Introduction To DCA

K.K M

mkk@hust.edu.cn

December 12, 2016

Contents

Definition of Distance

A statistical distance quantifies the distance between two statistical objects, which can be two random variables, or two probability distributions or samples, or the distance can be between an individual sample point and a population or a wider sample of points.

- $d(x, x) = 0$ ▶ Identity of indiscernibles
- $d(x, y) \geq 0$ ▶ Non negative
- $d(x, y) = d(y, x)$ ▶ Symmetry
- $d(x, k) + d(k, y) \geq d(x, y)$ ▶ Triangle inequality

Different Distances

- Minkowski distance
- Euclidean distance
- Manhattan distance
- Chebyshev distance
- Mahalanobis distance
- Cosine similarity
- Pearson correlation
- Hamming distance
- Jaccard similarity
- Levenshtein distance
- DTW distance
- KL-Divergence

Minkowski Distance

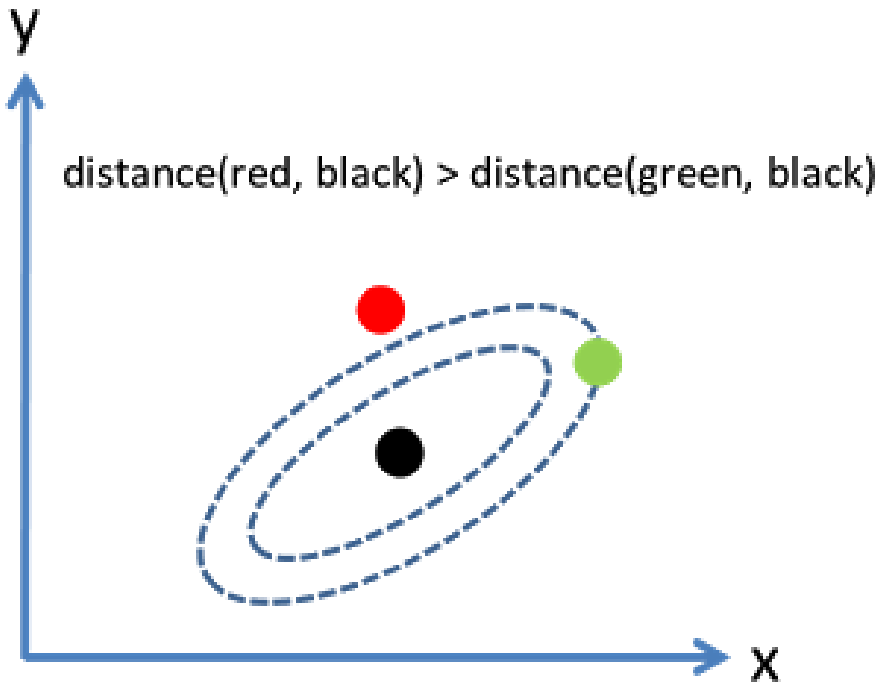
$P = (x_1, x_2, ..., x_n)$ and $Q = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$

- Minkowski distance: $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
  - Euclidean distance: $p = 2$
  - Manhattan distance: $p = 1$
  - Chebyshev distance: $p = \infty$

$$\lim_{p \to \infty} (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} = \max_{i=1}^{n} |x_i - y_i|$$

- z-transform: $(x_i, y_i) \mapsto (\frac{x_i - \mu_x}{\sigma_x}, \frac{y_i - \mu_y}{\sigma_y})$, x_i and y_i is independent

