Privacy Protection for Sparse Data

Jiashun Jin

Carnegie Mellon University

Feb. 24, 2010

Collaborators

Alphabetically:

Tony Cai	University of Pennsylvania
David Donoho	Stanford University
Stephen Fienberg	Carnegie Mellon University
Christopher Genovese	Carnegie Mellon University
Mark Low	University of Pennsylvania
Larry Wasserman	Carnegie Mellon University

Growing Concern of Individual Privacy

- Identity theft
- Breach in sensitive data (e.g. medical record)
- Hacker

Privacy Protection by Adding Noise

- With differential privacy as the emerging privacy protection technique
- With double exponential noise at the core
- Success has been found in various regression-type of settings
  - Dwork (06) (Differential privacy)
  - Zhou et al. (09) (PCA)
  - Chaudhuri & Moteleoni (08) (logistic reg.)

Linkage to Statistical Literature

- Tradeoff between utility and privacy is related to Statistical decision theory
- Differential privacy is reminiscent to lower bound argument in Statistics (e.g. Le Cam)
- Duality between two areas:
  - Confidentiality: adding noise
  - Statistics: noise removal
- Recent effort in linking two areas together:
  - Dwork and Lei (2009): differential privacy and robustness
  - Wasserman et al (2009): matrix masking and PCA

For Today

Forge linkage between Confidentiality and recent statistical literature in sparse inference

- Adding noise to sparse data
- Phase diagrams for when data mining is impossible/possible
- Individual privacy
- Application to restricted statistical queries

Sparsity

- A natural phenomenon found in many application areas
- Only **a small fraction** of the data contains relevant information or **signals**, others are irrelevant or noise
- How to exploit sparsity has been the theme of many active statistical areas
  - Wavelet
  - Variable selection
  - Cancer classification

An Example

A database contains diagnostic results for HIV

- Labels: 0 for Normal, 1 for HIV
- Sparsity: out of many such labels, the fraction of 1's is small (low risk population)
- **Goal:** add proper amount of noise so that
  - the 1's can not be successfully identified
  - valid data mining is still possible

Problem: how much noise to add?

Two Interconnected Problems

Sanitized Data (Gaussian Noise Added):

$$X_i = \beta_i + z_i, \quad \beta_i = 0/1, \quad z_i \stackrel{iid}{\sim} N(0, \sigma^2), \quad 1 \leq i \leq p$$

For which (ϵ, σ),

Impossibility/Possibility: valid inference is impossible/possible

No Recover/Recovery: individual 1's can't/can be identified

For double exponential noise, see Donoho and Jin (2004)

Model and Re-normalization

Original Setting:

$$X_i = \beta_i + z_i, \qquad \underline{z_i \overset{iid}{\sim} N(0, \sigma^2)}, \qquad 1 \leq i \leq p$$

For convenient, divide both sides by σ:

$$X_i = \beta_i + z_i, \qquad \underline{z_i \overset{iid}{\sim} N(0, 1)}, \qquad 1 \leq i \leq p$$

where

$$\beta_i = \begin{cases} \tau, & \text{with prob.} \quad \epsilon, \\ 0, & \text{with prob.} \quad 1 - \epsilon, \end{cases} \qquad \underline{\tau = \frac{1}{\sigma}}$$

Note:

- Driving parameters change from (ϵ, σ) to (ϵ, τ)!
- Marginally,

$$X_i \overset{iid}{\sim} (1 - \epsilon) N(0, 1) + \epsilon N(\tau, 1)$$

Impossibility/Possibility

The study of impossibility/possibility turns out to be closely related to the study of the following testing problem:

$$H_0: \qquad X_j \stackrel{iid}{\sim} N(0,1)$$

vs.

$$H_1^{(p)}: \qquad X_i \stackrel{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon N(\tau, 1)$$

Problem: for which (ϵ, τ) H_0 and $H_1^{(p)}$ separate completely, and for which they are inseparable

Calibrations of (ϵ, τ)

For asymptotics, let $p \to \infty$, and link (ϵ, τ) to p by parameters (ϑ, r)

- To model sparsity:

