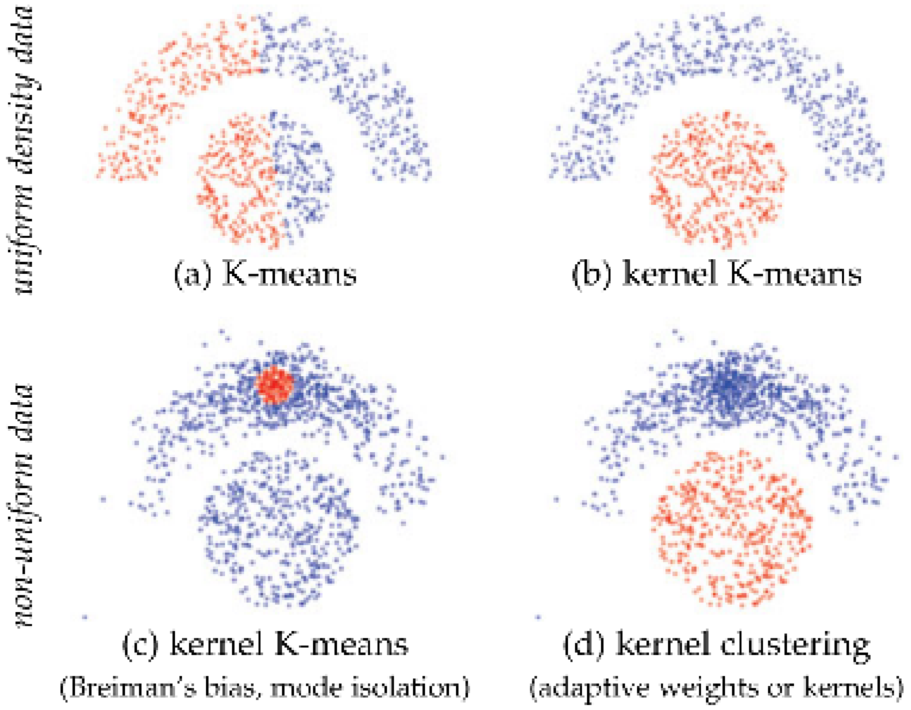
Kernel Clustering: Density Biases and Solutions

张培

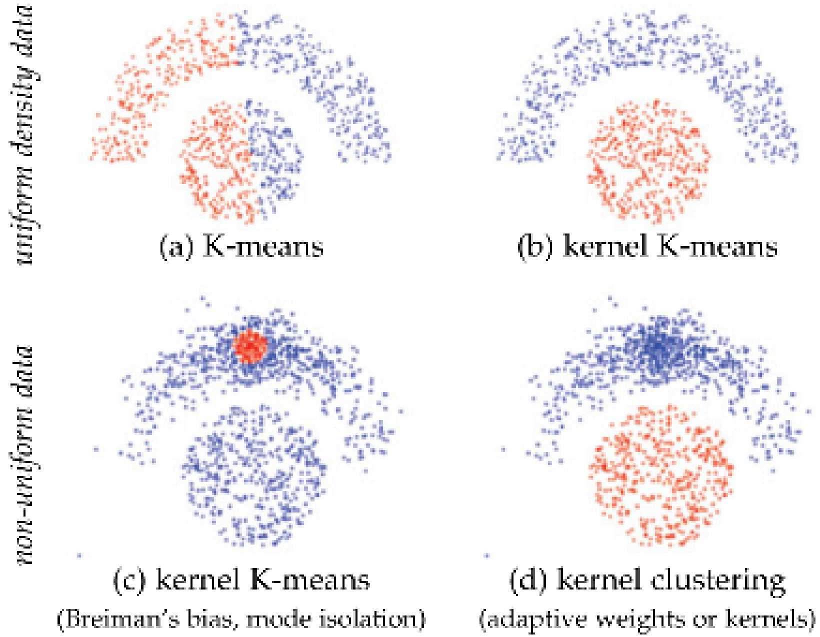
2019.11.9

Catalogue

- Kernel K-means
- Notions related to Kernel K-means
- Dicrete Gini Criterion
- Kernel K-means and Continuous Gini Criterion
- Method
  - Adaptive weight
  - Adaptive kernel
- Two Biases
  - Breiman' bias
  - Bias to "sparsest " subset in Normalized Cut



