Statistical Analysis of Big Data

Ruibin Xi

Peking University

December 12, 2015

Big Data

Big data are large and complex data sets that traditional data processing applications are inadequate. Three V's

- Volume
- Velocity
- Variety

Big Data

Big data are large and complex data sets that traditional data processing applications are inadequate. Three V's

- Volume
- Velocity
- Variety
- Veracity

Big Data

Big data is not new

- large data
- very large data
- massive data
- big data

Big Data

Big data sets bring many challenges

- computational time,
- memory,
- storage
- online analytical processing (OLAP).

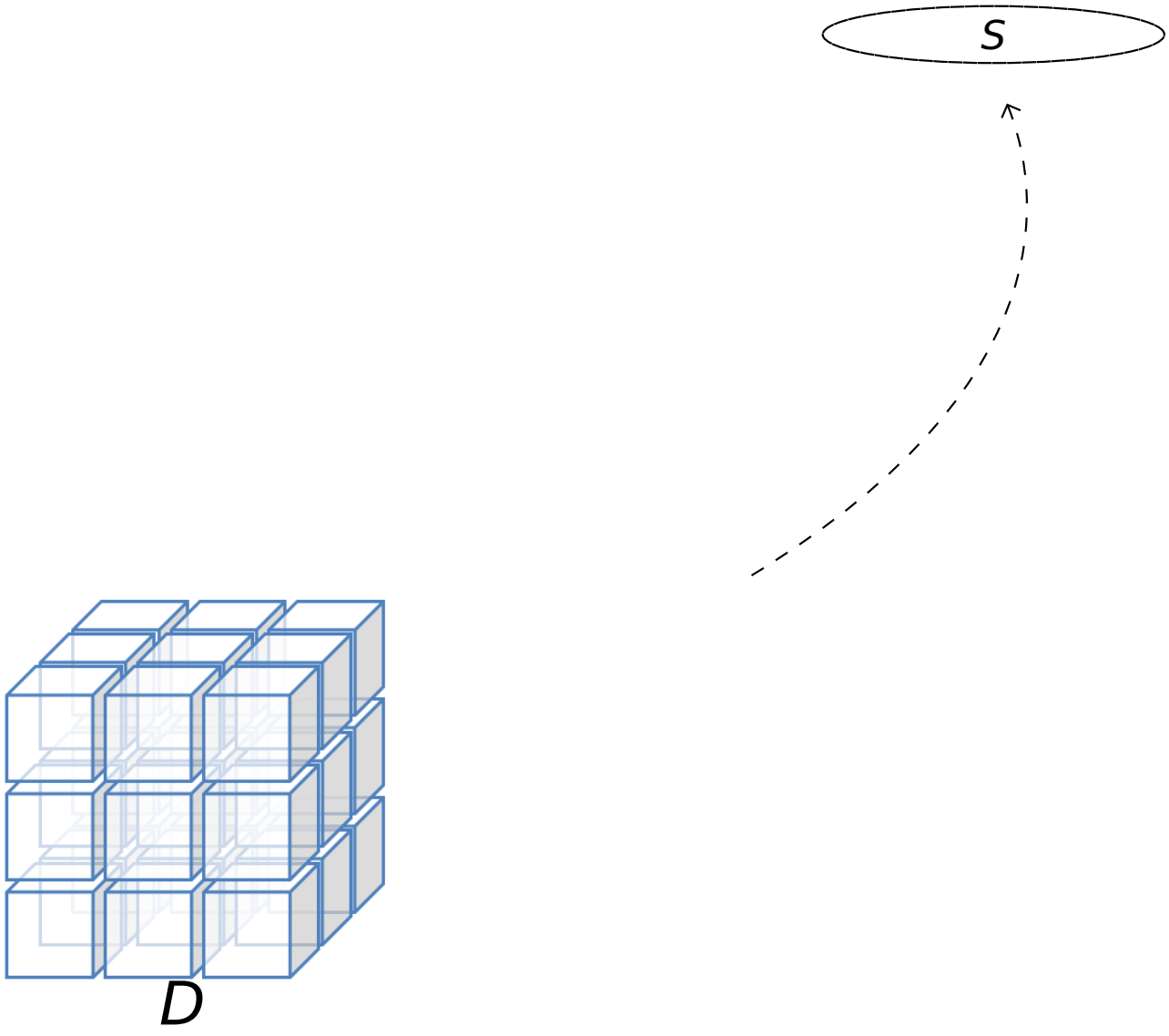
Big Data

Available strategies for analyzing big data

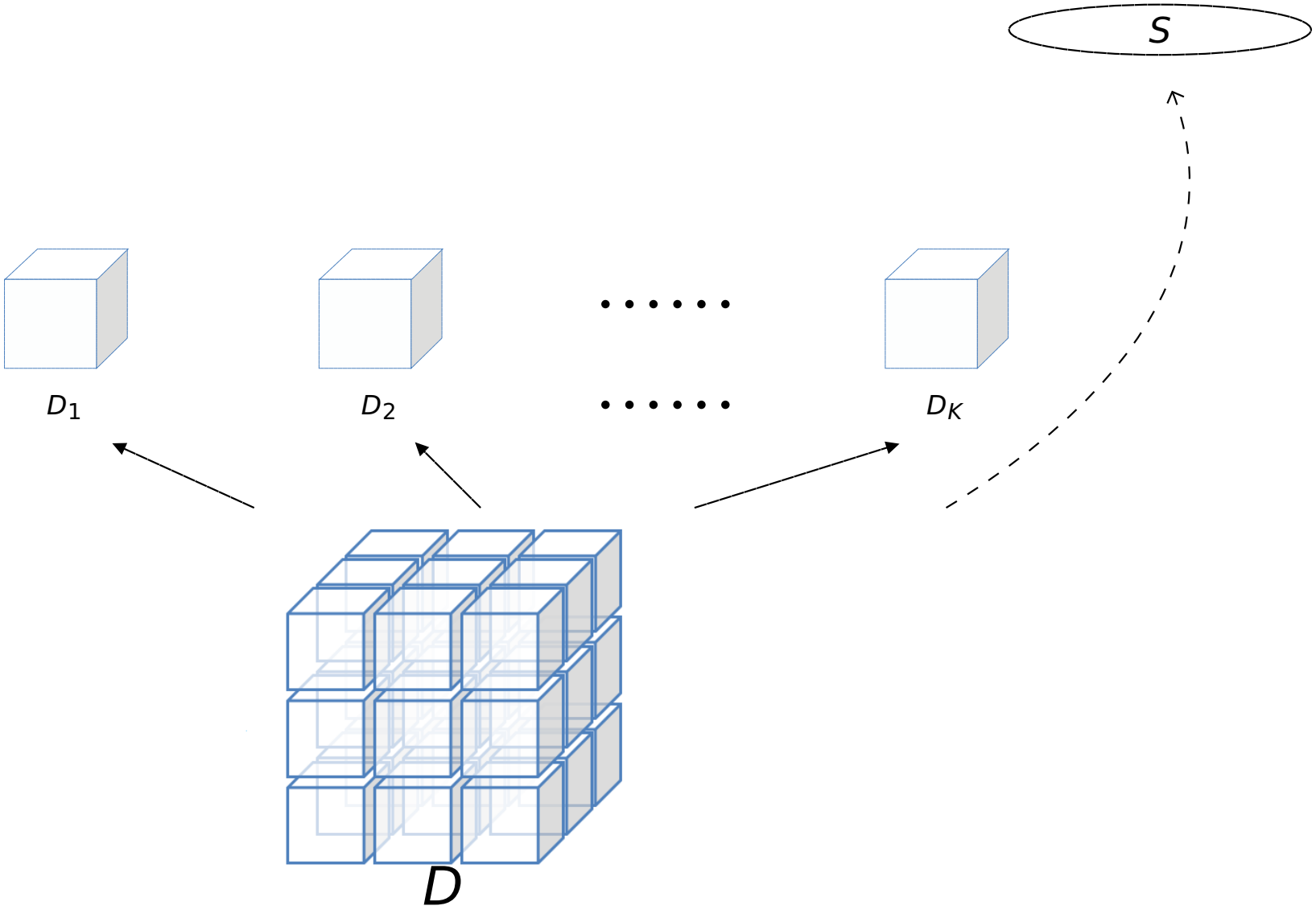
- subsampling methods (Kleiner et al., 2014; Ma et al., 2013; Liang et al., 2013; Maclaurin and Adams, 2014)
- divide and conquer strategy (Lin and Xi, 2011; Chen and Xie, 2014; Song and Liang, 2014; Neiswanger et al.,2013)

