Clustering high dimensional mixed data:
joint analysis of phenotypic and genotypic data

Damien McParland*, Catherine Phillips, Lorraine Brennan, Helen Roche and Claire Gormley*

*School of Mathematics and Statistics & the Insight Centre for Data Analytics
University College Dublin.
Modelling high dimensional data of mixed type: continuous, binary, nominal.
What’s coming up...

- Modelling **high dimensional** data of **mixed type**: continuous, binary, nominal.

- **Clustering** using finite mixture models.
What’s coming up...

- Modelling **high dimensional** data of **mixed type**: continuous, binary, nominal.

- **Clustering** using finite mixture models.

- Bayesian estimation, with **variable** and **model selection**.
What’s coming up...

- Modelling **high dimensional** data of **mixed type**: continuous, binary, nominal.

- **Clustering** using finite mixture models.

- Bayesian estimation, with **variable** and **model selection**.

- Motivating application: LIPGENE-SU.VI.MAX study. ‘**Diet, genomics and the metabolic syndrome**: an integrated nutrition, agro-food, social and economic analysis.’
Modelling high dimensional data of mixed type: continuous, binary, nominal.

Clustering using finite mixture models.

Bayesian estimation, with variable and model selection.

Motivating application: LIPGENE-SU.VI.MAX study. ‘Diet, genomics and the metabolic syndrome: an integrated nutrition, agro-food, social and economic analysis.’

Aim: uncover any sub-phenotypes, identify discriminating variables, considering all data.
The LIPGENE-SU.VI.MAX study.
The metabolic syndrome (MetS)

- Complex disorder that can lead to increased risk of type 2 diabetes and cardiovascular disease.

- The World Health Organisation estimates global diabetes prevalence will double by 2030.

- Diagnosed if have 3 abnormalities:
  - Fasting glucose > 5.5 mmol l\(^{-1}\)
  - Serum TAG > 1.5 mmol l\(^{-1}\)
  - Serum HDL-c < 1.04 mmol l\(^{-1}\) (Men), < 1.29 mmol l\(^{-1}\) (Women)
  - Blood pressure: Systolic BP > 130 mm Hg, Diastolic BP > 85 mm Hg
  - Waist circumference: > 94 cm (Men), > 80 cm (Women)
The metabolic syndrome (MetS)

- Complex disorder that can lead to increased risk of type 2 diabetes and cardiovascular disease.
- The World Health Organisation estimates global diabetes prevalence will double by 2030.
- Diagnosed if have ≥ 3 abnormalities:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fasting glucose concentration</td>
<td>≥5.5 mmol l⁻¹</td>
</tr>
<tr>
<td>Serum TAG concentration</td>
<td>≥ 1.5 mmol l⁻¹</td>
</tr>
<tr>
<td>Serum HDL-c concentration</td>
<td>&lt; 1.04 mmol l⁻¹ (Men)</td>
</tr>
<tr>
<td></td>
<td>&lt; 1.29 mmol l⁻¹ (Women)</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>Systolic BP ≥ 130 mm Hg</td>
</tr>
<tr>
<td></td>
<td>Diastolic BP ≥ 85 mm Hg</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>&gt; 94 cm (Men)</td>
</tr>
<tr>
<td></td>
<td>&gt; 80 cm (Women)</td>
</tr>
</tbody>
</table>
The LIPGENE-SU.VI.MAX study

- Pan-European, prospective population based study focusing on interaction of nutrients and genotype in MetS.

- Initial data collected on $N = 505$ participants.
The LIPGENE-SU.VI.MAX study

- Pan-European, prospective population based study focusing on interaction of nutrients and genotype in MetS.

- Initial data collected on $N = 505$ participants.

- **Continuous** phenotypic variables:
  - Anthropometric (eg. waist circumference) and biochemical (eg. plasma fatty acid levels) measurements. ($A = 26$)
The LIPGENE-SU.VI.MAX study

- Pan-European, prospective population based study focusing on interaction of nutrients and genotype in MetS.

- Initial data collected on $N = 505$ participants.

- **Continuous** phenotypic variables:
  - Anthropometric (eg. waist circumference) and biochemical (eg. plasma fatty acid levels) measurements. ($A = 26$)

- **Nominal** genetic SNP data.
  - $B = 341$ nominal SNPs with 3 levels.
  - Eg. $rs512535 \in \{AA, GG, AG\}$
The LIPGENE-SU.VI.MAX study

- Pan-European, prospective population based study focusing on interaction of nutrients and genotype in MetS.

