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. The most informative questions per respondent are chosen sequentially using active learning with variance minimization. We begin with Gaussian responses standard in matrix completion and derive a simple active strategy with closed-form posterior updates. Next we model responses more realistically as ordinal logit; posterior inference and question selection are adapted to the nonconjugate setting. We simulate our matrix sampling procedure on data from real-world surveys. Our active question selection achieves efficiency gains over baselines and can benefit from available side information about respondents.
entropy, traded off with dropout probability. They also review the literature on respondent burden in detail and cite adaptive information-maximizing strategies in other fields.

1.2 Optimal design and active learning

Classical research on surveys addresses how to configure survey parameters to minimize variance of estimates. Each member of the population is sampled with some inclusion probability; optimal inclusion probabilities have been derived for a variety of survey designs [44]. More broadly, guidance for variance minimization comes from the classical field of optimal experimental design. Optimal design proceeds by minimizing a measure of inverse Fisher information, which is the asymptotic variance of the maximum likelihood estimator and a lower bound on the variance of unbiased estimators. Different measures of the inverse information matrix define different design criteria: A-optimality minimizes the trace, D-optimality minimizes the determinant and E-optimality minimizes the maximum eigenvalue [48]. Gonzalez and Eltinge use A- and D-optimality to obtain a scalar objective function involving all the variances they want to control [20].

Closely related to optimal design for surveys is optimal subsampling, useful when training a model on the full dataset is too computationally demanding. For instance, prior to logistic regression, Wang et al derive optimal subsampling weights using A-optimality to minimize the variance of the subsample maximum likelihood estimate [55].

Optimal design occupies a place in the broader field of active learning, which encompasses many strategies for selecting informative training points when labeling them is expensive. Sequential active learning strategies iteratively select the next query point. A common baseline is uncertainty sampling, which simply chooses the point with greatest predictive uncertainty [48]. Uncertainty sampling is simple but myopic; it does not consider the global effect of item selection on the model or the goal at hand. MacKay distinguishes between various goals of active learning, such as obtaining maximal information about model parameters, or maximizing model performance in a region of input space [31]. For the latter objective, a principled approach is to choose the point that minimizes the variance component of generalization error – the predictive variance integrated over the input distribution. The seminal work by Cohn et al derives, for several models, closed-form expressions for this integrated variance given a new training point, which can be optimized to suggest the next query point [11].

The extension to a multi-step search horizon is studied by Garnett et al [18]. They consider how to optimally query points in a dataset to estimate the proportion of a binary class. Casting the problem in a Bayesian decision-theoretic framework, they use a branch-and-bound strategy to prune the search space. In their experiments, a multi-step lookahead provides marginal gains over the greedy one-step lookahead; both outperform random and uncertainty sampling.

Theoretical work has established the lower sample complexity of active learning relative to passive sampling in certain settings, usually within binary classification [48]. Dasgupta et al show this for data distributed uniformly on the unit sphere, where the base learner is a linear separator through the origin [12]. However, the advantage of active learning disappears when the learner is inhomogeneous, and is recovered by weakening the definition of sample complexity [6]. In addition, Attenberg and Provost point out several practical challenges to the adoption of active learning [5]. These challenges include choosing an initial base learner and query selection strategy within the label budget; poor query selection by non-robust strategies, especially with rare classes or concepts; and artificial advantages given to active learning in research experiments.

1.3 Active learning for matrix factorization

The spectrum of active learning strategies has been specialized to matrix factorization to select the most informative entries of the response matrix. Often research asks, in a recommender systems context, which movies to prompt users to rate; the researcher seeks accurate predictions of user ratings or rankings of unseen movies. One baseline strategy simply prompts for the most popular items, since users are more likely familiar with them and more likely to remain attentive [14]. Many variants of uncertainty sampling have been proposed. They represent unobserved matrix entries with disparate models, such as the graphical lasso and ensembles [9]. Sutherland et al select matrix elements with highest posterior variance in a probabilistic matrix factorization model [52].

Other strategies quantify the global effect of item selection and try to maximize this effect. Rubens and Sugiyama find the item with largest influence empirically by perturbing ratings and computing the change in predictions [41]. Karimi et al seek the item that would change user factors
most if its rating were known [25]. Instead of influence, Silva and Carin maximize mutual information between selected and unobserved instances [49]. They also suggest an efficient alternative: sort both user and item factors by posterior variance, match them into user-item pairs, and query the corresponding matrix entries [49]. Our eventual strategy recalls elements of theirs, as they approximate posteriors with variational Bayes and reduce approximate posterior variance to the trace.

Still other strategies directly target the predictive error of matrix factorization [19, 26]. As responses are not known before querying, direct minimization of RMSE or MAE relies on assumptions about the empirical rating distribution, such as stationarity. Beyond prediction, active learning for recommender systems could target objectives like profitability or user satisfaction [42, 52].

1.4 Computerized adaptive testing

Active learning for matrix factorization could be posed as adaptive item selection that places respondents on latent scales with maximal precision. The largely separate literature of item response theory has long pursued this goal. Computerized adaptive testing (CAT) algorithms customize the questions asked of individual test-takers adaptively, to optimally determine their latent ability parameters with some question budget. Montgomery and Cutler advocate for applying CAT methods to public opinion surveys [34]. They model answers to political knowledge questions using logistic regression with a one-dimensional latent ability parameter. Their item selection strategy, which minimizes expected posterior variance of ability, can shorten a battery by 40% while retaining the measurement accuracy of the fixed battery.

The generalization of latent ability to higher dimensions is handled by Segall’s work on multidimensional adaptive testing (MAT) in a Bayesian setting [47]. The probability of a correct response is a logistic function of the inner product of multivariate normal ability parameters and fixed ability discrimination parameters. Since the logistic form prevents a closed-form exact posterior update, Segall uses a Laplace approximation to the posterior of user ability parameters. Segall selects the question that maximizes the determinant of the precision matrix (D-optimal), equivalently minimizing the size of the posterior credibility region.

Approaching the same MAT problem with optimal design, Mulder and Van der Linden directly optimize the Fisher information of ability parameters [35]. They note that the trace of inverse information includes the determinant as a factor, so A- and D-optimality should act similarly. Their simulations show A- and D-optimality outperform a random selection baseline, while E-optimality is worse than random. In a separate paper addressing the Bayesian MAT framework, Mulder and Van der Linden analyze additional item selection criteria based on KL divergence and mutual information [36].

Like [34], we argue the item response theory literature can inform optimal design of adaptive surveys. The principled item selection for MAT in [47] is possibly the closest work to ours. This work relies on the latent low-rank structure of the response matrix, and is clarified by the formalism of matrix factorization. In turn, we offer to matrix factorization an active learning technique that demonstrably outperforms random sampling.

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; then $u_i \in \mathbb{R}^r$ and $v_j \in \mathbb{R}^r$ for all $i$ and $j$, and $r$ is small.

Matrix completion performs matrix factorization on a response matrix with some unobserved entries. Let $I$ be an indicator matrix for whether the corresponding responses in $R$ exist, so that $I_{ij} = 1$ implies user $i$ responded to question $j$ with value $R_{ij}$. Matrix completion finds $U$ and $V$ that minimize the reconstruction error $u_i^T v_j$ for $R_{ij}$ on observed entries, where $I_{ij} = 1$.

Matrix completion that enforces a hard rank constraint is nonconvex and generally intractable [16]. It is common to work with a convex relaxation that instead regularizes the nuclear norm, or sum of singular values [50]. This optimization problem seeks a matrix $Z$, in place of $UV^T$. 
that minimizes reconstruction error, and encourages a low-rank solution by favoring sparsity in the singular values.

\[
\min_{\mathbf{Z}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} I_{ij}(R_{ij} - Z_{ij})^2 + \lambda \|\mathbf{Z}\|_* \tag{1}
\]

This nuclear norm regularized problem enjoys theoretical guarantees: recovery of the complete matrix occurs with high probability when \(O(n \text{ polylog}(n))\) entries are observed at random, with or without noise [39, 37]. Moreover, (1) has an efficient solution in the SoftImpute algorithm by Mazumder et al [33]. SoftImpute iteratively computes the SVD of \(\mathbf{Z}\), 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 \(\mathbf{Z} = \mathbf{U} \mathbf{D} \mathbf{V}^T\) for some matrices \(\mathbf{U}, \mathbf{D}, \mathbf{V}\).

An alternate formulation of matrix completion, introduced by [40], penalizes the Frobenius norm of \(\mathbf{U}\) and \(\mathbf{V}\):

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

Nonconvex in \(\mathbf{U}\) and \(\mathbf{V}\), (2) is solved via gradient descent. This formulation is useful for largescale problems with low rank, since it is less expensive to operate on \(\mathbf{U}\) \((n \times r)\) and \(\mathbf{V}\) \((k \times r)\) than \(\mathbf{Z}\) \((n \times k)\). Due to an identity relating the nuclear norm and sum of Frobenius norms, if \(\lambda_U = \lambda_V\) and the solution to (1) has rank at most \(r\), then this is also a solution to (2) [16, 33].

However, the user and question factors resulting from either optimization are point estimates, as are the imputed survey responses. We would like some measure of uncertainty over the imputed responses. Further, we seek a strategy for actively selecting users and questions to survey next based on the uncertainty reduction achieved. For this uncertainty quantification, we turn to Bayesian probabilistic matrix factorization (BPMF) [43].

2.2 Bayesian probabilistic matrix factorization

BPMF, introduced by Salakhutdinov and Mnih, models user and question factors as independent and normally distributed around a prior mean and covariance. Responses add zero-mean, constant-variance Gaussian noise to the inner product of user and question factors. Following the notation in [43],

\[
\begin{align*}
\mathbf{u}_i & \sim \mathcal{N}(\mu_U, \Lambda_U^{-1}) \\
\mathbf{v}_j & \sim \mathcal{N}(\mu_V, \Lambda_V^{-1}) \\
R_{ij} \mid \mathbf{U}, \mathbf{V} & \sim \mathcal{N}(\mathbf{u}_i^T \mathbf{v}_j, \alpha^{-1})
\end{align*}
\]

The authors of [43] note an interesting connection to the Frobenius norm regularized problem: for prior hyperparameters \(\mu_U = \mu_V = 0\), \(\Lambda_U = \alpha_U I\) and \(\Lambda_V = \alpha_V I\), the MAP estimate of \(\mathbf{U}\) and \(\mathbf{V}\) conditional on \(\mathbf{R}\) is the solution to (2) with \(\lambda_U = \alpha_U / \alpha\) and \(\lambda_V = \alpha_V / \alpha\).

