数理統計学チーム　　下平 英寿
Mathematical Statistics Team
Hidetoshi Shimodaira

Mathematical Statistics Team

Shimodaira Lab

Graduate School of Informatics, Kyoto University

Hidetoshi Shimodaira (Team Leader @ Kyoto Univ), Kazuki Fukui (RA), Akifumi Okuno (RA), Tetsuya Hada (RA), Masaaki Inoue (RA), Geewook Kim (RA), Thong Pham (Postdoctoral Researcher @ AIP), Tomoharu Iwata (Visiting Scientist @ NTT), Yoshikazu Terada (Visiting Scientist @ Osaka Univ), Shinpei Imori (Visiting Scientist @ Hiroshima Univ), Kei Hirose (Visiting Scientist @ Kyushu Univ)

Statistics and Machine Learning: Methodology and Applications

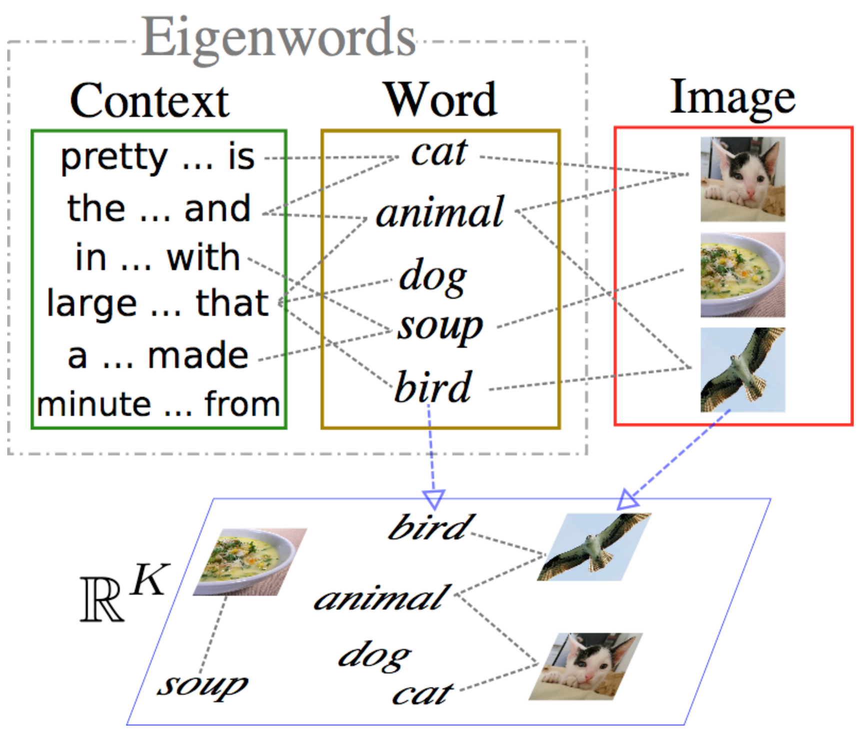
- Inductive inference, Resampling methods, Information geometry
- Generalization error under Missing, Covariate-shift, etc.
- Multi-modal data representation, Graph Embedding, and Multivariate Analysis
- Network growth mechanism (Preferential attachment and fitness)
- Phylogenetics, Gene expression (hierarchical clustering)
- Image, word embedding (search and reasoning)

Multimodal Eigenwords (Fukui, Oshikiri and Shimodaira, Textgraphs 2017)

- A multimodal word embedding that jointly embeds words and corresponding visual information
- We employed **Cross-Domain Matching Correlation Analysis** (CDMCA; Shimodaira 2016) for extending a CCA-based word embedding (Dhillon et al. 2015) to deal with complex associations

Our proposed method:

- Feature learning via graph embedding
- Feature vectors reflect both semantic and visual similarities
- The method enables **multimodal vector arithmetic** between images and words



Multimodal vector arithmetic:

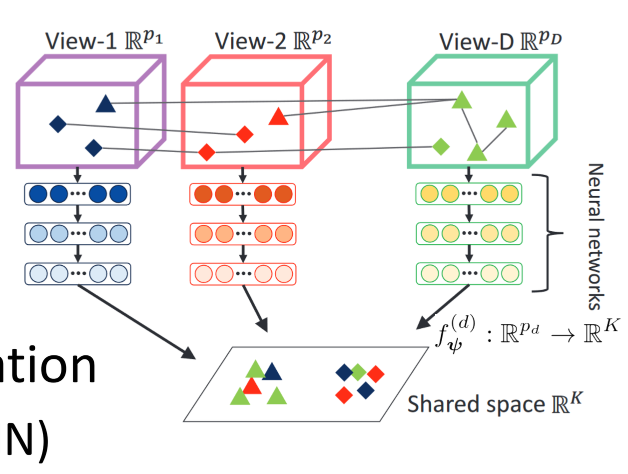
Probabilistic Multi-view Graph Embedding (PMvGE)

(Okuno, Hada and Shimodaira, arXiv:1802.04630)

Multi-view feature learning with many-to-many associations via neural networks for predicting new associations

$$w_{ij}^{(de)} \sim \mathrm{Po}\left(\alpha^{(de)} \exp\left(\langle f_\psi^{(d)}(\boldsymbol{x}_i^{(d)}), f_\psi^{(e)}(\boldsymbol{x}_j^{(e)})\rangle\right)\right)$$

Strength of association (≧0)　Neural network　data vector (in view-d)



Theorem (universal approximation):

Inner product + Neural networks

≈ Arbitrary similarity + Arbitrary transformation

(Mercer's theorem + universal approximation of NN)

Advantages:

(1) PMvGE with neural networks is proved to be **highly expressive**.

(2) PMvGE **approximately non-linearly generalizes CDMCA** (Shimodaira 2016, Neural Networks) which already generalizes various existing methods such as CCA, LPP, and Spectral graph embedding.

(3) Likelihood-based estimation of neural networks is **efficiently computed** by mini-batch Stochastic Gradient Descent.

Selective inference (Terada and Shimodaira, arXiv:1711.00949)

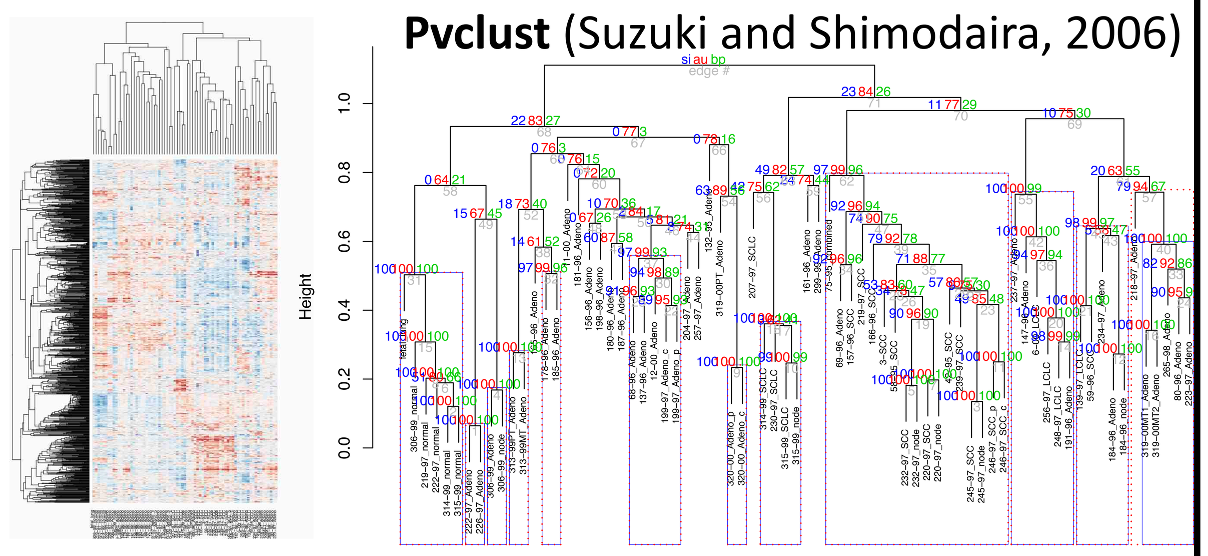
- **Motivation : Assessing the confidence of each obtained cluster**
- Consider approx. unbiased p-values as frequentist confidence measures
  Null hypothesis = obtained cluster is NOT true
- Example : lung data (73 tumors, 916 genes; Garber et al., 2001)

