1 Data

1.1 Examples

Health Records Data

Data: $n = 85,021$ patients

Variables $y_i$: fasting blood glucose, white blood cell count, red blood cell count, hemoglobin, platelets, low density lipoproteins, total cholesterol, triglycerides, triketopurine, high density lipoproteins, serum creatinine, serum glutamic oxaloacetic transaminase, total bilirubin, gender, height, weight, blood pressure and waist

Outcome: diabetes

Goal: clustering, classification & prediction

Zodiac Data

Data: gene-gene interactions of $n = 19,304$ genes with all other genes.

Data on gene-gene interactions from Zodiac (Zhu et al., 2015) with TCGA data.
1.3 Predictive recursion clustering

**Predictive recursion clustering** (PRC)

**Predictive recursion clustering**: Zuanetti et al. (2018 StatComp); use predictive recursion like Newton et al. (98), approximating the posterior predictive

\[ p(\theta_{i+1} \mid y_1, \ldots, y_{i-1}) \approx g_i(\theta) \]

\[ g_i(\theta) = (1 - w_i)g_{i-1}(\theta) + w_i \frac{f(y_i \mid \theta)g_{i-1}(\theta)}{c(y_i, g_{i-1})}, \]

exact for \( i = 1 \) (and \( w_i = 1/(1 + \alpha) \)), and approx beyond.

**Clustering**: \( g_i \) builds up as a mixture model, which implicitly defines a random partition (with some computational simplifications, like dropping terms with very small weight, etc.)

---

2 SIGN algorithm for BNP clustering

**SIGN algorithm for BNP clustering**

1. split data into “shards”;
2. subset posterior on shards;
3. split clusters into shards → Step 2 with clusters as the new units;
4. stop when only one shard is left.

---

**Cluster of clusters**

In **Step 2,3,..**: clustering of clusters; similar notion in Argiento et al. (2014) and Malsiner-Walli, Frühwirth-Schnatter & Grün (2017) (for mixture of non-Gaussian dists).

**Notation**: let \( \bar{s}_i = j \) if \( i \)th cluster (of original units) joins the \( j \)th cluster of clusters.

**Prior prob**: exchangeable prior \( \propto p(\rho) = f(n_1, \ldots, n_C) \) for \( C \) clusters with cardinalities \( n_c \) \(\Rightarrow\)

\[ p(\bar{s}_i = c \mid \bar{s}^{-i}) \propto \frac{p(\rho+c)}{p(\rho-i)} \]

for any BNP prior – easy;

**Transdim MCMC**: Neal’s (2000) Algorithm 8 for new singleton clusters

---

**Approximation**

**Partition**: \( s_i = k \) if \( i \)th unit in \( k \)th cluster; alternatively use indicators \( \delta_{ij} = I(s_i = s_j) \).

**Subset posteriors**: Let \( [n] = A \cup B \) denote two shards;

\( \delta_A = (\delta_{ij}, i, j \in A) \), same for \( \delta_B \),

\( \delta_{AB} = (\delta_{ij}, i \in A, j \in B) \)

\[ p(\delta \mid y) \approx q(\delta \mid y) = p(\delta_A \mid y_A)p(\delta_B \mid y_B)p(\delta_{AB} \mid \delta_A, \delta_B, y) \]

**Summary**: judge approximation by

\[ F_{.1} = \% \text{ pairs with } |E_q(\delta_{ij}) - E(\delta_{ij} \mid y)| > 0.1 \]
### 3 Simulation

#### Slide 11

**Simulations**

**Simulation I**: simple mixture of \( C_0 = 5 \) (true) normals; comparison with PY mixture (full MCMC) and DBSCAN (Ester et al., 1996)

<table>
<thead>
<tr>
<th></th>
<th>SIGN</th>
<th>SIGN-VI</th>
<th>PYM</th>
<th>DBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C )</td>
<td>4.94 (0.31)</td>
<td>4.90 (0.30)</td>
<td>5.08 (0.27)</td>
<td>3.64 (0.63)</td>
</tr>
<tr>
<td>MISC</td>
<td>0.08 (0.03)</td>
<td>0.09 (0.03)</td>
<td>0.04 (0.01)</td>
<td>0.41 (0.11)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.01 (0.01)</td>
<td>0.01 (0.01)</td>
<td>0.01 (0.00)</td>
<td>0.49 (0.18)</td>
</tr>
</tbody>
</table>

MISC=misclassification rate; MSE=estimation of cluster-specific means

**Estimated partition at the end of each step**: need estimated clusters after each step; use Dahl (2006) summary.

Alternatively, variation of information loss (SIGN-VI)

---

#### Slide 12

**Simulation II: two spirals**

- full MCMC
- SIGN
- DBSCAN

\[ \hat{C} = 14.10(1.37) \]

The reduced \( \hat{C} \) under the SIGN approximation is typical. Clusters from early steps can be merged, but never split.

