Topics in Two-Sample Testing

Nelson Ray
(joint work with Susan Holmes)

Stanford University

March 3, 2013

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [?]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [?]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [?]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [**?**]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [**?**]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

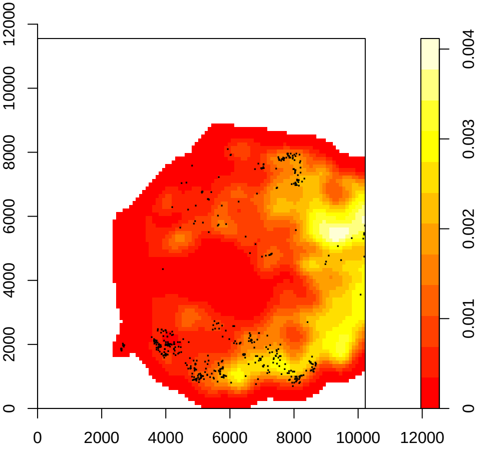
Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [?]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [?]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation t-test
- Permutation dependence: Stein's method for rates of convergence bounds
- Simulations to verify bounds in proof (experimental mathematics)
- Kernel-based two sample tests for non-vectorial data
- Multiple Kernel Learning for heterogeneous data

Breast Cancer Data: Spatial



Breast Cancer Data: Survival

Pathology no.	Initial Diagnosis Date	Relapse or Disease Free	RDF (R=relapsed; F=DF)	Recurrence Date	Las
98_17969D	1997-08-25	Disease Free	F	Disease Free	
97_24046C8	1997-08-25	Disease Free	F	Disease Free	
98_8501C1	1998-04-03	Disease Free	F	Disease Free	
98_8501A1	1998-04-03	Disease Free	F	Disease Free	
98_9134D4	1998-04-09	Left in-situ BrCa in 1999 (2nd primary cancer, not a metastasis from the right BrCa in 1997)	F	Disease Free	
98_9134B	1998-04-09	Left in-situ BrCa in 1999 (2nd primary cancer, not a metastasis from the right BrCa in 1997)	F	Disease Free	
98_14783B1	1998-06-10	bone, brain, lymph nodes, pericardium, liver metastasis	R	2004-07-30	
98_14783A	1998-06-10	bone, brain, lymph nodes, pericardium, liver metastasis	R	2004-07-30	
98_16169C2	1998-06-24	Disease Free	F	Disease Free	
98_16169A	1998-06-24	Disease Free	F	Disease Free	
98_16169B	1998-06-24	Disease Free	F	Disease Free	
98_16253C1	1998-06-25	Disease Free	F	Disease Free	
60C1	1998-07-10	Disease Free	F	Disease Free	

Breast Cancer Data: Medical

Pathology no.	Age at time of diagnosis	Gender	SLN tumor status	Diagnosis	ER status	PR status	Her-2 overexpressio
98_17969D	68	F	+	Invasive ductal carcinoma (IDC)	-	-	-
97_24046C8	68	F	+	Invasive ductal carcinoma (IDC)	-	-	-
98_8501C1	51	F	+	IDC & DCIS	+	+	?
98_8501A1	51	F	+	IDC & DCIS	+	+	?
98_9134D4	70	F	+	IDC	+	+	n/a
98_9134B	70	F	+	IDC	+	+	n/a
98_14783B1	67	F	+	IDC & DCIS	+	+	+
98_14783A	67	F	+	IDC & DCIS	+	+	+
98_16169C2	79	F	+mic	IDC	+	+	+
98_16169A	79	F	+mic	IDC	+	+	+
98_16169B	79	F	+mic	IDC	+	+	+
98_16253C1	70	F	+mic	IDC & DCIS	+	-	-
60C1	51	F	- (rare keratin+ cells)	IDC & DCIS	+	+	+
				IDC; DCIS;			

Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?
- Two-sample tests

Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?
- Two-sample tests

Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?
- Two-sample tests

Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?
- Two-sample tests

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$ from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ from $q(\mathbf{z})$ testing
$\mathcal{H}_A$: $p \neq q$ against $\mathcal{H}_0$: $p = q$

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$.
2. Assign label $y_i = 1$ to the first group and $y_i = -1$ to the second group.
3. Apply a binary classification learning machine f to the training data to score the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.
4. Calculate a univariate two-sample test statistic $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.
5. Determine the permutation null distribution of the above statistic to yield a p-value.

Permutation t-test Connection

With univariate data and linear scoring functions/kernels, Friedman's test reduces to the permutation t-test (normal convergence result). With multivariate/non-vectorial/heterogeneous data and arbitrary kernels, null distribution is consistent with the Normal.

Permutation t-test Connection

With univariate data and linear scoring functions/kernels, Friedman's test reduces to the permutation t-test (normal convergence result). With multivariate/non-vectorial/heterogeneous data and arbitrary kernels, null distribution is consistent with the Normal.

Other Work

- Fisher (1935) [**?**] proposed distribution-free randomization test.
- Lehmann [**?**] proved a normal convergence result for the randomization distribution.
- Bentkus et al. [**?**], Shao [**?**] proved Berry-Esseen bounds for Student's t-statistic in independent case.

