REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
05-08-2021	Final Report	30-Dec-2017 - 29-Dec-2020

4. TITLE AND SUBTITLE
Final Report: Reliability and robustness for fast Bayesian inference of complex data

5a. CONTRACT NUMBER
W911NF-18-1-0063

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Massachusetts Institute of Technology (MIT
77 Massachusetts Avenue
NE18-901
Cambridge, MA 02139 -4307

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
71204-CS-YIP.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tamara Broderick
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-324-6749

RPPR Final Report
as of 19-Oct-2021

Agency Code: 21XD

Proposal Number: 71204CSYIP **Agreement Number: W911NF-18-1-0063**
INVESTIGATOR(S):

Name: Tamara A Broderick
Email: tbroderick@csail.mit.edu
Phone Number: 6173246749
Principal: Y

Organization: **Massachusetts Institute of Technology (MIT)**
Address: 77 Massachusetts Avenue, Cambridge, MA 021394307
Country: USA
DUNS Number: 001425594 EIN: 042103594

Report Date: 29-Mar-2021 Date Received: 05-Aug-2021
Final Report for Period Beginning 30-Dec-2017 and Ending 29-Dec-2020
Title: Reliability and robustness for fast Bayesian inference of complex data
Begin Performance Period: 30-Dec-2017 **End Performance Period:** 29-Dec-2020
Report Term: 0-Other

Submitted By: Tamara Broderick Email: tbroderick@csail.mit.edu
Phone: (617) 324-6749

STEM Degrees: **STEM Participants:**

Major Goals: The overall goal of this project is to develop inference methods that practitioners can trust. Importantly, these methods should be able to run quickly on modern problems of interest, and they should be able to work on complex data. That is, practitioners should be able to trust these methods as they actually work in practice. Especially in light of existing heuristics that can yield arbitrarily wrong results, an important part of reliability comes from supporting theory and evaluation.

Specifically our goal has been to develop algorithms and theory for Bayesian inference and evaluation. In particular, we aim to guarantee that the quantities of interest to practitioners are well-recovered by our algorithms and that our algorithms yield fast, practical results. We endeavor to engage with challenging real-world data problems.

Accomplishments: See pdf in Upload section.

Training Opportunities: Graduate students and postdocs in the PI's group meet individually on (at least) a weekly basis with the PI. In addition, they attend a weekly group meeting -- featuring research presentations and skillshares. Finally, they attend a weekly reading group in which they present and discuss related research.

Results Dissemination: The PI and her group members have presented the research work here in many university seminars, workshops, conferences, and other meetings. Many of these are recorded for online viewing.

RPPR Final Report
as of 19-Oct-2021

Honors and Awards: 2020 Summer School of Machine Learning at Skoltech (SMILES): Students Choice Award (for top-3 best lecturers of the school)

2020 Ruth and Joel Spira Award for Distinguished Teaching

2019 AISTATS Notable Paper Award

2019 Junior Bose Award (for outstanding contributions to education in the MIT School of Engineering)

2018 NSF CAREER Award

2018 Alfred P. Sloan Research Fellowship

2018 Amazon Research Award

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Trevor Campbell
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Jonathan Huggins
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: William Stephenson
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Raj Agrawal
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: PD/PI
Participant: Tamara Broderick
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 19-Oct-2021

Participant Type: Graduate Student (research assistant)
Participant: Brian Trippe
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Sameer Deshpande
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Tin Nguyen
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Hannah Diehl
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 19-Oct-2021

