Necessary and Sufficient Conditions for High-Dimensional Salient Feature Subset Recovery

Vincent Tan, Matt Johnson, Alan S. Willsky

Stochastic Systems Group,

Laboratory for Information and Decision Systems,

Massachusetts Institute of Technology

ISIT (Jun 14, 2010)

Motivation: A Real-Life Example

- The Manchester Asthma and Allergy Study (MAAS) is a birth-cohort study of more than $n \approx 1000$ children.

Motivation: A Real-Life Example

- The Manchester Asthma and Allergy Study (MAAS) is a birth-cohort study of more than $n \approx 1000$ children.

www.maas.org.uk

Motivation: A Real-Life Example

- The Manchester Asthma and Allergy Study (MAAS) is a birth-cohort study of more than $n \approx 1000$ children.

www.maas.org.uk

- Number of variables $d \approx 10^6$.

Motivation: A Real-Life Example

- The Manchester Asthma and Allergy Study (MAAS) is a birth-cohort study of more than $n \approx 1000$ children.

www.maas.org.uk

- Number of variables $d \approx 10^6$.
- But only $k \approx 30$ are salient for assessing susceptibility to asthma.

Motivation: A Real-Life Example

- The Manchester Asthma and Allergy Study (MAAS) is a birth-cohort study of more than $n \approx 1000$ children.

www.maas.org.uk

- Number of variables $d \approx 10^6$.
- But only $k \approx 30$ are salient for assessing susceptibility to asthma.
- Identification of these salient features is important.

Some Natural Questions

Natural questions:

Some Natural Questions

Natural questions:

- What is the meaning of saliency?
- How to define it information-theoretically?

Some Natural Questions

Natural questions:

- What is the meaning of saliency?
- How to define it information-theoretically?
- How many samples n are required to identify the k salient features out of the d features?

$$\text{Num features } d \gg \text{Num salient features } k$$

- What are the fundamental limits for recovery of the salient set?

Some Natural Questions

Natural questions:

- What is the meaning of saliency?
- How to define it information-theoretically?
- How many samples n are required to identify the k salient features out of the d features?

$$\text{Num features } d \gg \text{Num salient features } k$$

- What are the fundamental limits for recovery of the salient set?
- Are there any efficient algorithms for special classes of features?

Main Contributions

- We define the salient set by appealing to the error exponents in hypothesis testing.

Main Contributions

- We define the salient set by appealing to the error exponents in hypothesis testing.
- Sufficiency: We show that if for all n sufficiently large,

$$n > \max\left\{C_1 \cdot k \cdot \log\left(\frac{d-k}{k}\right), \exp(C_2 \cdot k)\right\},$$

then the error probability is arbitrarily small.

Main Contributions

- We define the salient set by appealing to the error exponents in hypothesis testing.
- Sufficiency: We show that if for all n sufficiently large,

$$n > \max\left\{C_1 \cdot k \cdot \log\left(\frac{d-k}{k}\right), \exp(C_2 \cdot k)\right\},$$

  then the error probability is arbitrarily small.
- Necessity: Under certain conditions, for $\lambda \in (0,1)$, if

$$n < \lambda \cdot C_3 \cdot \log\left(\frac{d}{k}\right)$$

  then the error probability $\geq 1 - \lambda$.

System Model

- Let the alphabet for each variable be $\mathcal{X}$, a finite set.
- High-dimensional setting where d and k grow with n.
- Two sequences of unknown d-dimensional distributions

$$P^{(d)}, Q^{(d)} \in \mathcal{P}(\mathcal{X}^d), \qquad d \in \mathbb{N}.$$

System Model

- Let the alphabet for each variable be $\mathcal{X}$, a finite set.
- High-dimensional setting where d and k grow with n.
- Two sequences of unknown d-dimensional distributions

$$P^{(d)}, Q^{(d)} \in \mathcal{P}(\mathcal{X}^d), \qquad d \in \mathbb{N}.$$

- IID samples $(\mathbf{x}^n, \mathbf{y}^n) := (\{\mathbf{x}^{(l)}\}_{l=1}^n, \{\mathbf{y}^{(l)}\}_{l=1}^n)$ drawn from $P^{(d)} \times Q^{(d)}$.
- Each pair of samples $\mathbf{x}^{(l)}, \mathbf{y}^{(l)} \in \mathcal{X}^d$.

