and SCHNEIDER, J. (2013). Active learning and search on low-rank matrices. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 212–220. ACM, New York.

THOMAS, N., RAGHUNATHAN, T. E., SCHENKER, N., KATZOFF, M. J. and JOHNSON, C. L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Surv. Methodol.* **32** 217–231.

TOURANGEAU, R., KREUTER, F. and ECKMAN, S. (2015). Motivated misreporting: Shaping answers to reduce survey burden. In *Survey Measurements*. *Techniques*, *Data Quality and Sources of Error* 24–41. Campus Verlag, Frankfurt.

TRAN, D., KUCUKELBIR, A., DIENG, A. B., RUDOLPH, M., LIANG, D. and BLEI, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. Preprint. Available at arXiv:1610.09787.

UDELL, M. and TOWNSEND, A. (2019). Why are big data matrices approximately low rank? *SIAM J. Math. Data Sci.* **1** 144–160. MR3949704 https://doi.org/10.1137/18M1183480

WAGNER, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opin. Q.* **74** 223–243.

XU, M., JIN, R. and ZHOU, Z.-H. (2013). Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems* 2301–2309.

YAMMARINO, F. J., SKINNER, S. J. and CHILDERS, T. L. (1991). Understanding mail survey response behavior a meta-analysis. *Public Opin. Q.* **55** 613–639.

ZHANG, C., TAYLOR, S. J., COBB, C. and SEKHON, J. (2020). Supplement to "Active matrix factorization for surveys." https://doi.org/10.1214/20-AOAS1322SUPPA, https://doi.org/10.1214/20-AOAS1322SUPPB

ZHOU, T., SHAN, H., BANERJEE, A. and SAPIRO, G. (2012). Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proceedings of the* 2012 *SIAM International Conference on Data Mining* 403–414. SIAM, Philadelphia, PA.