1. Statistical Aggregation

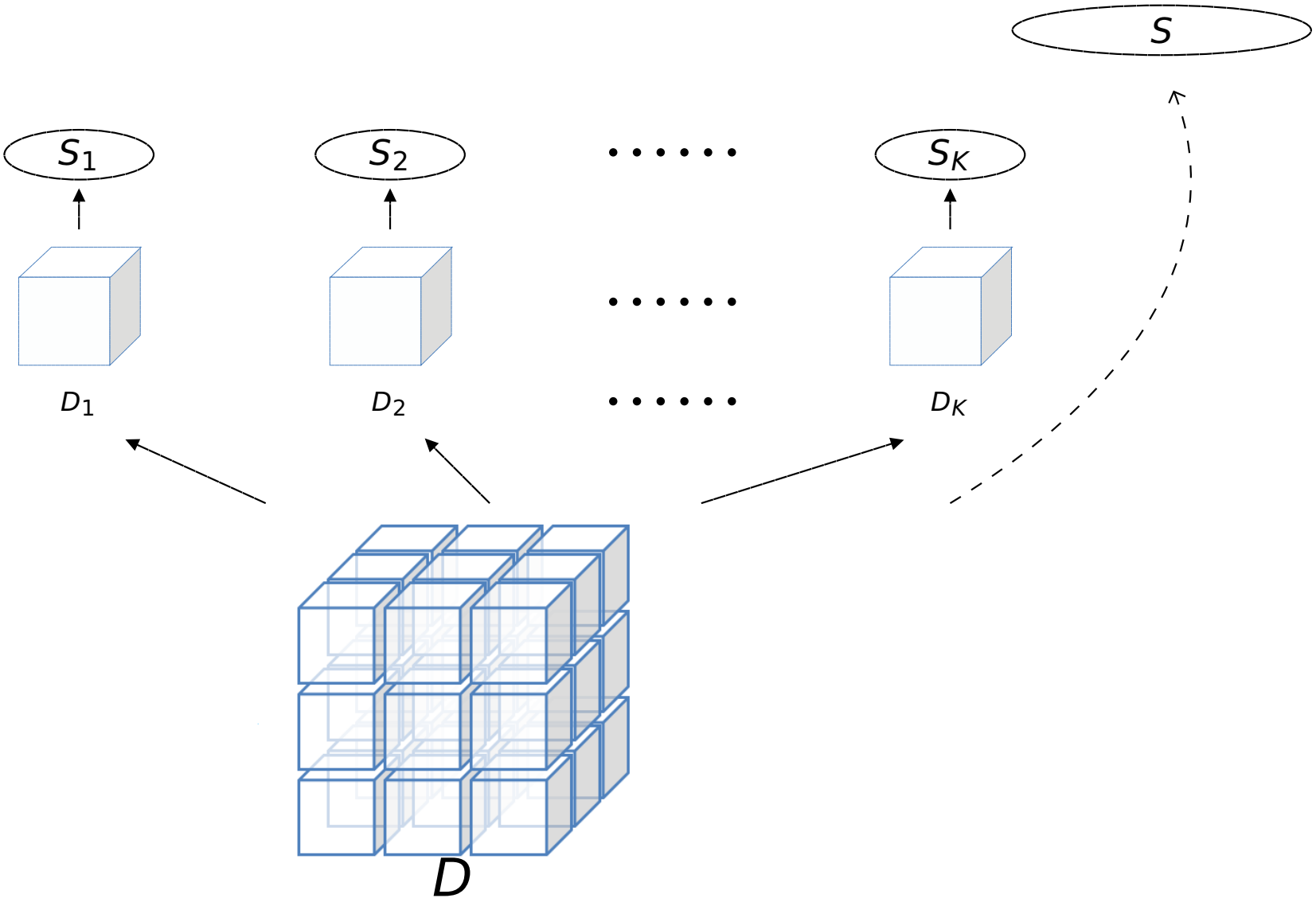
Statistical Aggregation



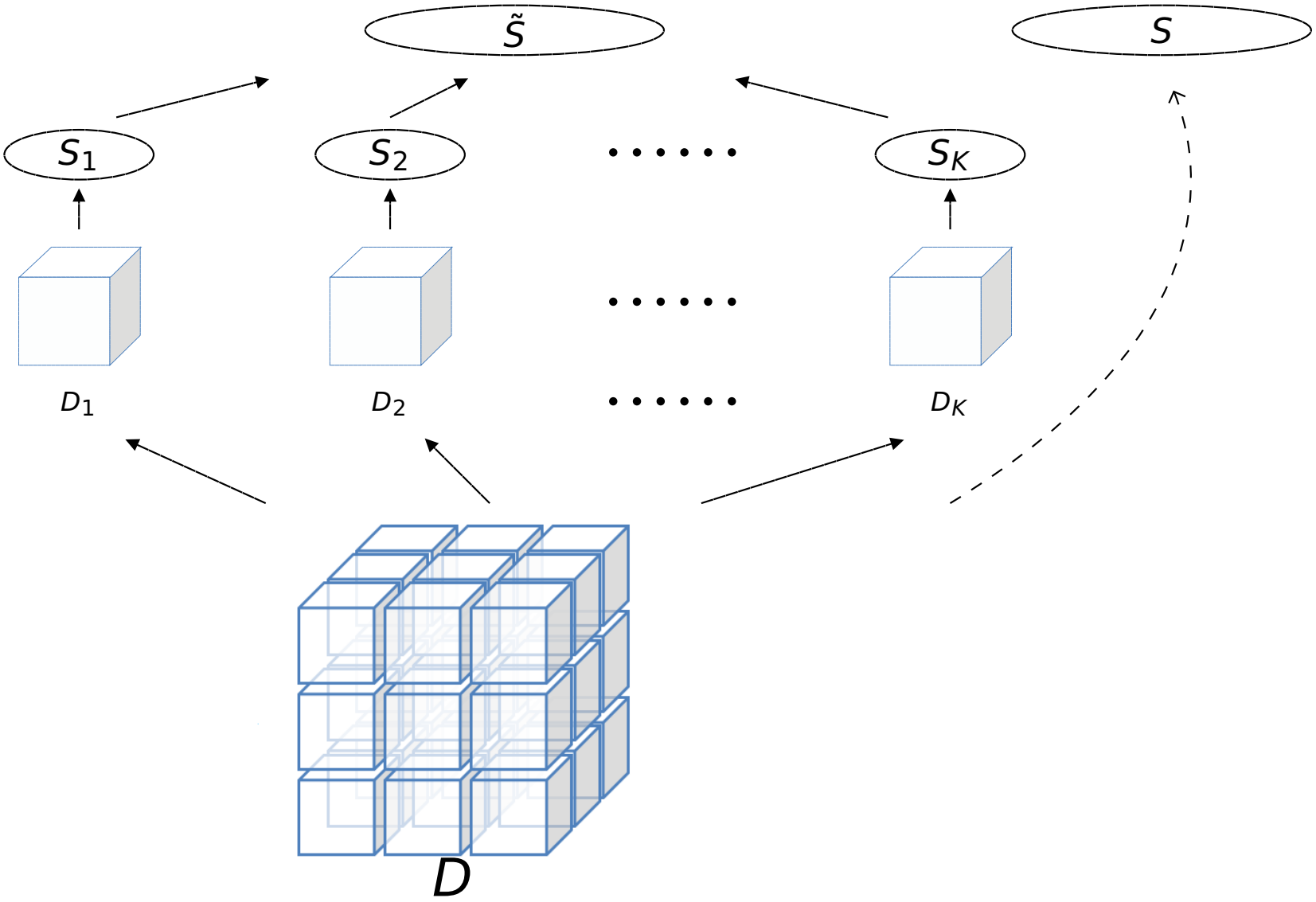
Statistical Aggregation



Statistical Aggregation



Statistical Aggregation

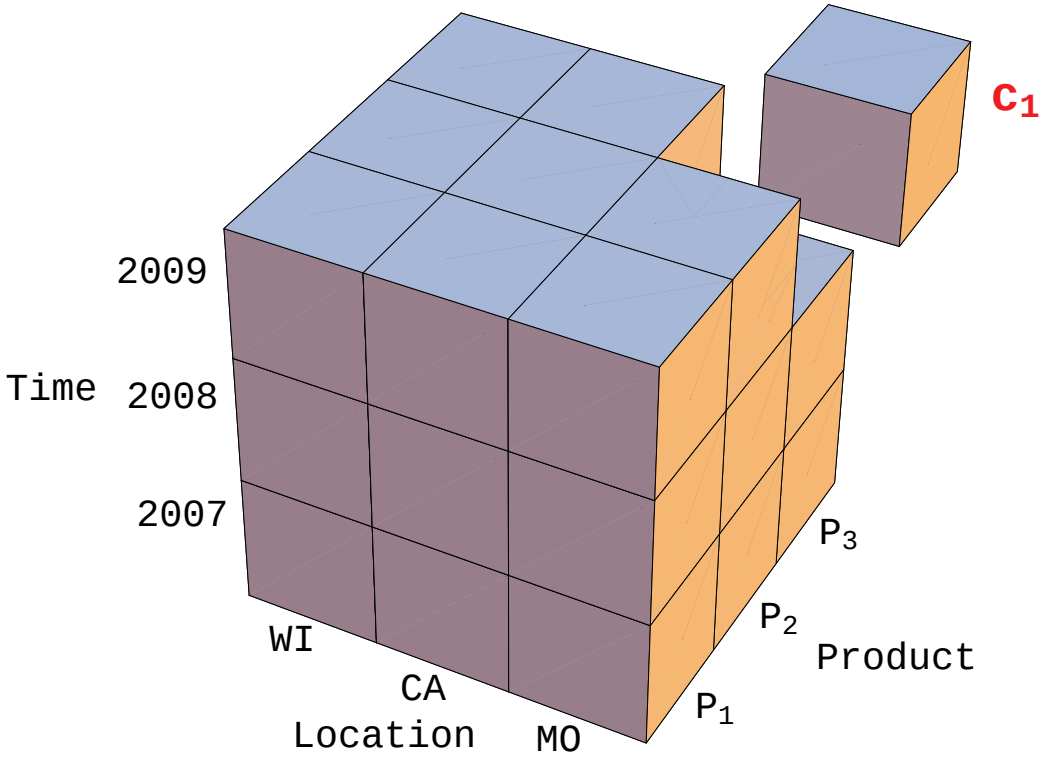


Statistical Aggregation

1. Partition: how to choose the number of subset K.
2. Compression: the dimension of S_k should be independent of the sample size of D_k
3. Aggregation:
   - easy to compute.
   - $\tilde{S}$ and S are asymptotically equivalent.

