Mixture Models and Its Applications

Weixin Yao

Associate Professor, Department of Statistics, University of California

1 Introduction to Mixture models

Example 1.1 *The Fishery Data (Titterington et al. 1985): a data set of the length of 256 snappers, is an example from biology. Multiple groups exist since the fish begong to different age groups. As age is hard to measure, no observations concerning the age group a fish belongs to are available.*

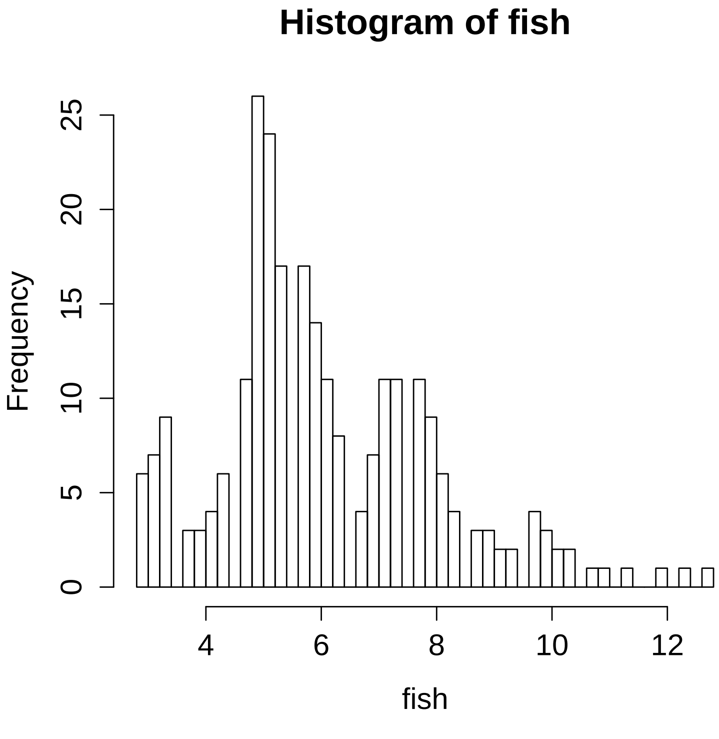


Figure 1: Fishery data

Example 1.2 *Old Faithful dataset: measurements give time in minutes between eruption of the Old Faithful geyser in Yellowstone National Park, USA.*

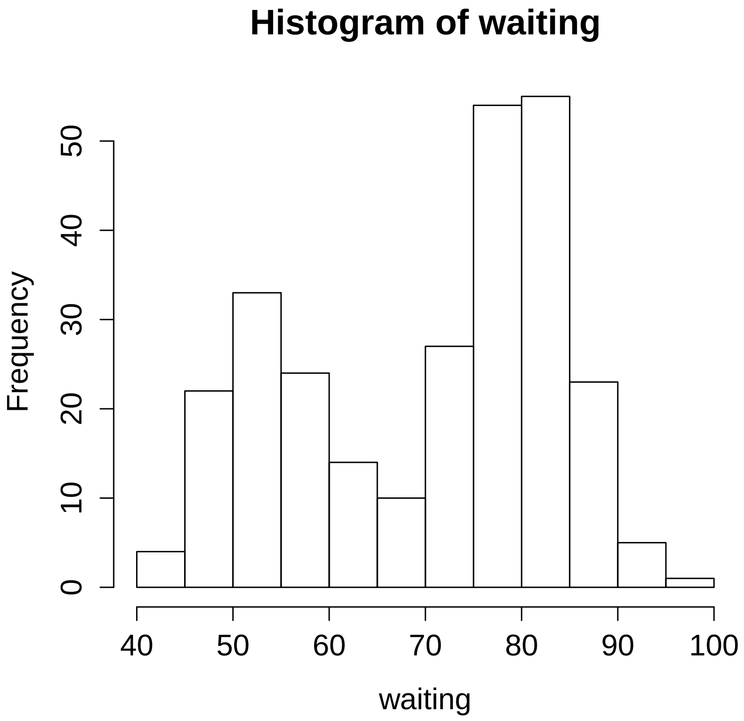
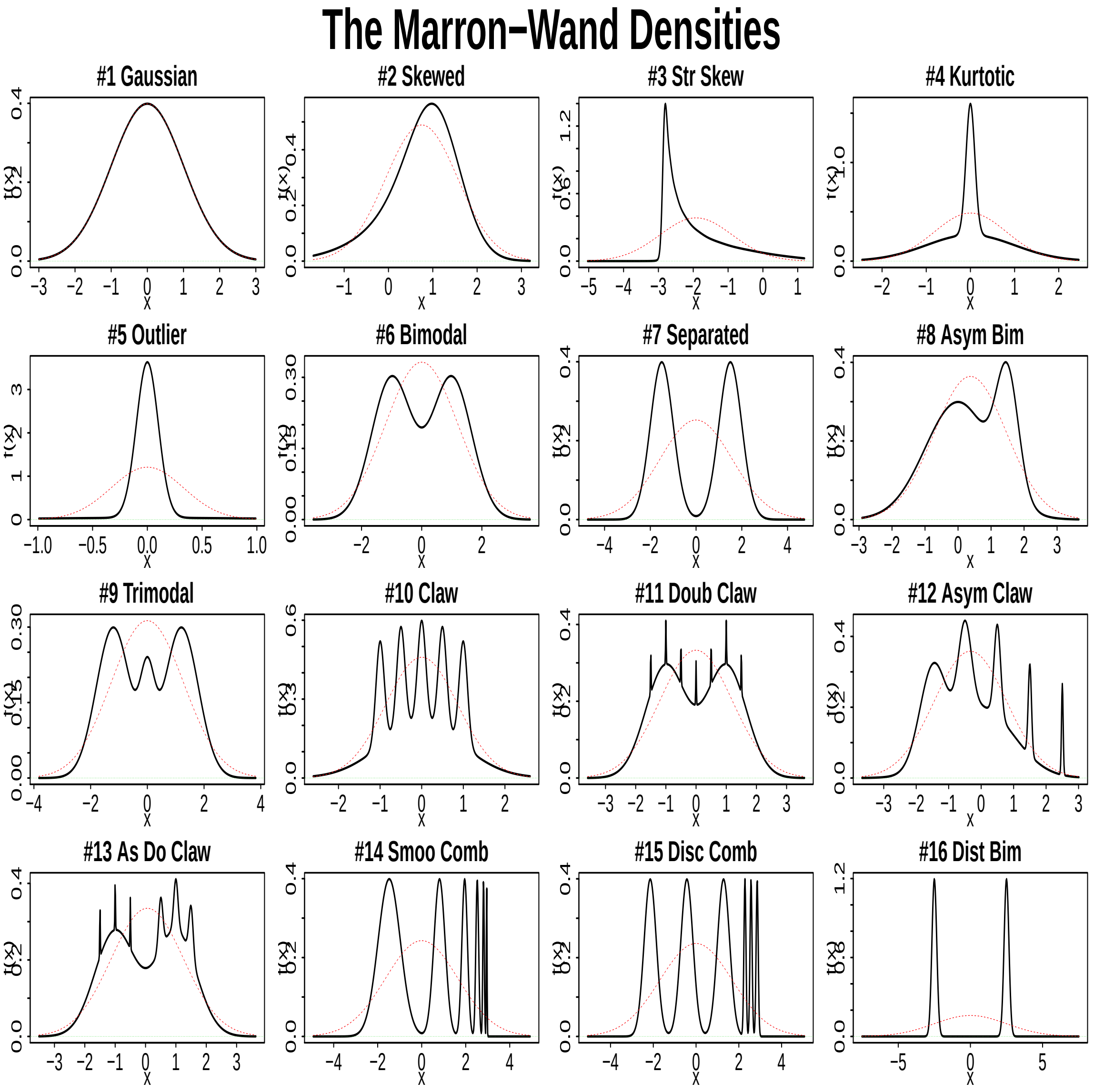


