Kernels on Histograms through the Transportation Polytope

Marco Cuturi, cuturi@ism.ac.jp

The Institute of Statistical Mathematics, Tokyo

August 1st - MLSS Workshop

Outline

Outline

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Some solutions

- Use fingerprints (feature vectors defined by experts)
- Slice the objects if they are discrete and convolute kernels on subparts [Haussler, 1998]
- Use the power of statistical models [Jaakkola, 1998]
- In the most simple setting, use bags-of-components to describe your object, that is histograms.

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Some solutions

- Use fingerprints (feature vectors defined by experts)
- Slice the objects if they are discrete and convolute kernels on subparts [Haussler, 1998]
- Use the power of statistical models [Jaakkola, 1998]
- In the most simple setting, use bags-of-components to describe your object, that is histograms.

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Some solutions

- Use fingerprints (feature vectors defined by experts)
- Slice the objects if they are discrete and convolute kernels on subparts [Haussler, 1998]
- Use the power of statistical models [Jaakkola, 1998]
- In the most simple setting, use bags-of-components to describe your object, that is histograms.

Problem

- Kernel methods, SVM's are well understood for vector inputs with a challenge: which $\gamma, \Sigma, a, b, d \ldots$?
- However, there is no natural dot-product or norm for strings, graphs, texts... so we cannot use Gaussian $e^{-\gamma\|\cdot-\cdot\|^2}$, polynomial $(< \cdot, \cdot > +b)^d$ etc..

Some solutions

- Use fingerprints (feature vectors defined by experts)
- Slice the objects if they are discrete and convolute kernels on subparts [Haussler, 1998]
- Use the power of statistical models [Jaakkola, 1998]
- In the most simple setting, use bags-of-components to describe your object, that is histograms.

Bag-of-components, histogram representations

Example

A long sequence seen as a bag of n-grams:

$$AABHLKFHGH \cdots HAABGJYHLK \cdots AT$$
$$\mapsto \{(AAB, 2), (HLK, 2), (FHG, 1) \cdots\}$$

Bag-of-components, histogram representations

Example

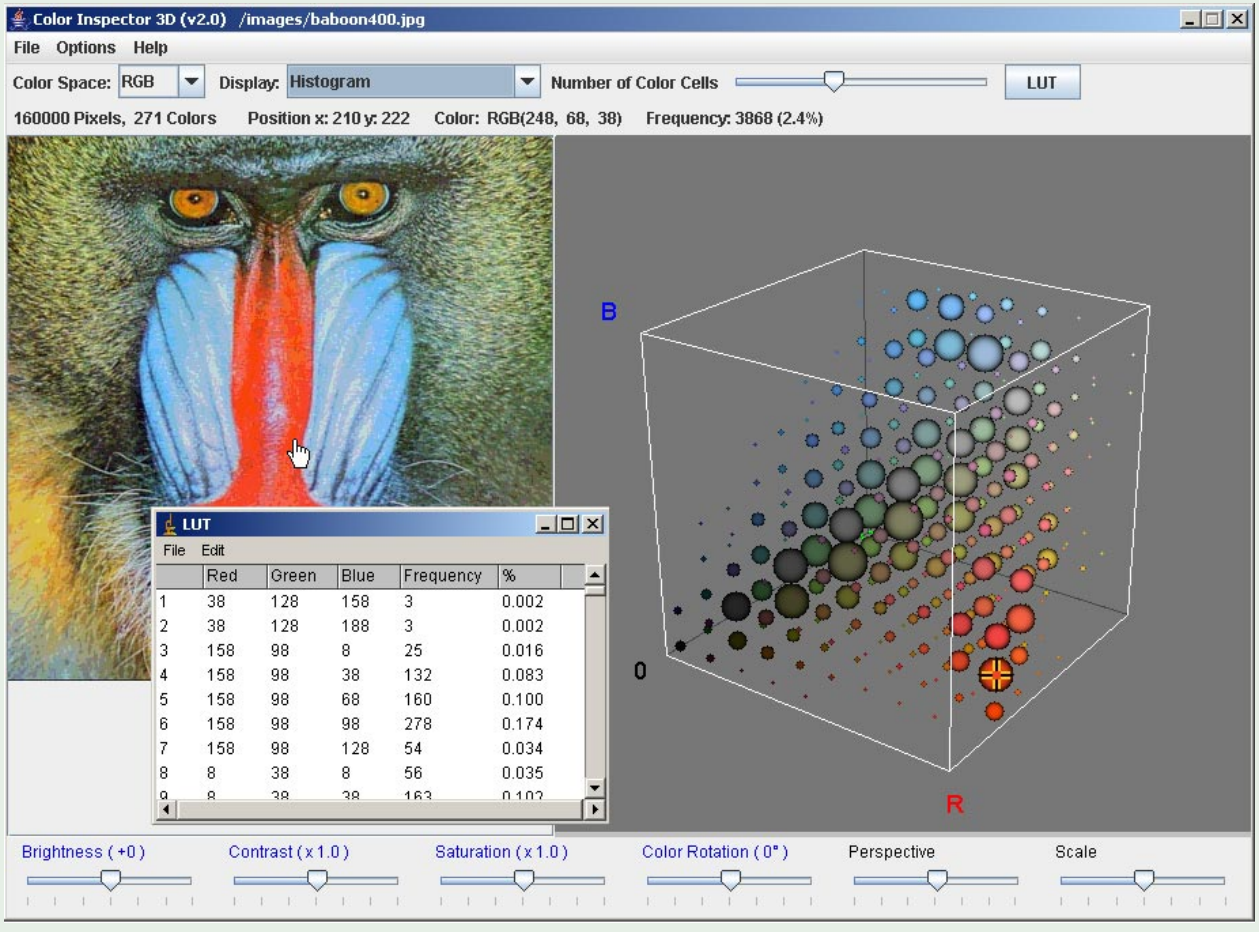
A text as a bag of words:

$$\texttt{the cat eats the mouse} \mapsto \{(\texttt{the}, 2), (\texttt{cat}, 1), \cdots\}.$$

Bag-of-components, histogram representations

Example

An image as an histogram of RGB colors, or any other features:



Probability representation

A composite object z (a text) can be seen as an aggregation of basic components (words) taken from $\mathcal{X}$ (a dictionary).

$$z \mapsto \mu \in M_1^+(\mathcal{X})$$

when $\mathcal{X}$ is finite $\text{card}(\mathcal{X}) = d$, this is a multinomial distribution from Σ_d, that is a vector $\theta \in [0,1]^d$ such that $\sum_{i=1}^d \theta_i = 1$.

kernels for structured objects $\longrightarrow$ kernels for histograms

In practice: complex objects $\longrightarrow$ histograms $\overset{k_{\Sigma_d}+\text{SVM}}{\longrightarrow}$ results.

Probability representation

A composite object z (a text) can be seen as an aggregation of basic components (words) taken from $\mathcal{X}$ (a dictionary).

$$z \mapsto \mu \in M_1^+(\mathcal{X})$$

when $\mathcal{X}$ is finite $\text{card}(\mathcal{X}) = d$, this is a multinomial distribution from Σ_d, that is a vector $\theta \in [0,1]^d$ such that $\sum_{i=1}^d \theta_i = 1$.

kernels for structured objects $\rightarrow$ kernels for histograms

In practice: complex objects $\rightarrow$ histograms $\overset{k_{\Sigma_d}+\text{SVM}}{\longrightarrow}$ results.

$\Sigma_d \subset \mathbb{R}_+^d \subset \mathbb{R}^d$

- Sure, so Gaussian, polynomials etc.. can be used.
- In fact most early applications of SVM [Joachims '98, Chapelle '99] used this framework.
- However... counts are positive [Hein, Bousquet '05], which make multinomial vectors lie in the positive orthant,
- and normalized so their mass is 1, so we can use the information geometry of the manifold of probability measures [Lafferty & Lebanon '05, Lebanon '06, Kondor Jebara '03, C Fukumizu Vert 05 etc.]

$\Sigma_d \subset \mathbb{R}_+^d \subset \mathbb{R}^d$

- Sure, so Gaussian, polynomials etc.. can be used.
- In fact most early applications of SVM [Joachims '98, Chapelle '99] used this framework.
- However... counts are positive [Hein, Bousquet '05], which make multinomial vectors lie in the positive orthant,
- and normalized so their mass is 1, so we can use the information geometry of the manifold of probability measures [Lafferty & Lebanon '05, Lebanon '06, Kondor Jebara '03, C Fukumizu Vert 05 etc.]

$\Sigma_d \subset \mathbb{R}_+^d \subset \mathbb{R}^d$

- Sure, so Gaussian, polynomials etc.. can be used.
- In fact most early applications of SVM [Joachims '98, Chapelle '99] used this framework.
- However... counts are positive [Hein, Bousquet '05], which make multinomial vectors lie in the positive orthant,
- and normalized so their mass is 1, so we can use the information geometry of the manifold of probability measures [Lafferty & Lebanon '05, Lebanon '06, Kondor Jebara '03, C Fukumizu Vert 05 etc.]