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 0-Other
Journal: In process of preparation
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/5/21 12:00AM Date Published:
Publication Location:
Article Title: An Automatic Finite-Sample Robustness Metric, Can Dropping a Little Data Change Conclusions?
Authors: Tamara Broderick, Ryan Giordano, Rachael Meager
Keywords: sensitivity, robustness, influence function, dropping data
Abstract: We propose a method to assess the sensitivity of econometric analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so we provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. We call our resulting metric the Approximate Maximum Influence Perturbation. Our approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes). At minimal computational cost, our metric provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness our metric finds is conclusive. While we find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 0-Other
Journal: In process of preparation
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/5/21 12:00AM Date Published:
Publication Location:
Article Title: Approximate Cross-Validation for Structured Models
Authors: Soumya Ghosh, William Stephenson, Tin Nguyen, Sameer Desphande, Tamara Broderick
Keywords: approximate cross validation, time series, spatiotemporal
Abstract: Many modern data analyses benefit from explicitly modeling dependence structure in data – such as measurements across time or space, ordered words in a sentence, or genes in a genome. Cross-validation is the gold standard to evaluate these analyses but can be prohibitively slow due to the need to re-run already-expensive learning algorithms many times. Previous work has shown approximate cross-validation (ACV) methods provide a fast and provably accurate alternative in the setting of empirical risk minimization. But this existing ACV work is restricted to simpler models by the assumptions that (i) data are independent and (ii) an exact initial model fit is available. In structured data analyses, (i) is always untrue, and (ii) is often untrue. In the present work, we address (i) by extending ACV to models with dependence structure. To address (ii), we verify – both theoretically and empirically – that ACV quality deteriorates smoothly with noise in the initial fit.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 19-Oct-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 3-Accepted
Journal: AISTATS
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/5/21 12:00AM Date Published:
Publication Location:
Article Title: Validated Variational Inference via Practical Posterior Error Bounds
Authors: Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, Tamara Broderick
Keywords: Bayes, variational, workflow, posterior
Abstract: Variational inference has become an increasingly attractive fast alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, a major obstacle to the widespread use of variational methods is the lack of post-hoc accuracy measures that are both theoretically justified and computationally efficient. In this paper, we provide rigorous bounds on the error of posterior mean and uncertainty estimates that arise from full-distribution approximations, as in variational inference. Our bounds are widely applicable, as they require only that the approximating and exact posteriors have polynomial moments. Our bounds are also computationally efficient for variational inference because they require only standard values from variational objectives, straightforward analytic calculations, and simple Monte Carlo estimates. We show that our analysis naturally leads to a new and improved workflow for validated variational inference.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 3-Accepted
Journal: AISTATS
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/5/21 12:00AM Date Published: 8/27/21 12:48AM
Publication Location:
Article Title: Approximate Cross-Validation in High Dimensions with Guarantees
Authors: William Stephenson, Tamara Broderick
Keywords: approximate cross validation, infinitesimal jackknife, high dimensions
Abstract: Leave-one-out cross-validation (LOOCV) can be particularly accurate among cross-validation (CV) variants for machine learning assessment tasks – e.g., assessing methods' error or variability. But it is expensive to re-fit a model N times for a dataset of size N. Previous work has shown that approximations to LOOCV can be both fast and accurate – when the unknown parameter is of small, fixed dimension. But these ap-proximations incur a running time roughly cubic in dimension – and we show that, besides computational issues, their accuracy dramatically deteriorates in high dimensions. We find that all but proposals perform so poorly as to be unusable for approximating LOOCV. Crucially, though, we are able to show, both empirically and theoretically, that one approximation can perform well in high dimensions – in cases where the high-dimensional parameter exhibits sparsity. Under interpretable assumptions, our theory demonstrates that the problem can be reduced to a small support.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 19-Oct-2021

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 0-Other
Journal: In process of preparation
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/5/21 12:00AM Date Published:
Publication Location:
Article Title: Approximate Cross-Validation with Low-Rank Data in High Dimensions
Authors: William Stephenson, Madeleine Udell, Tamara Broderick
Keywords: approximate cross validation, low rank
Abstract: Many recent advances in machine learning are driven by a challenging trifecta: large data size N, high dimensions, and expensive algorithms. In this setting, cross-validation (CV) serves as an important tool for model assessment. Recent advances in approximate cross validation (ACV) provide accurate approximations to CV with only a single model fit, avoiding traditional CV's requirement for repeated runs of expensive algorithms. Unfortunately, these ACV methods can lose both speed and accuracy in high dimensions — unless sparsity structure is present in the data. Fortunately, there is an alternative type of simplifying structure that is present in most data: approximate low rank (ALR). Guided by this observation, we develop a new algorithm for ACV that is fast and accurate in the presence of ALR data.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Partners

,

Prof. Caroline Uhler, MIT, MA, USAProf. Rachael Meager, London School of Economics, UKSaifuddin Syed, Univer

I certify that the information in the report is complete and accurate:
Signature: Tamara Broderick
Signature Date: 8/5/21 9:25PM

Interim Project Report for Grant W911NF-18-1-0063
Report Period - Start Date: 2017-12-30; End Date: 2020-12-29