Mahalanobis Distance



- Mahalanobis Distance

  Covariance matrix $C = LL^T$

  Transformation $x \xrightarrow{L^{-1}(x-\mu)} x'$

Example for the Mahalanobis distance

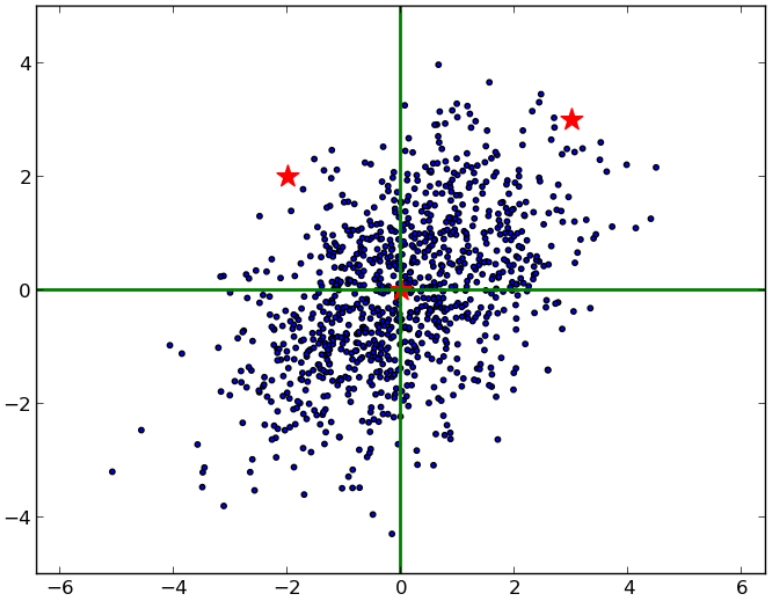


Figure: Euclidean distance

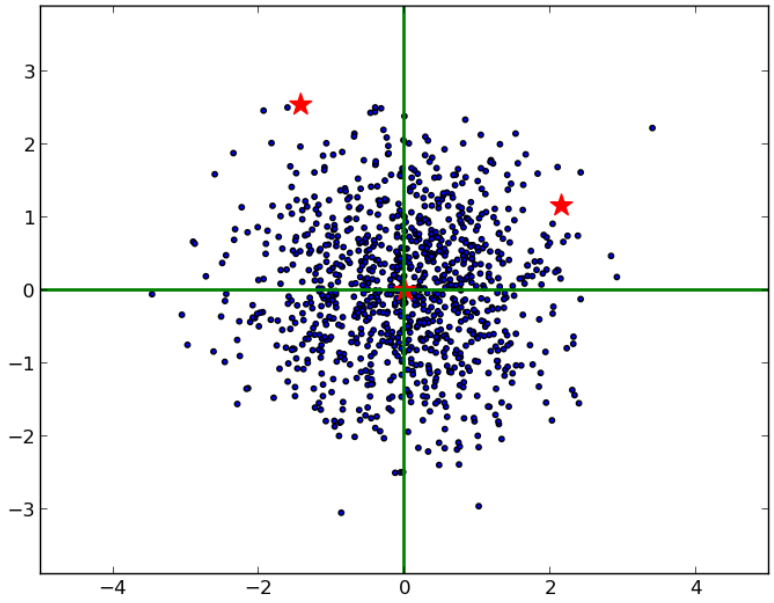


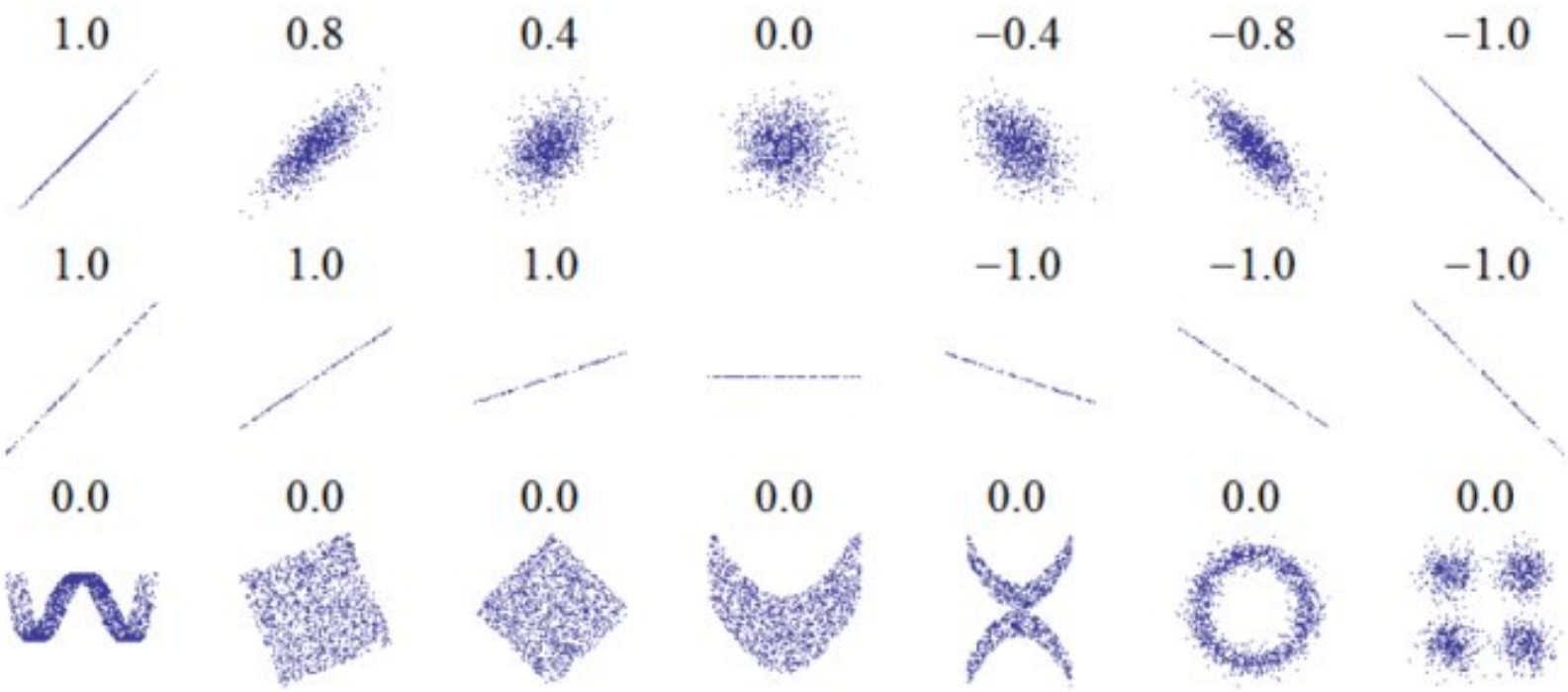
Figure: Mahalanobis distance

Pearson Correlation

- Inner Product: $Inner(x, y) = \langle x, y \rangle = \sum_i x_i y_i$
- Cosine similarity: $CosSim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x,y \rangle}{\|x\| \, \|y\|}$
- Pearson correlation

$$\begin{aligned} Corr(x, y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \, \|y - \bar{y}\|} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

Different Values for the Pearson Correlation



Limits of Pearson Correlation

Q: Why don't use Pearson correlation in pairwise associations?

A: Pearson correlation is a misleading measure for direct dependence as it only reflects the association between two variables while ignoring the influence of the remaining ones.

