Copula Mixture Model for Dependency-seeking Clustering

Melanie Rey, Volker Roth

Department of Mathematics and Computer Science, University of Basel, Switzerland

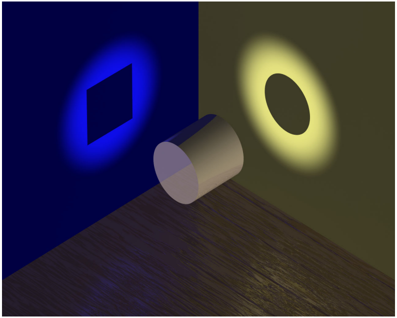
December 16, 2011

Dependency-seeking Clustering

- Clustering co-occurring samples from different data sources called views:

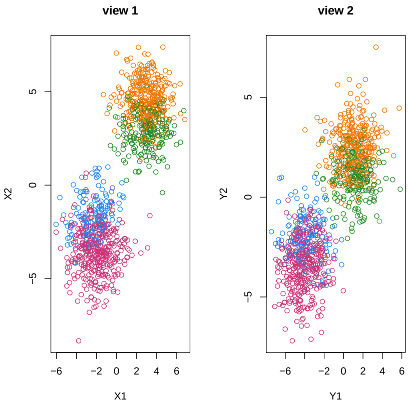
$$
\begin{array}{ccc}
(x^1, \ldots, x^p)_1 & \longleftrightarrow & (y^1, \ldots, y^q)_1 \\
\vdots & & \vdots \\
(x^1, \ldots, x^p)_n & \longleftrightarrow & (y^1, \ldots, y^q)_n
\end{array}
$$

- The aim is to cluster the points according to their between-views dependence structure.



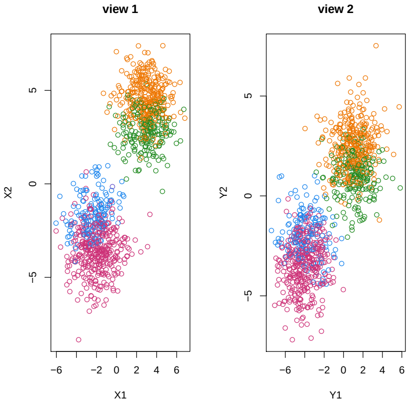
Dependency-seeking Clustering

$$(1) + (2) : cor(X_2, Y_2) = 0.45.$$

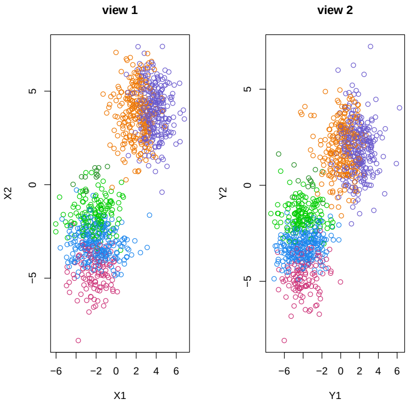


Dependency-seeking Clustering

$(1) + (2) : cor(X_2, Y_2) = 0.45.$



$(1) : cor(X_2, Y_2) = 0.8,$
$(2) : cor(X_1, Y_1) = 0.45.$



Probabilistic CCA

- The probabilistic interpretation of CCA [Bach, 2005]:

$$Z \sim \mathcal{N}_d\,(0, I_d),$$
$$(X, Y)\,|Z \sim \mathcal{N}_{p+q}\,(WZ + \mu, \Psi)\,,$$

where Ψ has a block-diagonal form:

$$\Psi = \begin{pmatrix} \Psi_x & 0 \\ 0 & \Psi_y \end{pmatrix}.$$

Dependency-seeking Clustering

- Probabilistic dependency-seeking clustering [Klami, 2006]:

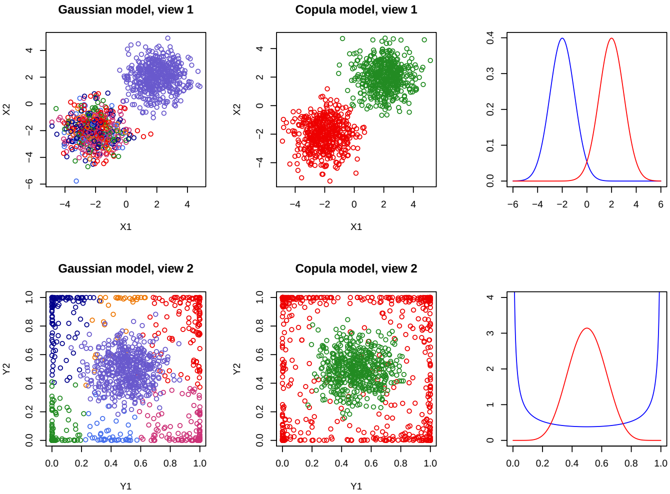
$$Z \sim \mathrm{Mult}(\theta),$$
$$(X, Y) \,|\, Z \sim \mathcal{N}_{p+q}(\mu_z, \Psi).$$

  where Ψ has a block-diagonal form:

$$\Psi = \begin{pmatrix} \Psi_x & 0 \\ 0 & \Psi_y \end{pmatrix}.$$

- Ψ block diagonal $\longrightarrow$ independent views conditioned on cluster assignment $\longrightarrow$ cluster structure captures dependencies.

Clustering of non-Gaussian data



Meta-Gaussian distribution

- Specify dependence by a Gaussian copula with block-diagonal correlation matrix P:

$$C_P^{\mathcal{G}}(u) = \Phi_P\left(\Phi^{-1}\left(u^1\right), \ldots, \Phi^{-1}\left(u^d\right)\right).$$

- Margins are arbitrary continuous distributions:

$$X^j|\theta = X^j|\theta^j \sim F^j_{X|\theta},\ j = 1, \ldots, p,$$
$$Y^j|\theta = Y^j|\theta^j \sim F^j_{Y|\theta},\ j = 1, \ldots, q.$$

- Use Sklar's Theorem to construct $F_{\theta,P}$.

Meta-Gaussian density

- Consider F with copula C and margins $F^1, \dots, F^d$. If F has a density then it can be expressed as:

$$f(x^1, \dots, x^d) = c\left(F^1(x^1), \dots, F^d(x^d)\right) \prod_{j=1}^{d} f^j(x^j),$$

  where $c(u^1, \dots, u^d) = \frac{\partial C(u^1, \dots, u^d)}{\partial u^1 \dots \partial u^d}$ is the copula density of C.
- Gaussian copula density has a simple form and $f_{\theta,P}$ is:

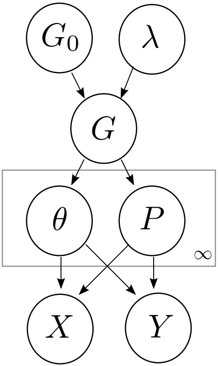
$$f_{(X,Y)|\theta,P}(x, y) = |P|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\tilde{x}^T (P^{-1} - I)\tilde{x}\right\} \prod_{j=1}^{p+q} f^j(x^j),$$

  where $\tilde{x}^j = \Phi^{-1}\left(F^j(x^j)\right)$.

Mixture of Copula Model

The joint density of X and Y is a Dirichlet process prior mixture:

$$f_{(X,Y)}(x,y) =$$
$$\int\int f_{(X,Y)|\theta,P}(x,y)\,\mathrm{d}\mu_{\theta,P}\,\mathrm{d}\mu_G(\lambda,G_0).$$



The priors

- Assume *a priori* independence for θ and P:
  $\rightarrow$ specify the priors separately
- Specify prior distributions for P_x and P_y, where $P = \begin{pmatrix} P_x & 0 \\ 0 & P_y \end{pmatrix}$, assuming *a priori* independence.
- For P_x and P_y we choose the marginally uniform prior [Barnard, 2000]:

$$f\left(R, d+1\right) \propto |R|^{\frac{d(d-1)}{2}-1} \left(\prod_{i=1}^{d} |R_{ii}| \right)^{-\frac{(d+1)}{2}}.$$