Figure 2: Faithful data

Example 1.3 *Mixture model is very flexible and can be used to represent a wide variety of density shapes.*



Model Setting: Suppose $(y_i, z_i), i = 1, \ldots, n$ satisfy the following model assumption

$$P(Z_i = j) = \pi_j, j = 1, \ldots, m$$
$$p(y \mid z = j) = p_j(y \mid \lambda_j),$$

where $\sum_{j=1}^m \pi_j = 1$. If $(z_1, \ldots, z_n)$ are missing, we only observe $(y_1, \ldots, y_n)$ from the marginal density of Y. Then, Y has the m-component finite *mixture* density

$$p(y; \boldsymbol{\theta}) = \sum_{j=1}^m p(y \mid Z = j) P(Z = j) = \sum_{j=1}^m \pi_j p_j(y \mid \lambda_j)$$

where the $\pi_j's$ are called component weights/mixing probabilities, $p_j(\mu \mid \lambda_j)$s are called component density, λ_js are component parameters, and θ is the collection of π_js and λ_js.

Define the latent variable Φ such that $P(\Phi = \lambda_j) = \pi_j$. Then the associated discrete distribution for Φ is

$$Q = \begin{pmatrix} \pi_1 & \cdots & \pi_m \\ \lambda_1 & \cdots & \lambda_m \end{pmatrix}.$$

Therefore,

$$p(y; \boldsymbol{\theta}) = p(y; Q) = \mathrm{E}(p(y; \Phi)) = \sum_{j=1}^m \pi_j p_j(y \mid \lambda_j).$$

Q is also called mixing distribution.

Example 1.4 *For the normal mixture*

$$\sum_{j=1}^{m} \pi_j N(\mu_j, \sigma_j^2),$$

where $N(\mu_j, \sigma_j^2)$ is the normal distribution with mean μ_j and variance σ_j^2, we have the following resutls:

$$\begin{aligned} E(Y) &= E[E(Y \mid Z)] = \sum_{j=1}^{m} \pi_j \mu_j \equiv \bar{\mu}, \\ Var(Y) &= E[Var(Y \mid Z)] + Var[E(Y \mid Z)] = \sum_{j=1}^{m} \pi_j \sigma_j^2 + \sum_{j=1}^{m} \pi_j (\mu_j - \bar{\mu})^2, \end{aligned}$$

2 Some of its applications

Mixture models can be applied in many fields such as astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing, physics, and social sciences.

Some Examples:

Clustering: Based on the mixture model fit, we can classify the labels of the points based on the classification probabilities.

Dealing with outlier modeling: For liner regression model

$$y = x\beta + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. If there are some outliers, we can fit a contaminated normal mixture (normal scale mixture model) for error ϵ

$$\pi_1 N(0, \sigma^2) + \pi_2 N(0, k\sigma^2),$$

where k is large and π_2 is supposed to be small. The second component represents the outlier component. Then we can use the EM algorithm to estimate β.

Density estimation: Mixture models can be used to approximate any density when the number of components is large. The kernel density estimation is a special case.

