Differentially Private Model Selection with Penalized and Constrained Likelihood

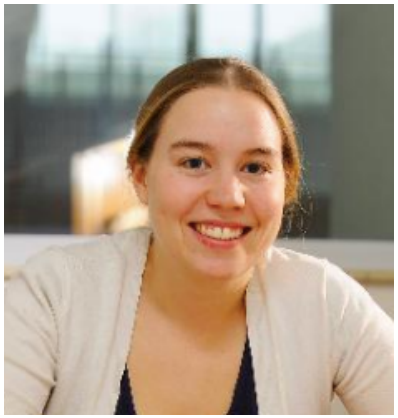
Jing Lei

Department of Statistics and Data Science, Carnegie Mellon University

Symposium on Data Science and Statistics

Reston VA, 2018.05.17

Collaborators



Anne-Sophie Charest
(U. Laval)

Sesa Slavkovic
(Penn State)

Adam Smith
(Boston Univ.)

Steve Fienberg

Privacy in the age of information

- Detailed personal data is being collected and used on a daily basis
  - Search queries are used to determine ads placement.
  - Emails in Gmail are used for targeted Ads.
  - YouTube & Amazon use viewing/buying records for recommendations.
  - Social networks: Facebook, LinkedIn, etc.
  - Hospitals collect health records.
- We want to make good use of these data, but individual privacy is a big concern.

Famous privacy stories: Netflix

- Netflix launched machine learning competitions to predict users' movie ratings.
- Released training data: anonymized user-movie ratings.
- User identity recovered by matching with IMDB data.
- The second Netflix competition ended in a privacy law-suit.

Famous privacy stories: NYC taxi

- In response to a public records request, NYC officials released start-end data for 173 million taxi trips.
- The two taxi ID numbers are converted to one-way cryptographic hashes.
- All ID's fully recovered by matching the hashing output from the two ID systems.

Other examples

- Anonymized medical records can be re-identified by matching demographic information in public data base.
- AOL released search queries of anonymized users, but queries contain identifying information.

The need of strong privacy protection

- In these examples, there was some protection, but apparently not enough.
  - Unknown auxiliary data (IMDB, public demographic record, etc)
  - Powerful and smart attackers (matched hashing, search query mining, etc)
- Call for *ad omnia* privacy protection.

Basic setup

- Data $D = \{z_1, \ldots, z_n\} \in \mathscr{X}^n$, where $\mathscr{X}$ is the sample space.
- Statistic $f : D \mapsto f(D) \in \mathbb{R}$, e.g., sample mean, standard deviation, regression coefficients, p-value, etc.
- If f is deterministic, then not private against knowledgeable attackers (eg. attacker knows all but one record).
- In order to be private, f must be random.
- Assume that $f(D)$ is a random variable taking values in $\mathbb{R}^d$.
  - noise perturbed statistic
  - sampling from a predictive distribution

Differential Privacy [Dwork et al 06]

Let f be a randomized statistic. We say f satisfies ε-Differential Privacy if

$$e^{-\varepsilon} \leq \frac{\Pr(f(D) \in S)}{\Pr(f(D') \in S)} \leq e^{\varepsilon},$$

for all pairs (D, D') differing in one entry and all measurable sets $S \subseteq \mathbb{R}$.

This is a property of f only, regardless of the dataset.

Differential privacy in statistics

- Point estimation: [Dwork & L. 09], [Smith 11], [Chaudhuri et al 11], [L. 11], [Bassily et al 14], [Karwar & Slavkovic 16] ...
- Nonparametric estimation: [Wasserman & Zhou 11], [Hall et al 12].
- Minimax theory: [Chaudhuri & Hsu 11] [Duchi et al 14], [Barber & Duchi 14]
- Hypothesis testing: [Fienberg et al 13], [Johnson & Shmatikov 13], [Uhler et al 13], [Yu et al 13] ...
- + vast literature in machine learning and theoretical computer science

This Work: Linear Model Selection

Data: $D = (\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i) : 1 \leq i \leq n\}$

Model:

$$Y_i = \beta^T X_i + Z_i \,,$$

where $\beta \in \mathbb{R}^d$, $X_i \in \mathbb{R}^d$, $X_i \overset{iid}{\sim} P_X$, $Z_i \overset{iid}{\sim} N(0, \sigma^2)$.