1 Statement of the problem studied

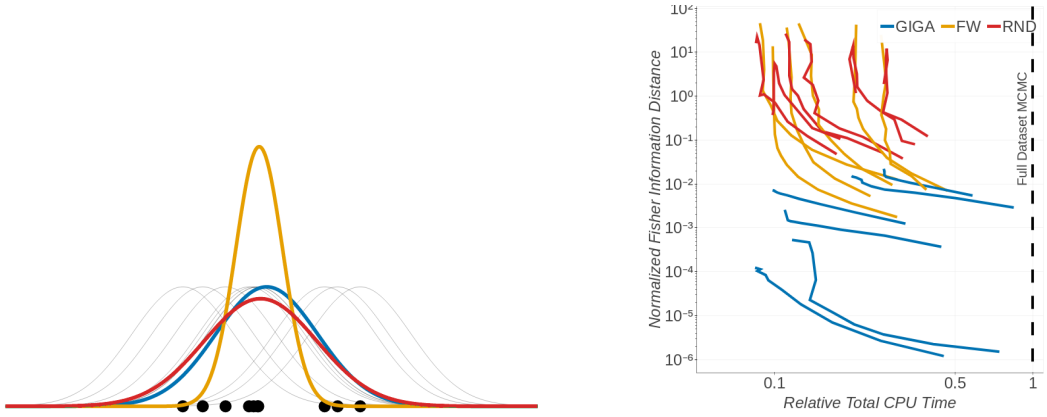
The overall goal of this project is to develop inference methods that practitioners can trust. Importantly, these methods should be able to run quickly on modern problems of interest, and they should be able to work on complex data. That is, practitioners should be able to trust these methods as they actually work in practice. Especially in light of existing heuristics that can yield arbitrarily wrong results, an important part of reliability comes from supporting theory and evaluation.

Specifically our goal has been to develop algorithms and theory for Bayesian inference and evaluation. In particular, we aim to guarantee that the quantities of interest to practitioners are well-recovered by our algorithms and that our algorithms yield fast, practical results. We endeavor to engage with challenging real-world data problems.

2 Summary of the most important results

Bayesian coreset construction via greedy iterative geodesic ascent [5] Coherent uncertainty quantification is a key strength of Bayesian methods. But modern algorithms for approximate Bayesian posterior inference often sacrifice accurate posterior uncertainty estimation in the pursuit of scalability. This work shows that previous Bayesian coreset construction algorithms – which build a small, weighted subset of the data that approximates the full dataset – are no exception. We demonstrate that these algorithms scale the coreset log-likelihood suboptimally, resulting in underestimated posterior uncertainty. To address this shortcoming, we develop greedy iterative geodesic ascent (GIGA), a novel algorithm for Bayesian coreset construction that scales the coreset log-likelihood optimally. GIGA provides geometric decay in posterior approximation error as a function of coreset size, and maintains the fast running time of its predecessors. Our paper concludes with validation of GIGA on both synthetic and real datasets, demonstrating that it reduces posterior approximation error by orders of magnitude compared with previous coreset constructions (Fig. 1).

Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach [10] Bayesian inference typically requires the computation of an approximation to the posterior distribution. An important requirement for an approximate Bayesian inference algorithm is to output high-accuracy posterior mean and uncertainty estimates. Classical Monte Carlo methods, particularly Markov Chain Monte Carlo, remain the gold standard for approximate Bayesian inference because they have a robust finite-sample theory and reliable convergence diagnostics. However, alternative methods, which are more scalable or apply to problems where Markov Chain Monte Carlo cannot be used, lack the same finite-data approximation theory and tools for evaluating their accuracy. In this work, we develop a flexible new approach to bounding the error of mean and uncertainty estimates of scalable inference algorithms. Our strategy is to control the estimation errors in terms of Wasserstein distance, then bound the Wasserstein distance via a generalized notion of Fisher distance. Unlike computing the Wasserstein distance, which requires access to the normalized posterior distribution, the Fisher distance is tractable to compute because it requires access only to the gradient of the log posterior density. We demonstrate the usefulness of our Fisher distance approach by deriving bounds on the Wasserstein error of the Laplace



(a)

Figure 1: *Left*: Gaussian inference for an unknown mean, showing data (black points and likelihood densities), exact posterior (blue), and optimal coreset posterior approximations of size 1 from solving the ideal coreset construction problem (red) and the modified problem (orange). The orange coreset posterior has artificially low uncertainty. The exact and approximate log-posteriors are scaled down (by the same amount) for visualization. *Right*: Comparison of the median Fisher information distance to the true posterior for GIGA, FW, and RND on the logistic and Poisson regression models over 20 random trials. Distances are normalized by the median value of RND for comparison. On the right, computation time is normalized by the median value required to run MCMC on the full dataset. GIGA consistently outperforms FW and RND.

approximation and Hilbert coresets. We anticipate that our approach will be applicable to many other approximate inference methods such as the integrated Laplace approximation, variational inference, and approximate Bayesian computation.

