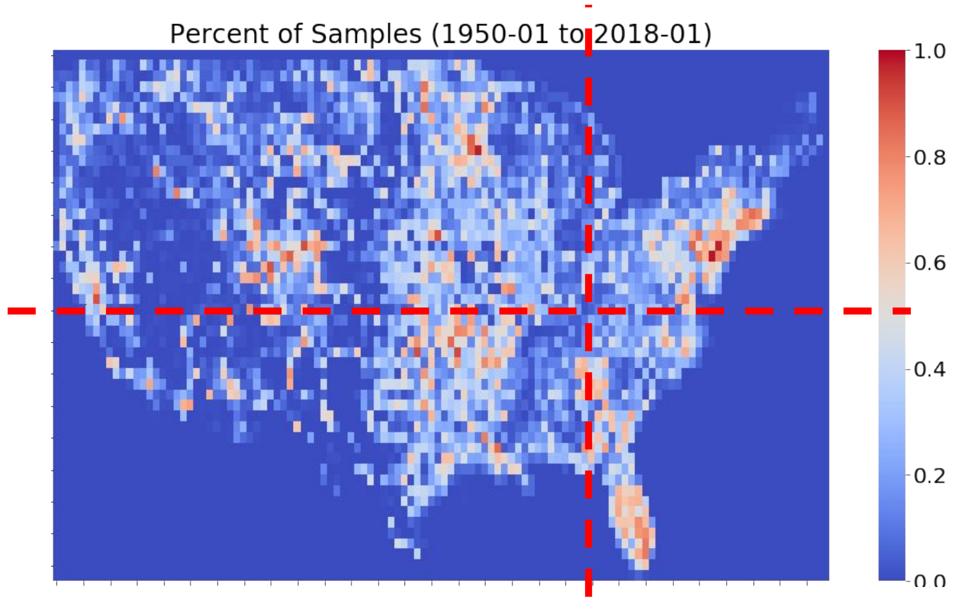
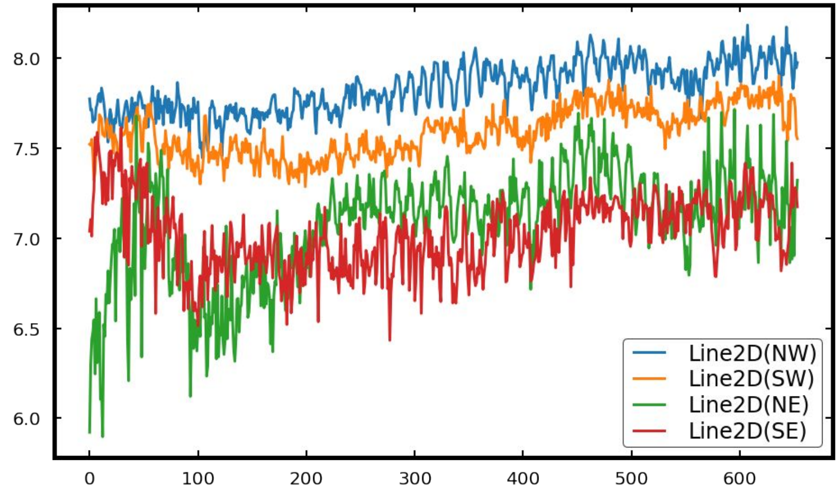
A comparison of algorithms for Spatial-Temporal Data Imputation