K-means & Kernel K-means



(k-means criterion)

$$\sum_k \sum_{p\in S^k} \|f_p - m_k\|^2. \qquad (2)$$

$$F(S,m) = \sum_k \sum_{p\in S^k} \|\phi_p - m_k\|^2, \qquad (3)$$

(kernel k-means criterion)

$$F(S) \overset{c}{=} -\sum_k \frac{\sum_{pq\in S^k} k(f_p, f_q)}{|S^k|}. \qquad (8)$$

Related to Graph Clustering Criteria

Kernel K-means Criterion :

$$F(S) \stackrel{c}{=} -\sum_k \frac{\sum_{pq \in S^k} k(f_p, f_q)}{|S^k|}. \quad (8)$$

Average association Criterion :

$$-\sum_k \frac{\sum_{pq \in S^k} A_{pq}}{|S^k|}. \quad (9)$$

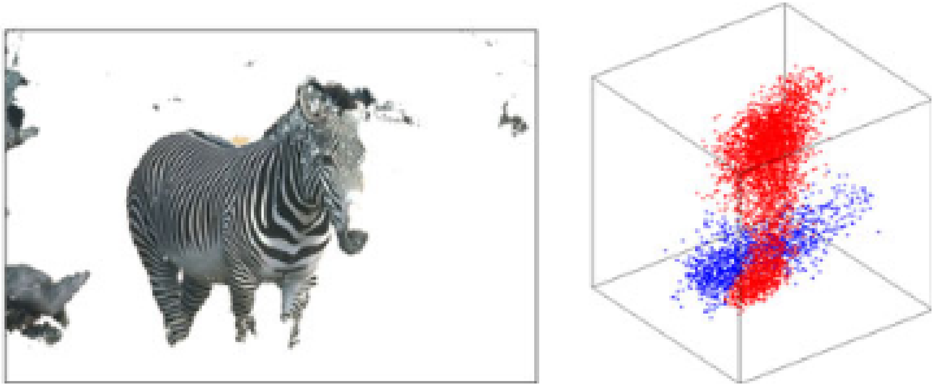
[15] showed that dropping p.s.d. assumption is not essential:
for arbitrary association A there is a p.s.d. kernel k such that objective (8) is equivalent to (9) up to a constant.

[3] authors experimentally observed that :
the average association (9) or kernel K-means (8) objectives have a bias to separate small dense group of data points from the rest.

[15] V. Roth, J. Laub, M. Kawanabe, and J. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," IEEE Trans.Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1540—1551, Dec. 2003.

Example

(a) Breiman's bias



(b) good clustering

(a) shows the result for kernel K-means with a fixed-width Gaussian kernel isolating a small dense group of pixels from the rest.

(b) shows the result for an adaptive kernel

Relation between kernel clustering and probability density estimation

Kernel K-means Criterion :

$$F(S) \stackrel{c}{=} -\sum_k \frac{\sum_{pq\in S^k} k(f_p, f_q)}{|S^k|}. \quad (8)$$

Kernel Density Estimate / Parzen density estimate :

$$\mathcal{P}_\Sigma(x|S^k) := \frac{\sum_{q\in S^k} k(x, f_q)}{|S^k|}, \quad (11)$$

If kernel k has form (12) up to a positive multiplicative constant then kernel K-means objective (8) can be expressed (15) in terms of kernel densities (11) for points in each cluster.

$$k(x,y) = |\Sigma|^{-\frac{1}{2}}\,\psi\left(\Sigma^{-\frac{1}{2}}(x-y)\right), \quad (12)$$

$$F(S) \stackrel{c}{=} -\sum_k \sum_{p\in S^k} \mathcal{P}_\Sigma(f_p|S^k). \quad (15)$$

For shortness, we use adjective r-small to describe bandwidths providing accurate density estimation.

$$\sqrt{\Sigma_{ii}} = \frac{r_i}{N\sqrt[4]{n}}, \qquad \Sigma_{ij} = 0 \text{ for } i \neq j, \quad (14)$$

Discrete Gini criterion

Probabilistic K-Means:

$$-\sum_{k}\sum_{p\in S^k}\log P(f_p|\theta_k). \qquad (16)$$

in case of highly descriptive model P, e.g., GMM or histograms

Entropy criterion:

$$\sum_{k}|S^k|\cdot H(S^k), \qquad (17)$$

(18) has a form similar to the entropy criterion in (17), except that entropy H is replaced by the Gini impurity.