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad 0 < \vartheta < 1,$$

  - $0 < \vartheta < 1/2$: moderately sparse
  - $1/2 < \vartheta < 1$: very sparse

- For recovery, or hypothesis testing when it is very sparse, interesting range of τ is

$$\tau_p = \sqrt{2r \log p}, \qquad r > 0$$

- For hypothesis testing when it is moderately sparse, interesting range of τ is

$$\tau_p = O(p^{\vartheta - 1/2}) \qquad \text{(which is algebraically small)}$$

Detection Boundary (Very Sparse)

$$H_0: \quad X_i \stackrel{iid}{\sim} N(0,1); \qquad H_1^{(p)}: \quad X_i \stackrel{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1)$$

Theorem 1. If $\epsilon_p = p^{-\vartheta}$ and $\tau_p = \sqrt{2r \log p}$, where $1/2 < \vartheta < 1$ and $r > 0$, then:

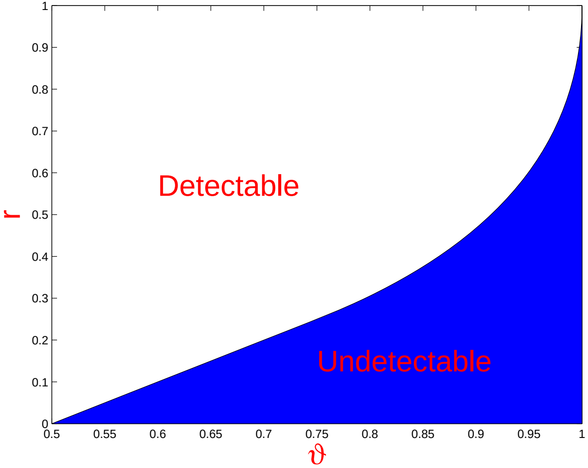
If $r > \rho(\vartheta)$, $\quad H_0$ and $H_1^{(p)}$ separate asymptotically,

If $r < \rho(\vartheta)$, $\quad H_0$ and $H_1^{(p)}$ merge asymptotically.

where

$$\rho(\vartheta) = \begin{cases} \vartheta - \frac{1}{2}, & \frac{1}{2} < \vartheta < \frac{3}{4}, \\ (1 - \sqrt{1-\vartheta})^2, & \frac{3}{4} < \vartheta < 1. \end{cases}$$

We call $r = \rho(\vartheta)$ the "detection boundary."



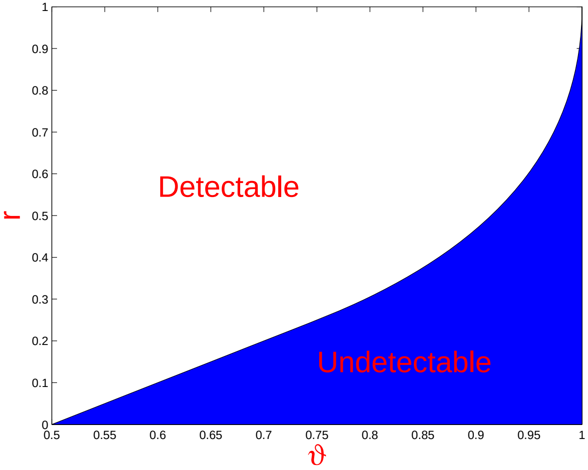
$$\epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r \log p}, \quad 1/2 < \vartheta < 1 \qquad \text{(very sparse)}$$

Detection Boundary (Moderately Sparse)

$$H_0: \qquad X_i \stackrel{iid}{\sim} N(0,1); \qquad H_1^{(p)}: \qquad X_i \stackrel{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1)$$

Theorem 2. If $\epsilon_p = p^{-\vartheta}$ and $0 < \vartheta < 1/2$. Then

$\tau_p \cdot p^{1/2-\vartheta} \to \infty$: H_0 and $H_1^{(p)}$ separate asymptotically
$\tau_p \cdot p^{1/2-\vartheta} \to 0$: H_0 and $H_1^{(p)}$ merge asymptotically



$$\epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r \log p}, \quad 1/2 < \vartheta < 1 \qquad \text{(very sparse)}.$$

Note: the detection boundary reaches 0 to the left.