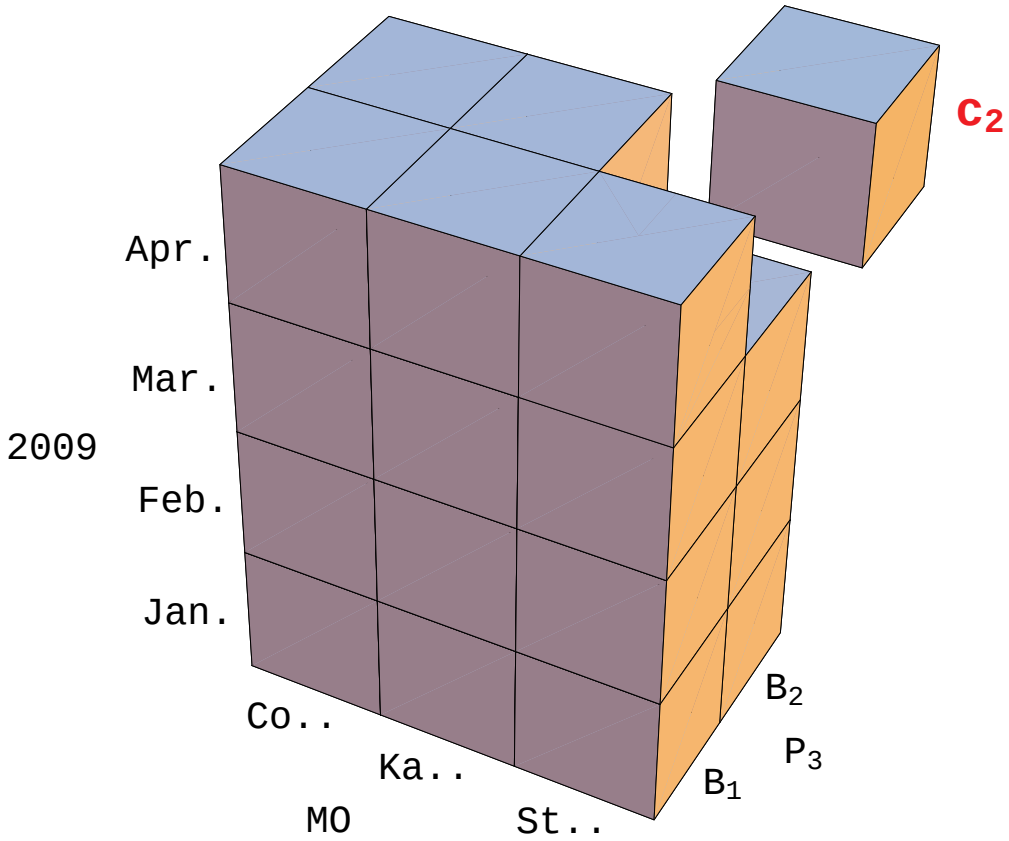
Data Cubes

- model the big data as a multidimensional hyper-rectangle
- support OLAP computing in data warehouse
- cells: determined by values of attributes such as *location* and *time*
- base cells: cells without sub-cells



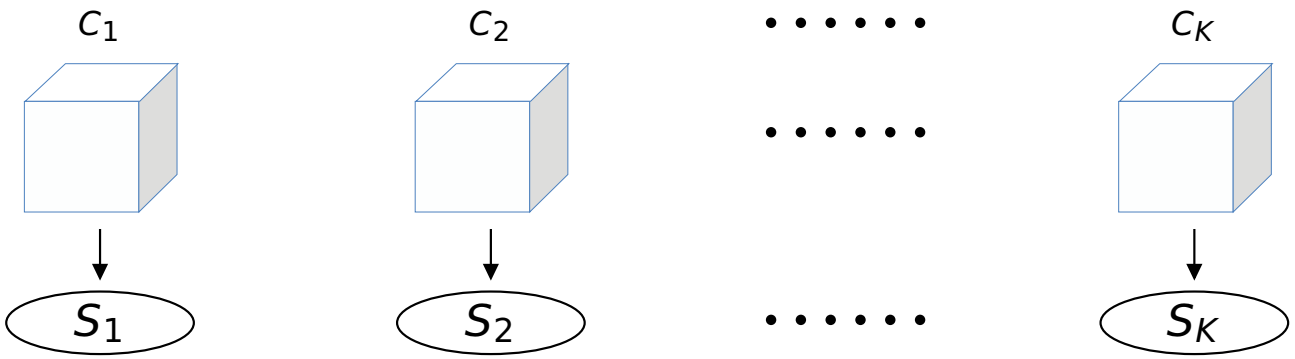
Data Cubes

- model the big data as a multidimensional hyper-rectangle
- support OLAP computing in data warehouse
- cells: determined by values of attributes such as *location* and *time*
- base cells: cells without sub-cells