Partial Correlation

For M given samples in L measured variables:

$$\mathbf{x}^1 = (x_1^1, \ldots, x_L^1)^T, \ldots, \mathbf{x}^M = (x_1^M, \ldots, x_L^M) \in \mathbb{R}^L$$

Pearson correlation coefficient:

$$r_{ij} = \frac{\hat{C}_{ij}}{\sqrt{\hat{C}_{ii}\hat{C}_{jj}}}$$

where $\hat{C}_{ij} = \frac{1}{M}\sum_{m=1}^{M}(x_i^m - \bar{x}_i)(x_j^m - \bar{x}_j)$ is empirical covariance matrix.
Scale ecah of variables to zero-mean and unit-standard deviation,

$$x_i \mapsto \frac{(x_i - \bar{x}_i)}{\sqrt{\hat{C}_{ii}}}$$

Simplify the correlation coefficient: $r_{ij} \equiv \overline{x_i x_j}$

Partial Correlation of Three-variable System

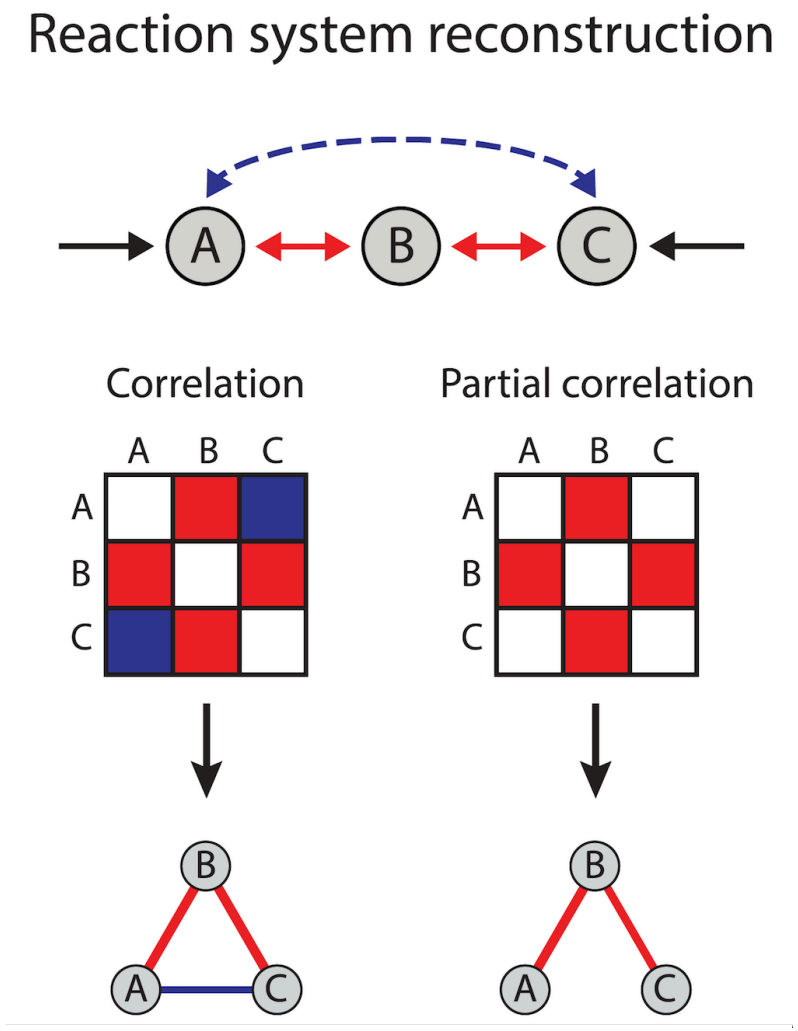
The partial correlation between X and Y given a set of n controlling variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$, written $r_{XY \cdot Z}$, is

$$r_{XY \cdot \mathbf{z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - (\sum_{i=1}^{N} r_{X,i})^2} \sqrt{N \sum_{i=1}^{N} r_{Y,i}^2 - (\sum_{i=1}^{N} r_{Y,i})^2}}$$

In particularly, where $\mathbf{Z}$ is a single variable, which of a three random variables between x_A and x_B given x_C is defined as

$$r_{AB \cdot C} = \frac{r_{AB} - r_{BC} r_{AC}}{\sqrt{1 - r_{AC}^2} \sqrt{1 - r_{BC}^2}} \equiv -\frac{(\hat{C}^{-1})_{AB}}{\sqrt{(\hat{C}^{-1})_{AA} (\hat{C}^{-1})_{BB}}}$$

Reaction System Reconstruction



Entropy

The thermodynamic definition of entropy was developed in the early 1850s by Rudolf Clausius:

$$\Delta S = \int \frac{\delta Q_{rev}}{T}$$

In 1948, Shannon defined the entropy H of a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ and probability mass function $p(x)$ as:

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Joint & Conditional Entropy

- Joint Entropy: $H(X, Y)$
- Conditional Entropy: $H(Y|X)$

$$
\begin{aligned}
H(X, Y) - H(X) &= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_{x} p(x) \log p(x) \\
&= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_{x} \left(\sum_{y} p(x, y) \right) \log p(x) \\
&= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_{x,y} p(x, y) \log p(x) \\
&= -\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)} \\
&= -\sum_{x,y} p(x, y) \log p(y|x) \\
&= H(Y|X)
\end{aligned}
$$

