

Técnicas de Caracterização de Carga de Trabalho

Professor Alexandre Magno de Sousa
Curso de Engenharia de Computação
Departamento de Computação e Sistemas

Sumário

- Terminologia
- Componentes e seleção de parâmetros
- Técnicas de caracterização de carga.
- Técnicas de clusterização.
- Interpretação e problemas com clusterização.
- Bibliografia.

Terminologia

- Usuário = entidade que realiza as solicitações de serviço.
- Componentes do workload: aplicações; sites; sessões do usuário.
- Parâmetros (características) do Workload: quantidades medidas, solicitações de serviço ou demandas do recurso.
- Exemplos: tipos de transações, instruções, tamanhos de pacotes, fonte e destino de um pacote, padrões de referências de página.

Componentes e Seleção de Parâmetros

- Na escolha dos parâmetros para caracterizar a carga, é melhor utilizar parâmetros que dependem da carga mais do que do sistema.
 - Exemplo: o tempo de resposta de uma transação não é um parâmetro adequado para um workload uma vez que ele depende do sistema em que a transação é executada.
- # de solicitações mais do que a quantidade do recurso demandado é preferível.

Componentes e Seleção de Parâmetros

- Existem diversas características de solicitações de serviço que são de interesse.
 - Exemplo: tempo de chegada, tipo de recurso demandado, duração da solicitação, quantidade de recurso demandado pela solicitação.
- Essas características têm impacto no desempenho e devem ser incluídas.
- As que têm pouco impacto, devem ser excluídas.
- Se o tamanho do pacote não tem impacto no tempo de envio pelo roteador, deve ser omitido, no lugar, o # de pct e tempo de chegada deve ser considerado.

Componentes e Seleção de Parâmetros

- Técnicas:
 - Média.
 - Especificação da dispersão.
 - Histogramas de único e múltiplos parâmetros.
 - Análise do componente principal (PCA).
 - Modelos de markov.
 - Clusterização.

Média e Dispersão

- Média: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 - Obs.: existem casos que a mediana, moda, média geométrica ou média harmônica devem ser utilizadas (veja cap. 12 Jain).
- A variabilidade é especificada pela variância:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Desvio padrão: $s = \sqrt{s^2}$
- Coeficiente de Variação (CV): $CV = \frac{s}{\bar{x}}$

Média e Dispersão

- Exemplo: as demandas de recurso de vários programas executados no campus de 6 universidades foram medidas por 6 meses:

Dados	Média	CV
Tempo de CPU	2.19 segundos	40.23
# de writes	8.20	53.59
Writes bytes	10.21 Kbytes	82.41
# de reads	22.64	25.65
Reads bytes	49.70 Kbytes	21.01

Média e Dispersão

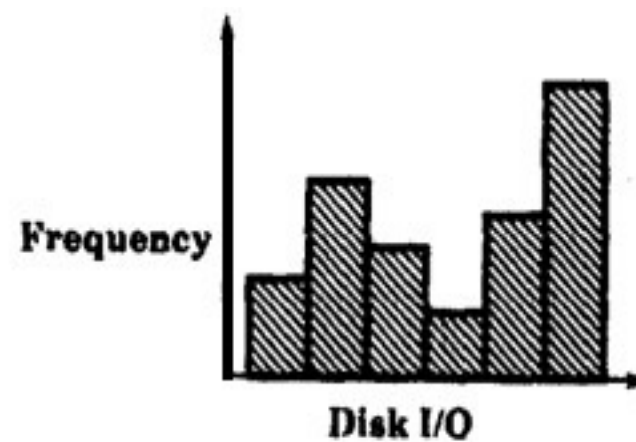
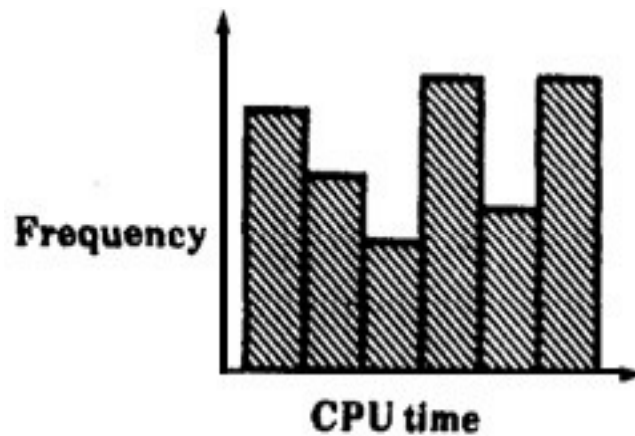
- Observe que o *CV* dos valores medidos é muito alto, indicando que combinar todos os programas para obter a média não é uma boa ideia.
- Os programas devem ser divididos em classes. Os dados da tabela a seguir mostram a demanda para todas sessões de edição. Veja que o *CV* agora é baixo.

Dados	Média	CV
Tempo de CPU	2.57 segundos	3.54
# de writes	19.74	4.33
Writes bytes	13.46 Kbytes	3.87
# de reads	37.77	3.73
Reads bytes	36.93 Kbytes	3.16

Histogramas de Parâmetro Único

- Um histograma mostra a frequência relativa de vários valores de um parâmetro.
- Para parâmetros de valores contínuos, requer a divisão da faixa de parâmetros em várias subfaixas pequenas chamadas células e a contagem das observações em cada célula (alta variabilidade).
- Esses dados podem ser utilizados na medição ou na simulação para gerar o workload de testes.
- Em modelagem analítica histogramas podem ser utilizados para ajustar a distribuição de probabilidade.
- O uso de histogramas de parâmetro único (HPU) é que eles ignoram a correlação entre diferentes parâmetros.