- Initial data collected on $N = 505$ participants.

- **Continuous** phenotypic variables:
  - Anthropometric (eg. waist circumference) and biochemical (eg. plasma fatty acid levels) measurements. ($A = 26$)

- **Nominal** genetic SNP data.
  - $B = 341$ nominal SNPs with 3 levels.
  - Eg. $rs512535 \in \{AA, GG, AG\}$

- **Binary** genetic SNP data.
  - $C = 371$ SNPs with 2 levels.
  - Eg. $rs17777371 \in \{GG, CG/CC\}$
The LIPGENE-SU.VI.MAX study

- Pan-European, prospective population based study focusing on interaction of nutrients and genotype in MetS.

- Initial data collected on $N = 505$ participants.

- **Continuous** phenotypic variables:
  - Anthropometric (eg. waist circumference) and biochemical (eg. plasma fatty acid levels) measurements. ($A = 26$)

- **Nominal** genetic SNP data.
  - $B = 341$ nominal SNPs with 3 levels.
  - Eg. $\text{rs512535} \in \{\text{AA, GG, AG}\}$

- **Binary** genetic SNP data.
  - $C = 371$ SNPs with 2 levels.
  - Eg. $\text{rs17777371} \in \{\text{GG, CG/CC}\}$

- **Aim:** model $J = A + B + C = 738$ variables simultaneously.
The LIPGENE-SU.VI.MAX study

- Seven year follow up data: continuous phenotypic data only collected.
- Participants were then diagnosed as having the MetS or not.
The LIPGENE-SU.VI.MAX study

- Seven year follow up data: continuous phenotypic data only collected.

- Participants were then diagnosed as having the MetS or not.

Questions of interest:

1. In the initial data, are there clusters or *sub-phenotypes*?
The LIPGENE-SU.VI.MAX study

- Seven year follow up data: continuous phenotypic data only collected.
- Participants were then diagnosed as having the MetS or not.
- Questions of interest:
  1. In the initial data, are there clusters or *sub-phenotypes*?
  2. If so, are there discriminating variables?
The LIPGENE-SU.VI.MAX study

- Seven year follow up data: continuous phenotypic data only collected.

- Participants were then diagnosed as having the MetS or not.

- Questions of interest:
  1. In the initial data, are there clusters or sub-phenotypes?
  2. If so, are there discriminating variables?
  3. If so, are discriminating variables genetic, phenotypic, or both?
The LIPGENE-SU.VI.MAX study

- Seven year follow up data: continuous phenotypic data only collected.

- Participants were then diagnosed as having the MetS or not.

Questions of interest:

1. In the initial data, are there clusters or *sub-phenotypes*?
2. If so, are there discriminating variables?
3. If so, are discriminating variables genetic, phenotypic, or both?
4. Is there a correspondence between the initial clusters and the 7-yr follow up diagnosis?
Clustering data of mixed type.
Early attempts employed latent variable models and location models:
State of the art

- Early attempts employed latent variable models and location models:

- Non-model based approaches:
  Huang (1997), Ahmad & Dey (2007), . . .
State of the art

- Early attempts employed latent variable models and location models:
  Everitt (1988), Hunt & Jorgensen (1999) ...

- Non-model based approaches:
  Huang (1997), Ahmad & Dey (2007), ...

- Clustering mixed categorical data:
  Cai et al. (2011), Morlini (2011), Browne & McNicholas (2012), McParland et al. (2014) ...
State of the art

- Early attempts employed latent variable models and location models:

- Non-model based approaches:
  Huang (1997), Ahmad & Dey (2007), . . .

- Clustering mixed categorical data:

- Clustering mixed continuous & categorical data:
  McParland & Gormley (2016) & associated R package clustMD
State of the art

- Early attempts employed latent variable models and location models:

- Non-model based approaches:
  Huang (1997), Ahmad & Dey (2007), . . .

- Clustering mixed categorical data:

- Clustering mixed continuous & categorical data:
  McParland & Gormley (2016) & associated R package clustMD

- Copula based approaches:
  Marbec et al. (2014), Kosmidis & Karlis (2015), . . .
Clustering data of mixed type.

- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.

- (Categorical) data are high dimensional.
Clustering data of mixed type.

- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.

- (Categorical) data are high dimensional.

