Gaussian Information Bottleneck

Gal Chechik http://robotics.stanford.edu/~gal
Amir Globerson, Naftali Tishby, Yair Weiss

Full NIPS paper at
http://robotics.stanford.edu/~gal/ps_files/chechik_nips2003.pdf

Preview

Information Bottleneck/distortion

- Was mainly studied in the discrete case (categorical variables)
- Solutions are characterized analytically by self consistent equations, but obtained numerically (local maxima).

We describe a complete analytic solution for the Gaussian case.

- Reveal the connection with known statistical methods
- Analytic characterization of the compression-information tradeoff curve

IB with continuous variables

- Extracting relevant features of continuous variables:
  - Result of analogue measurements: *gene expression vs heat or chemical conditions*
  - Continuous low dim manifolds: *face expressions, postures*

- IB formulation is not limited to discrete variables

$$\min_{p(t|x),\, Y \to X \to T} L = I(X;T) - \beta I(T;Y)$$

  - Use continuous mutual information and entropies

$$h(X) = -\int f(x) \log f(x) dx$$

  - In our case the problem contains an inherent scale, which makes all quantities well defined.

- The general continuous solutions are characterized by the self consistent equations
  - but this case is very difficult to solve

Gaussian IB

Definition:

Let X and Y be jointly Gaussian (multivariate)

Search for another variable T that minimizes

$$\min_{T} L = I(X;T) - \beta I(T;Y)$$

The optimal T is jointly Gaussian with X and Y.

Equivalent formulation:

T can always be represented as

$$T = A X + \xi \qquad (\text{with } \xi \sim N(0,\Sigma_\xi),\ A = \Sigma tx\ \Sigma x^{-1})$$

Minimize L over the A and ξ.

The goal:

Find optimum for all beta values

Before we start:

What types of solutions do we expect?

- Second order correlation only:

  probably eigenvectors of some correlation matrices…

  - but which?

- The parameter β effects the model complexity:

  Probably deterine the number of eigen vectors and their scale…

  - but how?

Derive the solution

Using the entropy of a Gaussian $h(X)=\frac{1}{2}\log\left((2\pi e)^d\,|\Sigma_x|\right)$ we write the target function

$$L=(1-\beta)\log\left|A\Sigma_x A^T+\Sigma_\xi\right|-\log\left|\Sigma_\xi\right|+\beta\log\left|A\Sigma_{x|y}A^T+\Sigma_\xi\right|$$

Although L is a function of A and Σ_ξ, there is always an equivalent solution A' with spherized noise Σ_ξ=I, that lead to same L value.

Differentiate L w.r.t. A (matrix derivatives)

$$\frac{dL}{dA}=(1-\beta)\left(A\Sigma_x A^T+I\right)^{-1}2A\Sigma_x+\beta(A\Sigma_{x|y}A^T+I)2A\Sigma_{x|y}$$

The scalar T case

- When A is a single row vector

$$0 = \text{scalar}\;)^{-1}\, 2A\Sigma_x + \text{scalar}\;)\,2A\Sigma_{x|y}$$

can be written as

$$\underbrace{\left(\frac{\beta-1}{\beta}\right)\left(\frac{A\Sigma_{x|y}A^T+I}{A\Sigma_x A^T+I}\right)} A = A\underbrace{\left(\Sigma_{x|y}\Sigma_x^{-1}\right)}$$

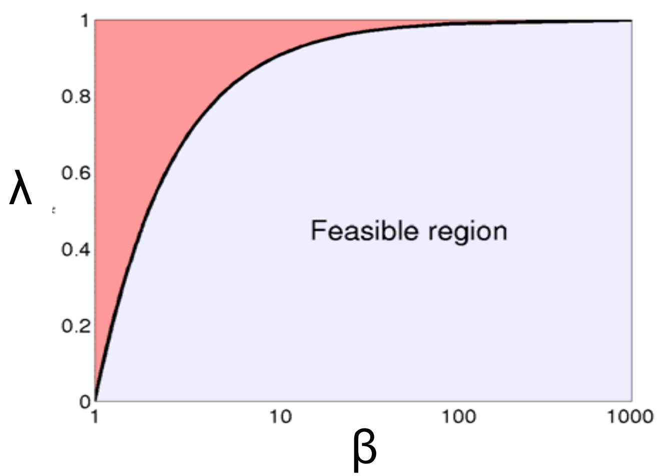
$$\lambda \quad A = A \quad M$$

- This has two types of solution:
  - A degenerates to zero
  - A is an eigenvector of $M=\Sigma_{x|y}\,\Sigma_x^{-1}$