Mengqi Liu

Project Overview

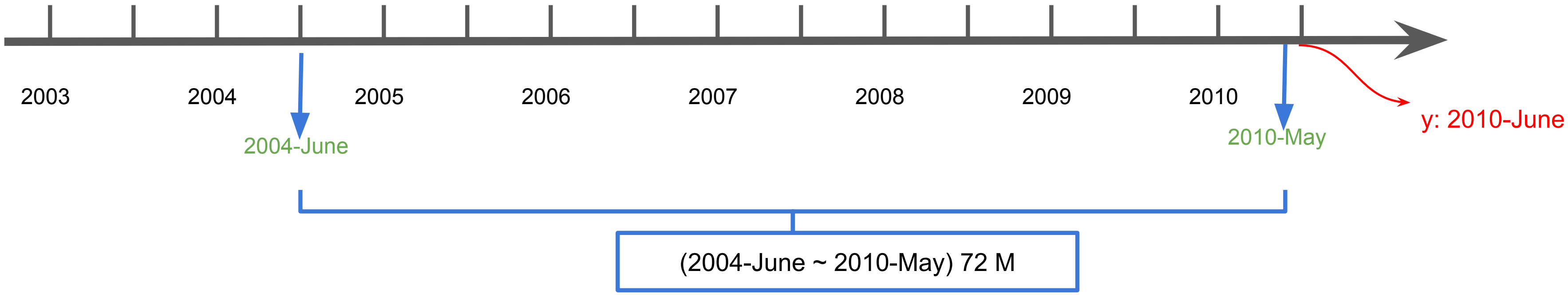
- **Objective:** impute missing values in spatial-temporal data
- **Challenge:**
    - Data does not fit MAR (missing at random)
    - In total, there are 56.7586% grid (see next slides) has value.
    - Grids in different area has different distribution

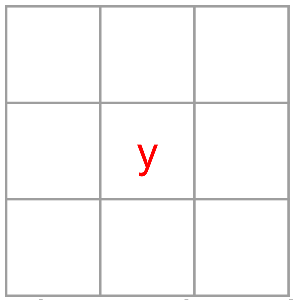




Task Description - Input & Output

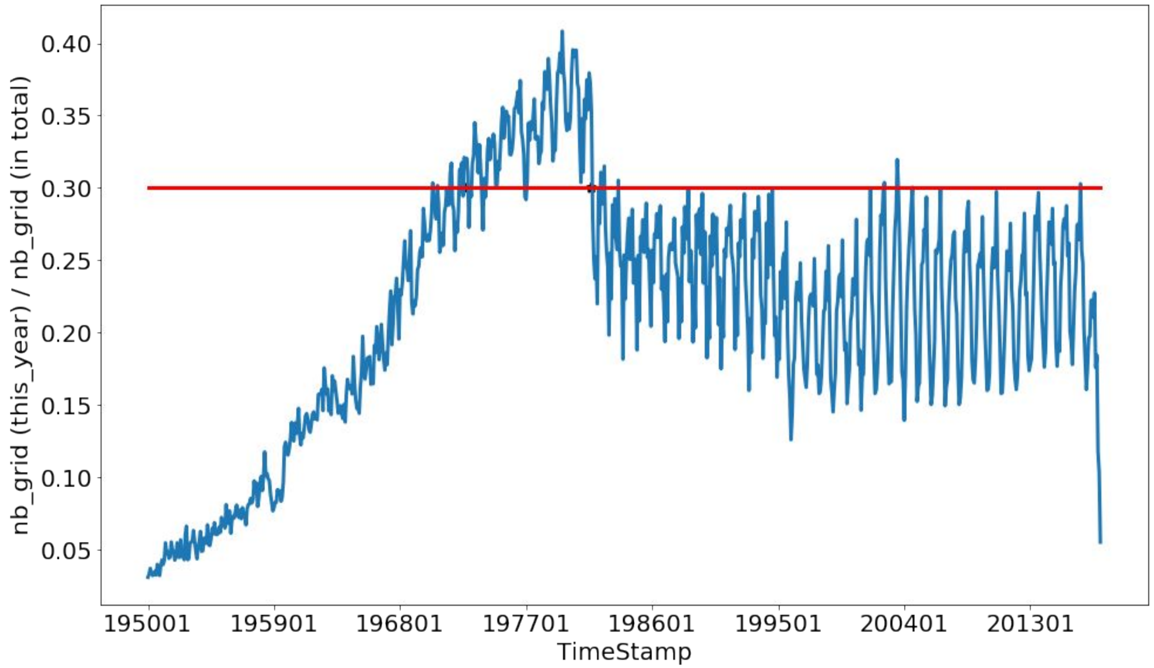
- Temporal Settings in previous experiments (X for input, y for output)