Data Cubes

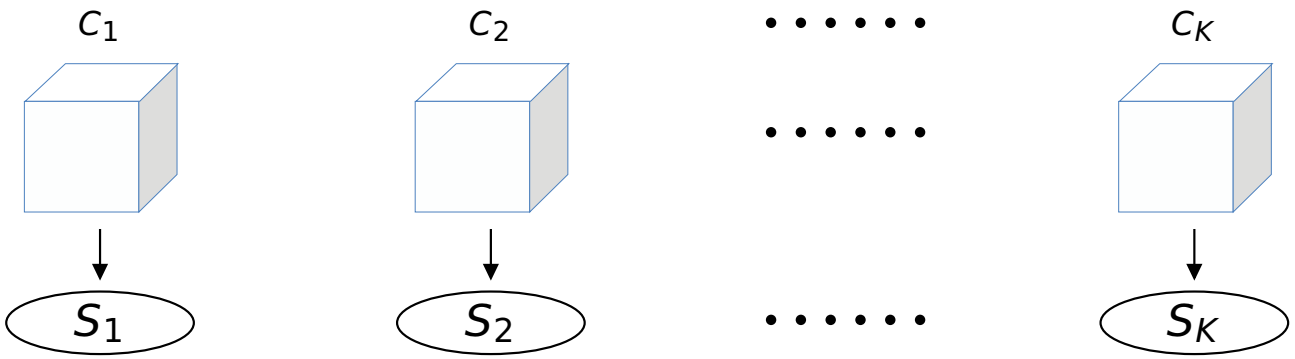
1. Compression



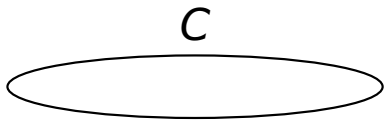
2. Aggregation

Data Cubes

1. Compression

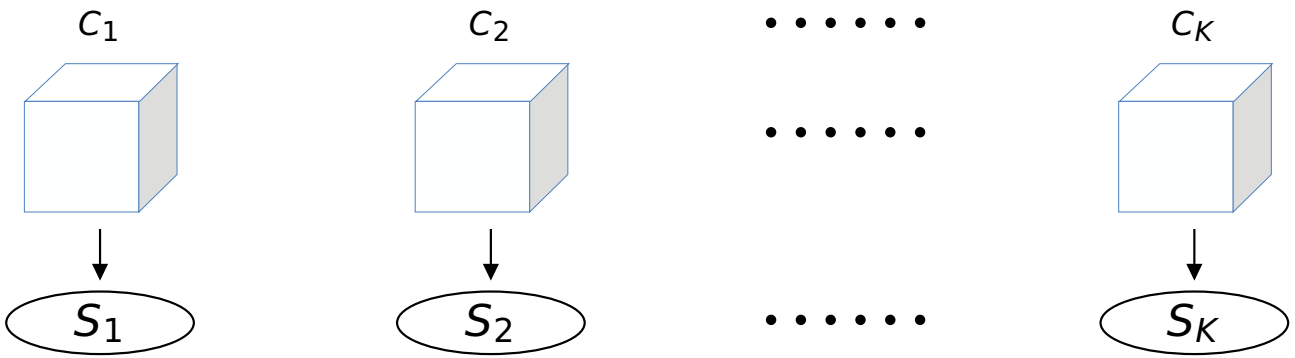


2. Aggregation



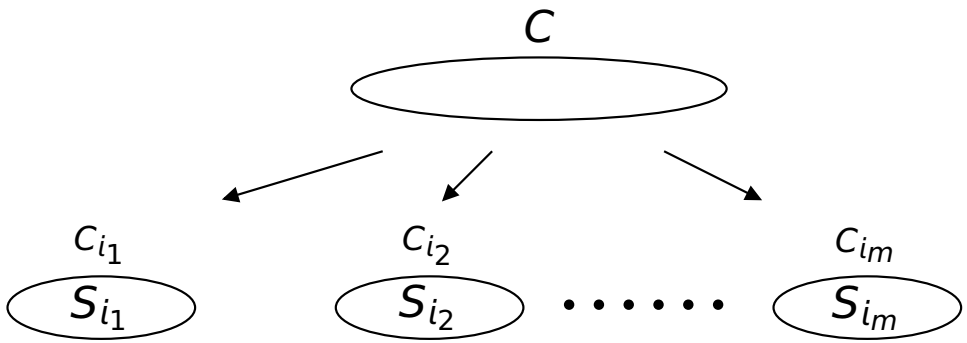
Data Cubes

1. Compression



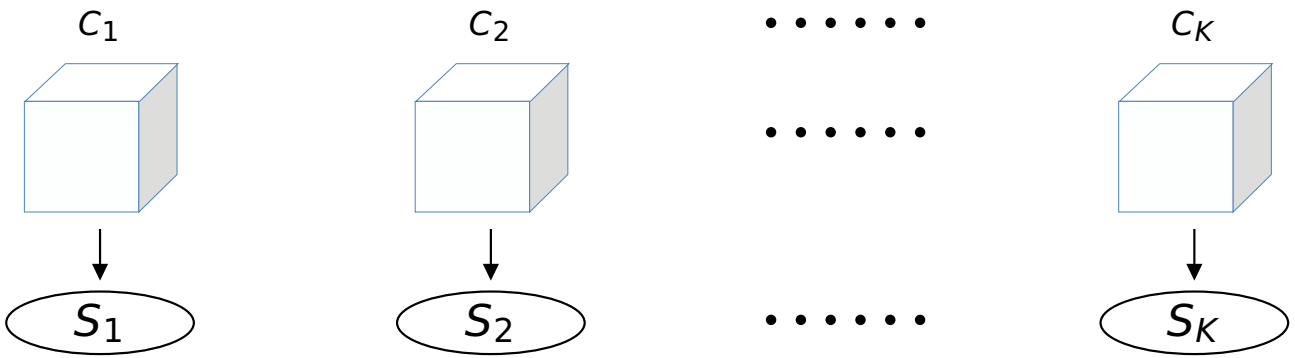
2. Aggregation

2.1 Drill down: ↓



Data Cubes

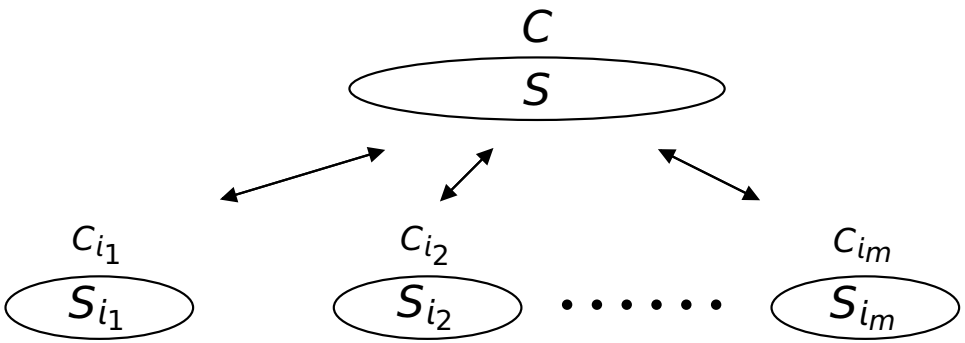
1. Compression



2. Aggregation

2.1 Drill down: ↓

2.2 Roll up: ↑



Data Cubes

Earlier work in data cubes supports the following statistical analysis.

- Simple statistical analysis like *minimum*, *maximum*, *count*, *sum* and *average* (Gray et al., 1997; Agarwal et al., 1996; Zhao et al., 1997).
- The ordinary least quare (OLS) estimation in linear regression (Chen et al. 2006): *regression cube*.

Regression Cube

Recall that the linear regression model is given by

$$E(y) = \mathbf{x}\boldsymbol{\beta}.$$

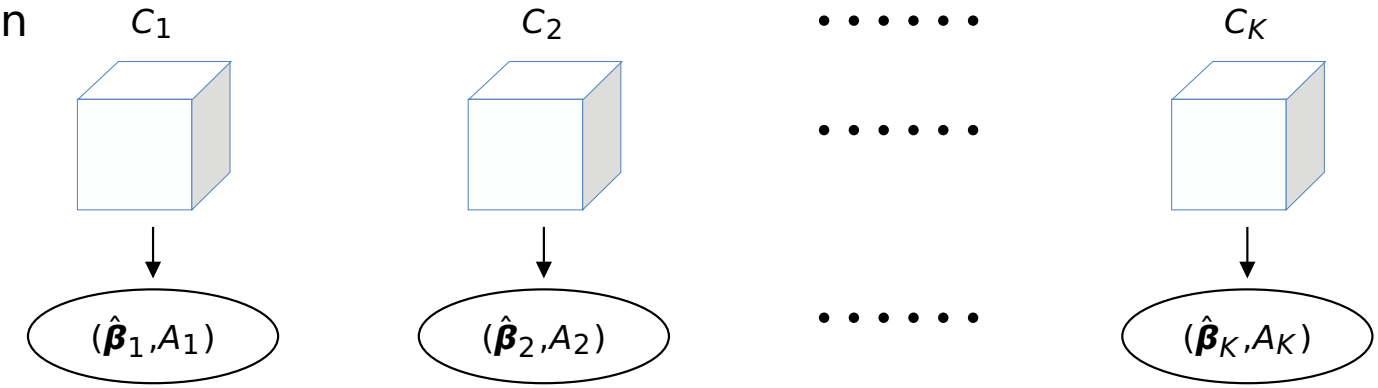
- $y \in \mathbb{R}$: the response variable.
- $\mathbf{x} = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$: predictors.
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$: unknown regression parameters.
- $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$: observations.
- $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $A = X^T X$.

The OLS estimator is given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = A^{-1} X^T \mathbf{y}.$$

Regression Cube

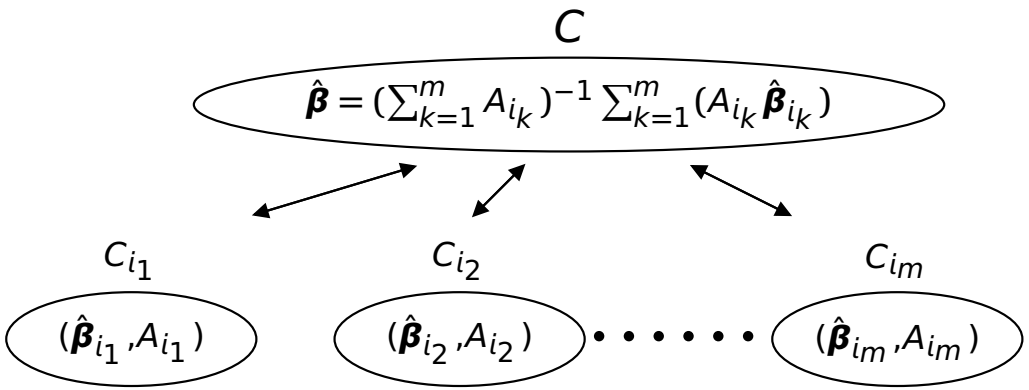
1. Compression



2. Aggregation

   2.1 Drill down: ↓

   2.2 Roll up: ↑



2. Parametric Statistical Aggregation

2.1 Aggregation of the Estimating Equation Estimation

Estimating Equation (EE) Estimation

If $\mathbf{z}_i$ $(i = 1, \cdots, N)$ are independent random variables, the EE estimator is the solution to

$$\sum_{i=1}^{N} \boldsymbol{\psi}(\mathbf{z}_i, \boldsymbol{\beta}) = 0$$

for some score function $\boldsymbol{\psi}$.

EE Estimation

Many estimators in regression are EE estimators. In regression, we have $\mathbf{z}_i = (y_i, \mathbf{x}_i^T)$.

- the OLS estimator in linear regression,
- the MLE in logistic regression,

$$\sum_{i=1}^{N} \mathbf{x}_i [y_i - p(\mathbf{x}_i^T \boldsymbol{\beta})] = 0.$$

- the quasi-likelihood estimator (QLE) in Chen et al. (1999),

$$\sum_{i=1}^{N} \mathbf{x}_i [y_i - \mu(\mathbf{x}_i^T \boldsymbol{\beta})] = 0$$

  where $\mu(t)$ is some known function.