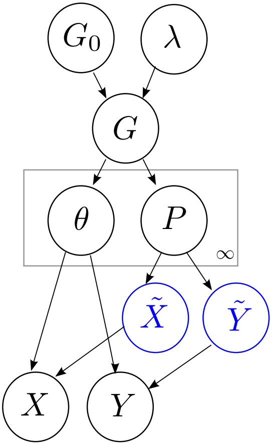
Inference

- MCMC algorithm for DP with non-conjugate prior [Neal, 1998].
- Simplifies when using data augmentation: introduce the normal scores $(\tilde{X}, \tilde{Y})$

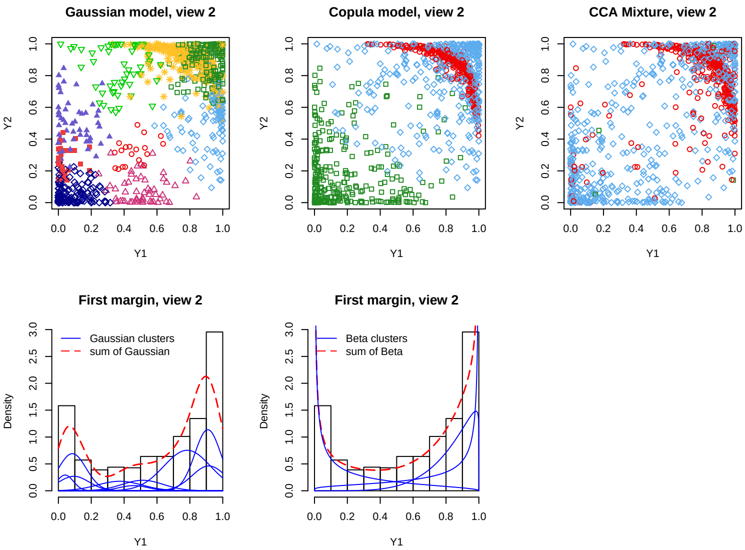
$$\tilde{X}^j = \Phi^{-1}\left(F^j(X^j)\right),$$
$$\tilde{Y}^j = \Phi^{-1}\left(F^j(Y^j)\right).$$

We then have:

$$(\tilde{X}, \tilde{Y}) \sim \mathcal{N}_{p+q}(0, P)$$



Simulations

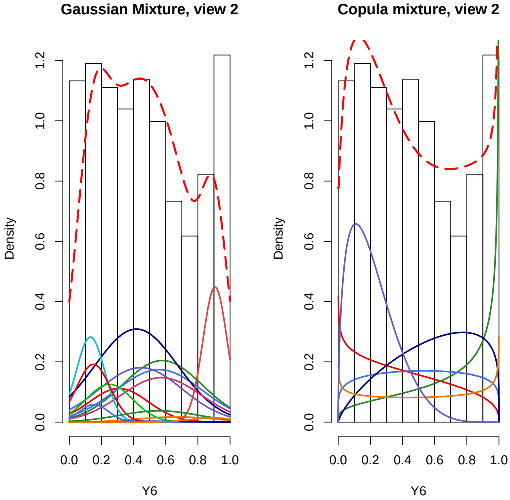


Real data experiments

Two data sets containing information about the regulation of heat shock in yeast, [Gasch, 2000], [Harbison, 2004].

- **First view** : gene expressions for yeast measured at 4 time points
  $\rightarrow$ **Gaussian**
- **Second view**: probability scores of binding interactions for 8 different regulators
  $\rightarrow$ **Beta**

Real data experiments



Conclusion

- Dependency-seeking clustering as alternative to CCA for multi-view analysis.
- Gaussian model produces misleading results when Gaussian assumption violated.
- Increase flexibility using a copula mixture model.

- Thank you !

melanie.rey@unibas.ch