Relative Entropy & Mutual Information

- Relative Entropy(KL-Divergence): $D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$
  - $D(p\|q) \neq D(q\|p)$
  - $D(p\|q) \geq 0$
- Mutual Information:

$$I(X,Y) = D(\, p(x,y) \,\|\, p(x)p(y) \,) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual Information

$$
\begin{aligned}
H(Y) - I(X,Y) &= -\sum_{y} p(y) \log p(y) - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= -\sum_{y} \left(\sum_{x} p(x,y) \right) \log p(y) - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= -\sum_{x,y} p(x,y) \log p(y) - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} \\
&= H(Y|X)
\end{aligned}
$$

$$
H(X,Y) - H(X) = H(Y|X) \;\Rightarrow\; I(X,Y) = H(X) + H(Y) - H(X,Y)
$$

MI & DI

- $MI_{ij} = \sum_{\sigma,\omega} f_{ij}(\sigma,\omega) \ln \left(\frac{f_{ij}(\sigma,\omega)}{f_i(\sigma) f_j(\omega)} \right)$
- $DI_{ij} = \sum_{\sigma,\omega} P_{ij}^{dir}(\sigma,\omega) \ln \left(\frac{P_{ij}^{dir}(\sigma,\omega)}{f_i(\sigma) f_j(\omega)} \right)$

where $P_{ij}^{dir}(\sigma,\omega) = \frac{1}{z_{ij}} \exp(e_{ij}(\sigma,\omega) + \tilde{h}_i(\sigma) + \tilde{h}_j(\omega))$

Principle of Maximum Entropy

- The principle was first expounded by E.T.Jaynes in two papers in 1957 where he emphasized a natural correspondence between statistical mechanics and information theory.
- The principle of maximum entropy states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy.
- maximize $S = -\int_x P(x) \ln P(x) dx$

Example

- Suppose we want to estimate a probability distribution $p(a, b)$, where $a \in \{x, y\}$ ans $b \in \{0, 1\}$
- Furthermore the only fact known about p is that $p(x, 0) + p(y, 0) = 0.6$

p(a,b)	0	1	
x	?	?	
y	?	?	
	0.6		1.0

Unbiased Principle

$p(a,b)$	0	1	
x	0.5	0.1	
y	0.1	0.3	
	0.6		**1.0**

Figure: One way to satisfy constraints

$p(a,b)$	0	1	
x	0.3	0.2	
y	0.3	0.2	
	0.6		**1.0**

Figure: The most uncertain way to satisfy constraints

Maximum-Entropy Probability Model

- Continuous random variables: $\mathbf{x} = (x_1, \ldots, x_L)^T \in \mathbb{R}^L$
- Constraints:
  - $\int_x P(x)dx = 1$
  - $\langle x_i \rangle = \int_x P(x) x_i dx = \frac{1}{M} \sum_{m=1}^{M} x_i^m = \overline{x_i}$
  - $\langle x_i x_j \rangle = \int_x P(x) x_i x_j dx = \frac{1}{M} \sum_{m=1}^{M} x_i^m x_j^m = \overline{x_i x_j}$
- Maximize: $S = -\int_x P(x) \ln P(x) dx$

Lagrange Multipliers Method

Convert a constrained optimization problem into an unconstrained one by means of the Lagrangian $\mathcal{L}$.

$$\mathcal{L} = S + \alpha(\langle 1 \rangle - 1) + \sum_{i=1}^{L} \beta_i(\langle x_i \rangle - \overline{x_i}) + \sum_{i,j=1}^{L} \gamma_{ij}(\langle x_i x_j \rangle - \overline{x_i x_j})$$

$$\frac{\delta \mathcal{L}}{\delta P(x)} = 0 \;\Rightarrow\; -\ln P(x) - 1 + \alpha + \sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j = 0$$

Pairwise maximum-entropy probability distribution

$$P(\mathbf{x}; \beta, \gamma) = \exp\left(-1 + \alpha + \sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j\right) = \frac{1}{Z} e^{-\mathcal{H}(x;\beta,\gamma)}$$

Lagrange Multipliers Method

- Partition function as normalization constant

$$Z(\beta, \gamma) := \int_x \exp\left(\sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j\right) dx \equiv \exp(1-\alpha)$$