Other Work

- Fisher (1935) [**?**] proposed distribution-free randomization test.
- Lehmann [**?**] proved a normal convergence result for the randomization distribution.
- Bentkus et al. [**?**], Shao [**?**] proved Berry-Esseen bounds for Student's t-statistic in independent case.

Other Work

- Fisher (1935) [**?**] proposed distribution-free randomization test.
- Lehmann [**?**] proved a normal convergence result for the randomization distribution.
- Bentkus et al. [**?**], Shao [**?**] proved Berry-Esseen bounds for Student's t-statistic in independent case.

Stein's Method and the Randomization Distribution

Let $\Phi(t)$ denote the standard normal CDF.
Can we get a bound on

$$\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)|?$$

$\mathcal{O}(N^{-1/4})$ with mild conditions on the data and $\mathcal{O}(N^{-1/2})$ with an additional condition

Stein's Method and the Randomization Distribution

Let $\Phi(t)$ denote the standard normal CDF.
Can we get a bound on

$$\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)|?$$

$\mathcal{O}(N^{-1/4})$ with mild conditions on the data and $\mathcal{O}(N^{-1/2})$ with an additional condition

Other Results

Theorem (Berry-Esseen)

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2 > 0$, and $\mathbb{E}|X_i|^3 = \rho < \infty$. Let $F_n(x)$ denote the CDF of standardized sample mean of the X_i. Then

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{0.33477(\rho + 0.429\sigma^3)}{\sigma^3\sqrt{n}}$$

$$= \frac{C}{\sqrt{n}} f(\rho, \sigma).$$

Note that ρ and σ are fixed as $n \to \infty$.

Other Results

Theorem (Berry-Esseen)

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2 > 0$, and $\mathbb{E}|X_i|^3 = \rho < \infty$. Let $F_n(x)$ denote the CDF of standardized sample mean of the X_i. Then

$$\begin{aligned}\sup_x |F_n(x) - \Phi(x)| &\leq \frac{0.33477(\rho + 0.429\sigma^3)}{\sigma^3\sqrt{n}} \\ &= \frac{C}{\sqrt{n}} f(\rho, \sigma).\end{aligned}$$

Note that ρ and σ are fixed as $n \to \infty$.

Other Results

Theorem (Berry-Esseen)

Suppose $X_1, \dots, X_n$ are i.i.d. random variables with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2 > 0$, and $\mathbb{E}|X_i|^3 = \rho < \infty$. Let $F_n(x)$ denote the CDF of standardized sample mean of the X_i. Then

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{0.33477(\rho + 0.429\sigma^3)}{\sigma^3\sqrt{n}} = \frac{C}{\sqrt{n}} f(\rho, \sigma).$$

Note that ρ and σ are fixed as $n \to \infty$.

Other Results

Theorem (Hoeffding, Stein)

Let $A = \{a_{ij}\}_{i,j\in\{1,\ldots,n\}}$ be a square array of numbers such that $\sum_j a_{ij} = 0$ for all i, $\sum_i a_{ij} = 0$ for all j, and $\sum_i \sum_j a_{ij}^2 = n - 1$. Then with $F_n(x) = P(\sum_i a_{i\Pi(i)} \leq x)$,

$$|F_n(x) - \Phi(x)| \leq \frac{C}{\sqrt{n}} \left(\sqrt{\sum_{i,j} a_{ij}^4} + \sqrt{\sum_{i,j} |a_{ij}|^3} \right)$$

$$= \frac{C}{\sqrt{n}} f(A).$$

Given a sampling scheme for A, $f(A)$ must be $\mathcal{O}(1)$ to have rate $\mathcal{O}(n^{-1/2})$.

Other Results

Theorem (Hoeffding, Stein)

Let $A = \{a_{ij}\}_{i,j\in\{1,\ldots,n\}}$ *be a square array of numbers such that* $\sum_j a_{ij} = 0$ *for all* i, $\sum_i a_{ij} = 0$ *for all* j, *and* $\sum_i \sum_j a_{ij}^2 = n-1$. *Then with* $F_n(x) = P(\sum_i a_{i\Pi(i)} \leq x)$,

$$\begin{aligned} |F_n(x) - \Phi(x)| &\leq \frac{C}{\sqrt{n}} \left(\sqrt{\sum_{i,j} a_{ij}^4} + \sqrt{\sum_{i,j} |a_{ij}|^3} \right) \\ &= \frac{C}{\sqrt{n}} f(A). \end{aligned}$$

Given a sampling scheme for A, $f(A)$ must be $\mathcal{O}(1)$ to have rate $\mathcal{O}(n^{-1/2})$.

Other Results

Theorem (Hoeffding, Stein)

Let $A = \{a_{ij}\}_{i,j\in\{1,\dots,n\}}$ *be a square array of numbers such that* $\sum_j a_{ij} = 0$ *for all* i, $\sum_i a_{ij} = 0$ *for all* j, *and* $\sum_i \sum_j a_{ij}^2 = n - 1$. *Then with* $F_n(x) = P(\sum_i a_{i\Pi(i)} \le x)$,

$$
\begin{aligned}
|F_n(x) - \Phi(x)| &\le \frac{C}{\sqrt{n}} \left(\sqrt{\sum_{i,j} a_{ij}^4} + \sqrt{\sum_{i,j} |a_{ij}|^3} \right) \\
&= \frac{C}{\sqrt{n}} f(A).
\end{aligned}
$$

Given a sampling scheme for A, $f(A)$ must be $\mathcal{O}(1)$ to have rate $\mathcal{O}(n^{-1/2})$.

