Medical image analysis using high-dimensional information-theoretic criteria

Nicolas Rougon

ARTEMIS Department; CNRS UMR 8145
Institut TELECOM SudParis

Nicolas.Rougon@it-sudparis.eu

IPTA'2010 - Paris, 7-10 July 2010

Motivation

- Information theory provides powerful criteria for solving low-level problems in medical image understanding based on **statistical descriptions of visual features**
  - Registration & Motion estimation
    - frame-to-frame similarity $\longrightarrow$ informations
  - Segmentation & Tracking
    - object / background discrepancy $\longrightarrow$ divergences
    - frame / labelling similarity $\longrightarrow$ informations
  - Classification & Retrieval
    - Class separation $\longrightarrow$ divergences
    - Class consistency $\longrightarrow$ informations

Motivation

- **In medical imaging:** a *huge*, still expanding literature body on mutual information and its variations since 1995

Tutorial

MICCAI 2009 tutorial
Information theoretic similarity measures for image registration and segmentation
`ubimon.doc.ic.ac.uk/MICCAI09/a1882.html`

Review article

J.P.W. Pluim, J.B.A. Maintz and M.A. Viergever
Mutual-information-based registration of medical images: a survey
IEEE Transactions on Medical Imaging, 22(7):986-1004, July 2003

Motivation

- **Impact on the medical imaging domain**
  - In 2000, recognized by IEEE as *a landmark in the profession, with enduring importance and influence far beyond its peer*
  - In 2005, princeps papers recognized by ISI as one of the 10 most cited papers of the last decade published in Engineering [Collignon *et al.*, 1995] [Viola and Wells, 1995]

Motivation

- **Software implementations**
  - AIR (UCLA, Loni): www.loni.ucla.edu/Software/AIR
  - DROP (TU Muenchen): www.mrf-registration.net
  - Elastix (ISI, Utrecht): elastix.isi.uu.nl
  - FSL/FLIRT/FNIRT (FMRIB, Oxford): www.fmrib.ox.ac.uk/fsl
  - ImageFusion™: www.radionics.com/products/functional/imagefusion.shtml
  - IRTK (Imperial College, London): www.doc.ic.ac.uk/~dr/software
  - Slicer (BWH): www.slicer.org
  - SPM (University College, London): fil.ion.ucl.ac.uk/spm
  - ...

Motivation

- **In computer vision:** growing interest for information-theoretic region-based segmentation via curve evolution since 2001
  - Statistical region competition paradigm [Zhu and Yuille *et al.*, 1996]
  - Levelset framework
  - Optimization
    - standard variational calculus [Freedman, 2004] [Kim *et al.*, 2005] [Rougon *et al.*, 2006] [Michailovich *et al.*, 2007]
    - shape calculus [Aubert *et al.*, 2003] [Jehan-Besson *et al.*, 2003] [Herbulot *et al.*, 2006] [Boltz *et al.*, 2008]

Motivation

- **Strong assets**
  - Solid theoretical foundations
  - Blindness: no limitations on sensors / data
    In particular: deal with multimodal / uncommensurable data
  - Robustness against image degradations
    (noise, artifacts, local distortions)
  - No supervision
  - “Easy” implementation
  - Versatility and pervasiveness

Motivation

- **BUT**
  - Notoriously difficult to reliably estimate in high-dimensions
  - No straightforward generalization to the multivariate case

Restriction to bivariate information-theoretic measures acting on low-dimensional (usually scalar) random variables

- A stringent restriction
  - Local intensity information only is ambiguous
    $\longrightarrow$ spatial information (*i.e.* geometry) required
  - Non-local, complex image features can be informative
    $\longrightarrow$ *e.g.* wavelet coefficients, image tags, nonlocal textural features, etc.
  - Multiple image analysis ?

Motivations > Bibliography I

- G. Aubert,M. Barlaud, O. Faugeras and S. Jehan-Besson
  Image segmentation using active contours: Calculus of variations or shape gradients ?
  SIAM Journal on Applied Mathematics, 63(6):2128-2154, 2003
- S. Boltz, A. Herbulot, E. Debreuve, M. Barlaud and G. Aubert
  Motion and appearance nonparametric joint entropy for video segmentation
  International Journal of Computer Vision, 80(2):242-259, 2008
- A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens and G. Marchal
  Automated multimodality image registration using information theory
  Proceedings IPMI'95, pp. 263-274, June 1995
- D. Freedman and T. Zhang
  Active contours for tracking distributions
  IEEE Transactions on Image Processing, 13(4):518-526, 2004
- A. Herbulot, S. Jehan-Besson, M. Barlaud and G. Aubert
  Segmentation of vectorial image feaures using shape gradients and information measures
  Journal of Mathematical Imaging and Vision, 25(3):365-386, 2006
- S. Jehan-Besson, M. Barlaud and G. Aubert
  DREAM2S: Deformable Regions driven by an Eulerian Accurate Minimization Method for image and video segmentation
  International Journal of Computer Vision, 53(1):47-70, 2003

Motivations > Bibliography II

- J. Kim, J.W. III Fisher, A. Yezzi, M. Cetin and A.S. Willsky
  A nonparametric statistical method for image segmentation using information theory and curve evolution
  IEEE Transactions on Image Processing, 14(10):1486-1502, 2005

- F. Maes, A. Collignon, D. Vandermeulen, G. Marchal and P. Suetens
  Multimodality image registration by maximization of mutual information
  IEEE Transactions on Medical Imaging, 16(2):187-198, 1997

- O. Michailovich, Y. Rathi and A. Tannenbaum
  Image segmentation using active contours driven by the Bhattacharyya flow
  IEEE Transactions on Image Processing, 16(11):2787-2801, 2007

- N. Rougon, A. Discher and F. Prêteux
  Region-based statistical segmentation using informational active contours
  Proceedings SPIE Conference on Mathematics of Data/Image Pattern Recognition, Compression, and Encryption with Applications IX, Vol. 6315, pp. 63150I:1-12, August 2006

- P. Viola and W.M. Wells III
  Alignment by maximization of mutual information
  Proceedings ICCV'95, pp. 15-23, June 1995

- S.C. Zhu and A. Yuille
  Region competition: Unifying snakes, region growing and Bayes/MDL for multiband image segmentation
  IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(9):884-900, 1996

Outline

1. Foundations
2. kNN entropy estimators
3. Entropic graphs
4. Multivariate information measures

Outline

Notations

- X (continuous/discrete) random variable (RV) over some d-dimensional state space $\mathcal{X}$ with density p^X
- $X(\Omega)$ set of i.i.d. samples of X, denoted $X(s)$ or X_s, indexed by $s \in \Omega$
- Ω the image domain/grid, an image region or a point distribution
- $X|Y$ conditional RV with density $p^{X|Y}$
- (X, Y) joint RV with density $p^{X,Y}$

Differential entropy

Differential entropy

$$H(X) = -\int_{\mathcal{X}} p^X(x) \log p^X(x)\, dx$$

- Discrete framework: Shannon entropy. Does not converge to differential entropy in the continuous limit

- Joint entropy

$$H(X,Y) = -\int_{\mathcal{X}^2} p^{X,Y}(x,y) \log p^{X,Y}(x,y)\, dxdy$$

- Conditional entropy

$$H(X|Y) = -E_Y\left(\int_{\mathcal{X}} p^{X|Y}(x|y) \log p^{X|Y}(x|y)\, dx\right)$$

Differential entropy estimation

Notice that:

$$H(X) = -\mathbb{E}_X \left[\log p^X\right]$$

Ahmad-Lin entropy estimator [Ahmad and Lin, 1976]

If p^X is known

$$H^{\mathrm{AL}}(X) = -\frac{1}{|\Omega|} \sum_{s \in \Omega} \log p^X(X_s)$$

is a consistent estimator of differential entropy $H(X)$

- Holds if p^X is replaced by a consistent estimator $\hat{p}^X$ of p^X
- Ergodic approximation: expectation over the state space $\mathcal{X}$ is replaced by averaging over the sample set $X(\Omega)$
- Levelset segmentation: Ω is a free boundary domain and p^X is domain-dependent $\longrightarrow$ shape calculus required

