On Consistent Bayesian Inference from Synthetic Data

Ossi Räisä Joonas Jälkö Antti Honkela
Department of Computer Science
University of Helsinki
{ossi.raisa, joonas.jalko, antti.honkela}@helsinki.fi

Abstract

Generating synthetic data, with or without differential privacy, has attracted significant attention as a potential solution to the dilemma between making data easily available, and the privacy of data subjects. Several works have shown that consistency of downstream analyses from synthetic data, including accurate uncertainty estimation, requires accounting for the synthetic data generation. There are very few methods of doing so, most of them for frequentist analysis. In this paper, we study how to perform consistent Bayesian inference from synthetic data. We prove that mixing posterior samples obtained separately from multiple large synthetic datasets converges to the posterior of the downstream analysis under standard regularity conditions when the analyst's model is compatible with the data provider's model. We also present several examples showing how the theory works in practice, and showing how Bayesian inference can fail when the compatibility assumption is not met, or the synthetic dataset is not significantly larger than the original.

1 Introduction

Synthetic data has the potential of opening privacy-sensitive datasets for widespread analysis. The most convenient and straightforward way for downstream analysts to analyse synthetic data is using the same method that would be used with real data. However, ignoring the additional stochasticity arising from the synthetic data generation will yield biased results and overconfident uncertainty estimates (Raghunathan et al. 2003; Räisä et al. 2023; Wilde et al. 2021). This is especially problematic when *differential privacy* (DP) (Dwork et al. 2006b) is used to guarantee the privacy of the synthetic data, as it requires adding extra noise to the synthetic data generation process. This problem creates the need for *noise-aware* analyses that account for the synthetic data generation.

For frequentist downstream analyses, it is possible to account for the extra uncertainty using multiple synthetic datasets (Raghunathan et al. 2003; Räisä et al. 2023) while reusing the existing analysis method that would be used with real data. For Bayesian analyses, the only clearly noise-aware method (Wilde et al. 2021) requires both public data to correct the analysis, and a more complicated analysis process.

We study the applicability of multiple synthetic datasets to synthetic data, aiming the bring the simplicity of the frequentist methods using multiple synthetic datasets to Bayesian downstream analysis. The frequentist methods using multiple synthetic datasets were derived from methods in missing data imputation (Rubin 1987), so our starting point is the method of Gelman et al. (2014) for Bayesian inference with missing data. They proposed inferring the downstream posterior by imputing multiple completed datasets, inferring the analysis posterior for each completed dataset separately, and mixing the posteriors together. We investigate whether this method is also applicable to synthetic data.

NeurIPS 2023 Workshop on Synthetic Data for Empowering ML Research.

Contributions

1. We study inferring the downstream analysis posterior by generating multiple synthetic datasets, inferring the analysis posterior for each synthetic dataset as if it were the real dataset, and mixing the posteriors together. We find two important conditions for consistent Bayesian inference in this setting: synthetic datasets that are larger than the original one, and a notion of compatibility between the data provider's and analyst's models called *congeniality* (Meng 1994), which we introduce in Section 2.1.
2. We prove that when congeniality is met and the Bernstein–von Mises, or a similar theorem, applies, this method converges to the true posterior as the number of synthetic datasets and the size of the synthetic datasets grow. These are presented in Section 2.
3. We evaluate this method on logistic regression and Gaussian mean estimation examples in Section 3. We verify that the methods works in practice when the assumptions are met, and examine what can happen when congeniality is not met.

2 Bayesian Inference from Synthetic Data

When the downstream analysis is Bayesian, the analyst wants to obtain the posterior $p(Q|X, I_A)$ of some quantity $Q \in \mathcal{Q}$ given real data X and the background knowledge I_A, such as priors, of the analyst. We assume the analyst has a method to sample $p(Q|X, I_A)$ if they had access to real data, and study what they can do when they only have access to synthetic data. We introduce Bayesian inference and additional background material in Supplemental Section A.

In the DP case, the exact posterior is unobtainable, so we assume that X is only available through a noisy summary $\tilde{s}$ (Ju et al. 2022; Räisä et al. 2023), so the posterior is $p(Q|\tilde{s}, I_A)$. To unify these notations, we use Z to denote the observed values, so $Z = X$ in the non-DP case, $Z = \tilde{s}$ in the DP case, and the posterior of interest is $p(Q|Z, I_A)$. We summarise these random variables and their dependencies in Figure 1, and give an introduction to DP in Supplemental Section A.2.

In order to introduce the synthetic data into the posterior of interest, we can decompose the posterior as

$$p(Q|Z, I_A) = \int p(Q|Z, X^*, I_A)p(X^*|Z, I_A)\,\mathrm{d}X^*, \tag{1}$$

where we abuse notation by using X^* as the variable to integrate over, so inside the integral X^* is not a random variable. The decomposition in (1) means that we could sample $p(Q|Z, I_A)$ by

- θ: data generating model parameters
- X: real data
- X^*: hypothetical data
- Z: observed summary of X ($Z = X$ without DP)
- X^{Syn}: synthetic data, $X^{Syn} \sim p(X^*|Z, I_S)$
- $Q \in \mathcal{Q}$: estimated quantity in downstream analysis
- I_S: synthetic data provider's background information
- I_A: analyst's background information

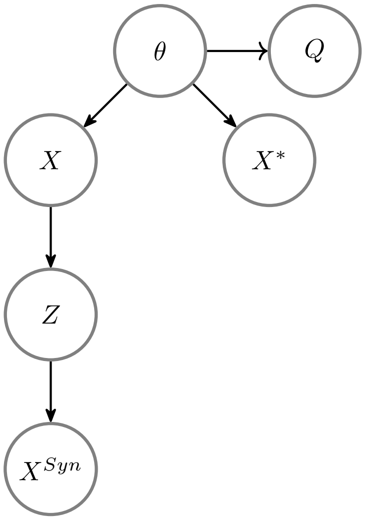


Figure 1: Left: random variables in noise-aware uncertainty estimation from synthetic data. Right: a Bayesian network describing the dependencies of the random variables. This network is conditional on either I_S or I_A, depending on whether viewed by the data provider or the analyst.

first sampling the synthetic data from the posterior predictive $X^{Syn} \sim p(X^*|Z, I_A)$, and then sampling $Q \sim p(Q|Z, X^* = X^{Syn}, I_A)$. The posterior predictive decomposes as $p(X^*|Z, I_A) = \int p(X^*|\theta, I_A) p(\theta|Z, I_A) \,\mathrm{d}\theta$, where θ is the data generating model parameters, so it is sampled by sampling $\theta \sim p(\theta|Z, I_A)$, and then sampling $X^{Syn} \sim p(X^*|\theta, I_A)$.

Note that X^* and X^{Syn} are not the same random variable. X^* represents a hypothetical real dataset that could be obtained if more data was collected, as seen in Figure 1, and it is not the synthetic dataset. The synthetic dataset X^{Syn} is a sample from the conditional distribution of X^* given Z. For this reason, $p(Q|Z, X^*, I_A) \neq p(Q|Z, I_A)$. To make our notation less cluttered, we write $p(\,\cdot\,|X^*, \,\cdot\,)$ in place of $p(\,\cdot\,|X^* = X^{Syn}, \,\cdot\,)$ in probabilities when the meaning is clear.

There are still two major issues with the decomposition in (1):

1. Sampling $p(Q|Z, X^*, I_A)$ requires access to Z, which defeats the purpose of using synthetic data.
2. X^* needs to be sampled conditionally on the analyst's background information I_A, while the synthetic data provider could have different background information I_S.

To solve the first issue, in Section 2.2 we show that if we replace $p(Q|Z, X^*, I_A)$ inside the integral of (1) with $p(Q|X^*, I_A)$, the resulting distribution converges to the desired posterior,

$$\int p(Q|X^*, I_A) p(X^*|Z, I_A) \,\mathrm{d}X^* \to p(Q|Z, I_A) \tag{2}$$

in total variation distance as the size of X^* grows. It should be noted that many synthetic data sets $X^{Syn} \sim p(X^*|Z, I_A)$ will be needed to account for the integral over X^*.

The second issue is known as *congeniality* in the multiple imputation literature (Meng 1994; Xie and Meng 2016). We look at congeniality in the context of Bayesian inference from synthetic data in Section 2.1, and find that we can obtain $p(Q|Z, I_A)$ under appropriate assumptions on the relationship between I_A and I_S.

Exactly sampling the LHS of (2) requires generating a synthetic dataset for each sample of $p(Q|Z, I_A)$, which is not practical. However, we can compute a Monte-Carlo approximation for $p(Q|Z, I_A)$ by generating m synthetic datasets $X_1^{Syn}, \dots, X_m^{Syn} \sim p(X^*|Z, I_A)$, drawing multiple samples from each of the $p(Q|X^* = X_i^{Syn}, I_A)$, and mixing these samples, which allows us to obtain more than one sample of $p(Q|Z, I_A)$ per synthetic dataset. We look at some properties of this in Supplemental Section E, but we use the integral form in (2) in the rest of our theory.

2.1 Congeniality

In the decomposition (1) of the analyst's posterior, X^* should be sampled conditionally on the analyst's background information I_A, while in reality the synthetic data provider could have different background information I_S.

A similar distinction has been studied in the context of missing data (Meng 1994; Xie and Meng 2016), where the imputer of missing data has a similar role as the synthetic data provider. Meng (1994) found that combining inferences from many completed datasets requires that the probability models of both parties are compatible in a certain sense, which they defined as *congeniality*.

As our examples with Gaussian distributions in Supplemental Section C show, some notion of congeniality is also required in our setting. However, because we study synthetic data instead of imputation, and Bayesian instead of frequentist downstream analysis, we need a different formal definition. As the analyst only makes inferences on Q, it suffices that both the analyst and synthetic data provider make the same inferences of Q:

Definition 2.1. *The background information sets I_S and I_A are congenial for observation Z if*

$$p(Q|X^*, I_S) = p(Q|X^*, I_A) \tag{3}$$

for all X^* and

$$p(Q|Z, I_S) = p(Q|Z, I_A). \tag{4}$$

In the non-DP case, (4) is redundant, as it is implied by (3), but in the DP case, both are needed, as the parties may draw different conclusions on X given $Z = \tilde{s}$.

