1. SDE

<Learning a Kernel Matrix for Nonlinear Dimensionality Reduction> 2004
http://delivery.acm.org/10.1145/1020000/1015345/p85-weinberger.pdf?ip=218.94.142.142&id=1015345&acc=ACTIVE%20SERVICE&key=BF85BBA5741FDC6E%2E180A41DAF8736F97%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1570933790_8e1e25c6d1e42a5696e9c380d5a25130

2. MVU

<Colored Maximum Variance Unfolding>  NIPS 2008

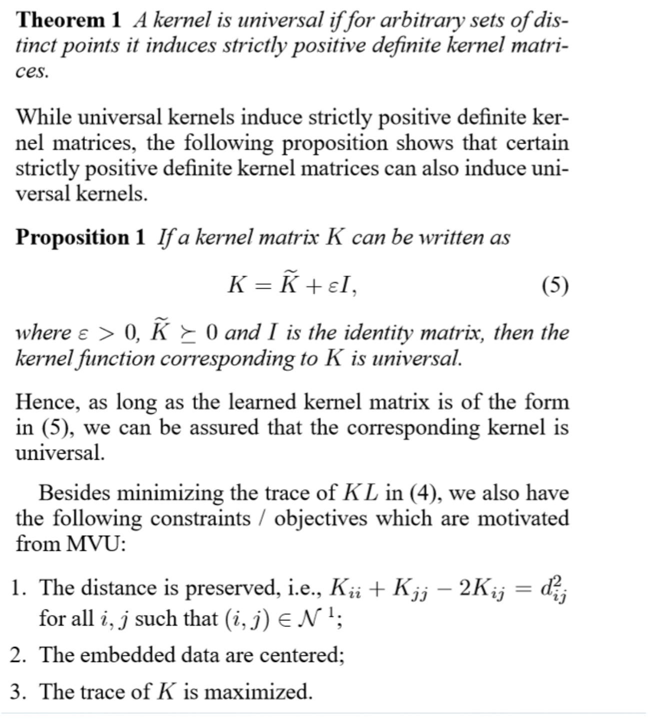
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0942CB4CF66CE16B557B4E72D2ABF165?doi=10.1.1.449.7685&rep=rep1&type=pdf

核心假设: maximum variance unfolding (MVU), has gained popularity as a method for dimensionality reduction. This method is based on a simple heuristic: maximizing the overall variance of the embedding while preserving the local distances between neighboring observations.

3. MMDE

（当矩阵 K 进行了中心化的时候，核函数矩阵对角线元素表示方差，映射后的方差 \phi 乘 phi）（这个 K 当作一个重构矩阵来理解，算法近邻，保距离的时候用到了原样本，用分解的方式求结果，这个思想有点像流形空间）

<Transfer Learning via Dimensionality Reduction> AAAI 2018
http://www.aaai.org/Papers/AAAI/2008/AAAI08-108.pdf



**Theorem 1** *A kernel is universal if for arbitrary sets of distinct points it induces strictly positive definite kernel matrices.*

While universal kernels induce strictly positive definite kernel matrices, the following proposition shows that certain strictly positive definite kernel matrices can also induce universal kernels.

**Proposition 1** *If a kernel matrix $K$ can be written as*

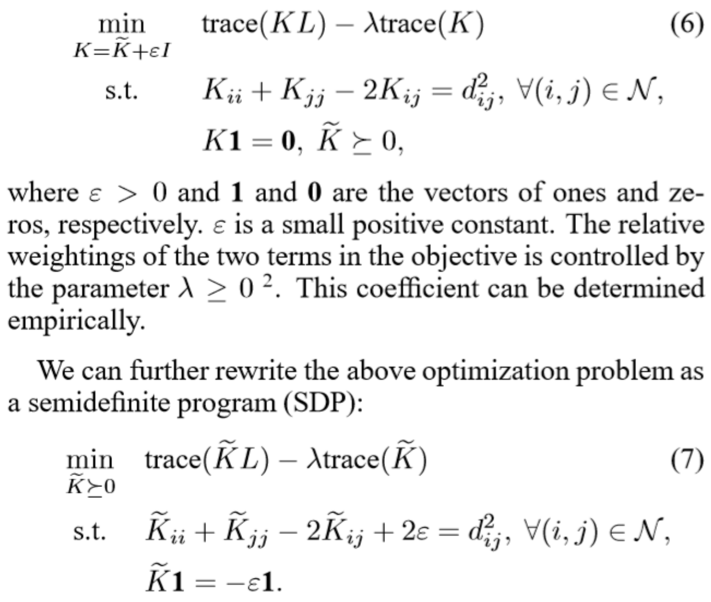
$$K = \widetilde{K} + \varepsilon I, \qquad (5)$$

*where $\varepsilon > 0$, $\widetilde{K} \succeq 0$ and $I$ is the identity matrix, then the kernel function corresponding to $K$ is universal.*

Hence, as long as the learned kernel matrix is of the form in (5), we can be assured that the corresponding kernel is universal.

Besides minimizing the trace of $KL$ in (4), we also have the following constraints / objectives which are motivated from MVU:

1. The distance is preserved, i.e., $K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2$ for all $i, j$ such that $(i, j) \in \mathcal{N}$ [1];
2. The embedded data are centered;
3. The trace of $K$ is maximized.

