Scalable robust hypothesis tests using graphical models

Umamahesh Srinivas

iPAL Group Meeting

October 22, 2010

Binary hypothesis testing problem

Random vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ generated from either of two hypotheses

$$H_0 : \quad \mathbf{x} \sim g(\mathbf{x}|H_0)$$
$$H_1 : \quad \mathbf{x} \sim g(\mathbf{x}|H_1)$$

Given: Training sets $\mathcal{T}_0$ and $\mathcal{T}_1$, K samples each

Goal: Classify new sample as coming from H_0 or H_1

Assumptions:

Conditional densities $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ known exactly

Samples in $\mathcal{T}_0$ and $\mathcal{T}_1$ generated i.i.d. from $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ respectively

Likelihood ratio test (LRT)

$$L(\mathbf{x}) := \frac{g(\mathbf{x}|H_1)}{g(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtreqless}} \tau \quad (1)$$

Binary hypothesis testing problem

Random vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ generated from either of two hypotheses

$$H_0 : \quad \mathbf{x} \sim g(\mathbf{x}|H_0)$$
$$H_1 : \quad \mathbf{x} \sim g(\mathbf{x}|H_1)$$

Given: Training sets $\mathcal{T}_0$ and $\mathcal{T}_1$, K samples each

Goal: Classify new sample as coming from H_0 or H_1

Assumptions:

Conditional densities $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ known exactly

Samples in $\mathcal{T}_0$ and $\mathcal{T}_1$ generated i.i.d. from $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ respectively

Likelihood ratio test (LRT)

$$L(\mathbf{x}) := \frac{g(\mathbf{x}|H_1)}{g(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau \qquad (1)$$

Binary hypothesis testing problem

Random vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ generated from either of two hypotheses

$$H_0 : \quad \mathbf{x} \sim g(\mathbf{x}|H_0)$$
$$H_1 : \quad \mathbf{x} \sim g(\mathbf{x}|H_1)$$

Given: Training sets $\mathcal{T}_0$ and $\mathcal{T}_1$, K samples each

Goal: Classify new sample as coming from H_0 or H_1

Assumptions:

Conditional densities $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ known exactly

Samples in $\mathcal{T}_0$ and $\mathcal{T}_1$ generated i.i.d. from $g(\mathbf{x}|H_0)$ and $g(\mathbf{x}|H_1)$ respectively

Likelihood ratio test (LRT)

$$L(\mathbf{x}) := \frac{g(\mathbf{x}|H_1)}{g(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtreqless}} \tau \tag{1}$$

Need for robustness

Assumption of knowledge of true densities unrealistic:

- Limited training
- Training data acquired in the presence of noise
- Dynamically evolving conditional densities
- Secondary physical effects on signal not modeled

Robust hypothesis test[1] (RHT):

- Uncertainty in knowledge of true densities modeled as class of distributions in the proximity of some nominal density
- Minimum level of performance guaranteed for all models in the vicinity of nominal density

[1] Huber, 1965

Need for robustness

Assumption of knowledge of true densities unrealistic:

- Limited training
- Training data acquired in the presence of noise
- Dynamically evolving conditional densities
- Secondary physical effects on signal not modeled

Robust hypothesis test[1] (RHT):

- Uncertainty in knowledge of true densities modeled as class of distributions in the proximity of some nominal density
- Minimum level of performance guaranteed for all models in the vicinity of nominal density

[1] Huber, 1965

Measures of model proximity

Contamination model:

$$\mathcal{F}_k^c = \{f(\mathbf{x}) : f(\mathbf{x}) = (1-\epsilon_k) f_k(\mathbf{x}) + \epsilon_k h(\mathbf{x})\},\ k = 0, 1,$$

where $f_k(\mathbf{x})$ are the nominal densities, $0 \leq \epsilon_0, \epsilon_1 \leq 1$, and $h(\mathbf{x})$ is an *unknown* probability density.

Total variation:

$$\mathcal{F}_k^{TV} = \{f(\mathbf{x}) : d_{TV}(f_k, f) = \int |f_k(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} < \epsilon\},\ k = 0, 1.$$

Kullback-Leibler divergence:

$$\mathcal{F}_k^{KL} = \{f(\mathbf{x}) : D(f_k|f) = \int f_k(\mathbf{x}) \ln\left(\frac{f_k(\mathbf{x})}{f(\mathbf{x})}\right) d\mathbf{x} < \epsilon\},\ k = 0, 1.$$

Problem set-up

$\mathcal{D}$: convex set of pointwise randomized decision functions $\delta(\cdot)$.

