Insight

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.

What's coming up...

- 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.'*

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.'*
- 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.

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 concentration	$\geq$5.5 mmol l^{-1}
Serum TAG concentration	$\geq$ 1.5 mmol l^{-1}
Serum HDL-c concentration	$<$ 1.04 mmol l^{-1} (Men) $<$ 1.29 mmol l^{-1} (Women)
Blood pressure	Systolic BP $\geq$ 130 mm Hg Diastolic BP $\geq$ 85 mm Hg
Waist circumference	$>$ 94 cm (Men) $>$ 80 cm (Women)

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. `rs512535` $\in$ {AA, GG, AG}
- Binary genetic SNP data.
  - $C = 371$ SNPs with 2 levels.
  - Eg. `rs17777371` $\in$ {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.

State of the art

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

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), . . .

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:
  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) . . .
- 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:
  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) ...
- 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 $\rightarrow$ item response theory model.
  - Nominal data $\rightarrow$ mutinomial probit model.

Clustering data of mixed type.

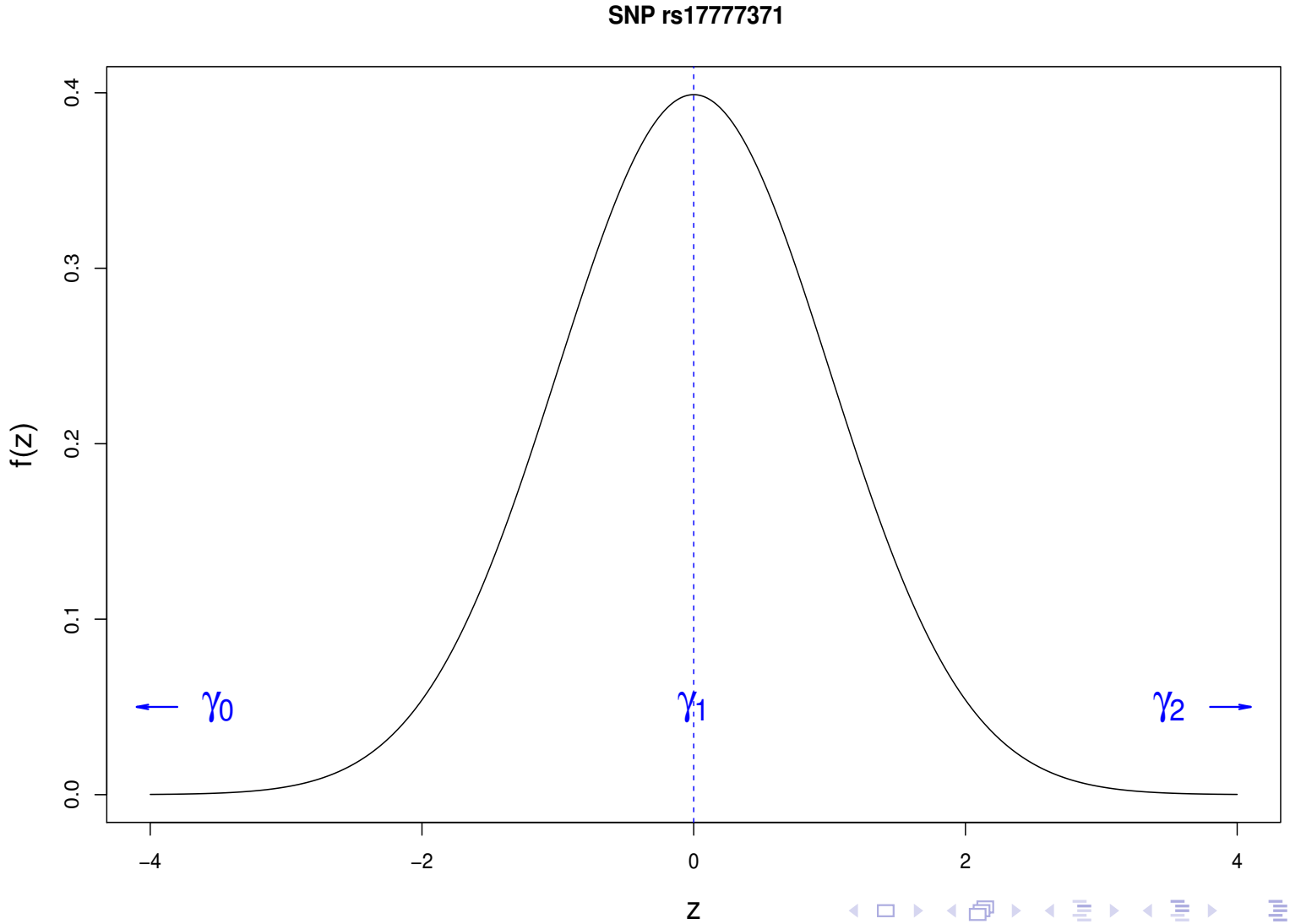
- Discovering clustering structure when we have mixed data i.e. binary, nominal and continuous variables.
- (Categorical) data are high dimensional.
- Draw on ideas from item response theory and latent variable models.
- Three data types:
  - Binary data $\rightarrow$ item response theory model.
  - Nominal data $\rightarrow$ mutinomial probit model.
  - Continuous data $\rightarrow$ factor analysis.

Binary data: item response theory model.