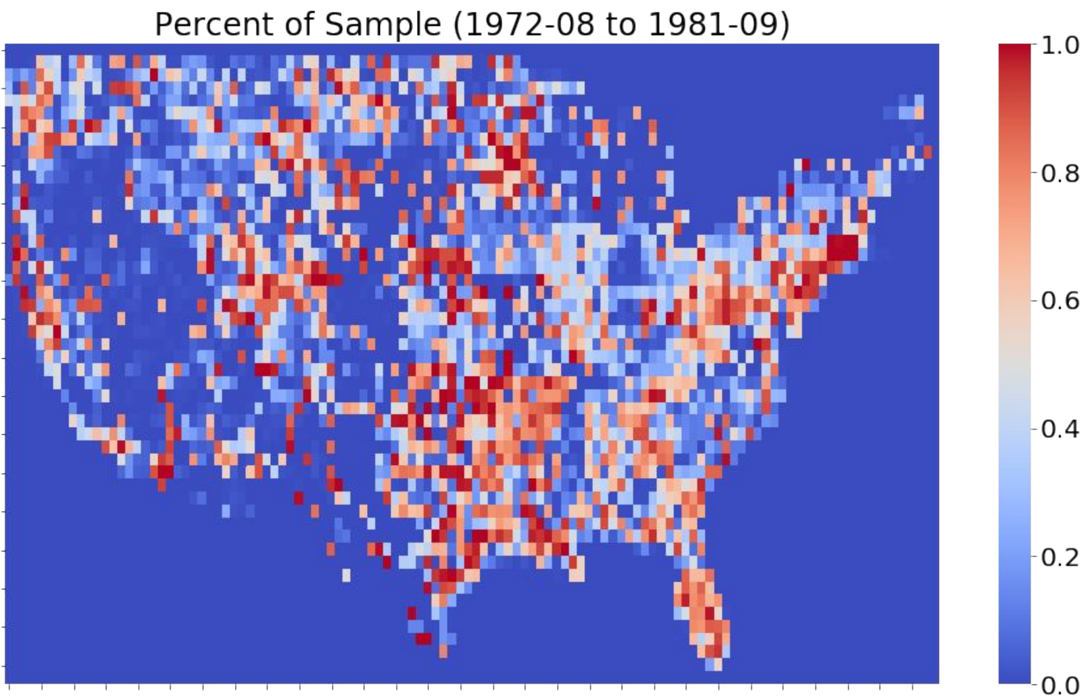




- Spatial Setting:
  - 3x3 grids (including current grid) in the previous time steps to predict the pH value of current grid at current time step.
  - grid size: ½ latitude x ½ longitude
- Model: single XGBoost for continental US

Data Used in Experiment





- Most of the time step doesn’t have much data
- Only use the continuous time steps that has above 30% of data (grid) that has value

Prediction Method - XGB [1]

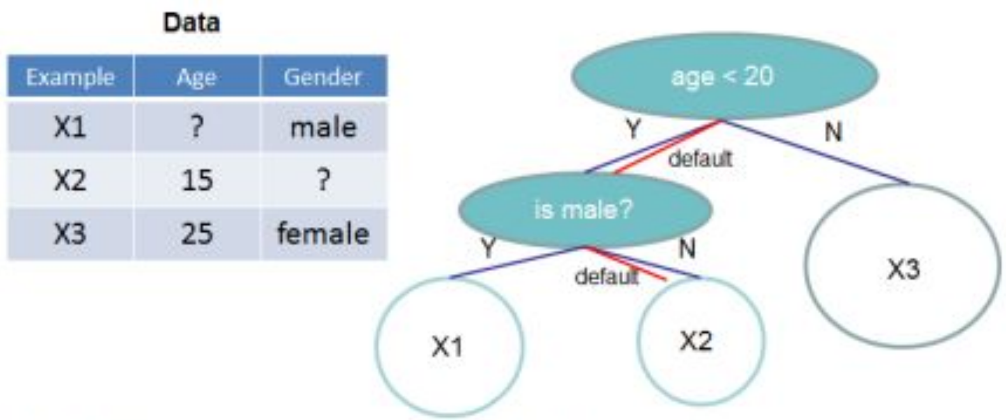


Figure 4: Tree structure with default directions. An example will be classified into the default direction when the feature needed for the split is missing.