Definition of Saliency

Motivated by asymptotics of binary hypothesis testing

$$H_0 : \mathbf{z}^n \sim P^{(d)}, \qquad H_1 : \mathbf{z}^n \sim Q^{(d)}$$

Definition of Saliency

Motivated by asymptotics of binary hypothesis testing

$$H_0 : \mathbf{z}^n \sim P^{(d)}, \qquad H_1 : \mathbf{z}^n \sim Q^{(d)}$$

Chernoff-Stein Lemma: If $\Pr(\hat{H}_1 | H_0) < \alpha$, then

$$\Pr(\hat{H}_0 | H_1) \doteq \exp(-nD(P^{(d)} \,||\, Q^{(d)}))$$

Chernoff Information:

$$\Pr(\mathrm{err}) = \pi_0 \Pr(\hat{H}_1 | H_0) + \pi_1 \Pr(\hat{H}_0 | H_1) \doteq \exp(-nC(P^{(d)}, Q^{(d)}))$$

where

$$C(P, Q) := -\min_{t \in [0,1]} \log \sum_{\mathbf{z}} P(\mathbf{z})^t Q(\mathbf{z})^{1-t}$$

Definition of Saliency

$P_A^{(d)}$ is the marginal of $P^{(d)}$ on the subset $A \subset \{1, \ldots, d\}$.

Definition of Saliency

$P_A^{(d)}$ is the marginal of $P^{(d)}$ on the subset $A \subset \{1, \ldots, d\}$.

- Definition: A set $S_d \subset \{1, \ldots, d\}$ is KL-divergence-salient if

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P_{S_d}^{(d)} \,||\, Q_{S_d}^{(d)}).$$

Definition of Saliency

$P_A^{(d)}$ is the marginal of $P^{(d)}$ on the subset $A \subset \{1, \ldots, d\}$.

- Definition: A set $S_d \subset \{1, \ldots, d\}$ is KL-divergence-salient if

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P_{S_d}^{(d)} \,||\, Q_{S_d}^{(d)}).$$

- Definition: A set $S_d \subset \{1, \ldots, d\}$ is Chernoff Information-salient if

$$C(P^{(d)}\,,\, Q^{(d)}) = C(P_{S_d}^{(d)}\,,\, Q_{S_d}^{(d)}).$$

Characterization of Saliency

Lemma

The following are equivalent:

- S_d *is KL-divergence-salient.*
- S_d *is Chernoff Information-salient.*

Characterization of Saliency

Lemma

The following are equivalent:

- S_d *is KL-divergence-salient.*
- S_d *is Chernoff Information-salient.*
- *Conditionals are identical*

$$P^{(d)} = P^{(d)}_{S_d} \cdot W_{S_d^c|S_d} \qquad Q^{(d)} = Q^{(d)}_{S_d} \cdot W_{S_d^c|S_d}.$$

Characterization of Saliency

Lemma

The following are equivalent:

- S_d *is KL-divergence-salient.*
- S_d *is Chernoff Information-salient.*
- *Conditionals are identical*

$$P^{(d)} = P^{(d)}_{S_d} \cdot W_{S_d^c|S_d} \qquad Q^{(d)} = Q^{(d)}_{S_d} \cdot W_{S_d^c|S_d}.$$

What are the scaling laws on (n, d, k) so that the error probability can be made arbitrarily small?

Definition of Achievability

- A decoder is a set-valued function that maps samples to subsets of size k, i.e.,

$$\psi_n : (\mathcal{X}^d)^n \times (\mathcal{X}^d)^n \to \binom{\{1, \ldots, d\}}{k}.$$

- The decoder is given the true value of k, the number of salient features.

- We are working on relaxing this assumption.

Definition of Achievability

- A decoder is a set-valued function that maps samples to subsets of size k, i.e.,

$$\psi_n : (\mathcal{X}^d)^n \times (\mathcal{X}^d)^n \to \binom{\{1,\ldots,d\}}{k}.$$

- The decoder is given the true value of k, the number of salient features.
- We are working on relaxing this assumption.