Mixture of regression models: $(\mathbf{x}_i, y_i, z_i), i = 1, \ldots, n$ are from the following model

$$y_i = \begin{cases} \mathbf{x}_i^T \beta_1 + \epsilon_1, & \text{if } z_i = 1; \\ \mathbf{x}_i^T \beta_2 + \epsilon_2, & \text{if } z_i = 2. \end{cases}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$. If $P(z_i = j) = \pi_j$, then without observing z

$$f(y \mid x, \theta) = \sum_{j=1}^{2} \pi_j \phi(y; x^T \beta_j, \sigma^2)$$

3 Estimation

Method of moments: Sample mean=population mean, sample variance=population variance,...

Comment: Computationally too difficult.

Maximum likelihood estimate (MLE): Given observations $(y_1, \ldots, y_n)$ from the mixture density

$$p(y; \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j f_j(y; \lambda_j),$$

where $\boldsymbol{\theta} = (\pi_1, \lambda_1, \ldots, \pi_m, \lambda_m)$, the log likelihood for $\boldsymbol{\theta}$ is

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \pi_j f(y_i; \lambda_j).$$

Its score function is

$$\begin{aligned}
\partial \log L(\boldsymbol{\theta})/\partial \lambda_j &= \sum_{i=1}^{n} \frac{\pi_j f'(y_i; \lambda_j)}{\sum_{j=1}^{m} \pi_j f(y_i; \lambda_j)}, j = 1, \ldots, m, \\
\partial \log L(\boldsymbol{\theta})/\partial \pi_j &= \sum_{i=1}^{n} \frac{f(y_i; \lambda_j) - f(y_i; \lambda_m)}{\sum_{j=1}^{m} \pi_j f(y_i; \lambda_j)}, j = 1, \ldots, m-1.
\end{aligned}$$

No explicit solution!

4 EM algorithm

EM(Expectation-Maximization) algorithm appears naturally in problems where

- some parts of the data are missing, and analysis of the incomplete data is somewhat complicated or nonlinear;
- it is possible to 'fill in' the missing data, and analysis of the complete data is relatively simple.

The notation of 'missing data' does not have to involve actually missing data, but any incomplete information.

Given the observed data y, suppose the the complete data are $x = (y, z)$, where z is missing. Our problem is to estimate θ from the likelihood based on y:

$$L(\theta; y) = p_\theta(y).$$

If the mle by maximize the above likelihood is difficult but the mle from the complete data likelihood

$$L_c(\theta; x) = p_\theta(x)$$

is easy if z is known, then we can use the EM algorithm to find the mle based on the observed data x.

EM algorithm: Start with an initial value θ^0, iterate the following two steps

- E step: compute the conditional expected value of the log complete likelihood

$$Q(\theta \mid \theta^{(k)}) = E\{\log L_c(\theta; x) \mid y, \theta^k\}$$

- M step: maximize $Q(\theta \mid \theta^{(k)})$ with respect to θ to given an updated value $\theta^{(k+1)}$.

Theorem 4.1 *One of the most important properties of the EM algorithm is that its step always increases the likelihood:*

$$L(\theta^{(k+1)}; y) \geq L(\theta^{(k)}; y).$$

Pf: Note that

$$\log L(\theta; x) = \log L(\theta; y) + \log L(\theta; z \mid y).$$

Let $h(\theta \mid \theta^{(k)}) = \mathrm{E}\{\log L(\theta; z \mid y) \mid y, \theta^{(k)}\}$. Hence,

$$\log L(\theta; y) = Q(\theta \mid \theta^{(k)}) - h(\theta \mid \theta^{(k)}).$$

From the information inequality, for any two densities $f(x) \neq g(x)$ we have

$$E_g \log f(X) \leq E_g \log g(X).$$

Applying this to the conditional density of $x \mid y$,

$$h(\theta^{(k+1)} \mid \theta^{(k)}) \leq h(\theta^{(k)} \mid \theta^{(k)}),$$

and at the next iterate $\theta^{(k+1)}$ we have

$$Q(\theta^{(k+1)} \mid \theta^{(k)}) \geq Q(\theta^{(k)} \mid \theta^{(k)}),$$

Hence,

$$\log L(\theta^{(k+1)}; y) \geq \log L(\theta^{(k)}; y). \ \square$$

This makes EM a numerically stable procedure as it climbs the likelihood surface; in contrast, no such guarantee exists for the Newton-Raphson algorithm. Another practical advantage of the EM algorithm is that is usually handles parameter constraints automatically. This is because each M step produces an MLE-type estimate. For example, estimates of probabilities are naturally constrained to be between zero and one. Note that the EM algorithm can only guarantee to converge to the local maximum instead of the global maximum. In complex cases it is important to try several starting values, or to start with a sensible estimate.

5 EM algorithm for mixture models

The log-likelihood based on the observed data $Y = (y_1, \ldots, y_n)$ is

$$\log L(\theta; Y) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j p_j(y_i \mid \lambda_j) \right\}.$$

Because of the constraints on $\pi_j's$, a simplistic application of the Newton-Raphson algorithm is prone to failure.