Exchangeable Pair

Assume $M = N$. Fix data $\{u_1, \ldots, u_N, u_{N+1}, \ldots, u_{2N}\}$. Π is a uniformly random permutation, and let

$$T = T\left(\{u_{\Pi(i)}\}_{i=1}^{N}, \{u_{\Pi(i)}\}_{i=N+1}^{2N}\right).$$

Let $(I, J) = (i, j)$ w.p. $\frac{1}{N^2}$ for $1 \leq i \leq N$ and $N + 1 \leq j \leq 2N$. Then

$$T' = T\left(\{u_{\Pi\circ(I,J)(i)}\}_{i=1}^{N}, \{u_{\Pi\circ(I,J)(i)}\}_{i=N+1}^{2N}\right).$$

T and T' form an exchangeable pair.

Exchangeable Pair

Assume $M = N$. Fix data $\{u_1, \ldots, u_N, u_{N+1}, \ldots, u_{2N}\}$. Π is a uniformly random permutation, and let

$$T = T\left(\{u_{\Pi(i)}\}_{i=1}^{N}, \{u_{\Pi(i)}\}_{i=N+1}^{2N}\right).$$

Let $(I, J) = (i, j)$ w.p. $\frac{1}{N^2}$ for $1 \leq i \leq N$ and $N + 1 \leq j \leq 2N$. Then

$$T' = T\left(\{u_{\Pi\circ(I,J)(i)}\}_{i=1}^{N}, \{u_{\Pi\circ(I,J)(i)}\}_{i=N+1}^{2N}\right).$$

T and T' form an exchangeable pair.

Exchangeable Pair

Assume $M = N$. Fix data $\{u_1, \ldots, u_N, u_{N+1}, \ldots, u_{2N}\}$. Π is a uniformly random permutation, and let

$$T = T\left(\{u_{\Pi(i)}\}_{i=1}^{N}, \{u_{\Pi(i)}\}_{i=N+1}^{2N}\right).$$

Let $(I, J) = (i, j)$ w.p. $\frac{1}{N^2}$ for $1 \leq i \leq N$ and $N + 1 \leq j \leq 2N$. Then

$$T' = T\left(\{u_{\Pi\circ(I,J)(i)}\}_{i=1}^{N}, \{u_{\Pi\circ(I,J)(i)}\}_{i=N+1}^{2N}\right).$$

T and T' form an exchangeable pair.

Main Theorem

Theorem

If T, T' are mean 0, exchangeable random variables with variance $\mathbb{E}[T^2]$ satisfying

$$\mathbb{E}[T' - T|T] = -\lambda(T - R)$$

for some $\lambda \in (0,1)$ and some random variable R, then $\sup_{t\in\mathbb{R}} |P(T \leq t) - \Phi(t)|$ *is bounded by*

$$\underbrace{(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T' - T|^3}{\lambda}}}_{\leq N^{-1/4} f_1(\mathbf{u})} + \underbrace{\frac{1}{2\lambda}\sqrt{\mathrm{var}(\mathbb{E}[(T' - T)^2|T])}}_{\leq N^{-1} f_2(\mathbf{u})}$$

$$\underbrace{|\mathbb{E}T^2 - 1|}_{\leq N^{-1} f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\leq N^{-1/2} f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\leq N^{-1/2} f_5(\mathbf{u})} \leq N^{-1/4} f_6(\mathbf{u})$$

Main Theorem

Theorem

If T, T' are mean 0, exchangeable random variables with variance $\mathbb{E}[T^2]$ satisfying

$$\mathbb{E}[T' - T | T] = -\lambda(T - R)$$

for some $\lambda \in (0,1)$ and some random variable R, then $\sup_{t\in\mathbb{R}} |P(T \le t) - \Phi(t)|$ *is bounded by*

$$\underbrace{(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T' - T|^3}{\lambda}}}_{\le N^{-1/4} f_1(\mathbf{u})} + \underbrace{\frac{1}{2\lambda}\sqrt{\mathrm{var}(\mathbb{E}[(T' - T)^2|T])}}_{\le N^{-1} f_2(\mathbf{u})}$$

$$\underbrace{|\mathbb{E}T^2 - 1|}_{\le N^{-1} f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\le N^{-1/2} f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\le N^{-1/2} f_5(\mathbf{u})} \quad \le N^{-1/4} f_6(\mathbf{u})$$