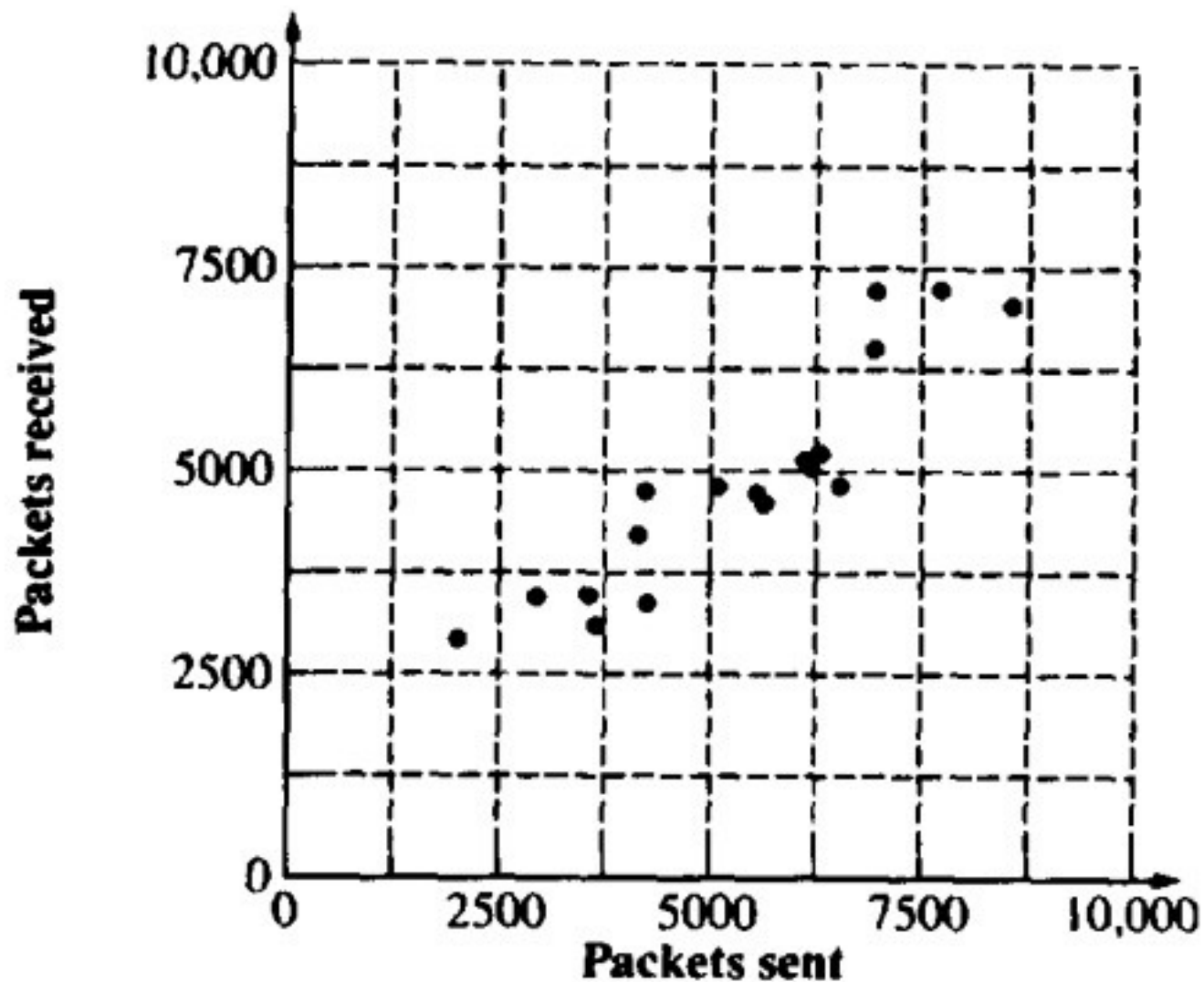
Histogramas de Parâmetro Único

[illegible]

Histogramas de Múltiplos Parâmetros

- Se existir correlação entre diferentes parâmetros do workload, um histograma de múltiplos parâmetros (HMP) deve ser utilizado.
- Uma matrix $n \times n$ é utilizada para descrever a distribuição de n parâmetros.
- É difícil construir histogramas para mais de dois parâmetros.
- Em alguns casos, HPU são muito detalhados.
- HMP são ainda mais! → raramente são utilizados

Histogramas de Múltiplos Parâmetros



Análise do Componente Principal

- Utilizada para classificar componentes pela soma do peso dos valores dos parâmetros.
- Utilizando a_j como peso para o j -ésimo parâmetro x_j , a soma dos pesos é dada por:

$$y = \sum_{j=1}^n a_j x_j$$

- Esta soma pode ser utilizada para classificar os componentes dentro de uma classe (e.g. baixa/alta demanda).
- Utilizada na avaliação de desempenho de software: na maioria dos casos, o peso é dado pela pessoa que executa.

Análise do Componente Principal

- Método para determinar o peso: análise do componente principal (PCA).
- Permite encontrar o peso a_j de modo que y_j fornece a discriminação máxima entre componentes.
- Fator principal: y_j
- Estatisticamente, dado um conjunto de n parâmetros $\{x_1, x_2, \dots, x_n\}$, a PCA produz um conjunto de valores $\{y_1, y_2, \dots, y_n\}$ de modo que mantém:

Análise do Componente Principal

1º) Os y s são combinações lineares de x s

$$y = \sum_{j=1}^n a_{ij} x_j$$

cada a_{ij} é chamado de carga da x_j .

2º) Os y 's formam um conj. ortogonal, isto é, o produto interno é zero:

$$\langle y_i, y_j \rangle = \sum_{k=1}^n a_{ik} a_{kj} = 0$$

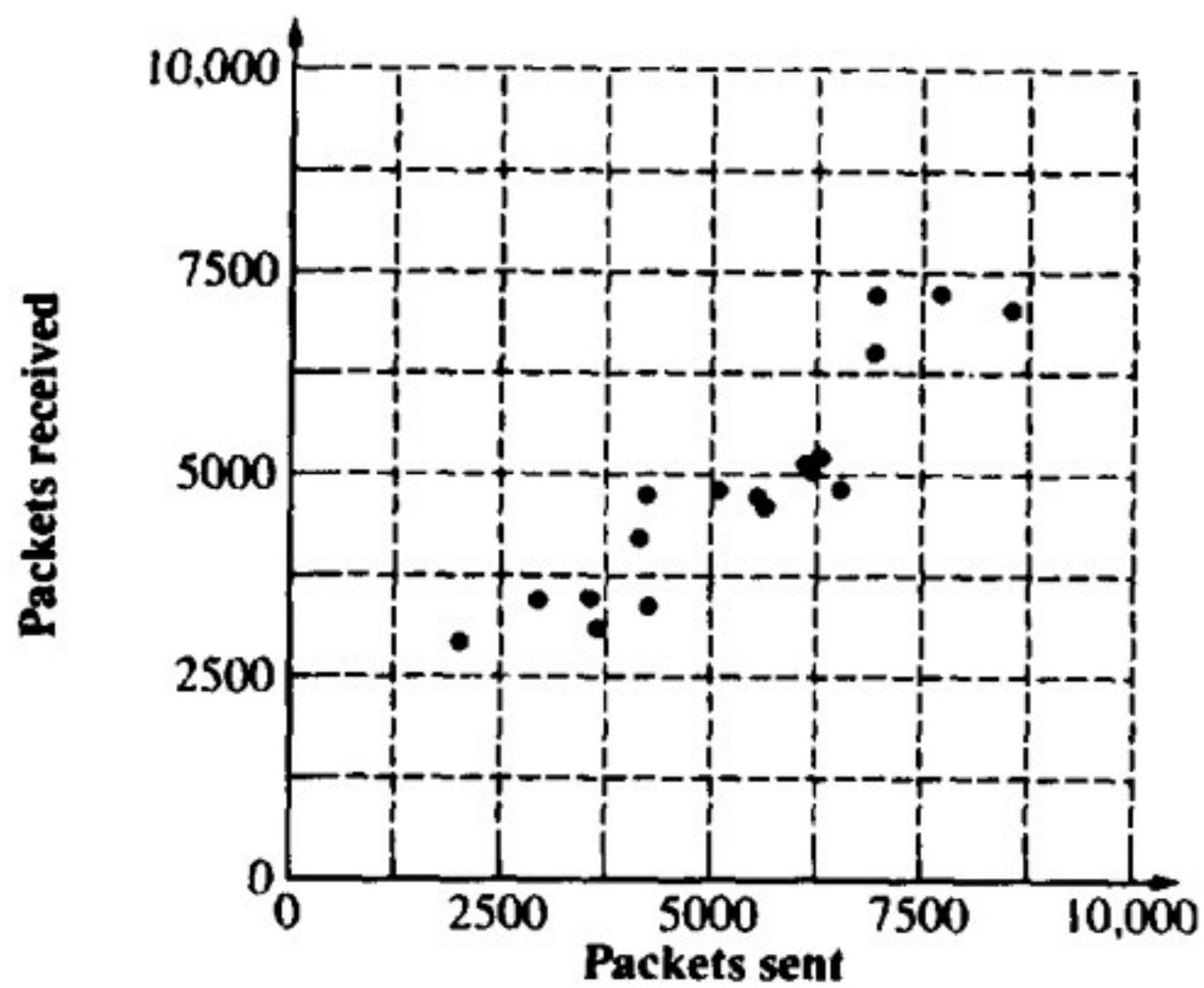
os y s não são correlacionados.