Combining congeniality and (2),

$$\begin{aligned}\int p(Q|X^*, I_A)p(X^*|Z, I_S)\,\mathrm{d}X^* &= \int p(Q|X^*, I_S)p(X^*|Z, I_S)\,\mathrm{d}X^* \\ &\to p(Q|Z, I_S) = p(Q|Z, I_A),\end{aligned} \tag{5}$$

where the convergence is in total variation distance as the size of X^* grows. In the following, we assume congeniality, and drop I_A and I_S from our notation.

2.2 Consistency Proof

To recap, we want to prove that the posterior from synthetic data,

$$\bar{p}_n(Q) = \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^*, \tag{6}$$

converges in total variation distance to $p(Q|Z)$ as the size n of X_n^* grows. We prove this in Theorem 2.4, which requires that both $p(Q|Z, X_n^*)$ and $p(Q|X_n^*)$ approach the same distribution as n grows. We formally state this in Condition 2.2. In Lemma 2.3, we show that Condition 2.2 is a consequence of the Bernstein–von Mises theorem (Theorem A.6) under some additional assumptions, so we expect it to hold in typical settings.

To make the notation more compact, let $\bar{Q}_n^+ \sim p(Q|Z, X_n^*)$, and let $\bar{Q}_n \sim p(Q|X_n^*)$.

Condition 2.2. *For the observed Z and all Q, there exist distributions D_n such that*

$$\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \xrightarrow{P} 0 \quad \text{and} \quad \mathrm{TV}\left(\bar{Q}_n, D_n\right) \xrightarrow{P} 0 \tag{7}$$

as $n \to \infty$, where the convergence in probability is over sampling $X_n^* \sim p(X_n^*|Z, Q)$.

The Bernstein–von Mises theorem (Theorem A.6) implies Condition 2.2 with some additional assumptions:

Lemma 2.3. *If the assumptions of Theorem A.6 (Condition A.5) hold for the downstream analysis for all Q_0, and the following assumptions:*

(1) Z and X^* are conditionally independent given Q; and

(2) $p(Z|Q) > 0$ for all Q,

hold, then Condition 2.2 holds.

Proof. The full proof is in Supplemental Section B.1. Proof idea: when Z and X^* are conditionally independent given Q,

$$p(Q|Z, X^*) \propto p(X^*|Q)p(Z|Q)p(Q) \tag{8}$$

so $p(Q|Z, X^*)$ can be equivalently seen as the result of Bayesian inference with observed data X^* and prior $p(Q|Z)$. As the only difference to $p(Q|X^*)$ is the prior, the Bernstein–von Mises theorem implies that both $p(Q|Z, X^*)$ and $p(Q|X^*)$ converge in total variation distance to the same distribution. □

Assumption (1) of Lemma 2.3 will hold if the downstream analysis treats its input data as an i.i.d. sample from some distribution. Assumption (2) holds when the likelihood is always positive, and in the DP case when the density of the privacy mechanism is also positive everywhere, which is the case for common DP mechanisms like the Gaussian and Laplace mechanisms (Dwork and Roth 2014).

Next is the main theorem of this work: (2) holds under Condition 2.2.

Theorem 2.4. *Under congeniality and Condition 2.2,* $\mathrm{TV}\left(p(Q|Z), \bar{p}_n(Q)\right) \to 0$ *as $n \to \infty$.*

Proof. The full proof is in Supplemental Section B.1. Proof idea: the proof consists of three steps. The first two are in Lemma B.1 and the third is in Lemma B.2 in the Supplement. The first step is showing that $\mathrm{TV}(\bar{Q}_n, \bar{Q}_n^+) \xrightarrow{P} 0$ when $X_n^* \sim p(X_n^*|Z, Q)$ for fixed Z and Q. This is a simple consequence of the triangle inequality and Condition 2.2, as total variation distance is a metric. In the second step, we show that $\mathrm{TV}(\bar{Q}_n, \bar{Q}_n^+) \xrightarrow{P} 0$ also holds when $X_n^* \sim p(X_n^*|Z)$. In the final step, we show that this implies the claim. □

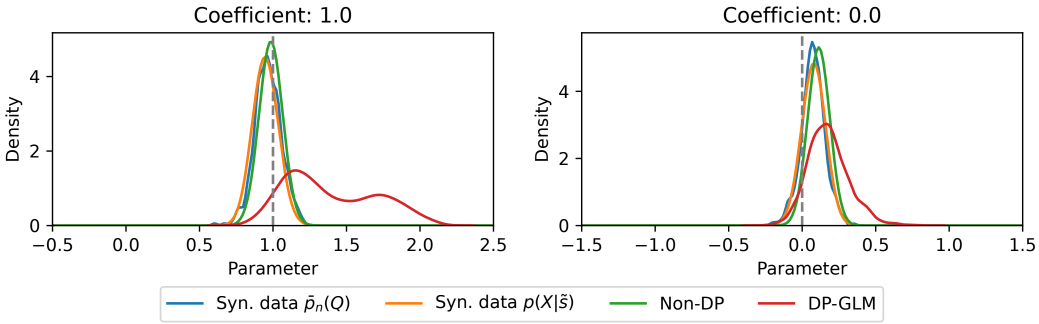


Figure 2: Posteriors in the DP logistic regression experiment, where Q are the regression coefficients. The mixture of posteriors from synthetic data, $\bar{p}_n(Q)$, (with $n_{X^*}/n_X = 20$, $m = 400$) is very close the to the private posterior $p(Q|\tilde{s})$ computed using (1). Computing the posterior without synthetic data with DP-GLM gives a somewhat wider posterior. The true parameter values are highlighted by the grey dashed lines and shown in the panel titles. The privacy bounds are $\epsilon = 1$, $\delta = n_X^{-2} = 2.5 \cdot 10^{-7}$.

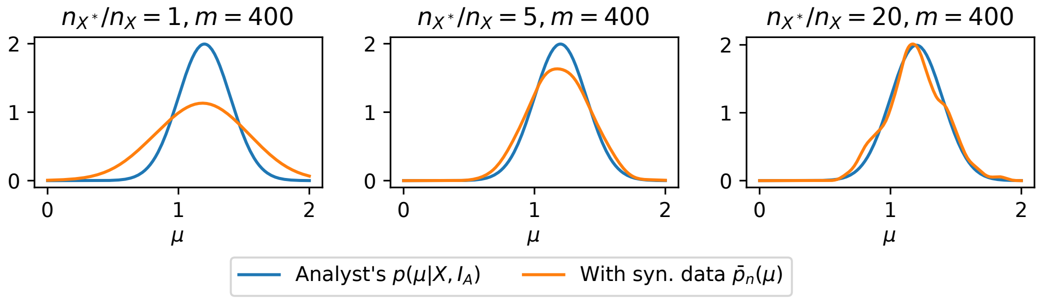


Figure 3: Convergence of the mixture of posteriors from synthetic data on Gaussian mean (μ) estimation with known variance, with different sizes of the synthetic dataset (n_{X^*}), while the size of the real data (n_X) is constant. $n_{X^*} = n_X$ is clearly not enough, but $n_{X^*} = 20n_X$ is already relatively good.

3 Examples

In this section, we present two examples: logistic regression on a toy dataset, with DP synthetic data, and univariate Gaussian mean estimation.

Logistic Regression Example The data for the logistic regression example is a simple 3-d binary toy dataset, ($n_X = 2000$), with DP synthetic data, under the same setting as used by Räisä et al. (2023) for frequentist logistic regression. We change the downstream task to Bayesian logistic regression to evaluate our theory.

Under DP, Z is a noisy summary $\tilde{s}$ of the real data. We need synthetic data sampled from the posterior predictive $p(X^*|\tilde{s})$, which is exactly what the NAPSU-MQ algorithm of Räisä et al. (2023) provides. In NAPSU-MQ, $\tilde{s}$ is the values of user-selected marginal queries with added Gaussian noise. We used the open-source implementation of NAPSU-MQ[1] by Räisä et al. (2023), and describe NAPSU-MQ in Supplemental Section A.2.

Because of the simplicity of this model, it is possible to use the exact posterior decomposition (1) as a baseline, by using $p(X|\tilde{s})$ instead of $p(X^*|\tilde{s})$ to generate synthetic data. We give a detailed description of this process in Supplemental Section D. We have also included the DP-GLM algorithm (Kulkarni et al. 2021) that does not use synthetic data, and the non-DP posterior from the real data as baselines. We obtained the code for DP-GLM from Kulkarni et al. (2021) upon request.

[1] `https://github.com/DPBayes/NAPSU-MQ-experiments`

Figure 2 compares the mixture of posteriors from synthetic data $\bar{p}_n(Q)$ from (6) that uses $p(Q|X^*)$, with $n_{X^*}/n_X = 20$ and $m = 400$ synthetic datasets, to the baselines. $\bar{p}_n(Q)$ is very close to the posterior $p(Q|\tilde{s})$ from (1). The DP-GLM posterior that does not use synthetic data is somewhat wider. The privacy bounds are $\epsilon = 1$, $\delta = n_X^{-2} = 2.5 \cdot 10^{-7}$.

We ran the experiment 100 times and also with $\epsilon = 0.1$ and $\epsilon = 0.5$, and plot coverages and widths of credible intervals in Figure S4 in the Supplement. With $\epsilon = 1$ and $\epsilon = 0.5$, the coverages are accurate and DP-GLM consistently produces wider intervals. With $\epsilon = 0.1$, the mixture of synthetic data posteriors likely needs more and larger synthetic datasets to converge, as it produced wider and slightly overconfident intervals for one coefficient.

Gaussian Example In Figure 3, we examine how changing the size of the synthetic dataset affects $\bar{p}_n(Q)$ when estimating the mean μ of a univariate Gaussian with known variance. We see that when the synthetic dataset has the same size as the real data ($n_{X^*}/n_X = 1$), $\bar{p}_n(Q)$ has a much higher variance than $p(\mu|X, I_A)$, but when the synthetic dataset gets larger, $\bar{p}_n(Q)$ gets closer to $p(\mu|X, I_A)$. See Supplemental Section C for more details on this setting, and additional results.

In the Supplement (Section C), we also look at what happens when congeniality is not met. In the Gaussian mean estimation example, if the data provider and analyst do not have equal known variances, $\bar{p}_n(Q)$ no longer converges to $p(\mu|X, I_A)$, as seen on the right in Figure S1. However, $\bar{p}_n(Q)$ still converges to $p(\mu|X, I_S)$, so the result is still sensible. If the task were to infer the variance σ^2 of the Gaussian instead, and the parties had unequal known means, the situation is different, as seen in Figure S3. In this case, the limit of $\bar{p}_n(Q)$ can be different either parties' posterior from real data, with the difference depending on how different the known means for the parties are.