Discrete Gini criterion:

$$\sum_{k}|S^k|\cdot G(S^k), \qquad (18)$$

$$\begin{aligned} Gini(p) &= \sum_{k=1}^{K} p_k(1-p_k) \\ &= \sum_{k=1}^{K}(p_k-p_k^2) \\ &= 1-\sum_{k=1}^{K}p_k^2 \end{aligned}$$

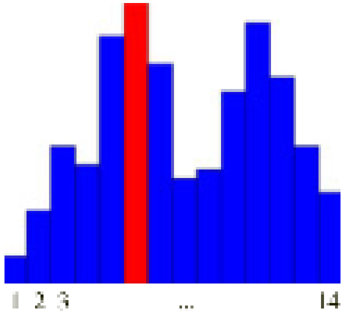
$$G(S^k) := 1-\sum_{l\in\mathcal{L}}\mathcal{P}(l\,|\,S^k)^2, \qquad (19)$$

Theoretical Properties of the Discrete Gini Criterion

" Technical Note: Some Properties of Splitting Criteria "

The gini prefers splits that put the largest class into one pure node, and all others into the other.
Entropy put its emphasis on balancing the sizes at the two children nodes.

Theorem 1 (Breiman). *For $K = 2$ the minimum of the Gini criterion (18) for discrete Gini impurity (19) is achieved by assigning all data points with the highest-probability feature value in $\mathcal{L}$ to one cluster and the remaining data points to the other cluster, as in example for $\mathcal{L} = \{1, \ldots, 14\}$ on the left.*



Continuous Gini Criterion

Discrete Gini criterion:

$$\sum_k |S^k| \cdot G(S^k), \qquad (18)$$

$$G(S^k) := 1 - \sum_{l \in \mathcal{L}} \mathcal{P}(l \,|\, S^k)^2, \qquad (19)$$

Continuous Gini Criterion:

$$\sum_k w_k \cdot G(s, k), \qquad (20)$$

$$G(s, k) := 1 - \int \rho_k^s(x)^2 \, \mathrm{d}x. \qquad (21)$$

continuous probability density function: $\rho_k^s(x) := \rho(x \,|\, s(x) = k)$

Kernel K-Means and Continuous Gini Criterion

连续型随机变量的期望: $E(X)=\int_{-\infty}^{+\infty} x \underbrace{f_X}_{\text{X的概率密度函数}}(x)\,dx$

无意识统计学家法则(LOTUS):

已知随机变量X的概率分布，但不知g(X)的分布，使用LOTUS可以计算出g(X)的期望:

$$E[g(X)]=\int_{-\infty}^{+\infty} g(x)\underbrace{f_X}_{\text{X的概率密度函数(PDF)}}(x)\,dx$$

就是可以用已知的X的PDF代替未知的g(X)的PDF.

Kernel K-Means and Continuous Gini Criterion

Monte-Carlo

代取独立同分布随机变量 $\{X_i\}$，在 $[a,b]$ 上服从分布律 f_X (PDF)，

令 $g^*(x) = \dfrac{g(x)}{f_X(x)}$，则 $g^*(X_i)$ 也是一组独立同分布的随机变量，且由LOTUS：

$$E[g^*(X_i)] = \int_a^b g^*(x) f_X(x)\,dx = \int_a^b \frac{g(x)}{f_X(x)} f_X(x)\,dx = \underline{\int_a^b g(x)\,dx}$$

$$\hat{E}[g^*(X_i)] = \frac{1}{N}\sum_{i=1}^{N} g^*(X_i)$$

MC方法之平均值法，就是用 $\hat{E}[g^*(X_i)]$ 作为 $E[g^*(X_i)]$ 的近似值。

由MC：

$$\frac{\sum_{i=1}^{N} g^*(X_i)}{N} \approx \int_a^b g^*(x) f_X(x)\,dx$$

则

$$\frac{\sum_{p\in S^k} P_\Sigma(f_p|s^k)}{|S^k|} \approx \int P_\Sigma(x|s^k)\,\rho_k^s(x)\,dx$$