Main Theorem

Theorem

If T, T' are mean 0, exchangeable random variables with variance $\mathbb{E}[T^2]$ satisfying

$$\mathbb{E}[T' - T|T] = -\lambda(T - R)$$

for some $\lambda \in (0,1)$ and some random variable R, then $\sup_{t\in\mathbb{R}} |P(T \le t) - \Phi(t)|$ *is bounded by*

$$\underbrace{(2\pi)^{-1/4}\sqrt{\frac{\mathbb{E}|T' - T|^3}{\lambda}}}_{\le N^{-1/4} f_1(\mathbf{u})} + \underbrace{\frac{1}{2\lambda}\sqrt{\mathrm{var}(\mathbb{E}[(T' - T)^2|T])}}_{\le N^{-1} f_2(\mathbf{u})}$$

$$\underbrace{|\mathbb{E}T^2 - 1|}_{\le N^{-1} f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\le N^{-1/2} f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\le N^{-1/2} f_5(\mathbf{u})} \le N^{-1/4} f_6(\mathbf{u})$$

Main Theorem (Improved Rate)

Theorem

If in addition $|T' - T| \le \delta$, $\sup_{t\in\mathbb{R}} |P(T \le t) - \Phi(t)|$ *is bounded by*

$$\underbrace{\frac{.41\delta^3}{\lambda}}_{\le N^{-1/2}c_1''^*} + \underbrace{3\delta(\sqrt{\mathbb{E}T^2} + \mathbb{E}|R|)}_{\le N^{-1}f_1'(\mathbf{u})^*} + \underbrace{\frac{1}{2\lambda}\sqrt{\mathrm{var}(\mathbb{E}[(T'-T)^2|T])}}_{\le N^{-1}f_2(\mathbf{u})}$$

$$\underbrace{|\mathbb{E}T^2 - 1|}_{\le N^{-1}f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\le N^{-1/2}f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\le N^{-1/2}f_5(\mathbf{u})} \le N^{-1/2}f_6'(\mathbf{u})^*$$

* if $\delta < c_1' N^{-1/2}$

Main Theorem (Improved Rate)

Theorem

If in addition $|T' - T| \leq \delta$, $\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)|$ *is bounded by*

$$\underbrace{\frac{.41\delta^3}{\lambda}}_{\leq N^{-1/2} c_1''^*} + \underbrace{3\delta(\sqrt{\mathbb{E}T^2} + \mathbb{E}|R|)}_{\leq N^{-1} f_1'(\mathbf{u})^*} + \underbrace{\frac{1}{2\lambda}\sqrt{\mathrm{var}(\mathbb{E}[(T' - T)^2 | T])}}_{\leq N^{-1} f_2(\mathbf{u})}$$

$$\underbrace{|\mathbb{E}T^2 - 1|}_{\leq N^{-1} f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\leq N^{-1/2} f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\leq N^{-1/2} f_5(\mathbf{u})} \quad \leq N^{-1/2} f_6'(\mathbf{u})^*$$

* if $\delta < c_1' N^{-1/2}$

Main Theorem (Improved Rate)

Theorem

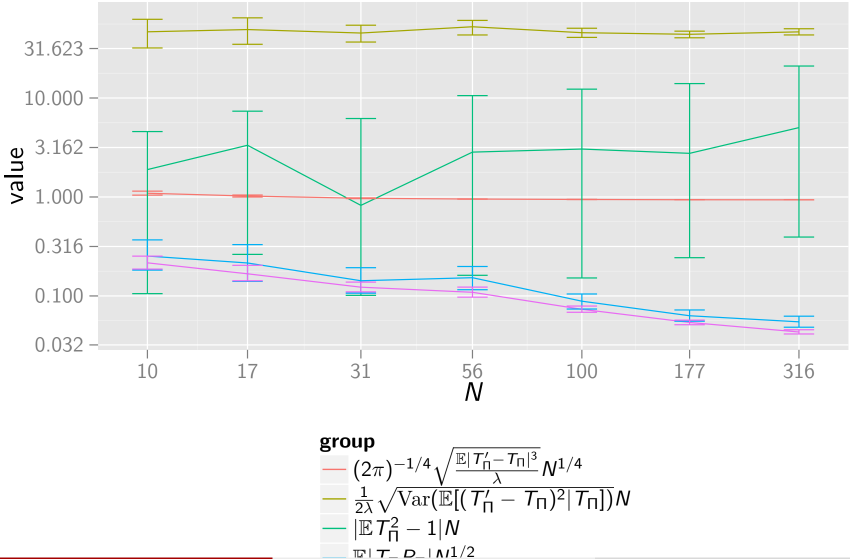
If in addition $|T' - T| \leq \delta$, $\sup_{t\in\mathbb{R}} |P(T \leq t) - \Phi(t)|$ *is bounded by*

$$\underbrace{\frac{.41\delta^3}{\lambda}}_{\leq N^{-1/2} c_1''^{*}} + \underbrace{3\delta(\sqrt{\mathbb{E}T^2} + \mathbb{E}|R|)}_{\leq N^{-1} f_1'(\mathbf{u})^{*}} + \underbrace{\frac{1}{2\lambda}\sqrt{\operatorname{var}(\mathbb{E}[(T' - T)^2 | T])}}_{\leq N^{-1} f_2(\mathbf{u})}$$