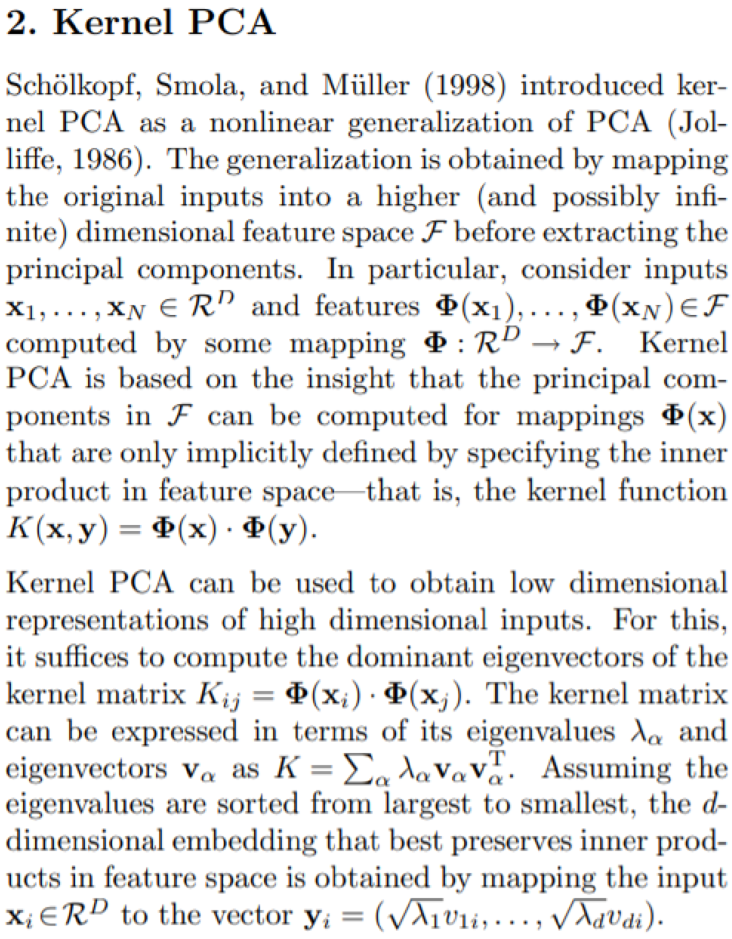


$$\begin{aligned} \min_{K=\widetilde{K}+\varepsilon I} \quad & \text{trace}(KL) - \lambda\text{trace}(K) \qquad (6) \\ \text{s.t.} \quad & K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2,\ \forall (i,j) \in \mathcal{N}, \\ & K\mathbf{1} = \mathbf{0},\ \widetilde{K} \succeq 0, \end{aligned}$$

where $\varepsilon > 0$ and $\mathbf{1}$ and $\mathbf{0}$ are the vectors of ones and zeros, respectively. $\varepsilon$ is a small positive constant. The relative weightings of the two terms in the objective is controlled by the parameter $\lambda \geq 0$ [2]. This coefficient can be determined empirically.

We can further rewrite the above optimization problem as a semidefinite program (SDP):

$$\begin{aligned} \min_{\widetilde{K}\succeq 0} \quad & \text{trace}(\widetilde{K}L) - \lambda\text{trace}(\widetilde{K}) \qquad (7) \\ \text{s.t.} \quad & \widetilde{K}_{ii} + \widetilde{K}_{jj} - 2\widetilde{K}_{ij} + 2\varepsilon = d_{ij}^2,\ \forall (i,j) \in \mathcal{N}, \\ & \widetilde{K}\mathbf{1} = -\varepsilon\mathbf{1}. \end{aligned}$$



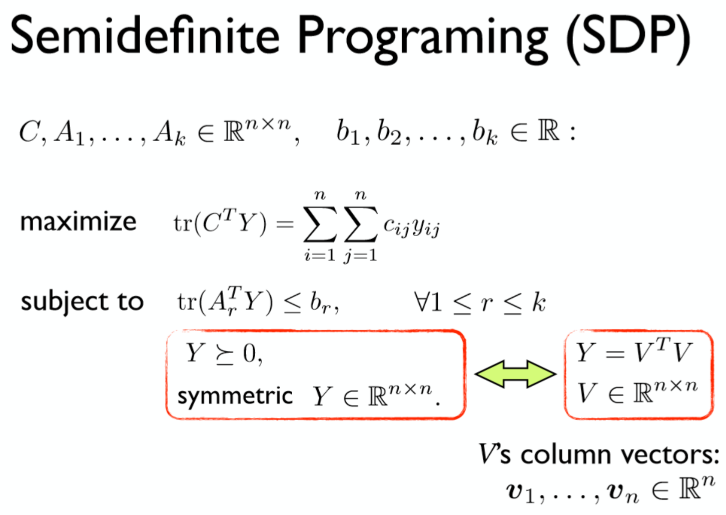
## 2. Kernel PCA

Schölkopf, Smola, and Müller (1998) introduced kernel PCA as a nonlinear generalization of PCA (Jolliffe, 1986). The generalization is obtained by mapping the original inputs into a higher (and possibly infinite) dimensional feature space $\mathcal{F}$ before extracting the principal components. In particular, consider inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{R}^D$ and features $\mathbf{\Phi}(\mathbf{x}_1), \ldots, \mathbf{\Phi}(\mathbf{x}_N) \in \mathcal{F}$ computed by some mapping $\mathbf{\Phi}: \mathcal{R}^D \to \mathcal{F}$. Kernel PCA is based on the insight that the principal components in $\mathcal{F}$ can be computed for mappings $\mathbf{\Phi}(\mathbf{x})$ that are only implicitly defined by specifying the inner product in feature space—that is, the kernel function $K(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{y})$.