$\Sigma_d \subset \mathbb{R}_+^d \subset \mathbb{R}^d$

- Sure, so Gaussian, polynomials etc.. can be used.
- In fact most early applications of SVM [Joachims '98, Chapelle '99] used this framework.
- However... counts are positive [Hein, Bousquet '05], which make multinomial vectors lie in the positive orthant,
- and normalized so their mass is 1, so we can use the information geometry of the manifold of probability measures [Lafferty & Lebanon '05, Lebanon '06, Kondor Jebara '03, C Fukumizu Vert 05 etc.]

Dozens of classic probability metrics

- Some well-known divergences and metrics can be used directly as kernels: Jensen-Divergence, Hellinger distance, χ_2 distance etc... (long list, but not KL)
- Some kernels aim at incorporating a prior information (a kernel κ) on $\mathcal{X}$... as if they were actually computing these metrics in a feature space (the rkhs $\mathcal{H}_\kappa$).

The transport or Monge-Kantorovich distance

- Very old distance (goes back to Monge, late 18th), known as 'Earth Movers' in vision.
- It is not directly a (n.d) kernel (remember the $\exp(-\text{dist})$ distances ?) but we can take a closer look at its possible kernel.

Dozens of classic probability metrics

- Some well-known divergences and metrics can be used directly as kernels: Jensen-Divergence, Hellinger distance, χ_2 distance etc... (long list, but not KL)
- Some kernels aim at incorporating a prior information (a kernel κ) on $\mathcal{X}$... as if they were actually computing these metrics in a feature space (the rkhs $\mathcal{H}_\kappa$).

The transport or Monge-Kantorovich distance

- Very old distance (goes back to Monge, late 18th), known as 'Earth Movers' in vision.
- It is not directly a (n.d) kernel (remember the edit distance ?) but we can take a closer look at it to build a kernel.

Dozens of classic probability metrics

- Some well-known divergences and metrics can be used directly as kernels: Jensen-Divergence, Hellinger distance, χ_2 distance etc... (long list, but not KL)
- Some kernels aim at incorporating a prior information (a kernel κ) on $\mathcal{X}$... as if they were actually computing these metrics in a feature space (the rkhs $\mathcal{H}_\kappa$).

The transport or Monge-Kantorovich distance

- Very old distance (goes back to Monge, late 18th), known as Earth Movers' in vision.
- It is not directly a (n.d) kernel (remember the edit distance?) but we can take a closer look at it to build a kernel.

Dozens of classic probability metrics

- Some well-known divergences and metrics can be used directly as kernels: Jensen-Divergence, Hellinger distance, χ_2 distance etc... (long list, but not KL)
- Some kernels aim at incorporating a prior information (a kernel κ) on $\mathcal{X}$... as if they were actually computing these metrics in a feature space (the rkhs $\mathcal{H}_\kappa$).

The transport or Monge-Kantorovich distance

- Very old distance (goes back to Monge, late 18th), known as Earth Movers' in vision.
- It is not directly a (n.d) kernel (remember the edit distance?) but we can take a closer look at it to build a kernel.

Outline

MK distance between two histograms

- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that

$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$

- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$

MK distance between two histograms

- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that

$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$

- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$

MK distance between two histograms

- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that

$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$

- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$

MK distance between two histograms

- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that

$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$

- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$

MK distance between two histograms

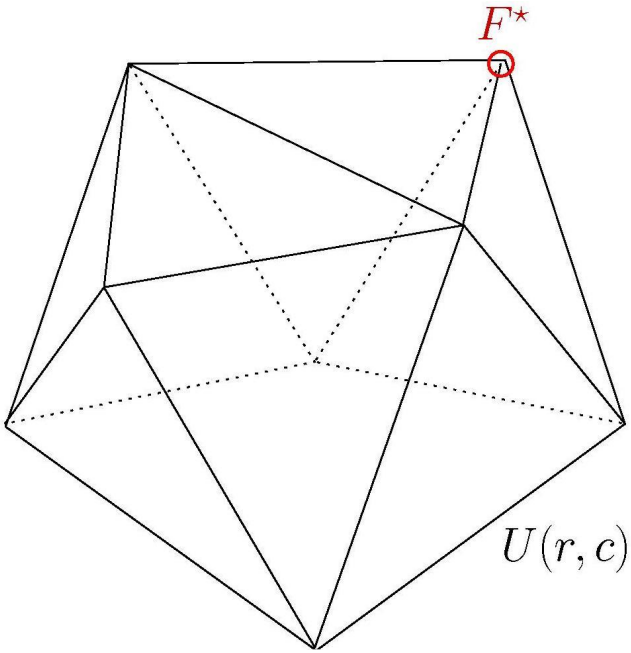
- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that
$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$
- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$