Interpretation

- Critical σ^2 (recall that $\sigma^2 = \frac{1}{\tau^2}$):
  - $O(\frac{1}{\log(p)})$ for very sparse case ($\vartheta > 1/2$)
  - $O(p^{1-2\vartheta})$ for moderately sparse case ($\vartheta < 1/2$)
  - sparsifying data helps privacy protection
- Undetectable region: valid inference impossible
  - impossible to tell whether $\epsilon_p = 0$ or not
  - analyst unable to tell whether this is sanitized data, or pure white noise
  - impossible to accurately estimate ϵ_p

Sketch of Proofs

- For (ϑ, r) in the undetectable region, show that as $p \to \infty$, the Hellinger distance between the joint density under H_0 and that under $H_1^{(p)}$ $\to 0$
- For (ϑ, r) in the detectable region, show that as $p \to \infty$, the Neyman-Pearson's Likelihood ratio test (LRT) has level $\to 0$ and power $\to 1$
- LRT needs (ϑ, r); prefer to some method that does not depend on (ϑ, r)

(Tukey's) Higher Criticism

Observe $X_1, X_2, \ldots, X_p$

- Convert each X_i to a p-value by $\pi_j = P(N(0,1) \geq X_j)$
- Sorting all p-values in the ascending order: $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$
- Higher Criticism is defined as

$$HC_p^* = \max_{\{1 \leq j \leq p/2\}} \frac{(j/p) - \pi_{(j)}}{\sqrt{\pi_{(j)}(1 - \pi_{(j)})}}$$

Donoho and Jin (2004)

Higher Criticism, II

$$H_0: \quad X_j \stackrel{iid}{\sim} N(0,1); \qquad\qquad H_1^{(p)}: \quad X_j \stackrel{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1)$$

Higher Criticism Test (HCT): rejecting H_0 if and only if

$$HC_n^* \geq \sqrt{2(1+\delta)\log\log n}, \qquad \text{say,} \quad \delta = 0.1$$

Theorem 3. Fix (ϑ, r) in the "interior" of the detectable region. As $p \to \infty$, the level of HCT $\to 0$ and the power of HCT $\to 1$.

"Interior":

$$\begin{cases} r > \rho(\vartheta), & \text{if } 1/2 < \vartheta < 1 \quad \text{(very sparse)} \\ \frac{\tau_p \cdot p^{1/2-\vartheta}}{\sqrt{2\log\log p}} \to \infty, & \text{if } 0 < \vartheta < 1/2 \quad \text{(moderately sparse)} \end{cases}$$

Detectable Region

- Higher Criticism test yields full power detection
- It is possible to consistently estimate (ϵ_p, τ_p)
- In broader models where nonzero β_j maybe unequal, it is possible to have a nonzero confidence for ϵ_p

Remaining Problem: Identifying nonzero β_j and Individual Privacy

See details in Cai et al. (2007), Meinshausen and Rice (2006)

Hamming Distance

- For any estimator $\hat{\beta}$, the hamming distance is

$$Hamm_p(\vartheta, r) = E_{\epsilon_p,\tau_p}\left[\sum_{j=1}^{p} 1_{\{\mathrm{sgn}(\hat{\beta}_j)\neq \mathrm{sgn}(\beta_j)\}}\right]$$

- For an appropriately chosen threshold $t = t_p$,

$$\beta_j = \begin{cases} \tau_p, & X_j \geq t_p \\ 0, & X_j < t_p \end{cases}$$

Problem: what is the best t_p?

Intruder's Option

$$X_i \stackrel{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1), \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}$$

Theorem 4. The best threshold for the intruder is

$$t_p = \begin{cases} t_B(\vartheta, r), & r > \vartheta, \\ \sqrt{2\vartheta \log p} \ \dagger, & r < \vartheta; \end{cases} \quad t_B(\vartheta, r) = \frac{\vartheta + r}{2\sqrt{r}}\sqrt{2\log p}$$

which gives the optimal Hamming distance

$$Hamm_p(\vartheta, r) \begin{cases} = L(p) \cdot p^{1-\frac{(\vartheta+r)^2}{4r}}, & r > \vartheta, \\ \sim p^{1-\vartheta}, & 0 < r < \vartheta \end{cases}$$

where $L(p)$ denotes a multi-log(p) term.