Aggregation of the EE Estimation

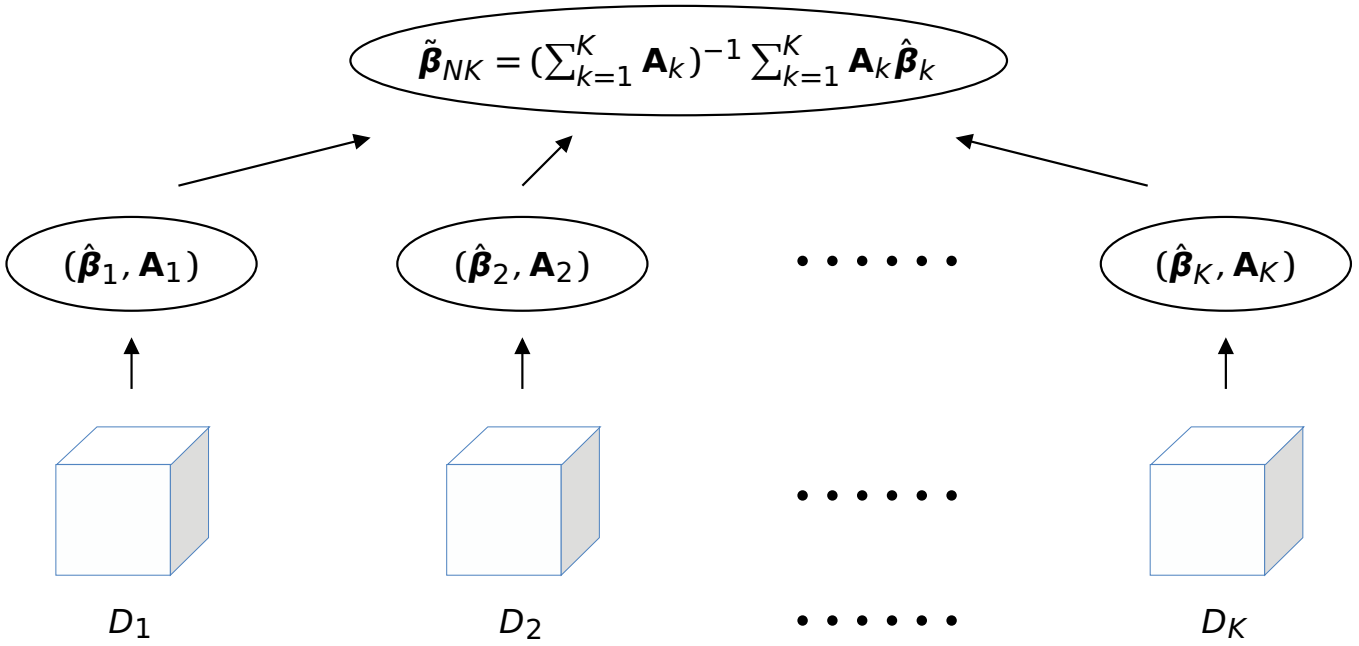
- Partition the entire data set into K subsets.
- $\mathbf{z}_{k1}, \cdots, \mathbf{z}_{kn}$: the observations in the kth subset.
- The EE estimate $\hat{\boldsymbol{\beta}}_k$ for the kth subset is the solution

$$\mathbf{M}_k(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{z}_{ki}, \boldsymbol{\beta}) = 0.$$

- Let $\mathbf{A}_k = -\frac{\partial \mathbf{M}_k}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_k)$.
- Define the aggregated EE estimator (AEE) as

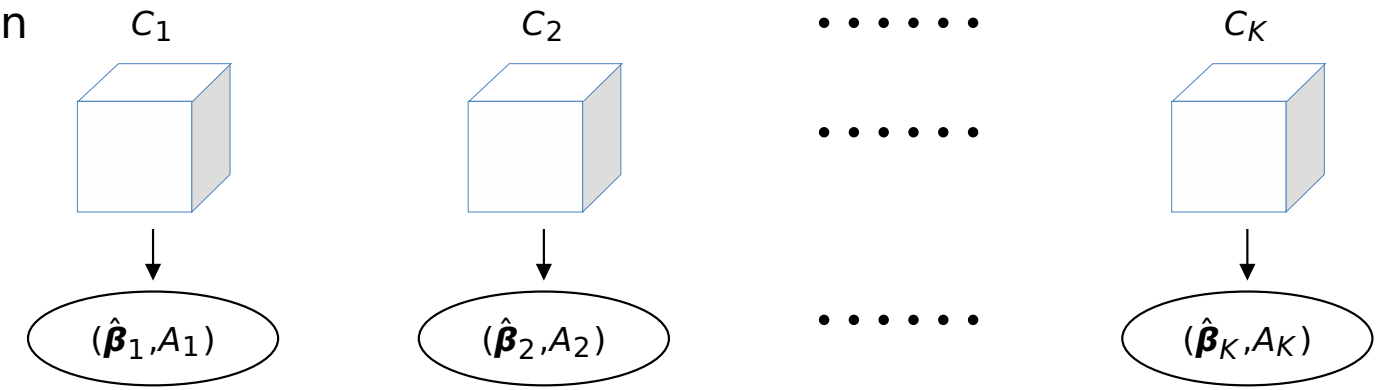
$$\tilde{\boldsymbol{\beta}}_{NK} = \left(\sum_{k=1}^{K} \mathbf{A}_k \right)^{-1} \sum_{k=1}^{K} \mathbf{A}_k \hat{\boldsymbol{\beta}}_k.$$

Aggregation of the EE Estimation



The EE Estimation in Data Cubes

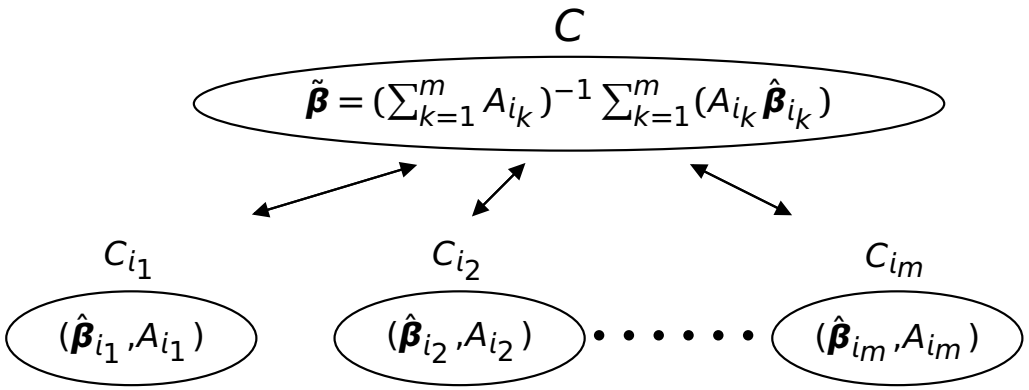
1. Compression



2. Aggregation

   2.1 Drill down: ↓

   2.2 Roll up: ↑



Properties of the AEE estimator

- If K goes to infinity slowly, then under some regularity conditions, we have for any $\delta > 0$

$$P(\sqrt{N}\|\tilde{\boldsymbol{\beta}}_{NK} - \hat{\boldsymbol{\beta}}_N\| > \delta) = o(1).$$

- The critical regularity condition is about the convergence rate of the EE estimator.

Asymptotic Property of the Aggregated QLE (AQLE)

Recall that the QLE $\hat{\boldsymbol{\beta}}_N$ is the solution to the equation

$$\sum_{i=1}^{N}[y_i - \mu(\mathbf{x}_i^T\boldsymbol{\beta})]\mathbf{x}_i = 0, \tag{1}$$

where y_i and $\mathbf{x}_i$ satisfies $E(y_i) = \mu(\mathbf{x}_i^T\boldsymbol{\beta}_0)$ for some known function $\mu(t)$.

- If $K = o(N^{1/4})$, the AQLE $\tilde{\boldsymbol{\beta}}_{NK}$ and the QLE $\hat{\boldsymbol{\beta}}_N$ are asymptotically equivalent.