Definition: The tuple of model parameters (n, d, k) is achievable for $\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}$ if there exists a sequence of decoders $\{\psi_n\}$ such that

$$q_n(\psi_n) := \Pr(\psi_n(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) < \epsilon, \qquad \forall\, n > N_\epsilon.$$

Three Assumptions on Distributions $P^{(d)}, Q^{(d)}$

- Saliency: For every $P^{(d)}, Q^{(d)}$ there exists a salient set S_d of known size k, i.e.,

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}).$$

Three Assumptions on Distributions $P^{(d)}, Q^{(d)}$

- Saliency: For every $P^{(d)}, Q^{(d)}$ there exists a salient set S_d of known size k, i.e.,

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}).$$

- η-Distinguishability: There exists a constant $\eta > 0$ such that

$$D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}) - D(P^{(d)}_{T_d} \,||\, Q^{(d)}_{T_d}) \geq \eta$$

for all $T_d \neq S_d$ such that $|T_d| = k$.

Three Assumptions on Distributions $P^{(d)}, Q^{(d)}$

- Saliency: For every $P^{(d)}, Q^{(d)}$ there exists a salient set S_d of known size k, i.e.,

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}).$$

- η-Distinguishability: There exists a constant $\eta > 0$ such that

$$D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}) - D(P^{(d)}_{T_d} \,||\, Q^{(d)}_{T_d}) \geq \eta$$

for all $T_d \neq S_d$ such that $|T_d| = k$.

- L-Boundedness: There exists a constant $L \in (0, \infty)$ such that

$$\log \left[\frac{P^{(d)}_{S_d}(\mathbf{z}_{S_d})}{Q^{(d)}_{S_d}(\mathbf{z}_{S_d})} \right] \in [-L, L]$$

for all states $\mathbf{z}_{S_d} \in \mathcal{X}^k$.

Achievability Result

Theorem

If there exists an $\delta > 0$ such that for some $B > 0$

$$n > \max\left\{ \frac{k}{B} \log\left(\frac{d-k}{k}\right), \exp\left(\frac{2k \log|\mathcal{X}|}{1-\delta}\right)\right\},$$

then there exists a sequence of decoders ψ_n^* that satisfies

$$q_n(\psi_n^*) = O(\exp(-nE)),$$

for some exponent $E > 0$.

Proof Idea and a Corollary

- Use the exhaustive search decoder. Search for the size-k set with the largest empirical KL-divergence.
- Large deviation bounds, e.g., Sanov's theorem.

Proof Idea and a Corollary

- Use the exhaustive search decoder. Search for the size-k set with the largest empirical KL-divergence.
- Large deviation bounds, e.g., Sanov's theorem.
- Positivity of error exponents under the specified conditions.
- Similar to main result in [Ng, UAI 1998] but we focus on exact subset recovery and not generalization error.

Proof Idea and a Corollary

- Use the exhaustive search decoder. Search for the size-k set with the largest empirical KL-divergence.
- Large deviation bounds, e.g., Sanov's theorem.
- Positivity of error exponents under the specified conditions.
- Similar to main result in [Ng, UAI 1998] but we focus on exact subset recovery and not generalization error.

Corollary

Let $k = k_0$ be a constant and $R \in (0, B/k_0)$. Then if

$$n > \frac{\log d}{R}$$

Proof Idea and a Corollary

- Use the exhaustive search decoder. Search for the size-k set with the largest empirical KL-divergence.
- Large deviation bounds, e.g., Sanov's theorem.
- Positivity of error exponents under the specified conditions.
- Similar to main result in [Ng, UAI 1998] but we focus on exact subset recovery and not generalization error.

Corollary

Let $k = k_0$ be a constant and $R \in (0, B/k_0)$. Then if

$$n > \frac{\log d}{R} \qquad \Rightarrow \qquad q_n(\psi_n^*) = O\left(\exp(-nE)\right).$$

Converse Result

We assume that the salient set S_d is chosen uniformly at random over all subsets of size k.

Converse Result

We assume that the salient set S_d is chosen uniformly at random over all subsets of size k.

Theorem

If for some $\lambda \in (0,1)$,

$$n < \frac{\lambda \cdot k \cdot \log\left(\frac{d}{k}\right)}{H(P^{(d)}) + H(Q^{(d)})}$$

then

$$q_n(\psi_n) \geq 1 - \lambda$$

for all decoders ψ_n.