- Hamiltonian

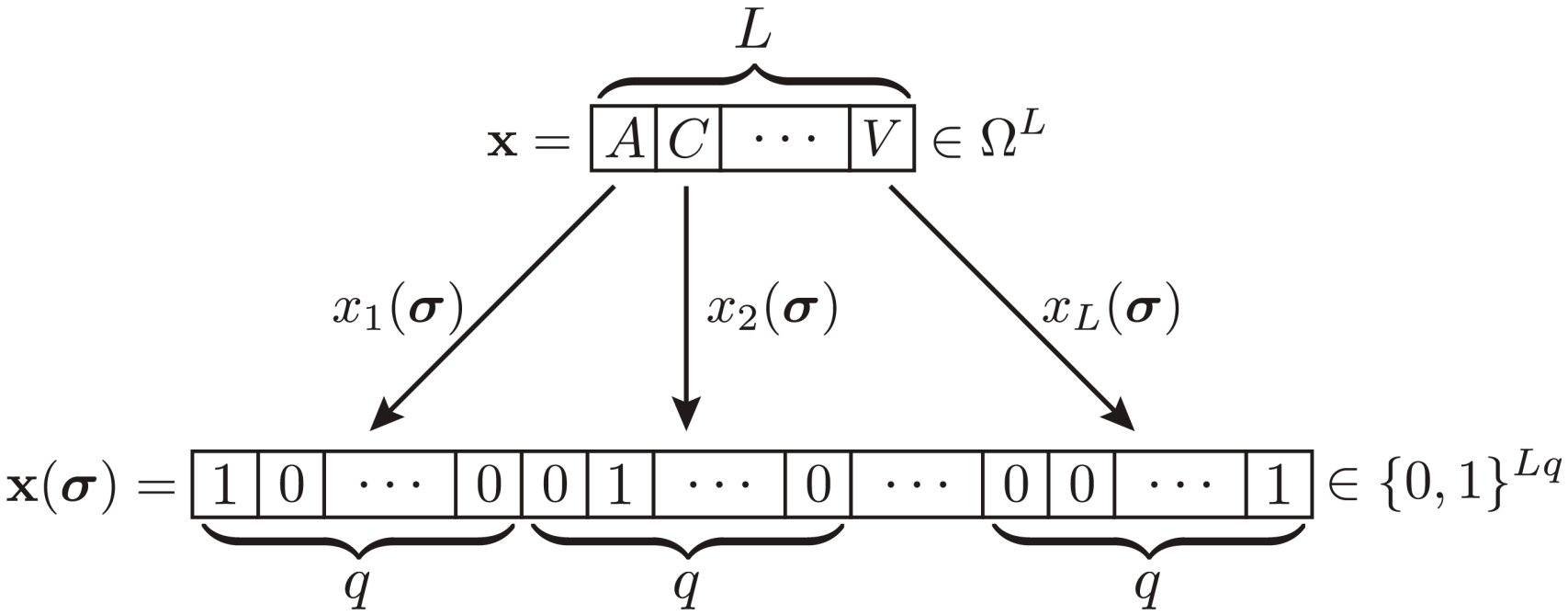
$$\mathcal{H}(x) := -\sum_{i=1}^{L} \beta_i x_i - \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j$$

Categorical Random Variables

- jointly distributed categorical variables $\mathbf{x} = (x_1, \ldots, x_L)^T \in \Omega^L$, each x_i is defined on the finite set $\Omega = \{\sigma_1, \ldots, \sigma_q\}$.
- In the concrete example of modeling protein co-evolution, this set contains the 20 amino acids represented by a 20-letter alphabet plus one gap element.

  $\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}$ and $q = 21$

Binary Embedding of Amino Acid Sequence



Pairwise MEP Distribution on Categorical Variables

- single and pairwise marginal probability

$$\langle x_i(\sigma)\rangle = \sum_{x(\sigma)} P(x(\sigma))x_i(\sigma) = \sum_{x} P(x_i = \sigma) = P_i(\sigma)$$

$\langle x_i(\sigma)x_j(\omega)\rangle = \sum_{x(\sigma)} P(x(\sigma))x_i(\sigma)x_j(\omega) = \sum_{x} P(x_i = \sigma, x_j = \omega) = P_{ij}(\sigma,\omega)$

- single-site and pair frequency counts

$$\overline{x_i(\sigma)} = \frac{1}{M}\sum_{m=1}^{M} x_i^m(\sigma) = f_i(\sigma)$$

$$\overline{x_i(\sigma)x_j(\omega)} = \frac{1}{M}\sum_{m=1}^{M} x_i^m(\sigma)x_j^m(\omega) = f_{ij}(\sigma,\omega)$$

Pairwise MEP Distribution on Categorical Variables

- Constraints

$$\sum_{x} P(x) = \sum_{x(\sigma)} P(x(\sigma)) = 1$$

$$P_i(\sigma) = f_i(\sigma),\ P_{ij}(\sigma,\omega) = f_{ij}(\sigma,\omega)$$

- Maximize $S = -\sum_{x} P(x) \ln P(x) = \sum_{x(\sigma)} P(x(\sigma)) \ln P(x(\sigma))$
- Lagrangian

$$\mathcal{L} = S + \alpha(\langle 1 \rangle - 1) + \sum_{i=1}^{L} \sum_{\sigma \in \Omega} \beta_i(\sigma)(P_i(\sigma) - f_i(\sigma))$$

$$+ \sum_{i,j=1}^{L} \sum_{\sigma,\omega \in \Omega} \gamma_{ij}(\sigma,\omega)(P_{ij}(\sigma,\omega) - f_{ij}(\sigma,\omega))$$

Pairwise MEP Distribution on Categorical Variables

Distribution

$$P(x(\sigma); \beta, \gamma)$$
$$= \frac{1}{Z} \exp\left(\sum_{i=1}^{L} \sum_{\sigma \in \Omega} \beta_i(\sigma) x_i(\sigma) + \sum_{i,j=1}^{L} \sum_{\sigma,\omega \in \Omega} \gamma_{ij}(\sigma, \omega) x_i(\sigma) x_j(\omega) \right)$$

Let

$$h_i(\sigma) := \beta_i(\sigma) + \gamma_{ii}(\sigma, \sigma), \; e_{ij}(\sigma, \omega) := 2\gamma_{ij}(\sigma, \omega)$$

Then

$$P(x_i, \ldots, x_L) \equiv \frac{1}{Z} \exp\left(\sum_{i=1}^{L} h_i(x_i) + \sum_{1 \le i < j \le L} e_{ij}(x_i, x_j) \right)$$

Gauge Fixing

$L(q-1)+\frac{L(L-1)}{2}(q-1)^2$ independent constraints compared to $Lq+\frac{L(L-1)}{2}q^2$ free parameters to be estimated.