3º) Os y s são um conj. ordenado: y_1 explica o maior percentual da variância na demanda do recurso, y_2 explica um percentual menor do que y_1 , e assim por diante. Os 1ºs fatores são utilizados para classificar.

Análise do Componente Principal

- Exemplo: o # de pcts enviados e recebidos, definidos por x_s e x_r , por várias estações em uma LAN foram medidos. Os dados são apresentados na 1ª e 2ª colunas na próxima tabela. Um histograma dos dados é apresentado em seguida, existe uma correlação considerável entre as duas variáveis.

Observation No.	Variables		Normalized Variables		Principal Factors	
	x_s	x_r	x'_s	x'_r	y_1	y_2
1	7718	7258	1.359	1.717	2.175	-0.253
2	6958	7232	0.922	1.698	1.853	-0.549
3	8551	7062	1.837	1.575	2.413	-0.186
4	6924	6526	0.903	1.186	1.477	-0.200
5	6298	5251	0.543	0.262	0.570	0.199
6	6120	5158	0.441	0.195	0.450	0.174
7	6184	5051	0.478	0.117	0.421	0.255
8	6527	4850	0.675	-0.029	0.457	0.497
9	5081	4825	-0.156	-0.047	-0.143	-0.077
10	4216	4762	-0.652	-0.092	-0.527	-0.396
11	5532	4750	0.103	-0.101	0.002	0.145
12	5638	4620	0.164	-0.195	-0.022	0.254
13	4147	4229	-0.692	-0.479	-0.828	-0.151
14	3562	3497	-1.028	-1.009	-1.441	-0.013
15	2955	3480	-1.377	-1.022	-1.696	-0.251
16	4261	3392	-0.627	-1.085	-1.211	0.324
17	3644	3120	-0.981	-1.283	-1.601	0.213
18	2020	2946	-1.914	-1.409	-2.349	-0.357
$\sum x$	96,336	88,009	0.000	0.000	0.000	0.000
$\sum x^2$	567,119,488	462,661,024	17.000	17.000	32.565	1.435
Mean	5352.0	4899.4	0.000	0.000	0.000	0.000
Standard Deviation	1741.0	1379.5	1.000	1.000	1.384	0.290



Análise do Componente Principal

1º) Calcule a média e o desvio padrão das variáveis:

$$\bar{x}_s = \frac{1}{n} \sum_{i=1}^n x_{si} = \frac{96336}{18} = 5352.0$$

$$\bar{x}_r = \frac{1}{n} \sum_{i=1}^n x_{ri} = \frac{88009}{18} = 4889.4$$

$$s_{xs}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{si} - \bar{x}_s)^2$$

$$s_{xs}^2 = \frac{1}{n-1} \left[\left(\sum_{i=1}^n x_{si}^2 \right) - n \bar{x}_s^2 \right]$$

$$s_{xs}^2 = \frac{567119488 - 18 \times 5352^2}{17} = 1741.0^2$$

$$s_{xr}^2 = \frac{462661024 - 18 \times 4889.4^2}{17} = 1379.5^2$$

Análise do Componente Principal

2º) Normalize as variáveis com a média zero e desvio padrão 1:

$$x'_s = \frac{x_s - \bar{x}_s}{s_{xs}} = \frac{x_s - 5352}{1741}$$

$$x'_r = \frac{x_r - \bar{x}_r}{s_{xr}} = \frac{x_r - 4889}{1380}$$

Os valores normalizados são apresentados na 4ª e 5ª coluna da tabela.

Análise do Componente Principal

3º) Calcule a correlação entre as variáveis:

$$R_{x_s, x_r} = \frac{\frac{1}{(n-1)} \sum_{i=1}^n (x_{si} - \bar{x}_s)(x_{ri} - \bar{x}_r)}{S_{xs} S_{xr}} = 0.970$$

4º) prepare a matriz de correlação:

$$C = \begin{bmatrix} 1.000 & 0.970 \\ 0.970 & 1.000 \end{bmatrix}$$

5º) Calcule os autovalores da matriz:

$$|\lambda I - C| = \begin{bmatrix} \lambda - 1 & -0.970 \\ -0.970 & \lambda - 1 \end{bmatrix} = 0 \quad \text{ou} \quad (\lambda - 1)^2 - 0.970^2 = 0$$

Os autovalores são: $\lambda_1 = 1.970$, $\lambda_2 = 0.030$

Análise do Componente Principal

- 6º) Calcule os autovetores da matriz: $C q_1 = \lambda_1 q_1$

$$\text{ou} \quad \begin{bmatrix} 1.000 & 0.970 \\ 0.970 & 1.000 \end{bmatrix} \times \begin{bmatrix} q_{11} \\ q_{21} \end{bmatrix} = 1.970 \begin{bmatrix} q_{11} \\ q_{21} \end{bmatrix}$$

$$\text{ou} \quad q_{11} = q_{21} , \text{ assim } q_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\text{para } C q_2 = \lambda_2 q_2 , \quad q_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Análise do Componente Principal

7º) Calcular os fatores principais pela multiplicação dos autovetores pelos vetores normalizados:

$$\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} x'_s \\ x'_r \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1.359 \\ 1.717 \end{bmatrix} = \begin{bmatrix} 2.175 \\ -0.253 \end{bmatrix}$$