This work provides the missing link between our Hilbert coresets framework and direct theoretical error bounds on posterior point estimates and uncertainties.

Scalable Gaussian process inference with finite-data mean and variance guarantees [11] Gaussian processes (GPs) offer a flexible class of priors for nonparametric Bayesian regression, but popular GP posterior inference methods are typically prohibitively slow or lack desirable finite-data guarantees on quality. We develop an approach to scalable approximate GP regression with finite-data guarantees on the accuracy of pointwise posterior mean and variance estimates. Our main contribution is a novel objective for approximate inference in the nonparametric setting: the *preconditioned Fisher (pF) divergence*. We show that unlike the Kullback–Leibler divergence (used in variational inference), the pF divergence bounds the 2-Wasserstein distance, which in turn provides tight bounds the pointwise difference of the mean and variance functions. We demonstrate that, for sparse GP likelihood approximations, we can minimize the pF divergence efficiently. Our experiments show that optimizing the pF divergence has the same computational requirements as variational sparse GPs while providing comparable empirical performance – in addition to our novel finite-data quality guarantees.

Minimal I-MAP MCMC for scalable structure discovery in causal DAG models [1] Learning a Bayesian network (BN) from data can be useful for decision-making or discovering causal relationships. However, traditional methods often fail in modern applications, which exhibit a larger number of observed variables than data points. The resulting uncertainty about the underlying network as well as the desire to incorporate prior information recommend a Bayesian approach to learning the BN, but the highly combinatorial structure of BNs poses a striking challenge for inference. The current state-of-the-art methods such as order MCMC are faster than previous methods but prevent the use of many natural structural priors and still have running time exponential in the maximum indegree of the true directed acyclic graph (DAG) of the BN. We here propose an alternative posterior approximation based on the observation that, if we incorporate empirical conditional independence tests, we can focus on a high-probability DAG associated with each order of the vertices. We show that our method allows the desired flexibility in prior specification, removes timing dependence on the maximum indegree, and yields provably good posterior approximations; in addition, we show that it achieves superior accuracy, scalability (Fig. 2), and sampler mixing on several datasets.

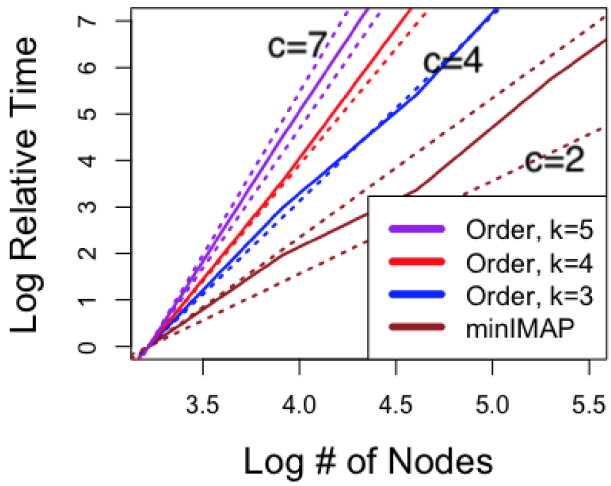


Figure 2: Average iteration times for different sized networks. The times are relative to the average iteration time for $p = 25$ nodes; c denotes the slope of the dotted lines and estimates the computational complexity $O(p^c)$. We compare our method, minIMAP, to order MCMC.

A Swiss Army Infinitesimal Jackknife [9] The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data. The ubiquitous tools of cross-validation (CV) and the bootstrap are examples of this technique. These methods are powerful in large part due to their model agnosticism but can be slow to run on modern, large data sets due to the need to repeatedly re-fit the model. In this work, we use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster than repeated re-fitting by an order of magnitude. This linear approximation is sometimes known as the "infinitesimal jackknife" in the statistics literature, where it is mostly used as a theoretical tool to prove asymptotic results. We provide explicit finite-sample error bounds for the infinitesimal jackknife in terms of a small number of simple, verifiable

assumptions. Our results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the infinitesimal jackknife on a wide variety of problems. As a corollary, we state mild regularity conditions under which our approximation consistently estimates true leave-k-out cross-validation for any fixed k. These theoretical results, together with modern automatic differentiation software, support the application of the infinitesimal jackknife to a wide variety of practical problems in machine learning, providing a "Swiss Army infinitesimal jackknife." We demonstrate the accuracy of our methods on a range of simulated and real datasets (Fig. 3).