This figure is taken from the original paper of XGB.

Experiment (Impute by grids at the same timestep)

RMSE from XGBoost			
None	1.7363	**EM**	0.3523
Fast KNN	0.2428	**Mean**	0.2338
MICE	0.2338	**Median**	0.2378
Mode	0.3058	**Random**	0.3686

Experiment (Impute by all timestep of current grid)

RMSE from XGBoost			
None	1.7363	**EM**	0.5353
Fast KNN	0.2544	**Mean**	0.2729
MICE	0.2729	**Median**	0.2729
Mode	0.2898	**Random**	0.8618

Reference

[1] Chen, T., He, T., & Benesty, M. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.

EM algorithm for Composite Likelihood

with application to two-way data array

Huy Dang

November 27, 2018

Motivation and Definition

Motivation: High dimensional response variables make likelihood inferences difficult by rendering the computation of likelihoods infeasible.

Thus, a class of likelihoods, called *Composite likelihoods /Pseudo-likelihoods* is often used in place of the full likelihood.

Definition: (Varin et. al., 2011) Consider a vector of random variable Y from the density $f(y;\theta)$ for some unknown p-dim parameter $\theta \in \Theta$. Let $(\mathcal{A}_1, \cdots, \mathcal{A}_K)$ be a set of marginal or conditional events with associated likelihoods $\mathcal{L}_k(\theta;y) \propto f(y \in \mathcal{A}_k;\theta)$.

Composite likelihood is defined as the weighted product:

$$\mathcal{L}_C(\theta;y) = \prod_{k=1}^{K} \mathcal{L}_k(\theta;y)^{w_k}$$

Examples

Examples:

- Composite conditional likelihood: pairwise cond. densities

$$\mathcal{L}_C(\theta; y) = \prod_{r=1}^{m} \prod_{s=1}^{m} f(y_r | y_s; \theta)$$

- Composite marginal likelihood:

$$\mathcal{L}_C(\theta; y) = \prod_{r=1}^{m} f(y_r | \theta)$$

Properties: There are results on asymptotic properties, efficiency, robustness of composite likelihood based estimators. But they vary case by case, and are somewhat limited.

Problem Statement

My simplified version: 2-way data array. U and V are i.i.d row and column discrete latent variables.

	V_1	V_2	$\cdots$	V_s
U_1	Y_{11}	Y_{12}	$\cdots$	Y_{1s}
U_2	Y_{21}	Y_{22}	$\cdots$	Y_{2s}
$\vdots$	$\vdots$		$\vdots$	
U_r	Y_{r1}	Y_{r2}	$\cdots$	Y_{rs}

$\lambda_u = P(U_i = u), u = 1, \cdots, k_1$

$\rho_v = P(V_j = v), v = 1. \cdots, k_2$

$Y_{ij} | U_i = u, V_j = v \sim N(\psi_{uv}, \sigma^2)$

	1	2	$\cdots$	k_2
1	ψ_{11}	ψ_{12}	$\cdots$	ψ_{1k_2}
2	ψ_{21}	ψ_{22}	$\cdots$	ψ_{2k_2}
$\vdots$	$\vdots$		$\vdots$	
k_1	$\psi_{k_1 1}$	$\psi_{k_1 2}$	$\cdots$	$\psi_{k_1 k_2}$

In reality: Problems can be made more complicated by allowing V to be generated from a Markov chain with k_2 states. It accommodates certain types of data: genomics, economics, etc.