4 Discussion

Limitations A clear limitation of mixing posteriors from multiple synthetic datasets is the computational cost of analysing many large synthetic datasets, which may be substantial for more complex Bayesian downstream models, where even a single analysis can be computationally expensive. However, the separate analyses can be run in parallel. We also expect that the information gained from sampling the posteriors from a few synthetic datasets could be used to speed up sampling the others, as they likely won't bee too far from the sampled ones.

Under DP, noise-aware synthetic data generation is needed, which limits the settings in which the method can currently be applied. However, if new noise-aware methods are developed in the future, the method can immediately be used with them.

Condition 2.2 limits the applicability our theory to downstream analyses where the prior's influence vanishes as the sample size grows. This does not always happen for some models, such as some infinite-dimensional models, models where the number of parameters increases with dataset size, and models with a support that heavily depends on the parameters. The method also requires congeniality, which basically requires the analyst's prior to be compatible with the data provider's. Our Gaussian examples show that it is sometimes possible to recover useful inferences even without congeniality, but not always, so an important direction for future research is separating these two cases, and finding out what can be done in the latter case.

Conclusion We considered the problem of consistent Bayesian inference of downstream analyses using multiple, potentially DP, synthetic datasets, and studied an inference method that mixes the posteriors from multiple large synthetic datasets. We proved, under congeniality and the general and well-understood regularity conditions of the Bernstein–von Mises theorem, that the method is asymptotically exact as the sizes of the synthetic datasets grow. We studied the method in two examples: non-private Gaussian mean (or variance) estimation and DP logistic regression. In the former, we were able to use the analytically tractable structure of the setting to derive additional properties of the method, in particular examining what can happen without congeniality. In both settings, we experimentally validated our theory, and showed that the method works in practice. This greatly expands the understanding of Bayesian inference from synthetic data, filling a major gap in the synthetic data analysis literature.

Acknowledgments and Disclosure of Funding

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI and Grant 356499), the Strategic Research Council at the Research Council of Finland (Grant 358247) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources. We thank Tejas Kulkarni for providing the DP-GLM code.

References

Balle, B. and Y.-X. Wang (2018). "Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 394–403.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). "Hybrid Monte Carlo". In: *Physics letters B* 195.2, pp. 216–222.

Dwork, C. (2008). "Differential Privacy: A Survey of Results". In: *International Conference on Theory and Applications of Models of Computation*. Springer, pp. 1–19.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006a). "Our Data, Ourselves: Privacy Via Distributed Noise Generation". In: *Advances in Cryptology - EUROCRYPT*. Vol. 4004. Lecture Notes in Computer Science. Springer, pp. 486–503.

Dwork, C., F. McSherry, K. Nissim, and A. D. Smith (2006b). "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Third Theory of Cryptography Conference*. Vol. 3876. Lecture Notes in Computer Science. Springer, pp. 265–284.

Dwork, C. and A. Roth (2014). "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4, pp. 211–407.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*. Third edition. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton: CRC Press.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). "Adaptive Rejection Metropolis Sampling Within Gibbs Sampling". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 44.4, pp. 455–472.

Hoffman, M. D. and A. Gelman (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Jennrich, R. I. (1969). "Asymptotic Properties of Non-Linear Least Squares Estimators". In: *The Annals of Mathematical Statistics* 40.2, pp. 633–643.

Ju, N., J. Awan, R. Gong, and V. Rao (2022). "Data Augmentation MCMC for Bayesian Inference from Privatized Data". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 12732–12743.

Kelbert, M. (2023). "Survey of Distances between the Most Popular Distributions". In: *Analytics* 2.1, pp. 225–245.

Kulkarni, T., J. Jälkö, A. Koskela, S. Kaski, and A. Honkela (2021). "Differentially Private Bayesian Inference for Generalized Linear Models". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 5838–5849.

Meng, X.-L. (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input". In: *Statistical Science* 9.4.

Neal, R. M. (2011). "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003). "Multiple Imputation for Statistical Disclosure Limitation". In: *Journal of Official Statistics* 19.1, p. 1.

Räisä, O., J. Jälkö, S. Kaski, and A. Honkela (2023). "Noise-Aware Statistical Inference with Differentially Private Synthetic Data". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3620–3643.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley \& Sons.

Rubin, D. B. (1993). "Discussion: Statistical Disclosure Limitation". In: *Journal of Official Statistics* 9.2, pp. 461–468.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Repr. 2000. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wilde, H., J. Jewson, S. J. Vollmer, and C. Holmes (2021). “Foundations of Bayesian Learning from Synthetic Data”. In: *The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 541–549.

Xie, X. and X.-L. Meng (2016). “Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God’s, Imputer’s and Analyst’s Models Are Uncongenial?” In: *Statistica Sinica*.

A Background

A.1 Bayesian Inference

Bayesian inference is a paradigm of statistical inference where the data analyst's uncertainty in a quantity Q after observing data X is represented using the posterior distribution $p(Q|X)$ (Gelman et al. 2014). The posterior is given by Bayes' rule:

$$p(Q|X) = \frac{p(X|Q)p(Q)}{\int p(X|Q')p(Q')\,\mathrm{d}Q'}, \tag{9}$$

where $p(X|Q)$ is the likelihood of observing the data X for a given value of Q, and $p(Q)$ is the analyst's prior of Q. Computing the denominator is typically intractable, so analysts often use numerical methods to sample $p(Q|X)$ (Gelman et al. 2014).

Total Variation Distance A crucial component of our theory is the notion of *total variation distance* between random variables, which is used to measure the difference between two random variables or probability distributions.

Definition A.1. *The total variation distance between random variables (or distributions) P_1 and P_2 is*

$$\mathrm{TV}(P_1, P_2) = \sup_A |\Pr(P_1 \in A) - \Pr(P_2 \in A)|, \tag{10}$$

where A is any measurable set.

As a slight abuse of notation, we allow the arguments of $\mathrm{TV}(\cdot, \cdot)$ to be random variables, probability distributions, or probability density functions interchangeably.

Lemma A.2 (Kelbert (2023))**.** *Properties of total variation distance:*

1. *For probability densities p_1 and p_2,*

$$\mathrm{TV}(p_1, p_2) = \frac{1}{2}\int |p_1(x) - p_2(x)|\,\mathrm{d}x. \tag{11}$$

2. *Total variation distance is a metric.*

3. *Pinsker's inequality: for distributions P_1 and P_2,*

$$\mathrm{TV}(P_1, P_2) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P_1 \,||\, P_2)} \tag{12}$$

4. *Invariance to bijections: if f is a bijection and P_1 and P_2 are random variables,*

$$\mathrm{TV}(f(P_1), f(P_2)) = \mathrm{TV}(P_1, P_2) \tag{13}$$

We also occasionally write $\mathrm{TV}(p_1, p_2)$ for probability densities p_1 and p_2 as

$$\mathrm{TV}(p_1, p_2) = \sup_h \left|\int h(x)p_1(x)\,\mathrm{d}x - \int h(x)p_2(x)\,\mathrm{d}x\right| \tag{14}$$

where h is an indicator function of some measurable set A.

Bernstein–von Mises Theorem It turns out that in many typical settings, the prior's influence on the posterior vanishes when the dataset X is large. A basic example of this is the Bernstein–von Mises theorem (van der Vaart 1998), which informally states that under some regularity conditions, the posterior approaches a Gaussian that does not depend on the prior as the size of the dataset increases.

The version of the Bernstein–von Mises theorem we use is from van der Vaart (1998). To state the regularity conditions, we need two definitions:

Definition A.3. *A parametric probability density p_Q is differentiable in quadratic mean at Q_0 if there exists a measurable vector-valued function $\dot{\ell}_{Q_0}$ such that, as $Q \to Q_0$,*

$$\int \left(\sqrt{p_Q(x)} - \sqrt{p_{Q_0}(x)} - \frac{1}{2}(Q - Q_0)^T \dot{\ell}_{Q_0}(x)\sqrt{p_{Q_0}(x)}\right)^2 \mathrm{d}x = o(||Q - Q_0||_2^2). \tag{15}$$

Definition A.4. *A randomised test is a function $\phi\colon \mathcal{X} \to [0, 1]$.*

The interepretation of $\phi(X)$ is the probability of rejecting some null hypothesis after observing data X.

Now we can state the regularity conditions of the theorem:

Condition A.5 (van der Vaart (1998))**.** *For true parameter value Q_0 and observed data X_n:*

1. *The datapoints of X_n are i.i.d.*
2. *The likelihood $p(x|Q)$ for a single datapoint x is differentiable in quadratic mean at Q_0.*
3. *The Fisher information matrix of $p(x|Q)$ is nonsingular at Q_0.*
4. *For every $\beta > 0$, there exists a sequence of randomised tests ϕ_n such that*
$$p(X_n|Q_0)\phi_n(X_n) \to 0, \qquad \sup_{||Q-Q_0||_2 \geq \beta} p(X_n|Q)(1 - \phi_n(X_n)) \to 0. \tag{16}$$
5. *The prior $p(Q)$ is absolutely continuous (as a measure) in a neighbourhood of Q_0 with a continuous positive density at Q_0.*

Now we can state the theorem:

Theorem A.6 (Bernstein–von Mises (van der Vaart 1998))**.** *Let n denote the size of the dataset X_n. Under regularity conditions stated in Condition A.5, for true parameter value Q_0, the posterior $\bar{Q}(X_n) \sim p(Q|X_n)$ satisfies*

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}(X_n) - Q_0), \mathcal{N}(\mu(X_n), \Sigma)\right) \xrightarrow{P} 0 \tag{17}$$

as $n \to \infty$ for some $\mu(X_n)$ and Σ, that do not depend on the prior, where the convergence in probability is over sampling $X_n \sim p(X_n|Q_0)$.

A.2 Differential Privacy and Noise-Aware Synthetic Data

Differential privacy (DP) (Dwork et al. 2006b) quantifies the privacy loss from releasing the results of analysing data. The quantification is done by looking at the output distributions of the analysis algorithm for two datasets that differ in a single data subject (Dwork and Roth 2014):

Definition A.7. *An algorithm $\mathcal{M}$ is (ϵ, δ)-DP if*

$$\Pr(\mathcal{M}(X) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(X') \in S) + \delta \tag{18}$$

for all measurable sets S and all datasets X, X' that differ in one data subject.