Task: find $J_\beta = \{j : \beta_j \neq 0\}$.

Model Selection For Linear Regression

- Classical model selection ($d << n$): minimize some criteria among a set of candidate models.

  AIC, BIC, C_p, CV, GCV, etc.

- High dimensional ($d \asymp n$ or $d >> n$): minimize penalized residual sum of squares over the parameter space.

  LASSO, SCAD, ElasticNet, ...

- To achieve differential privacy, we combine these two approaches, with additional post-randomization.

- We give sufficient conditions on (n, d) and P_X for consistent and differentially private model selection.

Information Criteria

- Let $M \subseteq \{1,\ldots,d\}$ represent a model $\Theta_M := \{\beta \in \mathbb{R}^d : J_\beta \subseteq M\}$.
- Information Criteria

$$IC(M;D) = \text{Goodness of fit} + \text{Model Complexity}$$

- Goodness of fit: $\min_{\beta\in\Theta_M} \sum_{i=1}^n (Y_i - X_i^T\beta)^2 =: Q(M,D)$.
- Model Complexity: $\phi_n|M|$.
  - AIC: $\phi_n = 2$, BIC: $\phi_n = \log n$.
  - Our choice of ϕ_n: more similar to BIC.

Step 1: Truncation & Standardization

- Assume $|X_{ij}| \leq 1$ for all $1 \leq i \leq n$, $1 \leq j \leq d$.
- $|Y_i| \leq r$, for all $1 \leq i \leq n$.
- r is a tuning parameter
  - r too small: more bias
  - r too large: hard to control privacy
- Can be achieved by standard d.p. pre-processing [Dwork & L. 09, Smith 11].

Step 2: Penalized Constrained Least Square

- Assume σ^2 is known (e.g., σ^2=1)
- Constrained GoF with ℓ_1 constraint parameter R

$$Q_R(M,D) = \min_{\beta \in \Theta_M, \|\beta\|_1 \le R} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 .$$

- Private model selection with privacy parameter ε

$$\hat{M} = \arg\min_{m \in \mathscr{M}} \left\{ Q_R(M,D) + \phi_n |M| + \frac{2(r+R)^2}{\varepsilon} W_M \right\}$$

where W_M ($M \in \mathscr{M}$) are independent double-exponential random variables with mean 0 and variance 2.

Remarks

- Privacy is achieved by randomization with additive noise W_M.
- The additive noise is calibrated by $\frac{2(r+R)^2}{\varepsilon}$
- Recall r upper bounds $|Y_i|$, R upper bounds $\|\beta\|_1$.
  - $(r+R)$ large $\Rightarrow$ less bias but needs more noise for privacy
  - $(r+R)$ small $\Rightarrow$ more bias but less sensitive
- ϕ_n is the penalty coefficient.
- Can be extended to the case of unknown σ^2 using local sensitivity [Nissim et al 07].

Choice of algorithm parameters

- Choice of R
  - The ideal choice is $R = \|\beta^*\|_1$, where β^* is the true coefficient.
  - Practically, use a d.p. version of $\max_M \|\hat{\beta}_M\|_1$.
- Choice of ϕ_n
  - $\phi_n = \hat{\sigma}^2 \log n$, where $\hat{\sigma}^2$ is a d.p. estimate of $\hat{\sigma}^2$.
  - mimics BIC.
- Choice of ε
  - $\varepsilon = 1$: posterior probability changes less than three-fold
  - $\varepsilon = 0.1$: less than 10%
  - $\varepsilon \geq 10$ is practically meaningless.