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

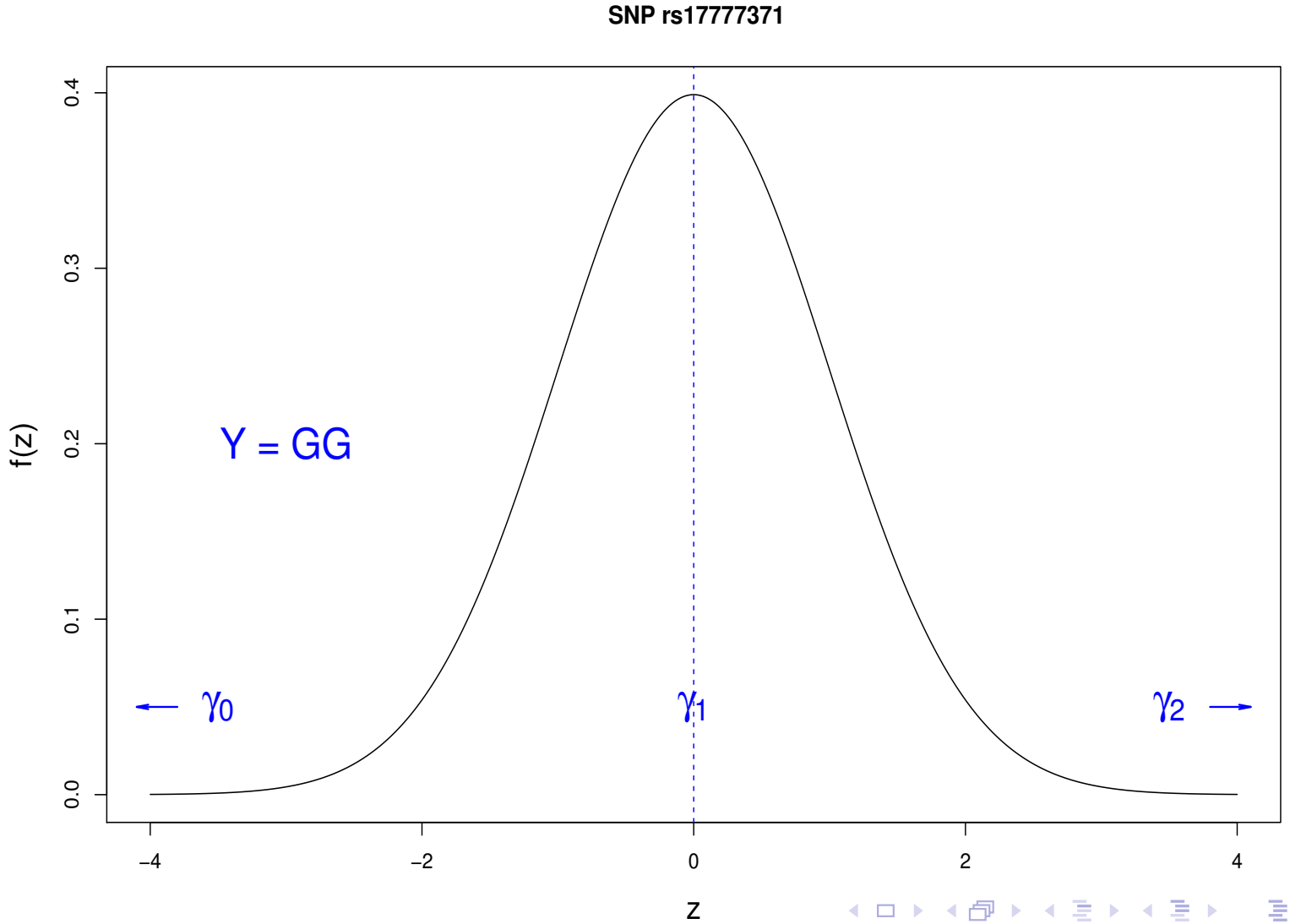
Binary data: item response theory model.

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



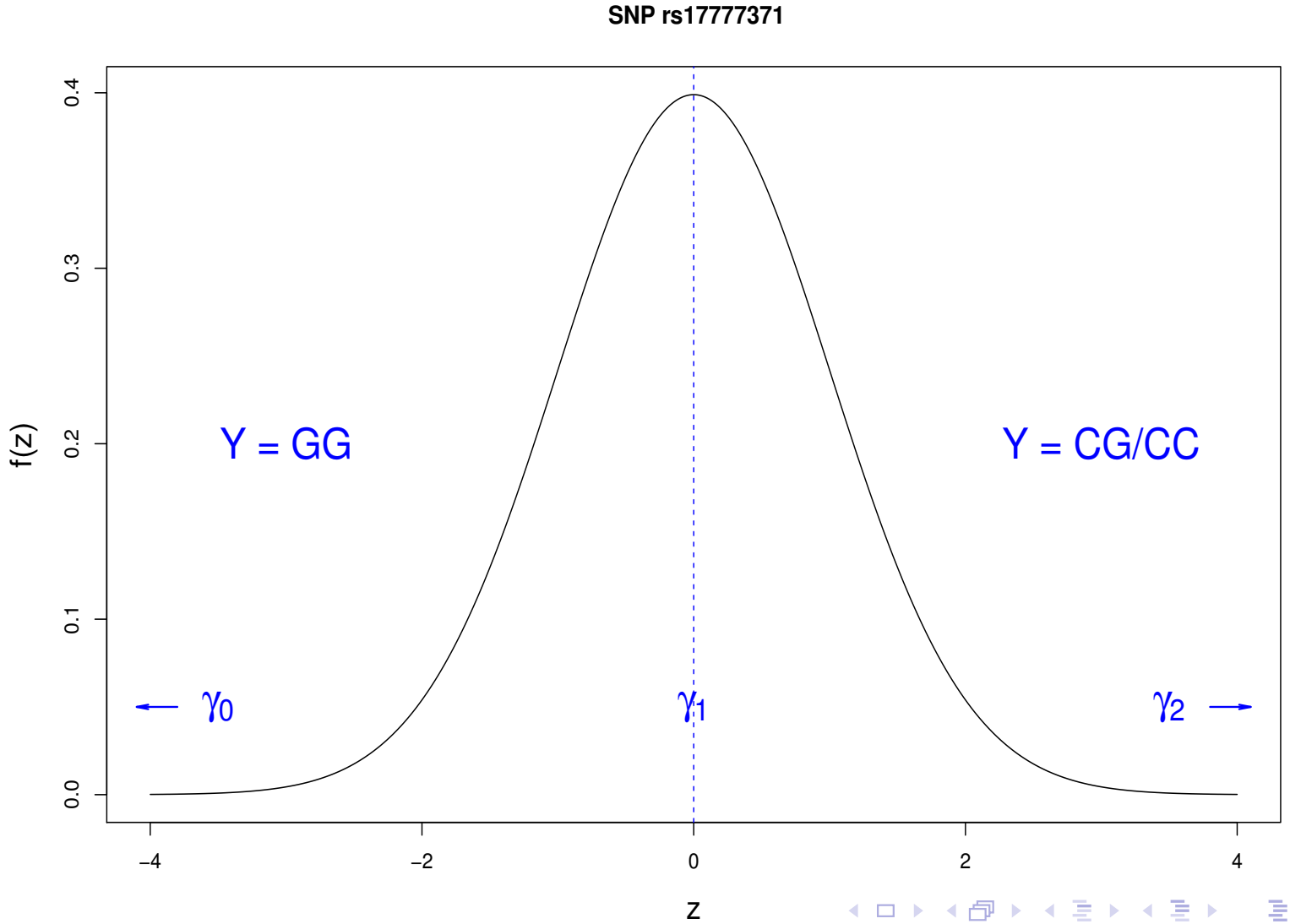
Binary data: item response theory model.

- 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}.



Binary data: item response theory model.

- The binary response y_{ij} serves as an indicator of z_{ij}:

$$\text{if } y_{ij} = k \qquad \text{then} \qquad \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}$$

Binary data: item response theory model.

- The binary response y_{ij} serves as an indicator of z_{ij}:

$$\text{if } y_{ij} = k \qquad \text{then} \qquad \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}$$

- For threshold parameters $\underline{\gamma}_j$ for variable j with $K_j = 2$ levels:

$$\gamma_{j,0} \leq \gamma_{j,1} \leq \gamma_{j,2}$$

Binary data: item response theory model.

- The binary response y_{ij} serves as an indicator of z_{ij}:

$$\text{if } y_{ij} = k \qquad \text{then} \qquad \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}$$

- For threshold parameters $\underline{\gamma}_j$ for variable j with $K_j = 2$ levels:

$$\gamma_{j,0} \leq \gamma_{j,1} \leq \gamma_{j,2}$$

- Identifiability:

$$\gamma_{j,0} = -\infty \qquad \gamma_{j,1} = 0 \qquad \gamma_{j,K_j} = \infty$$

Item response theory model: factor analytic structure.