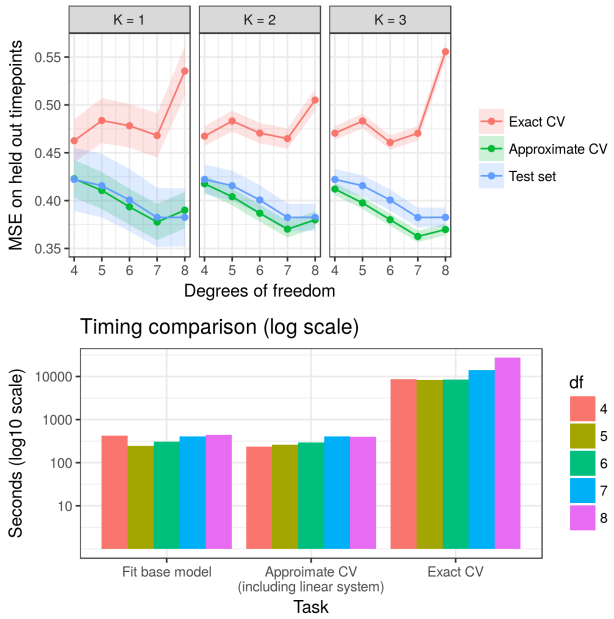


Figure 3: We consider a genomics application in which we use CV to choose the degree of a spline smoother when clustering time series of gene expression data. We show that our method is more accurate than exact CV in this example and orders of magnitude faster – essentially the same speed as running the algorithm once.

The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions [3] Discovering interaction effects on a response of interest is a fundamental problem faced in biology, medicine, economics, and many other scientific disciplines. In theory, Bayesian methods for discovering pairwise interactions enjoy many benefits such as coherent uncertainty quantification, the ability to incorporate background knowledge, and desirable shrinkage properties. In practice, however, Bayesian methods are often computationally intractable for even moderate-dimensional problems. Our key insight is that many hierarchical models of practical interest admit a particular Gaussian process (GP) representation; the GP allows us to capture the posterior with a vector of $O(p)$ kernel hyper-parameters rather than $O(p^2)$ interactions and main effects. With the implicit representation, we can run Markov chain Monte Carlo (MCMC) over model hyper-parameters in time and memory linear in p per iteration. We focus on sparsity-inducing models and show on datasets with a variety of covariate behaviors that our method: (1) reduces runtime by orders of magnitude over naive applications of MCMC, (2) provides lower

Table 1: Building dataset results. MAIN (respectively, PAIR) MSE refers to total error in estimating main (respectively, pairwise) effects. The main and pairwise MSE added together yield the total MSE. The second and fourth columns show (# of effects correctly selected) : (# of incorrect effects selected) for main and pairwise effects, respectively. Larger green values are better while larger red values are worse. We compare our method to two variants of LASSO designed to solve this problem. In particular, notice that our method picks up at least as many correct effects as any other method and zero incorrect effects (the best possible number of incorrect effects).

METHOD	MAIN MSE	# MAIN	PAIR MSE	# PAIR
OUR METHOD	0.1	**3 : 0**	7.0	**3 : 0**
PLASSO	5.0	**2 : 5**	9.3	**3 : 21**
HLASSO	1.5	**3 : 19**	7.8	**3 : 18**

Type I and Type II error relative to state-of-the-art LASSO-based approaches, and (3) offers improved computational scaling in high dimensions relative to existing Bayesian and LASSO-based approaches.

Data-dependent compression of random features for large-scale kernel approximation [2] Kernel methods offer the flexibility to learn complex relationships in modern, large data sets while enjoying strong theoretical guarantees on quality. Unfortunately, these methods typically require cubic running time in the data set size, a prohibitive cost in the large-data setting. Random feature maps (RFMs) and the Nyström method both consider low-rank approximations to the kernel matrix as a potential solution. But, in order to achieve desirable theoretical guarantees, the former may require a prohibitively large number of features J_+, and the latter may be prohibitively expensive for high-dimensional problems. We propose to combine the simplicity and generality of RFMs with a data-dependent feature selection scheme to achieve desirable theoretical approximation properties of Nyström with just $O(\log J_+)$ features. Our key insight is to begin with a large set of random features, then reduce them to a small number of weighted features in a data-dependent, computationally efficient way, while preserving the statistical guarantees of using the original large set of features. We achieve this compression exactly by repurposing the efficient Bayesian coreset construction algorithms we developed in previous work for this project [5] – but now applying it to reduce the number of features rather than the cardinality of a data set. We demonstrate the efficacy of our method with theory and experiments—including on a data set with over 50 million observations. In particular, we show that our method achieves small kernel matrix approximation error and better test set accuracy with provably fewer random features than state-of-the-art methods. See, e.g., Figure 4.

LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations [15] Due to the ease of modern data collection, applied statisticians often have access to a large set of covariates that they wish to relate to some observed outcome. Generalized linear models (GLMs) offer a particularly interpretable framework for such an analysis. In these high-dimensional problems, the number of covariates is often large relative to the number of observations, so we face

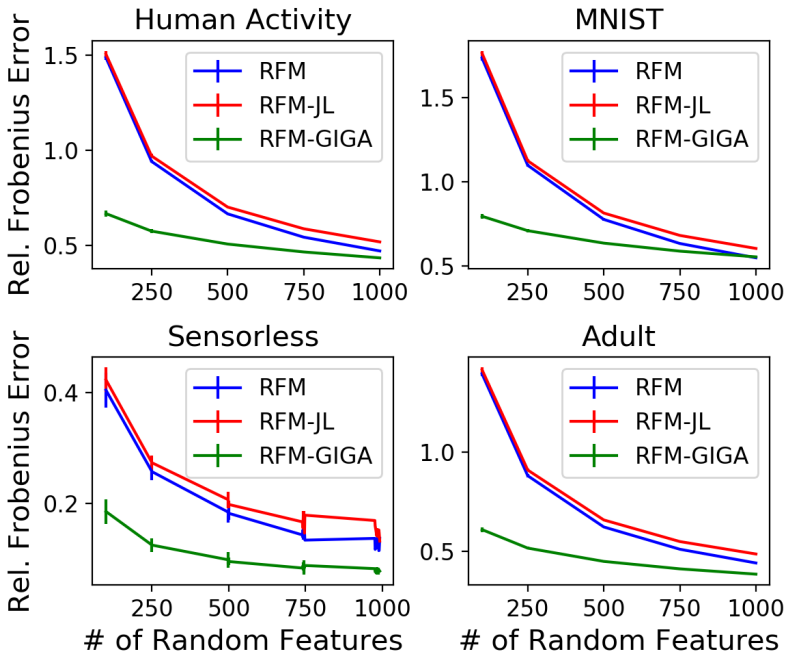


Figure 4: Kernel matrix approximation error is on the vertical axes. Lower is better. The horizontal axis gives the number of features. Points average 20 runs; error bar is one standard deviation. RFM stands for random feature maps. RFM-JL represents a widely used compression method for random feature maps. We find that while RFM-JL does not decrease Frobenius error for different numbers of features, our method (RFM-GIGA) decreases error substantially.

non-trivial inferential uncertainty; a Bayesian approach allows coherent quantification of this uncertainty. Unfortunately, existing methods for Bayesian inference in GLMs require running times roughly cubic in parameter dimension, and so are limited to settings with at most tens of thousand parameters. We propose to reduce time and memory costs with a low-rank approximation of the data in an approach we call LR-GLM. When used with the Laplace approximation or Markov chain Monte Carlo, LR-GLM provides a full Bayesian posterior approximation and admits running times reduced by a full factor of the parameter dimension. We rigorously establish the quality of our approximation and show how the choice of rank allows a tunable computational–statistical trade-off. Experiments support our theory and demonstrate the efficacy of LR-GLM on real large-scale datasets.