The hypothesis is tested **only for** the clusters appeared in the obtained tree

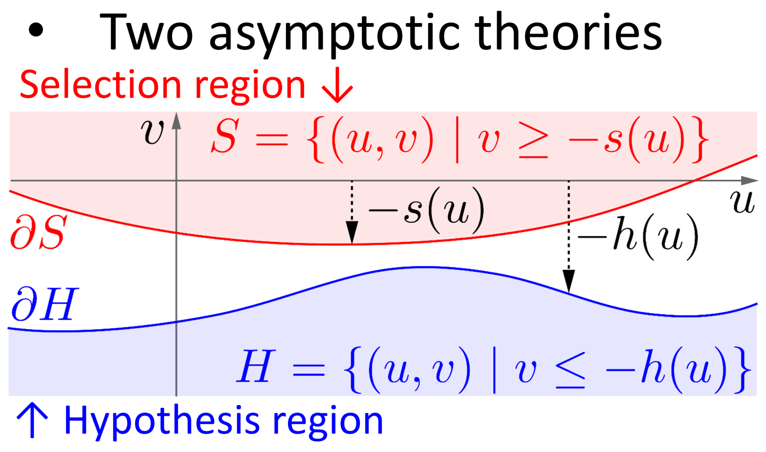
⬇

Need selective inference for pvclust

Pvclust (Suzuki and Shimodaira, 2006)



- Two asymptotic theories



(1) Large sample theory with "**nearly parallel surfaces**"

$$h(u) = h_0 + h_i u_i + h_{ij} u_i u_j + h_{ijk} u_i u_j u_k + \cdots, \quad n \to \infty$$

$h_0 = O(1)$, $h_i = O(n^{-1/2})$, $h_{ij} = O(n^{-1/2})$, $h_{ijk} = O(n^{-1})$, … Smooth surface

(u,v) are $O(\sqrt{n})$, but $h_0 = O(n^{1/2})$, $h_i = O(1)$ are multiplied by $O(n^{-1/2})$

(2) Asymptotic theory of "**nearly flat surfaces**"

$\sup_{u \in \mathbb{R}^m} |h(u)| = O(\lambda)$, $\int_{\mathbb{R}^m} |h(u)|\, du < \infty$, $\int_{\mathbb{R}^m} |\mathcal{F}h(\omega)|\, d\omega < \infty$

(Shimodaira 2008)　non-smooth surface　$\lambda \to 0$

Theorem (large sample theory):

Boundary surfaces of H and S are **smooth** and "**nearly parallel**",

⇒ The proposed p-value is **second order accurate with error** $O(n^{-1})$

Theorem (nearly flat surfaces):

Boundary surfaces are "**nearly flat**", approaching flat but **allowing non-smooth** such as cones and polyhedra

⇒ the proposed p-value is justified **as unbiased with error** $O(\lambda^2)$

Key point: Multiscale Bootstrap (Shimodaira, 2002; 2004)

✓ Low computational cost : $O(B)$

✓ **Double bootstrap** method has **same accuracy** but high comp. cost $O(B^2)$

Algorithm : Computing approx. unbiased selective *p*-values

1: Specify several $n' \in \mathbb{N}$ values, and set $\sigma^2 = n/n'$. Set the number of bootstrap replicates B, say, 1000.