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

Item response theory model: factor analytic structure.

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

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$\underline{\mu}$		C-vector of negative item difficulty parameters
Λ		$C \times Q$ matrix of item discrimination parameters
$\underline{\theta}_i$	$\sim$	$MVN_Q(\underline{0}, \mathbf{I})$
$\underline{\epsilon}_i$	$\sim$	$MVN_C(\underline{0}, \mathbf{I})$

Item response theory model: factor analytic structure.

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

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$\underline{\mu}$		C-vector of negative item difficulty parameters
Λ		$C \times Q$ matrix of item discrimination parameters
$\underline{\theta}_i$	$\sim$	$MVN_Q(\underline{0}, \mathbf{I})$
$\underline{\epsilon}_i$	$\sim$	$MVN_C(\underline{0}, \mathbf{I})$

- Dimension Q of the latent trait $\underline{\theta}_i$ is unknown, but $Q \ll C$.

$$\underline{z}_i | \underline{\theta}_i \sim \text{MVN}_C(\underline{\mu} + \Lambda \underline{\theta}_i, \mathbf{I})$$

Nominal data: multinomial probit model.

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

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.

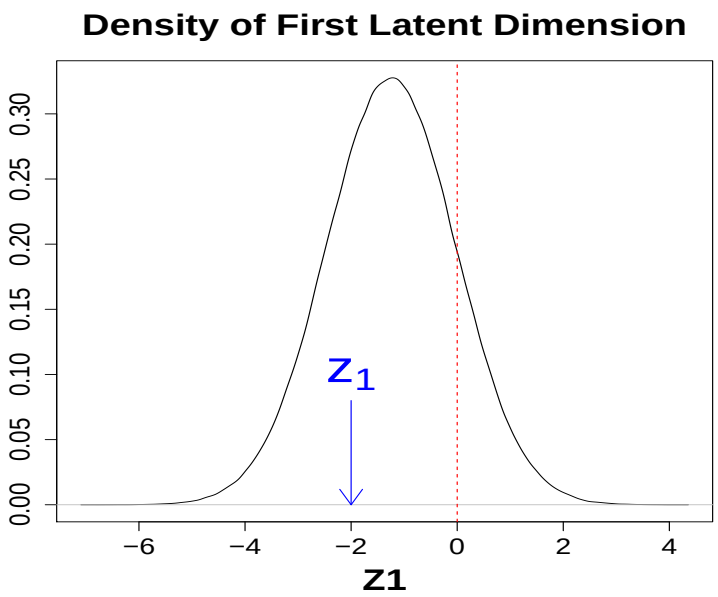
Nominal data: multinomial probit model.

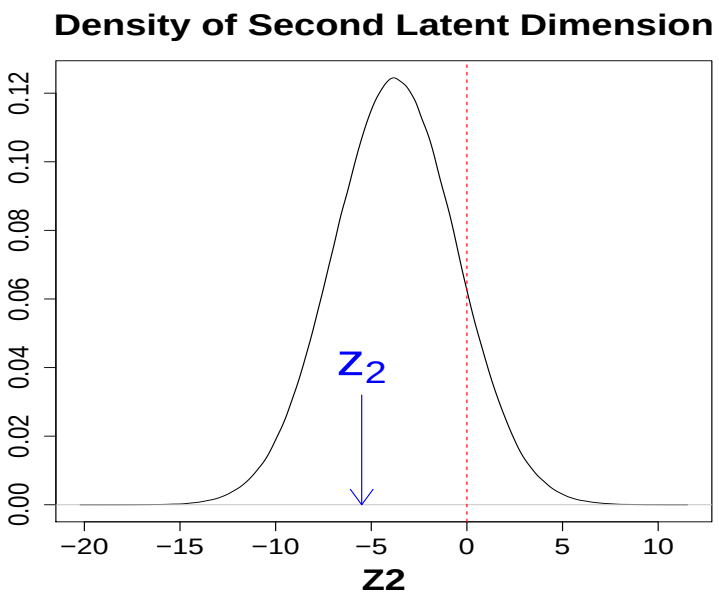
- Underlying y_{ij} are $K_j - 1$ latent Gaussian variables $\{z_{ij}^k\}$.
- Each observed nominal SNP y_{ij} has $K_j = 3$ levels.
- Example: SNP `rs512535` $\in$ {AA, GG, AG}. Thus,

$$\underline{z}_{ij} = \{z_{ij}^1, z_{ij}^2\}$$

Nominal data: multinomial probit model.

Damien:

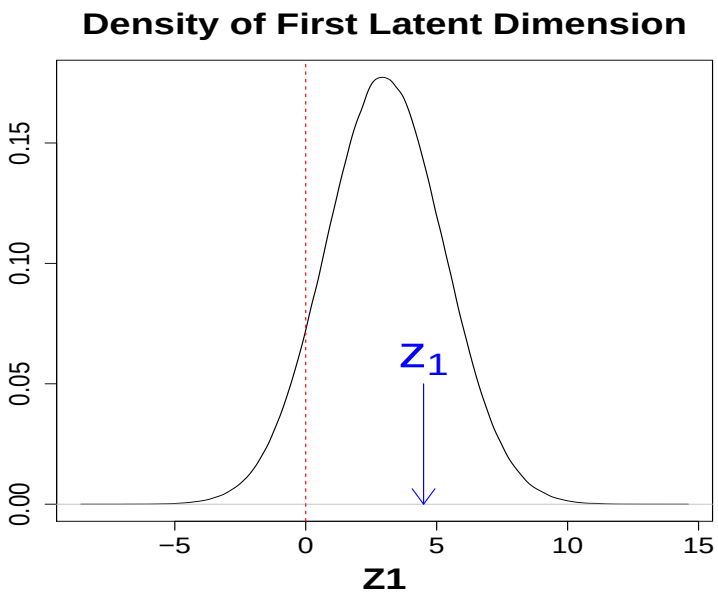


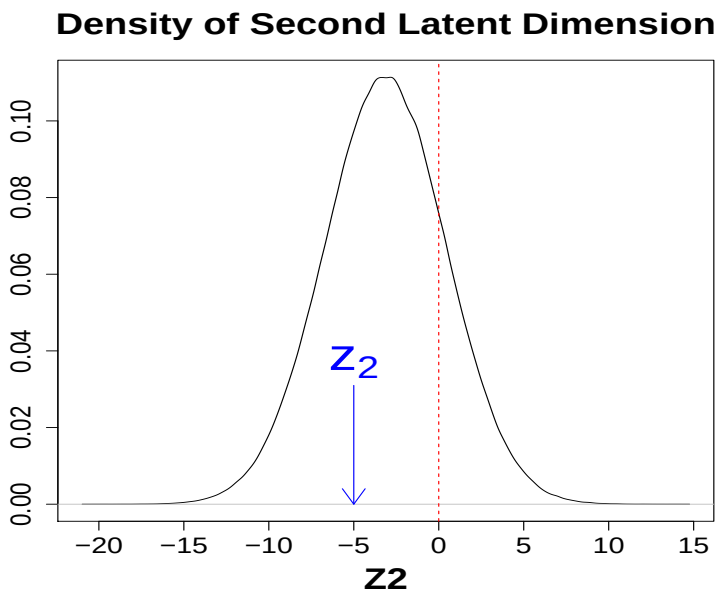


$\Rightarrow$ AA

Nominal data: multinomial probit model.