Kullback-Leibler divergence

- **Discrepancy** of a RV X w.r.t. a reference RV Y

Kullback-Leibler divergence

$$D_{\mathrm{KL}}(X \parallel Y) = \int_{\mathcal{X}} p^X(x) \log \frac{p^X(x)}{p^Y(x)}\, dx$$

- Non-symmetric
- Interpretation as an entropy w.r.t. the measure $p^Y dx$

- Symmetrized KL divergence

$$D_{\mathrm{KL}}(X, Y) = \frac{1}{2}\left[D_{\mathrm{KL}}(X \parallel Y) + D_{\mathrm{KL}}(Y \parallel X)\right]$$

Kullback-Leibler divergence

- **Discrepancy** of a RV X w.r.t. a reference RV Y

Kullback-Leibler divergence

$$D_{\mathrm{KL}}(X \parallel Y) = \int_{\mathcal{X}} \frac{p^X(x)}{p^Y(x)} \log \frac{p^X(x)}{p^Y(x)} p^Y(x)\, dx$$

- Non-symmetric
- Interpretation as an entropy w.r.t. the measure $p^Y dx$

- Symmetrized KL divergence

$$D_{\mathrm{KL}}(X, Y) = \frac{1}{2}\Big[D_{\mathrm{KL}}(X \parallel Y) + D_{\mathrm{KL}}(Y \parallel X)\Big]$$

Kullback-Leibler divergence estimation

Notice that:

$$D_{\mathrm{KL}}(X \parallel Y) = \mathbb{E}_X \left[\log \frac{p^X}{p^Y} \right]$$

Ahmad-Lin KL divergence estimator

If p^X and p^Y are known,

$$D_{\mathrm{KL}}^{\mathrm{AL}}(X \parallel Y) = -\frac{1}{|\Omega|} \sum_{s \in \Omega} \log \frac{p^X(X_s)}{p^Y(X_s)}$$

is a consistent estimator of KL divergence $D_{\mathrm{KL}}(X \parallel Y)$.

- Holds if p^X, p^Y are replaced by consistent estimators $\hat{p}^X, \hat{p}^Y$

Mutual information

- **Similarity** between by 2 jointly observed RVs

Mutual information

Information shared by 2 jointly observed RVs

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

- Reduction of uncertainty on a RV brought by another one

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

- Deviation from independence

$$\begin{aligned} I(X,Y) &= \int_{\mathcal{X}^2} p^{X,Y}(x,y) \log \frac{p^{X,Y}(x,y)}{p^X(x)p^Y(y)}\, dxdy \\ &= D_{\mathrm{KL}}\left((X,Y) \parallel X \times Y\right) \end{aligned}$$

Normalized information measures

Normalized Mutual Information [Studholme *et al.*, 1999]

$$NMI(X, Y) = \frac{H(X) + H(Y)}{H(X, Y)} = 1 + \frac{I(X, Y)}{H(X, Y)}$$

Robust w.r.t. incomplete statistics

- Registration: partial overlap of image supports

Entropy Correlation Coefficient [Maes *et al.*, 1997]

$$ECC(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)} = 2\left(1 - \frac{1}{NMI(X, Y)}\right)$$

Also called Symmetric Uncertainty Coefficient

[Melbourne *et al.*, 2009]

Exclusive information

Exclusive information [Rougon *et al.*, 2003] [Zhang *et al.*, 2005]

Information contained solely in either of 2 jointly observed RVs

$$\begin{aligned} Z(X,Y) &= I(X,Y) - H(X,Y) \\ &= H(X) + H(Y) - 2H(X,Y) \\ &= -\left[H(X|Y) + H(Y|X)\right] \end{aligned}$$

- Faster optimization than MI

Shannon information theory

Entropy $H(X)$

$$-\mathbb{E}_X \left[f(p^X) \right]$$

▼

Divergence $D(X \parallel Y)$

$$-H(X \parallel Y)$$

▼

Mutual information $I(X,Y)$

$$D((X,Y) \parallel X \times Y)$$

- The information metrics f is the Kullback metrics $f(x) = \log(x)$
- This design can be generalized by considering other metrics f

  $\rightarrow$ Ali-Silvey class

f-entropy of a probability measure

μ, ν: probability measures (PMs). ν chosen as a reference.

Integral f-entropy

$$H_{f,\nu}(\mu) = -\int f\left(\frac{d\mu}{d\nu}\right) d\nu$$

f continuous, convex over $\mathbb{R}^+$

Non-integral f-entropy

$$H_{\psi,\nu}(\mu) = -\log \psi^{-1}\left(\int \frac{d\mu}{d\nu}\, \psi\left(\frac{d\mu}{d\nu}\right) d\nu\right)$$

ψ continuous, monotonic over $\mathbb{R}^+$

f, ψ : information metrics

f-entropy of a RV

- RVs are a special case: ν Borel $(d\nu = dx)$ and $d\mu = p^X(x)d\nu$

Integral f-entropy

$$H_f(X) = -\int f\left(p^X(x)\right)\,dx$$

f continuous, convex over $\mathbb{R}^+$

Non-integral f-entropy

$$H_{\psi,\nu}(\mu) = -\log \psi^{-1}\left(\int p^X(x)\,\psi\left(p^X(x)\right)\,dx\right)$$

ψ continuous, monotonic over $\mathbb{R}^+$

f, ψ : information metrics.

Some integral f-entropies

- $f(x) = x \log x \rightarrow$ Differential entropy
- $f(x) = \frac{x - x^\alpha}{\alpha - 1} \quad (\alpha \neq 1)$

Havrda-Charvát (Tsallis) entropy

$$H_\alpha(X) = \frac{1}{\alpha - 1}\left(1 - \int \left[p^X(x)\right]^\alpha dx\right)$$

Properties:

- $\lim_{\alpha \to 1} H_\alpha(X) = H(X)$
- Non-additivity: if X, Y independent

  $H_\alpha(X, Y) = H_\alpha(X) + H_\alpha(Y) + (1 - \alpha) H_\alpha(X) H_\alpha(Y)$

Some non-integral f-entropies

- $\psi(x) = \log x \longrightarrow$ Differential entropy

  The *only* both integral and non-integral f-entropy is $H(X)$
- $\psi(x) = x^{r-1} \quad (r \neq 1)$

Renyi entropy

$$H_r(X) = \frac{1}{1-r} \log \int \left[p^X(x)\right]^r dx$$

$r = \frac{1}{2} \longrightarrow$ Bhattacharyaa entropy

Properties:

- $\lim_{r \to 1} H_r(X) = H(X)$
- Additivity: if X, Y independent:

  $H_r(X, Y) = H_r(X) + H_r(Y)$

f-divergence of PMs

Integral / Non-integral f-divergence

$$D_f(\mu \parallel \nu) = -H_{f,\nu}(\mu)$$

f-divergence of RVs

- RVs are a special case: $d\mu = p^X(x)dx$ and $d\nu = p^Y(x)dx$

Integral f-divergence

$$D_f(X \parallel Y) = \int p^Y(x) f\left(\frac{p^X(x)}{p^Y(x)}\right) dx$$

Non-integral f-divergence

$$D_\psi(X \parallel Y) = \log \psi^{-1}\left(\int p^X(x)\psi\left(\frac{p^X(x)}{p^Y(x)}\right) dx\right)$$

- $D_f(X, Y) = \mathbb{E}_Y\left[f\left(\frac{p^X}{p^Y}\right)\right]$
- $D_\psi(X, Y) = \mathbb{E}_X\left[\psi\left(\frac{p^X}{p^Y}\right)\right]$