MK distance between two histograms

- Consider two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N.
- $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins.
- $U(r, c) = \{F \in \mathbb{N}_{d \times d} \,|\, F\mathbb{1}_d = r,\ F^\top \mathbb{1}_d = c\}$, transportation matrices between r and c.
- $F = f_{ij}$ is such that

$$\left(\begin{array}{cccc|c} f_{11} & f_{12} & \cdots & f_{1d} & r_1 \\ f_{21} & f_{22} & \cdots & f_{2d} & r_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{d1} & f_{d2} & \cdots & f_{dd} & r_d \\ \hline c_1 & c_2 & \cdots & c_d & N \end{array}\right)$$

- Then, $d_{\mathrm{MK}}(r, c) = \min_{F \in U(r,c)} < F, T >$.
- We write $F^\star = \operatorname{argmin}_{F \in U(r,c)} < F, T >$



Within the whole transportation polytope, and for a given T, the MK distance only considers $< F^\star, T >$. How can we use the extra information of $U(r,c)$?

Permanent

For a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is

$$\text{per}\, M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

Permanent kernel between clouds of points

- Two finite clouds of points, $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, x_i and $y_i \in \mathcal{X}$.
- Consider further that $\mathcal{X}$ has a kernel κ.
- Then $k_{\text{per}} : (x, y) \mapsto \text{per}[h(x_i, y_j)]_{i,j=1,\ldots,n}$ is p.d.
- If T is a distance matrix and $\kappa = e^{-\frac{\cdot}{t}}$,

$$k_{\text{per}}(x, y) = \sum_{\sigma \in S_n} e^{-\frac{\sum_i T(x_i, y_{\sigma(i)})}{t}}$$

Permanent

For a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is

$$\text{per}\, M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

Permanent kernel between clouds of points

- Two finite clouds of points, $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, x_i and $y_i \in \mathcal{X}$.
- Consider further that $\mathcal{X}$ has a kernel κ.
- Then $k_{\text{per}} : (x, y) \mapsto \text{per}([\kappa(x_i, y_j)]_{1 \leq i,j \leq n})$ is p.d.
- If T is a distance matrix and $\kappa = e^{-T}$,

$$k_{\text{per}}(x, y) = \sum_{\sigma \in S_n} e^{-\sum_{i=1}^{n} t(x_{\sigma_i}, y_i)}.$$

Permanent

For a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is

$$\operatorname{per} M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

Permanent kernel between clouds of points

- Two finite clouds of points, $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, x_i and $y_i \in \mathcal{X}$.
- Consider further that $\mathcal{X}$ has a kernel κ.
- Then $k_{\text{per}} : (x, y) \mapsto \operatorname{per}([\kappa(x_i, y_j)]_{1 \leq i,j \leq n})$ is p.d.
- If T is a distance matrix and $\kappa = e^{-T}$,

$$k_{\text{per}}(x, y) = \sum_{\sigma \in S_n} e^{-\sum_{i=1}^{n} t(x_{\sigma_i}, y_i)}.$$

Permanent

For a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is

$$\text{per}\, M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

Permanent kernel between clouds of points

- Two finite clouds of points, $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, x_i and $y_i \in \mathcal{X}$.
- Consider further that $\mathcal{X}$ has a kernel κ.
- Then $k_{\text{per}} : (x, y) \mapsto \text{per}([\kappa(x_i, y_j)]_{1 \leq i,j \leq n})$ is p.d.
- If T is a distance matrix and $\kappa = e^{-T}$,

$$k_{\text{per}}(x, y) = \sum_{\sigma \in S_n} e^{-\sum_{i=1}^{n} t(x_{\sigma_i}, y_i)}.$$