$$1=\sum_{\sigma\in\Omega}P_i(\sigma),\ i=1,\ldots,L$$
$$P_i(\sigma)=\sum_{\omega\in\Omega}P_{ij}(\sigma,\omega),\ i,j=1,\ldots,L$$

Two guage fixing ways

- gap to zero: $e_{ij}(\sigma_q,\cdot)=e_{ij}(\cdot,\sigma_q)=0,\ h_i(\sigma_q)=0$
- zero-sum guage: $\sum_\sigma e_{ij}(\sigma,\omega)=\sum_\sigma e_{ij}(\omega',\sigma)=0,\ \sum_\sigma h_i(\sigma)=0$

Closed-Form Solution for Continuous Variables

rewirte the exponent of the pairwise maximum- entropy probability distribution

$$P(\boldsymbol{x};\beta,\tilde{\gamma}) = \frac{1}{Z}\exp\left(\beta^{\boldsymbol{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\tilde{\gamma}\boldsymbol{x}\right)$$
$$= \frac{1}{Z}\exp\left(\frac{1}{2}\beta^{\boldsymbol{T}}\tilde{\gamma}^{-\mathbf{1}}\beta - \frac{1}{2}(\boldsymbol{x}-\tilde{\gamma}^{-\mathbf{1}}\beta)^{T}\tilde{\gamma}(\boldsymbol{x}-\tilde{\gamma}^{-\mathbf{1}}\beta)\right)$$

where $\tilde{\gamma} := -2\gamma$, let $\boldsymbol{z} = (z_1,\ldots,z_L)^T := x - \tilde{\gamma}^{-1}\beta$, then

$$P(x) = \frac{1}{\tilde{Z}}\exp\left(-\frac{1}{2}(x-\tilde{\gamma}^{-1}\beta)^{T}\tilde{\gamma}(x-\tilde{\gamma}^{-1}\beta)\right) \equiv \frac{1}{\tilde{Z}}e^{-\frac{1}{2}z^{T}\tilde{\gamma}z}$$

normalization condition

$$1 = \int_x P(x)dx \equiv \frac{1}{\tilde{Z}}\int_x e^{-\frac{1}{2}z^{T}\tilde{\gamma}z}dz$$

Closed-Form Solution for Continuous Variables

use the point symmetry of the integrand

$$\int_z e^{-\frac{1}{2}z^T\tilde{\gamma}z} z_i dz = 0$$

then

$$\langle x_i \rangle = \int_x P(x) x_i dx \equiv \frac{1}{\tilde{Z}} \int_z e^{-\frac{1}{2}z^T\tilde{\gamma}z} \left(z_i + \sum_{j=1}^{L} (\tilde{\gamma}^{-1})_{ij}\beta_j \right) dz = \sum_{j=1}^{L} (\tilde{\gamma}^{-1})_{ij}\beta_j$$

$$\langle x_i x_j \rangle = \int_x P(x) x_i x_j dx \equiv \frac{1}{\tilde{Z}} \int_z e^{-\frac{1}{2}z^T\tilde{\gamma}z} (z_i - \langle x_i \rangle)(z_j - \langle x_j \rangle) dz = \langle z_i z_j \rangle + \langle x_i \rangle \langle x_j \rangle$$

so

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \equiv \langle z_i z_j \rangle$$

Closed-Form Solution for Continuous Variables

the term $\langle z_i z_j \rangle$ is solved using a spectral decomposition

$$\langle z_i z_j \rangle = \frac{1}{\tilde{Z}} \sum_{l,n=1}^{L} (v_l)_i (v_n)_j \int_y \exp\left(-\frac{1}{2}\sum_{k=1}^{L} \lambda_k y_k^2\right) y_l y_n dy$$

$$= \sum_{k=1}^{L} \frac{1}{\lambda_k} (\mathbf{v}_k)_i (\mathbf{v}_k)_j \equiv (\tilde{\gamma}^{-1})_{ij}$$

with $C_{ij} = (\tilde{\gamma}^{-1})_{ij}$, the Lagrange multipliers β and γ are

$$\beta = C^{-1}\langle x \rangle, \ \gamma = -\frac{1}{2}\tilde{\gamma} = -\frac{1}{2}C^{-1}$$

the real-valued maximum-entropy distribution

$$P(x; \langle x \rangle, C) = (2\pi)^{-L/2} \det(C)^{-1/2} \exp\left(-\frac{1}{2}(x - \langle x \rangle)^T C^{-1} (x - \langle x \rangle)\right)$$

Pair Interaction Strength

The pair interaction strength is evaluated by the already introduced partial correlation coefficient between x_i and x_j given the remaining variables $\{x_r\}_{r\in\{1,\ldots,L\}\setminus\{i,j\}}$.

$$\rho_{ij\cdot\{1,\ldots,L\}\setminus\{i,j\}} \equiv \frac{\gamma_{ij}}{\sqrt{\gamma_{ii}\gamma_{jj}}} = \begin{cases} -\dfrac{(C^{-1})_{ij}}{\sqrt{(C^{-1})_{ii}(C^{-1})_{jj}}} & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

Mean-Field Approximation

the empirical covariance matrix

$$\hat{C}_{ij}(\sigma,\omega) = f_{ij}(\sigma,\omega) - f_i(\sigma)f_j(\omega)$$

application of the closed-form solution for continuous variables to the categorical variables for $C^{-1}(\sigma,\omega) \approx \hat{C}^{-1}(\sigma,\omega)$ yields the so-called mean-field (MF) approximation

$$\gamma_{ij}^{MF}(\sigma,\omega) = -\frac{1}{2}(C^{-1})_{ij}(\sigma,\omega) \;\Rightarrow\; e_{ij}^{MF}(\sigma,\omega) = -(C^{-1})_{ij}(\sigma,\omega)$$

The same solution has been obtained by using a perturbation ansatz to solve the q-state Potts model termed (mean-field) Direct Coupling Analysis (DCA or mfDCA)

What is the Maximum-Likelihood

- A well-known approach to estimate the parameters of a model is maximum-likelihood inference.
- The likelihood is a scalar measure of how likely the model parameters are, given the observed data, and the maximum-likelihood solution denotes the parameter set maximizing the likelihood function.