$\rightarrow$ Ahmad-Lin-like estimators

Some f-divergences

- $f(x) = x \log x$ and $\psi(x) = \log x \rightarrow$ KL divergence

- $f(x) = \frac{x - x^{\alpha}}{\alpha - 1} \quad (\alpha \neq 1)$

- $\psi(x) = x^{r-1} \quad (r \neq 1)$

I_α (Chernoff) divergence

$$D_\alpha(X \parallel Y) = \frac{1}{\alpha - 1}\left(\mathcal{I}_q^{X\parallel Y} - 1\right)$$

Renyi divergence

$$D_r(X \parallel Y) = \frac{1}{r-1} \log \mathcal{I}_q^{X\parallel Y}$$

where:

$$\mathcal{I}_q^{X\parallel Y} = \int_\Omega \left[p^X(x)\right]^q \left[p^Y(x)\right]^{1-q} dx$$

f-information of PMs

Integral / Non-integral f-information

$$I_f(\mu, \nu) = D_f\big((\mu, \nu) \parallel \mu \times \nu\big)$$

f-information of RVs

Integral f-information

$$I_f(X,Y) = \int p^X(x)p^Y(y)\, f\left(\frac{p^{X,Y}(x,y)}{p^X(x)p^Y(y)}\right) dxdy$$

Non-integral f-information

$$I_\psi(X,Y) = \log \psi^{-1}\left(\int p^{X,Y}(x,y)\, \psi\left(\frac{p^{X,Y}(x,y)}{p^X(x)p^Y(y)}\right) dxdy\right)$$

Note that $\frac{p^{X,Y}}{p^X p^Y}$ is the likelihood ratio between (X,Y) and $X \times Y$

- $I_f(X,Y) = \mathbb{E}_{X\times Y}\left[f\left(\frac{p^{X,Y}}{p^X p^Y}\right)\right]$
- $I_\psi(X,Y) \sim \mathbb{E}_{(X,Y)}\left[\psi\left(\frac{p^{X,Y}}{p^X p^Y}\right)\right]$

$\longrightarrow$ Ahmad-Lin-like estimators

Some f-informations

- $f(x) = x \log x$ and $\psi(x) = \log x \rightarrow$ Mutual information
- $f(x) = \frac{x-x^\alpha}{\alpha-1} \quad (\alpha \neq 1)$
- $\psi(x) = x^{r-1} \quad (r \neq 1)$

I_α information

$$I_\alpha(X, Y) = \frac{1}{\alpha - 1}\left(\mathcal{I}_\alpha^{X,Y} - 1\right)$$

Renyi information

$$I_r(X, Y) = \frac{1}{r - 1} \log \mathcal{I}_r^{X,Y}$$

where:

$$\mathcal{I}_q^{X,Y} = \int \left[p^{X,Y}(x,y)\right]^q \left[p^X(x)p^Y(y)\right]^{1-q} dxdy$$

Non-parametric density estimation

- Given a sample set $X(\Omega)$, the density $p^X(x)$ is the proportion $\frac{k(x)}{|\Omega|}$ of samples x per volume unit
- Consider open balls $\mathcal{B}_r(x)$ of radius r centered at x

▶ See [Beirlant *et al.*, 1997] for a review

Empirical estimates

- Pointwise estimates

Normalized histogram

$$p^X_{\mathrm{emp}}(x) = \frac{1}{|\Omega|} \int_\Omega \delta(x - X(s))\, ds$$

- Biaised (even in low dimension)

Kernel density estimates (KDE)

- Count the # of samples within a given d-dimensional kernel centered at sample x:

Generalized KDE

$$p^X_{\mathrm{KDE}}(x) = \frac{1}{|\Omega|} \int_\Omega K_\sigma \left(x - X(s)\right)\, ds$$

K_σ symmetric kernel (unit mass) with bandwidth $\sigma(\Omega, x, s)$

- Consistent but biaised
- Bias depends on σ

Fixed bandwidth KDE

Parzen-Rozenblatt estimator

$$p^X_{\text{parzen}}(x) = K_\sigma \star \frac{1}{|\Omega|} \int_\Omega \delta(x - X(s))\, ds$$

$\equiv$ smoothed normalized histogram

- Prototype: Gaussian kernel
- Biais critically depends on σ $\longrightarrow$ bandwidth estimation ?
  - Plug-in estimator

$$\sigma = 0.9 \min\left(\hat{\sigma}, \frac{\hat{p}}{1.34}\right) |\Omega|^{-\frac{1}{5}}$$

$\longrightarrow$ oversmoothing when p^X has several modes

$\hat{\sigma}$ empirical standard deviation of $X(\Omega)$
$\hat{p}$ interquartile range of $X(\Omega)$

  - Cross-validation techniques
- On discrete square grids, sensitivity to interpolation artefacts [Maes, 1998] [Pluim *et al.*, 2000]

Fixed bandwidth KDE

- Regionalized KDE
  - Improving estimation of minor modes of p^X which tend to be oversmoothed by global Parzen windowing

Localized Parzen-Rozenblatt estimator

$$p^X_{\text{parzen}}(x, s) = \frac{1}{|\Omega|} \int_\Omega W_{\sigma_s}(s' - s)\, K_\sigma(x - X(s'))\, ds'$$

K_σ symmetric kernel with bandwidth σ

W_{σ_s} symmetric spatial kernel with bandwidth σ_s
e.g. Gaussian kernel [Hermosillo, 2002] **or**
spatial window with finite extension σ_s
e.g. B-spline control patch [Loeckx *et al.*, 2007]

- Biais dependence on σ is reduced but persists

Fixed bandwidth KDE

- **Differentiable kernels** yield differentiable density estimates
  $\rightarrow$ variational optimization of information-theoretic measures

- Non rigid registration: $X^{\varphi} = X \circ \varphi$ with transform $\varphi \in \mathcal{T}$
  - $\dim \mathcal{T} = \infty \quad \rightarrow \varphi = \mathbb{I}d + \mathbf{u}$

$$\partial_{\mathbf{u}} p^{X^{\mathbf{u}}}(x) = \frac{1}{|\Omega|} \int_{\Omega} K'_{\sigma}(x - X(s)) \nabla X^{\mathbf{u}}(s)\, ds$$

[Hermosillo, 2002] [d'Agostino *et al.*, 2003] [Rougon *et al.*, 2005]

  - $\dim \mathcal{T} < \infty \quad \rightarrow \varphi(s;\Theta) = \mathbf{B}(s)\Theta$

$$\partial_{\Theta} p^{X^{\varphi}}(x) = \frac{1}{|\Omega|} \int_{\Omega} K'_{\sigma}(x - X(s)) \left[\nabla X^{\varphi}(s)^T \mathbf{B}(s)\right] ds$$

[Viola, 1995] [Rueckert *et al.*, 1999]

Fixed bandwidth KDE

- **Differentiable kernels** are not required for the variational optimization of information-theoretic measures over shape spaces
  - Active region segmentation: X^Γ with region boundary $\Gamma \in \mathcal{S}$
    Level set framework: $\Gamma = \{ s \in \Omega \mid \phi(\mathbf{s}) = 0 \} \quad \rightarrow \quad X^\phi$
    - $\dim \mathcal{S} = \infty$

$$\partial_\phi p^{X^\phi}(x) = \frac{1}{|\Omega|} \left(p^{X^\phi}(x) - K_\sigma(x - X^\phi(s)) \right) \delta(\phi(s))$$

[Freedman, 2004] [Kim *et al.*, 2005] [Herbulot *et al.*, 2006] [Rougon *et al.*, 2006] [Michailovich *et al.*, 2007]

    - $\dim \mathcal{S} < \infty \quad \rightarrow \quad \phi(s; \Theta) = B(s)\Theta$

$$\partial_\Theta p^{X^\phi}(x) = \frac{1}{|\Omega|} \left(p^{X^\phi}(x) - K_\sigma(x - X^\phi(s)) \right) B(s)^T \, \delta(\phi(s))$$

[Bernard *et al.*, 2009]