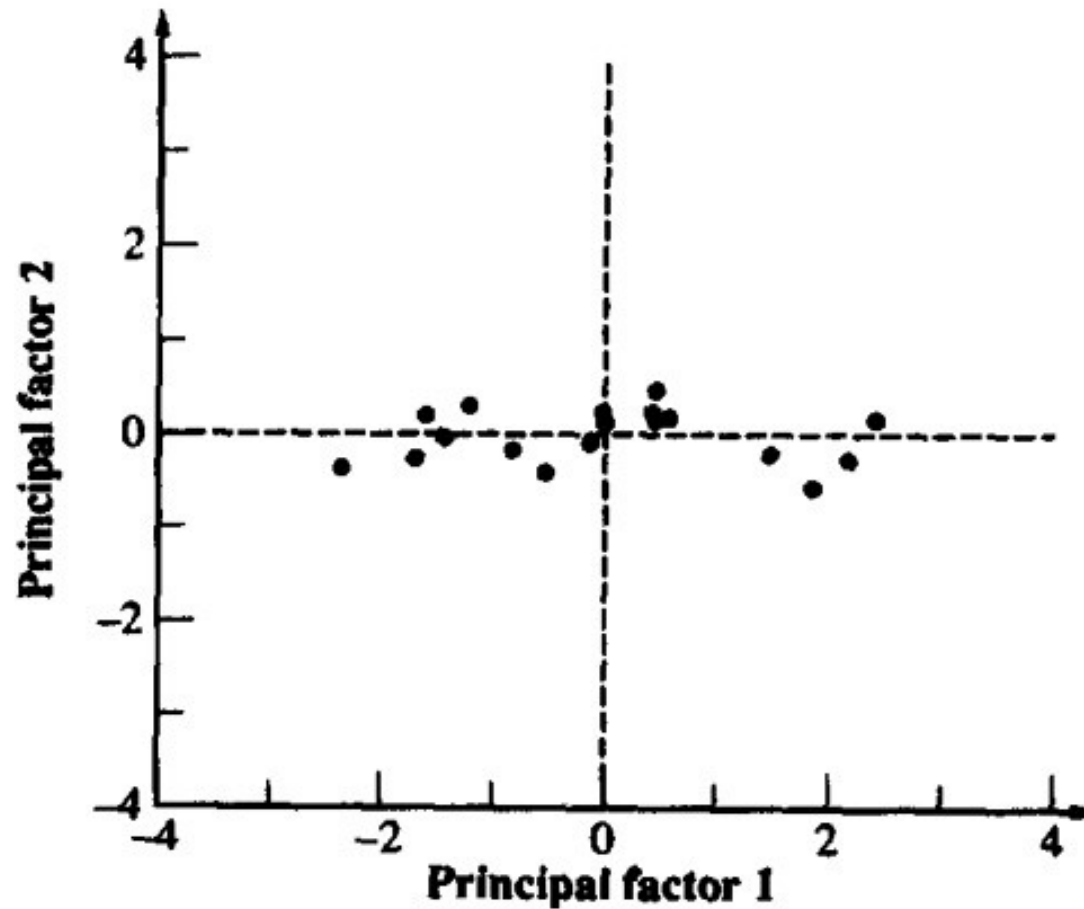
Esses valores são apresentados nas últimas duas colunas da tabela.

Análise do Componente Principal

8º) Calcular a soma e a soma dos quadrados dos fatores principais:

- A soma dos fatores deve ser zero.
- A soma dos quadrados dá o percentual da variação que deve ser explicado.
- No exemplo, a soma dos quadrados foram 32565 e 1435, o primeiro fator explica 95.7% e o segundo 4.3%, este último pode ser ignorado.

Análise do Componente Principal



Análise do Componente Principal

9º) Plotar os valores dos fatores principais:

- Note que a maioria da variação está no decorrer do eixo do primeiro fator.
- A variação do 2º fator é negligenciável.
- O 1º fator pode ser utilizado para classificar as estações em baixa, média e alta em termos de carga.
- Alternativamente, o par (y_1, y_2) também pode ser utilizado para classificar, as o ganho sobre utilizar apenas y_1 é muito pequeno.

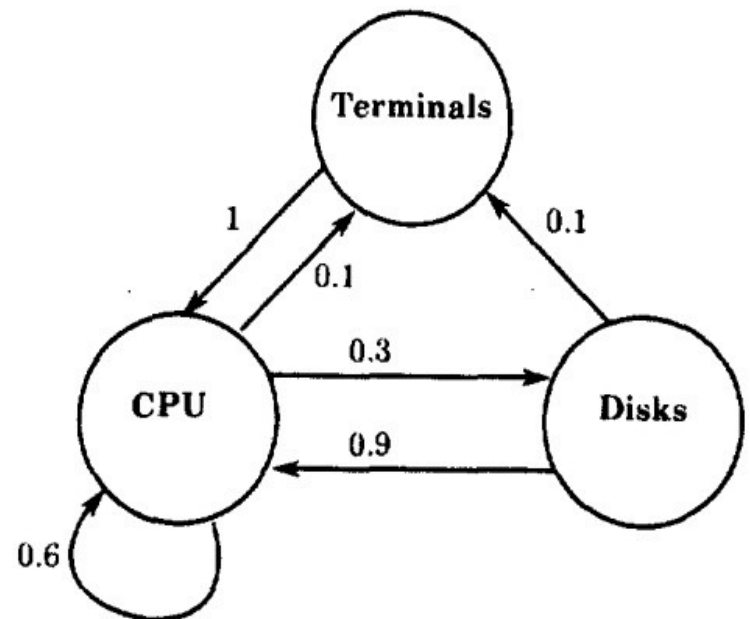
Modelos de Markov

- Além do # de solicitações de cada tipo, pode ser necessário a ordem em que acontecem.
- A próxima solicitação pode ser determinada pelas últimas solicitações.
- Nesse caso, as solicitações seguem um modelo de Markov.
- Esse modelo é mais utilizado para estados de sistemas do que para solicitações.
- Se o próximo estado do sistema depende apenas do estado corrente, o sistema segue um modelo de Markov.
- A matriz de transição também pode ser utilizada para transições entre aplicações.

Modelos de Markov

- Tais modelos podem ser descritos por uma matriz de transição.
- Essa matriz dá as probabilidades do próximo estado a partir do estado corrente.

From/To	CPU	Disk	Terminal
CPU	0.6	0.3	0.1
Disk	0.9	0	0.1
Terminal	1	0	0



Modelos de Markov

- Exemplo: o monitoramento de uma rede mostrou que a maioria dos pcts apresentam dois tamanhos (pequeno e grande). Os pcts pequenos representam 80% do tráfego. Um # diferente de matriz de transição de probabilidades resultariam em uma média global de 80% de pcts pequenos.