A Higher-Order Swiss Army Infinitesimal Jackknife [8] Cross validation (CV) and the bootstrap are ubiquitous model-agnostic tools for assessing the error or variability of machine learning and statistical estimators. However, these methods require repeatedly re-fitting the model with different weighted versions of the original dataset, which can be prohibitively time-consuming. For sufficiently regular optimization problems the optimum depends smoothly on the data weights, and so the process of repeatedly re-fitting can be approximated with a Taylor series that can be often evaluated relatively quickly. In our previous work for this project, we showed that the first-order approximation works well in theory and in practice. In the current work, we consider high-order

approximations, which we call the "higher order infinitesimal jackknife" (HOIJ). Under mild regularity conditions, we provide a simple recursive procedure to compute approximations of all orders with finite-sample accuracy bounds. Additionally, we show that the HOIJ can be efficiently computed even in high dimensions using forward mode automatic differentiation. We show that a linear approximation with bootstrap weights approximation is equivalent to those provided by asymptotic normal approximations. Consequently, the HOIJ opens up the possibility of enjoying higher-order accuracy properties of the bootstrap using local approximations. Consistency of the HOIJ for leave-one-out CV under different asymptotic regimes follows as corollaries from our finite sample bounds under additional regularity assumptions. The generality of the computation and bounds motivate the name "higher-order Swiss Army infinitesimal jackknife."

Local exchangeability [6] Exchangeability – in which the distribution of an infinite sequence is invariant to reorderings of its elements – implies the existence of a simple conditional independence structure that may be leveraged in the design of probabilistic models and efficient inference algorithms. In practice, however, this assumption is too strong an idealization; the distribution typically fails to be exactly invariant to permutations and de Finetti's representation theory does not apply. Thus there is the need for a distributional assumption that is both weak enough to hold in practice, and strong enough to guarantee a useful underlying representation. We introduce a relaxed notion of local exchangeability – where swapping data associated with nearby covariates causes a bounded change in the distribution. Next, we prove that locally exchangeable processes correspond to independent observations from an underlying measure-valued stochastic process, showing that de Finetti's theorem is robust to perturbation and providing further justification for the Bayesian modelling approach. We also provide an investigation of approximate sufficiency and sample continuity properties of locally exchangeable processes on the real line. The paper concludes with examples of popular statistical models that exhibit local exchangeability.

Validated Variational Inference via Practical Posterior Error Bounds [12] Variational inference has become an increasingly attractive fast alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, a major obstacle to the widespread use of variational methods is the lack of post-hoc accuracy measures that are both theoretically justified and computationally efficient. In fact, in this paper, we show cases where the Kullback-Leibler divergence can be small but posterior mean and variance estimates – even in unimodal, one-dimensional distributions – can be arbitrarily wrong. To address these issues, we provide rigorous bounds on the error of posterior mean and uncertainty estimates that arise from full-distribution approximations, as in variational inference. Our bounds are widely applicable, as they require only that the approximating and exact posteriors have polynomial moments. Our bounds are also computationally efficient for variational inference because they require only standard values from variational objectives, straightforward analytic calculations, and simple Monte Carlo estimates. We show that our analysis naturally leads to a new and improved workflow for validated variational inference. Finally, we demonstrate the utility of our proposed workflow and error bounds on a robust regression problem and on a real-data example with a widely used multilevel hierarchical model.

An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions? [4] We propose a method to assess the sensitivity of econometric analyses to the

removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so we provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. We call our resulting metric the Approximate Maximum Influence Perturbation. Our approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes). We provide explicit finite-sample error bounds on our approximation for linear and instrumental variables regressions. At minimal computational cost, our metric provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness our metric finds is conclusive. We demonstrate that the Approximate Maximum Influence Perturbation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification. Several empirical applications show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While we find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.

Approximate Cross-Validation in High Dimensions with Guarantees [13] Leave-one-out cross-validation (LOOCV) can be particularly accurate among cross- validation (CV) variants for machine learning assessment tasks – e.g., assessing methods' error or variability. But it is expensive to re-fit a model N times for a dataset of size N. Previous work has shown that approximations to LOOCV can be both fast and accurate – when the unknown parameter is of small, fixed dimension. But these approximations incur a running time roughly cubic in dimension – and we show that, besides computational issues, their accuracy dramatically deteriorates in high dimensions. Authors have suggested many potential and seemingly intuitive solutions, but these methods have not yet been systematically evaluated or compared. We find that all but one perform so poorly as to be unusable for approximating LOOCV. Crucially, though, we are able to show, both empirically and theoretically, that one approximation can perform well in high dimensions – in cases where the high-dimensional parameter exhibits sparsity. Under interpretable assumptions, our theory demonstrates that the problem can be reduced to working within an empirically recovered (small) support. This procedure is straightforward to implement, and we prove that its running time and error depend on the (small) support size even when the full parameter dimension is large.