The choice of ϵ and δ is a matter of policy (Dwork 2008). One should set $\delta \ll \frac{1}{n}$ for n datapoints, as $\delta \approx \frac{1}{n}$ permits mechanisms that clearly violate privacy (Dwork and Roth 2014).

A common primitive for making an algorithm DP is the *Gaussian mechanism* (Dwork et al. 2006a), which simply adds Gaussian noise to the output of a function:

Definition A.8. *The Gaussian mechanism with noise variance σ_{DP}^2 and function f outputs $f(X) + \mathcal{N}(0, \sigma_{DP}^2 I)$ for input X.*

For a given (ϵ, δ)-bound and function f, the required value for σ_{DP}^2 can be computed tightly using the analytical Gaussian mechanism (Balle and Wang 2018).

Noise-Aware Private Synthetic Data To solve the uncertainty estimation problem for frequentist analyses from DP synthetic data, Räisä et al. (2023) developed a noise-aware algorithm for generating synthetic data called NAPSU-MQ. NAPSU-MQ takes discrete data, summarises it with marginal queries, releases the query values under DP with the Gaussian mechanism, and finally generates multiple synthetic datasets. The downstream analysis is done on each synthetic dataset, and the results are combined using Rubin's rules for synthetic data (Raghunathan et al. 2003; Rubin 1993), which use the multiple analysis results to account for the extra uncertainty coming from the synthetic data generation.

The synthetic data is generated by sampling the posterior predictive distribution

$$p(X^*|\tilde{s}) = \int p(X^*|\theta)p(\theta|\tilde{s})\,\mathrm{d}\theta. \tag{19}$$

The conditioning on $\tilde{s}$ and including the Gaussian mechanism in the model is what makes NAPSU-MQ noise-aware, and allows Rubin's rules to accurately account for the synthetic data generation and DP noise in the downstream analysis.

A.3 Bayesian Inference with Gaussian Models

In this section, we collect well-known results on Bayesian inference of a Gaussian mean. See Gelman et al. (2014) for proofs.

Scaled inverse-chi-square distribution This parameterisation of the inverse gamma distribution is convenient in this setting.

$$\text{Inv-}\chi^2(\nu, s^2) = \text{Inv-Gamma}\left(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2}s^2\right). \tag{20}$$

If $\theta \sim \text{Inv-}\chi^2(\nu, s^2)$, $\theta > 0$,

$$p(\theta) = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} s^\nu \theta^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu s^2}{2\theta}} \tag{21}$$

$$\mathbb{E}(\theta) = \frac{\nu}{\nu - 2}s^2, \quad \nu > 2 \tag{22}$$

$$\mathrm{Var}(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}s^4, \quad \nu > 4. \tag{23}$$

Gaussian Model with Known Variance When the variance of the data is known to be σ_k^2, and only the mean is unknown, the conjugate prior is another Gaussian, and we get the following inference problem:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{24}$$

$$x_i|\mu \sim \mathcal{N}(\mu, \sigma_k^2). \tag{25}$$

The posterior with n datapoints with sample mean $\bar{X}$ is:

$$\mu|X \sim \mathcal{N}(\mu_n, \sigma_n^2) \tag{26}$$

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma_k^2}\bar{X}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}} \tag{27}$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma_k^2}. \tag{28}$$

Gaussian Model with Unknown Variance When the variance of the data is also unknown, the conjugate prior is a inverse-chi-squared for the variance, and Gaussian for the mean, which gives the following inference problem:

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \tag{29}$$

$$\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \tag{30}$$

$$x_i|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2). \tag{31}$$

The joint posterior of μ and σ^2 for n datapoints is:

$$\sigma^2|X \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) \tag{32}$$

$$\mu|\sigma^2, X \sim \mathcal{N}\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right) \tag{33}$$

$$\tag{34}$$

with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{35}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 \tag{36}$$

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{X} \tag{37}$$

$$\kappa_n = \kappa_0 + n \tag{38}$$

$$\nu_n = \nu_0 + n \tag{39}$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{X} - \mu_0)^2. \tag{40}$$

The marginal posterior of μ is

$$\mu | X \sim t_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right). \tag{41}$$

B Missing Proofs

B.1 Consistency Proof

For ease of reference, we repeat Theorem A.6 and Condition 2.2:

Theorem A.6 (Bernstein–von Mises (van der Vaart 1998))**.** *Let n denote the size of the dataset X_n. Under regularity conditions stated in Condition A.5, for true parameter value Q_0, the posterior $\bar{Q}(X_n) \sim p(Q|X_n)$ satisfies*

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}(X_n) - Q_0), \mathcal{N}(\mu(X_n), \Sigma)\right) \xrightarrow{P} 0 \tag{17}$$

as $n \to \infty$ for some $\mu(X_n)$ and Σ, that do not depend on the prior, where the convergence in probability is over sampling $X_n \sim p(X_n|Q_0)$.

Recall that $\bar{Q}_n^+ \sim p(Q|Z, X_n^*)$, and $\bar{Q}_n \sim p(Q|X_n^*)$.

Condition 2.2. *For the observed Z and all Q, there exist distributions D_n such that*

$$\mathrm{TV}\left(\bar{Q}_n^+, D_n\right) \xrightarrow{P} 0 \quad \text{and} \quad \mathrm{TV}\left(\bar{Q}_n, D_n\right) \xrightarrow{P} 0 \tag{7}$$

as $n \to \infty$, where the convergence in probability is over sampling $X_n^* \sim p(X_n^*|Z, Q)$.

Lemma 2.3. *If the assumptions of Theorem A.6 (Condition A.5) hold for the downstream analysis for all Q_0, and the following assumptions:*

(1) Z and X^* are conditionally independent given Q; and

(2) $p(Z|Q) > 0$ for all Q,

hold, then Condition 2.2 holds.

Proof. Under Assumption (1)

$$p(Q|Z, X_n^*) \propto p(X_n^*|Q)p(Z|Q)p(Q) \tag{42}$$

so we can view both $p(Q|Z, X_n^*)$ and $p(Q|X_n^*)$ as the posteriors for the same Bayesian inference problem with observed data X_n^*, and priors $p(Q|Z) \propto p(Z|Q)p(Q)$ and $p(Q)$, respectively. Due to Condition A.5 (5) and Assumption (2), $p(Q|Z)$ has an everywhere positive density. Recall that $\bar{Q}_n^+ \sim p(Q|Z, X_n^*)$ and $\bar{Q}_n \sim p(Q|X_n^*)$. Now, Theorem A.6 gives

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}_n^+ - Q_0), \mathcal{N}(\mu_n, \Sigma)\right) \xrightarrow{P} 0 \tag{43}$$

and

$$\mathrm{TV}\left(\sqrt{n}(\bar{Q}_n - Q_0), \mathcal{N}(\mu_n, \Sigma)\right) \xrightarrow{P} 0 \tag{44}$$

as $n \to \infty$, where μ_n, Σ are equal in the two cases because they do not depend on the prior. The probability is over $X_n^* \sim p(X_n^*|Q_0)$. Because of Assumption (1), $p(X_n^*|Q_0) = p(X_n^*|Z, Q_0)$, so the convergence also holds with probability over $X_n^* \sim p(X_n^*|Z, Q_0)$. These hold for any Q_0. Because the function $f_n(q) = \sqrt{n}(q - Q_0)$ is a bijection and total variation distance is invariant to bijections, Condition 2.2 holds with D_n being the pushforward distribution $D_n = f_n^{-1} \circ \mathcal{N}(\mu_n, \Sigma)$, with the Q of Condition 2.2 being Q_0. Note that D_n is allowed to depend on Q in Condition 2.2 due to the order of quantifiers. □

Lemma B.1. *Under Condition 2.2,*

$$\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) \xrightarrow{P} 0 \tag{45}$$

as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z)$.

Proof. Total variation distance is a metric, so

$$\mathrm{TV}\left(\bar{Q}_n^+, \bar{Q}_n\right) \leq \mathrm{TV}\left(\bar{Q}_n^+, D_n\right) + \mathrm{TV}\left(\bar{Q}_n, D_n\right) \tag{46}$$

so by Condition 2.2

$$\mathrm{TV}\left(\bar{Q}_n^+, \bar{Q}_n\right) \xrightarrow{P} 0 \tag{47}$$

as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z, Q)$.

It remains to show (47) with the probability over $X_n^* \sim p(X_n^*|Z)$ instead of $X_n^* \sim p(X_n^*|Z, Q)$. With $X_n^* \sim p(X_n^*|Z)$, for any $\epsilon > 0$,

$$\Pr_{X_n^*|Z}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = \int \Pr_{X_n^*|Z,Q}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) p(Q|Z)\, \mathrm{d}Q \tag{48}$$

(47) holds for any Q, so

$$\lim_{n\to\infty} \Pr_{X_n^*|Z,Q}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = 0 \tag{49}$$

The dominated convergence theorem then implies that

$$\lim_{n\to\infty} \Pr_{X_n^*|Z}(\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) > \epsilon) = 0 \tag{50}$$

so

$$\mathrm{TV}(\bar{Q}_n^+, \bar{Q}_n) \xrightarrow{P} 0 \tag{51}$$

as $n \to \infty$, with the probability over $X_n^* \sim p(X_n^*|Z)$. □

Lemma B.2. *Let $y_n \sim U_n$ be an arbitrary sequence of continuous random variables and let $S(y_n)$, $T(y_n)$ be continuous random variables that depend on y_n. Let the density functions of $S(y_n)$, $T(y_n)$ and U_n be $f_{S(y_n)}$, $f_{T(y_n)}$ and f_{U_n}, respectively. If*

$$\mathrm{TV}(S(y_n), T(y_n)) \xrightarrow{P} 0 \tag{52}$$

as $n \to \infty$, where the probability is over $y_n \sim U_n$, then

$$\mathrm{TV}\left(\int f_{S(y_n)}(x) f_{U_n}(y_n)\, \mathrm{d}y_n, \int f_{T(y_n)}(x) f_{U_n}(y_n)\, \mathrm{d}y_n\right) \to 0 \tag{53}$$

as $n \to \infty$.