Privacy guarantee

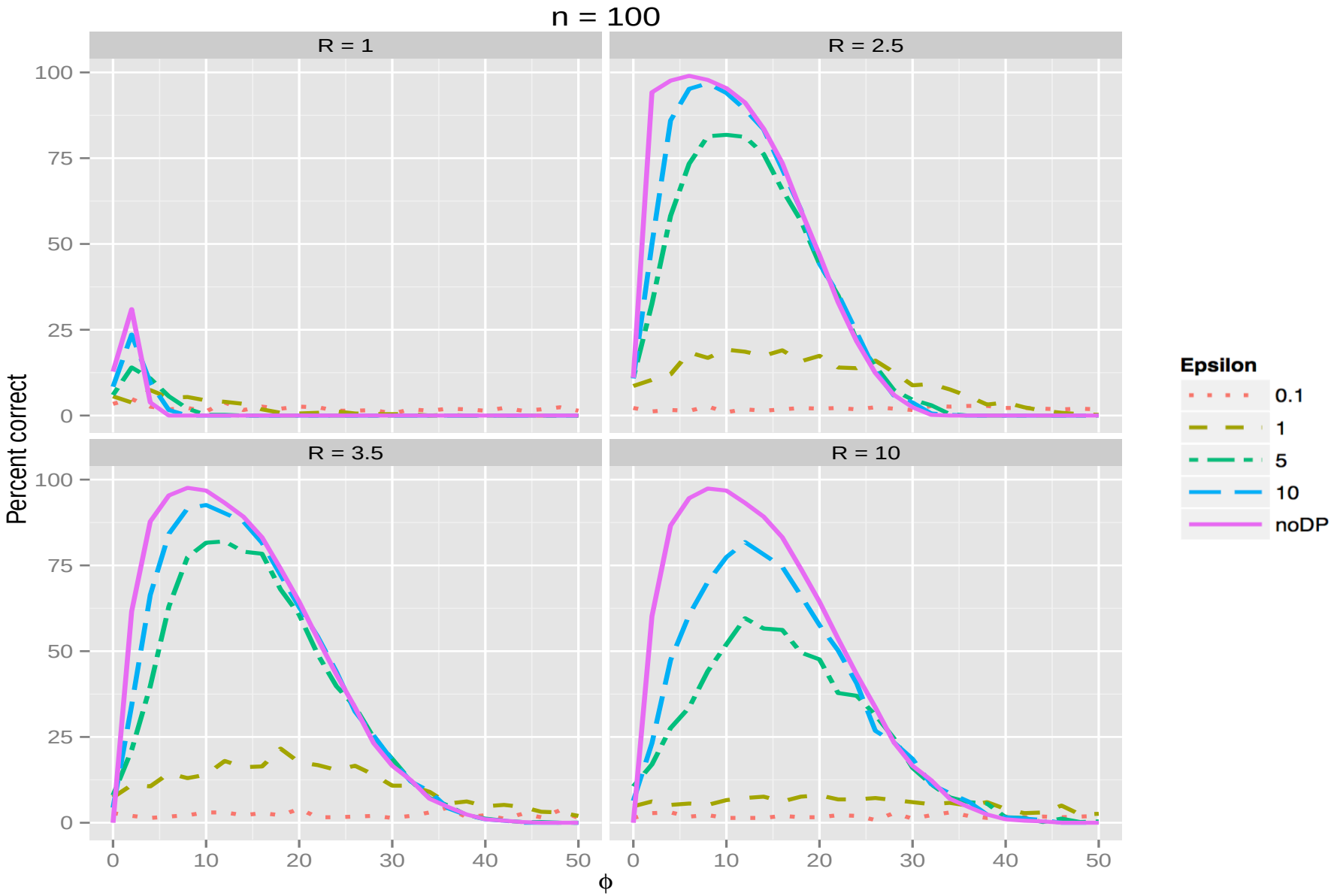
- The assumptions $|Y_i| \leq r$, $|X_{ij}| \leq 1$, $\|\hat{\beta}_M\| \leq R$ imply that the information criteria $Q_R(M,D) + \phi_n|M|$ are uniformly stable under perturbation of a single data entry (global sensitivity).
- The noise term $\frac{2(r+R)^2}{\varepsilon} W_M$ is calibrated to the sensitivity to ensure ε-differential privacy [Dwork et al 06].