2.2 Simulation Studies

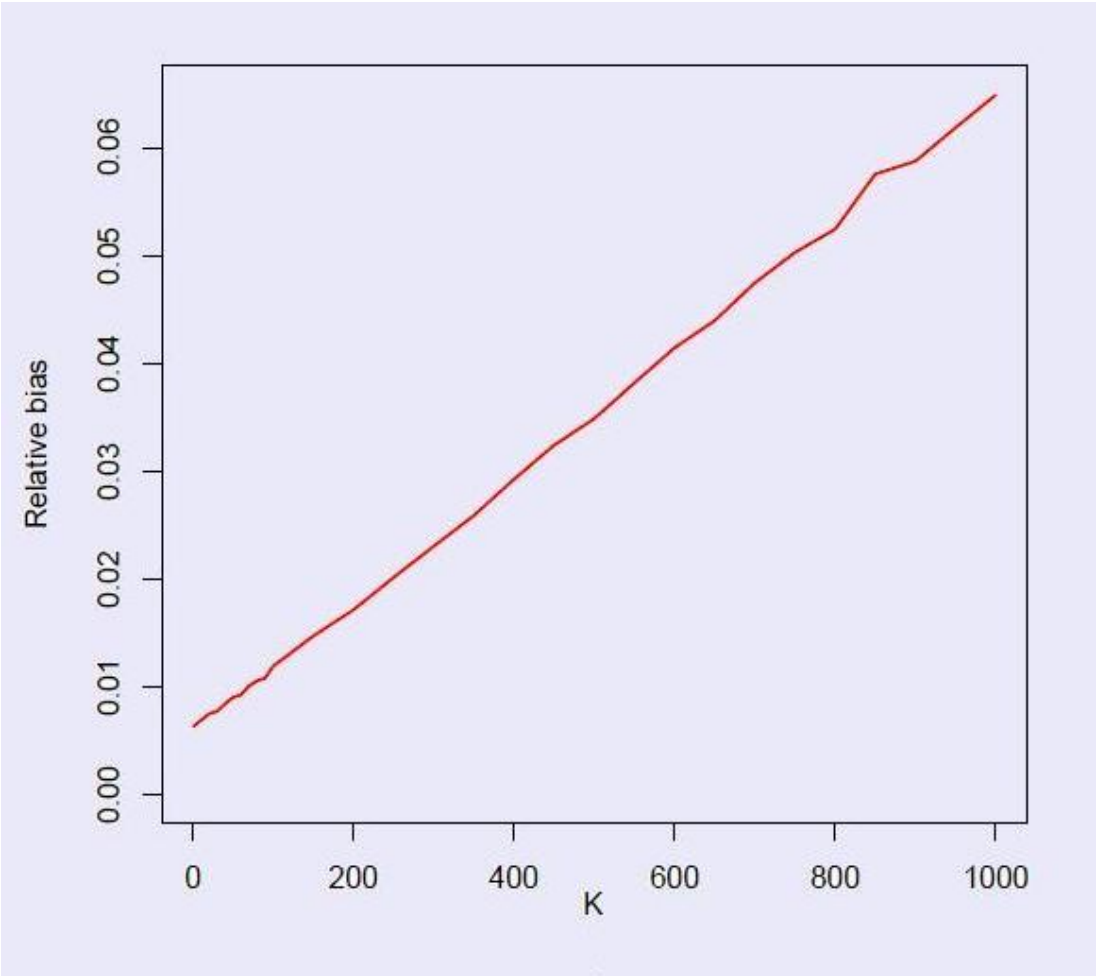
Simulation: Logistic Regression

We consider the MLE in logistic regression with five predictors $x_1, \cdots, x_5$.

- y_i $(i = 1, \cdots, N)$: the binary response;
- $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{i5})^T$ $(i = 1, \cdots, N)$: explanatory variable;
- In a logistic regression model,
  $$Pr(y_i = 1) = p(\mathbf{x}_i^T \boldsymbol{\beta}_0) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_0)}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad i = 1, \cdots, N;$$
- $N = 500,000$ and
  $\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \cdots, \beta_5) = (1, 2, 3, 4, 5, 6)$;
- The program is written in C.
- Computer: 3.4 GHz Pentium processor and 1 GB memory.

Simulation: Logistic Regression

Figure 1: Relative bias $\|\tilde{\boldsymbol{\beta}}_{NK} - \boldsymbol{\beta}_0\|/\|\boldsymbol{\beta}_0\|$ against K



Simulation: Logistic Regression

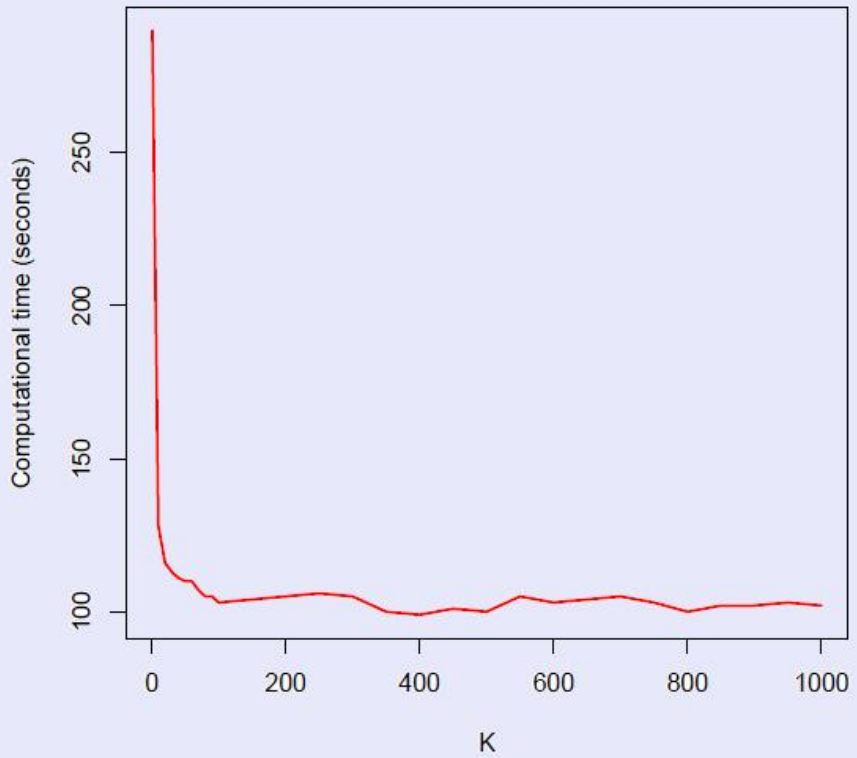


Figure 2: Computation time against number of partition K

2.3 Applications to Data Cubes

Logistic Regression in Data Cubes

We consider maximum likelihood estimation of logistic regression in data cube context.

- Use the same data set as in the first simulation study.
- Assume two dimensional attributes: location and time.
- $1000 = 50 \times 20$ base cells in total.
- Randomly generate 100 queries. For each query,

  1. generate D from $\{1, \cdots, 1000\}$
  2. randomly select D out of $1,000$ base cells

Logistic Regression in Data Cubes

Table: Comparison of computational time.

	AEE estimate	EE estimate
Compression	97 seconds	NA
Aggregation	0.0 second	6771 seconds

Table: Comparison of storage requirement.

	AEE estimate	EE estimate
Space	$42,000$	$3,500,000$

Homogeneity

Table: Success rates for different groups of stone size.

	Treatment A	Treatment B
Small Stone	93%(81/87)	87%(234/270)
Large Stone	73%(192/263)	69%(55/80)
Both	78%(273/350)	83%(289/350)

A homogeneity chi-square test can be developed based on $(\hat{\boldsymbol{\beta}}_{i_k}, A_{i_k})$. The test statistics is $\chi = 26.04$ which is highly significant.

3. Nonparametric Statistical Aggregation

3.1 Aggregation of U-statistics

U-statistics

- Kendall's τ rank correlation.
- Spearman's ρ.
- The Wilcoxon test statistic.
- Symmetry test statistic (Randles et al., 1980).

U-statistics

- $X_1, \cdots, X_N$ i.i.d random variables.
- $h(x_1, \cdots, x_m)$: a function defined on $\mathbb{R}^m$ that is symmetric in its arguments.
- A U-statistic with kernel h and degree m is defined as

$$U_N = \binom{N}{m}^{-1} \sum_{1 \le i_1 < \cdots < i_m \le N} h(X_{i_1}, \cdots, X_{i_m}). \quad (2)$$

The time complexity of computing the U-statistics is high when $m \geq 2$.

- The time complexity is $O(N^m)$.
- If $N = 1,000$, it takes about 1 hour in R to calculate the symmetry test statistic which is a U-statistic of degree 3.

Examples of U-statistics

- Kendall's τ rank correlation:
  $h(\mathbf{z}_1, \mathbf{z}_2) = 2I((x_1 - x_2)(y_1 - y_2) > 0) - 1$ for
  $\mathbf{z}_i = (x_i, y_i)^T \in \mathbb{R}^2$ $(i = 1, 2)$.
- Symmetry test statistic:

$$h(x, y, z) = \frac{1}{3}[\text{sign}(x+y-2z)+\text{sign}(x+z-2y)+\text{sign}(y+z-2x)],$$

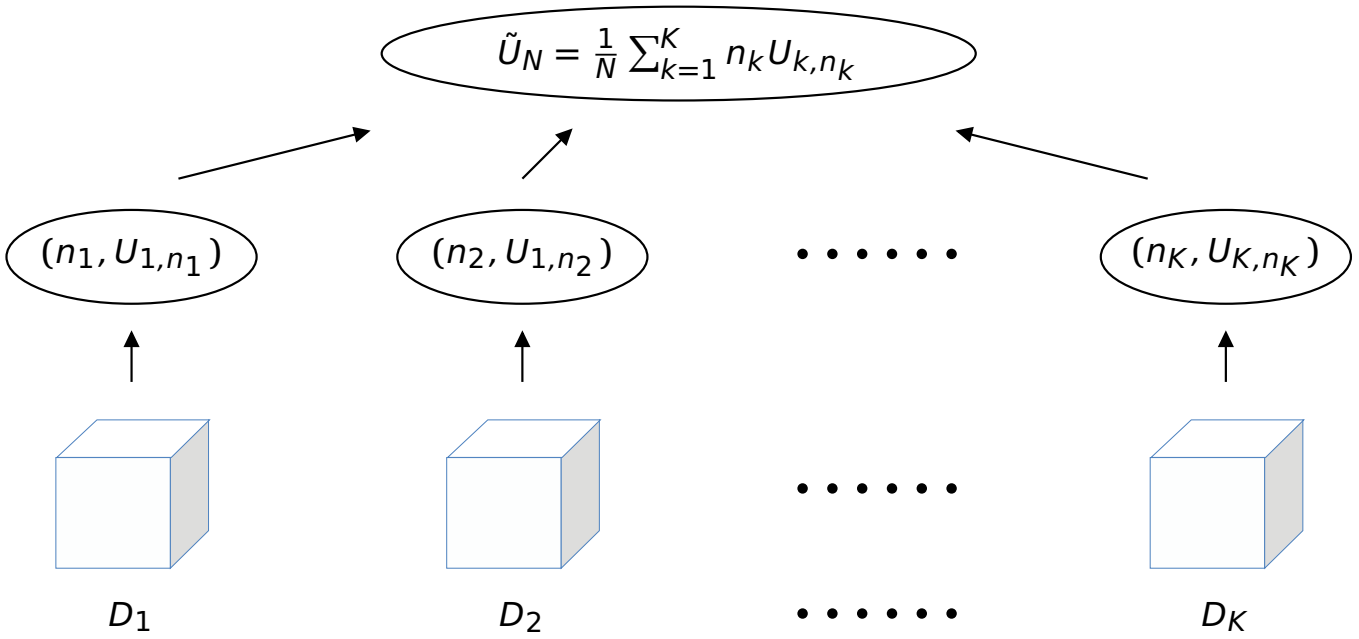
  where $\text{sign}(u) = -1,\ 0$, or 1 as $u <, =,$ or > 0.
- Sample variance: $h(x_1, x_2) = (x_1 - x_2)^2/2$.