Modelos de Markov

Pacote	Pct. Pequeno	Pct. Grande
Pequeno	0.75	0.25
Grande	1	0

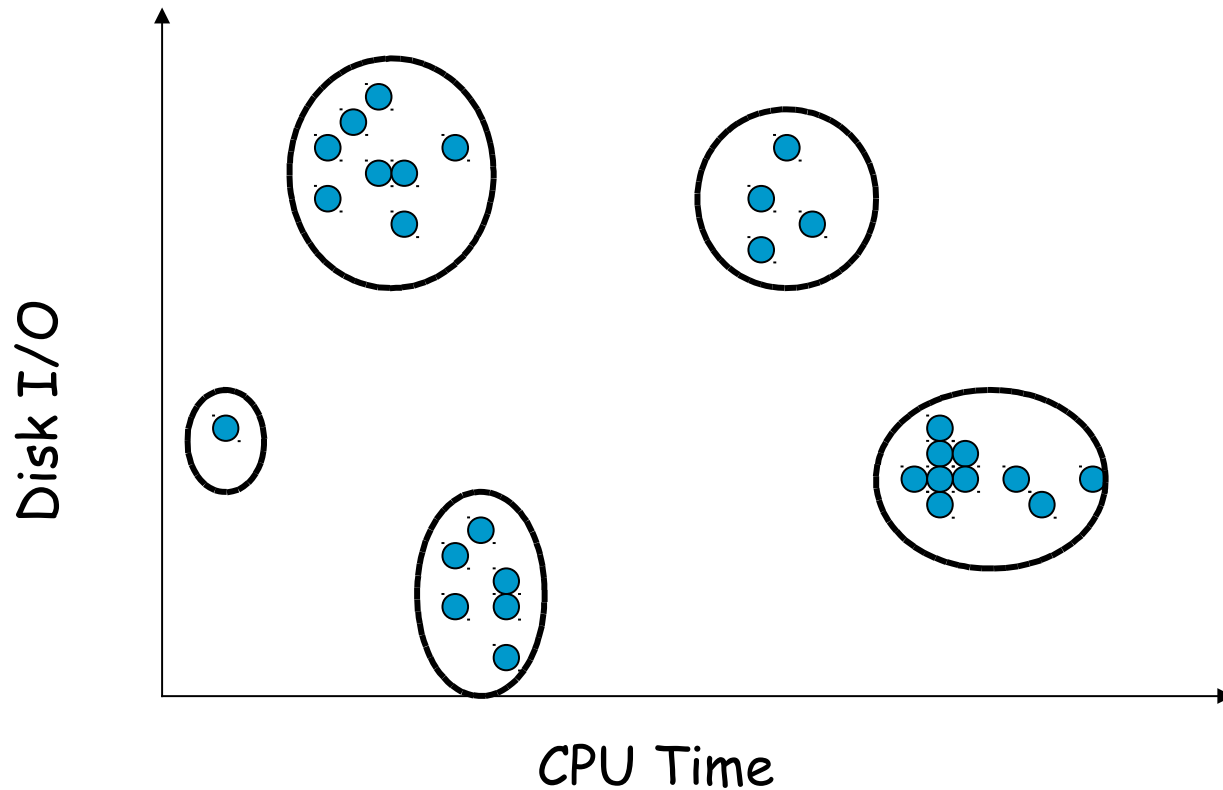
Pacote	Pct. Pequeno	Pct. Grande
Pequeno	0.80	0.20
Grande	0.80	0.20

Clusterização

- O workload medido consiste em um grande número de componentes.
 - Exemplo: diversos perfis de usuários podem ter sido medidos, é útil classificá-los em um # pequeno de classes ou clusters. Posteriormente, um membro de cada classe pode ser selecionado para representá-la e para estudar os efeitos de decisões de projeto do sistema.

Clusterização

- A figura a seguir mostra as demandas de CPU e I/O de discos de 30 requisições.



Clusterização

- Passos para caracterização do workload:
 - Amostragem (subconjunto dos componentes do workload).
 - Selecionar os parâmetros do workload (aplicação de PCA).
 - Transformar os parâmetros, caso necessário.
 - Remover outliers.
 - Escalar os dados das observações.
 - Selecionar uma métrica de distância.
 - Realizar a clusterização.
 - Interpretar os resultados.
 - Alterar os parâmetros, ou # de clusters e repetir os passos 3-7
 - Selecionar componentes representativos de cada cluster.

Clusterização: amostragem

- Geralmente, são muitos os dados medidos para serem usados na análise de clusterização.
- É necessário selecionar uma amostra.
 - Exemplo: milhares de sessões do usuários podem ter sido medidas, mas apenas algumas centenas serão utilizadas na análise de clusterização.
- Componentes não utilizados na clusterização apresentam comportamento similar aos utilizados na análise.
- O percentual de componentes não associados a algum cluster é uma medida da eficácia da amostragem.

Clusterização: amostragem

- Um método de amostragem é uma seleção aleatória.
- Esse resultado é um subconjunto representativo.
 - Exemplo: se a meta de estudo é examinar o impacto de um recurso em particular, por exemplo, de um disco, apenas os componentes que são clientes pesados fornecerão informação significativa.
- Além disso, os componentes mais frequentemente utilizados devem ser utilizados se a meta de estudo é identificar componentes para melhores interfaces e treinamento de uso.

Clusterização: seleção de parâmetros

- Alguns parâmetros são mais importantes porque pertencem ao recurso que é o gargalo ou ao recurso mais caro.
- Componentes menos importantes são omitidos para reduzir o custo da análise.
- Critério chave: impacto no desempenho e variância.
- Parâmetros que mudam pouco entre os clusters devem ser omitidos.

Clusterização: seleção de parâmetros

- Determinando o subconjunto mínimo de parâmetros:
 - refazer a clusterização com um parâmetro a menos e verificar o # de componentes que mudam seu relacionamento no cluster.
- Se a fração de tais componentes é pequena, os parâmetros podem ser removidos da lista.
- A **Análise do Componente Principal (PCA)** pode ser utilizada para identificar os parâmetros que tem a variância mais alta.

Clusterização: transformação

- Se a distribuição dos parâmetros é tendenciosa: trocar o parâmetro pela transformação ou pela função do parâmetro.
 - Exemplo: em um estudo, uma transformação log. do tempo de CPU foi utilizada porque o analista argumentou que 2 programas que gastam 1 e 2 s de tempo de CPU são quase tão diferentes como os que gastam 1 e 2 ms. Então, a razão do tempo de CPU mais do que a diferença foi considerada mais importante.
- Transformações são consideradas no capítulo 15 (JAIN).

Clusterização: outliers

- Outliers: dados com valores de parâmetros extremos, são valores muito mais altos do que a maioria.
- Cada outlier podem ter um efeito significativo nos valores máximos e mínimos dos parâmetros observados.
- Os valores são normalizados, sua inclusão ou exclusão afeta significativamente o resultado final da clusterização.
- Somente componentes periféricos que não consomem uma parcela significativa dos recursos devem ser removidos.
 - Exemplo: workload de I/Os no disco realizadas por programas e programas de backups.