For observation $\mathbf{x}$, we select H_1 with probability $\delta(\mathbf{x})$ and H_0 with probability $1 - \delta(\mathbf{x})$.

$$\text{False alarm: } P_F(\delta, f_0) = \int \delta(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \tag{2}$$

$$\text{Miss: } P_M(\delta, f_1) = \int (1 - \delta(\mathbf{x})) f_1(\mathbf{x}) d\mathbf{x} \,. \tag{3}$$

For equally likely hypotheses, the probability of error is given by

$$P_E(\delta, f_0, f_1) = \frac{1}{2} \left[P_F(\delta, f_0) + P_M(\delta, f_1) \right] \,. \tag{4}$$

Minimax RHT

$$(\delta_R, f_0^L(\mathbf{x}), f_1^L(\mathbf{x})) = \arg\min_{\delta \in \mathcal{D}} \max_{f_0, f_1 \in \mathcal{F}^c} P_E(\delta, f_0, f_1)\,, \qquad (5)$$

where

- δ_R is the robust test
- (f_0^L, f_1^L) are least favorable densities in $\mathcal{F}^c = \mathcal{F}_0^c \times \mathcal{F}_1^c$.

Solution to minimax RHT

$$f_0^L(\mathbf{x}) = \begin{cases} (1-\epsilon_0) f_0(\mathbf{x}) & \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < c'' \\ \frac{1}{c''}(1-\epsilon_0) f_1(\mathbf{x}) & \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \geq c'' \end{cases} \tag{6}$$

$$f_1^L(\mathbf{x}) = \begin{cases} (1-\epsilon_1) f_1(\mathbf{x}) & \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > c' \\ c'(1-\epsilon_1) f_0(\mathbf{x}) & \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \leq c' \end{cases} \tag{7}$$

$$\delta_R(\mathbf{x}) = \begin{cases} 1 & \frac{f_1^L(\mathbf{x})}{f_0^L(\mathbf{x})} \geq 1 \\ 0 & \frac{f_1^L(\mathbf{x})}{f_0^L(\mathbf{x})} < 1 \end{cases}, \tag{8}$$

where c' and c'' are defined such that f_0^L and f_1^L are valid probability distributions, leading to:

$$P_0\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < c''\right) + \frac{1}{c''} P_1\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \geq c''\right) = \frac{1}{1-\epsilon_0} \tag{9}$$

$$P_1\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > c'\right) + c' P_0\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \leq c'\right) = \frac{1}{1-\epsilon_1}. \tag{10}$$

P_k is the probability measure w.r.t $f_k(\mathbf{x})$.

Underlying intuition

Choice of c' and c'':

Consider

$$L(\mathbf{x}) = \frac{g(\mathbf{x}|H_1)}{g(\mathbf{x}|H_0)} = \prod_{i=1}^{n} \frac{g(x_i|H_1)}{g(x_i|H_0)}.$$

If any factor in the product approaches 0 or ∞, $L(\mathbf{x})$ is affected.

Introduce robustness by clipping the likelihood ratios to the range c', c''.

Least favorable densities:

Choose $f_0^L(\mathbf{x})$ "as close as possible" to $f_1(\mathbf{x})$, and $f_1^L(\mathbf{x})$ "as close as possible" to $f_0(\mathbf{x})$.

Scalability challenge

- RHT reduces to finding c' and c'' such that:

$$P_0\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < c''\right) + \frac{1}{c''}P_1\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \geq c''\right) = \frac{1}{1-\epsilon_0}$$
$$P_1\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > c'\right) + c' P_0\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \leq c'\right) = \frac{1}{1-\epsilon_1}\,.$$

- Highly nonlinear equations; require Monte Carlo methods (sample generation).
- Scales very poorly with dimension - computationally intractable.

Probabilistic graphical models

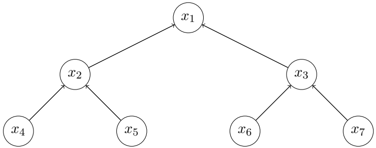
- Graph $G = (V, E)$ is defined by a set of nodes $V = \{1, \ldots, n\}$, and a set of edges $E \subset V \times V$ which connect pairs of nodes.
- Graphical model: Random vector defined on a graph such that each node represents one (or more) random variables, and edges reveal conditional dependencies.
- Underlying graph structure leads to factorization of joint probability distribution.
- Leverage efficient graph-based algorithms for statistical inference and learning.
- Trade-off between graph complexity and approximation accuracy.

Probabilistic graphical models

- Graph $G = (V, E)$ is defined by a set of nodes $V = \{1, \ldots, n\}$, and a set of edges $E \subset V \times V$ which connect pairs of nodes.
- Graphical model: Random vector defined on a graph such that each node represents one (or more) random variables, and edges reveal conditional dependencies.
- Underlying graph structure leads to factorization of joint probability distribution.
- Leverage efficient graph-based algorithms for statistical inference and learning.
- Trade-off between graph complexity and approximation accuracy.

Some graph structures

Tree:



$$f(\mathbf{x}) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2)f(x_5|x_2)f(x_6|x_3)f(x_7|x_3).$$

- Undirected acyclic graph with exactly $(n-1)$ edges.
- Chow-Liu (1965): optimal tree approximation reduces to a maximum weight spanning tree (MWST) problem.

Forest:

- Graph with $k < (n-1)$ edges.

Junction-tree:

- Tree-structured graph with edges between clusters of nodes.
- Clusters connected by an edge have at least one common node.

Block-tree graphs

Disjoint clusters of nodes, with only one path connecting any two clusters.