Some Properties of U-statistics

The fundamental theory of U-statistics was first developed by Hoeffding (1948).

- Nonparametric estimator.
- Unbiased estimator of $\theta = E[h(X_1, \cdots, X_m)]$.
- Asymptotic normality:
  $\sqrt{N}[U_N - \theta] \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1) \quad \text{as} \quad N \to \infty.$

Aggregation of U-statistics



Properties of the AU-statistic

1. Asymptotic Normality. If $K = o(N)$, $\tilde{U}_N$ and U_N are asymptotically equivalent.

2. The time complexity is much lower. If $m = 3$ and $K = \sqrt{N}$ and each subset has the same number of observations, then the time complexity of the AU-statistics is $O(N^2)$.

3.2 Simulation Studies

Symmetry Test Statistics

Randles et al. (1980) proposed a nonparametric method to test the symmetry of data distribution. Recall that the test statistic is a U-statistic of degree $m = 3$ with kernel function

$$h(x, y, z) = \frac{1}{3}[\text{sign}(x+y-2z)+\text{sign}(x+z-2y)+\text{sign}(y+z-2x)].$$

Simulation Setup

- 200 data sets were generated from $\mathcal{N}(0, 1)$.
  1. Each data set has 2,000 observations.
  2. Because the distribution is symmetric, we have $\theta = 0$.
- We used $K = 20, 60$, and 100 for AU-statistics.
- The program is written in C.

Performance of AU-statistics

Figure 3: Box plots of U-statistics and AU-statistics

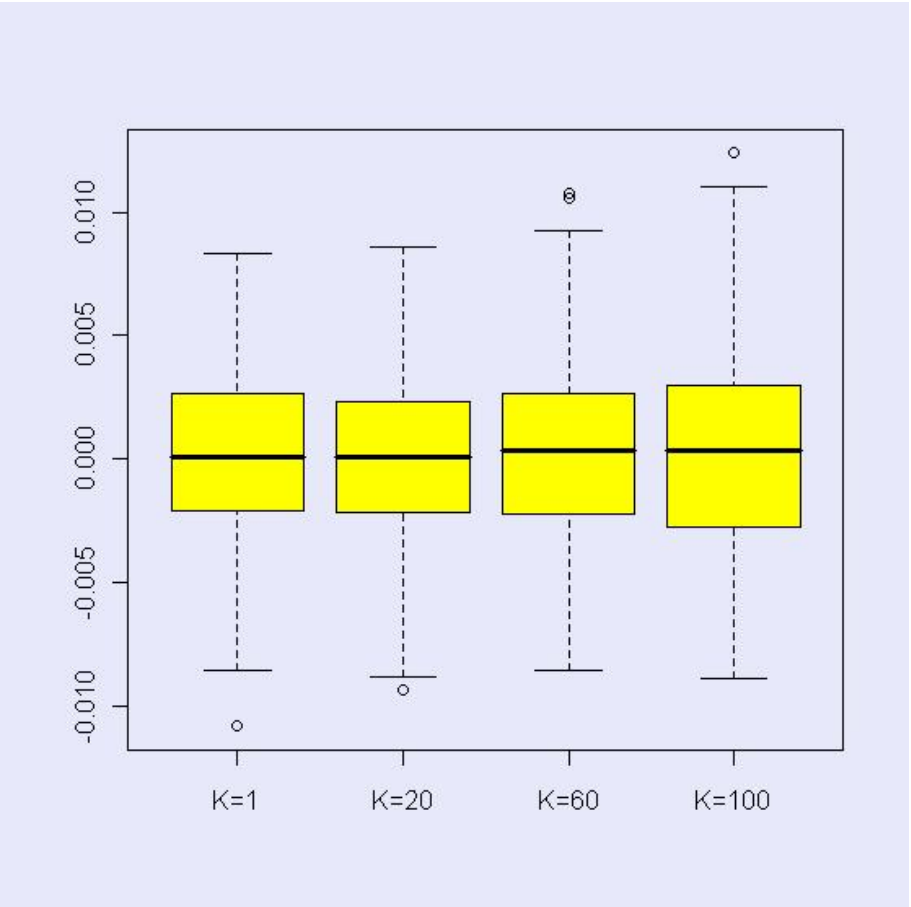


Table: Comparison of U-statistics and AU-statistics on computing the symmetry test statistic.

	K	Bias($\times 10^{-4}$)	Variance($\times 10^{-5}$)	Time (seconds)
U	1	-3.3	1.31	148
AU	20	-3.2	1.39	0.38
	60	-1.7	1.66	0.04
	100	-1.6	1.91	0.01

4. Applications to U-statistic Based Estimating Equation

U-statistic Based Estimating Equation

- $\mathbf{y}_{it} = (y_{1it}, y_{2it})$ $(t = 1, 2, 3)$: two different measurement on the ith subject at three time points $t = 1, 2, 3$.
- $E[(y_{1it} - y_{1jt})^2/2] = \sigma_{1t}^2$.
- $E[(y_{2it} - y_{2jt})^2/2] = \sigma_{2t}^2$.
- $E[(y_{1it} - y_{1jt})(y_{2it} - y_{2jt})/2] = \rho_t \sigma_{1t} \sigma_{2t}$.
- Interested in the correlation ρ_t.

U-statistic Based Estimating Equation

Let $\mathbf{y} = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{y}_{i3})$. The estimating equation is,

$$U_N(\boldsymbol{\theta}) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} H(\boldsymbol{\theta}, \mathbf{y}_i, \mathbf{y}_j) = 0,$$

where $\boldsymbol{\theta}$ is a parameter vector with ρ_t, σ_{1t}^2 and σ_{2t}^2 as its element and $H(\boldsymbol{\theta}, \cdot, \cdot)$ is a kernel function depending on $\boldsymbol{\theta}$.
This model was introduced by Kowalski and Tu (2007). The estimating equation is called U-statistic based generalized estimating equation (UGEE).

AU-statistic Based Estimating Equation

- Partition the sample set into K subsets.
- Denote $\{\mathbf{y}_{k,1}, \cdots, \mathbf{y}_{k,n_k}\}$ as all observations in the kth subset.
- For each data set, define

$$U_{k,n_k}(\boldsymbol{\theta}) = \binom{n_k}{2}^{-1} \sum_{1 \le i < j \le n_k} H(\boldsymbol{\theta}, \mathbf{y}_{k,i}, \mathbf{y}_{k,j}).$$

- Get the following AU-statistic based generalized estimating equation (AUGEE)

$$\tilde{U}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{K} n_k U_{k,n_k}(\boldsymbol{\theta}) = 0.$$

Properties of the AUGEE Estimator

In general, the UGEE and the AUGEE can be used to estimate the parameters in functional regression model.

- The AUGEE needs much less computation time than the UGEE in general.
- If $K = o(N)$, estimators from UGEE and AUGEE are asymptotically equivalent.

Simulation Setup

- Generate 100 data sets and each data set has 2000 observations.
- The true correlation $\rho_t = 0.2$ for $t = 1, 2, 3$.
- UGEE and AUGEE are used to estimate the coefficients.
- The program is in R.

- Box plots of estimates of correlation from different estimating equations. 1: UGEE; 2: AUGEE ($K = 5$); 3: AUGEE ($K = 10$) and 4: AUGEE ($K = 20$).