- Draw on ideas from item response theory and latent variable models.
Clustering data of mixed type.

- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.

- (Categorical) data are high dimensional.

- Draw on ideas from item response theory and latent variable models.

- Three data types:
  - Binary data $\rightarrow$ item response theory model.
Clustering data of mixed type.

- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.

- (Categorical) data are high dimensional.

- Draw on ideas from item response theory and latent variable models.

- Three data types:
  - Binary data → item response theory model.
  - Nominal data → mutinomial probit model.
Clustering data of mixed type.

- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.

- (Categorical) data are high dimensional.

- Draw on ideas from item response theory and latent variable models.

- Three data types:
  - Binary data → item response theory model.
  - Nominal data → mutinomial probit model.
  - Continuous data → factor analysis.
Corresponding to each observed binary SNP $y_{ij}$ is a latent Gaussian variable $z_{ij}$. 

**Binary data: item response theory model.**
Corresponding to each observed binary SNP $y_{ij}$ is a latent Gaussian variable $z_{ij}$. 

SNP rs17777371
Corresponding to each **observed** binary SNP $y_{ij}$ is a **latent** Gaussian variable $z_{ij}$. 

![Graph of Gaussian distribution with SNP rs17777371 and annotations $\gamma_0$, $\gamma_1$, and $\gamma_2$.]
Corresponding to each observed binary SNP $y_{ij}$ is a latent Gaussian variable $z_{ij}$.
Model $z_i = (z_{i1}, \ldots, z_{iC})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_j$: 
Item response theory model: factor analytic structure.

Model $\mathbf{z}_i = (z_{i1}, \ldots, z_{iC})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$:

$$\mathbf{z}_i = \mu + \Lambda \theta_i + \epsilon_i$$

where

- $\mu$ $-$ C-vector of negative item difficulty parameters
- $\Lambda$ $-$ $C \times Q$ matrix of item discrimination parameters
- $\theta_i$ $\sim$ $\text{MVN}_Q(0, I)$
- $\epsilon_i$ $\sim$ $\text{MVN}_C(0, I)$
Model $z_i = (z_{i1}, \ldots, z_{iC})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$:

$$z_i = \mu + \Lambda \theta_i + \epsilon_i$$

where

- $\mu$ is a $C$-vector of negative item difficulty parameters
- $\Lambda$ is a $C \times Q$ matrix of item discrimination parameters
- $\theta_i \sim MVN_Q(0, I)$
- $\epsilon_i \sim MVN_C(0, I)$

Dimension $Q$ of the latent trait $\theta_i$ is unknown, but $Q \ll C$.

$$z_i | \theta_i \sim MVN_C(\mu + \Lambda \theta_i, I)$$
Nominal data: multinominal probit model.

- Underlying $y_{ij}$ are $K_j - 1$ latent Gaussian variables $\{z_{ij}^k\}$. 
Underlying $y_{ij}$ are $K_j - 1$ latent Gaussian variables $\{z_{ij}^k\}$.

Each observed nominal SNP $y_{ij}$ has $K_j = 3$ levels.
Nominal data: multinomial probit model.

- Underlying $y_{ij}$ are $K_j - 1$ latent Gaussian variables $\{z_{ij}^k\}$.
- Each observed nominal SNP $y_{ij}$ has $K_j = 3$ levels.
- Example: SNP rs512535 $\in \{\text{AA, GG, AG}\}$. Thus,

$$z_{ij} = \{z_{ij}^1, z_{ij}^2\}$$
Nominal data: multinomial probit model.

Damien:

\[ \Rightarrow AA \]
Nominal data: multinomial probit model.

Lorraine:

Density of First Latent Dimension

Density of Second Latent Dimension

⇒ GG
Nominal data: multinominal probit model.

Claire:

Density of First Latent Dimension $Z_1$

Density of Second Latent Dimension $Z_2$

$\Rightarrow$ AG
Another view...
Another view...
Another view...

\[
\begin{align*}
Y &= AA \\
Y &= GG \\
Y &= AG
\end{align*}
\]
Another view...

\[ Y = AA \]

\[ Y = GG \]

\[ Y = AG \]
Multinomial probit model: factor analytic structure.