Clusterização: escala de dados

- Os valores dos parâmetros devem ser escalados de modo que os valores relativos sejam aproximadamente iguais.
- Existem 4 técnicas:
 - Normalizar para média zero e variância: o valor escalado do k-ésimo parâmetro x_{ik} é dado por

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

- Pesos: um peso pode ser dado de acordo com a importância relativa parâmetro ou pode ser inversamente proporcional ao desvio padrão dos valores dos parâmetro

$$x'_{ik} = w_{ik} \times x_{ik}$$

Clusterização: escala de dados

- Faixa de normalização: a faixa de valores é alterada de $[x_{\min,k}, x_{\max,k}]$ para $[0, 1]$ com

$$x'_{ik} = \frac{x_{ik} - x_{\min,k}}{x_{\max,k} - x_{\min,k}}$$

- Percentil de normalização: os dados são escalados para 95% dos valores que caem entre 0 e 1

$$x'_{ik} = \frac{x_{ik} - x_{2.5,k}}{x_{97.5,k} - x_{2.5,k}}$$

$x_{2.5,k}$ e $x_{97.5,k}$ são os 2.5 e 97.5-percentis.

Clusterização: métrica de distância

- A clusterização consiste em mapear cada componente em um espaço n -dimensional e identificar a proximidade de um com o outro.
- Existem 3 métodos:
 - Distância Euclidiana: a distância de entre dois componentes $\{x_{i1}, x_{i2}, \dots, x_{in}\}$ e $\{x_{j1}, x_{j2}, \dots, x_{jn}\}$ é definida como

$$d = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$$

- Distância Euclidiana de pesos: um peso a_k é dado para cada parâmetro

$$d = \sum_{k=1}^n \sqrt{a_k (x_{ik} - x_{jk})^2}$$

Clusterização: métrica de distância

– Distância Chi-quadrado:

$$d = \sum_{k=1}^n \frac{(x_{ik} - x_{jk})^2}{x_{ik}}$$

– Observações:

- A Distância Euclidiana é a métrica mais utilizada.
- A Distância Euclidiana de Pesos é utilizada se os parâmetros não foram escalados ou tem níveis diferentes de importância.
- Chi-quadrado é utilizada no ajuste de distribuição, é importante que os valores estejam normalizados.

Clusterização: técnicas

- **Objetivo:** particionar os componentes em grupos:
 - A variância dentro do grupo deve ser pequena (intra).
 - A variância entre grupos deve ser tão alta quanto possível (inter).
- Redundante: alcançar qualquer uma é suficiente!

$$S_{total}^2 = S_{intragrupos}^2 + S_{intergrupos}^2$$

- Duas classes:
 - Não hierárquicas: inicia com um conj. de k cluster, e os membros são movidos até que a variância intragrupos seja mínima.
 - Hierárquicas: abordagem aglomerativa e divisiva.

Clusterização: método da árvore geradora mínima

- Técnica de clusterização hierárquica aglomerativa.
- Passos:
 - 1º) Inicia com $k = n$ clusters.
 - 2º) Encontra o centróide do i -ésimo cluster, que tem valor de parâmetro igual a média de todos os pontos no cluster.
 - 3º) Calcula a matriz de distância entre clusters. Seu (i, j) -ésimo elemento é a distância entre o centróide do cluster i e j .
 - 4º) Encontra o menor elemento não-zero na matriz de distância. Seja d_{lm} a menor distância, misture o cluster l e m , também misture outros clusters que tenham a mesma distância.
 - 5º) Repita os passos 2 a 4 até que todos os componentes sejam parte de um só cluster.

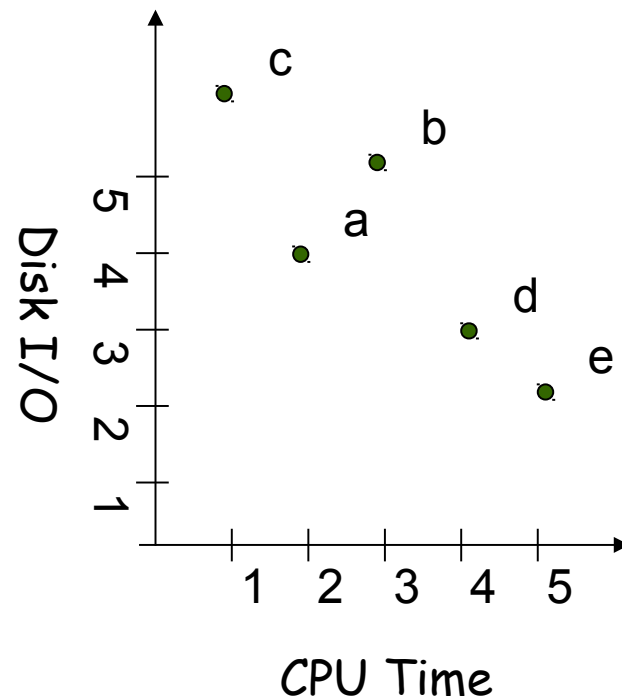
Clusterização: aplicação do método

- Exemplo: considere um workload de cinco componentes e dois parâmetros. O tempo de CPU e o # de I/Os do disco foram medidos para cinco programas. Os valores dos parâmetros depois da escala de dados são apresentados na seguinte tabela:

Program	CPU Time	Disk I/O
A	2	4
B	3	5
C	1	6
D	4	3
E	5	2

Clusterização: aplicação do método

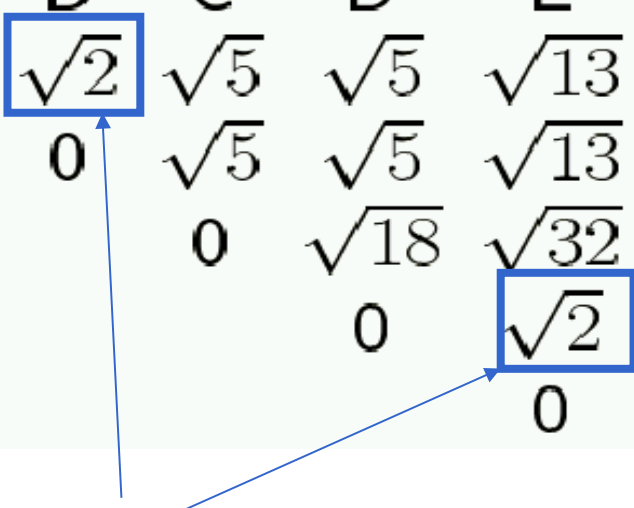
- **Passo 1:** considere 5 clusters com o i -ésimo cluster consistindo somente do i -ésimo programa.
- **Passo 2:** os centróides são $\{2, 4\}$, $\{3, 5\}$, $\{1, 6\}$, $\{4, 3\}$ e $\{5, 2\}$ que são apresentados na figura:



Clusterização: aplicação do método

- **Passo 3:** utilize distância Euclidiana para construir a matriz de distância

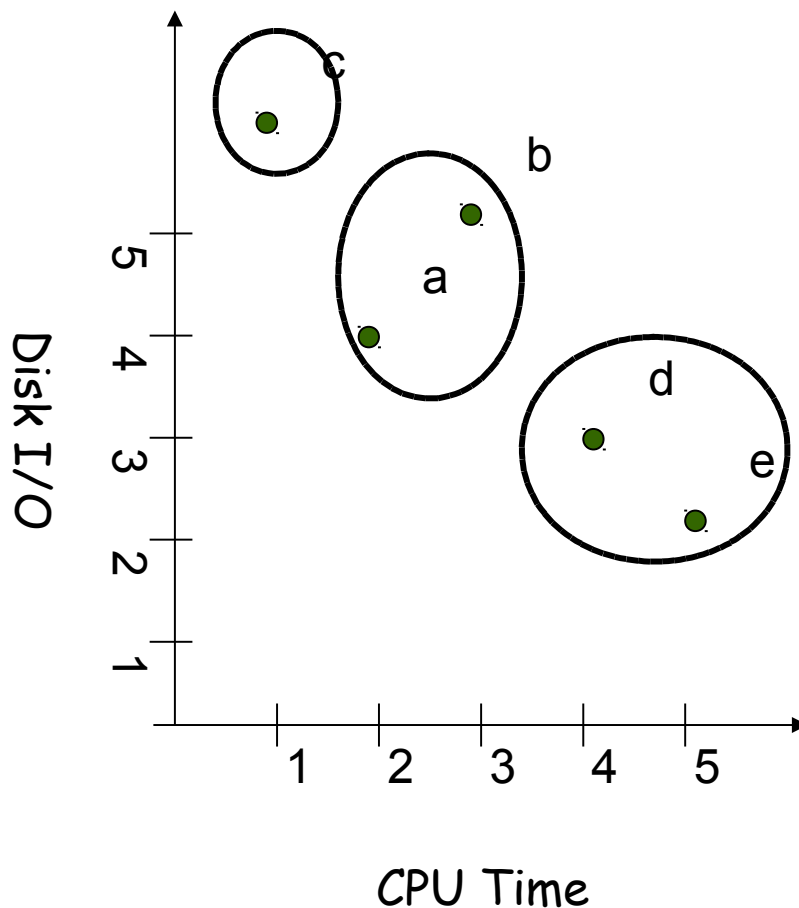
Program	Program				
	A	B	C	D	E
A	0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
B		0	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
C			0	$\sqrt{18}$	$\sqrt{32}$
D				0	$\sqrt{2}$
E					0



- **Passo 4:** juntar os pares com a menor distância intercluster.

Clusterização: aplicação do método

- **Passo 2:** o centróides do cluster do par AB é $\{ (2+3)/2, (4+5)/2 \} = \{2.5, 4.5\}$ e do par DE é $\{4.5, 2.5\}$.
- **Passo 3:** existem 3 clusters, a nova matriz de distância é:



Program	Program		
	AB	C	DE
AB	0	$\sqrt{4.5}$	$\sqrt{10.25}$
C		0	$\sqrt{24.4}$
DE			0

Menor distância

Clusterização: aplicação do método

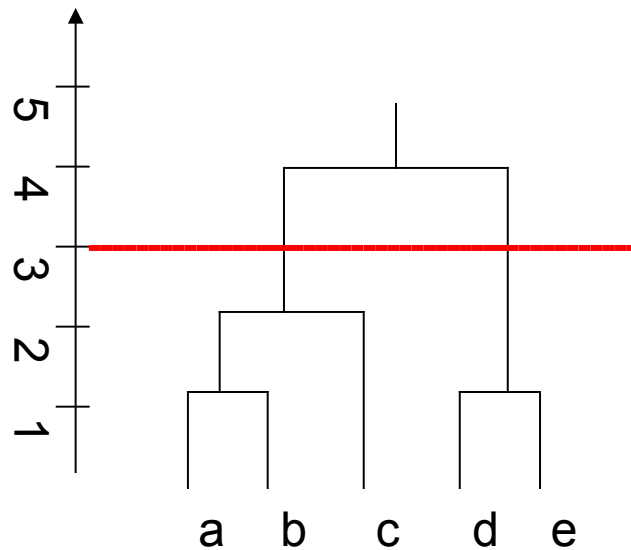
- **Passo 4:** a menor distância intercluster é entre AB e C, juntar estes dois.
- **Passo 2:** o centróide ABC é $\{(2+3+1)/3, (4+5+6)/3\} = \{2, 5\}$.
- **Passo 3:** a matriz de distância é:

	Program	
Program	ABC	DE
ABC	0	$\sqrt{12.5}$
DE		0

- **Passo 4:** a menor distância é $\text{sqrt}(12.5)$, assim ABC e DE resultam em um único cluster ABCDE.

Clusterização: aplicação do método

- Árvore geradora é chamada de dendrograma
 - Cada ramificação é um cluster, a altura é onde os clusters se misturam



Podem ser obtidos cluster de uma determinada altura. e.g.: na altura 3, são obtidos os clusters **ABC** e **DE**.

Clusterização: como definir o # de clusters?

- Deve-se examinar a variação de duas métricas:
 - Distância intraclusters: distância média entre pontos de um cluster e seu centróide.
 - Distância interclusters: distância média entre centróides.
- A variação é caracterizada pelo CV.
- Minimizar CV intracluster enquanto maximiza o CV intercluster.