2: For each n', perform bootstrap resampling to generate Y^* for B times and compute $\alpha_{\sigma^2}(H|y) = C_H/B$ and $\alpha_{\sigma^2}(S|y) = C_S/B$ by counting the frequencies $C_H = \#\{Y^* \in H\}$ and $C_S = \#\{Y^* \in S\}$. (We actually work on $\mathcal{X}^*_{n'}$ instead of Y^*.) Compute $\psi_{\sigma^2}(H|y) = \sigma\bar\Phi^{-1}(\alpha_{\sigma^2}(H|y))$ and $\psi_{\sigma^2}(S|y) = \sigma\bar\Phi^{-1}(\alpha_{\sigma^2}(S|y))$.

3: Estimate parameters $\beta_H(y)$ and $\beta_S(y)$ by fitting models　**Model fitting to psi**

$$\psi_{\sigma^2}(H|y) = \varphi_H(\sigma^2|\beta_H) \text{ and } \psi_{\sigma^2}(S|y) = \varphi_S(\sigma^2|\beta_S),$$

respectively. The parameter estimates are denoted as $\hat\beta_H(y)$ and $\hat\beta_S(y)$. If we have several candidate models, apply above to each and choose the best model based on AIC value.

4: Approximately unbiased p-values of selective inference (p_{SI}) and non-selective inference (p_{AU}) are computed by one of (A) and (B) below.

(A) Extrapolate $\psi_{\sigma^2}(H|y)$ and $\psi_{\sigma^2}(S|y)$ to $\sigma^2 = -1$ and 0, respectively, by

Extrapolation

$$z_H = \varphi_H(-1|\hat\beta_H(y)) \text{ and } z_S = \varphi_S(0|\hat\beta_S(y)),$$

and then compute p-values by

Selective *p*-value　**Non-selective p-value**

$$p_{\mathrm{SI}}(H|S,y) = \frac{\bar\Phi(z_H)}{\bar\Phi(z_H + z_S)} \text{ and } p_{\mathrm{AU}}(H|y) = \bar\Phi(z_H).$$

For non-smooth surfaces (such as cones and polyhedra)

(B) Specify $k \in \mathbb{N}$, $\sigma_0^2, \sigma_{-1}^2 > 0$ (e.g., $k = 3$ and $\sigma_{-1}^2 = \sigma_0^2 = 1$). Extrapolate $\psi_{\sigma^2}(H|y)$ and $\psi_{\sigma^2}(S|y)$ to $\sigma^2 = -1$ and 0, respectively, by

$$z_{H,k} = \varphi_{H,k}(-1|\hat\beta_H(y), \sigma_{-1}^2) \text{ and } z_{S,k} = \varphi_{S,k}(0|\hat\beta_S(y), \sigma_0^2),$$

where the Taylor polynomial approximation of φ_H at $\tau^2 > 0$ with k terms is:

$$\varphi_{H,k}(\sigma^2|\hat\beta_H(y), \tau^2) = \sum_{j=0}^{k-1} \frac{(\sigma^2 - \tau^2)^j}{j!} \frac{\partial^j \varphi_H(\sigma^2|\hat\beta_H(y))}{\partial(\sigma^2)^j}\bigg|_{\sigma^2=\tau^2},$$

and that of φ_S is defined similarly. Then compute p-values by

$$p_{\mathrm{SI},k}(H|S,y) = \frac{\bar\Phi(z_{H,k})}{\bar\Phi(z_{H,k} + z_{S,k})} \text{ and } p_{\mathrm{AU},k}(H|y) = \bar\Phi(z_{H,k}).$$

PAFit: an R Package for Estimating Preferential Attachment and Node Fitness in Temporal Complex Networks

(Pham, Sheridan and Shimodaira, arXiv:1704.06017)

- What drives the growth of real-world networks across diverse fields?
- Using two **interpretable and universal** mechanisms: (1) **Talent**: the intrinsic ability of a node to attract connections (fitness η_i) and (2) **Experience**: the preferential attachment (PA) function A_k governing the extent to which having more connections makes a node more/less attractive for forming new connections in the future.

Probability that node v_i gets a new edge at time t $= A_{k_i(t)}\eta_i$

Key finding: Although both talent and experience contribute to the growth process, the ratios of their contributions vary greatly.