Model $z_i = (z_{i1}, \ldots, z_{i(2B)})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$: 
Model \( z_i = (z_{i1}, \ldots, z_{i(2B)})^T \) as a linear function of a latent, low dimensional Gaussian variable \( \theta_i \):

\[
  z_i = \mu + \Lambda \theta_i + \epsilon_i
\]

where

- \( \mu \) is a \( 2B \) dimensional mean vector.
- \( \Lambda \) is a \( 2B \times Q \) loadings matrix.
- \( \theta_i \sim MVN_Q(0, I) \)
- \( \epsilon_i \sim MVN_{2B}(0, I) \)
Model $z_i = (z_{i1}, \ldots, z_{i(2B)})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$:

$$z_i = \mu + \Lambda \theta_i + \epsilon_i$$

where

- $\mu$ is a $2B$ dimensional mean vector.
- $\Lambda$ is a $2B \times Q$ loadings matrix.
- $\theta_i \sim MVN_Q(0, I)$
- $\epsilon_i \sim MVN_{2B}(0, I)$

Again, $Q << 2B$ and

$$z_i | \theta_i \sim MVN_{2B}(\mu + \Lambda \theta_i, I)$$
Continuous data: factor analysis model.

- Model $y_i = z_i = (z_{i1}, \ldots, z_{iA})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$: 
Continuous data: factor analysis model.

- Model $y_i = z_i = (z_{i1}, \ldots, z_{iA})^T$ as a linear function of a latent, low dimensional Gaussian variable $\theta_i$:

$$z_i = \mu + \Lambda \theta_i + \epsilon_i$$

where

- $\mu$: A dimensional mean vector.
- $\Lambda$: $A \times Q$ loadings matrix
- $\theta_i$: $MVN_Q(0, I)$
- $\epsilon_i$: $MVN_A(0, \Psi)$
Continuous data: factor analysis model.

Model \( y_i = z_i = (z_{i1}, \ldots, z_{iA})^T \) as a linear function of a latent, low dimensional Gaussian variable \( \theta_i \):

\[
z_i = \mu + \Lambda \theta_i + \epsilon_i
\]

where

- \( \mu \) A dimensional mean vector.
- \( \Lambda \) A \( \times \) Q loadings matrix
- \( \theta_i \) \( \text{MVN}_Q(0, I) \)
- \( \epsilon_i \) \( \text{MVN}_A(0, \Psi) \)

Again, Q \( \ll \) A and

\[
z_i | \theta_i \sim \text{MVN}_A(\mu + \Lambda \theta_i, \Psi)
\]
Similar model structure suggests a hybrid may be fruitful:

\[ y_{ij} = \begin{cases} z_{ij} & \text{if variable } j \text{ is continuous.} \\ k & \text{if variable } j \text{ is binary and } j < z_{ij} < k. \\ k & \text{if variable } j \text{ is nominal and } z_{ij} = \max_k \{ z_{ij} \} > 0. \end{cases} \]

Collect latent variables together into a single \( D = A + 2B + C \)-dimensional vector \( z_i \).

Model this joint latent vector using a factor analytic structure:

\( z_i \sim \text{MVN}(\mu_i; \Sigma) \).

Marginally, have a parsimonious covariance structure:

\( z_i \sim \text{MVN}(\mu; \Sigma_T) \).
Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

\[
y_{ij} = \begin{cases} 
  z_{ij} & \text{if variable } j \text{ is continuous.} 
\end{cases}
\]
Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

$$y_{ij} = \begin{cases} 
    z_{ij} & \text{if variable } j \text{ is continuous.} \\
    k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}.
\end{cases}$$
Similar model structure suggests a hybrid may be fruitful:

\[ y_{ij} = \begin{cases} 
  z_{ij} & \text{if variable } j \text{ is continuous.} \\
  k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}. \\
  k & \text{if variable } j \text{ is nominal and } z_{ij}^{k-1} = \max_{k} \{ z_{ij}^{k} \} > 0 .
\end{cases} \]
Similar model structure suggests a hybrid may be fruitful:

\[ y_{ij} = \begin{cases} 
  z_{ij} & \text{if variable } j \text{ is continuous.} \\
  k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}. \\
  k & \text{if variable } j \text{ is nominal and } z_{ij}^{k-1} = \max_k \{z_{ij}^k\} > 0.
\end{cases} \]

Collect latent variables together into a single

\[ D = A + 2B + C \] dimensional vector \( z_i \).
Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

\[ y_{ij} = \begin{cases} 
  z_{ij} & \text{if variable } j \text{ is continuous.} \\
  k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}. \\
  k & \text{if variable } j \text{ is nominal and } z_{ij}^{k-1} = \max_k \{z_{ij}^k\} > 0. 
\end{cases} \]

- Collect latent variables together into a single 
  \[ D = A + 2B + C \] dimensional vector \( z_i \).