Fixed bandwidth KDE

- **Curse of dimensionality:** as $d \nearrow$
  - the sample set $X(\Omega)$ gets sparser in the state space $\mathcal{X}$
  - $|\Omega|$ must $\nearrow$ exponentially to ensure that enough samples fall within the kernel
  - if $|\Omega|$ is fixed, the performance of p^X_{parzen} rapidly degrades [Cacoullos, 1966] [Epanechnikov, 1969]

Parzen density estimator is unrealiable for $d > 3$

Adaptive bandwidth KDE

- **Bandwidth adaptation strategies** [Terrell and Scott, 1992]
  - Sample-dependent: $\sigma(X_s) \longrightarrow$ sample point estimators
    - $\sigma(X_s) \propto p^X(X_s)^{-1/2}$ [Abramson, 1982]
    - See also [Comaniciu, 2003]
  - Estimation point-dependent: $\sigma(x) \longrightarrow$ balloon estimators

Direct approach

- adapt bandwidth around x to include enough samples
- see for instance [Sain, 2002]

Dual approach

- set the # of samples per volume unit to a given value k
- compute the containing ball $\mathcal{B}_k^X(x)$

$\longrightarrow$ kNN density estimators

Foundations > Bibliography I

- I.S. Abramson
  On bandwidth variation in kernel estimates - a square root law
  Annals of Statistics, 10(4):1217-1223, 1982
- E. d'Agostino, F. Maes, D. Vandermeulen and P. Suetens
  A viscous fluid model for multimodal non-rigid image registration using mutual information
  Medical Image Analysis, 7(4):565-575, 2003
- I. Ahmad and P. Lin
  A nonparametric estimation of the entropy for absolutely continuous distributions
  IEEE Transactions on Information Theory, 36:688-692, 1976
- [Beirlant *et al.*, 1997] J. Beirlant, E. Dudewicz, L. Györfi and E.C. van der Meulen
  Nonparametric entropy estimation: An overview
  International Journal of the Mathematical Statistics Sciences, 6(1):17-39, 1997
- O. Bernard, D. Friboulet, P. Thevenaz and M. Unser
  Variational B-Spline level-set: A linear filtering approach for fast deformable model evolution
  IEEE Transactions on Image Processing, 18(6):1179-1191, 2009
- T. Cacoullos
  Estimation of a multivariate density
  Annals of the Institute of Statistical Mathematics, 18:179-189, 1966
- D. Comaniciu
  An algorithm for data-driven bandwidth selection
  IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(2):281-288, 2003

Foundations > Bibliography II

- V.K. Epanechnikov
  Non-parametric estimation of a multivariate probability density
  Theory of Probability and its Applications, 14:153-158, 1969
- G. Hermosillo
  Variational methods for multimodal image matching
  Ph.D. Thesis, Université de Nice-Sophia-Antipolis, May 2002
- D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen and P. Suetens
  Nonrigid image registration using conditional mutual information
  Proceedings IPMI'2007 - LNCS 4584, pp. 725-737, July 2007
- F. Maes
  Segmentation and registration of multimodal medical images: from theory, implementation and validation to a useful tool in clinical practice
  Ph.D. Thesis, KU Leuven, 1998
- A. Melbourne, D. Hawkes and D.A tkinson
  Image registration using uncertainty coefficients
  Proceedings ISBI'2009, pp. 951-954, June 2009
- J.P.W. Pluim, J.B.A. Maintz and M.A. Viergever
  Interpolation artefacts in mutual information-based image registration
  Computer Vision and Image Understanding, 77(2):211-232, 2000

Foundations > Bibliography III

- N. Rougon, C. Petitjean and F. Prêteux
  Variational non rigid image registration using exclusive f-information
  Proceedings ICIP'2003, pp. 703-706, September 2003

- N. Rougon, C. Petitjean, F. Prêteux, P. Cluzel and P. Grenier
  A non-rigid registration approach for quantifying myocardial contraction in tagged MRI using generalized information measures
  Medical Image Analysis, 9(4):565-575, 2005

- D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach and D.J. Hawkes
  Nonrigid registration using free-form deformations: application to breast MR images
  IEEE Transactions on Medical Imaging, 18(8):712-721, 1999

- S.R. Sain
  Multivariate locally adaptive density estimation
  Computational Statistics and Data Analysis, 39(2):165-186, 2002

- C. Studholme, D. Hill and D. Hawkes
  An overlap invariant entropy measure of 3D medical image alignment
  Pattern Recognition, 32(1):71-86, 1999

- G.R. Terrell and D.W. Scott
  Variable kernel density estimation
  Annals of Statistics, 20(3):1236-1265, 1992

- P. Viola and W.M. Wells III
  Alignment by maximization of mutual information

Foundations > Bibliography IV

- J. Zhang and A. Rangarajan
  Multimodality image registration using an extensible information metric and high dimensional histogramming
  Proceedings IPMI'2005 - LNCS 4584, pp. 725-737, July 2005

Outline

kth nearest neighbor (kNN) density estimate

kNN density estimator [Loftsgaarden and Quesenberry, 1965]

$$p^X_{\text{knn}}(x) = \frac{k}{|\Omega|\ V_d\ (\rho^X_k(x))^d}$$

$\rho^X_k(x)$ Euclidean distance from x to its kth nearest neighbor in the sample set $X(\Omega)\backslash\{x\}$

V_d Volume of the unit ball of $\mathbb{R}^d$: $V_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$

- $\mathcal{B}^X_k(x)$: d-dimensional ball of radius $\rho^X_k(x)$ centered at x
- Consistent but biaised (even in low dimension)
- Bias depends on k

→ Not used for density estimation **but** yields consistent and asymptotically unbiased estimators of entropy in high dimensions

kNN *vs.* KDE

Relationship between KDE and kNN density estimates

$p^X_{\text{knn}}(x)$ can be interpreted as a **balloon estimate** with uniform kernel $U_{\rho^X_k(x)}$ over the ball $\mathcal{B}^X_k(x)$ with bandwidth $\sigma(x) = \rho^X_k(x)$

From kNN density to kNN entropy estimation

1. Consider the Ahmad-Lin entropy estimator

$$H^{\mathrm{AL}}(X) = -\frac{1}{|\Omega|}\sum_{s\in\Omega} \log p^X(X_s)$$

2. Plug p^X_{knn}. This yields a geometric entropy estimator:

$$H^{\mathrm{knn}}(X) = \frac{1}{|\Omega|}\sum_{s\in\Omega} \log \left[\rho^X_k(X_s)\right]^d + \log \frac{V_d|\Omega|}{k}$$

- Consistent but biaised
- Bias depends on k
- Bias $\searrow$ as $d \nearrow$
- H^{knn} performs significantly better than fixed bandwidth KDE as $d \nearrow$

3. Fix the bias

kNN entropy estimators

Kosachenko-Leonenko-Goria entropy estimator [Goria *et al.*, 2005]

Under weak conditions on X, the following estimator is a consistent and asymptotically unbiased estimator of differential entropy:

$$H^{\text{knn}}(X) = \frac{1}{|\Omega|} \sum_{s \in \Omega} \log \left[\rho_k^X(X_s) \right]^d + c_k(d)$$

where $c_k(d) = \log\left(V_d(|\Omega| - 1)\right) - \psi(k)$ and $\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$

- Conditions on k:
  - $\lim_{|\Omega| \to \infty} k = \infty$ and $\lim_{|\Omega| \to \infty} \frac{k}{|\Omega|} = 0$ *e.g.* $k = \sqrt{|\Omega|}$
  - Image analysis: $k >$ # degrees of freedom
- Low sensitivity to k

f-entropy kNN estimators

- Tsallis / Renyi entropies involve the integral:

$$\mathcal{I}_q^X = \int \left(p^X(x)\right)^q dx = \mathbb{E}_X\left[\left(p^X\right)^{q-1}\right]$$

Lemma [Leonenko *et al.*, 2008]