Approximate Cross-Validation with Low-Rank Data in High Dimensions [14] Many recent advances in machine learning are driven by a challenging trifecta: large data size N; high dimensions; and expensive algorithms. In this setting, cross-validation (CV) serves as an important tool for model assessment. Recent advances in approximate cross validation (ACV) provide accurate approximations to CV with only a single model fit, avoiding traditional CV's requirement for repeated runs of expensive algorithms. Unfortunately, these ACV methods can lose both speed and accuracy in high dimensions – unless sparsity structure is present in the data. Fortunately, there is an alternative type of simplifying structure that is present in most data: approximate low rank (ALR). Guided by this observation, we develop a new algorithm for ACV that is fast and accurate in the presence of ALR data. Our first key insight is that the Hessian matrix – whose inverse forms the computational bottleneck of existing ACV methods – is ALR. We show that, despite our use of the *inverse* Hessian, a low-rank approximation using the largest (rather than the smallest) matrix eigenvalues enables fast, reliable ACV. Our second key insight is that, in the presence of ALR data,

error in existing ACV methods roughly grows with the (approximate, low) rank rather than with the (full, high) dimension. These insights allow us to prove theoretical guarantees on the quality of our proposed algorithm – along with fast-to-compute upper bounds on its error. We demonstrate the speed and accuracy of our method, as well as the usefulness of our bounds, on a range of real and simulated data sets.

Approximate Cross-Validation for Structured Models [7] Many modern data analyses benefit from explicitly modeling dependence structure in data – such as measurements across time or space, ordered words in a sentence, or genes in a genome. Cross-validation is the gold standard to evaluate these analyses but can be prohibitively slow due to the need to re-run already-expensive learning algorithms many times. Previous work has shown approximate cross-validation (ACV) methods provide a fast and provably accurate alternative in the setting of empirical risk minimization. But this existing ACV work is restricted to simpler models by the assumptions that (i) data are independent and (ii) an exact initial model fit is available. In structured data analyses, (i) is always untrue, and (ii) is often untrue. In the present work, we address (i) by extending ACV to models with dependence structure. To address (ii), we verify – both theoretically and empirically – that ACV quality deteriorates smoothly with noise in the initial fit. We demonstrate the accuracy and computational benefits of our proposed methods on a diverse set of real-world applications.

References

[1] R. Agrawal, T. Broderick, and C. Uhler. Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning*, 2018.

[2] R. Agrawal, T. Campbell, J. H. Huggins, and T. Broderick. Data-dependent compression of random features for large-scale kernel approximation. In *AISTATS*, 2019.

[3] R. Agrawal, B. Trippe, J. Huggins, and T. Broderick. The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In K. Chaudhuri and R. Salakhutdinov, editors, *International Conference on Machine Learning (ICML)*, volume 97, pages 141–150, 2019. URL `http://proceedings.mlr.press/v97/agrawal19a.html`.

[4] T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *In preparation*, 2020.

[5] T. Campbell and T. Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, 2018.

[6] T. Campbell, S. Syed, C.-Y. Yang, M. I. Jordan, and T. Broderick. Local exchangeability. *arXiv preprint arXiv:1906.09507*, 2019.

[7] S. Ghosh, W. T. Stephenson, T. D. Nguyen, S. Deshpande, and T. Broderick. Approximate cross-validation for structured models. *NeurIPS*, 2020.

[8] R. Giordano, M. I. Jordan, and T. Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019.

[9] R. Giordano, W. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A Swiss Army infinitesimal jackknife. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[10] J. H. Huggins, T. Campbell, M. Kasprzak, and T. Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *Under review*, 2018.

[11] J. H. Huggins, T. Campbell, M. Kasprzak, and T. Broderick. Scalable Gaussian process inference with finite-data mean and variance guarantees. In *AISTATS*, 2019, to appear.

[12] J. H. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated variational inference via practical posterior error bounds. In *AISTATS*, 2020.

[13] W. Stephenson and T. Broderick. Approximate cross-validation in high dimensions with guarantees. In *AISTATS*, pages 2424–2434. PMLR, 2020.

[14] W. T. Stephenson, M. Udell, and T. Broderick. Approximate cross-validation with low-rank data in high dimensions. *NeurIPS*, 2020.

[15] B. Trippe, J. Huggins, R. Agrawal, and T. Broderick. LR-GLM: High-dimensional Bayesian inference using low-rank data approximations. In K. Chaudhuri and R. Salakhutdinov, editors, *International Conference on Machine Learning (ICML)*, volume 97, pages 6315–6324, 2019. URL `http://proceedings.mlr.press/v97/trippe19a.html`.