- Model this joint latent vector using a factor analytic 
  structure:

\[ z_i | \theta_i \sim \text{MVN}_D(\mu + \Lambda \theta_i, \Psi). \]
Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

\[ y_{ij} = \begin{cases} 
  z_{ij} & \text{if variable } j \text{ is continuous.} \\
  k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}. \\
  k & \text{if variable } j \text{ is nominal and } z_{ij}^{k-1} = \max_{k} \{ z_{ij}^{k} \} > 0. 
\end{cases} \]

- Collect latent variables together into a single \( D = A + 2B + C \) dimensional vector \( z_i \).

- Model this joint latent vector using a factor analytic structure:

\[ z_i | \theta_i \sim \text{MVN}_D(\mu + \Lambda \theta_i, \Psi). \]

- Marginally, have a parsimonious covariance structure:

\[ z_i \sim \text{MVN}_D(\mu, \Lambda \Lambda^T + \Psi) \]
Hybrid model: factor analysis for mixed data (FA-MD)

- Complex, augmented, likelihood function:

\[
\mathbb{P}(y_j | \mu, \Lambda, z_i, \Theta, \Gamma, \Psi) = \prod_{j \text{ cns}} N(\mu_j + \lambda_j^T \theta_i, \psi_j) \\
\times \prod_{j \text{ bin}} N^T(\mu_j + \lambda_j^T \theta_i, 1) \mathbb{I}\{z_{ij}\} \\
\times \prod_{j \text{ nom}} \prod_{k=1}^{K_j-1} N^T(\mu_{jk}^k + \lambda_j^k \theta_i, 1) \mathbb{I}\{z_{ij}^k\}
\]
- Facilitate clustering using a mixture modelling framework.
- Each of $G$ clusters modelled using an FA-MD model.
Mixture of factor analysers for mixed data (MFA-MD)

- Facilitate clustering using a mixture modelling framework.
- Each of $G$ clusters modelled using an FA-MD model.
- Clustering occurs at the latent variable level:

$$P(z_i) = \sum_{g=1}^{G} \pi_g \text{MVN}_D(\mu_g, \Lambda_g \Lambda_g^T + \Psi)$$

- Means and loadings are cluster specific; for parsimony $\Psi_g = \Psi_{g'}$. 
Variable selection, Bayesian inference and model selection.
Variable selection

- Highlight discriminating variables and ease computational burden.

\[ VR_j = \frac{s_{\text{within}}^2}{s_{\text{overall}}^2} = \frac{\sum G_g \sum n_{gi} (z_{ij} - z_{gj})^2}{\sum N_i (z_{ij} - z_j)^2} \]

Small values of \( VR_j \) indicate that variable \( j \) discriminates between clusters. If \( VR_j > \) then variable \( j \) is dropped from the model.
Variable selection

- Highlight discriminating variables and ease computational burden.
- Compare within cluster variance to overall variance for each variable.

\[ VR_j = \frac{s_{\text{within}}^2}{s_{\text{overall}}^2} = \frac{\sum_G \sum_{n_g} (z_{ij} - z_{gj})^2}{\sum_i (z_{ij} - z_{j})^2} \]  

Small values of \( VR_j \) indicate that variable \( j \) discriminates between clusters. If \( VR_j > \) then variable \( j \) is dropped from the model.
Variable selection

- Highlight discriminating variables and ease computational burden.
- Compare within cluster variance to overall variance for each variable.

\[
VR_j = \frac{S_{\text{within}}^2}{S_{\text{overall}}^2} = \frac{\sum_g \sum_i n_g (z_{ij} - \bar{z}_{gj})^2}{\sum_i (z_{ij} - \bar{z}_j)^2}
\]

Small values of \(VR_j\) indicate that variable \(j\) discriminates between clusters. If \(VR_j > \) then variable \(j\) is dropped from the model.
Variable selection

- Highlight discriminating variables and ease computational burden.

- Compare within cluster variance to overall variance for each variable.

\[ VR_j = \frac{S_{\text{within}}^2}{S_{\text{overall}}^2} = \frac{\sum_g \sum_i^{n_g} (z_{ij} - \bar{z}_{gj})^2}{\sum_i^N (z_{ij} - \bar{z}_j)^2} \]

- Small values of \( VR_j \) indicate that variable \( j \) discriminates between clusters.
Variable selection

- Highlight discriminating variables and ease computational burden.