Under weak conditions on X, the following estimator is a consistent and asymptotically unbiased kNN estimator of $\mathcal{I}_q^X$:

$$\widehat{\mathcal{I}_q^X} = \frac{1}{|\Omega|} \sum_{s \in \Omega} c_k(d) \left(\left[\rho_k^X(X_s)\right]^d \right)^{1-q}$$

where $c_k(d) = (V_d(|\Omega| - 1))^{1-q} \frac{\Gamma(k)}{\Gamma(k+1-q)}$

f-entropy kNN estimators

Tsallis / Renyi entropy kNN estimator [Leonenko *et al.*, 2008]

The following estimators are consistent and asymptitocally unbiased estimators of Tsallis / Renyi entropies, respectively:

$$H_\alpha^{\text{knn}}(X) = \frac{1}{\alpha - 1}\left(1 - \widehat{\mathcal{I}_\alpha^X}\right)$$

$$H_r^{\text{knn}}(X) = \frac{1}{1 - r}\log \widehat{\mathcal{I}_r^X}$$

KL divergence kNN estimators

KL divergence kNN estimator [Goria *et al.*, 2005]

Under weak conditions on X, the following estimator is a consistent and asymptotically unbiased estimator of KL divergence $D_{\mathrm{KL}}(X)$:

$$D_{\mathrm{KL}}^{\mathrm{knn}}(X \parallel Y) = \frac{1}{|\Omega|} \sum_{s \in \Omega} \log \left[\frac{\rho_k^Y(X_s)}{\rho_k^X(X_s)} \right]^d + c_k(d)$$

where $c_k(d) = \log \frac{|\Omega|}{|\Omega|-1}$

Note: This estimator involves the d-dimensional ball $\mathcal{B}_k^Y(X_s)$ comprising the k nearest samples of X_s in the sample set $Y(\Omega)$

- See also [Perez-Cruz, 2008] [Wang *et al.*, 2009]

f-divergence kNN estimators

- Chernoffs / Renyi divergences involve the integral:

$$\mathcal{I}_q^{X\|Y} = \int_\Omega [p^X(x)]^q [p^Y(x)]^{1-q}\, dx = \mathbb{E}_X\left[\left(\frac{p^X}{p^Y}\right)^{q-1}\right]$$

Lemma [Leonenko *et al.*, 2008]

Under weak conditions on X, the following estimator is a consistent and asymptotically unbiased kNN estimator of $\mathcal{I}_q^{X\|Y}$:

$$\widehat{\mathcal{I}_q^{X\|Y}} = \frac{1}{|\Omega|}\sum_{s\in\Omega} d_k(d)\left(\left[\frac{\rho_k^Y(X_s)}{\rho_k^X(X_s)}\right]^d\right)^{1-q}$$

where $d_k(d) = (V_d|\Omega|)^{1-q}\,\frac{\Gamma(k)}{\Gamma(k+1-q)}$

f-divergence kNN estimators

Chernoff / Renyi divergence kNN estimator [Goria *et al.*, 2005]

The following estimators are consistent and asymptitocally unbiased estimators of Chernoff / Renyi divergences, respectively:

$$D_\alpha^{\text{knn}}(X \parallel Y) = \frac{1}{\alpha - 1}\left(1 - \widehat{\mathcal{I}_\alpha^{X\|Y}}\right)$$

$$D_r^{\text{knn}}(X \parallel Y) = \frac{1}{1 - r}\log \widehat{\mathcal{I}_r^{X\|Y}}$$

Mutual information kNN estimators

Mutual information kNN estimator [Hamrouni and Rougon, 2010]

Under weak conditions on X, the following estimator is a consistent and asymptotically unbiased estimator of mutual information $I(X)$:

$$I^{\mathrm{knn}}(X,Y) = \frac{1}{|\Omega|}\sum_{s\in\Omega}\log\left[\frac{\rho_k^X(X_s)\,\rho_k^Y(Y_s)}{\left(\rho_k^{X,Y}(X_s),Y_s)\right)^2}\right]^d + c_k(d)$$

where $c_k(d) = \log\left(\frac{V_d^2}{V_{2d}}(|\Omega|-1)\right) - \psi(k)$

Note: I^{knn} involves the $2d$-dimensional ball $\mathcal{B}_k^{X,Y}(X_s,Y_s)$ containing the k nearest samples of (X_s,Y_s) in the sample set $X(\Omega)\times Y(\Omega)$

- Alternative kNN estimators: [Kraskov *et al.*, 2004] [Evans, 2007]

Optimizing kNN entropy estimators

- kNN estimators are not differentiable

  $\longrightarrow$ no direct variational optimization

- A plug-in approach is used [Boltz *et al.*, 2009]

  1. Compute the criterion derivative classically using differentiable Parzen estimators K_σ
  2. Replace K_σ by uniform kernel $U_{\rho_k^X(x)}$ over kNN ball $\mathcal{B}_k^X(x)$
     - integration over K_σ support $\rightarrow$ finite sum over $\mathcal{B}_k^X(x)$
     - K_σ derivative $\rightarrow$ indicator $1_{\mathcal{S}_k^X(x)}$ along kNN sphere $\mathcal{S}_k^X(x)$

Optimizing kNN entropy estimators

- Non rigid registration: $X^\varphi = X \circ \varphi$

Let $x = X^\varphi(s_0)$ and $y = X^\varphi(s)$

$$\partial_\varphi p^{X^\varphi}(x) = \frac{1}{|\Omega|} \int_\Omega K'_\sigma(x-y) \nabla X^\varphi(s)\, ds$$

translates into:

$$\partial_\varphi p^{X^\varphi}(x) = \frac{1}{|\Omega|} \sum_{s \in \Omega} U'_{\rho_k^X(x)}(x-y)\, \nabla \left(X^\varphi(s_0) - X^\varphi(s) \right)$$

with:

$$U'_{\rho_k^X(x)}(x-y) = \frac{-1}{V_d(\rho_k^X(x))^d} \frac{x-y}{|x-y|} 1_{\mathcal{S}_k^X(x)}(y)$$

Note: At most k points lie on the kNN sphere $\mathcal{S}_k^X(x)$

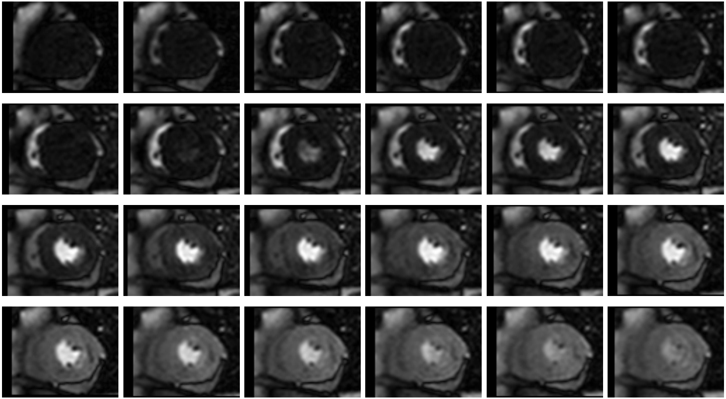
$\longrightarrow$ low computational cost

Efficient kNN search

- **Computational issues**
  - The numerical complexity of kNN-based estimation and optimization is dictated by the kNN search algorithm
  - Exact kNN search (kd-tree based) can be prohibitive
  - Approximate Nearest Neighbors (ANN) algorithm [Arya *et al.*, 1998]
    - $o(|\Omega|)$ and $o(d)$ complexity and memory usage
    - Free C^{++} ANN library: `www.cs.umd.edu/~mount/ANN`
    - GPU implementation in CUDA [Garcia *et al.*, 2008]