Kernel K-Means and Continuous Gini Criterion

由MC: $\dfrac{\sum_{i=1}^{N} g^*(X_i)}{N} \approx \int_a^b g^*(x) f_X(x)\,dx$

则 $\dfrac{\sum_{p\in S^k} P_\Sigma(f_p|S^k)}{|S^k|} \approx \int P_\Sigma(x|S^k)\,\rho_k^s(x)\,dx$

$\therefore \sum_{p\in S^k} P_\Sigma(f_p|S^k) \approx |S^k| \int P_\Sigma(x|S^k)\,\rho_k^s(x)\,dx$

由核密度估计表示的 kernel k-mean:

$$F(S) \triangleq -\sum_k \sum_{p\in S^k} P_\Sigma(f_p|S^k) \approx -\sum_k |S^k| \int P_\Sigma(x|S^k)\,\rho_k^s(x)\,dx \overset{c}{\approx} -\sum_k |S^k| \cdot \int \rho_k^s(x)^2\,dx$$

假设 $P_\Sigma(\cdot|S^k) \approx \rho_k^S(\cdot)$

即对核带宽的假设, r-small

由连续Gini准则:

$\sum_k w_k \cdot G(s,k)$

$G(s,k) = 1 - \int \rho_k^s(x)^2\,dx$

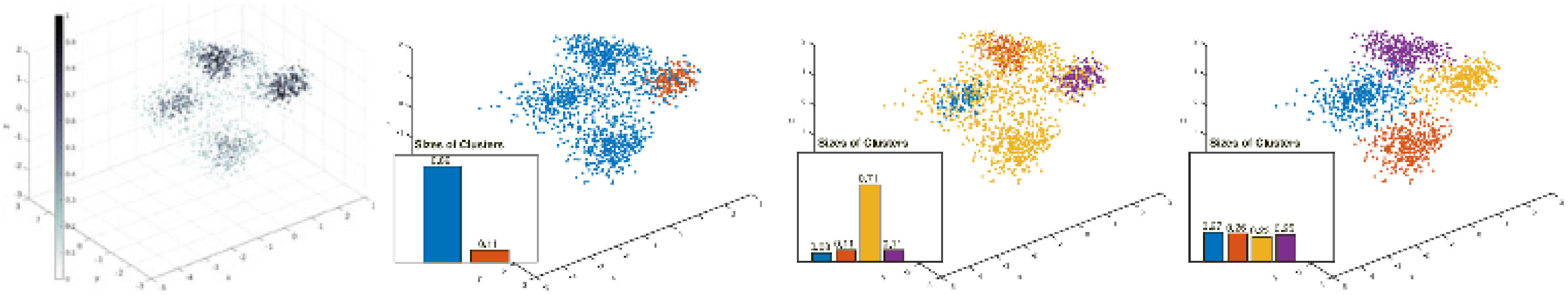
$\therefore \int \rho_k^s(x)^2\,dx = 1 - G(s,k)$

$$\therefore F(S) \overset{c}{\approx} -\sum_k |S^k| \int \rho_k^s(x)^2\,dx = -\sum_k |S^k|\,(1 - G(s,k)) = -\sum_k |S^k| + \sum_k |S^k| \cdot G(s,k) \overset{c}{=} \sum_k |S^k| \cdot G(s,k)$$

Breiman's Bias in Continuous Gini Criterion

Theorem 2 (Breiman's bias in continuous case). *For $K = 2$ the continuous Gini clustering criterion (20) achieves its optimal value at the partitioning of $\mathcal{R}^N$ into regions*

$$s_1 = \arg\max_x \rho(x) \quad \text{and} \quad s_2 = \mathcal{R}^N \setminus s_1.$$