Permanent

For a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is

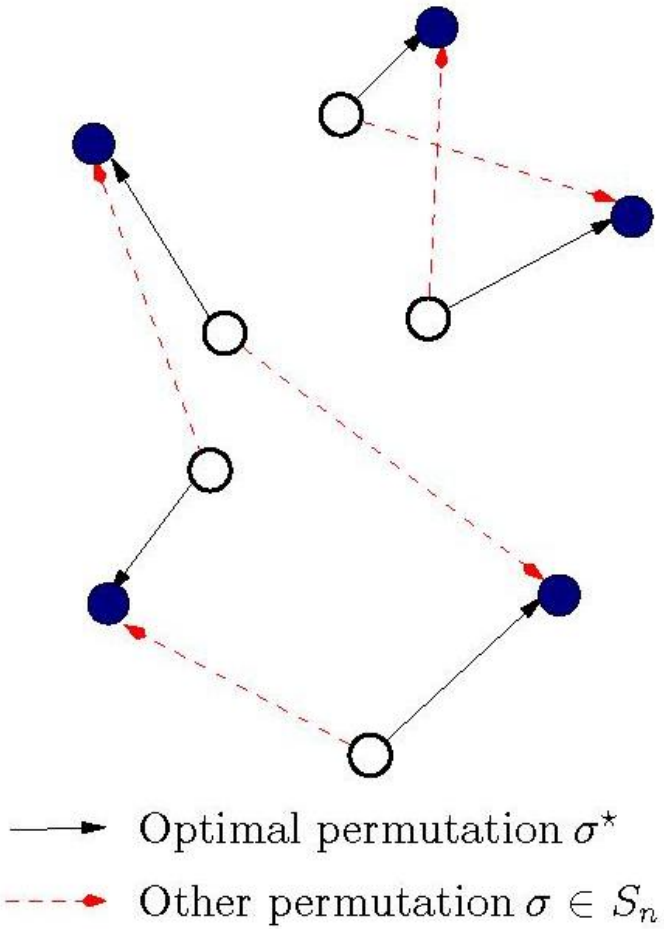
$$\operatorname{per} M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

Permanent kernel between clouds of points

- Two finite clouds of points, $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, x_i and $y_i \in \mathcal{X}$.
- Consider further that $\mathcal{X}$ has a kernel κ.
- Then $k_{\text{per}} : (x, y) \mapsto \operatorname{per}([\kappa(x_i, y_j)]_{1 \leq i,j \leq n})$ is p.d.
- If T is a distance matrix and $\kappa = e^{-T}$,

$$k_{\text{per}}(x, y) = \sum_{\sigma \in S_n} e^{-\sum_{i=1}^{n} t(x_{\sigma_i}, y_i)}.$$

Transportation of clouds of points



$$k_{\text{per}}(x, y) = e^{-d_\sigma^\star} + e^{-d_\sigma} + \dots$$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{vol}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between $T \in U(r, c)$ and pairs of Gelfand-Young Tableaux (p, q) of size n with identical shape γ.
- Hence $|U(r, c)| = \sum_{\gamma} K_{\gamma r} K_{\gamma c}$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{\text{Vol}}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between $T \in U(r, c)$ and pairs of Gelfand-Young Tableaux (P, Q) of the same shape $\nu \vdash n$.
- Hence $|U(r, c)| = \sum_{\nu} K_{\nu,r} K_{\nu,c}$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{\text{Vol}}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between every $T \in U(r, c)$ and pairs of Gen. Young Tableaux $(\gamma, t), (\gamma, c)$ with identical shape γ.
- Hence $|U(r, c)| = \sum_{\gamma} K_{\gamma, t} K_{\gamma, c}$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{\text{vol}}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between every $T \in U(r, c)$ and pairs of Gen. Young Tableaux $(\eta, r), (\eta, c)$ with identical shape η.
- Hence $|U(r, c)| = \sum_{\eta} K_{\eta,r} K_{\eta,c}$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{\text{vol}}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between every $T \in U(r, c)$ and pairs of Gen. Young Tableaux $(\eta, r), (\eta, c)$ with identical shape η.
- Hence $|U(r, c)| = \sum_{\eta} K_{\eta,r} K_{\eta,c}$

Volume of $U(r, c)$ for general histograms

- $|U(r, c)|$ quantifies the number of transportation plans
- Extremely difficult to compute exactly (active research topic) [Lorea & al. 2004,Barvinok 1993, Beck...]
- Some approximations exists [Diaconis & al. 1994]