Genovese, Jin, Wasserman (2009); $\dagger$*: not unique*

Phase Change

- Phase change in the optimal threshold:

$$\frac{t_p}{\tau_p} < 1, \quad \text{if } r > \vartheta; \qquad\qquad \frac{t_p}{\tau_p} > 1, \quad \text{otherwise}$$

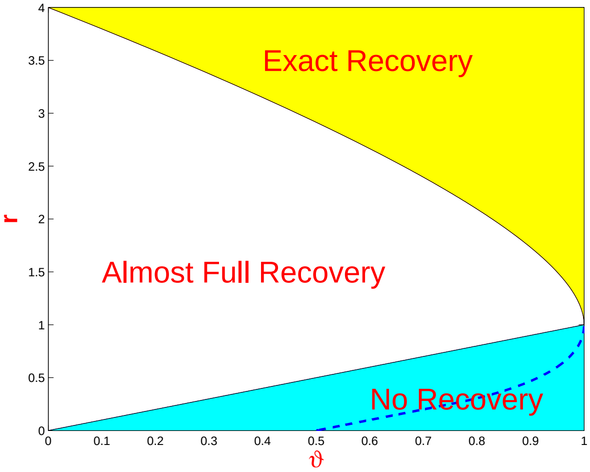
- Phase change in individual privacy:

$$P(\hat{\beta}_j = \tau_p | \beta_j = \tau_p) \begin{cases} \sim 1 & r > \vartheta \\ \text{algebraically small}, & r < \vartheta \end{cases}$$

- Phase change in the optimal rate:

$$\text{Hamm}_p(\vartheta, r) \begin{cases} \sim p\epsilon_p, & 0 < r < \vartheta & \text{(No Recovery)}, \\ \ll p\epsilon_p, & \vartheta < r < (1+\sqrt{1-\vartheta})^2 & \text{(Almost Full Recovery)} \\ o(1), & r > (1+\sqrt{1-\vartheta})^2 & \text{(Exact Recovery)} \end{cases}$$

Phase Diagram (Recovery)



$$\epsilon = p^{-\vartheta}, \qquad \tau = \sqrt{2r \log p}, \qquad 0 < \vartheta < 1, \qquad 0 < r < 1$$

Connection to Differential Privacy

- Except that in Region of Exact Recovery, we can finesse the data without noticing by either
  - replace a few signals by noise
  - replace a few noise by signal
- Idea: Hellinger distance between the joint densities **before** and **after** the finessing $= o(1)$
- Related to the optimal rate of estimating ϵ_p (see Cai *et al.* (2007)).

Application to Variable Selection

Linear model:

$$Y = W\beta + Z, \qquad Z \sim N(0, I_n)$$

- $W = W_{n,p}$; p: dimension; n: sample size
- β: p by 1 (unknown)
- Modern setting:

$$p \gg n, \qquad \beta \text{ is sparse}$$

Goal: decide which coordinates of β are nonzero and which are zero

Example: Statistical Queries

- Database allows for a total of n queries
- For the i-th query, the database randomly generates a weight vector

$$w_i = (w_{i1}, x_{i2}, \ldots, w_{ip})^T$$

and returns

$$y_i = w_i^T \beta + z_i, \qquad z_i \sim N(0,1), \qquad 1 \leq i \leq n$$

- In matrix form, $Y = W\beta + Z$

Dinur and Nissim (2004)

Asymptotic Model (Variable Selection)

Suppose

- $W = \big(w(i,j)\big)_{1 \leq i \leq n, 1 \leq j \leq p}$

$$w(i,j) \stackrel{iid}{\sim} N(0, \frac{1}{n})$$

- as before,

$$\beta_j = \begin{cases} \tau_p, & \text{prob. } \epsilon_p, \\ 0, & \text{prob. } 1 - \epsilon_p \end{cases}$$

