Data-driven goodness-of-fit tests

Mikhail Langovoy

Institute for Applied Mathematics,
University of Bonn

September 1, 2008

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

Outline

Outline

Two Approaches

- Constructing good tests is an essential problem of statistics.
- Two approaches:
  - direct: "distance" between theoretical and empirical distribution is proposed as statistic
  - aiming optimality: construct tests which are asymptotically efficient (Neyman, Le Cam, Wald)

Two Approaches

- Constructing good tests is an essential problem of statistics.
- Two approaches:
  - direct: "distance" between theoretical and empirical distribution is proposed as statistic
  - aiming optimality: construct tests which are asymptotically efficient (Neyman, Le Cam, Wald)

Examples of Distance-Based Statistics

- Kolmogorov - Smirnov:

$$D_n = \sqrt{n} \, \|F_n - F\|_\infty$$

- Cramer - von Mises:

$$\omega_n^2 = n \int_{-\infty}^{\infty} \left(F_n(t) - F_0(t)\right)^2 d\, F_0(t)$$

- many other

About Distance-Based Tests

- These tests works
- asymptotically optimal only in a few directions of alternatives

Another type: Neyman's Statistic

- hypothesis $H_0 : X \sim U[0,1]$
- $\{\phi_j\}_{j=0}^{\infty}$ orthonormal basis of $L_2([0,1],\lambda)$
- $$N_k = \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2$$

Another type: Neyman's Statistic

- hypothesis $H_0 : X \sim U[0, 1]$
- $\{\phi_j\}_{j=0}^{\infty}$ orthonormal basis of $L_2([0, 1], \lambda)$
- $$N_k = \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2$$

Another type: Neyman's Statistic

- hypothesis $H_0 : X \sim U[0,1]$
- $\{\phi_j\}_{j=0}^{\infty}$ orthonormal basis of $L_2([0,1],\lambda)$
- $$N_k = \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2$$

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

Idea of Selection Rule

- model dimension k was known fixed in advance
- important: select the right model dimension!
  - incorrect choice can decrease the power of a test
- **Solution**: incorporate the test statistic by some procedure choosing the right dimension automatically by the data

Idea of Selection Rule

- model dimension k was known fixed in advance
- important: select the right model dimension!
  - incorrect choice can decrease the power of a test
- **Solution**: incorporate the test statistic by some procedure choosing the right dimension automatically by the data

Idea of Selection Rule

- model dimension k was known fixed in advance
- important: select the right model dimension!
    - incorrect choice can decrease the power of a test
- **Solution**: incorporate the test statistic by some procedure choosing the right dimension automatically by the data

Bonuses of data-driven score tests

Data-driven score tests are

- asymptotically optimal in an infinite number of directions
- show good overall performance in practice

Outline

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

Penalty

Definition

- nested family of models M_k for $k = 1, \ldots, d(n)$
- $d(n)$ control sequence
- choose $\pi(\cdot, \cdot) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$
- assume
  - $\pi(1, n) < \pi(2, n) < \ldots < \pi(d(n), n)$ for all n
  - $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for $j = 2, \ldots, d(n)$

Call $\pi(j, n)$ a penalty attributed to model M_j and sample size n.

Penalty

Definition

- nested family of models M_k for $k = 1, \ldots, d(n)$
- $d(n)$ control sequence
- choose $\pi(\cdot, \cdot) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$
- assume
  - $\pi(1, n) < \pi(2, n) < \ldots < \pi(d(n), n)$ for all n
  - $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for $j = 2, \ldots, d(n)$

Call $\pi(j, n)$ a penalty attributed to model M_j and sample size n.

Penalty

Definition

- nested family of models M_k for $k = 1, \ldots, d(n)$
- $d(n)$ control sequence
- choose $\pi(\cdot, \cdot) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$
- assume
  - $\pi(1, n) < \pi(2, n) < \ldots < \pi(d(n), n)$ for all n
  - $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for $j = 2, \ldots, d(n)$

Call $\pi(j, n)$ a penalty attributed to model M_j and sample size n.

Penalty

Definition

- nested family of models M_k for $k = 1, \ldots, d(n)$
- $d(n)$ control sequence
- choose $\pi(\cdot, \cdot) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$
- assume
  - $\pi(1, n) < \pi(2, n) < \ldots < \pi(d(n), n)$ for all n
  - $\pi(j, n) - \pi(1, n) \to \infty$ as $n \to \infty$ for $j = 2, \ldots, d(n)$

Call $\pi(j, n)$ a penalty attributed to model M_j and sample size n.

Selection rule

- T_k : testing validity of M_k

Definition

A selection rule S for the sequence of statistics $\{T_k\}$ is

$$S = \min\{k : 1 \leq k \leq d(n);\ T_k - \pi(k,n) \geq T_j - \pi(j,n),\ j = 1, \ldots, d(n)\}$$

Call T_S a data-driven test statistic for testing validity of the initial model.