Let

$$z_{ij} = \begin{cases} 1, & \text{if } i\text{th observation is from the } j\text{th component;} \\ 0, & \text{ow.} \end{cases}$$

and $z_i = (z_{i1}, \ldots, z_{iJ})$. We define the 'complete data' $X = (x_1, \ldots, x_n)$, where $x_i = (y_i, z_i)$. Now the log likelihood of x is

$$\log L_c(\theta; X) = \sum_{i=1}^{n} \log L(\theta; x_i)$$

where the contribution of x_i to the log-likelihood is

$$\log L_c(\theta; x_i) = \sum_{j=1}^{m} z_{ij} \{\log p_j(y_i \mid \lambda_j) + \log \pi_j\}.$$

In E step, we need to calculate

$$Q(\theta \mid \theta^{(k)}) = E\{\log L_c(\theta; x) \mid y, \theta^k\}$$

which is simplified to calculating

$$\begin{aligned}
p_{ij}^{(k+1)} =& E\{z_{ij} \mid y_i, \theta^{(k)}\} \\
=& P(z_{ij} = 1 \mid y_i, \theta^{(k)}) \\
=& \frac{\pi_j^{(k)} p_j(y_i \mid \lambda_j^{(k)})}{\sum_{l=1}^{m} \pi_l^{(k)} p_l(y_i \mid \theta_l^{(k)})}.
\end{aligned}$$

In M step, we need to maximize

$$Q(\theta \mid \theta^{(k)}) = \sum_{j=1}^{m} p_{ij}^{(k+1)} \{\log p_j(y_i \mid \lambda_j) + \log \pi_j\}$$

EM algorithm: Starting with value $\theta^{(0)}$, in $(k+1)$th step,

E-step Finding the conditional probabilities

$$\begin{aligned}
p_{ij}^{(k+1)} &= E\{z_{ij} \mid y_i, \theta^{(k)}\} \\
&= P(z_{ij} = 1 \mid y_i, \theta^{(k)}) \\
&= \frac{\pi_j^{(k)} p_j(y_i \mid \lambda_j^{(k)})}{\sum_{l=1}^{m} \pi_l^{(k)} p_l(y_i \mid \theta_l^{(k)})}
\end{aligned}$$

This is the estimated probability of y_i coming from population j; in clustering problem it is the quantity of interest.

M-step Update each λ_j by

$$\lambda_j^{(k+1)} = \arg\max_{\lambda_j} \sum_{i=1}^{n} p_{ij}^{(k+1)} \log p_j(y_i \mid \lambda_j)$$

and π_j by

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ij}^{(k+1)}}{n}.$$

Comment: In EM step, if a hard label is drawn for each y_i from the multi-

nomial distribution with m categories specified by the $\{p_{ij}^{(k+1)}, j = 1, \ldots, m\}$, then such algorithm is called stochastic EM algorithm.

Example 5.1 *Suppose $y = (y_1, \ldots, y_n)$ are independent and identically distributed (iid) from the normal mixture*

$$\sum_{j=1}^{m} \pi_j N(\mu_j, \sigma_j^2) = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2) + \cdots + \pi_m N(\mu_m, \sigma_m^2).$$

The log-likelihood function is

$$\log L(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{m} \pi_j \phi(y_i; \mu_j, \sigma_j^2)\},$$

where $\boldsymbol{\theta} = (\pi_1, \mu_1, \sigma_1, \cdots, \pi_m, \mu_m, \sigma_m)$, and $\phi(y; \mu, \sigma)$ is the density of $N(\mu, \sigma^2)$.

The EM algorithm for $(k+1)^{th}$ step to find $\boldsymbol{\theta}$ is

E Step: Find the conditional probabilities

$$p_{ij}^{(k+1)} = \frac{\pi_j^{(k)} \phi(y_i; \mu_j^{(k)}, \sigma_j^{2(k)})}{\pi_1^{(k)} \phi(y_i; \mu_1^{(k)}, \sigma_1^{2(k)}) + \pi_2^{(k)} \phi(y_i; \mu_2^{(k)}, \sigma_2^{2(k)})}, \quad j = 1, \cdots, m.$$

M Step: Update the parameter estimate

$$\begin{aligned}
\mu_j^{(k+1)} &= \sum_{i=1}^{n} p_{ij}^{(k+1)} y_i / n_j^{(k+1)} \\
\sigma_j^{2(k+1)} &= \sum_{i=1}^{n} p_{ij}^{(k+1)} (y_i - \mu_j^{(k+1)})^2 / n_j^{(k+1)} \\
\pi_j^{(k+1)} &= n_j^{(k+1)} / n,
\end{aligned}$$

where $n_j^{(k+1)} = \sum_{i=1}^{n} p_{ij}^{(k+1)}$.

If we assume the variance is equal, i.e. $\sigma_1 = \sigma_2 = \cdots = \sigma$, then in the M step σ is updated by

$$\sigma^{2(k+1)} = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)}(y_i - \mu_j^{(k+1)})^2/n$$

Example 5.2 ***Mixtures of linear regressions****: Let Z be a latent class variable such that given $Z = j$, the response y depends on the $p-$dimensional predictor $\boldsymbol{x}$ in a linear way*

$$y = \boldsymbol{x}^T\boldsymbol{\beta}_j + \epsilon_j, j = 1, 2, \cdots, m, \tag{1}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\boldsymbol{x}$. Suppose $P(Z = j) = \pi_j, j = 1, 2, \cdots, m$, and Z is independent of $\boldsymbol{x}$, then the conditional density of Y given $\boldsymbol{x}$, without observing Z, is

$$f(y|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \boldsymbol{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{2}$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)^T$.