The eigenvector solution…

1) Is feasible only if:

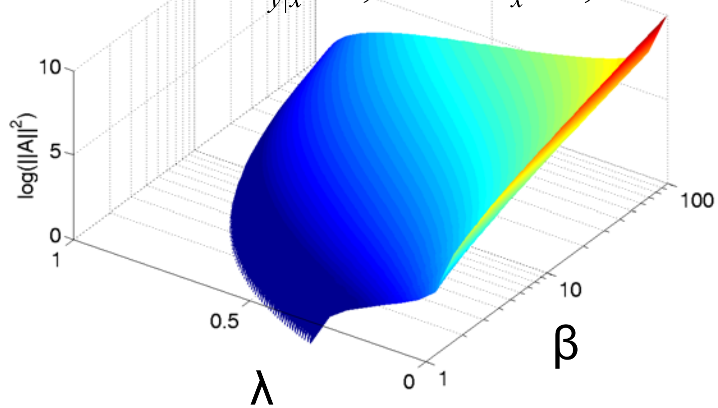
$$\beta \geq (1-\lambda)^{-1}$$



2) Has norm:

$$\frac{1}{\lambda r}(\beta(1-\lambda)-1)$$

$$\lambda = A\Sigma_{y|x}A^T; r = A\Sigma_x A^T;$$

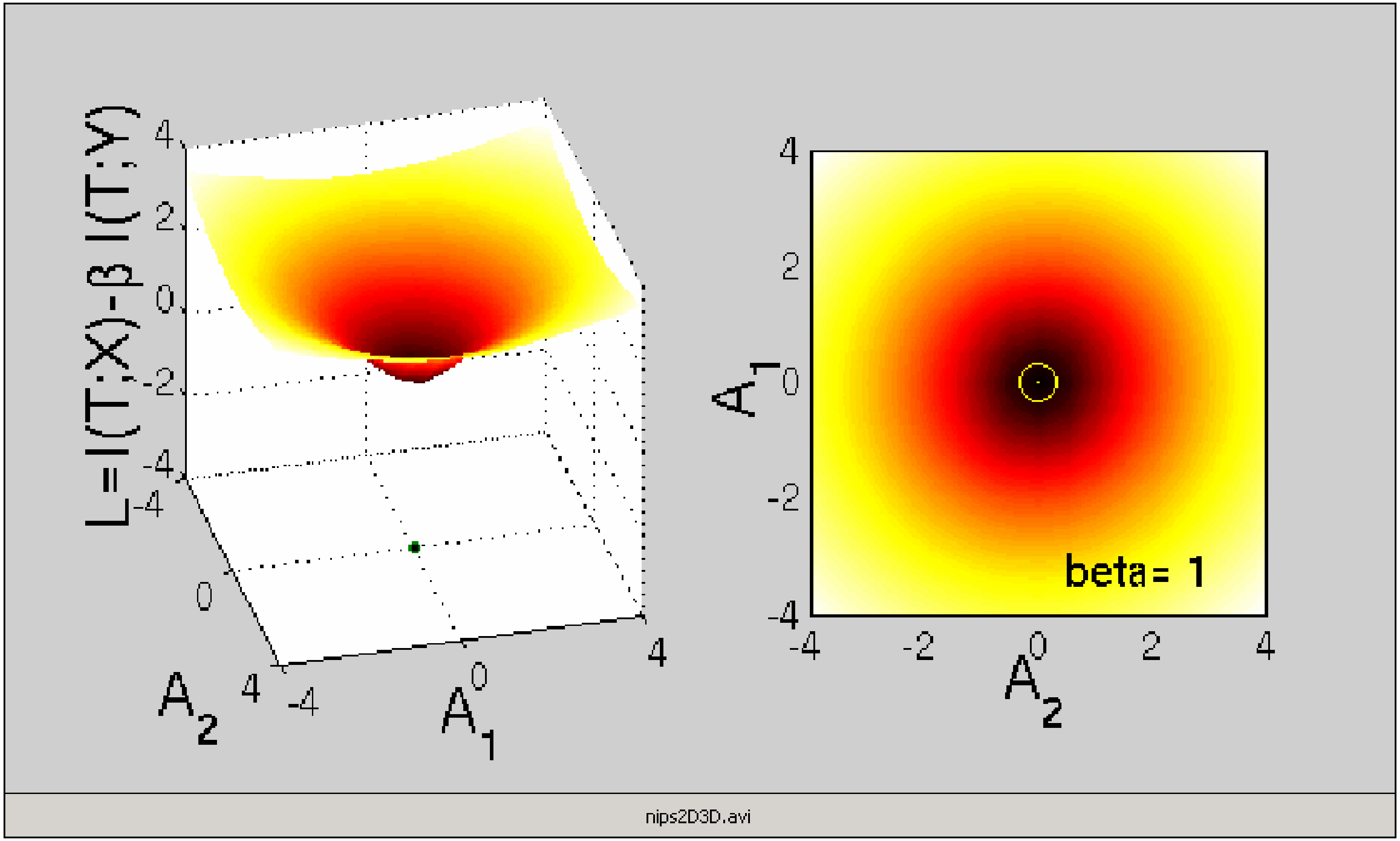


- The optimum is obtained with the smallest eigenvalues
- Conclusion:

$$A = \alpha v_1 \text{ with } \alpha = \begin{cases} \frac{\beta(1\text{-}\lambda_1)\text{-}1}{\lambda_1 r} & \beta > (1\text{-}\lambda_1)^{-1} \\ 0 & \text{other wise} \end{cases}$$