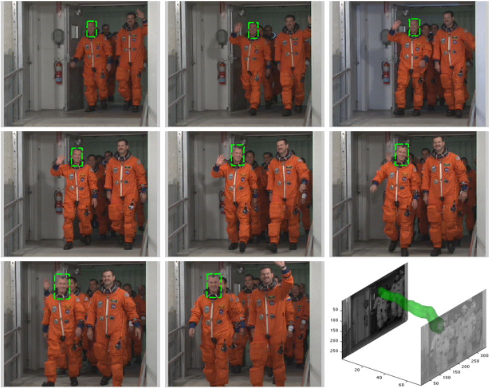
Example applications

- Groupwise rigid registration using MI kNN estimator [Hamrouni and Rougon *et al.*, 2010]



Example applications

- Tracking using KL divergence kNN estimator [Boltz *et al.*, 2009]



Source: E. Debreuve (`www.i3s.unice.fr/~debreuve`)

kNN entropy estimators > Bibliography I

- S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman and A. Wu
  An optimal algorithm for approximate nearest neighbor searching
  Journal of the ACM, 45(6):891-923, 1998
- S. Boltz, A. Herbulot, E. Debreuve, M. Barlaud and G. Aubert
  Motion and appearance nonparametric joint entropy for video segmentation
  International Journal of Computer Vision, 80(2):242-259, 2008
- S. Boltz, E. Debreuve and M. Barlaud
  High-dimensional statistical measure for region-of-interest tracking
  IEEE Transactions on Image Processing, 18(6):1266-1283,2009
- D. Evans
  A computationnally efficient estimator for mutual information
  Proceedings of the Royal Society A, 464:1203-1215, 2008
- V. Garcia, E. Debreuve and M. Barlaud
  Fast k nearest neighbor search using GPU
  Proceedings CVPR Workshop on Computer Vision on GPU (CVGPU), pp. 1-6, 2008
- M.N. Goria, N.N. Leonenko, V.V. Mergel and P.L. Novi Inverardi
  A new class of random vector entropy estimators and its applications in testing statistical hypotheses
  Journal of Nonparametric Statistics, 17(3):277-297, 2005

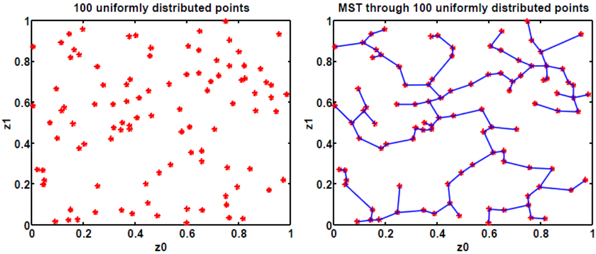
kNN entropy estimators > Bibliography II

- S. Hamrouni, N. Rougon and F. Prêteux
  Spatio-temporal registration of cardiac perfusion MRI exams using high-dimensional mutual information
  Proceedings IEEE EMBC'2010, September 2010.
- A. Kraskov, H. Stögbauer and P. Grassberger
  Estimating mutual information
  Physical Review E, 69(6):066138, 2004
- N.N. Leonenko, L. Pronzato and V. Savani
  A class of Rényi information estimators for multidimensional densities
  Annals of Statistics, 36(5):2153-2182, 2008
- D.O. Loftsgaarden and C.P. Quesenberry
  A nonparametric estimate of a multivariate density function
  Annals of Mathematical Statistics, 36(3):1049-1051, 1965
- [Perez-Cruz, 2008] F. Pérez-Cruz
  Estimation of information theoretic measures for continuous random variables
  Proceedings NIPS 2008, pp. 1257-1264, December 2008
- Q. Wang, S. Kulkarni and S. Verdú
  Divergence estimation for multidimensional densities via k-Nearest-Neighbor distances
  IEEE Transactions on Information Theory, 55(5):2392-2405, 2009

Outline

Spanning graphs

- Describing sample sets as spanning graphs
  [Readmond and Yukich, 1996] [Steele, 1997] [Yukich, 1998]
  - The sample set $X(\Omega)$ is viewed as the set of d-dimensional vertices $V = (X_s)_{s\in\Omega}$ of a graph $G = (V, E)$ with edges E
  - The graph set $\mathcal{G}$ is the set of spanning graphs (*i.e.* connected, acyclic) with vertices V



From [Neemuchwala and Hero, 2005]

Minimal spanning graphs

- Length of a spanning graph $G \in \mathcal{G}$:

$$\mathcal{L}_\gamma(G) = \sum_{e \in E} \|e\|^\gamma$$

$\|e\|$ Euclidean length of vertex e

γ power weighting constant $(0 < \gamma < d)$

- Minimal spanning graphs: graphs minimizing $\mathcal{L}_\gamma(G)$

  Denote by $L_\gamma(G)$ the minimal length:

$$L_\gamma(G) = \min_{G \in \mathcal{G}} \mathcal{L}_\gamma(G)$$

- Computing $L_\gamma(G)$ requires a combinatorial optimization over the set of spanning graphs $\mathcal{G}$

Entropic graphs

Theorem [Beardwood-Halton-Hammersley, 1959]

If $X(\Omega)$ is a set of i.i.d. samples drawn from a d-dimensional density p^X and $L_\gamma(G)$ is continuous quasi-additive, then:

$$\lim_{|\Omega|\to\infty} \frac{L_\gamma(G)}{|\Omega|^r} = \beta_{d,\gamma} \int \left(p^X(x)\right)^r dx \qquad \text{(a.s.)}$$

where $r = \frac{d-\gamma}{d}$ and $\beta_{d,\gamma}$ is a constant independent of p^X

Entropic graphs [Hero and Michel, 1999]

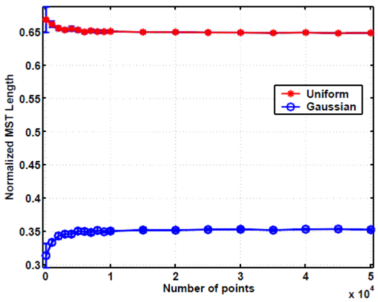
$$H_r^{\text{ent}}(X) = \frac{1}{1-r} \log \left(\frac{L_\gamma(G)}{|\Omega|^r} \right) - c$$

is a consistent and asymptotically unbiased estimator of Renyi entropy $H_r(X)$, where $c = \frac{1}{1-r} \log \beta_{d,\gamma}$ is a biais correction

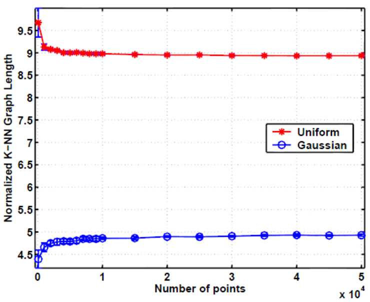
Computing entropic graphs

- $H_r(X)$ can be estimated from $L_\gamma(G)$ for any minimal spanning graph with continuous quasi-additive length functional

MST graph



kNN graph



From [Neemuchwala and Hero, 2005]

Computing entropic graphs

- Focus on efficiently computable graphs
  - Minimal Spanning Tree $\rightarrow$ kNN Kruskal algo.: $o(|E| \log |E|)$ [Hero and Michel, 1999] [Neemuchwala *et al.*, 2005b]
  - kNN graph $\rightarrow$ ANN algorithm: $o(|\Omega|)$ [Neemuchwala *et al.*, 2005]
- Once G is computed, $H_r(X)$ can be estimated for different values of $r \in [0,1]$ by adjusting $\gamma = d(1-r)$
- The topology of G is independent of γ in many cases
  $\rightarrow$ Single optimization required to estimate $H_r(X)$ for all r
- The biais correction c (which depends on graph type) can be estimated offline by Monte-Carlo simulation

Entropic graph estimators

- Entropic graphs yield geometric parameter-free estimators of:
  - Renyi divergence D_r and information I_r