The authors perform posterior inference via Gibbs sampling. They derive the following complete conditional for \(\mathbf{u}_i\) (the form for \(\mathbf{v}_j\) is analogous):

\[
P(\mathbf{u}_i \mid \mathbf{R}, \mathbf{V}, \mu_U, \Lambda_U, \alpha) \propto P(\mathbf{u}_i, \mathbf{R}_i \mid \mathbf{V}, \mu_U, \Lambda_U, \alpha) = \prod_{j=1}^{k} [P(R_{ij} \mid \mathbf{u}_i^T \mathbf{v}_j, \alpha^{-1})]^{I_{ij}} P(\mathbf{u}_i \mid \mu_U, \Lambda_U^{-1})
\]

The complete conditional is conjugate normal with mean \(\mu_i^*\) and precision \(\Lambda_i^*\):

\[
\begin{align*}
\mathbf{u}_i \mid \mathbf{R}, \mathbf{V}, \mu_U, \Lambda_U, \alpha & \sim \mathcal{N}(\mu_i^*, [\Lambda_i^*]^{-1}) \\
\Lambda_i^* &= \Lambda_U + \alpha \sum_{j=1}^{k} I_{ij} \mathbf{v}_j^T \mathbf{v}_j \\
\mu_i^* &= \Lambda_i^*[\Lambda_i^*]^{-1} \left( \alpha \sum_{j=1}^{k} I_{ij} R_{ij} \mathbf{v}_j + \Lambda_U \mu_U \right)
\end{align*}
\]
This can be recognized as the posterior for Bayesian linear regression with a Gaussian prior and Gaussian noise: \( u_i \) are the coefficients, \( V \) is the design matrix and \( \alpha^{-1} \) is the noise variance. The expression for \( \mu_i^t \) also arises in MAP estimation for BPMF, as the coordinate ascent update for \( u_i \), for prior hyperparameters \( \mu_U = 0, \Lambda_U = \alpha_U I \) (see section A.1).

### 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 \( i \) enter the survey pool, and we want to select questions optimally for learning \( u_i \). So far we assume we have no side information about users.

The BPMF model admits a convenient online formulation for updating our knowledge about \( u_i \) given responses from user \( i \). Suppose at time \( t \), \( u_i \sim \mathcal{N}(\mu_i^{(t)}, [\Lambda_i^{(t)}]^{-1}) \). At time \( t + 1 \) we gather the response by user \( i \) to question \( j \). Then the posterior for \( u_i \) is

\[
\begin{align*}
u_i^{(t+1)} | R^{(t+1)}, V, \mu_U, \Lambda_U, \alpha & \sim \mathcal{N}(\mu_i^{(t+1)}, [\Lambda_i^{(t+1)}]^{-1}) \\
\Lambda_i^{(t+1)} &= \Lambda_i^{(t)} + \alpha v_j v_j^T \\
\mu_i^{(t+1)} &= \left[\Lambda_i^{(t+1)}\right]^{-1} (\alpha R_{ij} v_j + \Lambda_i^{(t)} \mu_i^{(t)})
\end{align*}
\]

We now consider how to choose question \( j \) optimally. One standard approach in active learning is to minimize posterior variance, here the posterior variance of \( u_i \) is a matrix. Optimal design proposes minimizing various measures of the inverse information matrix, including the trace, determinant and maximum eigenvalue (\( A-, D- \) and \( E- \) optimality, respectively). Other optimality criteria are concerned with minimizing predictive variance over a certain input distribution.

For simplicity, let \( \mu, \Sigma \) denote the posterior mean and variance of \( u_i \). Inspired by the \( A- \) optimality criterion, we consider minimizing \( \text{tr} \Sigma \), the sum of the component-wise variances of \( u_i \). This criterion can be related to predictive variance as follows. Let \( \mathcal{P} \) be the uniform distribution on the unit sphere \( \{ v : \| v \|_2 = 1 \} \), and suppose \( \tilde{v} \sim \mathcal{P} \) independently of \( u_i \). We care about the prediction \( u_i^T \tilde{v} \). The predictive variance is (see A.2 for details)

\[
\text{Var} (u_i^T \tilde{v}) = \frac{1}{r} E \| u_i \|^2 = \frac{1}{r} (\text{tr} \Sigma + \| \mu \|^2)
\]

For any \( \mu \), minimizing \( \text{tr} \Sigma \) corresponds to minimizing a lower bound on the predictive variance along uniform latent directions. We cannot treat \( \mu \) as a constant since both \( \mu \) and \( \Sigma \) are affected by the choice of question \( j \).

Having justified our \( A \)-optimality strategy, we solve the following optimization at time \( t \):

\[
\min_j \text{tr} \left[\Lambda_j^{(t+1)}\right]^{-1}
\]

This is equivalent to solving the following, where \( \lambda_\ell (M) \) denotes the \( \ell \)th eigenvalue of \( M \):

\[
\min_j \sum_{\ell=1}^r \lambda_\ell \left(\left[\Lambda_j^{(t+1)}\right]^{-1}\right) = \min_j \sum_{\ell=1}^r \lambda_\ell \left(\left[\Lambda_j^{(t+1)}\right]^{-1}\right)
\]

This variance criterion penalizes small eigenvalues of the precision matrix, along the directions in which it has least information. Information is acquired by sampling \( v_j \) that lie in those directions. The optimal sampling strategy chooses questions as a function of their informativeness and their contribution to less explored directions. Provided questions exist in many directions with similar magnitudes, the strategy prefers new questions with factors \( v_j \) roughly orthogonal to those of older questions.

Algorithm 1 summarizes our active strategy. Note some limitations of this simple version. The algorithm is greedy with a one-step horizon; we could easily extend the active strategy to select multiple questions to ask in the next timestep. The optimal sequence of questions can be computed offline, as the optimality criterion does not depend on the actual responses. This results from the assumption that \( V \) is fixed and the Gaussian assumptions inherent in BPMF, but the implication is unrealistic. There is one fixed question order for all respondents. Finally, nonresponse is assumed ignorable; in other words, responses are missing at random. This assumption is unrealistic in practice, and modeling the missingness pattern could reduce bias.
Algorithm 1: Active question selection for single user

given: Question factors $v_1, \ldots, v_k$; prior parameters $\mu_U, \Lambda_U$ for user $i$; noise variance $\alpha^{-1}$

$O \leftarrow \emptyset$, $U \leftarrow \{1, \ldots, k\}$
$\mu_0 \leftarrow \mu_U$, $\Lambda_0 \leftarrow \Lambda_U$

for $t \leftarrow 1$ to $k$ do

// choose next question
$j \leftarrow \arg \min_{\ell \in U} \text{tr} \left[ \left( \Lambda_{t-1} + \alpha v_{\ell} v_{\ell}^T \right)^{-1} \right]$
$R_{ij} \leftarrow \text{get_response}(i, j)$

// update user posterior
$\Lambda_t \leftarrow \Lambda_{t-1} + \alpha v_j v_j^T$
$\mu_t \leftarrow \Lambda_t^{-1} \left( \alpha R_{ij} v_j + \Lambda_{t-1} \mu_{t-1} \right)$

// update observed, unobserved question sets
$O \leftarrow O \cup \{j\}$, $U \leftarrow U \setminus \{j\}$

3 Data collection and analysis

3.1 Datasets

We simulate the active strategy 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 adult U.S. citizens conducted by YouGov [3] that seeks to gauge voter opinions about prevailing political issues and elected officials, before and after an election. 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 surveys from 2012 and 2016, limiting consideration to Common Content questions that ask respondents to evaluate national political issues or entities on a binary or ordinal scale. For robustness checks we expand the question set to include voter demographics, party identification and other characteristics; we refer to this as the “full” CCES dataset. We exclude questions about voter actions in the past year and opinions of state or local representatives, as well as questions with a majority of responses missing.

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: about 4% in 2012 and less than 1% in 2016. The missingness distributions by question and by user are shown in Figure 1.

3.2 Simulating the active strategy

Our simulations of the active strategy begin by randomly splitting the respondent set into a training half and a simulation half. On the responses for the training half, we perform SoftImpute, learning the fixed $V$ for the active strategy. We choose SoftImpute for its efficiency and empirical stability of $V$ across simulations. The regularization parameter $\lambda$ is selected by grid search with warm starts as recommended in [33], based on mean absolute error on a 20% validation set within the training half.

| Dataset                        | Number of respondents | Number of questions |
|--------------------------------|-----------------------|---------------------|
| Facebook on-platform survey    | 11793                 | 53                  |
| CCES 2012                      | 54535                 | 31                  |
| CCES 2016                      | 64600                 | 43                  |
| CCES 2016 (full)               | 64600                 | 66                  |

Table 1: Dataset characteristics. CCES 2016 (full) refers to CCES 2016 with extra covariates.
On the simulation half, we hold out a random 20% of each user’s responses, and simulate running the survey on the remaining 80% of responses. We select the next question per respondent using the active strategy, reveal available responses to that question, and perform the BPMF posterior update for each user. Using the MAP estimate for all user factors along with the fixed $V$, we compute predictions on the holdout set and evaluate error. We repeat this process until all questions have been asked.

Error measures include mean squared error, mean absolute error, and bias of predictions, averaged over all questions. For interpretability, we also compute the proportion of predictions with the wrong sign. All simulations compare the active strategy to a baseline of asking questions in a random order per respondent and, in the case of CCES, existing question order.

We experimented with variations on the simulation procedure. For instance, rather than estimating question factors $v_j$ from SoftImpute, we attempted to use the columns of $V^T$ from solving the Frobenius norm regularized problem (2), reasoning that these correspond to the MAP estimate for BPMF for certain simple hyperparameters. However, this nonconvex optimization yielded highly variable question factors and orderings (Figure C.6). With SoftImpute, accounting for nonidentifiability, question factors are relatively stable across simulations. We also tried minimizing measures of posterior variance other than the trace, like the determinant and maximum eigenvalue, which correspond to D- and E-optimality respectively. These optimal design criteria lead to similar conclusions, with greater variability in predictive error for E-optimality (Figures C.2, C.3, C.4, C.5).