Utility analysis

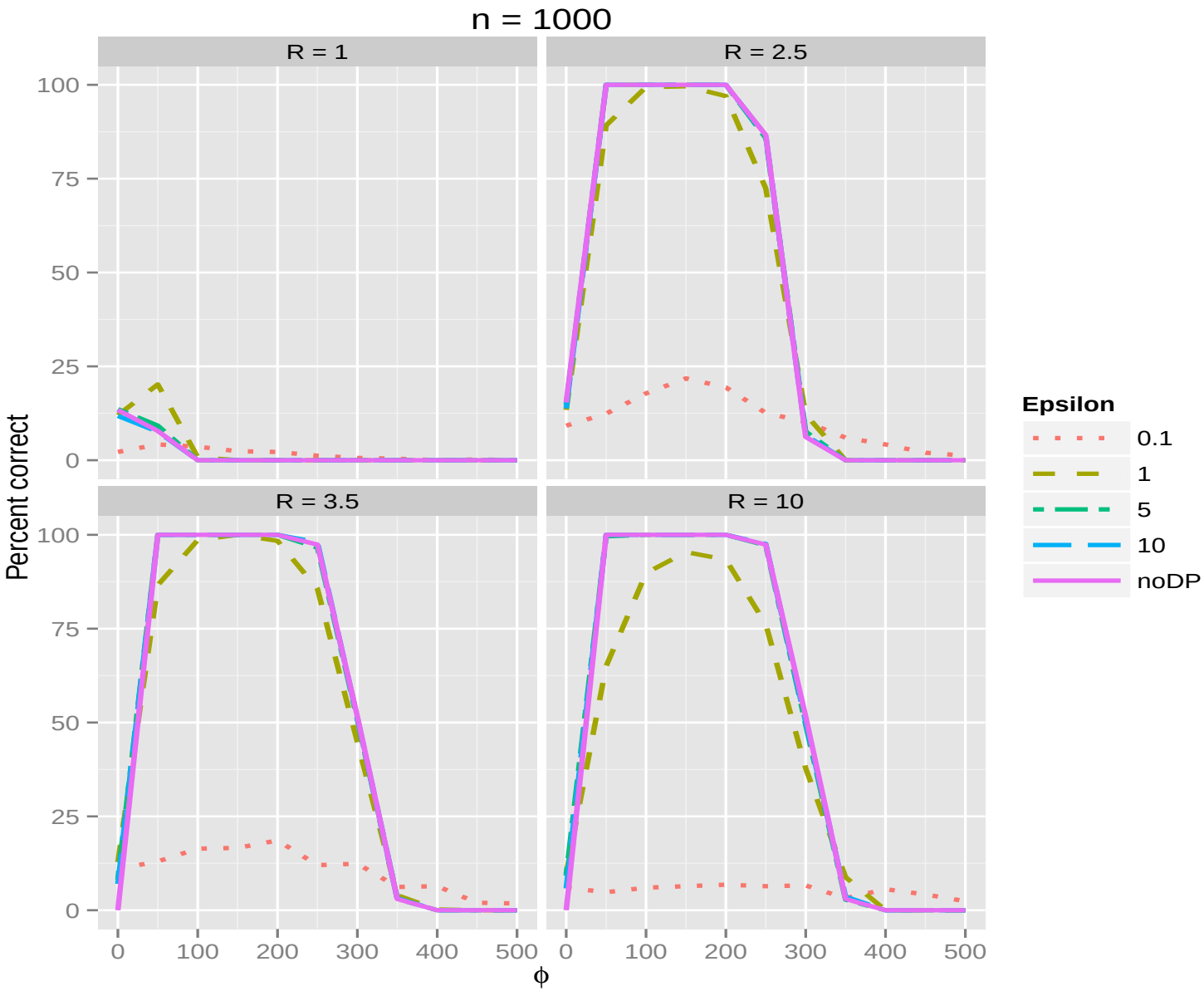
- β^* is true coefficient: $d_0 = \|\beta^*\|_0$, $b_0 = \min_{j:\beta_j^* \neq 0} |\beta_j|$
- $M^* = \{j : \beta_j^* \neq 0\} \in \mathscr{M}$
- $|\mathscr{M}| \leq n^{c_1}$ for some $c_1 > 0$
- $\max_{M \in \mathscr{M}} |M| \leq \bar{d} = o(n^{c_2})$ for some $c_2 > 0$
- $\inf_{1 \leq \|\beta\|_0 \leq \bar{d}+d_0} \frac{\beta^T \mathbf{X}^T \mathbf{X} \beta}{n\|\beta\|^2} := \kappa > 0$
- $2(1 \vee \sigma^2 \vee 4c_1 \varepsilon^{-1}(R+r)^2) \log n < \phi_n \leq \frac{1}{4 \vee (1+2d_0)} \kappa b_0^2 \sigma^2 n$
- $R \geq r\sqrt{\frac{\bar{d}}{\kappa}}$