Full and Composite Likelihood

Let $\boldsymbol{y}_i^{(r)}$ be the ith row of data, and $\boldsymbol{y}_j^{(c)}$ be the jth column.
The full likelihood is:

$$L(\theta; \boldsymbol{Y}) = p(\boldsymbol{Y}) = \sum_{\boldsymbol{u}} p(\boldsymbol{Y}|\boldsymbol{u})p(\boldsymbol{u})$$

where $p(\boldsymbol{Y}|\boldsymbol{u})$ is computed using a well-known recursion in HM literature (Baum et. al. 1970, Welch 2003).
Row Composite Likelihood: assuming that the rows are independent.

$$L_C(\theta; \boldsymbol{Y}) = \prod_i (\boldsymbol{y}_i^{(r)}) = \prod_i \sum_u \lambda_u p(\boldsymbol{y}_i^{(r)}|U_i = u)$$

where $p(\boldsymbol{y}_i^{(r)}|U_i = u)$ is computed using a well-known recursion in HM literature (Baum et. al. 1970, Welch 2003).
Flops(full) $= O(k_1^r k_2 s)$, Flops(Composite) $=$ O($\mathrm{k}_1 r k_2 s$)

EM Algorithm for Full Likelihood

$$L^*(\theta; \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{V}) = P(\boldsymbol{U} = \boldsymbol{u}) \cdot P(\boldsymbol{V} = \boldsymbol{v}) \cdot \prod_{i=1}^{r} \prod_{j=1}^{s} N(y_{ij}; \psi_{u_i v_j}, \sigma^2)$$

$$= \left(\prod_{i=1}^{r} \prod_{u=1}^{k_1} \lambda_u^{w_{iu}} \right) \left(\prod_{j=1}^{s} \prod_{v=1}^{k_2} \rho_v^{z_{jv}} \right) \left(\prod_{i=1}^{r} \prod_{j=1}^{s} \prod_{u=1}^{k_1} \prod_{v=1}^{k_2} N(y_{ij}; \psi_{uv}, \sigma^2) \right)^{w_{iu} z_{jv}}$$

where $w_{iu} = I(U_i = u)$; $z_{jv} = I(V_i = v)$

$$l^*(\theta; \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{r} \sum_{u=1}^{k_1} w_{iu} log(\lambda_u) + \sum_{j=1}^{s} \sum_{v=1}^{k_2} z_{jv} log(\rho_v)$$

$$+ \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{u=1}^{k_1} \sum_{v=1}^{k_2} w_{iu} z_{jv} log(N(y_{ij}; \psi_{uv}, \sigma^2))$$

The conditional expectation involves terms such as:
$E_{\theta^{(n-1)}}(w_{iu} | \boldsymbol{Y}) = P(U_i = u | \boldsymbol{Y}; \theta^{(n-1)}) = \frac{1}{p(\boldsymbol{Y})} \sum_{\boldsymbol{u}: u_i = u} p(\boldsymbol{Y} | \boldsymbol{u}) p(\boldsymbol{u})$

EM for Full (and Composite) Likelihood

For Composite Likelihood: Z_{ijv} in place of z_{jv}.

$$l_C^*(\theta; \boldsymbol{Y}_i^{(r)}, \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{r}\sum_{u=1}^{k_1} w_{iu} log(\lambda_u) + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{v=1}^{k_2} z_{ijv} log(\rho_v)$$
$$+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{u=1}^{k_1}\sum_{v=1}^{k_2} w_{iu} z_{jv} log(N(y_{ij}; \psi_{uv}, \sigma^2))$$

$$E_{\theta^{(n-1)}}(w_{iu}|\boldsymbol{Y}) = P(U_i = u|\boldsymbol{Y}_i^{(r)}) = \frac{1}{p(\boldsymbol{Y}_i^{(r)})} p(\boldsymbol{Y}_i^{(r)}|U_i = u)p(u)$$

Updates:

$$\lambda_u = \frac{1}{r}\sum_i \hat{w_{iu}}; \quad \rho_v = \frac{1}{s}\sum_j \hat{z_{jv}};$$

$$\mu_{uv} = \frac{(\widehat{w_{iu}z_{jv}})y_{ij}}{\sum_i \sum_j \widehat{w_{iu}z_{jv}}}; \quad \sigma^2 = \frac{1}{rs}\sum_i\sum_j\sum_u\sum_v (\widehat{w_{iu}z_{jv}})(y_{ij} - \mu_{uv})^2$$