Our main results use a rank-4 matrix decomposition ($r = 4$); this results in lower predictive error than $r = 2$, while keeping the dimensionality of latent space manageable. In the active strategy, we set the prior mean $\mu_U$ and prior precision $\Lambda_U$ using empirical Bayes. Specifically, we set $\mu_U$ and $\Lambda_U^{-1}$ to the sample mean and covariance of the rows of $UD$, the implied user factors from SoftImpute. It remains to set the noise variance $\alpha^{-1}$. 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 experimented with smaller values. Future work should estimate $\alpha$ from the training half of responses.

4 Results

4.1 Efficiency

Below we showcase simulated survey strategies on the 2016 CCES; similar results for the 2012 CCES and the Facebook survey appear in Appendix D and E, respectively. In Figure 2 we see the active strategy outperforms the random strategy and existing question order in terms of mean squared and mean absolute prediction error. The active strategy also correctly predicts the sign of survey responses more often than the baselines. Over the course of the survey, the active and sequential strategies suffer worse bias than the random strategy, as they select questions in a deterministic order.

Figure 3 clarifies the improved efficiency of the active strategy. It measures the question complexity of the active strategy relative to the random strategy—the number of questions each needs to attain a certain error level. The active strategy almost always requires fewer questions. Suppose we ask 20 questions in a random order for each respondent; the active strategy attains the same error with half as many questions.

We verify the active strategy works as intended in Figure C.1—it minimizes the active learning objective. While the random strategy decreases the objective function with every new question by virtue of gathering information and reducing posterior variance, it only catches up to the information gain of the active strategy at the end of the simulated survey.
Figure 2: We plot the error attained by each sampling strategy on the holdout set for the 2016 CCES, averaged over all questions. Error metrics include mean squared error, mean absolute error, the proportion of predictions with the wrong sign, and mean signed error or bias. The solid line depicts average error across 10 simulations of each strategy; shaded regions represent two standard deviations. In 500 simulations of each strategy, across survey questions and metrics, such intervals have reasonable coverage (approximately 95%, always between 93-98%) of data points.
We depict the complexity of the active strategy relative to random-order questions for the 2016 CCES. For the error metrics in Figure 2, the solid line plots the number of questions required by each strategy to attain the same level of error. For instance, the active strategy requires 20 questions to reach the same MSE as the random strategy with 30 questions. When the curve lies above the dashed $45^\circ$ line, active sampling outperforms random. These curves are obtained by fitting, for each strategy, a loess smoother $f_{\text{strategy}}(\epsilon)$ to predict number of questions asked for a given error level. The curves simply plot $(f_{\text{random}}(\epsilon), f_{\text{active}}(\epsilon))$ for the range of $\epsilon$ attained by both strategies.

4.2 Question order

To better understand how the active strategy operates, we examine the factors assigned to individual questions and their resultant order in the active survey. Figure 4a visualizes question factors as vectors in latent space. For ease of visualization, these results use a rank-2 decomposition, keeping all other simulation settings the same. The 2016 CCES questions are distributed unevenly in latent space. Longer vectors indicate questions that feature more prominently in the principal components. (As SoftImpute is solved via soft-thresholded SVD, it effectively performs PCA on some complete version of the incomplete responses.)

Based on our earlier analysis of how to sample latent space efficiently, we might expect the active strategy to begin with longer, relatively orthogonal vectors. However, it front-loads medium-length vectors in the second and fourth quadrants – the direction where the user factors have greatest prior variance. The active strategy spends its initial question budget sweeping around this direction of latent space, gaining precision. The longest question vectors get asked midway through the survey, and the shortest vectors toward the end. With a higher $\alpha$, which increases the relative information conveyed in each response, the active strategy reaches the longest question vectors in the first and third quadrants earlier (Figure C.7).

Visualizing the posterior of user factors during the survey reinforces this story. We can see from just 10 user trajectories in Figure 6a that the initial questions in the active ordering lie in the direction of greatest user variance. The active ordering first places users along this direction before seeking information along the orthogonal direction. Most precision gains occur along this direction, early in the survey, as the narrowing confidence region in Figure 6b shows. The confidence region provides another measure of error over the survey duration.

Cross-referencing Figure 4a against Figure 4b, which shows the top questions across simulations and their rank distribution, we can locate concepts in latent space. Questions 2 and 3 concerning support for abortion restrictions point in almost the same direction; question 6 concerning abortion rights points in the opposite direction. This general direction broadly indicates partisanship: questions in the top left quadrant address policies favored by Democrats, whereas questions in the bottom right quadrant address policies favored by Republicans. It makes sense that most prior variance lies along this latent direction. In the orthogonal direction, we have a “none-of-
the-above” question about immigration policies (12), which was answered affirmatively by 5% of respondents. Opposite that are questions about longer sentences for felons who have committed violent crimes (14) and background checks for all gun purchases (17), which are broadly popular policies supported by 84% and 90% of respondents, respectively. This direction seems to indicate bipartisan policies. With a 2-dimensional decomposition, the second direction is predetermined, and the latent concepts are limited to partisanship and bipartisanship.

To capture more meaningful concepts, our main results expand the dimensionality of latent space. In Figure 5 we examine the active ordering for a rank-4 decomposition. Across simulations, the order of the first 15 questions is almost constant. The foremost question is: “If you were in Congress would you vote FOR or AGAINST...” the Affordable Care Act of 2009 (also known as Obamacare)?” Questions about environment and abortion policies are prioritized over questions about the economy, crime and gun control. Unlike with the rank-2 decomposition, there is room in latent space to capture approval of elected representatives (6, 9, 12). (It must be noted that with a rank-2 decomposition for the 2012 CCES, approval of Congress is captured well – in the negatively bipartisan direction, with 79% of respondents somewhat or strongly disapproving. See Figure D.3a.)

To confirm these results are not simply an artifact of our question inclusion criteria for the CCES, we perform a series of robustness checks. We progressively add questions about respondent political affiliation, demographics, education and other characteristics. Questions with categorical responses, like race, are converted into indicators for each common response. This one-hot encoding artificially creates a separate survey question per response value, so we lose some fidelity with the CCES. See Appendix C.2 for box plots of question rank with the additional questions included.

The actively chosen ordering with augmented questions remains largely faithful to the active ordering in Figure 5. While questions about gender, party identification, parenthood and home ownership slot into the first 15 positions, the same questions about political issues dominate the top 15 in largely the same order. This provides more support for earlier inclusion of questions concerning the environment, abortion, and Obamacare. Respondent covariates displace questions about approval of senators. Interestingly, the active strategy leaves questions about race and education until later in the survey, possibly because these one-hot-encoded variables are not well captured by a low-rank matrix decomposition.

4.3 Changes in question order over time

We also compute the active ordering for the 2012 CCES using the same simulation parameters \( r = 4, \alpha = 1 \); see Figure D.4. As in 2016, the Obamacare question comes first, followed by a question about Obamacare repeal in third place to gather more information along this latent direction. In 2016 the active strategy does not address immigration until the 10th question; in 2012 four of the top 10 questions are about harsher immigration policies. Gay marriage also features more prominently in the 2012 active ordering, coming in 4th versus 14th in 2016.

Inspecting the 4-dimensional question factors for both CCES years (Table C.1 and D.1), the first two principal components seem to indicate partisanship and popularity, as in our earlier analysis of 2-dimensional question factors. However, in 2012 the third principal component aligns with support for isolationist or xenophobic policies, while no separate principal component loading heavily on the immigration questions appears in 2016. One possible explanation is that immigration, which was a separate enough concept in 2012 to command its own principal component, became absorbed into partisanship in 2016 as Trump claimed the immigration issue. Hence, the active strategy asks immigration questions early to determine user position along this concept in 2012 but not 2016.

The fourth principal component in 2012 may indicate fiscal conservatism and social liberalism, as it correlates support for gay marriage, ending the “Don’t ask, don’t tell” policy, and the Ryan and Simpson-Bowles budget plans reducing federal spending. In 2016, the third principal component seems to indicate approval of elected officials. The fourth principal component correlates support for greener environmental policies, support for abortion restrictions and opposition to gay marriage. This suggests a group of socially conservative or religious respondents who are concerned about the environment. The active strategy doubles down on both environmental and abortion questions in order to ascertain membership in this group, in addition to partisanship. Going beyond two latent dimensions helps to identify parts of the electorate that do not behave according to conventional partisan wisdom; the active strategy actively seeks information along these more subtle directions.
(a) We show the positions of the fixed question factors in latent space. The darker a vector is shaded, the earlier the corresponding question is asked by the active strategy. We label the first questions selected by the active strategy.

(b) We show the rank of each question across 10 simulations of the active strategy as a box plot. The active strategy produces a stable question ordering.

Figure 4: Active ordering for 2016 CCES using a rank-2 matrix decomposition.
Figure 5: Active ordering for 2016 CCES using a rank-4 matrix decomposition. We show the rank of each question across 10 simulations of the active strategy as a box plot. The top 15 questions are relatively stable.
Figure 6: We visualize the evolution of the user factors posterior over the course of the survey. As questions are asked, the posterior mean for a user is moved in the direction of the corresponding question factors by amounts depending on the response values. The confidence region also narrows in this direction, representing precision gained.

5 Side information

Active question ordering delivers a similar improvement for the Facebook survey – asking 20 questions achieves the same error as asking 30 questions in a random order – though the results are more variable. When we incorporate side information about respondents, both active and random strategies may see efficiency gains. Theoretical results have established that sufficiently informative side information improves the sample complexity of matrix completion [56, 10]. We give a simple proof-of-concept of the value of side information, by subgrouping respondents based on two covariates: country and length of time since joining Facebook. For each subgroup, we compute the mean and covariance of user factors in the training half, and use these as the empirical Bayes prior for all subgroup users in the simulation half. It is unclear whether the strategies using subgroup information attain lower predictive error than their fully pooled counterparts throughout the simulated survey. Future work should aim for shrinkage across subgroups, perhaps with a hierarchical model, and incorporate more covariate information.