    Recall: D_1 = KL divergence, I_1 = mutual information
  - Generalized Jensen divergence

$$\Delta_{r,p}(X, Y) = H_r(pX + qY) - [pH_r(X) + qH_r(Y)]$$

  - Generalized geometric-arithmetic mean divergence

$$D^{\mathrm{GA}}_{r,p}(X \parallel Y) = D_r(pX + qY \parallel X^p Y^q)$$

where $p \in [0,1]$ and $q = 1 - p$

Entropic graph estimators

Entropic graph estimator of Renyi information

$$I_r^{\text{ent}}(X,Y) = \frac{1}{r-1}\log\frac{1}{|\Omega|^r}\sum_{s\in\Omega}\sum_{j=1}^{k}\left(\frac{\rho_j^{X,Y}(X_s,Y_s)}{\rho_j^X(X_s)\,\rho_j^X(Y_s)}\right)^{2\gamma}$$

is a consistent and asymptotically unbiased estimator of Renyi information $I_r(X,Y)$

Note: $I_r^{\text{ent}}(X,Y)$ involve on 3 minimal graphs, with respective vertices $(X_s,Y_s)\in\mathbb{R}^{2d}$, $X_s\in\mathbb{R}^d$ and $Y_s\in\mathbb{R}^d$

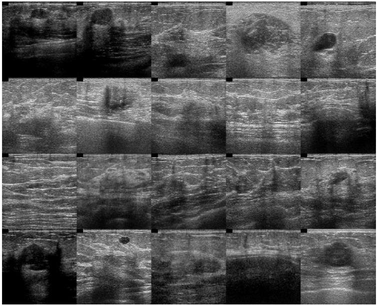
Optimizing entropic graph estimators

- Non rigid registration: $X^{\varphi} = X \circ \varphi$ with $\varphi(s;\Theta) = \mathbf{B}(s)\Theta$
  - Closed-form expression of the derivative of kNN entropic graph estimators w.r.t. Θ
    - Renyi joint-entropy [Oubel *et al.*, 2007]
    - Renyi information [Staring *et al.*, 2009]
  - Iterative variational optimization
    Assumption: invariance of graph topology for small $\delta\Theta$
    See also [Sabuncu and Ramadge, 2008]
  - Efficient stochastic gradient scheme [Staring *et al.*, 2099]
  - Much faster than finite difference or direct optimization
  - Yet computationally intensive

Example applications

- Breast ultrasound image registration
  [Neemuchwala *et al.*, 2007]

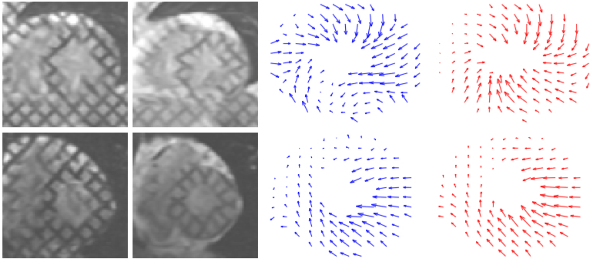
- Features: ICA coefficients ($d = 64$)
  Alternatives: wavelet coefficients, curvelet coefficients, image patches
- Transforms: isometries
- Generalized Jensen divergence
- kNN entropic graphs



Examples applications

- Cardiac tagged MR image registration
  [Oubel *et al.*, 2005] [Oubel *et al.*, 2006]

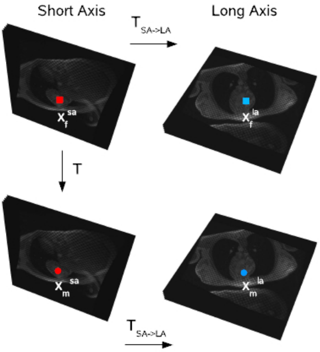
- Features: one-level Haar DWT ($d = 4$) or CWT ($d = 6$)
- Transforms: multilevel FFD
- Renyi information
- kNN entropic graphs



kNN entropic graph I_r *vs.* standard NMI

Examples applications

- Multiview cardiac tagged MR image registration [Oubel *et al.*, 2007]

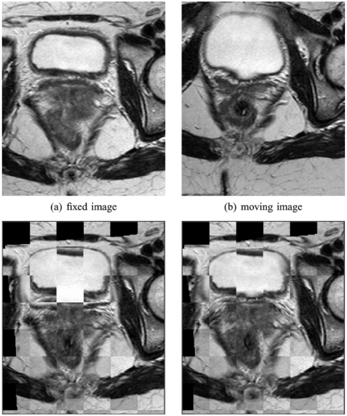


- Features: graylevel in SA and LA views ($d = 2$)
- Transforms: multilevel FFD
- Renyi joint entropy
- kNN entropic graphs
- Variational optimization

Examples applications

- Cervical MRI registration
  [Staring *et al.*, 2009]

- Features: intensity + 2nd-order 3D local image invariants at 2 scale levels ($d = 15$). PCA on image union $\rightarrow$ $d = 6$
- Transforms: FFD
- Renyi information
- kNN entropic graphs
- Variational optimization



(a) fixed image (b) moving image

(c) MI (d) α-MI

Entropic graphs > Bibliography I

- J. Beardwood, J.H. Halton and J.M. Hammersley
  The shortest path through many points
  Proceedings of the Cambridge Philosophical Society, 55(4):299-327, 1959
- A. Hero and O. Michel
  Asymptotic theory of greedy approximations to minimal k-point random graphs
  IEEE Transactions on Information Theory, 45(6):1921-1938, 1999
- H. Neemuchwala and A. Hero
  Entropic graphs for registration
  in Multi-Sensor Image Fusion and its Applications, R.S. Blum and Z. Liu (Eds), Marcel Dekker, New York, NY, chapter 6, pp. 185-235, 2005
- H. Neemuchwala, A. Hero and P. Carson
  Image matching using alpha-entropy measures and entropic graphs
  Signal Processing - Special Issue on Content-Based Image and Video Retrieval, 85(2):277-296, 2005
- H. Neemuchwala, A. Hero and P. Carson.
  Image registration methods in high-dimensional space
  International Journal of Imaging Systems and Technology, 16(5):130-145, 2007
- E. Oubel, C. Tobon-Gomez,A.O. Hero and A.F. Frangi
  Myocardial motion estimation in tagged MR sequences by using αMI-based non rigid registration
  Proceedings MICCAI'2005 - LNCS 3750, pp. 271-278, October 2005

Entropic graphs > Bibliography II

- E. Oubel, A.F. Frangi and A.O. Hero
  Complex wavelets for registration of tagged MRI sequences
  Proceedings ISBI'2006, pp. 622-625, April 2006
- E. Oubel, M. De Craene, M. Gazzola, A.O. Hero and A.F. Frangi
  Multiview registration of cardiac tagging sequences
  Proceedings ISBI'2007, pp. 388-391, April 2007
- M.R. Sabuncu and P.J. Ramadge
  Using spanning graphs for efficient image registration
  IEEE Transactions on Image Processing, 17(5):788-797, 2008
- M. Staring, U.A. van der Heide, S. Klein, M.A. Viergever and J.P.W. Pluim
  Registration of cervical MRI using multifeature mutual information
  IEEE Transactions on Medical Imaging, 28(9):1412-1421, 2009

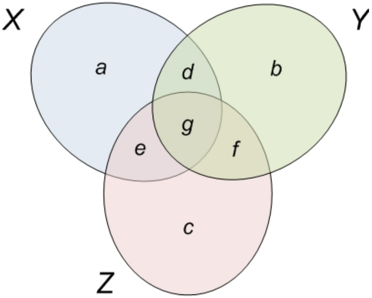
Outline

Multivariate information measures

- Multivariate information-theoretic measures are appealing for multiple image analysis (*e.g.* multi-view, multi-frame...)
  - groupwise registration
  - registration using class information
  - spatio-temporal segmentation
  - ...