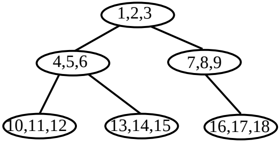


Figure: Example of a block-tree graph

Benefits:

- Favorable complexity-performance trade-off
- Low cost of sample generation
- Efficient greedy algorithms to compute block-trees.

Realizing RHT on block-tree graphs

Suppose $f(\mathbf{x})$ is Gaussian with mean zero.
State-space model on the block-tree graph[2] is given as:

$$x_{C_i} = A_i x_{C_{\Upsilon(i)}} + u_{C_i}\,, \tag{11}$$

$$A_i = E(x_{C_i} x^T_{C_{\Upsilon(i)}})[E(x_{C_{\Upsilon(i)}} x^T_{C_{\Upsilon(i)}})]^{-1} \tag{12}$$

$$E(u_{C_i} u^T_{C_i}) = E(x_{C_i} x^T_{C_i}) - A_i E(x_{C_{\Upsilon(i)}} x^T_{C_i})\,, \tag{13}$$

where u_{C_i} is white noise.

Computing c' and c'':

1. For each $f_k(\mathbf{x})$, compute block-tree graphs $\mathcal{G}_k$ using a specified value of m (number of nodes in a cluster). Using recursive sampling, generate sample sets $\mathcal{S}_k$, $k = 0, 1$.
2. Using $\mathcal{S}_0$ and $\mathcal{S}_1$, compute c' and c'' by Monte Carlo methods.

---
[2] Vats and Moura, 2010

Complexity benefits

- Assuming Gaussianity, generating a sample from $f(\mathbf{x})$ is $O(n^3)$ - inversion of an $n \times n$ matrix.
  For L generated samples, total complexity is $O(Ln^3)$.
- Using block-tree graph with cluster size m, computing block-tree graph has complexity $O(\log n) + O(mn^2) \approx O(mn^2)$, while generating samples has complexity $O(rm^3) = O(m^2 n)$.
  For L' generated samples, total complexity is $O(L'(mn^2 + m^2 n)) \approx O(L'mn^2)$.
- Reduction in complexity for sparse graphical models, since $m \ll n$ and $L' \ll L$.

Results

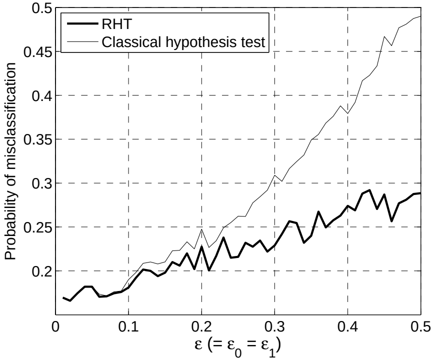


Figure: Error probability as a function of ϵ for classical hypothesis testing and RHT.

Results

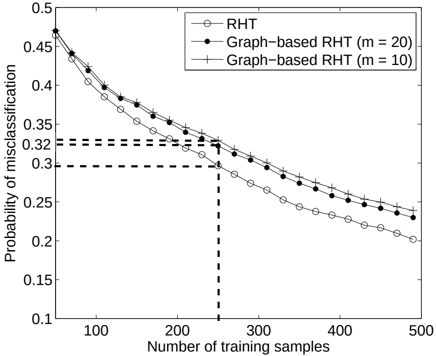


Figure: Error probability as a function of training size, for RHT and graph -based RHT (dense inverse covariance matrix).

Results

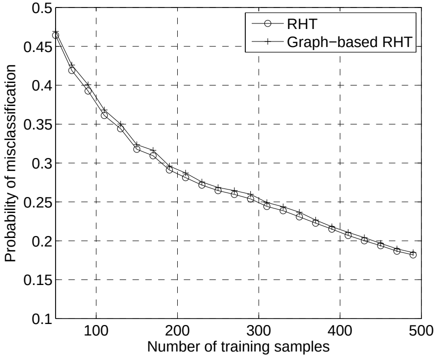


Figure: Error probability as a function of training size, for RHT and graph -based RHT (sparse inverse covariance matrix).

Results

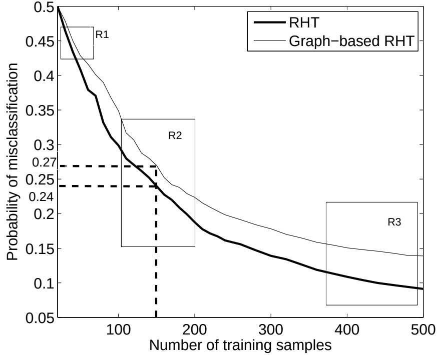


Figure: Automatic target recognition: Misclassification probability as a function of number training samples for graph-based RHT and RHT. Classification is performed on real-world SAR images.

Summary

- Real-world classification problems: high-dimensional data, limited training, noisy acquisition $\rightarrow$ need for robust hypothesis tests.
- Minimax test minimizes worst-case performance of making a decision via pursuit of least favorable densities.
- RHT is computationally intractable for high-dimensional data.
- Approximate densities by block-tree graphs and instantiate RHT - significant computational benefits with tolerable loss in classification performance.