As an alternative way of incorporating side information, we revisit the full CCES 2016 dataset used for robustness checks on question order. We reveal responses to all questions about respondent covariates before simulating a survey with the remaining questions, so that BPMF can update each user’s prior to include these “free” covariates. In terms of survey design, this can be regarded as requiring a set of covariate questions upfront, before applying an active ordering to opinion questions. Early in the survey, free covariates reduce error metrics for both random and active strategies, at the cost of introducing bias (Figure 7). The information advantage of free covariates disappears as more questions are asked, so that strategies making use of free covariates end up at a higher final error than their agnostic counterparts. The point where free covariates are no longer useful depends on the error metric. For MSE the active strategy breaks even around 8 questions; the random strategy breaks even around 17. The active ordering with free covariates is similar to that without (Figure C.11).
Figure 7: We plot error on the holdout set for the 2016 CCES, averaged over all questions. This time, we also make responses to covariate questions available for free (“free-cov”). At the beginning of the survey, information from free covariates narrows down user position in latent space, driving down predictive error. The free covariates participate in matrix factorization; the SoftImpute loss function now includes terms for reconstruction error on free covariates. This changes the question factors estimated from the training half. Meanwhile, predictive error is still evaluated on held-out survey responses only, which do not include covariates. Thus, we have biased question factors away from those that would optimize predictive error, in exchange for variance reduction from knowing the free covariates. This bias manifests as higher predictive error for the “free-cov” strategies at the end of the survey.
6 Ordinal logit response model

In this section we explore adjusting the model to better capture binary and ordinal response values. We replace the Gaussian likelihood for responses with an ordinal (ordered) logit likelihood. As we will see, question selection now depends on the respondent’s previous answers; the active question order becomes nondeterministic. The determinism of the active question order for BPMF is an artifact of the Gaussian likelihood.

To model quantized outputs in the response matrix, prior work has used the ordinal logit likelihood \[8\] and similar link functions \[27\]. These works formulate matrix completion as maximum likelihood with nuclear norm regularization. We continue with a Bayesian probabilistic matrix factorization approach. Posterior inference for \(U, V\) in the ordinal logit model involves non-conjugacy, so we resort to variational inference, also used for matrix completion in \[30, 46\]. We forfeit the closed-form posterior update exploited by the active strategy for BPMF, but optimal design via Fisher information offers a way forward.

6.1 Probabilistic matrix factorization model

The model becomes:

\[
\begin{align*}
\mathbf{u}_i & \overset{iid}{\sim} \mathcal{N}(\mu_U, \Lambda_U^{-1}) \\
\mathbf{v}_j & \overset{iid}{\sim} \mathcal{N}(\mu_V, \Lambda_V^{-1}) \\
R_{ij} \mid U, V & \overset{ind}{\sim} \text{OrdLogit}(\mathbf{u}_i^T \mathbf{v}_j, \beta_j)
\end{align*}
\]

By modeling \(R_{ij}\) as ordinal logit, we allow \(R_{ij}\) to take values in the range \(\{1, 2, \ldots, M_j + 1\}\) with probabilities \(\{\pi_{j,1}, \pi_{j,2}, \ldots, \pi_{j,M_j+1}\}\). Note response frequencies vary across questions, and questions need not have the same number of response values. The probabilities are defined by the logistic link and a series of cutpoints \(\beta_j = (\beta_{j,1}, \ldots, \beta_{j,M_j})\). For simplicity of presentation, we drop the indexing for question \(j\). Thus \(R_{ij}\) takes values in \(\{1, 2, \ldots, M + 1\}\) with probabilities \(\{\pi_1, \pi_2, \ldots, \pi_{M+1}\}\), parameterized by cutpoints \(\beta = (\beta_1, \ldots, \beta_M)\) as follows:

\[
\text{logit} \left( \sum_{k=1}^{\ell} \pi_k \right) = u_i^T v_j + \beta_\ell \quad (\ell = 1, \ldots, M)
\]

\[
\pi_{M+1} = 1 - \sum_{k=1}^{M} \pi_k
\]

6.2 Inference

We perform posterior inference on \(U, V\) in the above model when estimating user and question factors from the training half, and again after each iteration of the simulation, when the number of questions increases by one. From the updated posteriors per iteration, we compute predictive error on held-out survey responses. See Algorithm 2.

Given responses \(R\), we obtain approximate posteriors for \(U\) and \(V\) in the above model using mean-field variational inference in \texttt{edward} \[54\]. Our variational distributions are fully factorized Gaussian:

\[
\begin{align*}
q(U) &= \prod_{i=1}^{n} q(u_i) = \prod_{i=1}^{n} \prod_{j=1}^{r} \mathcal{N}(\mu_{ij}, \sigma_{ij}) \\
q(V) &= \prod_{j=1}^{k} q(v_j) = \prod_{j=1}^{k} \prod_{i=1}^{r} \mathcal{N}(\nu_{ji}, \tau_{ji})
\end{align*}
\]

Variational inference finds parameters \(\{\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}\}\) that maximizes the evidence lower bound, or equivalently minimizes the KL divergence between the variational distribution and the true posterior. To predict \(R_{ij}\), we set \(u_i\) and \(v_j\) equal to their variational means \(\mu_i\) and \(\nu_j\) and compute the mean of the resulting ordinal logit random variable.

As inputs to variational inference, we employ priors \(\mu_U = \mu_V \equiv 0\) and \(\Lambda_U = \Lambda_V \equiv I_r\). To derive cutpoints \(\beta\), we follow the inverse approach in the \texttt{rstanarm} package \[17\]. For each question,
we obtain a length-\((M+1)\) vector of probabilities \(\pi = (\pi_1, \ldots, \pi_M, 1 - \sum_{i=1}^{M} \pi_i)\) from the simplex, corresponding to the ordinal response values. We then apply the logit transform to the first \(M\) entries of \(\text{cumsum}(\pi) = (\pi_1, \pi_1 + \pi_2, \ldots, \sum_{i=1}^{M} \pi_i, 1)\), obtaining
\[
\{\beta_j\}_{j=1}^M = \log \frac{\sum_{i=1}^{j} \pi_i}{1 - \sum_{i=1}^{j} \pi_i}
\]
We draw \(\pi \sim \text{Dirichlet}(c_1, c_2, \ldots, c_{M+1})\), where the concentration parameters are prior counts of the response values. That is, we set \(c_\ell\) equal to the number of times a respondent answers \(\ell\) to this question in the training half.

### 6.3 Active learning formulation

We also update our item selection strategy for the ordinal logit response model. In this situation, we consider \(V\) to be fixed, using the posterior mean of \(q(V)\) given by \(\{\nu_j\}_{j=1}^J\). We are administering the survey to user \(i\); we seek the question \(j\) that maximizes information about \(u_i\).

We simplify the ordinal logit model to a single user:
\[
u_i \overset{iid}{\sim} N(\mu_U, \Lambda_U^{-1})
\]
\[
R_{ij} | U, V \overset{ind}{\sim} \text{OrdLogit}(u_i^T v_j, \beta)
\]

Unlike in the case of Gaussian likelihood, we do not have a conjugate, closed-form update for the posterior of \(u_i\), so we cannot minimize a measure of posterior variance directly. Instead, we work with the Fisher information, whose inverse is the asymptotic variance of the maximum likelihood estimate. This approach follows the optimal design literature, notably Segall [47], who applies it to logistic likelihood for binary responses. Our approach can be considered an ordinal logit generalized of [47].

Fisher information is computed around a value of \(u_i\). As our latest estimate of \(u_i\), we use the mean of the user variational distribution, \(\mu_i\), from the previous iteration of probabilistic matrix factorization on revealed responses. Repurposing this provisional estimate of \(u_i\) is more computationally efficient than computing a MAP estimate of \(u_i\) from the single-user model.

We denote the Fisher information gained from a response to question \(j\) as \(J^j(u_i; x)\), and the observed Fisher information from observing response \(x\) to question \(j\) as \(\bar{J}^j(u_i; x)\). Then
\[
J^j(u_i; x) = -\frac{\partial^2}{\partial u_i \partial u_i^T} \log \Pr(R_{ij} = x | u_i, v_j, \beta) = -\frac{\partial^2}{\partial u_i \partial u_i^T} \log \pi_{ijx}
\]
and
\[
\bar{J}(u_i) = E[J^j(u_i; R_{ij})] = \sum_{x=1}^{M+1} \pi_{ijx} J^j(u_i; x)
\]

In the ordinal logit model, \(J^j(u_i; x)\) and \(\bar{J}(u_i)\) involve complicated but closed-form expressions. The Hessians are computed with autodifferentiation in \texttt{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 find the question that minimizes our optimal design criterion, that is, which potential response would contribute the most information to \(u_i\) in expectation. Formally, this question is
\[
\min_{j \in \mathcal{U}} \text{tr} \left[ \Lambda_U + \sum_{\ell \in \mathcal{O}} J^\ell(u_i; R_{i\ell}) + \bar{J}(u_i) \right]^{-1}
\]
As before, each iteration of the simulation solves this optimization once per user, widening the survey by one question. The user-specific problems can be solved separately in parallel. This subprocedure is placed in context in Algorithm 2.
Algorithm 2: Active strategy simulation for ordinal logit model

given: Training responses $R^{train}$, cutpoints $\beta$

// estimate question factors, user prior
{$\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}$}$_{i,j} \leftarrow \text{ordinal_logit_matrix_factorization}(R^{train}, \beta)$
{$v_1, \ldots, v_k$} $\leftarrow$ ($v_1, \ldots, v_k$)
$\Lambda_U^{-1} \leftarrow \text{empirical_covariance}(\mu_1, \ldots, \mu_n)$

// simulation loop
$O_i \leftarrow \emptyset$, $U_i \leftarrow \{1, \ldots, k\} \forall i$
for $t \leftarrow 1$ to $k$
do
  // expand survey by one question
  for $i \leftarrow 1$ to $n$
do
    // choose next question for user $i$
    $j \leftarrow \arg\min_{j \in U} \tr \left[ \Lambda_U + \sum_{\ell \in O} J^t(u_i; R_{i\ell}) + I^t(u_i) \right]^{-1}$
    $R_{ij} \leftarrow \text{get_response}(i, j)$
  // update observed, unobserved question sets for user
  $O_i \leftarrow O_i \cup \{j\}$, $U_i \leftarrow U_i \setminus \{j\}$
  // update user factors and evaluate predictions
  {$\mu_{ij}, \sigma_{ij}, \nu_{ji}, \tau_{ji}$}$_{i,j} \leftarrow \text{ordinal_logit_matrix_factorization}(R, \beta)$
  {$u_1, \ldots, u_n$} $\leftarrow$ ($\mu_1, \ldots, \mu_n$)
  $R^{pred}_{ij} \leftarrow E \left[ \text{OrdLogit}(\mu_t, \nu_j, \beta) \right] \forall i,j$
compute_predictive_error($R, R^{pred}$)
do

6.4 Results

Simulations with the ordinal logit likelihood largely uphold prior BPMF results using the Gaussian likelihood, while showing improved modeling flexibility. In Figure F.1, the active strategy continues to outperform random in terms of predictive error. This comes at the cost of higher bias in the early stages of the survey. Matrix completion with the ordinal logit model attains lower final error levels than with BPMF, as a comparison with Figure 2 shows. The improvement comes from a combination of better modeling of ordinal responses and introducing additional parameters in the form of question-specific cutpoints. In addition, the advantage of the active strategy is more pronounced than before; see Figure F.2.