The unknown parameter $\boldsymbol{\theta}$, given observations $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, can be estimated by the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left[\sum_{j=1}^{m} \pi_j \phi(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2)\right]. \tag{3}$$

The EM algorithm for $(k+1)^{th}$ step to find $\boldsymbol{\theta}$ is

E Step: *Calculate the classification probabilities*

$$p_{ij}^{(k+1)} = \frac{\pi_j^{(k)}\phi(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(k)}, \sigma_j^{2(k)})}{\sum_{l=1}^{m}\pi_l^{(k)}\phi(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_l^{(k)}, \sigma_l^{2(k)})},\ i = 1, \ldots, n; j = 1, \ldots, m.$$

M Step: *Update the parameters*

$$
\begin{aligned}
\boldsymbol{\beta}_j^{(k+1)} &= \arg\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} p_{ij}^{(k+1)} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)^2 \\
&= (\sum_{i=1}^{n} p_{ij}^{(k+1)} \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \sum_{i=1}^{n} p_{ij}^{(k+1)} \boldsymbol{x}_i y_i \qquad (4) \\
&= (\boldsymbol{X}^T \boldsymbol{W}_j^{k+1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}_j^{(k+1)} \boldsymbol{y}, \qquad (5) \\
\pi_j^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(k+1)}, \\
\sigma_j^{2(k+1)} &= \frac{1}{\sum_{i=1}^{n} p_{ij}^{(k+1)}} \sum_{i=1}^{n} p_{ij}^{(k+1)} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(k+1)})^2,
\end{aligned}
$$

where $j = 1, \dots, m$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^T$, $\boldsymbol{y} = (y_1, \dots, y_n)^T$, *and* $\boldsymbol{W}_j^{(k+1)}$ *is a* $n \times n$ *diagonal matrix with diagonal elements* $\{p_{ij}^{(k+1)}, i = 1, \dots, n\}$.

If we assume the variance is equal, i.e. $\sigma_1 = \sigma_2 = \cdots = \sigma$, then in the M step σ is updated by

$$
\sigma^{2(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k+1)})^2.
$$

Example 5.3 *Consider a* CO_2*-GDP dataset published by World Resource Institute. As shown in Figure 3, the* CO_2*-GDP dataset contains two related variables of 171 countries in year 2005. The response variable is the* CO_2*-emission per capita in year 2005, and the predictor is the GDP per capita in the same year, measured by the current US dollars.*

The purpose of the analysis is to identify the group of countries through their development path as featured by the relationship of GDP and CO2-emission. We know that GDP is a measure of the size of a nation's economy, and Carbon dioxide (CO2) is an important greenhouse gas which causes the greenhouse effect and may relate to global warming. Development with *high GDP per capita and relative low CO2-emission is a desired goal and consensus for modern governments.*

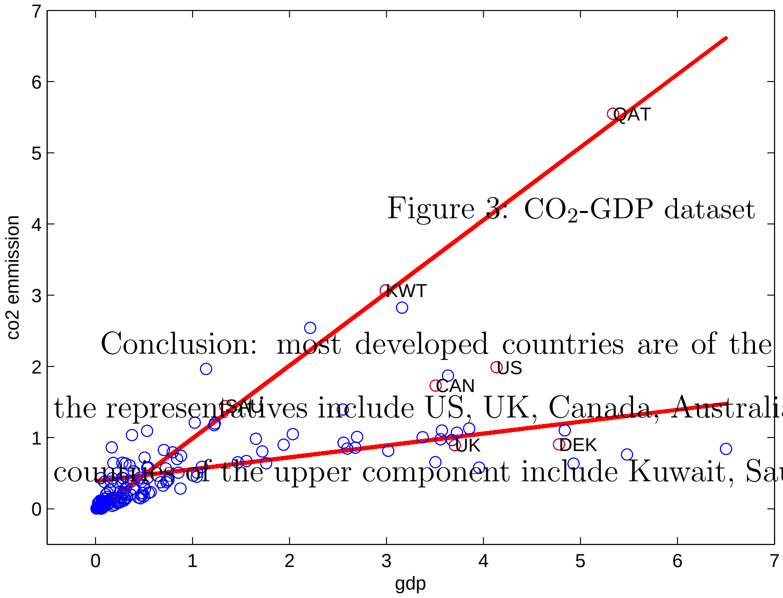


Figure 3: CO_2-GDP dataset

Conclusion: most developed countries are of the lower component, and the representatives include US, UK, Canada, Australia, etc. Representatives countries of the upper component include Kuwait, Saudi Arabia, Qatar, etc.

6 Starting value and stopping rule

The solution found by EM algorithm depends on the starting point, and there is no guarantee that the EM algorithm will converge to the global maximum. Therefore, it is prudent to run the algorithm from several starting-points and choose the best local optima found.

Possible starting values:

1. First use some clustering method (such as k-means clustering) to cluster data into m clusters. Then in each cluster, estimate the parameter values.

2. Random initial parameter values. We can first randomly partition the data or part of the data into m groups. Then estimate the component parameters based on each subgroup.

Stopping rules:

1. The changes of the parameter values are small enough, such as $||\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|| < \epsilon$.

2. The changes of likelihood values are small enough (better), such as $\log L(\boldsymbol{\theta}^{(k+1)}) - \log L(\boldsymbol{\theta}^{(k)}) < \epsilon$.

7 Unboundedness of mixture likelihood

Example 7.1 *Suppose $y = (y_1, \ldots, y_n)$ are iid from the normal mixture*

$$\sum_{i=1}^{m} N(\mu_j, \sigma_j^2).$$

The log-likelihood function is

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n} \log\{\sum_{j=1}^{m} \pi_j \phi(y_i; \mu_j, \sigma_j^2)\} \\ &= \sum_{i=1}^{n} \log\{\sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\{-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\}\end{aligned}$$

Then $\log L(\theta)$ will go to infinity if μ_j equals to one observation and σ_j goes to zero.

For mixture of regression, suppose $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ has the mixture log-likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{m} \pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2)\}.$$

Then $\log L(\boldsymbol{\theta})$ will go to infinity if $(\mathbf{x}_i, y_i)$ falls in one component line, say jth component, and σ_j goes to 0.