Likelihood Function

To use the method of maximum likelihood, one first specifies the joint density function for all observations. For an independent and identically distributed sample, this joint density function is

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

whereas θ be the function's variable and allowed to vary freely; this function will be called the likelihood:

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

log-likelihood

$$\ln \mathcal{L}(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

average log-likelihood $\hat{\ell} = \frac{1}{n} ln\mathcal{L}$

Maximum-Likelihood Inference

For a pairwise model with parameters $h(\sigma)$ and $e(\sigma,\omega)$, the likelihood function of given observed data $x^1,\ldots,x^M \in \Omega^L$, which are assumed to be independent and identically distributed as

$$I(h(\sigma),e(\sigma,\omega)|x^1,\ldots,x^M) = \prod_{m=1}^{M} P(x^m;h(\sigma),e(\sigma,\omega))$$

$$\begin{aligned}\{h^{ML}(\sigma),e^{ML}(\sigma,\omega)\} &= \underset{h(\sigma),e(\sigma,\omega)}{\arg\max}\ I(h(\sigma),e(\sigma,\omega)) \\ &= \underset{h(\sigma),e(\sigma,\omega)}{\arg\min}\ -\ln I(h(\sigma),e(\sigma,\omega)\end{aligned}$$

Maximum-entropy distribution as model distribution

$$\ln I(h(\sigma),e(\sigma,\omega)) = \sum_{m=1}^{M} \ln P(x^m;h(\sigma),e(\sigma,\omega))$$

$$= -M\left[\ln Z - \sum_{i=1}^{L}\sum_{\sigma} h_i(\sigma)f_i(\sigma) - \sum_{1\le i<j\le L}\sum_{\sigma,\omega} e_{ij}(\sigma,\omega)f_{ij}(\sigma,\omega)\right]$$

Maximum-Likelihood Inference

Taking the derivatives

$$\frac{\partial}{\partial h_i(\sigma)} \ln I = -M \left[\frac{\partial}{\partial h_i(\sigma)} \ln Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} - f_i(\sigma) \right] = 0$$

$$\frac{\partial}{\partial e_{ij}(\sigma,\omega)} \ln I = -M \left[\frac{\partial}{\partial e_{ij}(\sigma,\omega)} \ln Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} - f_{ij}(\sigma,\omega) \right] = 0$$

Partial derivatives of the partition function

$$\frac{\partial}{\partial h_i(\sigma)} \ln Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} = \frac{1}{Z} \partial_{h_i(\sigma)} \quad Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} = P_i(\sigma; h(\sigma), e(\sigma,\omega))$$

$$\frac{\partial}{\partial e_{ij}(\sigma,\omega)} \ln Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} = \frac{1}{Z} \partial_{e_{ij}(\sigma,\omega)} Z \Big|_{\{h(\sigma), e(\sigma,\omega)\}} = P_{ij}(\sigma,\omega; h(\sigma), e(\sigma,$$

Maximum-Likelihood Inference

The maximizing parameters, $h^{ML}(\sigma)$ and $e^{ML}(\sigma,\omega)$ are those matching the distributions single and pair marginal probabilities with the empirical single and pair frequency counts,

$$P_i(\sigma; h^{ML}(\sigma), e^{ML}(\sigma,\omega)) = f_i(\sigma),\ P_{ij}(\sigma,\omega; h^{ML}(\sigma), e^{ML}(\sigma,\omega)) = f_{ij}(\sigma,\omega)$$

In other words, matching the moments of the pairwise maximum-entropy probability distribution to the given data is equivalent to maximum-likelihood fitting of an exponential family.

Pseudo-Likelihood Maximization

Besag introduced the pseudo-likelihood as approximation to the likelihood function in which the global partition function is replaced by computationally tractable local estimates.

In this approach, the probability of the m-th observation, x^m, is approximated by the product of the conditional probabilities of $x_r = x_r^m$ given observations in the remaining variables $\mathbf{x}_{\backslash r} := (x_1, \ldots, x_{r-1}, x_{r+1}, \ldots, x_L)^T \in \Omega^{L-1}$

$$P(x^m; h(\sigma), e(\sigma, \omega)) \simeq \prod_{r=1}^{L} P(x_r = x_r^m | \mathbf{x}_{\backslash r} = \mathbf{x}_{\backslash r}^m; h(\sigma), e(\sigma, \omega))$$

Pseudo-Likelihood Maximization

Each factor is of the following analytical form,

$$P(x_r = x_r^m | \mathbf{x}_{\setminus r} = \mathbf{x}_{\setminus r}^m; h(\sigma), e(\sigma,\omega)) = \frac{\exp\left(h_r(x_r^m) + \sum_{j \neq r} e_{rj}(x_r^m, x_j^m)\right)}{\sum_{\sigma} \exp\left(h_r(\sigma) + \sum_{j \neq r} e_{rj}(\sigma, x_j^m)\right)}$$