Lorraine:

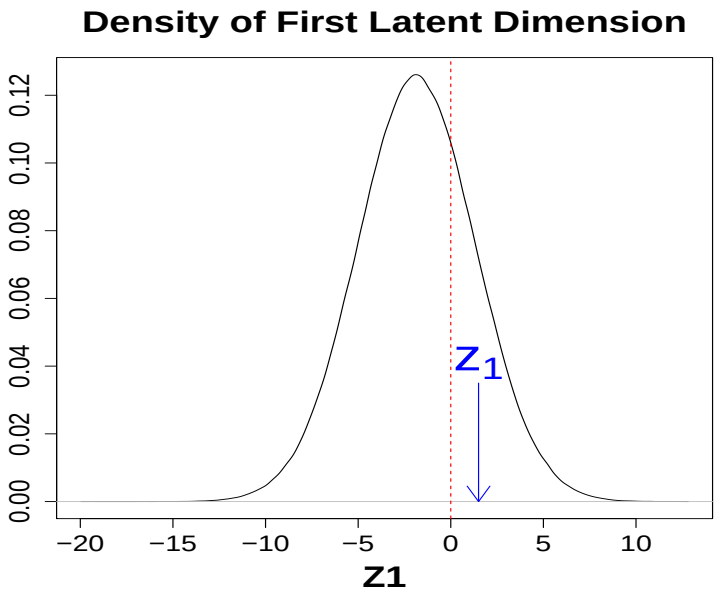


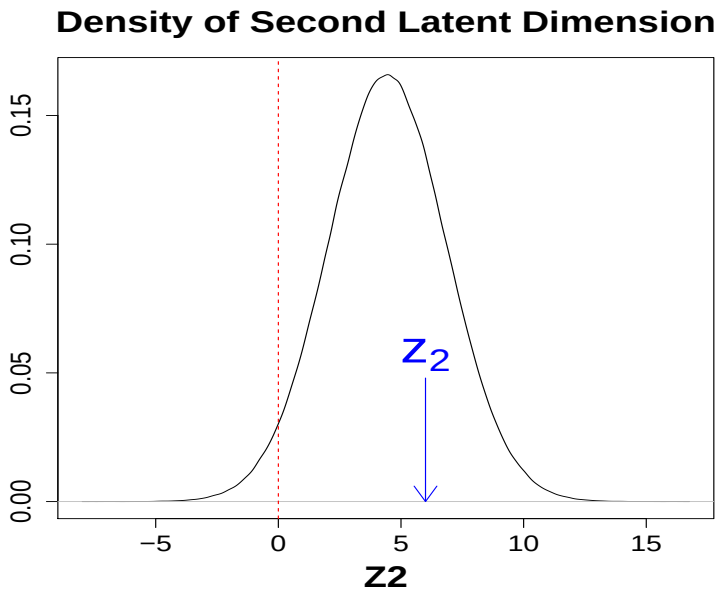


$\Rightarrow$ GG

Nominal data: multinomial probit model.

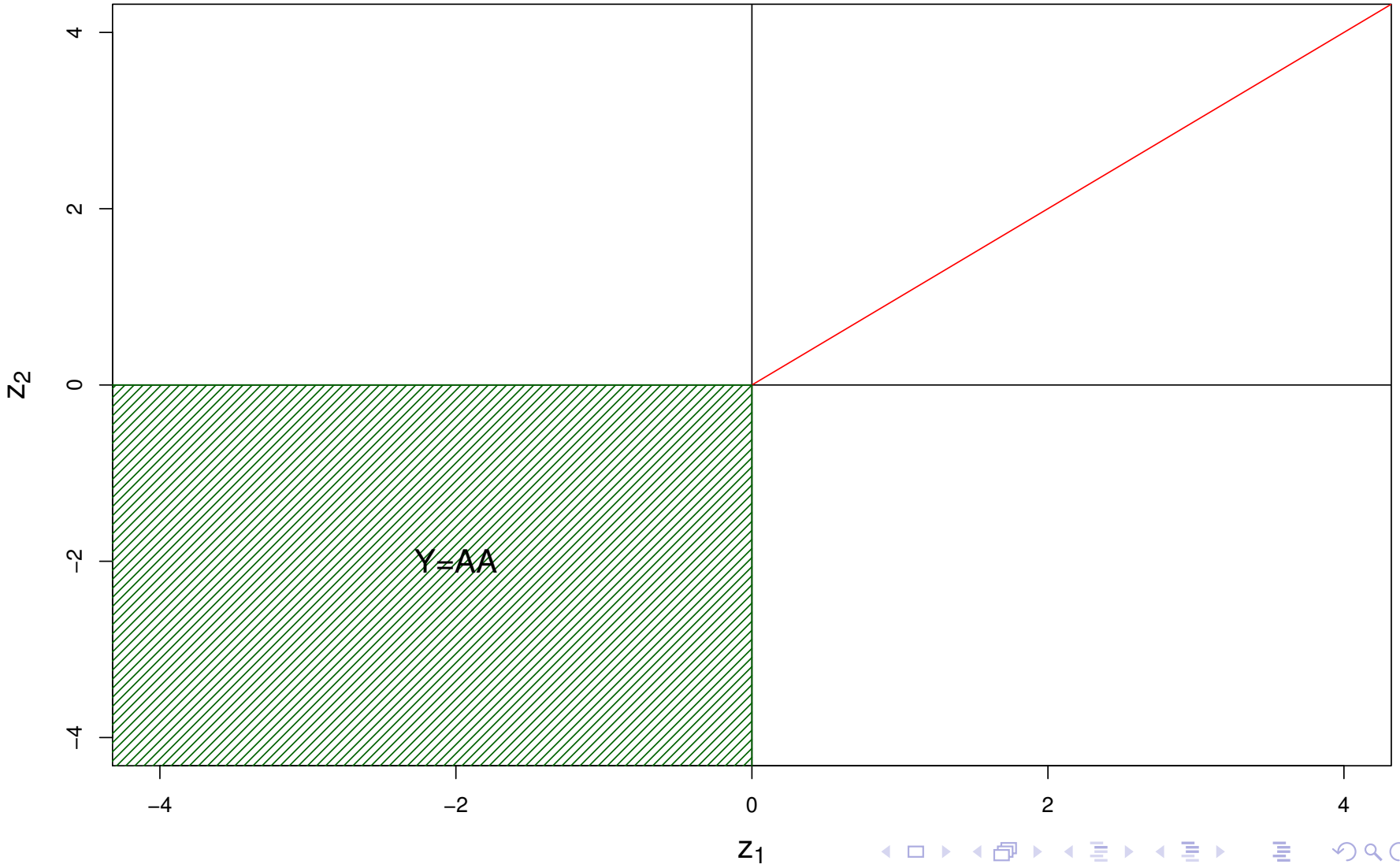
Claire:



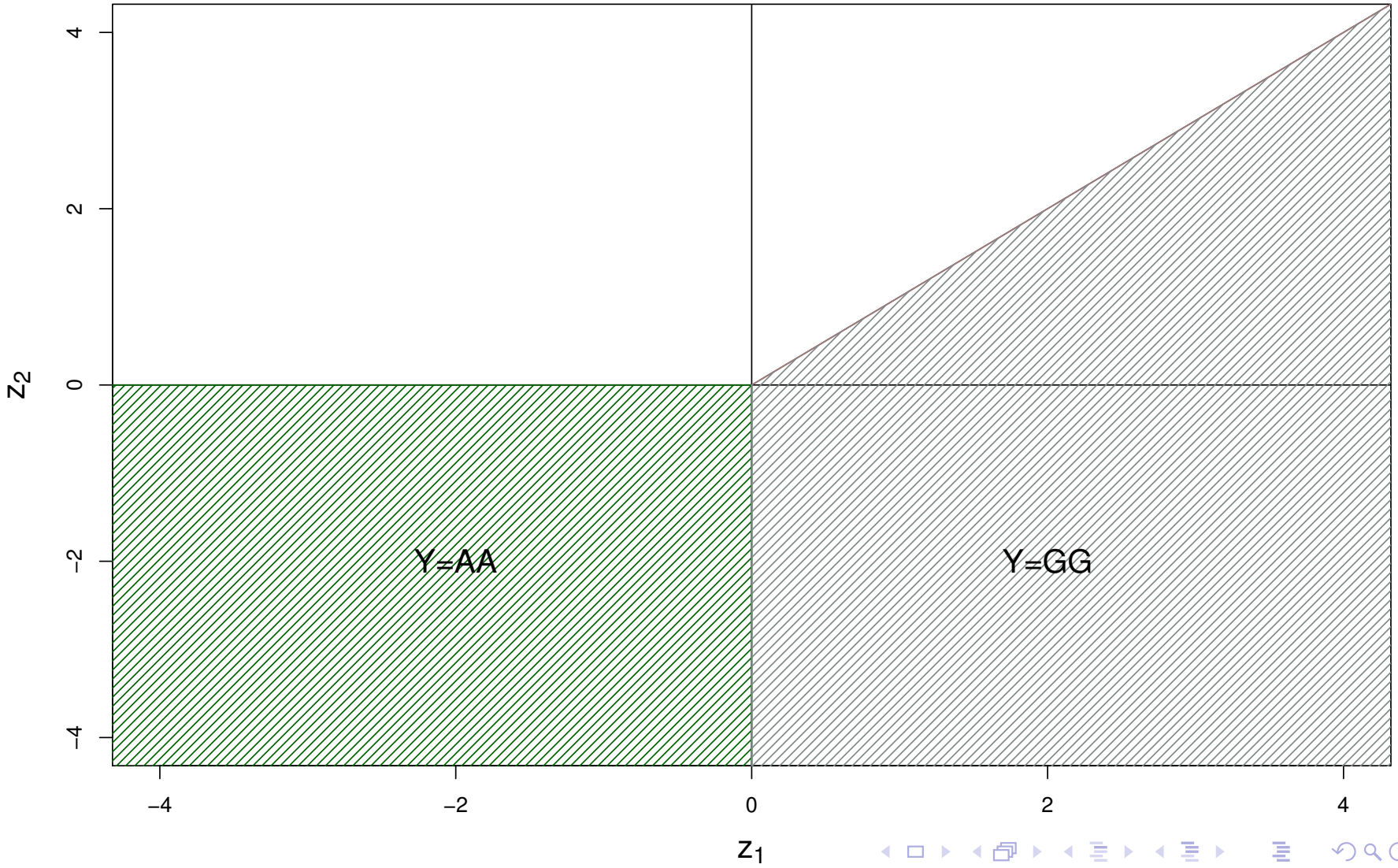


$\Rightarrow$ AG

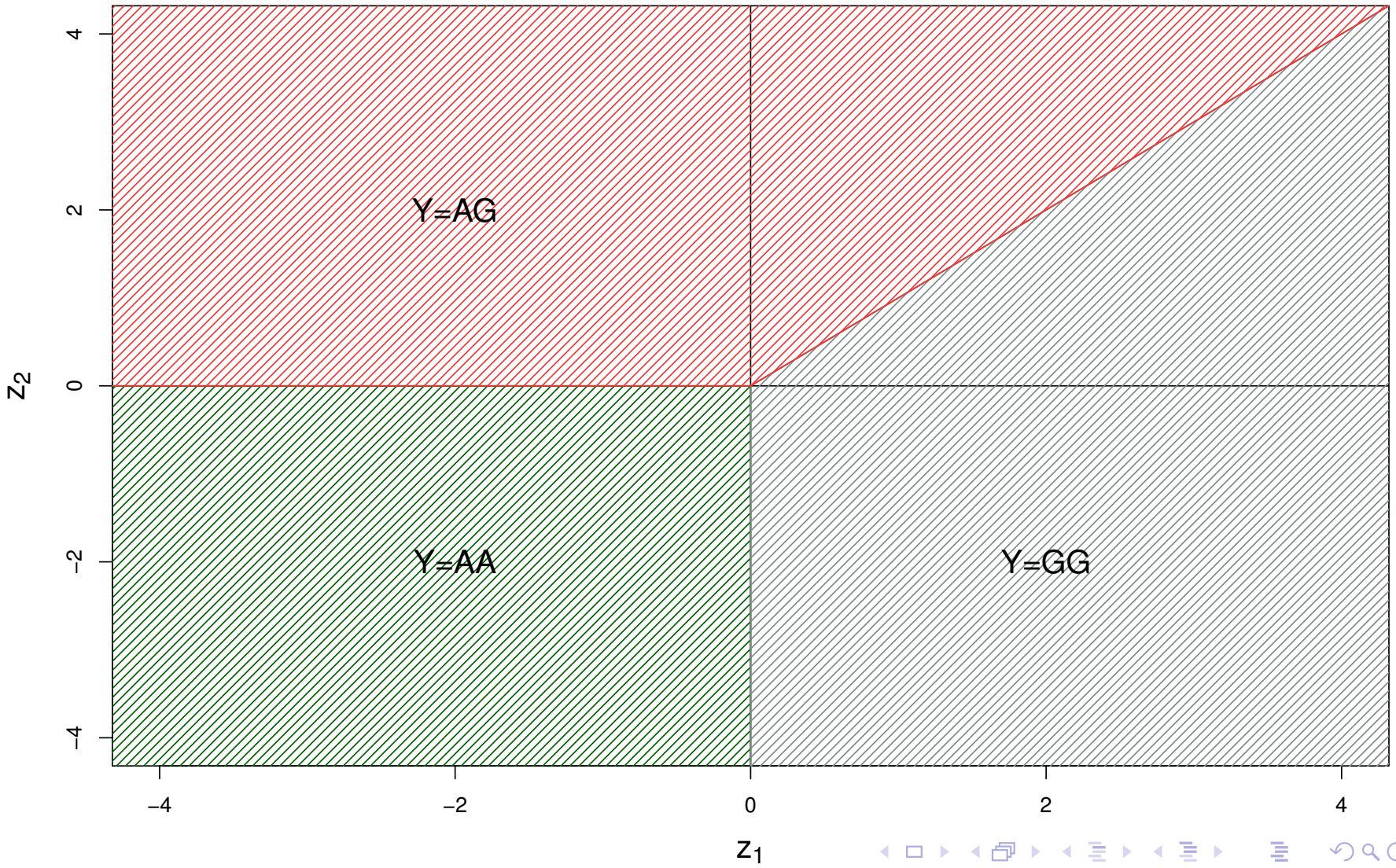
Another view...