Possible solutions:

1. Assume the component variance is equal, i.e., $\sigma_1 = \sigma_2 = \cdots = \sigma_m$.

2. Find the maximum interior mode, i.e., run the EM algorithm over a constrained parameter space

$$\Omega_C = \{\boldsymbol{\theta} \in \Omega : \sigma_h/\sigma_j \geq C > 0, 1 \leq h \neq j \leq m\}, \qquad (6)$$

   where $C \in (0, 1]$, Ω denotes the unconstrained parameter space. See Hathaway (1985, 1986) and Bezdak, Hathaway, and Huggins (1985) for more detail.

8 Shapes of some univariate normal mixtures

Consider two component normal mixture with equal component variance

$$f(y) = \pi_1\phi(y;\mu_1,\sigma^2) + \pi_2\phi(y;\mu_2,\sigma^2).$$

Let $\Delta = |\mu_2 - \mu_1|/\sigma$.

Results: When $\pi_1 = \pi_2 = 0.5$, then $f(y)$ has two modes if $\Delta > 2$ but only has one unique mode if $\Delta \leq 2$.

Example 8.1 *Let $\mu_1 = 0, \sigma = 1, \pi_1 = 0.5$. Figure 4 shows the plot of $f(y)$ versus y when $\Delta = 1, 2, 3, 4$.*

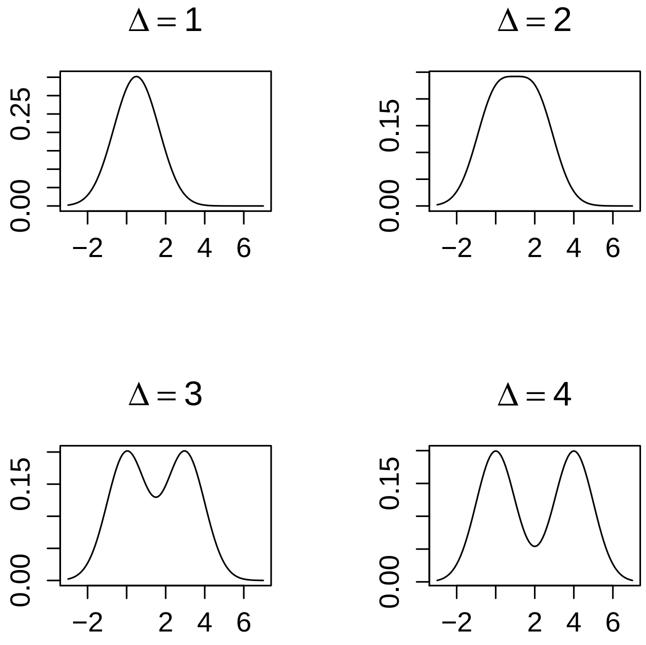


Figure 4: $\mu_1 = 0, \sigma = 1, \pi_1 = \pi_2 = 0.5, \Delta = \mu_2$

Example 8.2 *Let $\mu_1 = 0, \sigma = 1, \pi_1 = 0.75, \pi_2 = 0.25$. Figure 5 shows the plot of $f(y)$ versus y when $\Delta = 1, 2, 3, 4$.*

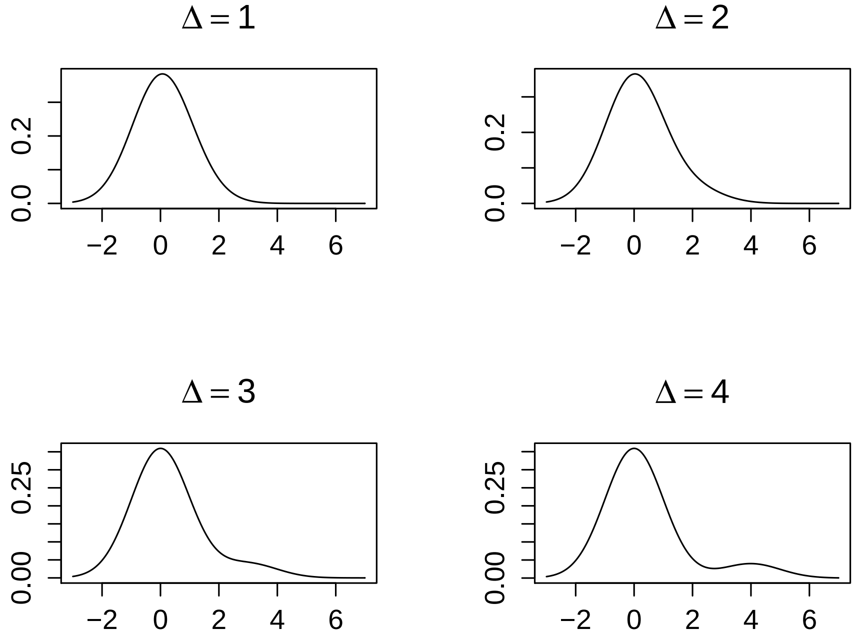


Figure 5: $\mu_1 = 0, \sigma = 1, \pi_1 = 0.75, \pi_2 = 0.25, \Delta = \mu_2$

9 Choose the number of components

Commonly used methods: Model selection by information criteria.

- BIC: $-2\log L + p\log(n)$.
- AIC: $-2\log L + 2p$.