In Figure F.3 we plot question position across randomly sampled users in one simulation of the active strategy. The nondeterministic question order creates more variability in question rank within a single ordinal logit simulation than there is across BPMF simulations in Figure 5. Still, the active orderings for both models prioritize questions about abortion and environmental policies. While ACA repeal is the top question for BPMF, the ordinal logit model places it anywhere in the first 25 questions. Other questions with highly variable positions concern approval of Obama, the national economy, and background checks for gun sales, all of which appear with little variability after the 25th question in the BPMF ordering. The ordinal logit model favors job approval questions more, possibly having detected greater prior variance in this latent direction.

Figure F.4 plots the active ordering more granularly for this subset of users, as individual paths through survey questions. We see that the variability in question rank in Figure F.3 arises not from a few common question orderings, but rather from diverse, personalized paths dependent on responses to previous questions. Using the ordinal logit likelihood, we have made our active strategy truly adaptive.

7 Order effects

Algorithmic ordering of survey questions may produce biased responses (compared to random ordering) when responses vary according to their position in the survey instrument – a phenomenon known as order effects. In this section we estimate the magnitude of order effects in the randomly-ordered Facebook survey in order to understand how large the bias introduced by the active strategy

17
is likely to be.

In order to estimate position effects, we fit a 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. Using this model, we estimate the difference in 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 8(a). We find evidence for a number of survey questions exhibiting position effects, where responses vary significantly depending on whether they are 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 whether the previous survey question a user answered affects the response on the following question. 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. We visualize the results of this model in Figure 8(b). About 10% of the possible question pairs exhibit a non-zero interaction effect. Some questions tend to be influential on the following question (columns with multiple points) while others tend to be more likely to be affected by the prior question (rows with multiple points). Similarly to position effects, the effects we observe are usually less than 0.2 standard deviations on the response scale.

While these estimates are rudimentary, they provide some sense of the size of the bias introduced by the active learning algorithm – it should be a small contribution to the total error compared to the variance reduction we achieve through receiving more informative responses. A promising area for future work would be to test the active learning strategy online compared to a random ordering and to directly estimate the bias.

8 Discussion and future work

With active learning that seeks to reduce variance in the latent space of concepts, we have identified survey orderings that achieve the same error with a smaller question budget. Exploring the CCES questions preferred by the active strategy, we develop insights about the most informative set of questions for predicting political opinion. The active strategy affords a new notion of feature importance for domains with wide item sets and low-dimensional latent structure. Matrix factorization gives us dimension reduction; active question order makes this dimension reduction interpretable.

However, the deterministic active question order arising from BPMF is not ideal; it is non-robust
to estimation error of question factors. With the ordinal logit model, we have made the active ordering dependent on collected responses. This is a side effect of changing the likelihood from Gaussian to ordinal logit, but our active strategy still assumes away uncertainty in question factors. Our variational inference for matrix factorization already quantifies uncertainty in question factors approximately. Future work should incorporate this uncertainty in the variance being minimized. The active strategy could take other forms of uncertainty into account. For instance, it is natural to consider concepts in latent space as changing over time. By explicitly modeling parameter drift, we could encourage the active ordering to evolve. Another way of inducing different question orderings is, broadly, better managing the exploration-exploitation tradeoff. Currently the active strategy always minimizes the optimality criterion. We could experiment with bandit algorithms that make use of estimated uncertainty in question factors.

Future work should devote special attention to the logistics of survey administration. The difficulty of implementing an active ordering depends on the variant used. The active question order from BPMF is relatively straightforward to use across survey modes, as it can be computed offline. An adaptive order computed on the fly using previous responses, as with the ordinal logit model, would require more infrastructure. Web and computer-assisted telephone surveys could lean on backend software to suggest the next question. In-person field surveys would need a mechanism of 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 would compel us to move beyond greedy question selection to optimal design with a multi-step horizon. It is also an opportunity to design logical groupings of questions, informed by domain knowledge of practitioners.

Our evaluation of the active strategy focuses on measures of predictive error. This is motivated by the wide range of downstream uses of survey data. Survey researchers may be interested in the distribution of responses to individual questions; the relationship between one question and another across individuals; the reduction of multiple questions to a single construct; or any of these analyses conditional on arbitrary subgroups of respondents. Our predictive error measures, which assign equal weight to the loss from each user-question pair, allow us to evaluate our imputations from various strategies agnostic to the downstream use case. In some cases, individual responses to a wide set of questions are equally useful. For instance, using the CCES, researchers have investigated whether alignment between the policy preferences of constituents and the votes of their legislators predicts constituent support for their legislators [2]. This research relies on knowing individual responses to “roll call” questions about policy preferences – the same questions we have included from the CCES. As another example, in survey experiments, responses from a baseline survey could serve as pre-treatment covariates for estimation of heterogeneous treatment effects [7]. A priori, the analyst cares equally about the predictive accuracy of imputed responses across respondents and questions. Still, future work should assess the frequency properties of common estimators applied to imputed responses, and explore how to specialize the active strategy to different estimators.

We have incorporated side information in two simple ways, but many avenues exist for more sophisticated modeling of side information. For instance, we could create individual user priors based on more covariates using multilevel regression. Alternatively, we could learn a Gaussian process that maps covariates into latent space, as a prior for user factors [1, 57]. A different tack is to add terms for user and question covariates to the response specification; they slot naturally into the nuclear norm minimization in [4] and the BPMF model in [38]. From a cost perspective, we could explore the tradeoff between gathering side information and survey responses, incorporating the information gain and acquisition cost of both. Our experiments including responses to covariate questions in the matrix factorization suggest an exchange rate between side information and question responses. Finally, it remains to address the bias of our matrix completion procedure. Standard matrix completion assumes entries are missing at random. This assumption is violated when users withhold responses that are noncommittal or unpopular. By tailoring questions to a user’s inferred position in latent space, the active strategy can exacerbate bias. One bias correction method for matrix completion is explicitly modeling the missing-data mechanism [32]. More common are weighting approaches that regularize a weighted nuclear norm [51] or de-bias the loss terms with inverse propensity weights [45, 4]. All of these approaches increase the importance of correctly predicting rare users and items. Another source of bias is model misspecification: a low-rank linear decomposition does not capture all of the response variance. Thus, predictive error plateaus after
a certain survey length. It could be worth exploring nonlinear matrix factorization in the form of Gaussian process latent variable models, which replace the dot product of user and question factors with a function of user factors, endowed with a Gaussian process prior [29].

9 Acknowledgements

We thank Karan Aggarwal, Eytan Bakshy, George Berry, Dennis Feehan, Avi Feller, Will Fithian, Mike Jordan,Frauke Kreuter, Jacob Montgomery, Adam Obeng, Rebecca Powell and Alex Theodoridis for their valuable input. CZ is supported by an NSF Graduate Research Fellowship. CZ and JS are supported by Office of Naval Research (ONR) Grant N00014-15-1-2367.

References

[1] Adams, R. P., Dahl, G. E., & Murray, I. (2010). Incorporating side information in probabilistic matrix factorization with gaussian processes. arXiv preprint arXiv:1003.4944.
[2] Ansolabehere, S., & Rivers, D. (2013). Cooperative survey research. Annual Review of Political Science, 16, 307-329.
[3] Ansolabehere, S., & Schaffner, B. (2016). CCES Common Content, 2012-CCES Dataverse. Dataverse. Harvard. Edu.
[4] Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2018). Matrix completion methods for causal panel data models (No. w25132). National Bureau of Economic Research.
[5] Attenberg, J., & Provost, F. (2011). Inactive learning?: difficulties employing active learning in practice. ACM SIGKDD Explorations Newsletter, 12(2), 36-41.
[6] Balcan, M. F., Hanneke, S., & Vaughan, J. W. (2010). The true sample complexity of active learning. Machine learning, 80(2-3), 111-139.
[7] Broockman, D. E., Kalla, J. L., & Sekhon, J. S. (2017). The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. Political Analysis, 25(4), 435-464.
[8] Cao, Y., & Xie, Y. (2015, December). Categorical matrix completion. In Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on (pp. 369-372). IEEE.
[9] Chakraborty, S., Zhou, J., Balasubramanian, V., Panchanathan, S., Davidson, I., & Ye, J. (2013, December). Active matrix completion. In Data Mining (ICDM), 2013 IEEE 13th International Conference on (pp. 81-90). IEEE.
[10] Chiang, K. Y., Hsieh, C. J., & Dhillon, I. S. (2015). Matrix completion with noisy side information. In Advances in Neural Information Processing Systems (pp. 3447-3455).
[11] Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. Journal of artificial intelligence research, 4(1), 129-145.
[12] Dasgupta, S., Kalai, A. T., & Monteleoni, C. (2005, June). Analysis of perceptron-based active learning. In International Conference on Computational Learning Theory (pp. 249-263). Springer, Berlin, Heidelberg.
[13] Early, K., Maukoff, J., & Fienberg, S. E. (2017). Dynamic question ordering in online surveys. Journal of Official Statistics, 33(3), 625-657.
[14] Elahi, M., Ricci, F., & Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. Computer Science Review, 20, 29-50.
[15] Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. Computers in human behavior, 26(2), 132-139.
[16] Fithian, W., & Mazumder, R. (2013). Flexible Low-Rank Statistical Modeling with Side Information. arXiv preprint arXiv:1308.4211.

[17] Gabry, J., & Goodrich, B. (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version, 2(1).

[18] Garnett, R., Krishnamurthy, Y., Xiong, X., Schneider, J., & Mann, R. (2012). Bayesian optimal active search and surveying. arXiv preprint arXiv:1206.6406.