Proof. Let h be an indicator function of x over any measurable set and let $\epsilon > 0$. Then

$$\left|\int h(x) \int f_{S(y_n)}(x) f_{U_n}(y_n)\, \mathrm{d}y_n\, \mathrm{d}x - \int h(x) \int f_{T(y_n)}(x) f_{U_n}(y_n)\, \mathrm{d}y_n\, \mathrm{d}x\right| \tag{54}$$

$$= \left|\int h(x) \int f_{U_n}(y_n) \big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\, \mathrm{d}y_n\, \mathrm{d}x\right| \tag{55}$$

$$= \left|\int f_{U_n}(y_n) \int h(x) \big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\, \mathrm{d}x\, \mathrm{d}y_n\right| \tag{56}$$

$$\leq \int f_{U_n}(y_n) \left|\int h(x) \big(f_{S(y_n)}(x) - f_{T(y_n)}(x)\big)\, \mathrm{d}x\right| \mathrm{d}y_n \tag{57}$$

$$= \int f_{U_n}(y_n) \left|\int h(x) f_{S(y_n)}(x)\, \mathrm{d}x - \int h(x) f_{T(y_n)}(x)\, \mathrm{d}x\right| \mathrm{d}y_n \tag{58}$$

Because $\mathrm{TV}(S(y_n), T(y_n)) \xrightarrow{P} 0$, for large enough n, there is a set Y_n with $\mathrm{TV}(S(y_n), T(y_n)) < \frac{\epsilon}{2}$ for all $y_n \in Y_n$, and $\Pr(y_n \in Y_n^C) < \frac{\epsilon}{2}$. As

$$\mathrm{TV}(S(y_n), T(y_n)) = \sup_h \left| \int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\,\mathrm{d}x \right| \leq 1 \tag{59}$$

now

$$\int f_{U_n}(y_n) \left| \int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\,\mathrm{d}x \right| \mathrm{d}y_n \tag{60}$$

$$\begin{aligned} &= \int_{Y_n} f_{U_n}(y_n) \left| \int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\,\mathrm{d}x \right| \mathrm{d}y_n \\ &+ \int_{Y_n^C} f_{U_n}(y_n) \left| \int h(x) f_{S(y_n)}(x)\,\mathrm{d}x - \int h(x) f_{T(y_n)}(x)\,\mathrm{d}x \right| \mathrm{d}y_n \end{aligned} \tag{61}$$

$$< \int_{Y_n} f_{U_n}(y_n) \frac{\epsilon}{2}\,\mathrm{d}y_n + \int_{Y_n^C} f_{U_n}(y_n)\,\mathrm{d}y_n \tag{62}$$

$$< \frac{\epsilon}{2} + \frac{\epsilon}{2} \tag{63}$$

$$= \epsilon \tag{64}$$

for large enough n. Now

$$\mathrm{TV}\left(\int f_{S(y_n)}(x) f_{U_n}(y_n)\,\mathrm{d}y, \int f_{T(y_n)}(x) f_{U_n}(y_n)\,\mathrm{d}y_n \right) \tag{65}$$

$$= \sup_h \left| \int h(x) \int f_{S(y_n)}(x) f_{U_n}(y_n)\,\mathrm{d}y_n\,\mathrm{d}x - \int h(x) \int f_{T(y_n)}(x) f_{U_n}(y_n)\,\mathrm{d}y_n\,\mathrm{d}x \right| \tag{66}$$

$$< \epsilon \tag{67}$$

for any $\epsilon > 0$ with large enough n. □

Theorem 2.4. *Under congeniality and Condition 2.2,* $\mathrm{TV}\,(p(Q|Z), \bar{p}_n(Q)) \to 0$ *as* $n \to \infty$.

Proof. The claim follows from Lemma B.2 with $y_n = X_n^*$, $U_n = p(X_n^*|Z)$, $S(y_n) \sim p(Q|X_n^*)$ and $T(y_n) \sim p(Q|Z, X_n^*)$. These meet the condition for Lemma B.2 due to Lemma B.1. □

C Gaussian Examples

C.1 Non-private Gaussian Mean Estimation

Our first Gaussian example is very simple: the analyst infers the mean μ of a Gaussian distribution with known variance from synthetic data that has been generated from the same model. The posteriors for this setting can be found in Supplemental Section A.3. To differentiate the variables for the analyst and data provider, we use bars for the data provider (like $\bar{\sigma}_0^2$) and hats for the analyst (like $\hat{\sigma}_0^2$).

When the synthetic data is generated from the known variance model with known variance $\bar{\sigma}_k^2$, we sample from the posterior predictive $p(X^*|X)$ as

$$\bar{\mu}|X \sim \mathcal{N}(\bar{\mu}_{n_X}, \bar{\sigma}_{n_X}^2), \quad X^*|\bar{\mu} \sim \mathcal{N}^{n_{X^*}}(\bar{\mu}, \bar{\sigma}_k^2) \tag{68}$$

$$\bar{\mu}_{n_X} = \frac{\frac{1}{\bar{\sigma}_0{}^2}\bar{\mu}_0 + \frac{n_X}{\bar{\sigma}_k^2}\bar{X}}{\frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2}}, \quad \frac{1}{\bar{\sigma}_{n_X}^2} = \frac{1}{\bar{\sigma}_0^2} + \frac{n_X}{\bar{\sigma}_k^2}. \tag{69}$$

$\mathcal{N}^{n_{X^*}}$ denotes a Gaussian distribution over n_{X^*} i.i.d. samples.

When downstream analysis is the model with known variance $\hat{\sigma}_k^2$, we have

$$\hat{\mu}|X^* \sim \mathcal{N}(\hat{\mu}_{n_{X^*}}, \hat{\sigma}_{n_{X^*}}^2), \quad \hat{\mu}_{n_{X^*}} = \frac{\frac{1}{\hat{\sigma}_0{}^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^*}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}}, \quad \frac{1}{\hat{\sigma}_{n_{X^*}}^2} = \frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}. \tag{70}$$

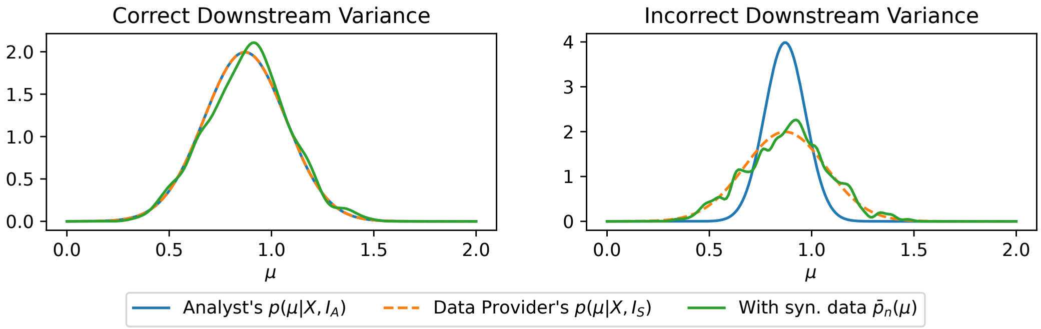


Figure S1: Simulation results for the Gaussian mean estimation example, showing that the mixture of posteriors from synthetic data in green converges. In the left panel, both the analyst and data provider have the correct known variance. The blue and orange lines overlap, as both parties have the same $p(\mu|X)$. On the right, the analyst's known variance is too small ($\hat{\sigma}_k^2 = \frac{1}{4}\bar{\sigma}_k^2$), so congeniality is not met, but the mixture of posteriors from synthetic data, $\bar{p}_n(\mu)$, still converges to the data provider's posterior. In both panels, $m = 400$ and $\frac{n_{X^*}}{n_X} = 20$.

Next we check where the mean and variance of $\mu^* \sim \bar{p}_n(\mu)$ converge when $n_{X^*} \to \infty$:

$$\begin{aligned}\mathbb{E}(\mu^*) &= \mathbb{E}(\mathbb{E}(\mu^*|X^*)) = \mathbb{E}\left(\hat{\mu}_{n_{X^*}}\right) && (71)\\ &= \mathbb{E}\left(\frac{\frac{1}{\hat{\sigma}_0^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^*}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}}\right) && (72)\\ &= \frac{\frac{1}{\hat{\sigma}_0^2}\hat{\mu}_0 + \frac{n_{X^*}}{\hat{\sigma}_k^2}\mathbb{E}(\bar{X}^*)}{\frac{1}{\hat{\sigma}_0^2} + \frac{n_{X^*}}{\hat{\sigma}_k^2}} && (73)\\ &\to \mathbb{E}(X^*) = \bar{\mu}_{n_X} && (74)\end{aligned}$$

as $n_{X^*} \to \infty$.

$$\begin{aligned}\mathrm{Var}(\mu^*) &= \mathbb{E}(\mathrm{Var}(\mu^*|X^*)) + \mathrm{Var}(\mathbb{E}(\mu^*|X^*)) && (75)\\ &= \mathbb{E}(\hat{\sigma}^2_{n_{X^*}}) + \mathrm{Var}(\hat{\mu}_{n_{X^*}}) && (76)\end{aligned}$$

$$\mathbb{E}(\hat{\sigma}^2_{n_{X^*}}) = \mathbb{E}\left(\frac{1}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) \to 0, n_{X^*} \to \infty \qquad (77)$$

$$\mathrm{Var}\left(\hat{\mu}_{n_{X^*}}\right) = \mathrm{Var}\left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}\bar{X}^* + \frac{\hat{\mu}_0}{\hat{\sigma}_0^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right) = \left(\frac{\frac{n_{X^*}}{\hat{\sigma}_k^2}}{\frac{n_{X^*}}{\hat{\sigma}_k^2} + \frac{1}{\hat{\sigma}_0^2}}\right)^2 \mathrm{Var}(\bar{X}^*) \qquad (78)$$

$$\mathrm{Var}(\bar{X}^*) = \mathbb{E}(\mathrm{Var}(\bar{X}^*|\bar{\mu})) + \mathrm{Var}(\mathbb{E}(\bar{X}^*|\bar{\mu})) = \frac{1}{n_{X^*}}\mathbb{E}(\mathrm{Var}(x_i^*)) + \mathrm{Var}(\bar{\mu}) \to \mathrm{Var}(\bar{\mu}) = \bar{\sigma}^2_{n_X} \qquad (79)$$

as $n_{X^*} \to \infty$.

Putting these together,

$$\begin{aligned}\mathbb{E}(\mu^*) &\to \bar{\mu}_{n_X} && (80)\\ \mathrm{Var}(\mu^*) &\to \bar{\sigma}^2_{n_X} && (81)\end{aligned}$$

as $n_{X^*} \to \infty$.