But generalizing bivariate information measures to the multivariate case is not straighforward

Co-information



Co-information is the information shared by all RVs

$$\longrightarrow g$$

$$\begin{aligned} I(X,Y|Z) - I(X,Y) = & - H(X) - H(Y) - H(Z) \\ & + H(X,Y) + H(X,Z) + H(Y,Z) \\ & - H(X,Y,Z) \end{aligned}$$

Co-information

Denote $\mathcal{X} = \{X^1, X^2, \ldots, X^n\}$ and $\mathcal{Y} \subseteq \mathcal{X}$ a RV subset of $\mathcal{X}$

Co-information [McGill, 1954]

$$I(X^1, \ldots, X^n) = \sum_{\mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{X} \setminus \mathcal{Y}|} H(\mathcal{Y})$$
$$= I(\mathcal{X} \setminus X^i \,|\, X^i) - I(\mathcal{X} \setminus X^i) \qquad \forall X^i \in \mathcal{X}$$

- Also called information interaction
  [Bell, 2003] [Kludas *et al.*, 2009] [Zhou and Li, 2010]
- Symmetric
- Stable and unambiguous: adding new RVs does not change existing interactions (can only add new ones)

Co-information

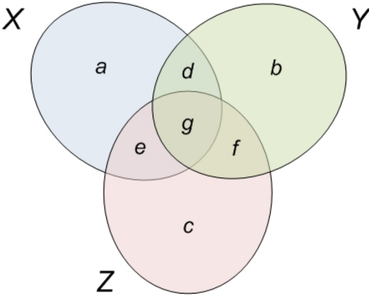
Denote $\mathcal{X} = \{X^1, X^2, \ldots, X^n\}$ and $\mathcal{Y} \subseteq \mathcal{X}$ a RV subset of $\mathcal{X}$

Co-information [McGill, 1954]

$$I(X^1, \ldots, X^n) = \sum_{\mathcal{Y} \subseteq \mathcal{X}} (-1)^{|\mathcal{X} \setminus \mathcal{Y}|} H(\mathcal{Y})$$
$$= I(\mathcal{X} \setminus X^i \,|\, X^i) - I(\mathcal{X} \setminus X^i) \qquad \forall X^i \in \mathcal{X}$$

- Can be positive (synergy) or negative (redundancy)
- Involve all p-dimensional ($p \leq d$) densities $p^{\mathcal{Y}}$ ($\mathcal{Y} \subseteq \mathcal{X}$)
  $\longrightarrow$ high estimation / optimization computational cost
- Often, too stringent similarity constraint
- **Currently unused in medical imaging**

Multi-information



Multi-information is the information shared by at least 2 RVs

$$\rightarrow \sum -a - b - c$$

$$H(X) + H(Y) + H(Z) - H(X, Y, Z)$$

Multi-information

Non rigid registration [Studholme *et al.*, 2006]

$\longrightarrow$ measurement of volume change in brain MRI

- X, Y: images
  Z: (overlapping) region labels
- Assumption: (X, Y) independent of Z
  $\equiv$ regionalized MI
- Regionalized KDE
- Transform: fluid
- Variational optimization

Multi-information

In the same spirit:

- Multi-channel MI [Holden *et al.*, 2004]

$$MI(X, X', Y, Y') = H(X, X') + H(X, Y') - H(X, X', Y, Y')$$

Note: X' (Y') are 1st- or 2nd-order scale-space derivatives of X (Y)

- Trivariate NMI [Papp *et al.*, 2009]

$$NMI(X, Y, Z) = \frac{H(X) + H(Y) + H(Z)}{H(X, Y, Z)}$$

Multi-information

Multi-information [Studeny and Vejnarova, 1998]

$$I(X^1, \ldots, X^n) = \sum_i H(X^i) - H(X^1, \ldots, X^n)$$
$$= \int p^{X^1,\ldots,X^n} \log \frac{p^{X^1,\ldots,X^n}}{p^{X^1} \ldots p^{X^n}} d\mathbf{x}$$

- Also called total correlation
- Symmetric, nonnegative
- Redundant information overcount
- Involves *nd*-dimensional joint density
- Straightforward generalization to arbitrary information metrics

Generalized multi-information

Integral generalized multi-information

$$I_f(X^1,\ldots,X^n) = \int p^{X^1}\ldots p^{X^n}\, f\left(\frac{p^{X^1,\ldots,X^n}}{p^{X^1}\ldots p^{X^n}}\right) d\mathbf{x}$$

f continuous, convex over $\mathbb{R}^+$

Non-integral generalized multi-information

$$I_\psi(X^1,\ldots,X^n) = \log\psi^{-1}\left(\int p^{X^1,\ldots,X^n}\,\psi\left(\frac{p^{X^1,\ldots,X^n}}{p^{X^1}\ldots p^{X^n}}\right) d\mathbf{x}\right)$$

ψ continuous, monotonic over $\mathbb{R}^+$

Generalized multi-information estimators

- Consistent and asymptitocally unbiased kNN entropy estimators / entropy graphs estimators are easily derived

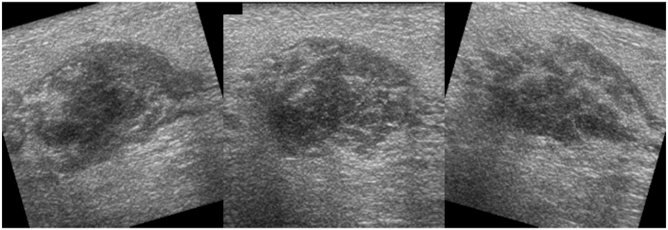
Entropic graph estimator of Renyi multi-information
[Neemuchwala *et al.*, 2007]

$$I_r^{\mathrm{ent}}(X^1,\ldots,X^n) = \frac{1}{r-1}\log\frac{1}{|\Omega|^r}\sum_{s\in\Omega}\sum_{j=1}^{k}\left(\frac{\rho_j^{X^1,\ldots,X^n}(x_s^1,\ldots,x_s^n)}{\rho_j^{X^1}(x_s^1)\ldots\rho_j^{X^n}(x_s^n)}\right)^{2\gamma}$$

Note: Involves $n+1$ minimal graphs

Example application

- Breast ultrasound multiple image registration
  [Neemuchwala *et al.*, 2007]



- Features: ICA coefficients ($d = 64$)
- Transforms: isometries
- Renyi multi-information
- kNN entropic graphs

Multivariate information measures > Bibliography I

- A.J. Bell
  The co-information lattice
  Proceedings ICA 2003, pp. 921-926, April 2003
- M. Holden, L.D. Griffin, N. Saeed and D.L.G. Hill
  Multi-channel mutual information using scale space
  Proceedings MICCAI'2004 - LNCS 3216, pp. 797-804, 2004
- J. Kludas, E. Bruno and S. Marchand-Maillet
  Can feature information interaction help for information fusion in multimedia problems ?
  Multimedia Tools and Applications, 42(1):57-71, 2009
- W.J. McGill
  Multivariate information transmission
  Psychometrika, 19(2):97-116, 1954
- L. Papp, M. Zuhayra and R. Koch
  Triple-modality normalized mutual information based medical image registration of cardiac PET/CT and SPECT images - Comparison with triple MI and dual NMI methods
  Proceedings Workshop Bildverarbeitung für die Medizin 2009, pp. 386-389, March 2009
- M. Studeny and J. Vejnarova
  The multi-information function as a tool for measuring stochastic dependence
  in Learning in Graphical Models, M.I. Jordan (Ed), Kluwer, Dordrecht, pp. 261-298, 1998

Multivariate information measures > Bibliography II

- C. Studholme, C. Dracapa, B. Iordanova and V. Cardenas
  Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change
  IEEE Transactions on Medical Imaging, 25(5):626-639, 2006
- Z-H. Zhou and N. Li
  Multi-information ensemble diversity
  Proceedings MCS'2010 - LNCS 5997, pp. 134-144, April 2010

Medical image analysis using high-dimensional information-theoretic criteria

Nicolas Rougon

ARTEMIS Department; CNRS UMR 8145
Institut TELECOM SudParis

Nicolas.Rougon@it-sudparis.eu

IPTA'2010 - Paris, 7-10 July 2010