Theorem: $P(\hat{M} = M^*) \to 1$.

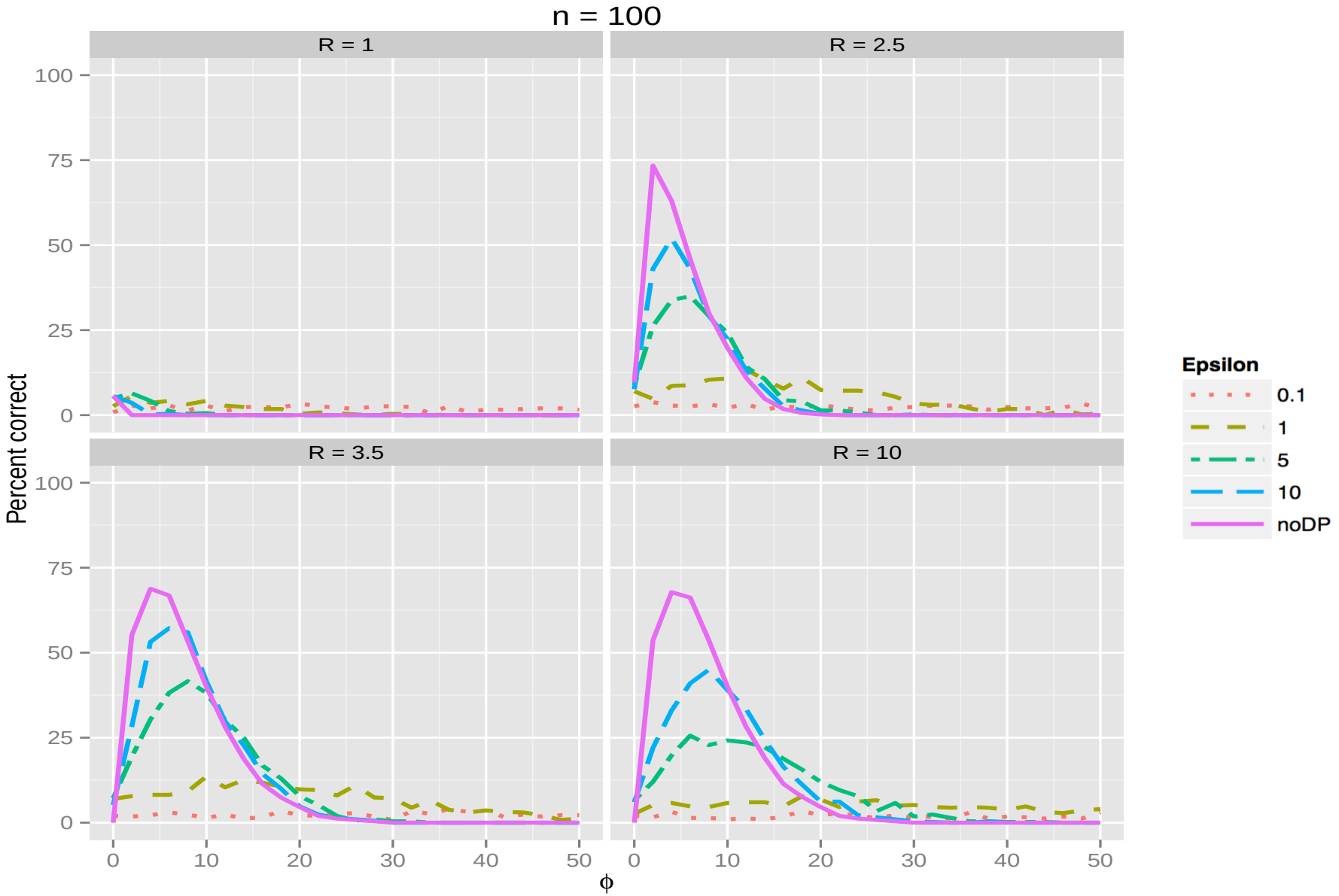
Simulation: $\beta = (1,1,1,0,0,0)$, $N(0,1)$ *noise*