Simulation

$$k_1 = 2, k_2 = 2, \rho = (0.39, 0.61), \lambda = (0.4, 0.6), \Psi = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \sigma^2 = 0.5$$

r = 10, s = 15

EM w/ Full Likelihood:

$$\hat{\rho} = (0.47, 0.53), \hat{\lambda} = (0.5, 0.5)$$

$$\hat{\Psi} = \begin{bmatrix} 0.9845 & 1.9034 \\ 3.0959 & 3.9929 \end{bmatrix}, \hat{\sigma}^2 = 0.2039$$

computation time:$4596.17s (76 mins)$

EM w/ Composite Likelihood:

$$\hat{\rho} = (0.47, 0.53), \hat{\lambda} = (0.5, 0.5)$$

$$\hat{\Psi} = \begin{bmatrix} 0.9845 & 1.8702 \\ 3.0959 & 3.97 \end{bmatrix}, \hat{\sigma}^2 = 0.1970$$

computation time:$1.27s$

r = 50, s = 100

EM w/ Full Likelihood: doesn't run

EM w/ Composite Likelihood:

$$\hat{\rho} = (0.52, 0.48), \hat{\lambda} = (0.48, 0.52)$$

$$\hat{\Psi} = \begin{bmatrix} 0.9807 & 1.9852 \\ 3.0093 & 4.0182 \end{bmatrix}, \hat{\sigma}^2 = 0.25$$

computation time:$551.58s$

Some comments

- When does it work? When does it misbehave?
- Starting value
- Possible next steps

Approximate Bayesian Computing for Differential Privacy

Jordan Awan

Department of Statistics, Penn State University

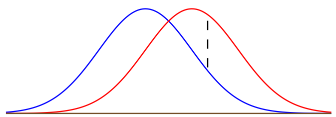
November 27, 2017

Differential Privacy

Definition (DMNS06, WZ10)

- Let $\mathcal{X}$ be a set,
- A *mechanism* $\mathcal{P} = \{P_{\underline{x}} \mid \underline{x} \in \mathcal{X}^n\}$ is a set of probability measures on a space $\mathcal{Z}$
- $\mathcal{P}$ satisfies ϵ-**Differential Privacy** (ϵ - DP) if for all $B \subset \mathcal{Z}$ and all $\underline{x}, \underline{x}'$ differing in one entry, we have

$$P_{\underline{x}}(B) \leq e^{\epsilon} P_{\underline{x}'}(B).$$



Problem Setup

- Collect sensitive data $\underline{X} \in \mathcal{X}^n$
- Output private summary $Z \sim P_{\underline{X}}(z)$

- Model $\underline{X} \sim f_\theta(\underline{x})$, with prior $\theta \sim \pi(\theta)$
- Want to infer about θ, given only Z.

$$\pi(\theta \mid Z) \propto \pi(\theta) \int_{\underline{x} \in \mathcal{X}^n} f_\theta(\underline{x}) P_{\underline{x}}(Z) \, d\underline{x}$$

- This integral is often intractable

Perspective originally from [WM10]

ABC

- Sample (approximately) a posterior distribution
- Does not require evaluating likelihood

Algorithm 1 ABC algorithm [MPR+11]

INPUT: $Z \in \mathcal{Z}$, ρ a pseudo-metric on $\mathcal{Z}$, and $c \geq 0$.

1: Draw $\theta \sim \pi$
2: Draw $Z' \sim f(z \mid \theta)$
3: If $\rho(Z', Z) \leq c$, accept θ, else reject θ,
4: Repeat 1-3 as desired.

OUTPUT: Accepted θ's

- If ρ is a metric, and $c = 0$, then samples are from $\pi(\theta \mid Z)$.