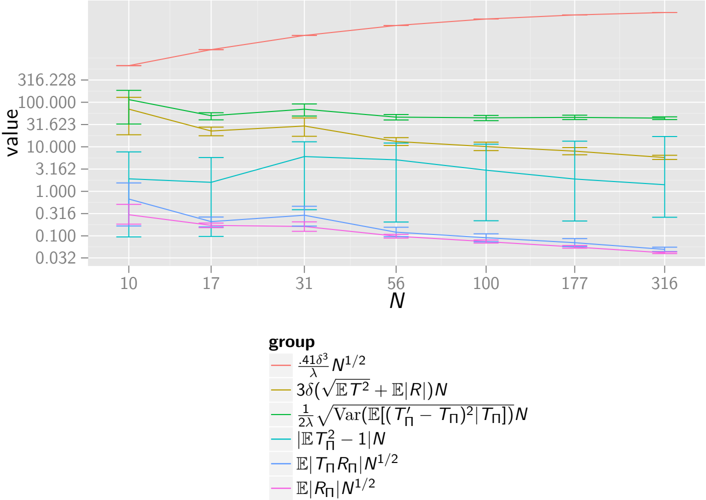
$$\underbrace{|\mathbb{E}T^2 - 1|}_{\leq N^{-1} f_3(\mathbf{u})} + \underbrace{\mathbb{E}|TR|}_{\leq N^{-1/2} f_4(\mathbf{u})} + \underbrace{\mathbb{E}|R|}_{\leq N^{-1/2} f_5(\mathbf{u})} \leq N^{-1/2} f_6'(\mathbf{u})^{*}$$

* if $\delta < c_1' N^{-1/2}$

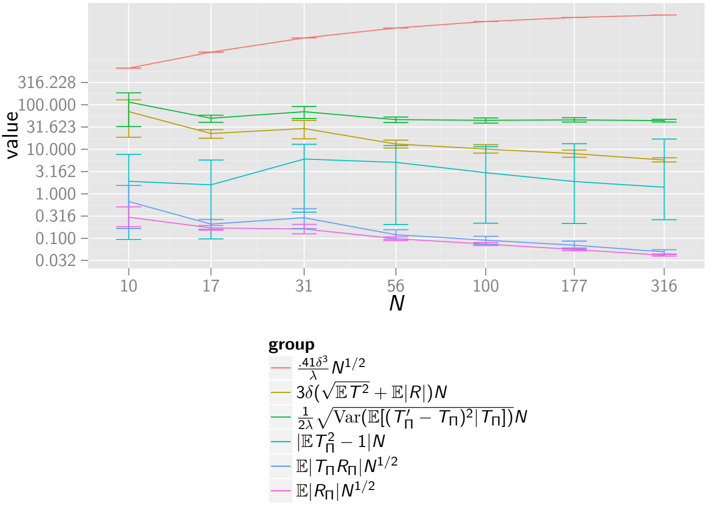
Simulated Bounds



Simulated Bounds (Improved Rate)



Simulated Bounds (Improved Rate)



Twitter Example

Twitter Data

Raw:

```
"BarackObama: We need to reward education reforms that are
driven not by Washington, but by principals and teachers and
parents. http://OFA.BO/6p2EMy"
"SarahPalinUSA: You betcha!! MT \"@AlaskaAces: Alaska Aces
are 2011 Kelly Cup Champs w/ 5-3 win over Kalamazoo Wings!
Aces win  ECHL Championship series 4-1\""
```

After pre-processing:

```
"we need to reward education reforms that are driven not by
washington but by principals and teachers and parents "
"you betcha mt alaskaaces alaska aces are  kelly cup champs
w  win over kalamazoo wings aces win  echl championship
series "
```

The Spectrum Kernel

Compares two strings based on the their length k contiguous subsequences.

- $\mathcal{X}$ is our input space, built up from an alphabet $\mathcal{A} = \{a, b, \ldots, z, \ \}$ with $|\mathcal{A}| = 27$.
- The k-spectrum ($k \geq 1$) of an input sequence is the set of all length k contiguous subsequences it contains.
- Define the feature map from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{A}|^k}$ by $\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$ where $\phi_a(x)$ is the number of times a occurs in x: $\{\#aaa, \#aab, \#aac, \ldots, \}$.
- $K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle$.

The Spectrum Kernel

Compares two strings based on the their length k contiguous subsequences.

- $\mathcal{X}$ is our input space, built up from an alphabet $\mathcal{A} = \{a, b, \dots, z, \ \}$ with $|\mathcal{A}| = 27$.
- The k-spectrum ($k \geq 1$) of an input sequence is the set of all length k contiguous subsequences it contains.
- Define the feature map from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{A}|^k}$ by $\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$ where $\phi_a(x)$ is the number of times a occurs in x: $\{\#aaa, \#aab, \#aac, \dots, \}$.
- $K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle$.

The Spectrum Kernel

Compares two strings based on the their length k contiguous subsequences.

- $\mathcal{X}$ is our input space, built up from an alphabet $\mathcal{A} = \{a, b, \ldots, z, \ \}$ with $|\mathcal{A}| = 27$.
- The k-spectrum $(k \geq 1)$ of an input sequence is the set of all length k contiguous subsequences it contains.
- Define the feature map from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{A}|^k}$ by $\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$ where $\phi_a(x)$ is the number of times a occurs in x: $\{\#aaa, \#aab, \#aac, \ldots, \}$.
- $K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle$.