Remarks

- Proof is a consequence of Fano's inequality.

Remarks

- Proof is a consequence of Fano's inequality.
- Bound never satisfied if variables in S_d^c are uniform and independent of S_d.

$$n < \underbrace{\frac{\lambda}{H(P^{(d)}) + H(Q^{(d)})}}_{O(d)} \cdot k \cdot \log\left(\frac{d}{k}\right)$$

- However, converse is interesting if distributions have additional structure on their entropies.

Remarks

- Proof is a consequence of Fano's inequality.
- Bound never satisfied if variables in S_d^c are uniform and independent of S_d.

$$n < \underbrace{\frac{\lambda}{H(P^{(d)}) + H(Q^{(d)})}}_{O(d)} \cdot k \cdot \log\left(\frac{d}{k}\right)$$

- However, converse is interesting if distributions have additional structure on their entropies.
- Assume most of the features are processed or redundant.
- Example: there could be two features, "BMI" and "isObese". One is a processed version of the other.

Converse Result

Corollary

Assume that there there exists a $M < \infty$ *such that the conditional entropies satisfy*

$$\max\left\{H\left(P^{(d)}_{S_d^c|S_d}\right), H\left(Q^{(d)}_{S_d^c|S_d}\right)\right\} \le M \cdot k.$$

Converse Result

Corollary

Assume that there there exists a $M < \infty$ such that the conditional entropies satisfy

$$\max\left\{H\left(P^{(d)}_{S_d^c|S_d}\right), H\left(Q^{(d)}_{S_d^c|S_d}\right)\right\} \le M \cdot k.$$

If the number of samples satisfies

$$n < \frac{\lambda}{2(M + \log|\mathcal{X}|)} \cdot \log\left(\frac{d}{k}\right)$$

then

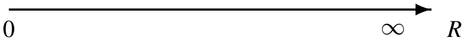
$$q_n(\psi_n) \ge 1 - \lambda.$$

Comparison to Achievability Result

- Assume $d = \exp(nR)$ and k is constant.

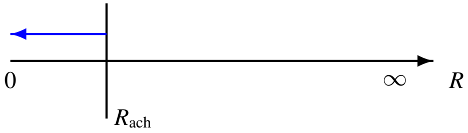
Comparison to Achievability Result

- Assume $d = \exp(nR)$ and k is constant.
- There is a rate R_{ach} so that if $R < R_{\text{ach}}$, then (n, d, k) is achievable.
- Conversely, there is another rate R_{conv} so that if $R > R_{\text{conv}}$, then recovery is not possible.



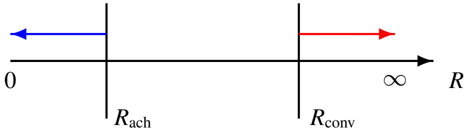
Comparison to Achievability Result

- Assume $d = \exp(nR)$ and k is constant.
- There is a rate R_{ach} so that if $R < R_{\text{ach}}$, then (n, d, k) is achievable.
- Conversely, there is another rate R_{conv} so that if $R > R_{\text{conv}}$, then recovery is not possible.



Comparison to Achievability Result

- Assume $d = \exp(nR)$ and k is constant.
- There is a rate R_{ach} so that if $R < R_{\text{ach}}$, then (n, d, k) is achievable.
- Conversely, there is another rate R_{conv} so that if $R > R_{\text{conv}}$, then recovery is not possible.



Conclusions

- Provided an information-theoretic definition of saliency motivated by error exponents in hypothesis testing.

Conclusions

- Provided an information-theoretic definition of saliency motivated by error exponents in hypothesis testing.
- Provided necessary and sufficient conditions for salient set recovery.
- Number of samples n can be much smaller than d, total number of variables.

Conclusions

- Provided an information-theoretic definition of saliency motivated by error exponents in hypothesis testing.
- Provided necessary and sufficient conditions for salient set recovery.
- Number of samples n can be much smaller than d, total number of variables.
- In the paper, we provide a computationally efficient and consistent algorithm to search for S_d when $P^{(d)}$ and $Q^{(d)}$ are Markov on trees.