flexible sampling of
discrete data correlations
without the marginal distributions

Alfredo Kalaitzis, Ricardo Silva

University College London, CSML

November 29, 2013
Learning the joint dependence of discrete variables is a fundamental problem in machine learning & statistics.

- Prediction
- Clustering
- Dimensionality reduction, etc.

Ways of constructing discrete distributions:
- Contingency tables (if small dimensional)
- Sparse structures
- (Hierarchical) LVMs
- Copulas
Copulas

\[ F(y_1, y_2) = F\left( F_1^{-1}(F_1(y_1)), F_2^{-1}(F_2(y_2)) \right), \]

where

\[ C(\cdot, \cdot) \equiv F\left( F_1^{-1}(\cdot), F_2^{-1}(\cdot) \right) \]

is the copula of \( F \).

- “←” reveals a modular approach to constructing joint distributions.
- Combining univariate marginals through a parametric distribution family, e.g. Gaussian.
- Complements graphical models and other modular parameterizations of joint distributions.
The Gaussian copula

Gaussian copula ("←"): 

1. \( z^{(i)} \sim \mathcal{N}_p(0, \mathbf{C}) \) and \( u_j^{(i)} = \Phi(z_j^{(i)}) \), thus \( u_j^{(i)} \sim \mathcal{U}[0, 1] \)

2. Given univariate CDFs, \( \{F_j\}_{1..p} \), define \( y_j^{(i)} = F_j^{-1}(u_j^{(i)}) \) to construct a model for the joint CDF of \( y \).

3. Think of \( Z \) is a continuous (augmented) representation of \( Y \).
What about CDFs of **discrete** data?

- **Bayesian inference** for $C$ given **discrete** data $Y$:

  $$p(C, \theta_F | Y) \propto p_{GC}(Y|C, \theta_F) \pi(C, \theta_F)$$

  - $p_{GC}$: PMF of a Gaussian copula and marginals given by $\theta_F$.
  - **But**, transforming **discrete** CDFs to PMFs is generally intractable due to jumps in the CDF of the **discrete** $Y_j$. 
The extended rank likelihood (XRL)

- **Jumps** induce hard **constraints**:
  \[ y_j^{(i)} < y_j^{(i')} \implies z_j^{(i)} < z_j^{(i')} \]

- (Hoff, 07) defines the event \( Z \in D \).
  \[ D \equiv \{ Z \in \mathbb{R}^{n \times p} : Z \text{ satisfies all constraints} \} \]

- \( p_{GC}(Y \mid C, \theta_F) = p_{GC}(Z \in D, Y \mid C, \theta_F) = p_{GC}(Y \mid Z \in D, C, \theta_F) p(Z \in D \mid C) \)  

- Now inference on \( C \) can be done by marginalizing over \( Z \):
  \[ p(C, Z \mid Y) \propto I(Z \in D) \mathcal{N}(Z \mid C) \pi(C). \]

  - **e.g.** by Gibbs sampling from the **truncated Gaussian** and \( p(C \mid Z) \) (an \( \mathcal{IW} \) density) and throwing away the \( Z \) sample.

- **Note**: \( C \) is no longer estimated based on \( Y \) or \( \theta_F \), but on the probability of an **event** that is a **superset** of observing the ranks.
XRL for *semiparametric* copula estimation

- XRL introduces a **summary statistic** of the data, independent of nuisance parameters $\theta_F$.
  - The XRL also known as a *generalized* marginal likelihood.
- **No assumptions** on $\{F_j\}_{1..p}$ required!
  - Much simpler to sample the XRL than the joint likelihood.
  - Avoids any “entanglements” between $\theta_F$ and $C$.
- **Extensibility**: further structure can be imposed on $C$ with appropriate priors.
- Hard constraints & univariate Gibbs sampling of $z_j^{(i)}$ can result in **slow mixing**.
  - Can we do better?
Exact HMC for truncated multivariate Gaussians

- Don’t sample each $Z_{i,j} | Z_{i,j}$, (univariate truncated Gaussians).
- Instead, sample $Z_{:,j} | Z_{:,j}$ jointly using a constrained HMC sampler (Pakman et al, 2012)
  - Shown to mix better than Gibbs sampling.

- Bottleneck in computing $O(n^2)$ bouncing times.
- We reduce this to $O(n)$ in practice.
 Searching for the Hough envelope

- The **special structure** of hard constraints: each component of the HMC particle has a sinusoid evolution:
  \[ x_i(t) = \mu_i + a_i \sin(t) + b_i \cos(t) \]
- Recall: \( Y_{L1,j} < Y_{L2,j} \) implies \( Z_{L1,j} < Z_{L2,j} \),
Application on the Bayesian Gaussian copula factor model

<table>
<thead>
<tr>
<th>Figure</th>
<th>vars</th>
<th>factors</th>
<th>ordinal levels</th>
<th>elapsed (mins):</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>20</td>
<td>5</td>
<td>2</td>
<td>115</td>
</tr>
<tr>
<td>(b)</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>(c)</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>203</td>
</tr>
</tbody>
</table>

1000 observations

![Graphs](image1.png) ![Graphs](image2.png) ![Graphs](image3.png)
Thank you

Research supported by EPSRC grant EP/J013293/1.