10 Label switching

The m-component mixture models we consider here have densities of the form

$$p(x;\boldsymbol{\theta}) = \pi_1 f(y;\lambda_1) + \pi_2 f(y;\lambda_2) + \cdots + \pi_m f(y;\lambda_m)\,.$$

For any permutation $\boldsymbol{\omega} = (\boldsymbol{\omega}(1),\ldots,\boldsymbol{\omega}(m))$ of the identity permutation $(1,\ldots,m)$, define the corresponding permutation of the parameter vector $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{\boldsymbol{\omega}} = (\pi_{\boldsymbol{\omega}(1)},\ldots,\pi_{\boldsymbol{\omega}(m)},\lambda_{\boldsymbol{\omega}(1)},\ldots,\lambda_{\boldsymbol{\omega}(m)})^T.$$

Supposing that $\mathbf{y} = (y_1,\ldots,y_n)$ is a random sample from the m-component mixture density, the likelihood for $\mathbf{y}$ is

$$L(\boldsymbol{\theta};\mathbf{y}) = \prod_{i=1}^{n}\{\pi_1 f(y_i;\lambda_1) + \pi_2 f(y_i;\lambda_2) + \cdots + \pi_m f(y_i;\lambda_m)\}\,. \qquad (7)$$

For any permutation $\boldsymbol{\omega}$, $L(\boldsymbol{\theta}^{\boldsymbol{\omega}};\mathbf{x})$ will be numerically the same as $L(\boldsymbol{\theta};\mathbf{x})$. Hence if $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE), $\hat{\boldsymbol{\theta}}^{\boldsymbol{\omega}}$ is the MLE for any permutation $\boldsymbol{\omega}$. (In a technical sense, this means that the subscripts we assign to the π's and λ's are not identifiable unless we put additional restrictions on the model.) This is the so-called *label switching.*

In a simulation study, given a sequence of raw unlabeled simulated estimates $(\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_N)$ of $\boldsymbol{\theta}$, in order to measure their bias and variation, one must first label these samples, i.e., find the labels $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N)$ such that $(\hat{\boldsymbol{\theta}}_1^{\boldsymbol{\omega}_1}, \ldots, \hat{\boldsymbol{\theta}}_N^{\boldsymbol{\omega}_N})$ have the same label meaning. Then one can use the labeled estimates to estimate the variation and the bias. Without "correct" labels, the estimates tend to have serious bias and the estimated variation might also be misleading.

Possible solutions:

1. Put an explicit parameter constraint (such as $\mu_1 < \mu_2 < \ldots < \mu_m$ or $\pi_1 < \pi_2 < \ldots < \pi_m$ for univariate data) so that only one permutation can satisfy it. Difficult to use for high dimension data. Different order constraint may generate markedly different results.

2. Label the samples based on minimizing the following negative log normal likelihood over $(\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}, \boldsymbol{\omega})$(Yao, 2009),

$$L(\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}, \boldsymbol{\omega}) = N \log(|\boldsymbol{\Sigma}|) + \sum_{t=1}^{N} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}}). \qquad (8)$$

Step 1: Update $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}$

$$\bar{\boldsymbol{\theta}} = \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\theta}_t^{\boldsymbol{\omega}_t},$$
$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}_t} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_t^{\boldsymbol{\sigma}_t} - \bar{\boldsymbol{\theta}})^T .$$

Step 2: For $t = 1, \ldots, N$, choose $\boldsymbol{\sigma}_t$ by

$$\boldsymbol{\sigma}_t = \arg\min_{\boldsymbol{\omega}} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}} - \bar{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_t^{\boldsymbol{\omega}} - \bar{\boldsymbol{\theta}}) . \quad \square$$

The New York Times

August 6, 2009

For Today’s Graduate, Just One Word: Statistics

By STEVE LOHR

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

“People think of field archaeology as Indiana Jones, but much of what you really do is data analysis,” she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by 2012, according to a projection by IDC, a research firm.

Yet data is merely the raw material of knowledge. “We’re rapidly entering a world where everything can be monitored and measured,” said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology’s Center for Digital Business. “But the big problem is going to be the ability of humans to use, analyze and make sense of the data.”

The new breed of statisticians tackle that problem. They use powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data. The applications are as diverse as improving Internet search and online advertising, culling gene sequencing information for cancer research and analyzing sensor and location data to optimize the handling of food shipments.

Even the recently ended Netflix contest, which offered $1 million to anyone who could significantly improve the company’s movie recommendation system, was a battle waged with the weapons of modern statistics.

Though at the fore, statisticians are only a small part of an army of experts using modern statistical techniques for data analysis. Computing and numerical skills, experts say, matter far more than degrees. So the new data sleuths come from backgrounds like economics, computer science and mathematics.

They are certainly welcomed in the White House these days. “Robust, unbiased data are the first step toward addressing our long-term economic needs and key policy priorities,” Peter R. Orszag, director of the Office of Management and Budget, declared in a speech in May. Later that day, Mr. Orszag confessed in a blog entry that his talk on the importance of statistics was a subject “near to my (admittedly wonkish) heart.”

I.B.M., seeing an opportunity in data-hunting services, created a Business Analytics and Optimization Services group in April. The unit will tap the expertise of the more than 200 mathematicians, statisticians and other data analysts in its research labs — but that number is not enough. I.B.M. plans to retrain or hire 4,000 more analysts across the company.

In another sign of the growing interest in the field, an estimated 6,400 people are attending the statistics profession’s annual conference in Washington this week, up from around 5,400 in recent years, according to the American Statistical Association. The attendees, men and women, young and graying, looked much like any other crowd of tourists in the nation’s capital. But their rapt exchanges were filled with talk of randomization, parameters, regressions and data clusters. The data surge is elevating a profession that traditionally tackled less visible and less lucrative work, like figuring out life expectancy rates for insurance companies.

Ms. Grimes, 32, got her doctorate in statistics from Stanford in 2003 and joined Google later that year. She is now one of many statisticians in a group of 250 data analysts. She uses statistical modeling to help improve the company’s search technology.