- for parameters $\vartheta, \theta \in (0,1)$ and $r > 0$,

$$n = p^{\theta}, \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r \log p}$$

Optimal Rate in Hamming Distance, II

Fix $0 < \vartheta, \theta, r < 1$,

$$n = p^{\theta}, \qquad \epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r \log p}$$

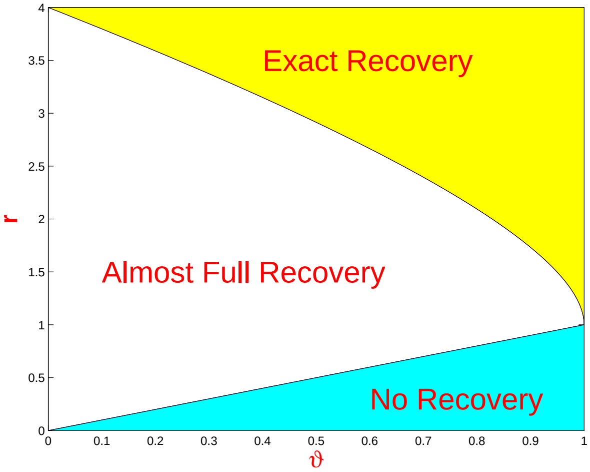
Theorem 5. Suppose $\theta > 2(1 - \vartheta)$. The optimal Hamming distance

$$\mathrm{Hamm}_p(\vartheta, r) \begin{cases} = L(p) \cdot p^{1 - \frac{(\vartheta + r)^2}{4r}}, & r > \vartheta, \\ \sim p^{1-\vartheta}, & 0 < r < \vartheta \end{cases}$$

where $L(p)$ denotes a multi-log(p) term.

Genovese, Jin, Wasserman (2009)

Phase Diagram (Recovery)



$$\epsilon = p^{-\vartheta}, \qquad \tau = \sqrt{2r \log p}, \qquad 0 < \vartheta < 1, \qquad 0 < r < 1$$

Sketch of Proofs

- Region of No Recovery: relate the variable selection to hypotheses testing

$$H_{0,i}: \qquad \beta_i = 0 \qquad \text{vs.} \qquad H_{1,i}: \qquad \beta_i = \tau_p$$

  Let f_{0i} be the density associated with H_{0i}, and f_{1i} be the density associated with H_{1i}. For any procedure, the Hamming distance $\geq \|(1-\epsilon_p)f_{0i} - \epsilon_p f_{1i}\|_1$
- Region of Almost Full Recovery/Exact Recovery: use the Lasso

Note: improves that in Wainwright (2006)

The Lasso

- A variable selection procedure proposed by Chen et al. (1995) and Tibshirani (1996).
- Look for solution $\hat{\beta}$ that minimizes

$$\|y - W\beta\|^2 + \lambda|\beta|_1,$$

with $\|\cdot\|$ for ℓ^2-norm and $|\cdot|_1$ for ℓ^2-norm.

Suppose $n = n_p = p^\theta$ and $\theta > 2(1 - \vartheta)$. Setting the tuning parameter

$$\lambda = 2 \cdot \max\left\{\frac{\vartheta + r}{2\sqrt{\vartheta r}},\ 1\right\} \cdot \sqrt{2\vartheta \log p}$$

yields the optimal rate in Hamming distance

Take-home messages

- Discussed adding noise approach to privacy protection for sparse data
- Introduced precise demarcation for
  - when data mining is impossible/possible
  - when accurately identifying individual signals is impossible/possible
- Tried to forge links between confidentiality and current statistical literature

Website

www.stat.cmu.edu/~jiashun/Research/

Donoho and Jin (2004):	Higher Criticism and Phase Diagram
Cai, Jin, and Low (2007):	Estimating ϵ_p
Fienberg and Jin (2009):	Multiplicity issues in Confidentiality
Genovese, Jin, Wasserman (2010):	Variable Selection and the Lasso

In preparation: linkage to confidentiality