μ^* also has a Gaussian distribution, which we will show next. In

$$\mu^* \sim \int p(\mu|X_n^*)p(X_n^*|X)dX^*, \qquad (82)$$

both $p(\mu|X_n^*) = \mathcal{N}(\hat{\mu}_{n_{X^*}}, \hat{\sigma}^2_{n_{X^*}})$ and $p(X_n^*|X)$ are Gaussian. $\hat{\mu}_{n_{X^*}}$ is a linear function of X_n^*, so $p(\hat{\mu}_{n_{X^*}}|X)$ is also Gaussian. $\hat{\sigma}^2_{n_{X^*}}$ does not depend on X^*, so μ^* is the sum of a random variable with distribution $\mathcal{N}(0, \hat{\sigma}^2_{n_{X^*}})$ and $\hat{\mu}_{n_{X^*}}$, which is also Gaussian, meaning that μ^* is Gaussian.

We test the theory with a numerical simulation in Figure S1. We generated the real data X of size $n_X = 100$ by i.i.d. sampling from $\mathcal{N}(1, 4)$. Both the analyst and data provider use $\mathcal{N}(0, 10^2)$ as the prior. The data provider uses the correct known variance ($\bar{\sigma}_k^2 = 4$), and the analyst either uses the correct known variance ($\hat{\sigma}_k^2 = 4$), or a too small known variance ($\hat{\sigma}_k^2 = 1$), which is an example of uncongeniality.

In the congenial case in the left panel of Figure S1, both parties have the same posterior given the real data X, and the mixture of posteriors from synthetic data is very close to that. In the uncongenial case in the right panel, where the analyst underestimates the variance, the parties have different posteriors given X, but the mixture of synthetic data posteriors is still close to the data provider's posterior.

In Figure 3, we examine the convergence of the mixture of posteriors from synthetic data under congeniality. We see that setting $n_{X^*} = n_X$ is not enough, as the mixture of posteriors is significantly wider than the analyst's posterior. The synthetic dataset needs to be larger than the original, with $n_{X^*} = 5n_X$ already giving a decent approximation and $n_{X^*} = 20n_X$ a rather good one. In Figure S2 in the Supplement, we also examine the effect of m on the mixture of synthetic data posteriors, and see that m must also be sufficiently large, otherwise the method produces very jagged posteriors.

The plots of $p(\mu^*)$ in Figures S1, 3, S2 and are density functions of a mixture of Gaussians, where each mixture component is the Gaussian posterior distribution from one synthetic dataset.

C.2 Gaussian with Known Mean, Unknown Variance

To asses the effects of uncongeniality when the downstream posterior is not Gaussian, we look at Bayesian estimation of the variance of a Gaussian, with known mean. In this case, the data provider's prior is

$$\bar{\sigma}^2 \sim \text{Inv-}\chi^2(\bar{\nu}_0, \bar{\sigma}_0^2) \tag{83}$$

and their known mean is $\bar{\mu}_k$. The synthetic data is generated from

$$\bar{v} = \frac{1}{n_X}\sum_{i=1}^{n_X}(x_i - \bar{\mu}_k)^2 \tag{84}$$

$$\bar{\nu}_{n_X} = \bar{\nu}_0 + n_X \tag{85}$$

$$\bar{\sigma}^2_{n_X} = \frac{\bar{\nu}_0\bar{\sigma}_0^2 + n_X\bar{v}}{\bar{\nu}_0 + n_X} \tag{86}$$

$$\bar{\sigma}^2|X \sim \text{Inv-}\chi^2(\bar{\nu}_{n_X}, \bar{\sigma}^2_{n_X}) \tag{87}$$

$$x_i^*|\bar{\sigma}^2 \sim \mathcal{N}(\bar{\mu}_k, \bar{\sigma}^2) \tag{88}$$

The analyst's prior is

$$\hat{\sigma}^2 \sim \text{Inv-}\chi^2(\hat{\nu}_0, \hat{\sigma}_0^2) \tag{89}$$

their known mean is $\hat{\mu}_k$, and the downstream posterior is

$$\hat{v} = \frac{1}{n_{X^*}}\sum_{i=1}^{n_{X^*}}(x_i^* - \hat{\mu}_k)^2 \tag{90}$$

$$\hat{\nu}_{n_{X^*}} = \hat{\nu}_0 + n_{X^*} \tag{91}$$

$$\hat{\sigma}^2_{n_{X^*}} = \frac{\hat{\nu}_0\hat{\sigma}_0^2 + n_{X^*}\hat{v}}{\hat{\nu}_0 + n_{X^*}} \tag{92}$$

$$\hat{\sigma}^2|X^* \sim \text{Inv-}\chi^2(\hat{\nu}_{n_{X^*}}, \hat{\sigma}^2_{n_{X^*}}) \tag{93}$$

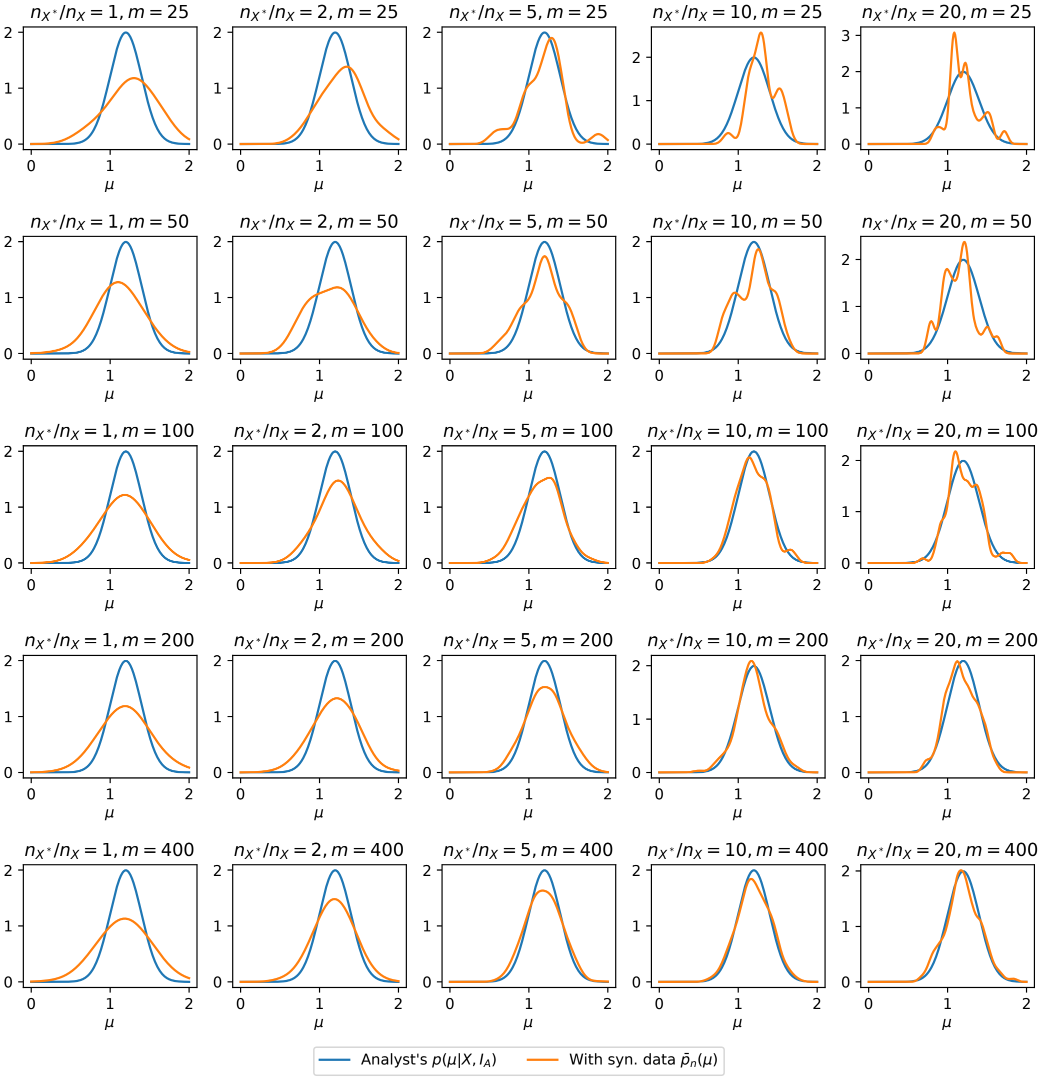


Figure S2: Convergence of the mixture of synthetic data posteriors (in orange) with different values of m and n_{X^*} in Gaussian mean estimation with known variance.

Denoting a sample from the mixture of synthetic data posteriors as σ_*^2, we have

$$\begin{aligned}
\mathbb{E}(\sigma_*^2) &= \mathbb{E}(\mathbb{E}(\sigma_*^2|X^*)) && (94)\\
&= \mathbb{E}\left(\frac{\hat{\nu}_{n_{X^*}}}{\hat{\nu}_{n_{X^*}}-2}\hat{\sigma}^2_{n_{X^*}}\right) && (95)\\
&= \frac{\hat{\nu}_0+n_{X^*}}{\hat{\nu}_0+n_{X^*}-2}\mathbb{E}\left(\hat{\sigma}^2_{n_{X^*}}\right) && (96)\\
&= \frac{\hat{\nu}_0+n_{X^*}}{\hat{\nu}_0+n_{X^*}-2}\frac{\hat{\nu}_0\hat{\sigma}_0^2+n_{X^*}\mathbb{E}(\hat{v})}{\hat{\nu}_0+n_{X^*}} && (97)\\
&= \frac{\hat{\nu}_0\hat{\sigma}_0^2+n_{X^*}\mathbb{E}(\hat{v})}{\hat{\nu}_0+n_{X^*}-2} && (98)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(\hat{v}) &= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}((x_i^* - \hat{\mu}_k)^2) && (99)\\
&= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}((x_i^*)^2 - 2x_i^*\hat{\mu}_k + \hat{\mu}_k^2) && (100)\\
&= \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \left(\mathbb{E}((x_i^*)^2) - 2\bar{\mu}_k\hat{\mu}_k + \hat{\mu}_k^2\right) && (101)\\
&= \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} \mathbb{E}((x_i^*)^2) && (102)\\
&= \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} (\mathbb{E}(x_i^*)^2 + \mathrm{Var}(x_i^*)) && (103)\\
&= \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \frac{1}{n_{X^*}} \sum_{i=1}^{n_{X^*}} (\bar{\mu}_k^2 + \mathrm{Var}(x_i^*)) && (104)\\
&= \bar{\mu}_k^2 + \hat{\mu}_k^2 - 2\bar{\mu}_k\hat{\mu}_k + \mathrm{Var}(x_i^*) && (105)\\
&= (\bar{\mu}_k - \hat{\mu}_k)^2 + \mathrm{Var}(x_i^*) && (106)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(x_i^*) &= \mathrm{Var}(\mathbb{E}(x_i^*|\bar{\sigma}^2)) + \mathbb{E}(\mathrm{Var}(x_i^*|\bar{\sigma}^2)) && (107)\\
&= \mathbb{E}(\bar{\sigma}^2) && (108)
\end{aligned}$$

Putting these together,

$$\mathbb{E}(\sigma_*^2) \to \mathbb{E}(\bar{\sigma}^2) + (\bar{\mu}_k - \hat{\mu}_k)^2, \tag{109}$$

as $n_{X^*} \to \infty$, so mixing the downstream posteriors can only recover the data provider's posterior when both parties have equal known means.