The Spectrum Kernel

Compares two strings based on the their length k contiguous subsequences.

- $\mathcal{X}$ is our input space, built up from an alphabet $\mathcal{A} = \{a, b, \ldots, z, \ \}$ with $|\mathcal{A}| = 27$.
- The k-spectrum ($k \geq 1$) of an input sequence is the set of all length k contiguous subsequences it contains.
- Define the feature map from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{A}|^k}$ by $\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$ where $\phi_a(x)$ is the number of times a occurs in x: $\{\#aaa, \#aab, \#aac, \ldots, \}$.
- $K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle$.

The Spectrum Kernel

Compares two strings based on the their length k contiguous subsequences.

- $\mathcal{X}$ is our input space, built up from an alphabet $\mathcal{A} = \{a, b, \ldots, z, \ \}$ with $|\mathcal{A}| = 27$.
- The k-spectrum ($k \geq 1$) of an input sequence is the set of all length k contiguous subsequences it contains.
- Define the feature map from $\mathcal{X}$ to $\mathbb{R}^{|\mathcal{A}|^k}$ by $\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$ where $\phi_a(x)$ is the number of times a occurs in x: $\{\#aaa, \#aab, \#aac, \ldots, \}$.
- $K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle$.

Support Vector Machines for Regression

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0, \quad h_m(x) \text{ basis functions}$$

To estimate β and β_0, minimize

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum_{m=1}^{M} \beta_m^2.$$

V is taken to be ϵ-insensitive loss:

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases}$$

The solution has the form $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i)$, where $K(x, y) = \langle h(x), h(y) \rangle$.

Support Vector Machines for Regression

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0, \quad h_m(x) \text{ basis functions}$$

To estimate β and β_0, minimize

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum_{m=1}^{M} \beta_m^2.$$

V is taken to be ϵ-insensitive loss:

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases}$$

The solution has the form $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i)$, where $K(x, y) = \langle h(x), h(y) \rangle$.

Support Vector Machines for Regression

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0, \quad h_m(x) \text{ basis functions}$$

To estimate β and β_0, minimize

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum_{m=1}^{M} \beta_m^2.$$

V is taken to be ϵ-insensitive loss:

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases}$$

The solution has the form $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i)$, where $K(x, y) = \langle h(x), h(y) \rangle$.

Support Vector Machines for Regression

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0, \quad h_m(x) \text{ basis functions}$$

To estimate β and β_0, minimize

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum_{m=1}^{M} \beta_m^2.$$

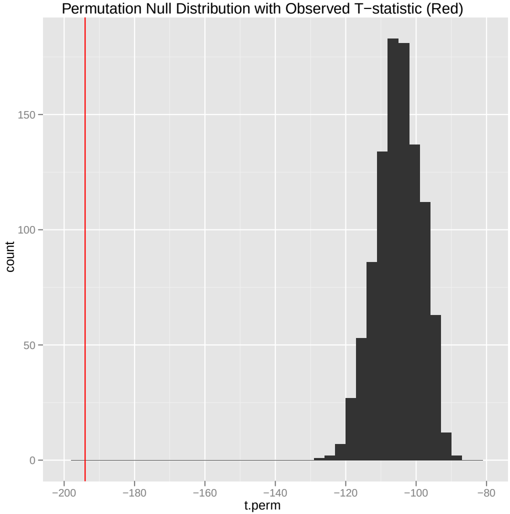
V is taken to be ϵ-insensitive loss:

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases}$$

The solution has the form $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i)$, where $K(x, y) = \langle h(x), h(y) \rangle$.

Twitter Example

$p < .001$:



Power Simulations at .05 Level

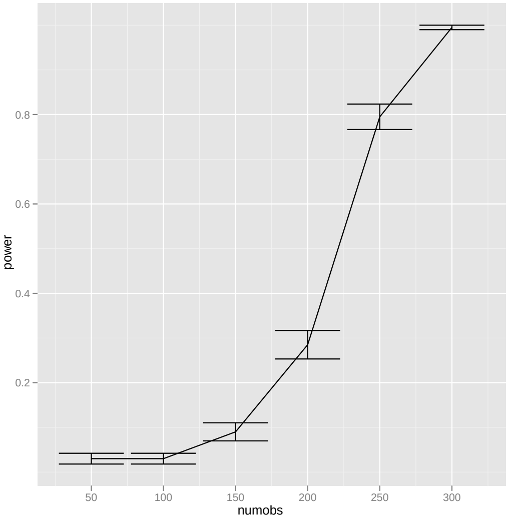


Image Data (Cars)

Caltech 101 Object Categories [**?**]
The cars are 300×197 grayscale.

Planes Before

The planes aren't.