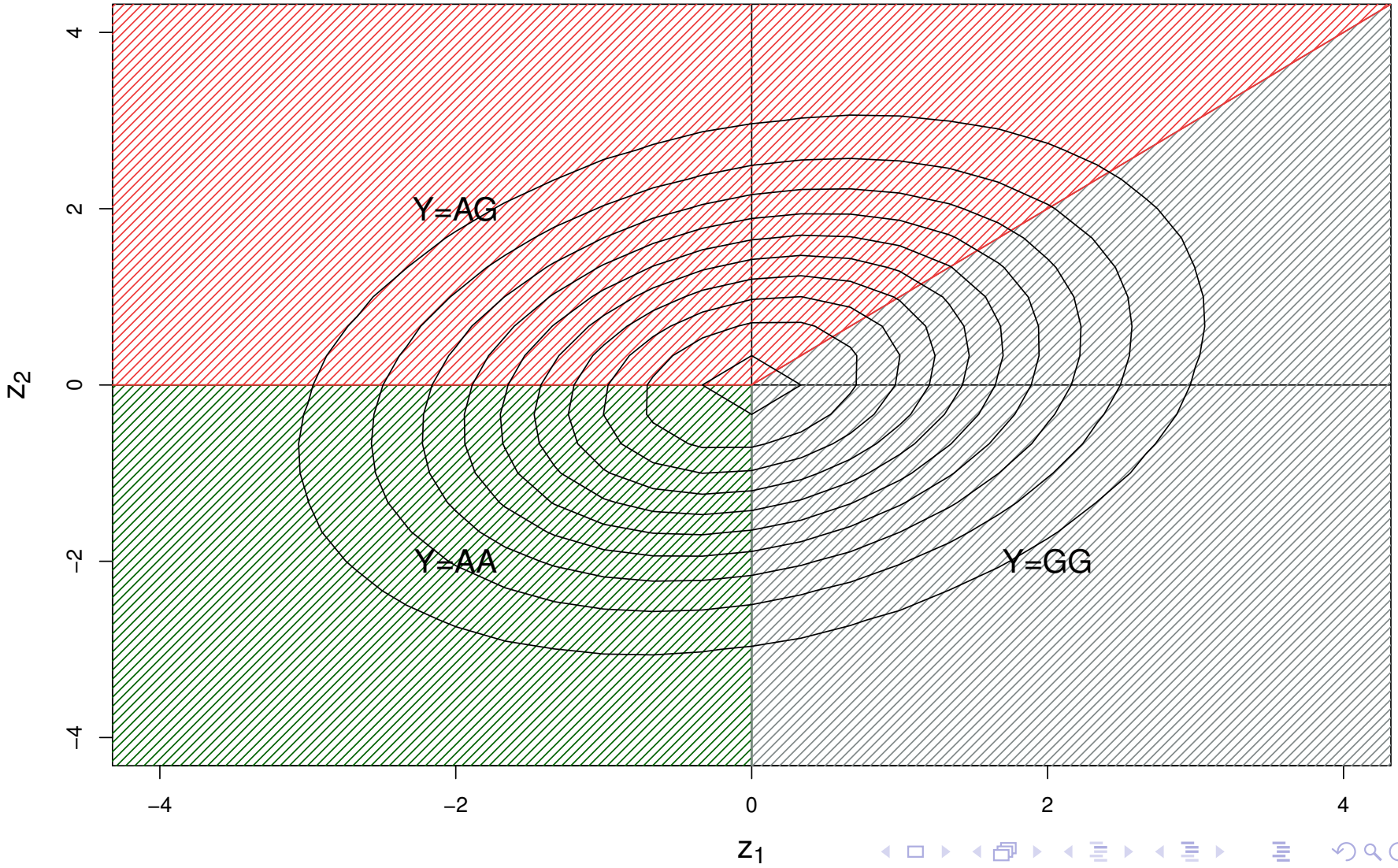
Another view...



Another view...



Another view...



Nominal data: multinomial probit model.

- Nominal response y_{ij} serves as an indicator of $(z_{ij}^1, z_{ij}^2)^T$:

Nominal data: multinomial probit model.

- Nominal response y_{ij} serves as an indicator of $(z_{ij}^1, z_{ij}^2)^T$:
  - if $y_{ij} = AA$ then $(z_{ij}^1, z_{ij}^2)^T < (0, 0)^T$.

Nominal data: multinomial probit model.

- Nominal response y_{ij} serves as an indicator of $(z_{ij}^1, z_{ij}^2)^T$:
  - if $y_{ij} = AA$ then $(z_{ij}^1, z_{ij}^2)^T < (0,0)^T$.
  - if $y_{ij} = GG$ then $z_{ij}^1 = \max_k \{z_{ij}^k\}$ and $z_{ij}^1 > 0$.

Nominal data: multinomial probit model.

- Nominal response y_{ij} serves as an indicator of $(z_{ij}^1, z_{ij}^2)^T$:
  - if $y_{ij} = AA$ then $(z_{ij}^1, z_{ij}^2)^T < (0,0)^T$.
  - if $y_{ij} = GG$ then $z_{ij}^1 = \max_k \{z_{ij}^k\}$ and $z_{ij}^1 > 0$.
  - if $y_{ij} = AG$ then $z_{ij}^2 = \max_k \{z_{ij}^k\}$ and $z_{ij}^2 > 0$.

Multinomial probit model: factor analytic structure.

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

Multinomial probit model: factor analytic structure.

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

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$$\begin{array}{lcl} \underline{\mu} & & 2B \text{ dimensional mean vector.} \\ \Lambda & & 2B \times Q \text{ loadings matrix} \\ \underline{\theta}_i & \sim & MVN_Q(\underline{0}, \mathbf{I}) \\ \underline{\epsilon}_i & \sim & MVN_{2B}(\underline{0}, \mathbf{I}) \end{array}$$

Multinomial probit model: factor analytic structure.