We verify this with a simulation shown in Figure S3. Both the data provider and analyst use the Gaussian with unknown variance, known mean as their model. Otherwise, the setting is identical with the other Gaussian examples. When both parties have the correct known mean, $\bar{p}_n(\sigma^2)$ converges as expected, but when the analyst has an incorrect known mean, $\bar{p}_n(\sigma^2)$ converges to neither party's posterior. However, after applying the mean correction from (109), $\bar{p}_n(\sigma^2)$ appears to have the same variance and shape as the data provider's posterior.

D Toy Data Logistic Regression Details

Setting Details In the toy data setting of Räisä et al. (2023), the real dataset consists of $n_X = 2000$ i.i.d. samples of three binary variables. The first two variables are sampled with independent coinflips, and the third is sampled from logistic regression on the other two, with coefficients $(1, 0)$. The prior for the downstream logistic regression is $\mathcal{N}(0, 10I)$.

We generate synthetic data with the NAPSU-MQ algorithm (Räisä et al. 2023), instructing the algorithm to generate m synthetic datasets of size n_{X^*}. DP-GLM doesn't use synthetic data, so we run it directly on the real data. For the privacy bounds, we vary ϵ, and set $\delta = n_X^{-2} = 2.5 \cdot 10^{-7}$, which is how DP mechanisms are typically evaluated.

Hyperparameters For NAPSU-MQ, we use the hyperparameters of Räisä et al. (2023), except we used NUTS (Hoffman and Gelman 2014) with 200 warmup samples and 500 kept samples for $\epsilon \in \{0.5, 1\}$, and 1500 kept samples for $\epsilon = 0.1$, as the posterior sampling algorithm. The prior is $\mathcal{N}(0, 10^2 I)$, and the marginal queries are the full set of 3-way marginals of all three variables.

The hyperparameters of DP-GLM are the L_2-norm upper bound R for the covariates of the logistic regression, a coefficient norm upper bound s, and the parameters of the posterior sampling algorithm

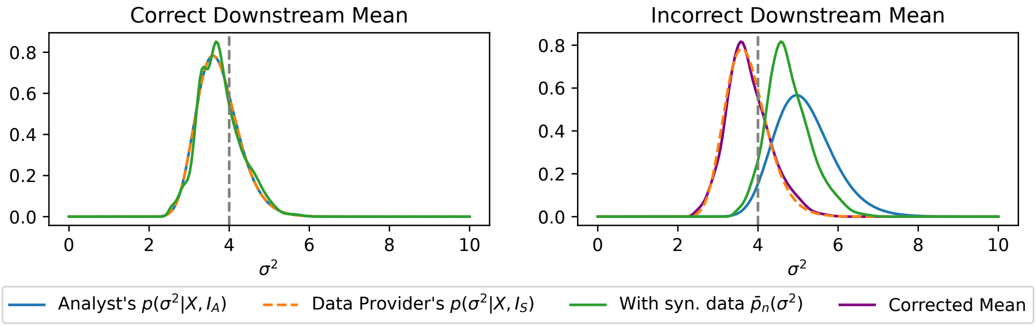


Figure S3: Posteriors from estimating a Gaussian variance σ^2 with known mean. On the right, both the data provider and the analyst use the correct known mean, so $\bar{p}_n(\sigma^2)$ converges as expected by our theory. On the right, the analyst mean is incorrect, so the model is not congenial. In this case, $\bar{p}_n(\sigma^2)$ does not converge to either to analyst's or the data provider's posterior. After correcting the mean of $\bar{p}_n(\sigma^2)$ as in (109), it appears to have the same variance and shape as the data provider's posterior. The gray line shows the true parameter value. In both panels, $m = 400$ and $\frac{n_{X^*}}{n_X} = 20$.

DP-GLM uses. We set $R = \sqrt{2}$ so that the covariates do not get clipped, and set $s = 5$ after some preliminary runs. The posterior sampling algorithm is NUTS (Hoffman and Gelman 2014) with 1000 warmup iterations and 1000 kept samples from 4 parallel chains.

Plotting Details The plotted density of DP-GLM in Figure 2 is a kernel density estimate from the posterior samples DP-GLM returns. The non-DP density is a Laplace approximation. Both synthetic data methods use Laplace approximations in the downstream analysis, so their posteriors are mixtures of these Laplace approximations for each synthetic dataset. This was also used in Figure S5.

Sampling the exact posterior In order to sample the exact posterior $p(Q|\tilde{s})$, we use another decomposition:

$$p(Q|\tilde{s}) = \int p(Q|\tilde{s}, X)p(X|\tilde{s})\,\mathrm{d}X = \int p(Q|X)p(X|\tilde{s})\,\mathrm{d}X, \tag{110}$$

where $p(Q|\tilde{s}, X) = p(Q|X)$ due to the independencies of the graphical model in Figure 1. It remains to sample $p(X|\tilde{s})$. This is not tractable in general, but is possible in the toy data setting due to using the full set of 3-way marginals that covers all possible values of a datapoint, and the simplicity of the toy data.

We can decompose

$$p(X|\tilde{s}) = \int p(s|\tilde{s})p(X|s)\,\mathrm{d}\theta\,\mathrm{d}X = \int p(X|s)\int p(s,\theta|\tilde{s})\,\mathrm{d}\theta\,\mathrm{d}X, \tag{111}$$

so we can sample $(s, \theta) \sim p(s, \theta|s)$ and then sample $X \sim p(X|s)$ to obtain a sample from $p(X|\tilde{s})$. Due to the simplicity of the toy data setting, sampling both $p(s, \theta|s)$ and $p(X|s)$ is possible.

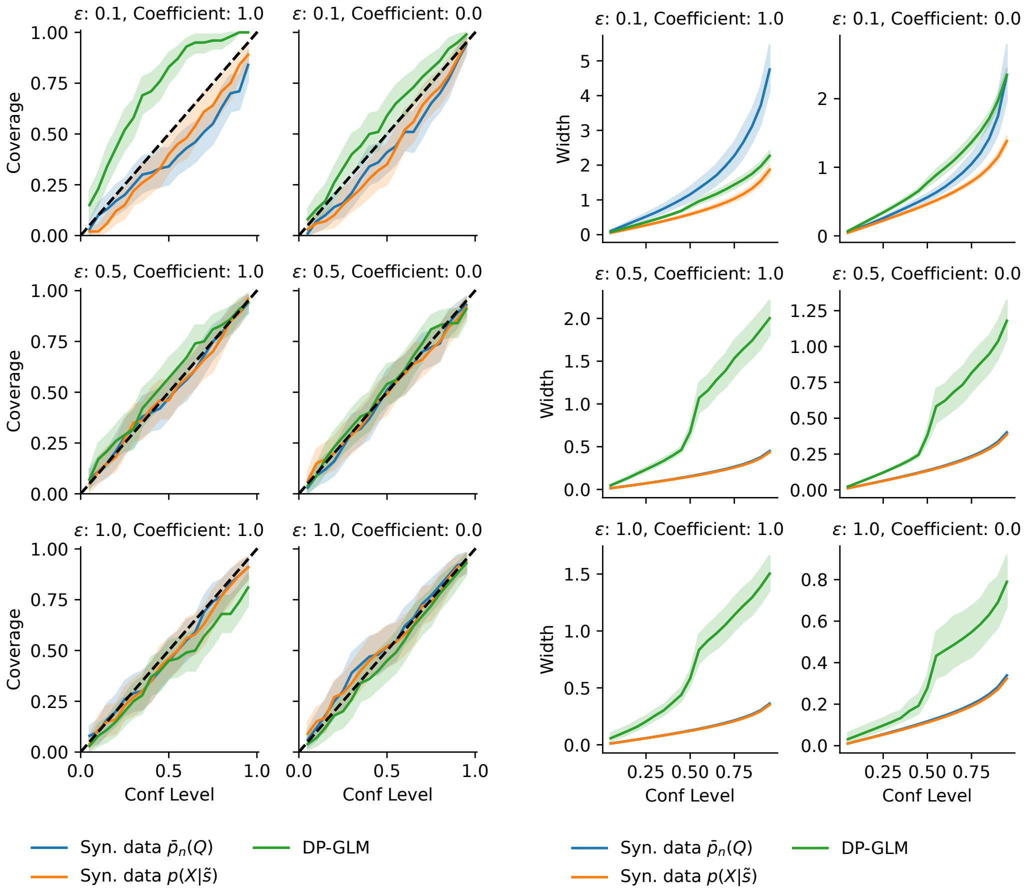
NAPSU-MQ uses the following Bayesian inference problem:

$$\theta \sim \text{Prior} \tag{112}$$
$$X \sim \text{MED}_\theta^n \tag{113}$$
$$s = a(X) \tag{114}$$
$$\tilde{s} \sim \mathcal{N}(s, \sigma_{DP}^2), \tag{115}$$

where a are the marginal queries, σ_{DP}^2 is the noise variance of the Gaussian mechanism, and MED_θ^n is the maximum entropy distribution (Räisä et al. 2023) with point probability

$$p(x) = \exp(\theta^T a(x) - \theta_0(\theta)), \tag{116}$$

where θ_0 is the log-normalising constant.



(a)

(b)

Figure S4: (a) Coverages of credible intervals in the toy data experiment. The mixture of synthetic data posteriors is accurate, except with $\epsilon = 0.1$, where it may not have converged yet. (b) Widths of credible intervals in the toy data experiment. DP-GLM produces much wider intervals than other methods, except with $\epsilon = 0.1$.

