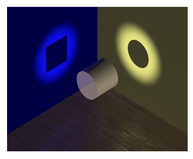
Copula Mixture Model for Dependency-seeking Clustering

Melanie Rey, Volker Roth

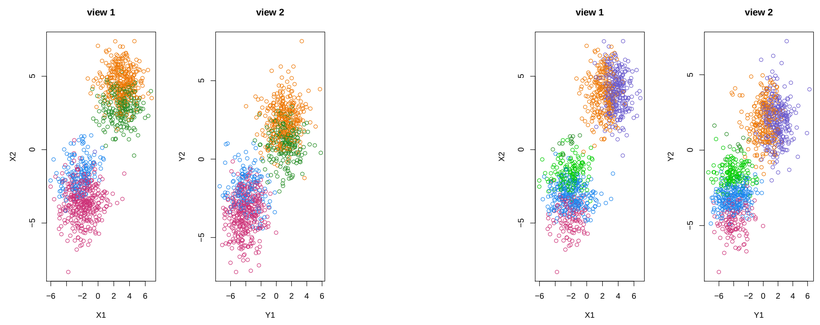
Department of Mathematics and Computer Science, University of Basel, Switzerland

Dependency-seeking Clustering

- Clustering of co-occurring samples from different data sources called views.



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



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

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

where $\mu = (\mu_x, \mu_y) \in \mathbb{R}^{p+q}$, $\mathbf{W} = \begin{pmatrix} \mathbf{W}_x \\ \mathbf{W}_y \end{pmatrix} \in \mathbb{R}^{(p+q)\times d}$, $1 \leq d \leq \min(p, q)$ and the covariance matrix Ψ has a block diagonal form:

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

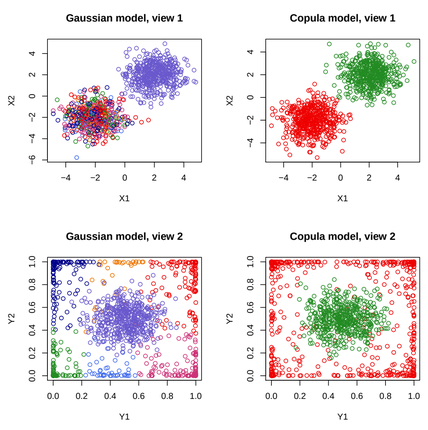
- Probabilistic dependency-seeking clustering model derived in [Klami, 2006] :

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

- The latent variable $\mathbf{Z}$ plays now the role of the cluster assignment.
- Ψ still has a block structure.
- This special form implies that given the cluster assignment the two views are independent and thereby enforces the cluster structure to capture all the dependencies.

Motivation for Copula Mixture

- The above model assumes that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally multivariate Gaussian.
- When applied to non-normally distributed data these models have to increase the number of clusters to achieve a reasonable fit.
- The components of these mixtures will not only be used to reflect differences in dependence structures but will also be used to approximate a non-Gaussian distribution.



Copula Mixture Model

- We assume that the joint density is a Dirichlet process prior mixture

$$\mathbf{f}_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}) = \int\int \mathbf{f}_{(\mathbf{X},\mathbf{Y})|\theta,\mathbf{P}}(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu_{\theta,\mathbf{P}} \mathrm{d}\mu_{\mathbf{G}}(\lambda, \mathbf{G}_0).$$

- We specify the margins and the dependence separately:
  - The margins can be arbitrary continuous cdfs:

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

  - The dependence structure is then specified by a Gaussian copula $\mathbf{C}^{\mathcal{G}}_{\mathbf{P}}$ with block diagonal correlation matrix $\mathbf{P}$.
  - Finally the conditional cdf is: $\mathbf{F}_{(\mathbf{X},\mathbf{Y})|\theta,\mathbf{P}}(\mathbf{x}, \mathbf{y}) = \mathbf{C}^{\mathcal{G}}_{\mathbf{P}}\left(\mathbf{F}^1_{\mathbf{X}|\theta}(\mathbf{x}^1), \ldots, \mathbf{F}^p_{\mathbf{X}|\theta}(\mathbf{x}^p), \mathbf{F}^1_{\mathbf{Y}|\theta}(\mathbf{x}^1), \ldots, \mathbf{F}^q_{\mathbf{Y}|\theta}(\mathbf{x}^q)\right).$

Gaussian Copula and Copula densities

- Consider a multivariate Gaussian variable $\mathbf{X} \sim \mathcal{N}_d(\mu, \Sigma)$, then by Sklar's Theorem there exists a unique copula $\mathbf{C}$ such that:

$$\Phi_{\mu,\Sigma}(\mathbf{x}^1, \ldots, \mathbf{x}^d) = \mathbf{C}\left(\Phi_{\mu_1,\Sigma_{11}}(\mathbf{x}^1), \ldots, \Phi_{\mu_d,\Sigma_{dd}}(\mathbf{x}^d)\right).$$