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

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$$\begin{array}{lcl} \underline{\mu} & & 2B \text{ dimensional mean vector.} \\ \Lambda & & 2B \times Q \text{ loadings matrix} \\ \underline{\theta}_i & \sim & MVN_Q(\underline{0}, \mathbf{I}) \\ \underline{\epsilon}_i & \sim & MVN_{2B}(\underline{0}, \mathbf{I}) \end{array}$$

- Again, $Q << 2B$ and

$$\underline{z}_i | \underline{\theta}_i \sim \text{MVN}_{2B}(\underline{\mu} + \Lambda \underline{\theta}_i, \mathbf{I})$$

Continuous data: factor analysis model.

- Model $\underline{y}_i = \underline{z}_i = (z_{i1}, \ldots, z_{iA})^T$ as a linear function of a latent, low dimensional Gaussian variable $\underline{\theta}_i$:

Continuous data: factor analysis model.

- Model $\underline{y}_i = \underline{z}_i = (z_{i1}, \ldots, z_{iA})^T$ as a linear function of a latent, low dimensional Gaussian variable $\underline{\theta}_i$:

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$\underline{\mu}$	A dimensional mean vector.
Λ	$A \times Q$ loadings matrix
$\underline{\theta}_i$	$MVN_Q(\underline{0}, \mathbf{I})$
$\underline{\epsilon}_i$	$MVN_A(\underline{0}, \Psi)$

Continuous data: factor analysis model.

- Model $\underline{y}_i = \underline{z}_i = (z_{i1}, \ldots, z_{iA})^T$ as a linear function of a latent, low dimensional Gaussian variable $\underline{\theta}_i$:

$$\underline{z}_i = \underline{\mu} + \Lambda \underline{\theta}_i + \underline{\epsilon}_i$$

where

$\underline{\mu}$	A dimensional mean vector.
Λ	$A \times Q$ loadings matrix
$\underline{\theta}_i$	$MVN_Q(\underline{0}, \mathbf{I})$
$\underline{\epsilon}_i$	$MVN_A(\underline{0}, \Psi)$

- Again, $Q << A$ and

$$\underline{z}_i | \underline{\theta}_i \sim \text{MVN}_A(\underline{\mu} + \Lambda \underline{\theta}_i, \Psi)$$

Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

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:

$$y_{ij} = \begin{cases} z_{ij} & \text{if variable } j \text{ is continuous.} \\ k & \text{if variable } j \text{ is binary and } \gamma_{j,k-1} < z_{ij} \leq \gamma_{j,k}. \end{cases}$$

Hybrid model: factor analysis for mixed data (FA-MD)

- Similar model structure suggests a hybrid may be fruitful:

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

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 $\underline{z}_i$.

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 $\underline{z}_i$.

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

$$\underline{z}_i | \underline{\theta}_i \sim \text{MVN}_D(\underline{\mu} + \Lambda \underline{\theta}_i, \Psi).$$

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 $\underline{z}_i$.

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

$$\underline{z}_i | \underline{\theta}_i \sim \text{MVN}_D(\underline{\mu} + \Lambda \underline{\theta}_i, \Psi).$$

- Marginally, have a parsimonious covariance structure:

$$\underline{z}_i \sim \text{MVN}_D(\underline{\mu}, \Lambda\Lambda^T + \Psi)$$

Hybrid model: factor analysis for mixed data (FA-MD)

- Complex, augmented, likelihood function:

$$\begin{aligned}\mathbb{P}(\underline{y}_i|\underline{\mu},\Lambda,\underline{z}_i,\Theta,\Gamma,\Psi) \quad = \quad & \prod_{j\ \text{cns}} N(\mu_j+\underline{\lambda}_j^T\underline{\theta}_i,\psi_j) \\ & \times \prod_{j\ \text{bin}} N^T(\mu_j+\underline{\lambda}_j^T\underline{\theta}_i,1)\mathbb{I}\{z_{ij}\} \\ & \times \prod_{j\ \text{nom}}\prod_{k=1}^{K_j-1} N^T(\mu_j^k+\underline{\lambda}_j^{kT}\underline{\theta}_i,1)\mathbb{I}\{z_{ij}^k\}\end{aligned}$$

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.

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:

$$\mathbb{P}(\underline{z}_i) = \sum_{g=1}^{G} \pi_g \, \text{MVN}_D(\underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi)$$

- 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.

Variable selection

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

Variable selection

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

$$VR_j = \frac{s^2_{within}}{s^2_{overall}} = \frac{\sum_g^G \sum_i^{n_g} (z_{ij} - \bar{z}_{gj})^2}{\sum_i^N (z_{ij} - \bar{z}_j)^2}$$

Variable selection

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

$$VR_j = \frac{s^2_{within}}{s^2_{overall}} = \frac{\sum_g^G \sum_i^{n_g} (z_{ij} - \bar{z}_{gj})^2}{\sum_i^N (z_{ij} - \bar{z}_j)^2}$$

- 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.

$$VR_j = \frac{s^2_{within}}{s^2_{overall}} = \frac{\sum_g^G \sum_i^{n_g} (z_{ij} - \bar{z}_{gj})^2}{\sum_i^N (z_{ij} - \bar{z}_j)^2}$$

- Small values of VR_j indicate that variable j discriminates between clusters.
- If $VR_j > \tau$ then variable j is dropped from the model.

Bayesian inference.

- For each participant, employ latent indicator variable:

$$\underline{\ell}_i \sim \text{Multinomial}(1, \underline{\pi})$$

Bayesian inference.

- For each participant, employ latent indicator variable:

$$\underline{\ell}_i \sim \text{ Multinomial}(1, \underline{\pi})$$

- Conjugate priors leads to Gibbs sampling.

Bayesian inference.

- For each participant, employ latent indicator variable:

$$\underline{\ell}_i \sim \text{ Multinomial}(1, \underline{\pi})$$

- Conjugate priors leads to Gibbs sampling.
- Identifiability issues:

Bayesian inference.

- For each participant, employ latent indicator variable:

$$\underline{\ell}_i \sim \text{Multinomial}(1, \underline{\pi})$$

- Conjugate priors leads to Gibbs sampling.
- Identifiability issues:
  1. rotational invariance $\Rightarrow$ Procrustean rotations employed.

Bayesian inference.

- For each participant, employ latent indicator variable:

$$\underline{\ell}_i \sim \text{Multinomial}(1, \underline{\pi})$$

- Conjugate priors leads to Gibbs sampling.
- Identifiability issues:
  1. rotational invariance $\Rightarrow$ Procrustean rotations employed.
  2. label switching $\Rightarrow$ minimise loss function.

Bayesian inference.

- Incorporating variable selection results in three stage fitting procedure:

Bayesian inference.

- Incorporating variable selection results in three stage fitting procedure:

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

Bayesian inference.