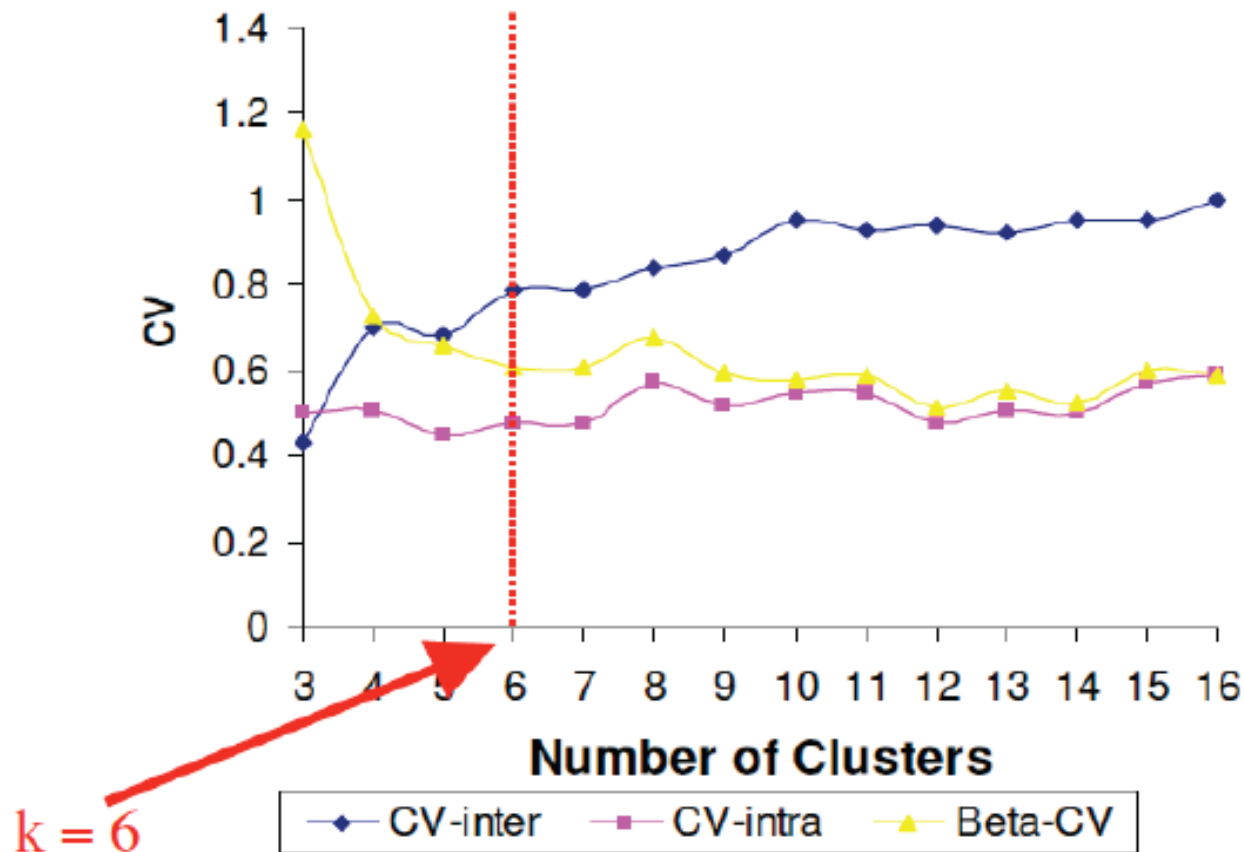
Clusterização: como definir o # de clusters?

- Defina:

$$\beta_{CV} = \frac{CV \text{ intra}}{CV \text{ inter}}$$

- Gere um gráfico de β_{CV} *versus* # de clusters k
- Escolha k depois que β_{CV} estabilizar.

Clusterização: como definir o # de clusters?



- Fonte: slides Virgílio Almeida DCC/UFMG.

Clusterização: interpretação

- Clusters com pouca população devem ser descartados:
 - se a demanda de recurso total tem impacto insignificante no desempenho.
- Um cluster que tem apenas 1 componente mas usa 50% do recursos do sistema não pode ser descartado!
- Interpretar o cluster em termos funcionais:
 - Se a maioria dos componentes de um cluster pertence a um único ambiente de aplicação, isso é útil para nomear o cluster adequadamente.
 - Uma vez que os componentes de um cluster tem demandas de recursos similares isso ajuda a rotular o cluster (e.g. CPU bound, I/O bound).
- Um ou mais componentes de cada cluster são selecionados para workload de testes em um estudo de desempenho.
 - Um # de representantes podem ser proporcionais ao tamanho do cluster, a demanda de recurso total do cluster, ou uma combinação dos dois.

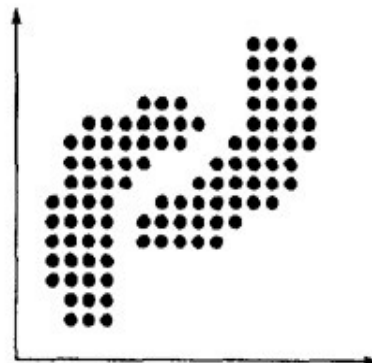
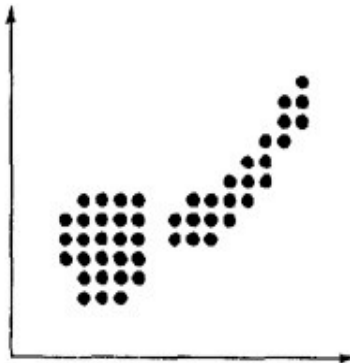
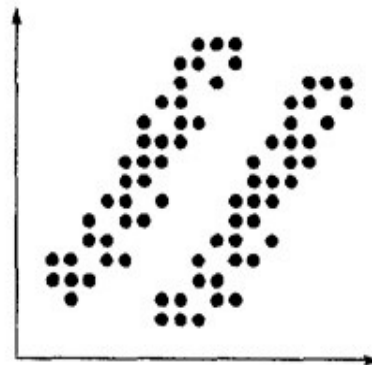
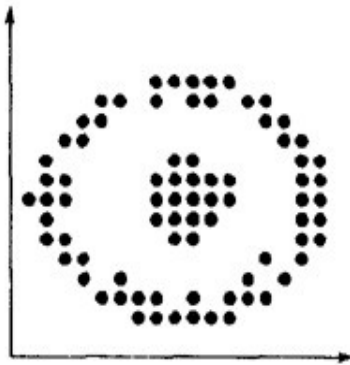
Clusterização: interpretação

- Clusters com pouca população devem ser descartados:
 - se a demanda de recurso total tem impacto insignifi
- Um cluster que tem apenas 1 componente mas u sistema não pode ser descartado!
- Interpretar o cluster em termos funcionais.
 - Se a maioria dos componentes de um cluster pertence a um único ambiente de aplicação, isso é útil para nomear o cluster adequadamente.
 - Uma vez que os componentes de um cluster tem demandas de recursos similares isso ajuda a rotular o cluster (e.g. CPU bound, I/O bound).
- Um ou mais componentes de cada cluster são selecionados para workload de testes em um estudo de desempenho.
 - Um # de representantes podem ser proporcionais ao tamanho do cluster, a demanda de recurso total do cluster, ou uma combinação dos dois.

O centróide do cluster pode ser utilizado para representá-lo!!!

Clusterização: problemas

- **Primeiro problema:** relacionado a definição de objetivos com o próprio cluster.
 - meta: minimizar variação intra ou maximizar variação intercluster.



Clusterização: problemas

- Resultados altamente variáveis:
 - não existem regras para selecionar parâmetros, métricas de distância, ou escala.
 - uma escolha diferente pode levar a um conjunto de clusters diferente e à diferentes conclusões.
- Dificuldade: rotular clusters pela funcionalidade.
 - Exemplo: em um estudo em que programas de edição aparecem em 23 diferentes clusters, pode ser mais significativo para caracterizar a média desses programas do que caracterizar 23 clusters que não tem significado funcional.
- Clusterização não ajuda quando a meta é comparar o workload de diferentes sítios.

Bibliografia

- The Art of Computer Systems Performance Analysis:
 - Capítulo 6