- Compare within cluster variance to overall variance for each variable.

\[
VR_j = \frac{S_{\text{within}}^2}{S_{\text{overall}}^2} = \frac{\sum_g \sum_i^{n_g} (z_{ij} - \bar{z}_g)^2}{\sum_i^N (z_{ij} - \bar{z}_j)^2}
\]

- Small values of \( VR_j \) indicate that variable \( j \) discriminates between clusters.

- If \( VR_j > \tau \) then variable \( j \) is dropped from the model.
For each participant, employ latent indicator variable:

\[ \ell_i \sim \text{Multinomial}(1, \pi) \]
Bayesian inference.

- For each participant, employ latent indicator variable:
  \[ \ell_i \sim \text{Multinomial}(1, \pi) \]

- Conjugate priors leads to Gibbs sampling.
Bayesian inference.

For each participant, employ latent indicator variable:

$$\ell_i \sim \text{Multinomial}(1, \pi)$$

Conjugate priors leads to Gibbs sampling.

Identifiability issues:
For each participant, employ latent indicator variable:

\[ \ell_i \sim \text{Multinomial}(1, \pi) \]

Conjugate priors leads to Gibbs sampling.

Identifiability issues:

1. rotational invariance \(\Rightarrow\) Procrustean rotations employed.
Bayesian inference.

For each participant, employ latent indicator variable:

\[ \ell_i \sim \text{Multinomial}(1, \pi) \]

Conjugate priors leads to Gibbs sampling.

Identifiability issues:

1. rotational invariance \( \Rightarrow \) Procrustean rotations employed.
2. label switching \( \Rightarrow \) minimise loss function.
Incorporating variable selection results in three stage fitting procedure:

1. Burn in phase: Gibbs sampling algorithm with all variables included.
2. Variable selection phase: remove variables for which $VR_j > \ldots$, burn in, repeat until no variables removed at successive checks.
3. Posterior sampling phase: Gibbs sampling algorithm with only discriminating variables included.
Incorporating variable selection results in three stage fitting procedure:

1. **Burn in phase:**
   Gibbs sampling algorithm with all variables included.
Incorporating variable selection results in three stage fitting procedure:

1. **Burn in phase:**
   Gibbs sampling algorithm with all variables included.

2. **Variable selection phase:**
   remove variables for which $VR_j > \tau$, burn in, repeat until no variables removed at successive checks.
Incorporating variable selection results in three stage fitting procedure:

1. **Burn in phase:**
   Gibbs sampling algorithm with all variables included.

2. **Variable selection phase:**
   remove variables for which $VR_j > \tau$, burn in, 
   repeat until no variables removed at successive checks.

3. **Posterior sampling phase:**
   Gibbs sampling algorithm with only discriminating variables included.
Model selection

- Both $G$ and $Q$ are unknown, but standard model selection tools are infeasible.

Likelihood evaluation requires integration of the multidimensional truncated Gaussian distribution, where truncation limits differ and are dependent across the dimensions. Also, different models may have different variable sets.

Let $y_i$ denote the $A$ continuous, $B$ nominal and $C$ binary discriminating variables. And $\_y_i$ the $\_A$ continuous, $\_B$ nominal and $\_C$ binary removed variables.
Model selection

- Both $G$ and $Q$ are unknown, but standard model selection tools are infeasible.

- Likelihood evaluation requires integration of the multidimensional truncated Gaussian distribution, where truncation limits differ and are dependent across the dimensions.
Both $G$ and $Q$ are unknown, but standard model selection tools are infeasible.

Likelihood evaluation requires integration of the multidimensional truncated Gaussian distribution, where truncation limits differ and are dependent across the dimensions.

Also, different models may have different variable sets.
Both $G$ and $Q$ are unknown, but standard model selection tools are infeasible.

Likelihood evaluation requires integration of the multidimensional truncated Gaussian distribution, where truncation limits differ and are dependent across the dimensions.

Also, different models may have different variable sets.

Let $\tilde{y}_i$ denote the $\tilde{A}$ continuous, $\tilde{B}$ nominal and $\tilde{C}$ binary discriminating variables.
Model selection

- Both $G$ and $Q$ are unknown, but standard model selection tools are infeasible.