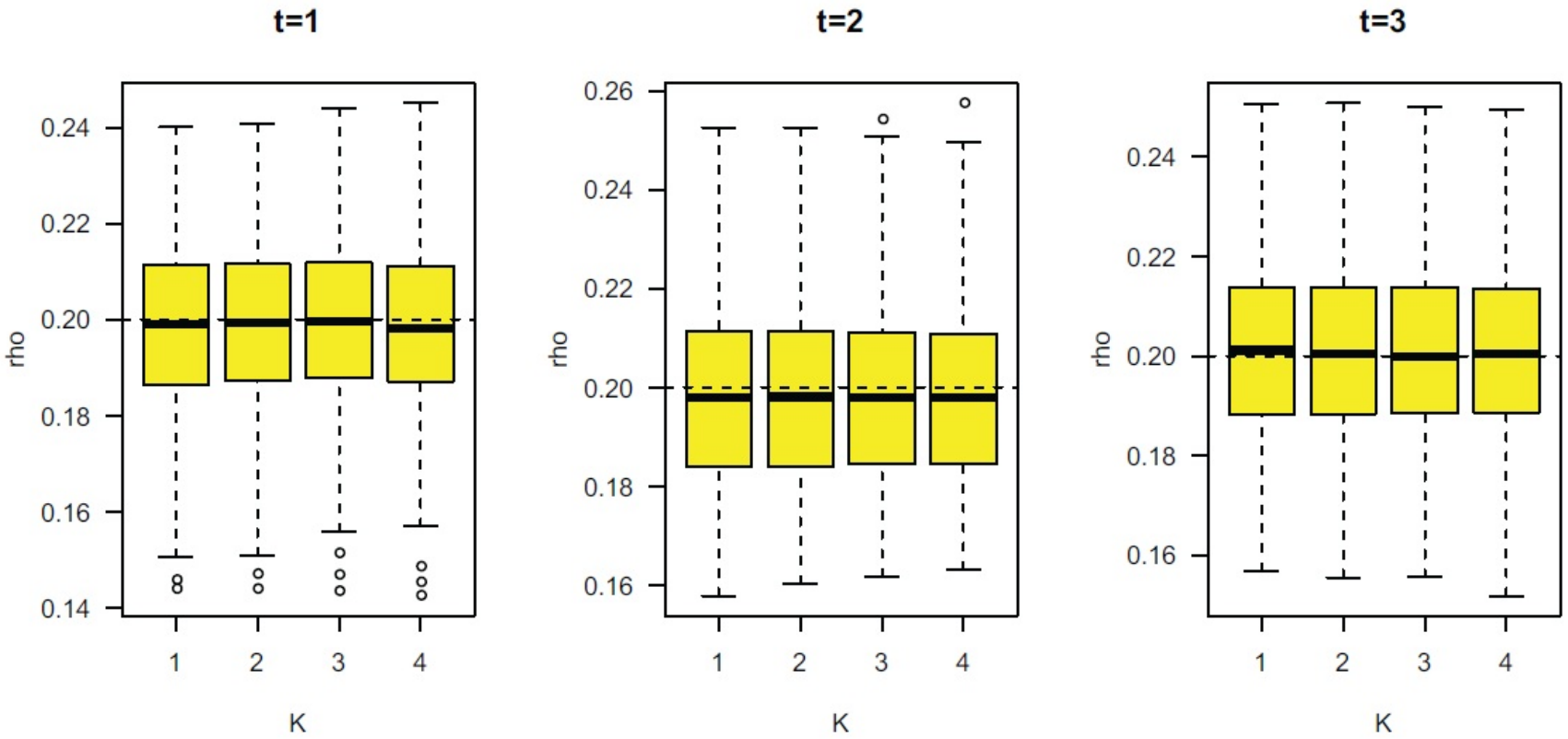


Table: Comparison of estimates from different estimating equations

	K	Mean	Variance($\times 10^{-4}$)	Time (seconds)
ρ_1 $(t=1)$	1	0.19813	3.71	71667.9
	5	0.19813	3.69	13050.8
	10	0.19823	3.74	6441.2
	20	0.19820	3.87	3223.2
ρ_2 $(t=2)$	1	0.19734	4.48	71667.9
	5	0.19740	4.47	13050.8
	10	0.19729	4.47	6441.2
	20	0.19730	4.47	3223.2
ρ_3 $(t=3)$	1	0.20021	3.30	71667.9
	5	0.20017	3.32	13050.8
	10	0.20010	3.33	6441.2
	20	0.19997	3.39	3223.2

A Real Example

Over-dispersion in next-generation sequencing data

- y_i: the number of reads of NGS data in a bin
- x_i: the percentage of GC-content in a bin
- 238,000 data points

$$\begin{aligned} E(y_i|\mathbf{x}_i) &= \mu_i = \exp(\mathbf{x}^T\boldsymbol{\beta}) \quad \mathbf{x}_i = (1, x_i, x_i^2) \\ E((y_i - \mu_i)^2) &= \lambda\mu_i \quad (\lambda > 0) \end{aligned}$$

A Real Example

Define

$$
\begin{aligned}
f_1(y_i, y_j) &= y_i + y_j \\
f_2(y_i, y_j) &= (y_i - y_j)^2 \\
h_1(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda) &= \mu_i + \mu_j \\
h_2(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda) &= \lambda(\mu_i + \mu_j) + (\mu_i - \mu_j)^2 \qquad (3) \\
\mathbf{f} &= (f_1, f_2) \\
\mathbf{h} &= (h_1, h_2).
\end{aligned}
$$

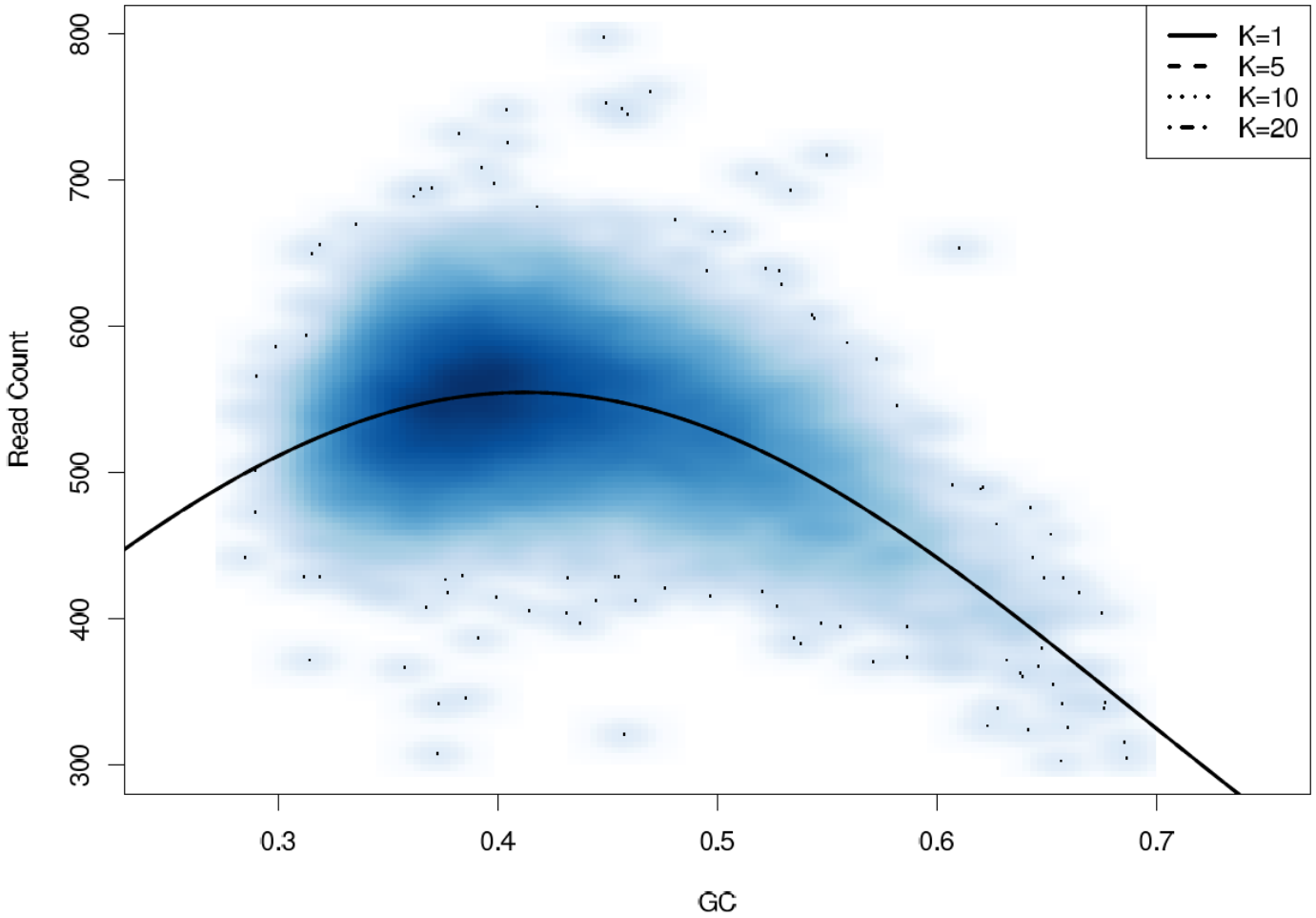
We have $E(\mathbf{f}(y_i, y_j)|\mathbf{x}_i, \mathbf{x}_j) = \mathbf{h}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\beta}, \lambda)$ and we may construct a UGEE using based on this equation.

A Real Example

Table: Comparison of estimates from different estimating equations

	K	Mean	Variance($\times 10^{-4}$)	Time (seconds)
β_0	1	5.000061	0.51	20647.9
	5	5.000005	0.51	4194.0
	10	4.999919	0.51	2091.3
	20	4.999733	0.51	1039.3
β_1	1	2.998614	7.58	20647.9
	5	2.998798	7.52	4194.0
	10	2.998992	7.50	2091.3
	20	2.999377	7.57	1039.3
β_2	1	−1.998244	6.15	20647.9
	5	−1.998412	6.12	4194.0
	10	−1.998500	6.10	2091.3
	20	−1.998635	6.19	1039.3
λ	1	2.010236	43.5	20647.9
	5	2.005958	43.6	4194.0
	10	1.999551	42.4	2091.3
	20	1.985545	42.9	1039.3

A Real Example



Thank you for your attention!