Planes After

Polynomial Kernel

Each $m \times n$ grayscale image is converted to a vector of length $p = mn$.
Given $X \in \mathbb{R}^{n \times p}$, the linear kernel is given by

$$K(x, x') = \langle x, x' \rangle = \langle \Phi(x), \Phi(x') \rangle.$$

The kernel matrix is given simply by $XX^T \succeq 0$. This corresponds to the identity mapping: $\Phi(x) = x$.
The homogeneous polynomial kernel,

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle x, x' \rangle^d,$$

corresponds to the mapping
$\Phi(x) = [x_1^d, \ldots, x_p^d, x_1^{d-1} x_2, \ldots, x_p^{d-1} x_{p-1}]^T \in \mathbb{R}^{d'}$, where $d' = \binom{d+N-1}{d}$.

Polynomial Kernel

Each $m \times n$ grayscale image is converted to a vector of length $p = mn$. Given $X \in \mathbb{R}^{n \times p}$, the linear kernel is given by

$$K(x, x') = \langle x, x' \rangle = \langle \Phi(x), \Phi(x') \rangle.$$

The kernel matrix is given simply by $XX^T \succeq 0$. This corresponds to the identity mapping: $\Phi(x) = x$.

The homogeneous polynomial kernel,

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle x, x' \rangle^d,$$

corresponds to the mapping
$\Phi(x) = [x_1^d, \ldots, x_p^d, x_1^{d-1}x_2, \ldots, x_p^{d-1}x_{p-1}]^T \in \mathbb{R}^{d'}$, where $d' = \binom{d+N-1}{d}$.

Polynomial Kernel

Each $m \times n$ grayscale image is converted to a vector of length $p = mn$. Given $X \in \mathbb{R}^{n \times p}$, the linear kernel is given by

$$K(x, x') = \langle x, x' \rangle = \langle \Phi(x), \Phi(x') \rangle.$$

The kernel matrix is given simply by $XX^T \succeq 0$. This corresponds to the identity mapping: $\Phi(x) = x$.
The homogeneous polynomial kernel,

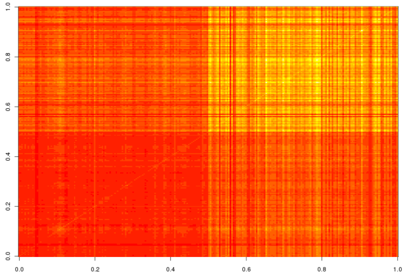
$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle x, x' \rangle^d,$$

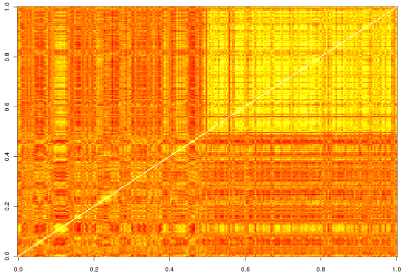
corresponds to the mapping
$\Phi(x) = [x_1^d, \ldots, x_p^d, x_1^{d-1} x_2, \ldots, x_p^{d-1} x_{p-1}]^T \in \mathbb{R}^{d'}$, where $d' = \binom{d+N-1}{d}$.

Standardization

In order to mitigate the effects of global differences in illumination, each vector is scaled so that it has mean zero and unit norm.

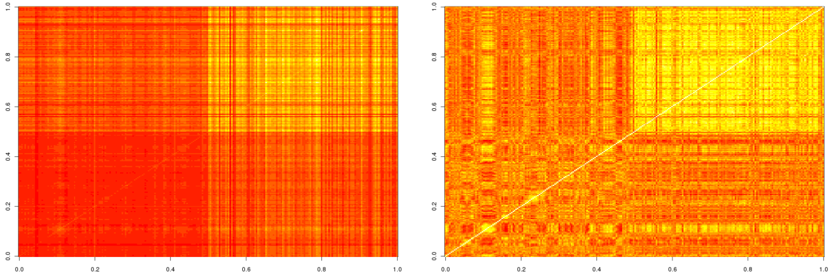
Unscaled linear kernel matrix, left; scaled, right