[19] Golbandi, N., Koren, Y., & Lempel, R. (2010, October). On bootstrapping recommender systems. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1805-1808). ACM.

[20] Gonzalez, J. M., & Eltinge, J. L. (2008, August). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In JSM proceedings.

[21] Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly, 75(5), 861-871.

[22] Heberlein, T. A., & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. American sociological review, 447-462.

[23] Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. Public opinion quarterly, 45(4), 549-559.

[24] Kallus, N., Mao, X., & Udell, M. (2018). Causal Inference with Noisy and Missing Covariates via Matrix Factorization. arXiv preprint arXiv:1806.00811.

[25] Karimi, R., Freudenthaler, C., Nanopoulos, A., & Schmidt-Thieme, L. (2011, August). Non-myopic active learning for recommender systems based on matrix factorization. In Information Reuse and Integration (IRI), 2011 IEEE International Conference on (pp. 299-303). IEEE.

[26] Karimi, R., Freudenthaler, C., Nanopoulos, A., & Schmidt-Thieme, L. (2011, November). Towards optimal active learning for matrix factorization in recommender systems. In Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on (pp. 1069-1076). IEEE.

[27] Klopp, O., Lafond, J., Moulines, É., & Salmon, J. (2015). Adaptive multinomial matrix completion. Electronic Journal of Statistics, 9(2), 2950-2975.

[28] Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied cognitive psychology, 5(3), 213-236.

[29] Lawrence, N. D., & Urtasun, R. (2009, June). Non-linear matrix factorization with Gaussian processes. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 601-608). ACM.

[30] Lim, Y. J., & Teh, Y. W. (2007, August). Variational Bayesian approach to movie rating prediction. In Proceedings of KDD cup and workshop (Vol. 7, pp. 15-21).

[31] MacKay, D. J. (1992). Information-based objective functions for active data selection. Neural computation, 4(4), 590-604.

[32] Marlin, B. M., & Zemel, R. S. (2009, October). Collaborative prediction and ranking with non-random missing data. In Proceedings of the third ACM conference on Recommender systems (pp. 5-12). ACM.

[33] Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of machine learning research, 11(Aug), 2287-2322.

[34] Montgomery, J. M., & Cutler, J. (2016). Computerized adaptive testing for public opinion surveys. Political Analysis, 21(2), 172-192.

[35] Mulder, J., & Van Der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. Psychometrika, 74(2), 273.
[36] Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with Kullback–Leibler information item selection. In Elements of adaptive testing (pp. 77-101). Springer, New York, NY.

[37] Negahban, S., & Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. Journal of Machine Learning Research, 13(May), 1665-1697.

[38] Porteous, I., Asuncion, A. U., & Welling, M. (2010, July). Bayesian Matrix Factorization with Side Information and Dirichlet Process Mixtures. In AAAI.

[39] Recht, B. (2011). A simpler approach to matrix completion. Journal of Machine Learning Research, 12(Dec), 3413-3430.

[40] Rennie, J. D., & Srebro, N. (2005, August). Fast maximum margin matrix factorization for collaborative prediction. In Proceedings of the 22nd international conference on Machine learning (pp. 713-719). ACM.

[41] Rubens, N., & Sugiyama, M. (2007, October). Influence-based collaborative active learning. In Proceedings of the 2007 ACM conference on Recommender systems (pp. 145-148). ACM.

[42] Rubens, N., Elahi, M., Sugiyama, M., & Kaplan, D. (2015). Active learning in recommender systems. In Recommender systems handbook (pp. 809-846). Springer, Boston, MA.

[43] Salakhutdinov, R., & Mnih, A. (2008, July). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In Proceedings of the 25th international conference on Machine learning (pp. 880-887). ACM.

[44] Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. Springer Science & Business Media.

[45] Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. arXiv preprint arXiv:1602.05352.

[46] Seeger, M., & Bouchard, G. (2012, March). Fast variational Bayesian inference for non-conjugate matrix factorization models. In Artificial Intelligence and Statistics (pp. 1012-1018).

[47] Segall, D. O. (2009). Principles of multidimensional adaptive testing. In Elements of adaptive testing (pp. 57-75). Springer, New York, NY.

[48] Settles, B. (2010). Active learning literature survey. University of Wisconsin, Madison, 52(55-66), 11.

[49] Silva, J., & Carin, L. (2012, August). Active learning for online bayesian matrix factorization. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 325-333). ACM.

[50] Srebro, N., Rennie, J., & Jaakkola, T. S. (2005). Maximum-margin matrix factorization. In Advances in neural information processing systems (pp. 1329-1336).

[51] Srebro, N., & Salakhutdinov, R. R. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In Advances in Neural Information Processing Systems (pp. 2056-2064).

[52] Sutherland, D. J., Póczos, B., & Schneider, J. (2013, August). Active learning and search on low-rank matrices. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 212-220). ACM.

[53] Tourangeau, R., Kreuter, F., & Eckman, S. (2015). Motivated misreporting: Shaping answers to reduce survey burden. Survey measurements. techniques, data quality and sources of error, 24-41.

[54] Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., & Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. arXiv preprint arXiv:1610.09787.
[55] Wang, H., Zhu, R., & Ma, P. (2018). Optimal subsampling for large sample logistic regression. Journal of the American Statistical Association, 1-16.

[56] Xu, M., Jin, R., & Zhou, Z. H. (2013). Speedup matrix completion with side information: Application to multi-label learning. In Advances in neural information processing systems (pp. 2301-2309).

[57] Zhou, T., Shan, H., Banerjee, A., & Sapiro, G. (2012, April). Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In Proceedings of the 2012 SIAM international Conference on Data mining (pp. 403-414). Society for Industrial and Applied Mathematics.
A Proofs

A.1 Relationship between BPMF posterior mean and MAP estimation

The complete conditional for user factors is

\[ u_i \mid R, V, \mu_U, \Lambda_U, \alpha \sim \mathcal{N}(\mu^*_i, [\Lambda^*_i]^{-1}) \]

\[ \Lambda^*_i = \Lambda_U + \alpha \sum_{j=1}^k I_{ij} v_j v_j^T \]

\[ \mu^*_i = [\Lambda^*_i]^{-1} \left( \alpha \sum_{j=1}^k I_{ij} R_{ij} v_j + \Lambda_U \mu_U \right) \]

We show this expression for \( \mu^*_i \) also arises in MAP estimation for BPMF, as the coordinate ascent update for \( u_i \).

We reproduce the MAP objective function, Eq (4) in [43]. We use prior hyperparameters \( \mu_U = \mu_V = 0, \Lambda_U = \alpha_U I, \Lambda_V = \alpha_V I \) and introduce \( \lambda_U = \alpha_U / \alpha, \lambda_V = \alpha_V / \alpha \). For these settings, the objective function is simply the Frobenius norm regularized problem.

\[ \ell(U, V, R) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^k I_{ij} (R_{ij} - u_i^T v_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^n \| u_i \|^2 + \frac{\lambda_V}{2} \sum_{j=1}^k \| v_j \|^2 \]

Taking the derivative with respect to \( u_i \) and setting to 0 yields

\[ \frac{\partial}{\partial u_i} \ell(U, V, R) = - \sum_{j=1}^k I_{ij} (R_{ij} - u_i^T v_j) v_j^T + \lambda_U u_i^T = 0 \]

\[ \sum_{j=1}^k I_{ij} R_{ij} v_j = \left( \sum_{j=1}^k I_{ij} v_j v_j^T + \lambda_U I \right) u_i \]

The coordinate ascent update is

\[ u^*_i = \left( \sum_{j=1}^k I_{ij} v_j v_j^T + \lambda_U I \right)^{-1} \left( \sum_{j=1}^k I_{ij} R_{ij} v_j \right) \]

\[ = \left( \alpha \sum_{j=1}^k I_{ij} v_j v_j^T + \alpha_U I \right)^{-1} \left( \alpha \sum_{j=1}^k I_{ij} R_{ij} v_j \right) \]

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

\[ = \mu^*_i \]

A.2 Relationship between trace minimization and predictive variance

Let \( \tilde{v} \) be drawn from \( \{ v : \| v \|_2 = 1 \} \), the uniform distribution on the unit sphere, independently of \( u_i \). We show the predictive variance is

\[ \text{Var} (u_i^T \tilde{v}) = \frac{1}{\nu} E \| u_i \|^2 = \frac{1}{\nu} \left( \text{tr} \Sigma + \| \mu \|^2 \right) \]
Proof:

\[
\text{Var}(u_i^T \tilde{v}) = E(u_i^T \tilde{v} \tilde{v}^T u_i) - (E(u_i^T \tilde{v}))^2
\]

\[
= E\left( u_i^T E(\tilde{v} \tilde{v}^T | u_i) u_i \right) - (E(u_i) E(\tilde{v}))^2
\]

\[
= E\left( u_i^T \frac{1}{r} \mathbf{I} u_i \right) - (\mu^T \mathbf{0})^2
\]

\[
= \frac{1}{r} E \|u_i\|^2_2
\]

\[
= \frac{1}{r} E \left[ \text{tr} \left( u_i u_i^T \right) \right]
\]

\[
= \frac{1}{r} \text{tr} \left( \Sigma + \mu \mu^T \right)
\]

\[
= \frac{1}{r} \left( \text{tr} \Sigma + \|\mu\|^2_2 \right)
\]

B Reweighting the evaluation loss function

B.1 Rationale

We developed the A-optimality method assuming \( \tilde{v} \) is uniformly distributed in latent space, drawn uniformly from either the elementary basis vectors or the unit sphere. However, the empirical \( \{v_j\}_{k=1}^k \) given by matrix factorization will not be distributed uniformly in general. We want to evaluate losses such as MSE and MAE on the entire instrument, yet a simple average over all questions will give more weight to high-density regions of question space. We investigate whether our results change if we assign equal importance to all latent concepts.