Outline

Framework

- $X_1, X_2, \ldots$ random variables, values in a measurable space $\mathbb{X}$
- $X_1, \ldots, X_m$ have joint distribution $P_m \in \mathbb{P}_m$ - for every m
- function $\mathcal{F}$ acting from $\otimes_{m=1}^{\infty} \mathbb{P}_m = (\mathbb{P}_1, \mathbb{P}_2, \ldots)$ to a known set Θ
- $\mathcal{F}(P_1, P_2, \ldots) = \theta$

Framework

- $X_1, X_2, \ldots$ random variables, values in a measurable space $\mathbb{X}$
- $X_1, \ldots, X_m$ have joint distribution $P_m \in \mathbb{P}_m$ - for every m
- function $\mathcal{F}$ acting from $\otimes_{m=1}^{\infty} \mathbb{P}_m = (\mathbb{P}_1, \mathbb{P}_2, \ldots)$ to a known set Θ
- $\mathcal{F}(P_1, P_2, \ldots) = \theta$

Framework

- $$H_0 : \quad \theta \in \Theta_0 \subset \Theta$$
- $$H_A : \quad \theta \in \Theta_1 = \Theta \setminus \Theta_0$$
- observations $Y_1, \ldots, Y_n$, values in a measurable space $\mathbb{Y}$
- not necessarily on the basis of $X_1, \ldots, X_m$!

Framework

- $$H_0: \quad \theta \in \Theta_0 \subset \Theta$$
- $$H_A: \quad \theta \in \Theta_1 = \Theta \setminus \Theta_0$$
- observations $Y_1, \ldots, Y_n$, values in a measurable space $\mathbb{Y}$
- not necessarily on the basis of $X_1, \ldots, X_m$!

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

Main concept

Definition

- observations $Y_1, \ldots, Y_n$, values in a measurable space $\mathbb{Y}$
- k fixed number
- $l = (l_1, \ldots, l_k)$ vector-function
- $l_i : \mathbb{Y} \to \mathbb{R}$ for $i = 1, \ldots, k$ are known Lebesgue measurable

Main concept

Definition

- $$L = \{E_0[l(Y)]^T l(Y)\}^{-1}$$
  - E_0 is with respect to P_0
  - P_0 is the d.f. of some (fixed and known) random variable Y
  - Y has values in $\mathbb{Y}$
- assume
  - $E_0\, l(Y) = 0$
  - L is well defined

Main concept

Definition

- $$L = \{E_0[l(Y)]^T l(Y)\}^{-1}$$
  - E_0 is with respect to P_0
  - P_0 is the d.f. of some (fixed and known) random variable Y
  - Y has values in $\mathbb{Y}$
- assume
  - $E_0\, l(Y) = 0$
  - L is well defined

Main concept

Definition

$$T_k = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\} L \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l(Y_j) \right\}^T$$

T_k - statistic of Neyman's type (NT-statistic)

Outline

1. Introduction
   - Main ideas
   - Data-driven tests
2. General notions
   - Selection rule
   - Framework
3. NT-statistics
   - Definitions
   - Examples
   - Theorems

New possibilities

$l_1, \ldots, l_k$ can be

- some score functions
- any other functions, depending on the problem
  - truncated, penalized or partial likelihood
  - possible to use $l_1, \ldots, l_k$ unrelated to any likelihood

New possibilities

$l_1, \ldots, l_k$ can be

- some score functions
- any other functions, depending on the problem
  - truncated, penalized or partial likelihood
  - possible to use $l_1, \ldots, l_k$ unrelated to any likelihood

New possibilities

$l_1, \ldots, l_k$ can be

- some score functions
- any other functions, depending on the problem
  - truncated, penalized or partial likelihood
  - possible to use $l_1, \ldots, l_k$ unrelated to any likelihood

Important applications

- statistical inverse problems
- rank tests for independence
- semiparametric regression

Important applications

- statistical inverse problems
- rank tests for independence
- semiparametric regression

Important applications

- statistical inverse problems
- rank tests for independence
- semiparametric regression

Statistical inverse problems

Deconvolution

- applications from signal processing to psychology
- basic statistical inverse problem

Deconvolution

- instead of X_i one observes Y_i

$$Y_i = X_i + \varepsilon_i$$

- ε_i's are i.i.d. with a known density h
- X_i and ε_i are independent for each i
- H_0 : $\quad X$ has density f_0

Deconvolution

- instead of X_i one observes Y_i

$$Y_i = X_i + \varepsilon_i$$

- ε_i's are i.i.d. with a known density h
- X_i and ε_i are independent for each i
- $H_0 : \quad X$ has density f_0

Deconvolution

- choose for every $k \leq d(n)$ an auxiliary parametric family $\{f_\theta\}$
- $\theta \in \Theta \subseteq \mathbb{R}^k$
- f_0 from this family coincides with f_0 from the null hypothesis H_0
- the true F possibly has no relation to the chosen $\{f_\theta\}$

Deconvolution

$$l(y) = \frac{\frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f_\theta(s)\, h(\,y - s)\, ds\right)\Big|_{\theta=0}}{\int_{\mathbb{R}} f_0(s)\, h(\,y - s)\, ds}$$

- define T_k as above $\Rightarrow$ T_k is an NT-statistic

Outline

Null distribution

Theorem

- $\{T_k\}$ *sequence of NT-statistics*
- S *selection rule, with penalty of proper weight*
- *large deviations of* T_k *are properly majorated*
- $$d(n) \le \min\{u_n \, , m_n\} \, .$$

Then $S = O_{P_0}(1)$ *and* $T_S = O_{P_0}(1)$.

Consistency condition

$\langle \mathbf{C} \rangle$ there exists integer $K = K(P) \geq 1$ such that

$$E_P \, l_1(Y) = 0, \; \ldots, E_P \, l_{K-1}(Y) = 0, \;\; E_P \, l_K = C_P \neq 0$$

Consistency theorem

Theorem

- $\{T_k\}$ *sequence of NT-statistics*
- *S selection rule, with penalty of proper weight*
- *the regularity assumptions are satisfied*
- $d(n) = o(r_n)$, $d(n) \leq \min\{u_n, m_n\}$.

Then T_S *is consistent against any (fixed) alternative P satisfying (C).*