Positive definiteness of the volume

- $k_{\text{vol}}(r, c) = |U(r, c)|$ is p.d.
- Proof: RSK one-to-one correspondence between every $T \in U(r, c)$ and pairs of Gen. Young Tableaux $(\eta, r), (\eta, c)$ with identical shape η.
- Hence $|U(r, c)| = \sum_{\eta} K_{\eta,r} K_{\eta,c}$

Weighted volume of $U(r, c)$

- $f(T) = \sum_{F \in U(r,c)} e^{-<F,T>}$: generating function
- Also very difficult to compute although some SIS approximations have been proposed

Positive definiteness of weighted volumes

- define $k_{\varphi,T} : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}$
- and $\varphi_a : \mathbb{N}_{d \times d} \to \mathbb{R}$ as

$$\varphi_a(F) = \prod_{i,j} \frac{(2a)^{a}}{(f_{ij}+a)^{(f_{ij}+a)}}$$

- for $0 \leq a \leq 2$ we have that $k_{\varphi_a,T}$ is p.d.
- In particular: $a \to 0$ Fisher-Yates, $a \to 1$ only uniform weights

Weighted volume of $U(r, c)$

- $f(T) = \sum_{F \in U(r,c)} e^{-<F,T>}$: generating function
- Also very difficult to compute although some SIS approximations have been proposed

Positive definiteness of weighted volumes

- define $k_{\varphi_a,T} : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}$
- and $\varphi_a : \mathbb{N}_{d \times d} \to \mathbb{R}$ as

$$\varphi_a(F) = \prod_{ij} \frac{(2f_{ij})!^a}{f_{ij}!} \prod_{i \neq j} f_{ij}!^{2a-1}$$

- for $0 \leq a \leq 2$ we have that $k_{\varphi_a,T}$ is p.d
- In particular: $a = 0$ Fisher-Yates, $a = \frac{1}{2}$ only the row weights

Weighted volume of $U(r, c)$

- $f(T) = \sum_{F \in U(r,c)} e^{-<F,T>}$: generating function
- Also very difficult to compute although some SIS approximations have been proposed

Positive definiteness of weighted volumes

- define $k_{\varphi,T} : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}$
- and $\varphi_a : \mathbb{N}_{d \times d} \to \mathbb{R}$ as

$$\varphi_a(F) = \prod_i \frac{(2f_{ii})!^a}{f_{ii}!} \prod_{i \neq j} f_{ij}!^{2a-1}$$

for $0 \leq a \leq 2$ we have that $k_{\varphi,T}$ is p.d

- In particular: $a = 0$ Fisher-Yates, $a = \frac{1}{2}$ only diagonal weights.

Weighted volume of $U(r, c)$

- $f(T) = \sum_{F \in U(r,c)} e^{-<F,T>}$: generating function
- Also very difficult to compute although some SIS approximations have been proposed

Positive definiteness of weighted volumes

- define $k_{\varphi,T} : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}$
- and $\varphi_a : \mathbb{N}_{d \times d} \to \mathbb{R}$ as

$$\varphi_a(F) = \prod_i \frac{(2f_{ii})!^a}{f_{ii}!} \prod_{i \neq j} f_{ij}!^{2a-1}$$

for $0 \leq a \leq 2$ we have that $k_{\varphi,T}$ is p.d

- In particular: $a = 0$ Fisher-Yates, $a = \frac{1}{2}$ only diagonal weights.

Weighted volume of $U(r, c)$

- $f(T) = \sum_{F \in U(r,c)} e^{-<F,T>}$: generating function
- Also very difficult to compute although some SIS approximations have been proposed

Positive definiteness of weighted volumes

- define $k_{\varphi,T} : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}$
- and $\varphi_a : \mathbb{N}_{d \times d} \to \mathbb{R}$ as

$$\varphi_a(F) = \prod_i \frac{(2f_{ii})!^a}{f_{ii}!} \prod_{i \neq j} f_{ij}!^{2a-1}$$

  for $0 \leq a \leq 2$ we have that $k_{\varphi,T}$ is p.d
- In particular: $a = 0$ Fisher-Yates, $a = \frac{1}{2}$ only diagonal weights.

2000 images, 10 classes, 20 pixels per image

σ	Gaussian	Permanent
.1	34.3 ($\pm$ 1.4)	32.3 ($\pm$ 1.2)
.2	33.45 ($\pm$ 1.0)	31.3 ($\pm$ 1.3)
.3	37.3 ($\pm$ 1.0)	33.2 ($\pm$ 1.2)

Table: Misclassification rate expressed in percents for the 2 considered kernels along with their standard errors averaged over 3 cross-validation folds.