Conceptually, we are computing a loss (say MSE) with respect to a distribution of question factors, and we want to adjust the empirical distribution to look more uniform. Given the per-question MSE, we want a weighted average that approximates the loss we would have computed if the question factors were distributed uniformly. That is, we find \( w \in \mathbb{R}^k \) such that

\[
E_{\tilde{p}}[w_j M\text{SE}_j] \approx E_p[M\text{SE}_j]
\]

To simplify the analysis, we assume that conditional on \( v_j \), our estimates of \( u_i \) are unbiased and \( E u_i^T v_j = 0 \). Recalling that for fixed \( \tilde{v} \),

\[
\text{Var}(u_i^T \tilde{v}) = \tilde{v}^T \Sigma \tilde{v} - (E u_i^T \tilde{v})^2
\]

our goal is to find \( w \in \mathbb{R}^k \) such that

\[
E_{\tilde{p}}[w_j v_j^T \Sigma v_j] \approx E_p[v_j^T \Sigma v_j]
\]

\[
E_{\tilde{p}}[w_j \text{tr} (\Sigma v_j v_j^T)] \approx E_p[\text{tr} (\Sigma v_j v_j^T)]
\]

\[
\text{tr} \left[ \Sigma \left( \frac{1}{k} \sum_{j=1}^k w_j v_j v_j^T \right) \right] \approx \frac{1}{r} \text{tr} \Sigma
\]

We choose \( \{w_j\}_{j=1}^k \) such that

\[
\frac{1}{k} \sum_{j=1}^k w_j v_j v_j^T \approx \frac{1}{r} \mathbf{I}_r
\]

This makes intuitive sense, as we want to reweight the empirical covariance matrix to approximate the identity.

We cannot in general find an exact solution \( \{w_j\}_{j=1}^k \), especially if the empirical \( \{v_j\}_{j=1}^k \) do not span latent space. However, with a sufficiently diverse set of question vectors, we can get close by solving the following convex problem:

\[
\min_{w_1, \ldots, w_k} \left\| \frac{1}{k} \sum_{j=1}^k w_j v_j v_j^T - \frac{1}{r} \mathbf{I}_r \right\|_F
\]

s.t. \( w_j \geq 0 \; \forall j \) and \( \sum_{j=1}^k w_j = 1 \)
B.2 Reweighted results

Empirical question factors are rescaled to unit norm and provided to the above optimization. The reweighted error metrics, shown in Figure B.1, barely change at all, reinforcing our prior conclusions.

Figure B.1: Error on the holdout set for the 2016 CCES, a weighted average over all questions. Per-question weights are computed via an optimization that seeks to distribute questions more uniformly in latent space.
C Additional results for 2016 CCES

Figure C.1: We plot the trace of posterior variance achieved by 10 simulations of the active and random strategies on 2016 CCES. Note the active strategy produces a single question ordering across all users, and thus a single sequence of objective function values per simulation. This sequence varies slightly across simulations due to randomness in estimated question factors. To compute the objective function for the random strategy in one simulation, we average the trace of posterior variance after each question across 100 random question orderings.
| Question | PC1  | PC2  | PC3  | PC4  | Question text |
|----------|------|------|------|------|---------------|
| CC16-302 | -0.06 | 0.11 | -0.11 | 0.02 | National Economy |
| CC16-303 | -0.05 | 0.04 | -0.08 | 0.00 | Past year - household income |
| CC16-304 | -0.05 | 0.08 | -0.13 | 0.00 | Next year - household income |
| CC16-307 | -0.09 | -0.22 | -0.08 | 0.07 | Police make R feel safe |
| CC16-320a | -0.16 | 0.09 | -0.13 | 0.02 | Approve of Job - Obama |
| CC16-320b | 0.06 | 0.05 | -0.23 | 0.04 | Approve of Job - Congress |
| CC16-320c | -0.04 | 0.04 | -0.25 | 0.11 | Approve of Job - Supreme Court |
| CC16-320d | -0.01 | -0.03 | -0.31 | 0.12 | Approve of Job - Governor |
| CC16-320e | 0.04 | 0.03 | -0.33 | 0.13 | Approve of Job - Legislature |
| CC16-320f | 0.05 | 0.05 | -0.34 | 0.18 | Approve of Job - Rep |
| CC16-320g | -0.01 | 0.07 | -0.33 | 0.18 | Approve of Job - Senator 1 |
| CC16-320h | 0.03 | 0.05 | -0.34 | 0.18 | Approve of Job - Senator 2 |
| CC16-330b | -0.22 | 0.04 | -0.00 | -0.11 | Gun Control - Ban assault rifles |
| CC16-330e | 0.19 | -0.09 | -0.09 | 0.05 | Gun Control - Make it easier for people to obtain concealed-carry permit |
| CC16-331-1 | -0.17 | 0.13 | -0.15 | -0.07 | Immigration - Grant legal status to all illegal immigrants who have held jobs an |
| CC16-331-2 | 0.11 | -0.19 | -0.02 | 0.12 | Immigration - Increase the number of border patrols on the U.S.-Mexican border |
| CC16-331-3 | -0.11 | 0.17 | -0.19 | 0.03 | Immigration - Grant legal status to people who were brought to the US illegally |
| CC16-331-7 | 0.17 | -0.15 | 0.06 | 0.09 | Immigration - Identify and deport illegal immigrants |
| CC16-331-9 | 0.21 | 0.28 | 0.03 | -0.11 | Immigration - None of these |
| CC16-332a | -0.20 | 0.12 | 0.09 | 0.25 | Abortion Policies - Always allow a woman to obtain an abortion as a matter of ch |
| CC16-332c | 0.05 | -0.27 | -0.07 | -0.22 | Abortion Policies - Prohibit all abortions after the 20th week of pregnancy |
| CC16-332d | 0.18 | -0.17 | -0.14 | -0.20 | Abortion Policies - Allow employers to decline coverage of abortions in insuranc |
| CC16-332e | 0.17 | -0.19 | -0.21 | -0.23 | Abortion Policies - Prohibit the expenditure of funds authorized or appropriated |
| CC16-332f | 0.21 | 0.17 | -0.11 | -0.20 | Abortion Policies - Make abortions illegal in all circumstances |
| CC16-333a | -0.25 | 0.05 | -0.09 | -0.26 | Environment Policies - Give Environmental Protection Agency power to regulate Ca |
| CC16-333b | -0.22 | 0.02 | -0.02 | -0.19 | Environment Policies - Raise required fuel efficiency for the average automobile |
| CC16-333d | -0.20 | 0.13 | -0.07 | -0.26 | Environment Policies - Strengthen enforcement of the Clean Air Act and Clean Wat |
| CC16-334a | -0.17 | 0.01 | -0.03 | 0.14 | Crime Policies - Eliminate mandatory minimum sentences for non-violent drug off |
| CC16-334b | -0.22 | -0.20 | 0.04 | 0.02 | Crime Policies - Require police officers to wear body cameras that record all of |
| CC16-334d | -0.13 | -0.32 | -0.01 | -0.03 | Crime Policies - Increase prison sentences for felons who have already committed |
| CC16-335 | -0.21 | 0.06 | 0.11 | 0.19 | Gay Marriage |
| CC16-351B | -0.10 | 0.04 | -0.05 | -0.26 | For or Against - Congress - Trans-Pacific Partnership Act |
| CC16-351E | -0.15 | 0.20 | -0.01 | 0.06 | For or Against - Congress - Education Reform |
| CC16-351F | -0.21 | 0.18 | -0.09 | 0.09 | For or Against - Congress - Highway and Transportation Funding Act |
| CC16-351G | -0.12 | -0.29 | 0.00 | 0.10 | For or Against - Congress - Iran Sanctions Act |
| CC16-351H | -0.14 | 0.10 | -0.08 | 0.08 | For or Against - Congress - Medicare Accountability and Cost Reform Act |
| CC16-351I | 0.14 | -0.27 | 0.01 | 0.02 | For or Against - Congress - Repeal Affordable Care Act |
| CC16-351K | -0.23 | 0.03 | -0.00 | -0.13 | For or Against - Congress - Minimum wage |

Table C.1: Question factors for 2016 CCES.
C.1 Tuning simulation parameters

Figure C.2: Error on the holdout set for the 2016 CCES, averaged over all questions. The active strategy minimizes the determinant of posterior variance (D-optimality).
Question rank across simulations, rank=4, alpha=1, all users

Figure C.3: Active ordering for 2016 CCES using a rank-4 matrix decomposition and D-optimality.
Figure C.4: Error on the holdout set for the 2016 CCES, averaged over all questions. The active strategy minimizes the maximum eigenvalue of posterior variance (E-optimality).
Figure C.5: Active ordering for 2016 CCES using a rank-4 matrix decomposition and E-optimality.
(a) Positions of the fixed question factors in latent space, for different simulations.

(b) Rank distribution for each question across 10 simulations of the active strategy.

Figure C.6: Active ordering for 2016 CCES using a rank-2 matrix decomposition, this time using Frobenius norm regularization instead of SoftImpute to estimate question factors from training responses.
(a) Positions of the fixed question factors in latent space.

(b) Rank distribution for each question across 10 simulations of the active strategy.

Figure C.7: Active ordering for 2016 CCES using a rank-2 matrix decomposition, this time setting $\alpha = 4$. This increases the precision and hence the information content of responses.
C.2 Robustness checks for question order

Figure C.8: Question rank across 10 simulations of the active strategy on the 2016 CCES. This iteration adds questions about political affiliation. Of these questions, the most informative is whether the respondent identifies as a Democrat; this appears sixth in the active order. Identifying as a Republican or independent and rating one’s ideology on an ordinal scale do not appear in the top 20.