The effect of β in the scalar case

- Plot the surface of the target L as a function of *A*, when *A* is a 1x2 vector:



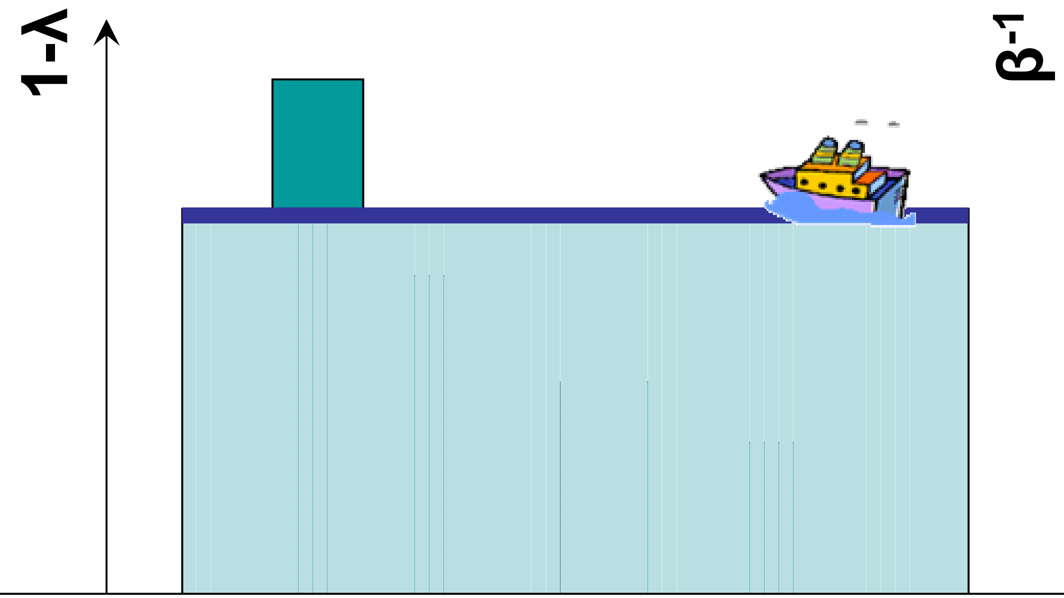
The multivariate case

- Back to
$$\frac{dL}{dA} = (1-\beta)\left(A\Sigma_x A^T + I\right)^{-1} 2A\Sigma_x + \beta(A\Sigma_{x|y} A^T + I)2A\Sigma_{x|y}$$
- The rows of A are in the span of several eigenvectors. An optimal solution is achieved with the smallest eigenvectors.
- As β increases A goes through a series of transitions, each adding another eigen vector