- Likelihood evaluation requires integration of the multidimensional truncated Gaussian distribution, where truncation limits differ and are dependent across the dimensions.

- Also, different models may have different variable sets.

- Let $\tilde{y}_i$ denote the $\tilde{A}$ continuous, $\tilde{B}$ nominal and $\tilde{C}$ binary discriminating variables.

- And $\check{y}_i$ the $\check{A}$ continuous, $\check{B}$ nominal and $\check{C}$ binary removed variables.
Approximate the observed likelihood:

\[ \tilde{\mathcal{L}}_i = f(\bar{y}_i) f(\bar{y}_j) \]
Model selection

- Approximate the observed likelihood:

\[
\tilde{L}_i = f(\tilde{y}_i)f(\tilde{y}_j) = \sum_{g=1}^{G} \pi_g \left\{ \text{MVN}_{\bar{A}}(\mu_g, \Lambda_g\Lambda_g^T + \Psi) \prod_{j=1}^{B+C} P(\tilde{y}_{ij}|i \in g) \right\}
\]
Approximate the observed likelihood:

\[
\tilde{L}_i = f(\tilde{y}_i)f(\tilde{\hat{y}}_i) = \sum_{g=1}^{G} \pi_g \left\{ \text{MVN}_{\hat{A}}(\mu_g, \Lambda_g \Lambda_g^T + \Psi) \prod_{j=1}^{\hat{B}+\hat{C}} P(\hat{y}_{ij} | i \in g) \right\} \\
\times \left\{ \text{MVN}_{\hat{A}}(\mu, \Lambda \Lambda^T + \Psi) \prod_{j=1}^{\hat{B}+\hat{C}} P(\hat{y}_{ij}) \right\}.
\]

For categorical variables, empirical probabilities are calculated from the observed data.

Incorporate \( \tilde{L} \) in BIC-MCMC (Frühwirth-Schnatter (2011)):

\[
\text{BIC-MCMC} = 2 \log \tilde{L} - \log(N).
\]
Approximate the observed likelihood:

\[ \tilde{L}_i = f(\bar{y}_i) f(\bar{y}_j) = \sum_{g=1}^{G} \pi_g \left\{ \text{MVN}_{\bar{A}}(\mu_g, \Lambda_g \Lambda_g^T + \Psi) \prod_{j=1}^{B+C} P(\bar{y}_{ij} | i \in g) \right\} \times \left[ \text{MVN}_{\bar{A}}(\mu, \Lambda \Lambda^T + \Psi) \prod_{j=1}^{B+C} P(\bar{y}_{ij}) \right]. \]

For categorical variables, empirical probabilities are calculated from the observed data.
Model selection

- Approximate the observed likelihood:

\[ \tilde{L}_i = f(\bar{y}_i)f(\bar{\bar{y}}_i) = \left[ \sum_{g=1}^{G} \pi_g \left\{ \text{MVN}(\mu_g, \Lambda_g\Lambda_g^T + \Psi) \prod_{j=1}^{\tilde{B}+\tilde{C}} P(\bar{y}_{ij} | i \in g) \right\} \right] \]

\[ \times \left[ \text{MVN}(\mu, \Lambda\Lambda^T + \Psi) \prod_{j=1}^{\tilde{B}+\tilde{C}} P(\bar{y}_{ij}) \right]. \]

- For categorical variables, empirical probabilities are calculated from the observed data.

- Incorporate \( \tilde{L} \) in BIC-MCMC (Frühwirth-Schnatter (2011)):

\[ \text{BIC-MCMC} = 2 \times \log \tilde{L} - \nu \times \log(N) \]
Application to the LIPGENE-SU.VI.MAX cohort.
The optimal model: $G = 2$ and $Q = 8$. Of the $J = 738$ original variables, 25 are retained: 12 phenotypic, 11 nominal SNPs and 2 binary SNPs.
The optimal model: $G = 2$ and $Q = 8$.