Toy Example

- $\theta \sim U[0, 1]$,
- $X \sim \text{Binom}(n, \theta)$,
- $Z = X + \text{DLap}(e^{-\epsilon})$

- Closed form of posterior
- Discrete: can use $c = 0$
- Simulation: $n = 100$, $\theta = .5$, $\epsilon = .1$
- $\approx 10^4$ accepted samples

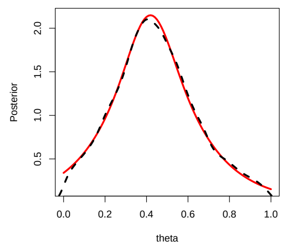


Figure: $c = 0$, AR: 1.7%

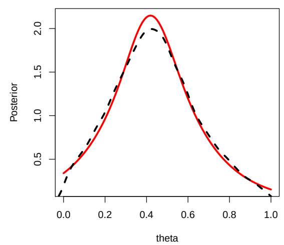
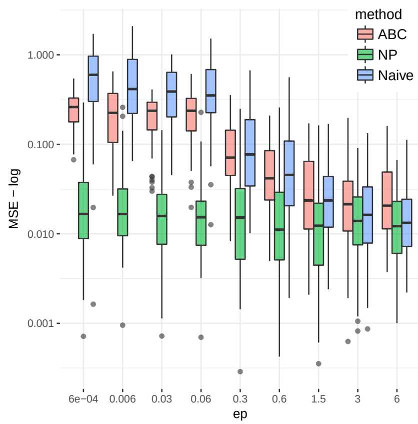


Figure: c as std error, AR: 20%

Problem based on [VS09] and [AS18]

Bigger Example

- Observe n iid copies of $D = (X, Y)$ (feature/class)
- $Y_i \sim \text{Bern}(p)$
- $X_i \mid (Y_i = j) \sim \text{Bern}(p_j)$



- Sufficient statistics:

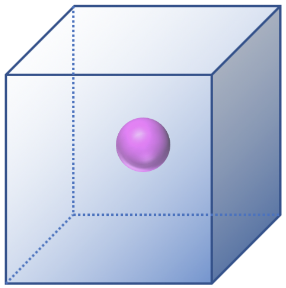
		X	
		1	2
Y	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

- Work with $m_{ij} = n_{ij} + e_{ij}$, where $e_{ij} \overset{\text{iid}}{\sim} \text{Dlap}(e^{-\epsilon/2})$.
- Posterior estimates of p, p_1, and p_2, given uniform priors

Problem based on [KKS16]

Acceptance Rate

- Each proposal in ABC is approximately uniform from $\mathcal{Z}$
- Suppose that $\mathcal{Z} = [a, b]^m$
- Acceptance region is a ball of radius $O\left(\frac{1}{\sqrt{n}}\right)$



- Acceptance rate is ratio of volumes $O\left(\frac{1}{n^{m/2}}\right)$

Image courtesy of Tobia Boschi

Conclusions

- Correct statistical inference by viewing private output as latent variable model
- Likelihood is often **computationally intractable**
- ABC offers an elegant method of sampling from posterior
  - ABC works well when Z is low-dimensional
  - Trade either accuracy, or computational efficiency when Z is higher-dimensional

References

[AS18] J. Awan and A. Slavković. Differentially Private Uniformly Most Powerful Tests for Binomial Data. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2018. To Appear.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *Calibrating Noise to Sensitivity in Private Data Analysis*, pages 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[KKS16] Vishesh Karwa, Dan Kifer, and Aleksandra Slavković. Private posterior distributions from variational approximations. *NIPS 2015 Workshop on Learning and Privacy with Incomplete Data and Weak Supervision*, 2016.

[MPR+11] Jean Michel Marin, Pierre Pudlo, Christian P. Robert, Université Paris Dauphine, Robin J. Ryder, and Université Paris Dauphine. Approximate bayesian computational methods. *Statistics and Computing*, pages 1–14, 2011.

[VS09] Duy Vu and Aleksandra Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143, Washington, DC, USA, 2009. IEEE Computer Society.

[WM10] Oliver Williams and Frank Mcsherry. Probabilistic inference and differential privacy. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2451–2459. Curran Associates, Inc., 2010.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *JASA*, 105:489:375–389, 2010.