$$A = \begin{cases} [0^T;\ldots;0^T] & 0 < \beta < \beta_i^c \\ [\alpha_1 \mathbf{v}_1^T; 0^T;\ldots;0^T] & \beta_1^c < \beta < \beta_2^c \\ [\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; 0^T;\ldots;0^T] & \beta_2^c < \beta < \beta_3^c \\ \vdots & \vdots \end{cases} \quad \text{with} \quad \begin{array}{l} \alpha_i = \frac{\beta(1-\lambda_i)-1}{\lambda_i} \\ r_i = \mathbf{v}_i^T \Sigma_x \mathbf{v}_i^T \\ \beta_i^c = (1-\lambda_i)^{-1} \end{array}$$

The multivariate case

- Reverse water filling effect: increasing complexity causes a series of phase transitions



$$A = \begin{cases} [0^T;\ldots;0^T] & 0 < \beta < \beta_i^c \\ [\alpha_1 \mathbf{v}_1^T;0^T;\ldots;0^T] & \beta_1^c < \beta < \beta_2^c \\ [\alpha_1 \mathbf{v}_1^T;\alpha_2 \mathbf{v}_2^T;0^T;\ldots;0^T] & \beta_2^c < \beta < \beta_3^c \\ \vdots & \vdots \end{cases} \text{with} \begin{array}{l} \alpha_i = \frac{\beta(1-\lambda_i)-1}{\lambda_i} \\ r_i = \mathbf{v}_i^T \Sigma_x \mathbf{v}_i^T \\ \beta_i^c = (1-\lambda_i)^{-1} \end{array}$$

The information curve

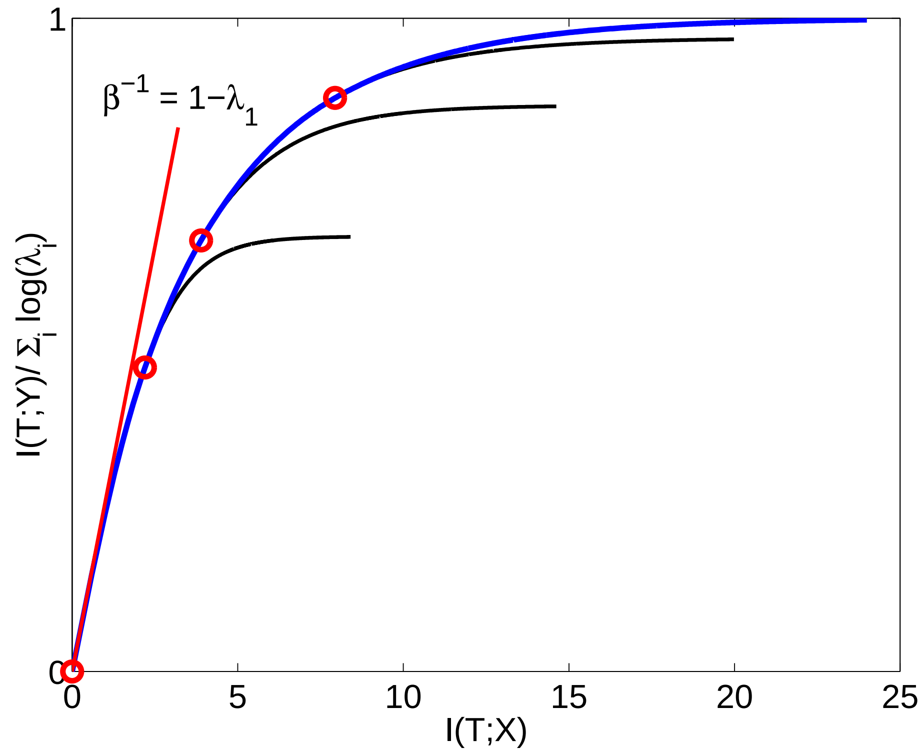
- Can be calculated analytically, as a function of the eigenvalue spectrum

$$I(T;Y) = I(T;X) - \frac{n_I}{2}\log\left(\prod_{i=1}^{n_I}(1-\lambda_i)^{-n_I} + \exp\left(\frac{2I(T;X)}{n_I}\right)\prod_{i=1}^{n_I}(\lambda_i)^{-n_I}\right)$$

n_I is the number of components required to obtain I(T;X).