---

### 4 Classification

#### Slide 13

**Prediction**

**Density estimation**: BNP clustering,

\[ p(\rho) p(\theta | \rho) p(y_i | \rho, \theta), \]

implies density estimation \( p(y_{n+1} | y) \).

**Regression & prediction**: to be useful for regression, need conditioning in \( p(\rho | x) \), on covariates \( x_i \);

\[ p(y_{n+1} | x_{n+1}, y, x) \] defines desired regression.

**Augmented model**: augment response to \( z_i = (x_i, y_i) \) and proceed as before

\[ p(\rho) p(\theta | \rho) p(x_i, y_i | \rho, \theta). \]

---

**PPMx**: define more general \( p(\rho | x) \), avoiding explicit modeling of a covariate distribution

---

#### Slide 14

**Classification**: AUC.

Comparison with full MCMC (“PPMx”), BART, random forest (RF), logistic regression (LR) and SVM

<table>
<thead>
<tr>
<th></th>
<th>Simulation III</th>
<th>Simulation IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGN</td>
<td>0.808 (0.067)</td>
<td>0.838 (0.067)</td>
</tr>
<tr>
<td>PPMx</td>
<td>0.824 (0.060)</td>
<td>0.841 (0.063)</td>
</tr>
<tr>
<td>BART</td>
<td>0.755 (0.062)</td>
<td>0.866 (0.050)</td>
</tr>
<tr>
<td>RF</td>
<td>0.793 (0.059)</td>
<td>0.838 (0.067)</td>
</tr>
<tr>
<td>LR</td>
<td>0.600 (0.091)</td>
<td>0.524 (0.073)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.622 (0.077)</td>
<td>0.585 (0.077)</td>
</tr>
</tbody>
</table>

---

#### Slide 15

**Results**

**Response \( y_i \)**: indicator for diabetes

**Covariates** \( x_i \): white blood cell count (WBC), red blood cell count (RBC), hemoglobin (HGB), platelets (PLT), fasting blood glucose (FBG), low density lipoproteins (LDL), total cholesterol (TC), triglycerides (Trig), triketopurine (Trik), high density lipoproteins (HDL), serum creatinine (SCr), serum glutamic oxaloacetic transaminase (SGOT), and total bilirubin (TB); sex, height, weight, blood pressure, and waist.

**Data**: \( n = 85,021 \) patients

**SIGN**: \( M_1 = 250 \) shards \( \rightarrow 1351 \) local clusters;

\( M_2 = 5 \) shards \( \rightarrow 25 \) regional clusters;

Algorithm stops at step \( K = 3 \)
<table>
<thead>
<tr>
<th></th>
<th>EHR</th>
<th>Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGN</td>
<td>0.880</td>
<td>0.825</td>
</tr>
<tr>
<td>PPMx</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BART</td>
<td>0.867</td>
<td>0.792</td>
</tr>
<tr>
<td>RF</td>
<td>0.869</td>
<td>0.786</td>
</tr>
<tr>
<td>LR</td>
<td>0.856</td>
<td>0.781</td>
</tr>
<tr>
<td>SVM</td>
<td>0.856</td>
<td>0.761</td>
</tr>
</tbody>
</table>

(“Bank” is another data set, on success of telemarketing)

**Slide 17**

GAN

**GAN:** Chinese policy requires “China first” publication; We use a “Generative Adversarial Network” (GAN) (Goodfellow et al. 2014) to generate a hypothetical repeat

- One network does density estimation $p(x_i, y_i)$ and predictive simulation of $n$ fake data, $i = n + 1, \ldots, n + n$; pass the augmented data to a second network:
- A second network tries to discriminate original versus fake data.
- Iterate until discrimination is impossible.

We comply with Chinese law, but statistical inference is identical

**Slide 20**

**Example:** Tumor heterogeneity

- experimental unites = mutations $i = 1, \ldots, n$;
- features = homogeneous subclones $F_j \subseteq [n]$,
  subsets of mutations;
- Each subclone is linked with a set of weights $w_j$, for observed tissue samples, use $d(w_j, w_\ell)$ to decide merging

**Double feature allocation**

Double feature allocation: two sets of experimental units, $i = 1, \ldots, n$ (e.g., patients) and $s = 1, \ldots, S$ (e.g., symptoms); each feature $F_j \subseteq [n]$ (e.g., disease) is associated with a subset $S_j \subseteq [S]$.

**Example:** EHR, features = ”disease”, $F_j \subseteq [n]$, subsets of patients;
Each disease is linked to a set $S_j$ of symptoms, $d(S_j, S_\ell) = |S_j \cap S_\ell|/|S_j \cup S_\ell|$.

**Slide 21**

**Summary**

- Model-based clustering (feature allocation, double FA) is more flexible than purely algorithmic methods, but computationally challenging for large $n$ (also for large $p$)
- Several algorithms, using predictive recursion, approximation, parallelization, subset posteriors (consensus MC)
• Approximate posterior uncertainties important for decision problems (e.g., phenotype discovery in EHR data)