For or Against - Congress - Repeal Affordable Care Act
Environment Policies - Strengthen enforcement of the Clean Air Act and Clean Water Act
Abortion Policies - Prohibit the expenditure of funds authorized or appropriated
Environment Policies - Give Environmental Protection Agency power to regulate Co
Abortion Policies - Allow employers to decline coverage of abortions in insurance
3 point party ID - Democrat
Approve of Job - Senator 1
Environment Policies - Require a minimum amount of renewable fuels (wind, solar,
Abortion Policies - Always allow a woman to obtain an abortion as a matter of ch
Approve of Job - Rep
Gay Marriages
Approve of Job - Senator 2
Immigration - Identify and deport illegal immigrants
For or Against - Congress - Minimum wages
Crime Policies - Increase the number of police on the street by 10 percent, even
Abortion Policies - Prohibit all abortions after the 20th week of pregnancy
Gun Control - ban assault rifles
Approve of Job - Legislature
Abortion Policies - Permit abortion only in case of rape, incest or when the woman’s life
Immigration - Grant legal status to all illegal immigrants who have held jobs an
Approve of Job - Governor
Environment Policies - Raise required fuel efficiency for the average automobile
Immigration - None of these
Approve of Job - Obama
Crime Policies - Increase prison sentences for felons who have already committed
3 point party ID - Republican
Immigration - Grant legal status to people who were brought to the US illegally
Gun Control - Make it easier for people to obtain concealed-carry permit
Immigration - Increase the number of Border Patrol on the U.S.-Mexican border
Abortion Policies - Make abortions illegal in all circumstances
7 point Party ID
Gun Control - Background checks for all sales, including at gun shows and even t
Approve of Job - Supreme Court
For or Against - Congress - Trans-Pacific Partnership Act
For or Against - Congress - Highway and Transportation Funding Act
For or Against - Congress - Iran Sanctions Act
Crime Policies - Eliminate mandatory minimum sentences for non-violent drug of
Approve of Job - Congress
Crime Policies - Require police officers to wear body cameras that record all of
Gun Control - Prohibit state and local governments from publishing the names and
3 point party ID - Independent
Police make it feel safe
Ideology
National Economy
Next year - household income
For or Against - Congress - Education Reform
For or Against - Congress - Medicare Accountability and Cost Reform Act
Past year - household income

Figure C.8: Question rank across 10 simulations of the active strategy on the 2016 CCES. This iteration adds questions about political affiliation. Of these questions, the most informative is whether the respondent identifies as a Democrat; this appears sixth in the active order. Identifying as a Republican or independent and rating one’s ideology on an ordinal scale do not appear in the top 20.
Figure C.9: Question rank across 10 simulations of the active strategy on the 2016 CCES. This iteration adds questions about political affiliation and demographics. Gender questions are asked early, and identification as a Democrat moves into fourth place. Questions about race do not appear in the top 25.
Figure C.10: Question rank across 10 simulations of the active strategy on the 2016 CCES. This iteration adds questions about political affiliation, demographics, education, financial well-being, and religiosity. Of the new questions, child and home ownership are prioritized, but only after gender and Democrat self-identification.

- For or Against - Congress - Repeal Affordable Care Act
- Environment Policies - Gave Environmental Protections Agency power to regulate CO
- Abortion Policies - Prohibit the expenditure of funds authorized or appropriated
- Gender - Female
- 5 point party ID - Democrat
- Abortion Policies - Allow employers to decline coverage of abortions in insurance
- Exempt - Mass
- Environment Policies - Strengthen enforcement of the Clean Air Act and Clean Water
- Abortion Policies - Always allow a woman to obtain an abortion as a matter of choice
- Child under 18 years - No
- Environment Policies - Require a minimum amount of renewable fuels (wind, solar,
- Abortion Policies - Prohibit all abortions after the 20th week of pregnancy
- Home ownership - Own
- Immigration - Grant legal status to all illegal immigrants who have held jobs an
- Child under 18 years - Yes
- For or Against - Congress - Minimum wage
- Immigration - Identify and deport illegal immigrants
- Home ownership - Rent
- Gay Marriage
- Approve of Job - Obama
- Approve of Job - Senator D
- Gun Control - Ban assault rifles
- Approve of Job - Rep
- 7 point Party ID
- Abortion Policies - Permit abortion only in case of rape, incest or when the wom
- Gun Control - Make it easier for people to obtain concealed-carry permit
- For or Against - Congress - Trans-Pacific Partnership Act
- Approve of Job - Senator E
- Immigration - Grant legal status to people who were brought to the US illegally
- 5 point party ID - Republican
- Crime Policies - Increase the number of police on the street by 10 percent, even
- Approve of Job - Legislators
- Immigration - Increase the number of border patrols on the U.S. Mexican border
- Political Interest
- Importance of religion (show version)
- Environment Policies - Raise required fuel efficiency for the average automobile
- Race - White
- Crime Policies - Increase prison sentences for felons who have already committed
- Approve of Job - Governor
- Gun Control - Background checks for all sales, including at gun shows and over 1
- Race - Black
- Abortion Policies - Make abortions illegal in all circumstances
- Approve of Job - Supreme Court
- Immigration - Name of those
- 3 point party ID - Independent
- Race - Mixed
- Crime Policies - Eliminate mandatory minimum sentences for non-violent drug of
- Race - Asian
- Gun Control - Prohibit state and local governments from publishing the names and
- National Economy
- Race - Hispanic
- Ideology
- Education - Post-grad
- Crime Policies - Require police officers to wear body cameras that record all of
- Approve of Job - Congress
- For or Against - Congress - Highway and Transportation Funding Act
- Next year - household income
- Education - High school Graduate
- For or Against - Congress - Iran Sanctions Act
- Police make it feel safe
- For or Against - Congress - Medicare Accountability and Cost Reform Act
- Education - 2-year
- Education - 4-year
- For or Against - Congress - Education Reform
- Past year - household income
- Education - some college

Question rank across simulations, rank=4, alpha=1, all users
Figure C.11: Question rank across 10 simulations of the active strategy on the 2016 CCES. All covariate questions are automatically included at the start of the survey; the active ordering takes information from those questions into account.
D Results for 2012 CCES

Figure D.1: Error on the holdout set for the 2012 CCES, averaged over all questions.
Figure D.2: Question complexity of the active strategy relative to random-order questions for the 2012 CCES.
Figure D.3: Active ordering for 2012 CCES using a rank-2 matrix decomposition.
Figure D.4: Active ordering for 2012 CCES using a rank-4 matrix decomposition.
| Question   | PC1  | PC2  | PC3  | PC4  | Question text                                                      |
|------------|------|------|------|------|-------------------------------------------------------------------|
| CC302      | -0.14| 0.13 | 0.03 | 0.12 | National Economy                                                  |
| CC305      | -0.23| -0.06| 0.27 | -0.01| All things considered do you think it was a mistake to invade Iraq? |
| CC306      | -0.09| 0.04 | 0.40 | -0.08| All things considered do you think it was a mistake to invade Afghanistan? |
| CC308a     | -0.24| 0.17 | 0.07 | 0.10 | Institution Approval - Obama                                      |
| CC308b     | 0.06 | 0.35 | 0.05 | 0.05 | Institution Approval - Congress                                   |
| CC308c     | -0.02| 0.19 | -0.02| 0.10 | Institution Approval - Supreme Court                              |
| CC308d     | 0.04 | 0.07 | -0.09| 0.02 | Institution Approval - Governor                                   |
| CC308e     | 0.05 | 0.13 | -0.06| 0.00 | Institution Approval - Legislature                                |
| CC321      | -0.18| -0.11| 0.01 | 0.03 | Climate                                                           |
| CC322-1    | -0.22| 0.19 | -0.16| 0.02 | Grant legal status to all illegal immigrants who have held jobs and paid taxes f |
| CC322-2    | 0.19 | -0.26| 0.14 | 0.24 | Increase the number of border patrols on the US-Mexican border.    |
| CC322-3    | 0.26 | -0.04| 0.23 | 0.10 | Allow police to question anyone they think may be in the country illegally. |
| CC322-4    | 0.13 | -0.35| 0.22 | 0.35 | Fine US businesses that hire illegal immigrants.                   |
| CC322-5    | 0.21 | 0.07 | 0.37 | 0.23 | Prohibit illegal immigrants from using emergency hospital care and public school |
| CC322-6    | 0.25 | 0.01 | 0.28 | 0.16 | Deny automatic citizenship to American-born children of illegal immigrants. |
| CC324      | 0.18 | 0.18 | -0.02| -0.15| Abortion                                                           |
| CC325      | -0.12| 0.06 | 0.07 | 0.08 | Jobs-Environment                                                  |
| CC326      | -0.25| 0.03 | 0.05 | 0.33 | Gay Marriage                                                      |
| CC327      | -0.15| 0.22 | 0.03 | 0.03 | Affirmative Action                                                |
| CC332A     | 0.17 | 0.38 | -0.01| 0.20 | Roll Call Votes - Ryan Budget Bill                                |
| CC332B     | 0.01 | 0.04 | -0.26| 0.57 | Roll Call Votes - Simpson-Bowles Budget Plan                      |
| CC332C     | -0.03| -0.19| -0.16| 0.04 | Roll Call Votes - Middle Class Tax Cut Act                        |
| CC332D     | 0.22 | 0.27 | -0.01| 0.03 | Roll Call Votes - Tax Hike Prevention Act                         |
| CC332E     | 0.28 | 0.09 | -0.13| -0.07| Roll Call Votes - Birth Control Exemption                         |
| CC332F     | 0.02 | 0.03 | -0.37| 0.32 | Roll Call Votes - U.S.-Korea Free Trade Agreement                 |
| CC332G     | 0.28 | -0.01| -0.19| 0.07 | Roll Call Votes - Repeal Affordable Care Ac                       |
| CC332H     | 0.09 | -0.36| -0.28| -0.07| Roll Call Votes - Keystone Pipeline                               |
| CC332I     | -0.29| 0.03 | 0.09 | 0.07 | Roll Call Votes - Affordable Care Act of 2010                    |

Table D.1: Question factors for 2012 CCES.
E Results for Facebook survey

Figure E.1: Error on the holdout set for the Facebook survey, averaged over all questions.
Figure E.2: Question complexity of the active strategy relative to random-order questions for the Facebook survey.

Figure E.3: Error on the holdout set for the Facebook survey, averaged over all questions. The “-subgroups” strategies initialize each user prior to the empirical Bayes estimate for the subgroup to which the user belongs.
Results for ordinal logit model

Figure F.1: Predictive error of the ordinal logit model with active and random question orderings on the holdout set for the 2016 CCES, averaged over all questions. We show mean error with two standard deviations across 6 (12) simulations of the active (random) strategy. The proportion of predictions with the wrong sign is now zero, as responses are now nonnegative ordinal values. That is, response preprocessing no longer rescales responses to $[-1, 1]$; instead we shift responses to begin at 0.
Figure F.2: Relative sample complexity of the active and random strategies for the 2016 CCES with ordinal logit model. The active strategy requires 10-15 questions to reach the same error as the random strategy with 30 questions.
Figure F.3: Active ordering for 2016 CCES with ordinal logit model using a rank-4 matrix decomposition. We show the rank of each question across 100 randomly sampled users in one simulation.
Figure F.4: Paths through questions on the 2016 CCES, chosen by the active strategy with ordinal logit model using a rank-4 matrix decomposition. We show paths for 100 randomly sampled users in one simulation as light gray lines. Darker points indicate questions appearing more frequently at a given position in the survey.