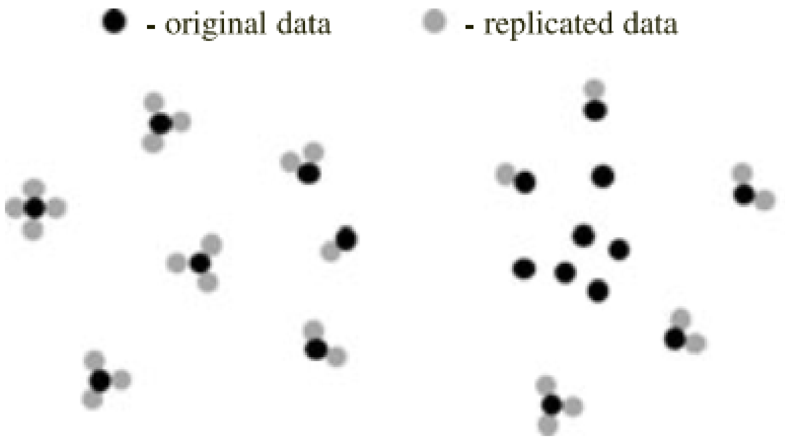
(a) density (b) Gaussian kernel, 2 clusters (c) Gaussian kernel, 4 clusters (d) KNN kernel, 4 clusters

Method1 : ADAPTIVE WEIGHTS

$$F_w(S, m) = \sum_k \sum_{p \in S^k} w_p \|\phi_p - m_k\|^2. \tag{41}$$

$$-\sum_k \frac{\sum_{pq \in S_k} w_p w_q A_{pq}}{\sum_{p \in S_k} w_p}. \tag{42}$$

$$\rho'_p \propto w_p \rho_p, \tag{43}$$



(a) adaptive weights (Sec. 3)

Method1 : ADAPTIVE KERNELS

$$\begin{pmatrix}\textbf{kernel}\\ \textbf{k-means}\\ \textbf{criterion}\end{pmatrix} \qquad F(S) \stackrel{c}{=} -\sum_k \frac{\sum_{pq\in S^k} k(f_p, f_q)}{|S^k|}. \qquad (8)$$

$$k_g(f_p, f_q) := \psi(d_g(f_p, f_q)) \equiv \psi(d_{pq}) \qquad (46)$$

Theorem 3. *Clustering (8) with (adaptive) geodesic kernel (46) is equivalent to clustering with fixed bandwidth kernel $k'(f'_p, f'_q) := \psi'(\|f'_p - f'_q\|)$ in new feature space $\mathcal{R}^{N'}$ for some radial basis function ψ' using the Euclidean distance and some constant N'.*

Proof. A powerful general result in [15], [35], [36] states that for any symmetric matrix (d_{pq}) with zeros on the diagonal there is a constant h such that squared distances

$$\tilde{d}^2_{pq} = d^2_{pq} + h^2[p \neq q], \qquad (50)$$

form *Euclidean matrix* $(\tilde{d}_{pq})$. That is, there exists some Euclidean embedding $\Omega \rightarrow \mathcal{R}^{N'}$ where for $\forall p \in \Omega$ there corresponds a point $f'_p \in \mathcal{R}^{N'}$ such that $\|f'_p - f'_q\| = \tilde{d}_{pq}$, see Fig. 6. Therefore,

$$\psi(d_{pq}) = \psi\left(\sqrt{\tilde{d}^2_{pq} - h^2\,[d_{pq} \geq h]}\right) \equiv \psi'(\tilde{d}_{pq}), \qquad (51)$$

for $\psi'(t) := \psi(\sqrt{t^2 - h^2[t \geq h]})$ and $k_g(f_p, f_q) = k'(f'_p, f'_q)$. $\square$