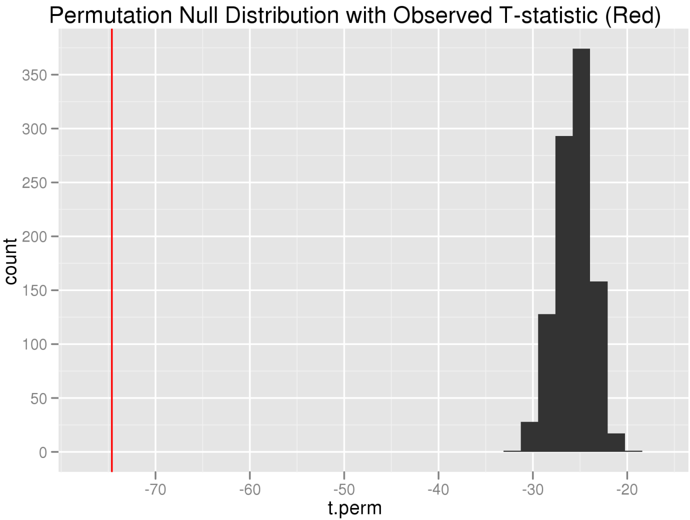


Standardization

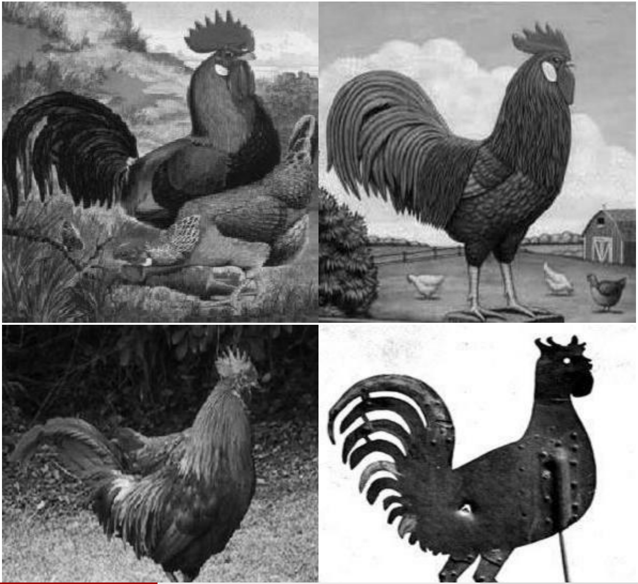
In order to mitigate the effects of global differences in illumination, each vector is scaled so that it has mean zero and unit norm.
Unscaled linear kernel matrix, left; scaled, right



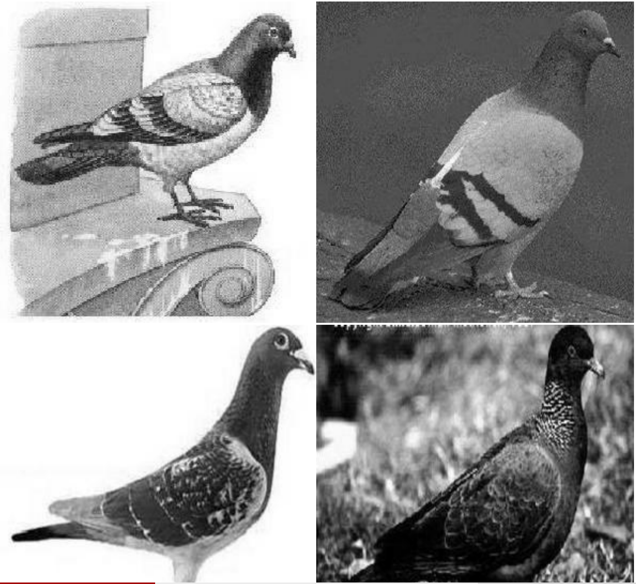
Car/Airplane Example (Linear Kernel)



Roosters

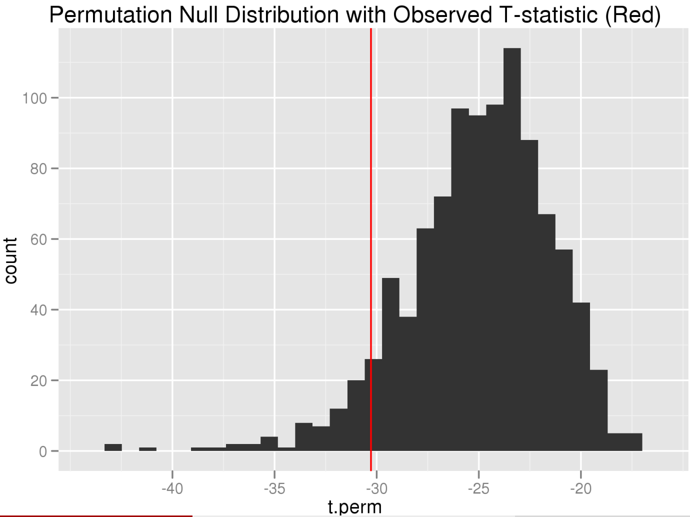


Pigeons



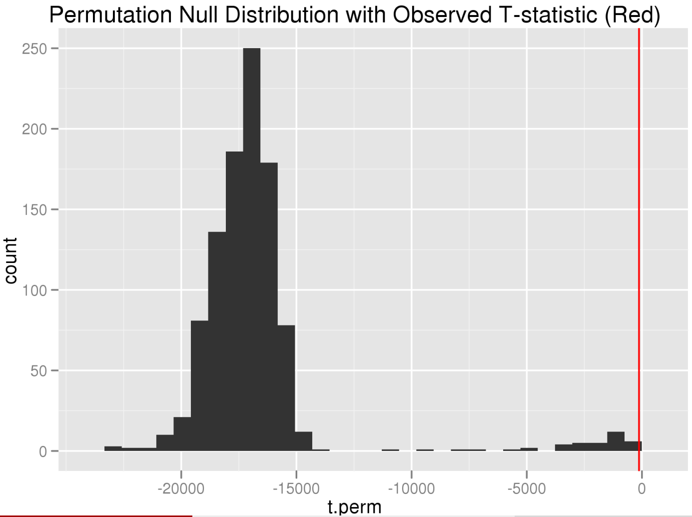
Rooster/Pigeon Example (Linear Kernel)

$p = .138$



Rooster/Pigeon Example (Inhomogeneous Degree 4)

$p < .001$



Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's T^2-test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's T^2-test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's T^2-test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's T^2-test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's T^2-test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

References I