Of the $J = 738$ original variables, 25 are retained: 12 phenotypic, 11 nominal SNPs and 2 binary SNPs.
Phenotypic cluster means

- Waist Circumference
- Triglycerides
- SBP
- Plasma C18.3−n6
- Plasma C18.1
- Plasma C16.1
- Glucose
- DBP
- Cholesterol
- BMI
- Apo B
- Apo A−1

Cluster 1 (µ1)
Cluster 2 (µ2)
Binary SNP cluster means

ADD1 (rs17777371)

OLR1 (rs1050289)

Cluster 1 (µ₁)
Cluster 2 (µ₂)
## SNP interpretations

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Associated biological pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD1</td>
<td>rs17777371</td>
<td>Blood pressure regulation</td>
</tr>
<tr>
<td>APOB</td>
<td>rs512535</td>
<td>Lipid metabolism</td>
</tr>
<tr>
<td>APOL1</td>
<td>rs136147</td>
<td>Lipid metabolism</td>
</tr>
<tr>
<td>CETP</td>
<td>rs4784744</td>
<td>Lipid metabolism</td>
</tr>
<tr>
<td>GYS1</td>
<td>rs2270938</td>
<td>Glucose homeostasis</td>
</tr>
<tr>
<td>SLC6A14</td>
<td>rs2071877</td>
<td>Amino acid transporter</td>
</tr>
</tbody>
</table>
Correspondence between sub-phenotypes and 7-year follow-up diagnosis
Correspondence between sub-phenotypes and 7-year follow-up diagnosis

<table>
<thead>
<tr>
<th>Initial data</th>
<th>Follow up data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy</td>
<td>MetS</td>
</tr>
<tr>
<td>Cluster 1 (‘Healthy’)</td>
<td>220</td>
<td>42</td>
</tr>
<tr>
<td>Cluster 2 (‘At risk’)</td>
<td>39</td>
<td>204</td>
</tr>
</tbody>
</table>
Correspondence between sub-phenotypes and 7-year follow-up diagnosis

Follow up data
<table>
<thead>
<tr>
<th>Healthy</th>
<th>MetS</th>
</tr>
</thead>
<tbody>
<tr>
<td>220</td>
<td>42</td>
</tr>
<tr>
<td>39</td>
<td>204</td>
</tr>
</tbody>
</table>

Rand index is 0.73 (adjusted Rand = 0.46).
Better than just using the phenotypic abnormality criterion?
Better than just using the phenotypic abnormality criterion?

<table>
<thead>
<tr>
<th>Initial data</th>
<th>Follow up data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>194</td>
</tr>
<tr>
<td>MetS</td>
<td>65</td>
</tr>
</tbody>
</table>

- Rand index: 0.69 (adjusted Rand: 0.38).
Better than just using the phenotypic abnormality criterion?

Follow up data

<table>
<thead>
<tr>
<th></th>
<th>Healthy</th>
<th>MetS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial data</td>
<td>Healthy</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>MetS</td>
<td>65</td>
</tr>
</tbody>
</table>

- Rand index: 0.69 (adjusted Rand: 0.38).
- Highlights the importance of utilising *both* phenotypic and genotypic factors.
- Suggests potential utility of early screening.
Quantifying sub-phenotype membership uncertainty

- Synonymous with concepts of precision medicine & nutrition.
Quantifying sub-phenotype membership uncertainty

- Synonymous with concepts of precision medicine & nutrition.
Assessing model fit

- Use Bayesian residuals & Bayesian latent residuals.
Assessing model fit

- Use Bayesian residuals & Bayesian latent residuals.

- Eg. Density estimates of the Bayesian latent residuals for the rs17777371 SNP for 50 randomly selected participants.
MFA-MD provides a method to cluster high dimensional data of mixed type in their innate form.

Proposed approach can incorporate variable and model selection.

Proposed method has applicability in any similar setting.
Discussion and further work

- MFA-MD provides a method to cluster high dimensional data of mixed type in their innate form.
- Proposed approach can incorporate variable and model selection.
- Proposed method has applicability in any similar setting.
- Highlighted influence of phenotypic and genotypic factors in the MetS.
- Highlighted the importance of early screening.
- Provides a tool to enable precision medicine.
Discussion and further work

- Include other variable types e.g. count

- More model flexibility eg \( Q_g \neq Q_{g'} \).

- Adapt to model longitudinal data.

- Variational approach to estimation should improve efficiency.
Discussion and further work

- Include other variable types e.g. count

- More model flexibility eg $Q_g \neq Q_{g'}$.

- Adapt to model longitudinal data.

- Variational approach to estimation should improve efficiency.

- Incorporate covariates such as gender etc.

- Improved approach to dealing with missing data in the LIPGENE-SU.VI.MAX cohort.


(Grant no: 09/RFP/MTH2367)