Method1 : ADAPTIVE KERNELS

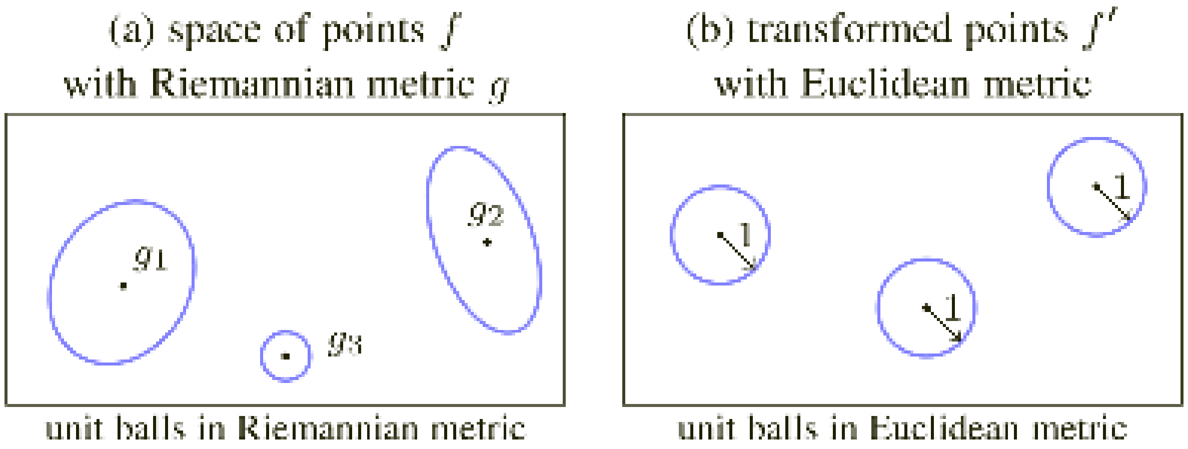


Fig. 6. Adaptive kernel (46) based on Riemannian distances (a) is equivalent to fixed bandwidth kernel after some *quasi-isometric* (50) embedding into Euclidean space (b), see Theorem 3, mapping ellipsoids (52) to balls (54) and modifying data density as in (57).

$$\rho_p \cdot |B_p| = |\Omega_p| = |\Omega'_p| = \rho'_p \cdot |B'_p|.$$

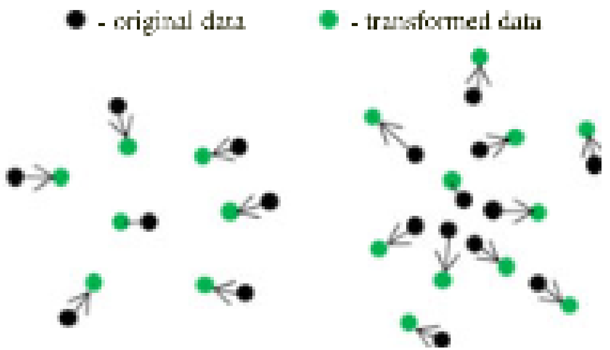
$$\rho'_p = \rho_p \frac{|B_p|}{|B'_p|} \propto \rho_p \, |\det g_p|^{-\frac{1}{2}},$$

$$\rho'_p \propto \rho_p \, \sigma_p^N .$$

$$\sigma_p \propto \sqrt[N]{\tau(\rho_p)/\rho_p}.$$

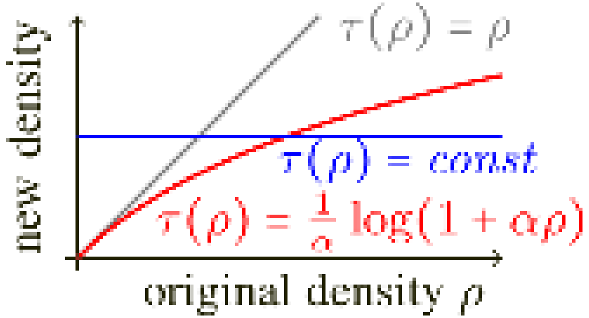
Method1: ADAPTIVE KERNELS

$$\sigma_p \propto \sqrt[N]{\tau(\rho_p)/\rho_p}. \qquad (59)$$



(b) adaptive kernels (Sec. 4.3)

density equalizing transforms:



estimating density by KNN approach:

$$\rho_p \approx \frac{K}{nV_K} \propto \frac{K}{n(R_p^K)^N}, \qquad (60)$$