For example, Ms. Grimes worked on an algorithm to fine-tune Google’s crawler software, which roams the Web to constantly update its search index. The model increased the chances that the crawler would scan frequently updated Web pages and make fewer trips to more static ones.

The goal, Ms. Grimes explained, is to make tiny gains in the efficiency of computer and network use. “Even an improvement of a percent or two can be huge, when you do things over the millions and billions of times we do things at Google,” she said.

It is the size of the data sets on the Web that opens new worlds of discovery. Traditionally, social sciences tracked people’s behavior by interviewing or surveying them. “But the Web provides this amazing resource for observing how millions of people interact,” said Jon Kleinberg, a computer scientist and social networking researcher at Cornell.

For example, in research just published, Mr. Kleinberg and two colleagues followed the flow of ideas across cyberspace. They tracked 1.6 million news sites and blogs during the 2008 presidential campaign, using algorithms that scanned for phrases associated with news topics like “lipstick on a pig.”

The Cornell researchers found that, generally, the traditional media leads and the blogs follow, typically by 2.5 hours. But a handful of blogs were quickest to quotes that later gained wide attention.

The rich lode of Web data, experts warn, has its perils. Its sheer volume can easily overwhelm statistical models. Statisticians also caution that strong correlations of data do not necessarily prove a cause-and-effect link.

For example, in the late 1940s, before there was a polio vaccine, public health experts in America noted that polio cases increased in step with the consumption of ice cream and soft drinks, according to David Alan Grier, a historian and statistician at George Washington University. Eliminating such treats was even recommended as part of an anti-polio diet. It turned out that polio outbreaks were most common in the hot months of summer, when people naturally ate more ice cream, showing only an association, Mr. Grier said.

If the data explosion magnifies longstanding issues in statistics, it also opens up new frontiers.

“The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd,” said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. “And that makes it easier for humans to do what they are good at — explain those anomalies.”

Andrea Fuller contributed reporting.

Talk about new york time magazines.

Why statistics is important?

1. Many election year polls. The Literary Digest story–1936 presidential poll. Candidates: Republican– Alfred Landon (Landon lecture series in K-state is named after him). Democratic: Franklin Roosevelt.

   The Literary Digest polled 10 millions from readers, registered automobile owners, and phone users and 2.4 million responded. The poll showed that Governor Alfred Landon of Kansas, was likely to be the overwhelming winner.

   Gallup's poll achieved national recognition by correctly predicting the result of the 1936 election using a smaller sample size of 50,000. (A much lower number, such as 1,500 persons, is adequate in most cases so long as they are appropriately chosen.)

   Problems: the polled persons are not appropriately chosen. The persons polled by the Literary Digest have incomes well above average.

2. Insurance: The rate that an insurance company charges you is based upon statistics from all drivers or homeowners in you area.

   - Well educated drives usually have less car insurance.
   - Married drivers usually have less insurance rate.

3. Business and Industry: Statisticians quantify unknowns in order to optimize resources. They need to

   - predict the demand for products and services
   - Which kind of advertisement is useful?
   - Which location needs their special focus?
   - Medical studies: scientists must show a statistically valid rate of effectiveness before any drug can be prescribed. Statistics are behind every medical study you hear about.

4. In biology, statistician are helping find the useful genes and determine the whole network of genes.

5. Quality Testing: companies make thousands of products every day and each company must make sure that a good quality item is sold. But a company can't test each and every time that they ship to you, the consumer. So the company uses statistics to test just a few, called a sample, of what they make. If the sample passes quality tests, then the company assumes that all the items made in the group are good.

6. Stock market: Stock analysts also use statistical computer models to forecast what is happening in the economy and to predict the stock trend in the future.
7. Iraq war: whether the improvement is significant. It is a battle between statisticians.
8. Women is unfairly paid in a company compared to men.
9. Netflix competition

Future of Statistics Big data: Apple, google, gmail predicts what you want based on email. Amazon can predict what you want to buy before you order so that they can manage their shipping or storage earlier.

Example 0.0.1 *52 prostate cancer patients and 50 normal controls have each had his genetic expression levels measured on* $N = 6033$ *genes. For each gene we can get a* t *statistics.*

Verify the accuracy and reliability of data during the process of data collection

1. Define population well and sample data from the right population. Using poll example.
2. Non-response. Non response rate is not very important. The nature of the nonrespondents is more important. You can't use the neighbor to replace the person who is not at home at 3pm. Using qinghua and beida can increase the response rate.
3. Clerical errors: check a sampling of the data collection sheet against the original source documents. If more than 10 percent sample entries are wrong, we can take action. The errors are found only by one collector or all of them. If all collector has high error rate, the collection procedures or questions designed might be problematic.
   For example:

   - interviewer shouldn't have shown the attitude toward an question.

   - how many grown-ups in the family should be replaced by how many persons in the family who is older than say 18.

   - The ranges of measurements can be checked. the weight can be 1040.

   - Question of ordering
     A. Do you think the united states should let communist newspaper reporters from other countries come in here and send back to their papers the news as they see it
     B. Do you think a communist country like Russia should let American newspaper

reporters come in and send back to America the news as they see it?
Another example
A. will you support an increase in state taxes for education?
B. Will you support an increase in state taxes?

-

4. Good design: record both age and year of birth. the hours per day assigned to different work tasks can't sum to more than 24.

5. The survey results could be checked against known facts. For example, the sample only contains 10% of women. The average income of collected sample is much lower than the reported average from other sources.

6. Some data can be verified by data capture programs. For example, when recording votes, record the home address of each votes and verify where the address is real in case collectors just made up.

7. Calculate the trends up to now and check the trend will vary unusually.