Kernel PCA can be used to obtain low dimensional representations of high dimensional inputs. For this, it suffices to compute the dominant eigenvectors of the kernel matrix $K_{ij} = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j)$. The kernel matrix can be expressed in terms of its eigenvalues $\lambda_\alpha$ and eigenvectors $\mathbf{v}_\alpha$ as $K = \sum_\alpha \lambda_\alpha \mathbf{v}_\alpha \mathbf{v}_\alpha^{\mathrm{T}}$. Assuming the eigenvalues are sorted from largest to smallest, the $d$-dimensional embedding that best preserves inner products in feature space is obtained by mapping the input $\mathbf{x}_i \in \mathcal{R}^D$ to the vector $\mathbf{y}_i = (\sqrt{\lambda_1} v_{1i}, \ldots, \sqrt{\lambda_d} v_{di})$.

3-1 SDP 问题

http://tcs.nju.edu.cn/slides/aa2018/SDP.pdf



## Semidefinite Programing (SDP)

$C, A_1, \ldots, A_k \in \mathbb{R}^{n\times n}, \quad b_1, b_2, \ldots, b_k \in \mathbb{R}:$

maximize $\quad \mathrm{tr}(C^T Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} y_{ij}$

subject to $\quad \mathrm{tr}(A_r^T Y) \le b_r, \qquad \forall 1 \le r \le k$

$Y \succeq 0,$ symmetric $Y \in \mathbb{R}^{n\times n}.$ $\quad\Longleftrightarrow\quad$ $Y = V^T V$, $V \in \mathbb{R}^{n\times n}$

$V$'s column vectors: $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$

4. TCA

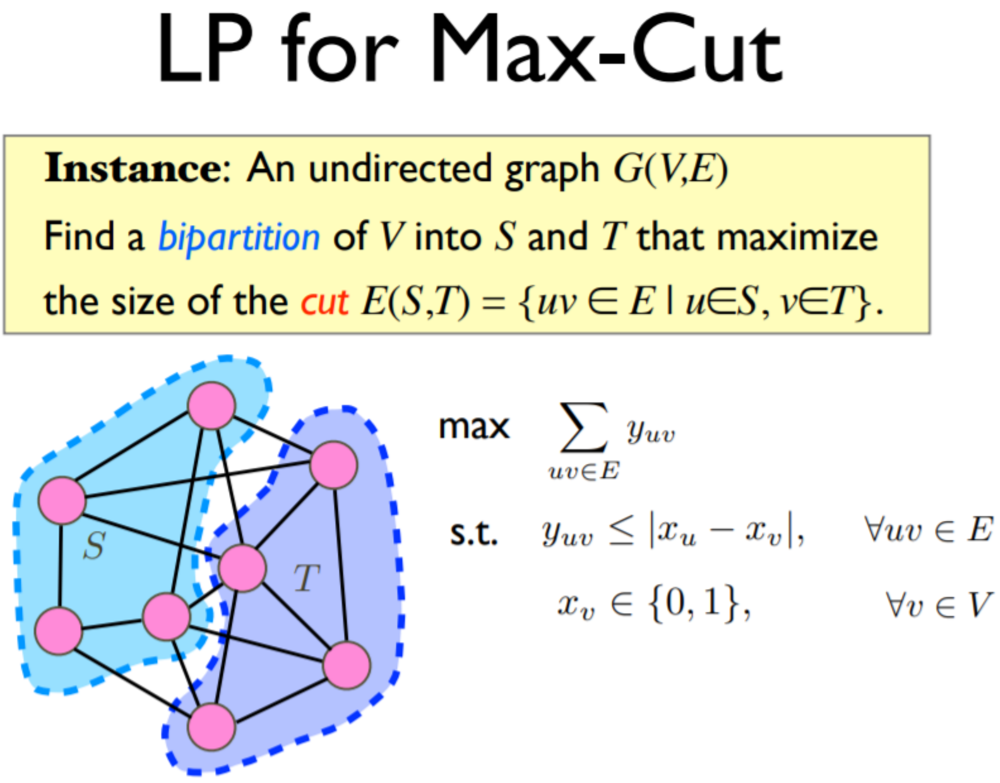
<Domain Adaptation via Transfer Component Analysis> IJCAI 2009 (期刊 2011)

1. 引入降维矩阵 W
2. 使得模型能够泛化
3. 在 MMDE 中 K 是变量，是求解出来的， 在 TCA 中 MMD -> tr(KL)这时候 K 以及\phi 还是变量，在引入降维矩阵 W 之后变成 tr(*W^TKLKW*),在这里实际上 K 也可以看作是变量，这里的变量有 W,K 两个，这里实际上是对问题进行了一个简化，引入了一个人为先验，假定是满足哪种映射（比如线性，高斯，..），这时的变量还有一个 W，也就是使得只是固定了映射类型，该类型定义了一个新的空间，W 是可以在这个空间中进行搜索的参数。实际上就是人为缩小了解空间，从而加速求解过程。

所以，在预先选定核函数实际上是引入人为先验进去。

4-1. 关于优化问题的理解（或 MMD 是否可以降低领域差异）

http://tcs.nju.edu.cn/slides/aa2018/SDP.pdf 定义优化问题，定义解空间，帮助求解（特例：当\phi(x) = x ，整个结果一定不会差于原始的）