- The curve is made of segments

- The tangent at critical points equals 1-λ



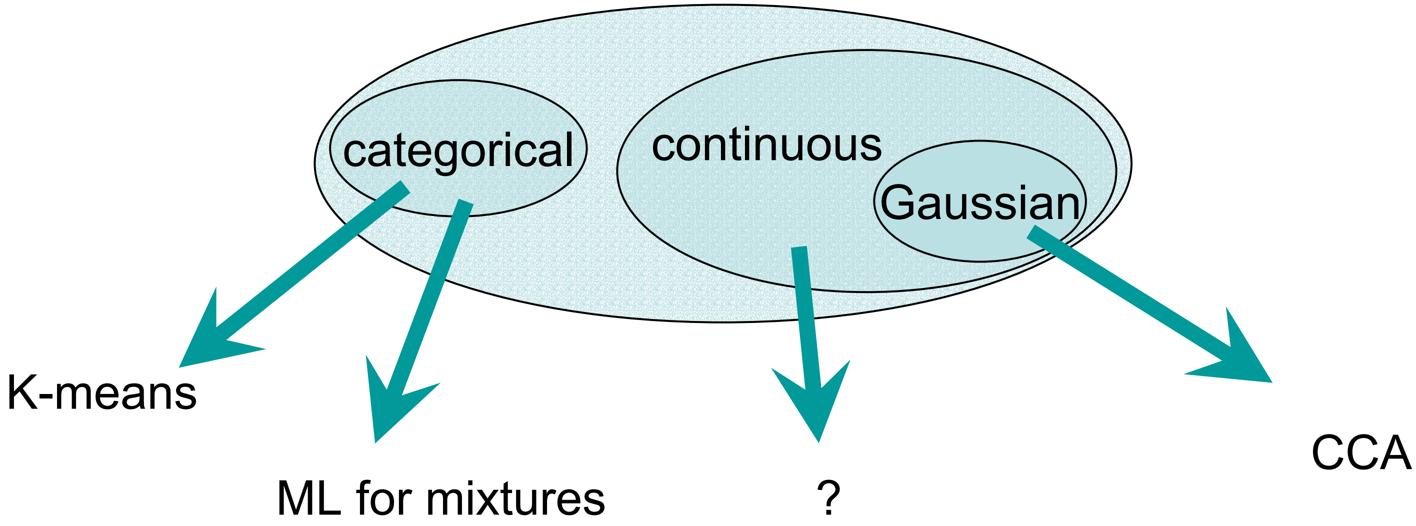
Relation to Canonical correlation analysis

- The eigenvectors used in GIB are also used in CCA [Hotelling 1935].
- Given two Gaussian variables $\{X,Y\}$, CCA finds basis vectors for both X and Y that maximize correlation on their projections (i.e. bases for which the correlation matrix is diagonal with maximal correlations on the diagonal)

  - GIB controls the level of compression, providing both the number and scale of the vectors (per β).
  - CCA is a normalized measure, invariant to rescaling of the projection.

What did we gain?

Specific cases coincide with known problems:



A unified approach allows to reuse algorithms and proofs.

What did we gain ?

Revealed connection allows to gain from both fields:

- CCA => GIB
  - Statistical significance for sampled distributions

    Slonim and Weiss showed a connection between the β and the number of samples. What will be the relation here?
- GIB => CCA
  - CCA as a special case of a generic optimization principle
  - Generalizations of IB, lead to generalizations of CCA
    - Multivariate IB => Multivariate CCA
    - IB with side information => CCA with side information (as in oriented PCA) generalized eigen value problems.
  - Iterative algorithms (avoid the costly calculation of covariance matrices)

Summary

- We solve analytically the IB problem for Gaussian variables
- Solutions described in terms of eigenvectors of a normalized cross correlation matrix, and its norm as a function of the regularization parameter beta.
- Solutions are related to canonical correlation analysis
- Possible extensions to general exponential families and multivariate CCA.