In the toy data setting, a is the full set of 3-way marginals for all of the 3 variables. In other words, $a(x)$ is the one-hot coding of x, so $s = a(X)$ is a vector of counts of how many times each of the 8 possible values is repeated in X. This means that sampling $p(X|s)$ is simple:

1. For each possible value of a datapoint, find the corresponding count from s, and repeat that datapoint according the that count.
2. Shuffle the datapoints to a random order.

As the downstream analysis $p(Q|X)$ doesn't depend on the order of the datapoints, the second step is not actually needed.

To sample $p(s, \theta|\tilde{s})$, we use a Metropolis-within-Gibbs sampler (Gilks et al. 1995) that sequentially updates s and θ while keeping the other fixed. The proposal for θ is obtained from Hamiltonian Monte Carlo (HMC) (Duane et al. 1987; Neal 2011). The proposal for s is obtained by repeatedly choosing a random index in s to increment and another to decrement. It is possible to obtain negative values in s from this proposal, but those will always be rejected by the acceptance test, as the likelihood for them is 0.

To initialise the sampler, we pick an initial value for θ from a Gaussian distribution, and pick the initial s by rounding $\tilde{s}$ to integer values, changing the rounded values such that they sum to n while ensuring that all values are non-negative.

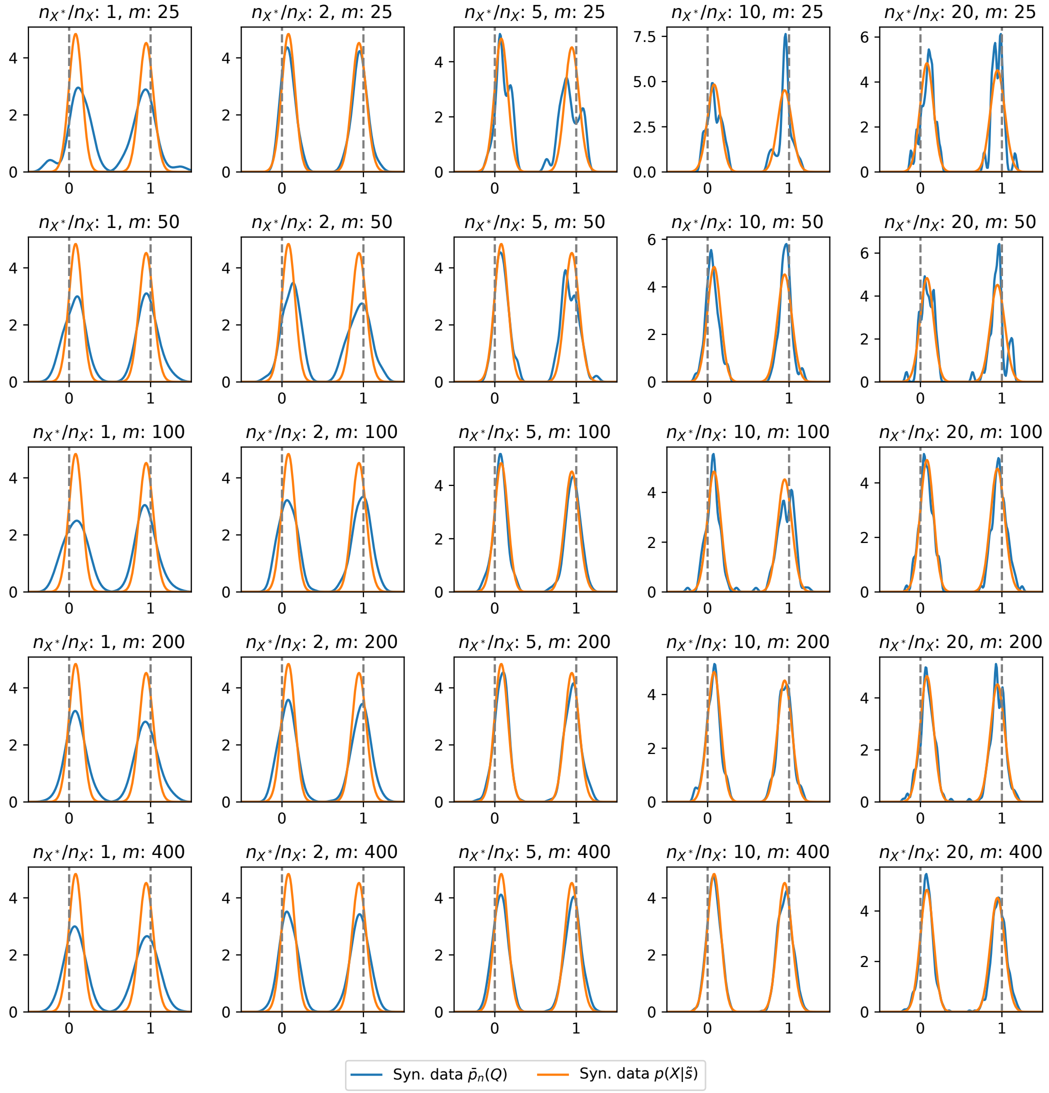


Figure S5: Convergence of the mixture of synthetic data posteriors (in blue) with different values of m and n_{X^*} in the toy data logistic regression experiment.

The step size for the HMC we used is 0.05, and the number of steps is 20. In the s proposal, we repeat the combination of an increment and a decrement 30 times. We take 20000 samples in total from 4 parallel chains, and drop the first 20% as warmup samples.

The method described in this section is similar to the noise-aware Bayesian inference method of Ju et al. (2022). The difference between the two is that Ju et al. (2022) use X instead of s as the auxiliary variable, and they sample the X proposals from the model, changing one datapoint at a time. This makes their algorithm more generalisable.

E Finite Number of Synthetic Datasets

In practice, we only have a finite number of synthetic datasets, so we must further approximate

$$p(Q|Z) \approx \int p(Q|X_n^*)p(X_n^*|Z)\,\mathrm{d}X_n^* \approx \frac{1}{m}\sum_{i=1}^{m} p(Q|X_n^* = X_{i,n}^*), \tag{117}$$

with $X_{i,n}^* \sim p(X_n^*|Z)$.

Total variation distance is a metric, so

$$\begin{aligned}
&\mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^*), p(Q|Z)\right) \\
&\leq \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^*), \bar{p}_n(Q)\right) + \mathrm{TV}\left(\bar{p}_n(Q), p(Q|Z)\right).
\end{aligned} \tag{118}$$

Theorem 2.4 gives

$$\lim_{n\to\infty} \mathrm{TV}\left(\bar{p}_n(Q), p(Q|Z)\right) = 0. \tag{119}$$

If, for all n, $p(Q|X_n^*)$ is continuous for all X_n^*, $p(Q|X_n^*) \leq h_n(X_n^*)$ for an integrable function $h_n(X_n^*)$, and $\mathcal{Q} \subset \mathbb{R}^d$ is compact, the uniform law of large numbers (Jennrich 1969, Theorem 2) gives

$$\sup_{Q\in\mathcal{Q}} \left| \frac{1}{m}\sum_{i=1}^{m} p(Q|X_n^* = X_{i,n}^*) - \bar{p}_n(Q) \right| \to 0 \tag{120}$$

almost surely as $m \to \infty$. If $\mathcal{Q} = \mathbb{R}^d$, we can represent $\mathcal{Q}$ as a countable union of compact sets $\mathcal{Q}_k$, apply the uniform law of large numbers on each $\mathcal{Q}_k$, and use the union bound to obtain (120) for $\mathcal{Q}$. A similar decomposition of $\mathcal{Q}$ can be done for many other constrained parameter sets encountered in practice.

This implies (van der Vaart 1998, Corollary 2.30)

$$\lim_{m\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^*), \bar{p}_n(Q)\right) = 0 \tag{121}$$

for almost all $X_{i,n}^*$, so

$$\lim_{n\to\infty}\lim_{m\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^*), p(Q|Z)\right) = 0 \tag{122}$$

almost surely when $X_{i,n}^* \sim p(X_n^*|Z)$.

Based on the experiment in Figure S2, it looks like

$$\lim_{m\to\infty}\lim_{n\to\infty} \mathrm{TV}\left(\frac{1}{m}\sum_{i=1}^{m} p(Q|X_{i,n}^*), p(Q|Z)\right) \neq 0 \tag{123}$$

because the distributions $p(Q|X_{i,n}^*)$ become narrower as n increases, so a fixed number of them is not enough to cover $p(Q|Z)$. This means that in practice, the number of synthetic datasets should be increased along with the size of the synthetic datasets.

F Relation to Missing Data Imputation

In the missing data setting, only a part X_{obs} of the complete dataset X is observed, while a part X_{mis} is missing (Rubin 1987). To facilitate downstream analysis, the missing data is imputed by sampling $X_{mis} \sim p(X_{mis}|X_{obs}, I_I)$. Analogously with synthetic data, I_I represents the imputer's background knowledge.

Like with synthetic data, we have the decomposition (Gelman et al. 2014)

$$p(Q|X_{obs}, I_A) = \int p(Q|X_{obs}, X_{mis}, I_A) p(X_{mis}|X_{obs}, I_A)\,\mathrm{d}X_{mis}. \tag{124}$$

If the analyst's and imputer's models are congenial in the sense that

$$p(Q|X_{obs}, I_A) = p(Q|X_{obs}, I_I) \tag{125}$$

and

$$p(Q|X, I_A) = p(Q|X, I_I) \tag{126}$$

for any complete dataset X, then

$$\begin{aligned} p(Q|X_{obs}, I_A) = p(Q|X_{obs}, I_I) &= \int p(Q|X_{obs}, X_{mis}, I_I) p(X_{mis}|X_{obs}, I_I) \,\mathrm{d}X_{mis} \\ &= \int p(Q|X_{obs}, X_{mis}, I_A) p(X_{mis}|X_{obs}, I_I) \,\mathrm{d}X_{mis}, \end{aligned} \tag{127}$$

so sampling $p(Q|X_{obs}, I_A)$ can be done by sampling $X_{mis} \sim p(X_{mis}|X_{obs}, I_I)$ multiple times, sampling $p(Q|X_{obs}, X_{mis}, I_A)$ for each X_{mis}, and combining the samples. Unlike with synthetic data, where sampling $p(Q|X, X^*, I_A)$ would require the original data and defeat the purpose of using synthetic data, in the missing data setting, sampling $p(Q|X_{obs}, X_{mis}, I_A)$ is simply the analysis for a complete dataset, so generating large imputed datasets is not required.