# LP for Max-Cut

**Instance**: An undirected graph $G(V,E)$

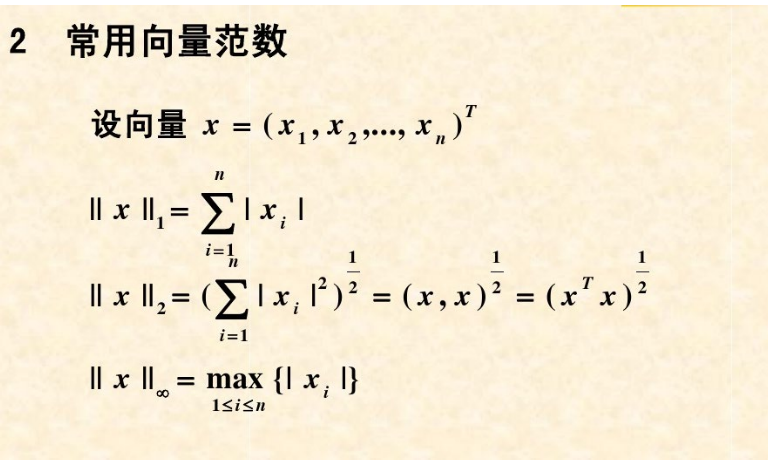
Find a *bipartition* of $V$ into $S$ and $T$ that maximize the size of the *cut* $E(S,T) = \{uv \in E \mid u \in S, v \in T\}$.

$$\begin{aligned} \max \quad & \sum_{uv \in E} y_{uv} \\ \text{s.t.} \quad & y_{uv} \le |x_u - x_v|, \qquad \forall uv \in E \\ & x_v \in \{0,1\}, \qquad \forall v \in V \end{aligned}$$

4-2 LDA 问题的求解

https://zhuanlan.zhihu.com/p/79696530

$$R(A,x) = \frac{x^H A x}{x^H x}$$



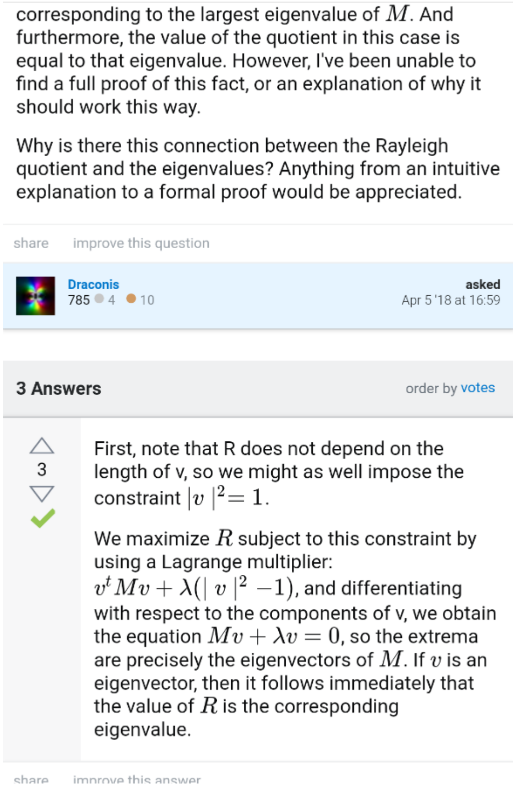
## 2 常用向量范数

设向量 $x = (x_1, x_2, \dots, x_n)^T$

$$\| x \|_1 = \sum_{i=1}^{n} | x_i |$$

$$\| x \|_2 = \left(\sum_{i=1}^{n} | x_i |^2\right)^{\frac{1}{2}} = (x, x)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}}$$