- The copula of $\mathcal{N}_d(\mu, \Sigma)$ is the same as the copula of $\mathcal{N}_d(0, \mathbf{P})$, where $\mathbf{P} = \mathcal{P}(\Sigma)$.
- The Gaussian copula is given by: $\mathbf{C}^{\mathcal{G}}_{\mathbf{P}}(\mathbf{u}) = \Phi_{\mathbf{P}}\left(\Phi^{-1}(\mathbf{u}^1), \ldots, \Phi^{-1}(\mathbf{u}^d)\right)$.
- Consider a multivariate cdf $\mathbf{F}$ with copula $\mathbf{C}$ and marginal cdfs $\mathbf{F}^1, \ldots, \mathbf{F}^d$:

$$\mathbf{F}(\mathbf{x}^1, \ldots, \mathbf{x}^d) = \mathbf{C}\left(\mathbf{F}^1(\mathbf{x}^1), \ldots, \mathbf{F}^d(\mathbf{x}^d)\right).$$

If $\mathbf{F}$ has a density then it can be expressed as:

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

where $\mathbf{c}(\mathbf{u}^1, \ldots, \mathbf{u}^d) = \frac{\partial \mathbf{C}(\mathbf{u}^1, \ldots, \mathbf{u}^d)}{\partial \mathbf{u}^1 \ldots \partial \mathbf{u}^d}$ is the copula density of $\mathbf{C}$.

- The Gaussian copula density has a simple form and the multivariate conditional density $\mathbf{f}$ becomes:

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

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

Bayesian inference

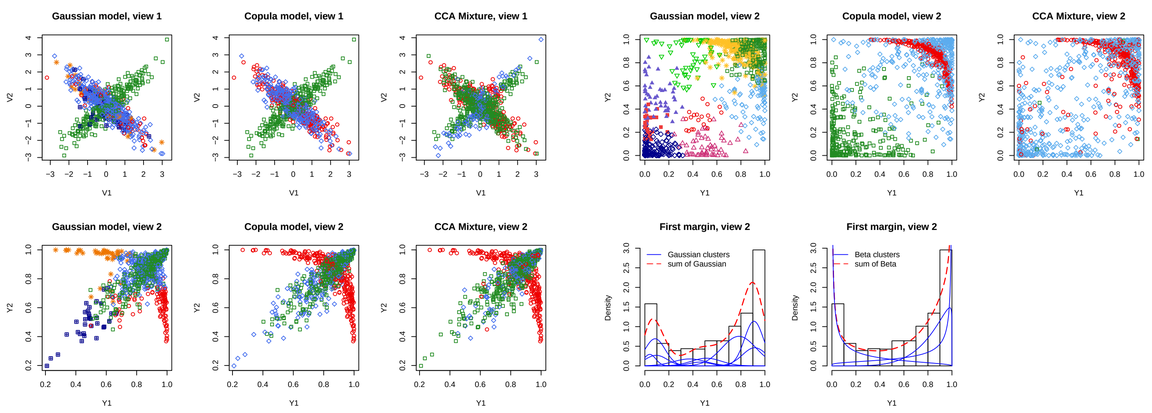
- Assume *a priori* independence of θ and $\mathbf{P}$, specify the priors separately.
- Specify prior distributions for the blocks $\mathbf{P}_x$ and $\mathbf{P}_y$, where $\mathbf{P} = \begin{pmatrix} \mathbf{P}_x & 0 \\ 0 & \mathbf{P}_y \end{pmatrix}$ and assume that $\mathbf{P}_x$ and $\mathbf{P}_y$ are *a priori* independent.
- For $\mathbf{P}_x$ and $\mathbf{P}_y$ we choose the marginally uniform prior [Barnard, 2000]:

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

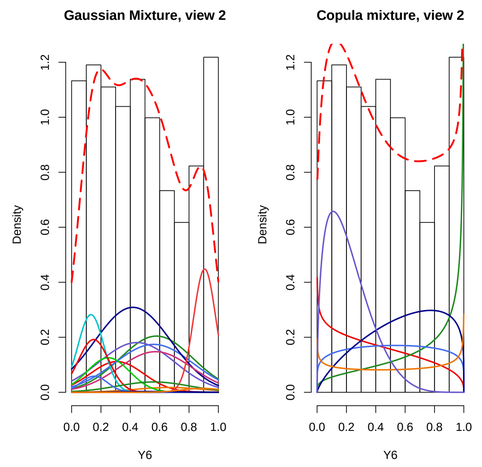
- Inference for θ and $\mathbf{P}$ is performed using MCMC.

Results

- Simulations



- Real data. Two data sets containing information about the regulation of heat shock in yeast. First view: series of gene expressions for yeast under heat shock measured at 4 time points. Second view: probability scores of binding interactions between the promoter region of different genes and several DNA-binding transcriptional regulators under heat shock.



<firstname.lastname>@unibas.ch