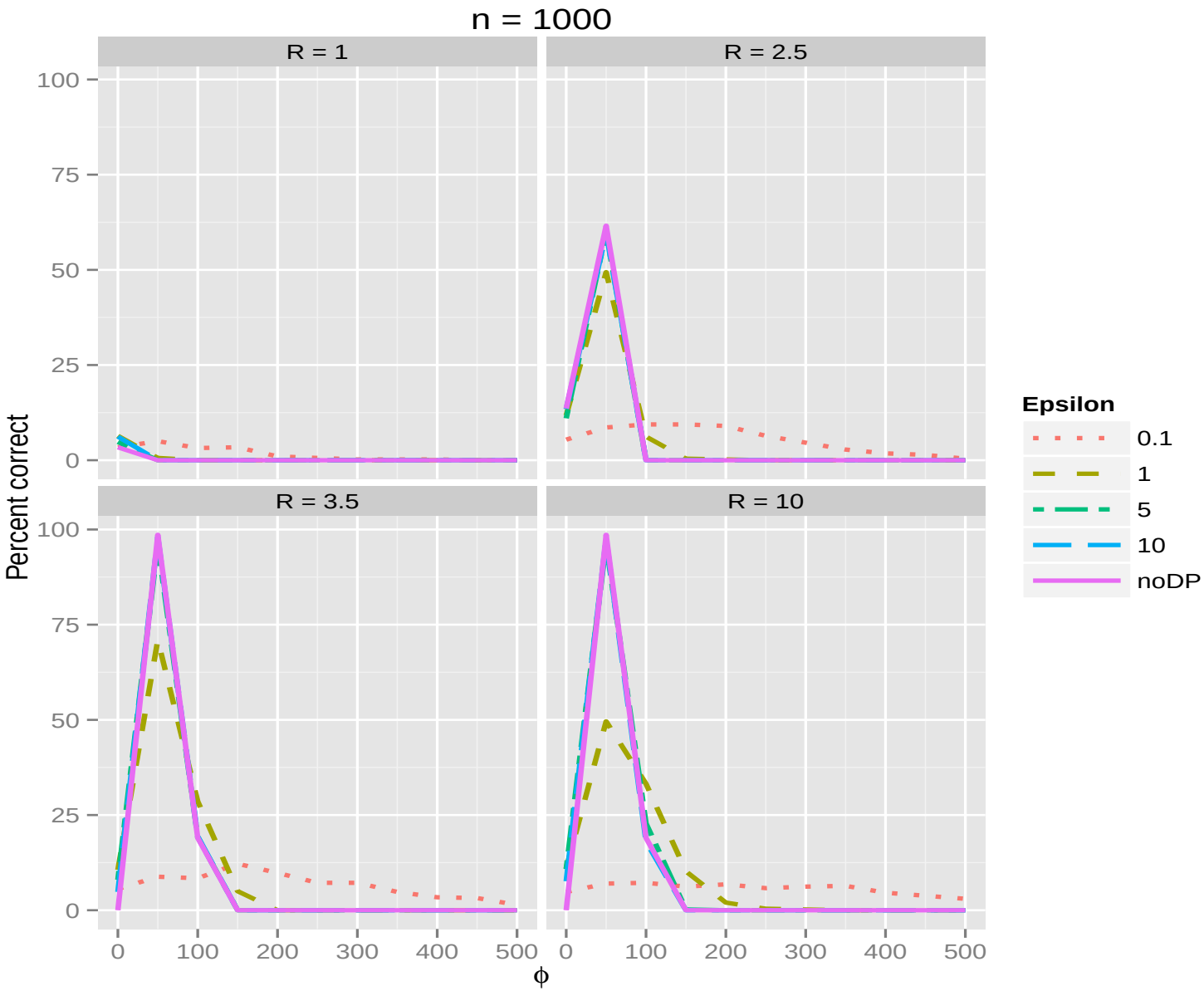
Simulation: $\beta = (1,1,1,0,0,0)$, $N(0,1)$ *noise*