$$\| x \|_\infty = \max_{1 \le i \le n} \{| x_i |\}$$

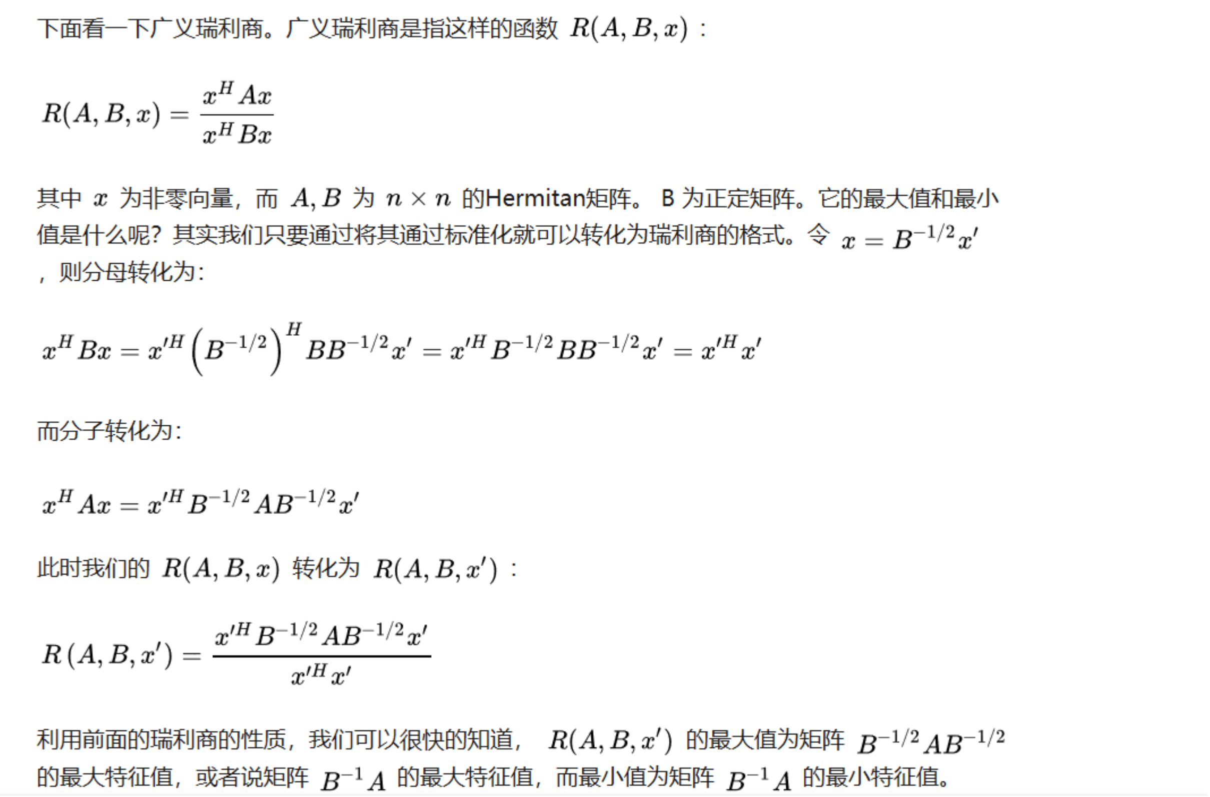


corresponding to the largest eigenvalue of $M$. And furthermore, the value of the quotient in this case is equal to that eigenvalue. However, I've been unable to find a full proof of this fact, or an explanation of why it should work this way.

Why is there this connection between the Rayleigh quotient and the eigenvalues? Anything from an intuitive explanation to a formal proof would be appreciated.

First, note that R does not depend on the length of v, so we might as well impose the constraint $|v|^2 = 1$.

We maximize $R$ subject to this constraint by using a Lagrange multiplier: $v^t M v + \lambda(|v|^2 - 1)$, and differentiating with respect to the components of v, we obtain the equation $Mv + \lambda v = 0$, so the extrema are precisely the eigenvectors of $M$. If $v$ is an eigenvector, then it follows immediately that the value of $R$ is the corresponding eigenvalue.



下面看一下广义瑞利商。广义瑞利商是指这样的函数 $R(A,B,x)$：

$$R(A,B,x)=\frac{x^HAx}{x^HBx}$$

其中 $x$ 为非零向量，而 $A,B$ 为 $n\times n$ 的Hermitan矩阵。 B 为正定矩阵。它的最大值和最小值是什么呢？其实我们只要通过将其通过标准化就可以转化为瑞利商的格式。令 $x=B^{-1/2}x'$ ，则分母转化为：

$$x^HBx=x'^H\left(B^{-1/2}\right)^HBB^{-1/2}x'=x'^HB^{-1/2}BB^{-1/2}x'=x'^Hx'$$

而分子转化为：

$$x^HAx=x'^HB^{-1/2}AB^{-1/2}x'$$

此时我们的 $R(A,B,x)$ 转化为 $R(A,B,x')$：

$$R(A,B,x')=\frac{x'^HB^{-1/2}AB^{-1/2}x'}{x'^Hx'}$$

利用前面的瑞利商的性质，我们可以很快的知道， $R(A,B,x')$ 的最大值为矩阵 $B^{-1/2}AB^{-1/2}$ 的最大特征值，或者说矩阵 $B^{-1}A$ 的最大特征值，而最小值为矩阵 $B^{-1}A$ 的最小特征值。

5. JDA
   <Transfer Feature Learning with Joint Distribution Adaptation>  ICCV 2013
   https://web.xidian.edu.cn/qlhuang/files/20150705_225137.pdf

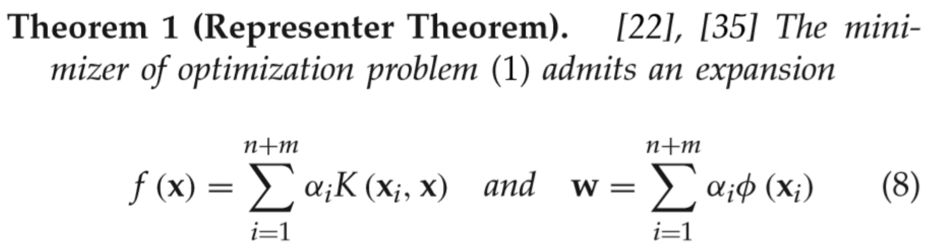
   第二十一讲 广义特征值与极小极大原理

6. BDA
   <Balanced Distribution Adaptation for Transfer Learning>  ICDM 2017

7. ARTL  TKDE 2014
   <Adaptation Regularization: A General Framework for Transfer Learning>

core idea：将特征表达的学习过程 or 分布差异降低的过程和分类器的学习统一在一起，分类器借助表示理论构建。

$$\begin{aligned} f &= \mathop{\arg\min}_{f\in\mathcal{H}_K}\sum_{i=1}^{n}\ell\left(f\left(\mathbf{x}_i\right),y_i\right)+\sigma\left\|f\right\|_K^2 \\ &\quad+\lambda D_{f,K}\left(J_s,J_t\right)+\gamma M_{f,K}\left(P_s,P_t\right), \end{aligned}$$



**Theorem 1 (Representer Theorem).** *[22], [35] The minimizer of optimization problem* (1) *admits an expansion*

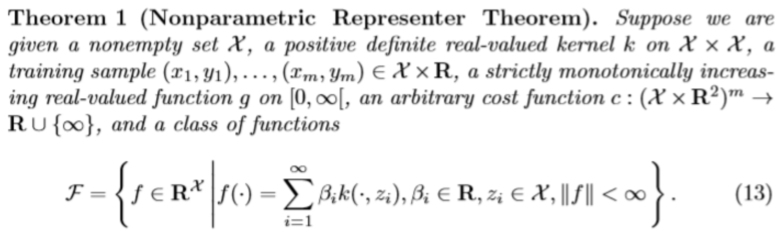
$$f(\mathbf{x}) = \sum_{i=1}^{n+m} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \quad \textit{and} \quad \mathbf{w} = \sum_{i=1}^{n+m} \alpha_i \phi(\mathbf{x}_i) \qquad (8)$$