By this approximation, the loglikelihood becomes the pseudo-loglikelihood,

$$\ln l_{PL}(h(\sigma), e(\sigma,\omega)) := \sum_{m=1}^{M} \sum_{r=1}^{L} \ln P(x_r = x_r^m | \mathbf{x}_{\setminus r} = \mathbf{x}_{\setminus r}^m; h(\sigma), e(\sigma,\omega))$$

An ℓ^2-regularizer is added to select for small absolute values of the inferred parameters,

$$\{h^{PLM}(\sigma), e^{PLM}(\sigma,\omega)\} = \underset{h(\sigma), e(\sigma,\omega)}{\arg\min} \{ -\ln l_{PL}(h(\sigma), e(\sigma,\omega)) + \lambda_h \|h(\sigma)\|_2^2 + \lambda_e \|e(\sigma,\omega)\|_2^2 \}$$

Scoring Functions for Pairwise Interaction Strengths

- Direct information: $DI_{ij} = \sum_{\sigma,\omega\in\Omega} P_{ij}^{dir}(\sigma,\omega) \ln \frac{P_{ij}^{dir}(\sigma,\omega)}{f_i(\sigma)f_j(\omega)}$
- Frobenius norm: $\|e_{ij}\|_F = \left(\sum_{\sigma,\omega\in\Omega} e_{ij}(\sigma,\omega)^2 \right)^{1/2}$
- Average product correction: $APC - FN_{ij} = \|e_{ij}\|_F - \frac{\|e_{i\cdot}\|_F \|e_{\cdot j}\|_F}{\|e_{\cdot\cdot}\|_F}$ where

$$\|e_{i\cdot}\|_F := \frac{1}{L-1} \sum_{j=1}^{L} \|e_{ij}\|_F$$

$$\|e_{\cdot j}\|_F := \frac{1}{L-1} \sum_{i=1}^{L} \|e_{ij}\|_F$$

$$\|e_{\cdot\cdot}\|_F := \frac{1}{L(L-1)} \sum_{i,j=1}^{L} \|e_{ij}\|_F$$

Summary

Continuous variables

$$\mathbf{x} = (x_1, \ldots, x_L) \in \mathbb{R}^L$$

↓ multivariate Gaussian

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathrm{T}} C^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle)\right)$$

↓ **Parameter inference**

- closed-form solution

$$\gamma_{ij} = -\frac{1}{2}\left(C^{-1}\right)_{ij}$$

- sparse maximum-likelihood: $C^{-1} \approx C^{-1}_{1,\lambda}$

$$C^{-1}_{1,\lambda} = \underset{\substack{\Theta \text{ pos. definite,} \\ \text{symmetric}}}{\arg\max} \left\{\ln\det(\Theta) - \mathrm{tr}(\hat{C}\Theta) - \lambda\|\Theta\|_1\right\}$$

↓

partial correlation

$$\rho_{ij \cdot \{1,\ldots,L\}\setminus\{i,j\}} \equiv \frac{\gamma_{ij}}{\sqrt{\gamma_{ii}\gamma_{jj}}}$$

Categorical variables

$$\mathbf{x} = (x_1, \ldots, x_L) \in \Omega^L$$

↓ **Pairwise maximum-entropy distribution**

$$P(x_1, \ldots, x_L) = \frac{1}{Z} \exp\left(\sum_i h_i(x_i) + \sum_{i<j} e_{ij}(x_i, x_j)\right)$$

↓ **Parameter inference**

- mean-field (MF) ≈

$$e^{\mathrm{MF}}_{ij}(\sigma, \omega) = -\left(C^{-1}\right)_{ij}(\sigma, \omega)$$

- sparse maximum-likelihood (SML) ≈

$$e^{\mathrm{SML}}_{ij}(\sigma, \omega) = -\left(C^{-1}_{1,\lambda}\right)_{ij}(\sigma, \omega)$$

- pseudolikelihood maximization (PLM)

$$\left\{\boldsymbol{h}^{\mathrm{PLM}}(\boldsymbol{\sigma}), \boldsymbol{e}^{\mathrm{PLM}}(\boldsymbol{\sigma}, \boldsymbol{\omega})\right\} = \underset{\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})}{\arg\min} \left\{-\ln l_{\mathrm{PL}} + \lambda_{\boldsymbol{h}}\|\boldsymbol{h}\|_2^2 + \lambda_{\boldsymbol{e}}\|\boldsymbol{e}\|_2^2\right\}$$

↓ **Pair scoring functions**

- direct information

$$\mathrm{DI}_{ij} = \sum_{\sigma,\omega} P^{\mathrm{dir}}_{ij}(\sigma, \omega) \ln\left(\frac{P^{\mathrm{dir}}_{ij}(\sigma, \omega)}{f_i(\sigma) f_j(\omega)}\right)$$

- Frobenius norm

$$\|e_{ij}\|_{\mathrm{F}} = \left(\sum_{\sigma,\omega} e_{ij}(\sigma, \omega)^2\right)^{1/2}$$

- average product-corrected Frobenius norm

$$\text{APC-FN}_{ij} = \|e_{ij}\|_{\mathrm{F}} - \frac{\|e_{i\cdot}\|_{\mathrm{F}}\|e_{\cdot j}\|_{\mathrm{F}}}{\|e_{\cdot\cdot}\|_{\mathrm{F}}}$$