Simulation: $\beta = (1.5, 1, 0.5, 0, 0, 0)$, $N(0,1)$ *noise*



Simulation: $\beta = (1.5, 1, 0.5, 0, 0, 0)$, $N(0,1)$ *noise*



Bay Area housing data

- $n = 235760$ houses in the Bay Area sold between 2003 and 2006, with price between 0.1 million and 0.9 million, size under 3000 sqft.
- Y is price.
- Covariates: year of transaction, latitude and longitude, county, house size, lot size, building age, number of bedrooms.
- Baseline estimator: least squares with BIC. Baseline R-squared$= 0.282$.

Results: average relative R-squared

	$\varepsilon = 1$				$\varepsilon = 5$			
ϕ	4	8	16	32	4	8	16	32
$R = 10$	.623	.623	.623	.623	.624	.624	.624	.623
$R = 25$	.995	.995	.995	.995	.998	.998	.998	.998
$R = 35$	.997	.997	.997	.996	1	1	1	.999
$R = 100$	.994	.993	.993	.993	.999	.999	.999	.999

Results: variable selection frequency

ϕ	bsqft	lsqft	time	lat	long	age
4	.85	.47	1	1	1	.84
8	.88	.49	1	1	1	.83
16	.85	.48	1	1	1	.86
32	.83	.45	1	1	1	.80
ϕ	nbr	ala	cc	mss	ns	sc
4	1	.60	.99	1	.92	.58
8	1	.60	.98	1	.91	.60
16	1	.58	.97	1	.91	.58
32	1	.56	.97	1	.90	.54

Conclusion

- D.p. model selection is possible, by privatizing standard methods.
- Good utility requires a large sample size.
- Side information (e.g., ℓ_1 norm of true coefficient) would be helpful.

Thank You!

Paper: https://doi.org/10.1111/rssa.12324

Slides: www.stat.cmu.edu/~jinglei/talk.shtml