7-0 希尔伯特空间及表示理论的理解

https://cosx.org/2014/05/svm-series-add-2-kernel-ii
http://sakigami-yang.me/2017/08/13/about-kernel-02/
https://zhuanlan.zhihu.com/p/54704957
https://zhuanlan.zhihu.com/p/77418542
https://zhuanlan.zhihu.com/p/29527729

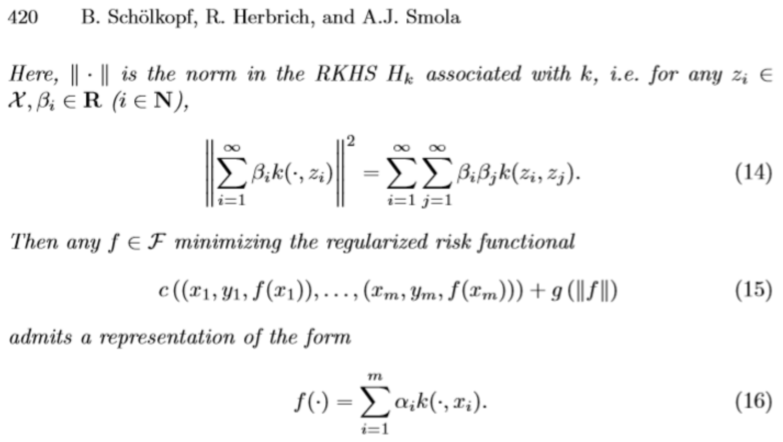
7-1 表示理论

<A Generalized Representer Theorem> COLT 2001
https://link.springer.com/content/pdf/10.1007%2F3-540-44581-1_27.pdf



**Theorem 1 (Nonparametric Representer Theorem).** *Suppose we are given a nonempty set $\mathcal{X}$, a positive definite real-valued kernel $k$ on $\mathcal{X} \times \mathcal{X}$, a training sample $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \mathbf{R}$, a strictly monotonically increasing real-valued function $g$ on $[0, \infty[$, an arbitrary cost function $c : (\mathcal{X} \times \mathbf{R}^2)^m \to \mathbf{R} \cup \{\infty\}$, and a class of functions*

$$\mathcal{F} = \left\{ f \in \mathbf{R}^{\mathcal{X}} \middle| f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \beta_i \in \mathbf{R}, z_i \in \mathcal{X}, \|f\| < \infty \right\}. \qquad (13)$$



420 B. Schölkopf, R. Herbrich, and A.J. Smola

*Here, $\|\cdot\|$ is the norm in the RKHS $H_k$ associated with $k$, i.e. for any $z_i \in \mathcal{X}, \beta_i \in \mathbf{R}$ $(i \in \mathbf{N})$,*

$$\left\| \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i) \right\|^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i \beta_j k(z_i, z_j). \qquad (14)$$

*Then any $f \in \mathcal{F}$ minimizing the regularized risk functional*

$$c\left((x_1, y_1, f(x_1)), \dots, (x_m, y_m, f(x_m))\right) + g\left(\|f\|\right) \qquad (15)$$

*admits a representation of the form*

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i). \qquad (16)$$

7-2 <Large Margin Transductive Transfer Learning> CIKM 2009 给出更加详细的推导 w 表示的过程

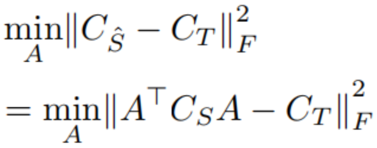
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=644189908FCFD3F226E41E0A30CCF732?doi=10.1.1.324.791&rep=rep1&type=pdf

7-3 Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples

http://delivery.acm.org/10.1145/1250000/1248632/7-2399-belkin.pdf?ip=218.94.142.142&id=1248632&acc=OPEN&key=BF85BBA5741FDC6E%2E180A41DAF8736F97%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1570891148_ead03bf76d5618ec5b439504792f9db6

8. CORAL

<Return of Frustratingly Easy Domain Adaptation>　AAAI 2015



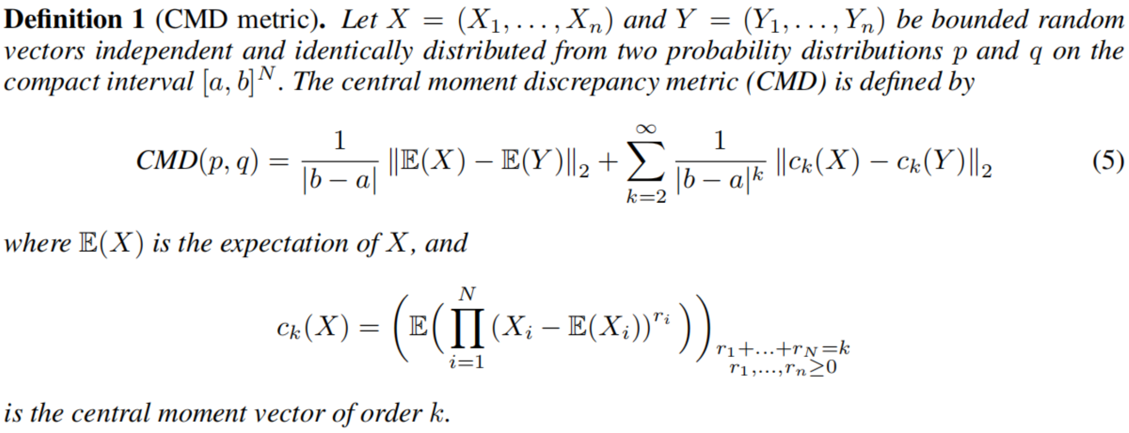
$$\begin{aligned}&\min_A \|C_{\hat{S}} - C_T\|_F^2\\ &= \min_A \|A^\top C_S A - C_T\|_F^2\end{aligned}$$

矩阵奇异值分解的理解

https://www.zhihu.com/question/22237507/answer/53804902
https://www.zhihu.com/question/19666954/answer/54788626

9. 任意阶 CMD

《CENTRAL MOMENT DISCREPANCY (CMD) FOR DOMAIN-INVARIANT REPRESENTATION LEARNING》　ICLR 2017



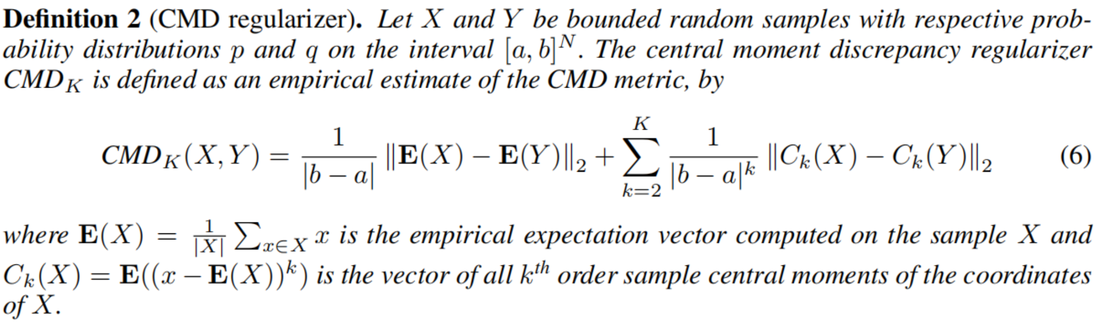
**Definition 1** (CMD metric). *Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be bounded random vectors independent and identically distributed from two probability distributions $p$ and $q$ on the compact interval $[a, b]^N$. The central moment discrepancy metric (CMD) is defined by*

$$CMD(p, q) = \frac{1}{|b-a|}\|\mathbb{E}(X) - \mathbb{E}(Y)\|_2 + \sum_{k=2}^{\infty}\frac{1}{|b-a|^k}\|c_k(X) - c_k(Y)\|_2 \tag{5}$$

*where $\mathbb{E}(X)$ is the expectation of $X$, and*

$$c_k(X) = \left(\mathbb{E}\left(\prod_{i=1}^{N}(X_i - \mathbb{E}(X_i))^{r_i}\right)\right)_{\substack{r_1+\ldots+r_N=k\\ r_1,\ldots,r_n\geq 0}}$$

*is the central moment vector of order $k$.*

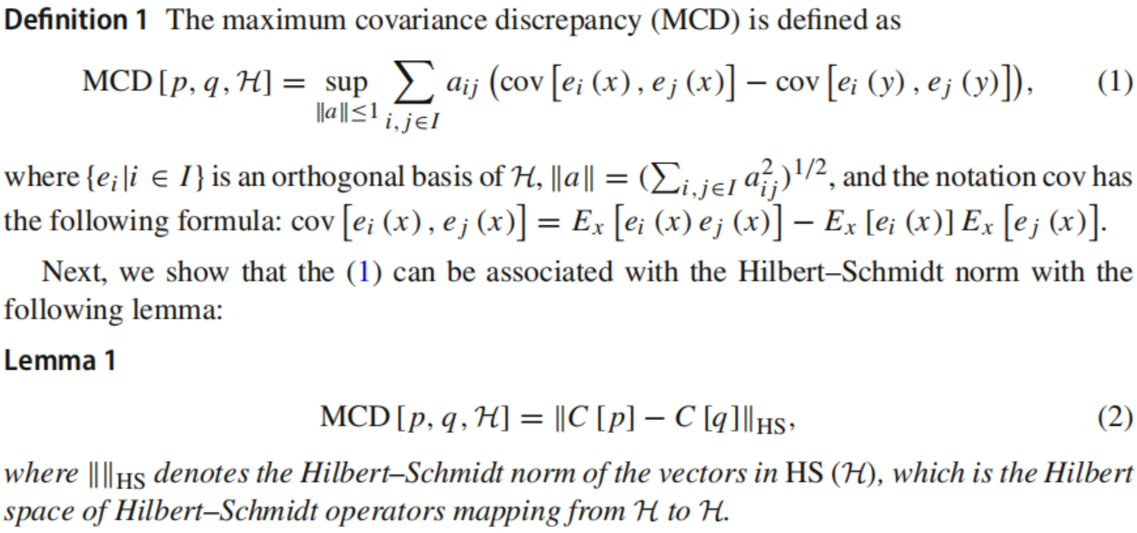


**Definition 2** (CMD regularizer)**.** *Let $X$ and $Y$ be bounded random samples with respective probability distributions $p$ and $q$ on the interval $[a,b]^N$. The central moment discrepancy regularizer $CMD_K$ is defined as an empirical estimate of the CMD metric, by*

$$CMD_K(X,Y) = \frac{1}{|b-a|}\|\mathbf{E}(X)-\mathbf{E}(Y)\|_2 + \sum_{k=2}^{K}\frac{1}{|b-a|^k}\|C_k(X)-C_k(Y)\|_2 \tag{6}$$

*where $\mathbf{E}(X)=\frac{1}{|X|}\sum_{x\in X}x$ is the empirical expectation vector computed on the sample $X$ and $C_k(X)=\mathbf{E}((x-\mathbf{E}(X))^k)$ is the vector of all $k^{th}$ order sample central moments of the coordinates of $X$.*

10. 希尔伯特空间下统一一阶矩和二阶矩



**Definition 1** The maximum covariance discrepancy (MCD) is defined as

$$\mathrm{MCD}[p,q,\mathcal{H}] = \sup_{\|a\|\le 1}\sum_{i,j\in I} a_{ij}\left(\mathrm{cov}\left[e_i(x), e_j(x)\right] - \mathrm{cov}\left[e_i(y), e_j(y)\right]\right), \tag{1}$$

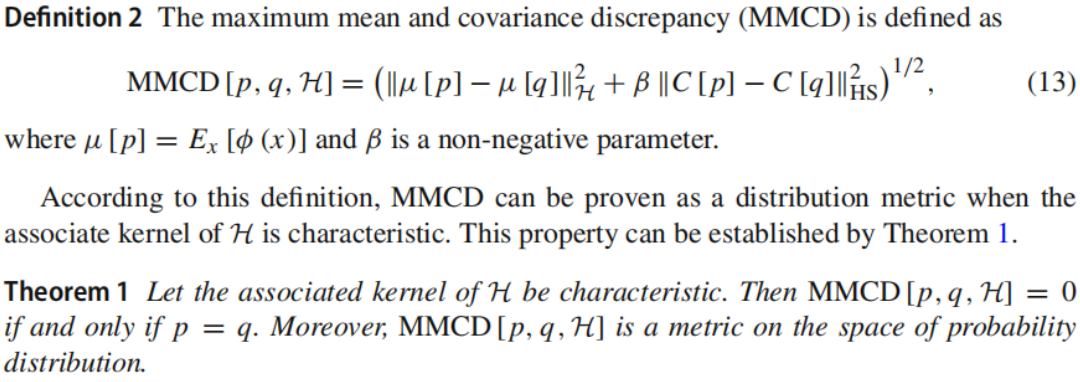
where $\{e_i | i \in I\}$ is an orthogonal basis of $\mathcal{H}$, $\|a\| = (\sum_{i,j\in I} a_{ij}^2)^{1/2}$, and the notation cov has the following formula: $\mathrm{cov}\left[e_i(x), e_j(x)\right] = E_x\left[e_i(x)\, e_j(x)\right] - E_x\left[e_i(x)\right] E_x\left[e_j(x)\right]$.

Next, we show that the (1) can be associated with the Hilbert–Schmidt norm with the following lemma:

**Lemma 1**

$$\mathrm{MCD}[p,q,\mathcal{H}] = \|C[p] - C[q]\|_{\mathrm{HS}}, \tag{2}$$

*where $\|\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm of the vectors in* HS $(\mathcal{H})$*, which is the Hilbert space of Hilbert–Schmidt operators mapping from $\mathcal{H}$ to $\mathcal{H}$.*



**Definition 2** The maximum mean and covariance discrepancy (MMCD) is defined as

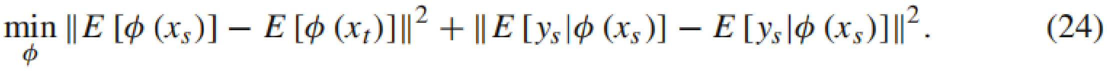
$$\mathrm{MMCD}[p,q,\mathcal{H}] = \left(\|\mu[p]-\mu[q]\|_{\mathcal{H}}^2 + \beta\,\|C[p]-C[q]\|_{\mathrm{HS}}^2\right)^{1/2}, \tag{13}$$

where $\mu[p] = E_x[\phi(x)]$ and $\beta$ is a non-negative parameter.

According to this definition, MMCD can be proven as a distribution metric when the associate kernel of $\mathcal{H}$ is characteristic. This property can be established by Theorem 1.

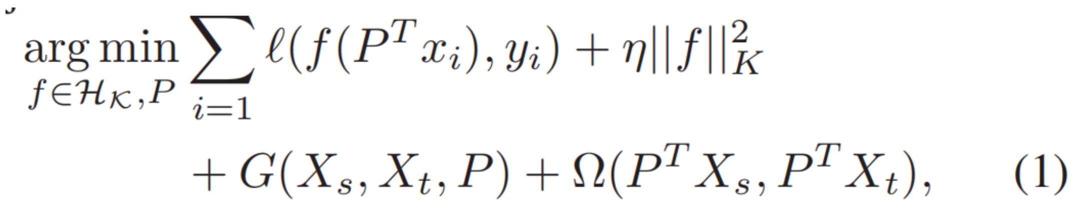
**Theorem 1** *Let the associated kernel of $\mathcal{H}$ be characteristic. Then* $\mathrm{MMCD}[p,q,\mathcal{H}] = 0$ *if and only if $p = q$. Moreover,* $\mathrm{MMCD}[p,q,\mathcal{H}]$ *is a metric on the space of probability distribution.*

优化目标：最小化边缘分布和条件分布下的 MMCD



$$\min_{\phi}\|E[\phi(x_s)] - E[\phi(x_t)]\|^2 + \|E[y_s|\phi(x_s)] - E[y_s|\phi(x_s)]\|^2. \tag{24}$$

11. ICME 2019



$$\arg\min_{f\in\mathcal{H}_K, P} \sum_{i=1} \ell(f(P^T x_i), y_i) + \eta ||f||_K^2 + G(X_s, X_t, P) + \Omega(P^T X_s, P^T X_t), \qquad (1)$$

一阶二阶按顺序进行，分别求出 P,Q，不是同时