- 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.

Bayesian inference.

- 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.

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.

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.

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 $\ddot{\underline{y}}_i$ denote the $\ddot{A}$ continuous, $\ddot{B}$ nominal and $\ddot{C}$ 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 $\underline{\ddot{y}}_i$ denote the $\ddot{A}$ continuous, $\ddot{B}$ nominal and $\ddot{C}$ binary discriminating variables.
- And $\underline{\dot{y}}_i$ the $\dot{A}$ continuous, $\dot{B}$ nominal and $\dot{C}$ binary removed variables.

Model selection

- Approximate the observed likelihood:

$$\tilde{\mathcal{L}}_i = f(\underline{\ddot{y}}_i) f(\underline{\dot{y}}_i)$$

Model selection

- Approximate the observed likelihood:

$$\begin{aligned}\tilde{\mathcal{L}}_i &= f(\ddot{\underline{y}}_i) f(\dot{\underline{y}}_i) \\ &= \left[\sum_{g=1}^{G} \pi_g \left\{ \text{MVN}_{\ddot{A}}(\mu_g, \Lambda_g \Lambda_g^T + \Psi) \prod_{j=1}^{\ddot{B}+\ddot{C}} P(\ddot{y}_{ij} | i \in g) \right\} \right]\end{aligned}$$

Model selection

- Approximate the observed likelihood:

$$
\begin{aligned}
\tilde{\mathcal{L}}_i &= f(\underset{\sim}{\ddot{y}}_i) f(\underset{\sim}{\dot{y}}_i) \\
&= \left[\sum_{g=1}^{G} \pi_g \left\{ \text{MVN}_{\ddot{A}}(\mu_g, \Lambda_g \Lambda_g^T + \Psi) \prod_{j=1}^{\ddot{B}+\ddot{C}} P(\ddot{y}_{ij} | i \in g) \right\} \right] \\
&\quad \times \left[\text{MVN}_{\dot{A}}(\mu, \Lambda\Lambda^T + \Psi) \prod_{j=1}^{\dot{B}+\dot{C}} P(\dot{y}_{ij}) \right].
\end{aligned}
$$

Model selection

- Approximate the observed likelihood:

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

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

Model selection

- Approximate the observed likelihood:

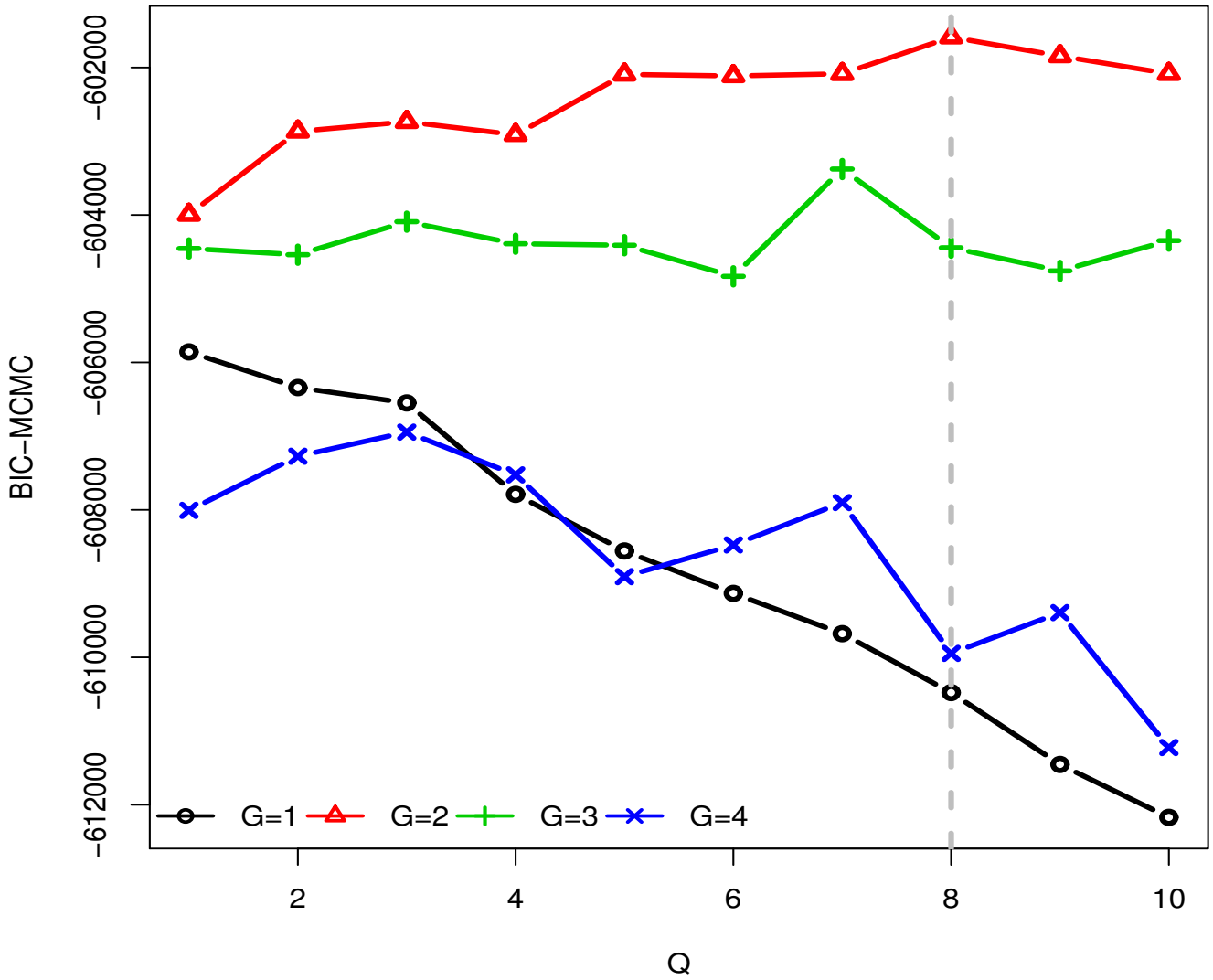
$$
\begin{aligned}
\tilde{\mathcal{L}}_i &= f(\ddot{\underline{y}}_i) f(\dot{\underline{y}}_i) \\
&= \left[\sum_{g=1}^{G} \pi_g \left\{ \mathrm{MVN}_{\ddot{A}}(\mu_g, \Lambda_g \Lambda_g^T + \Psi) \prod_{j=1}^{\ddot{B}+\ddot{C}} P(\ddot{y}_{ij} | i \in g) \right\} \right] \\
&\quad \times \left[\mathrm{MVN}_{\dot{A}}(\mu, \Lambda\Lambda^T + \Psi) \prod_{j=1}^{\dot{B}+\dot{C}} P(\dot{y}_{ij}) \right].
\end{aligned}
$$

- For categorical variables, empirical probabilities are calculated from the observed data.
- Incorporate $\tilde{\mathcal{L}}$ in BIC-MCMC (Frühwirth-Schnatter (2011)):

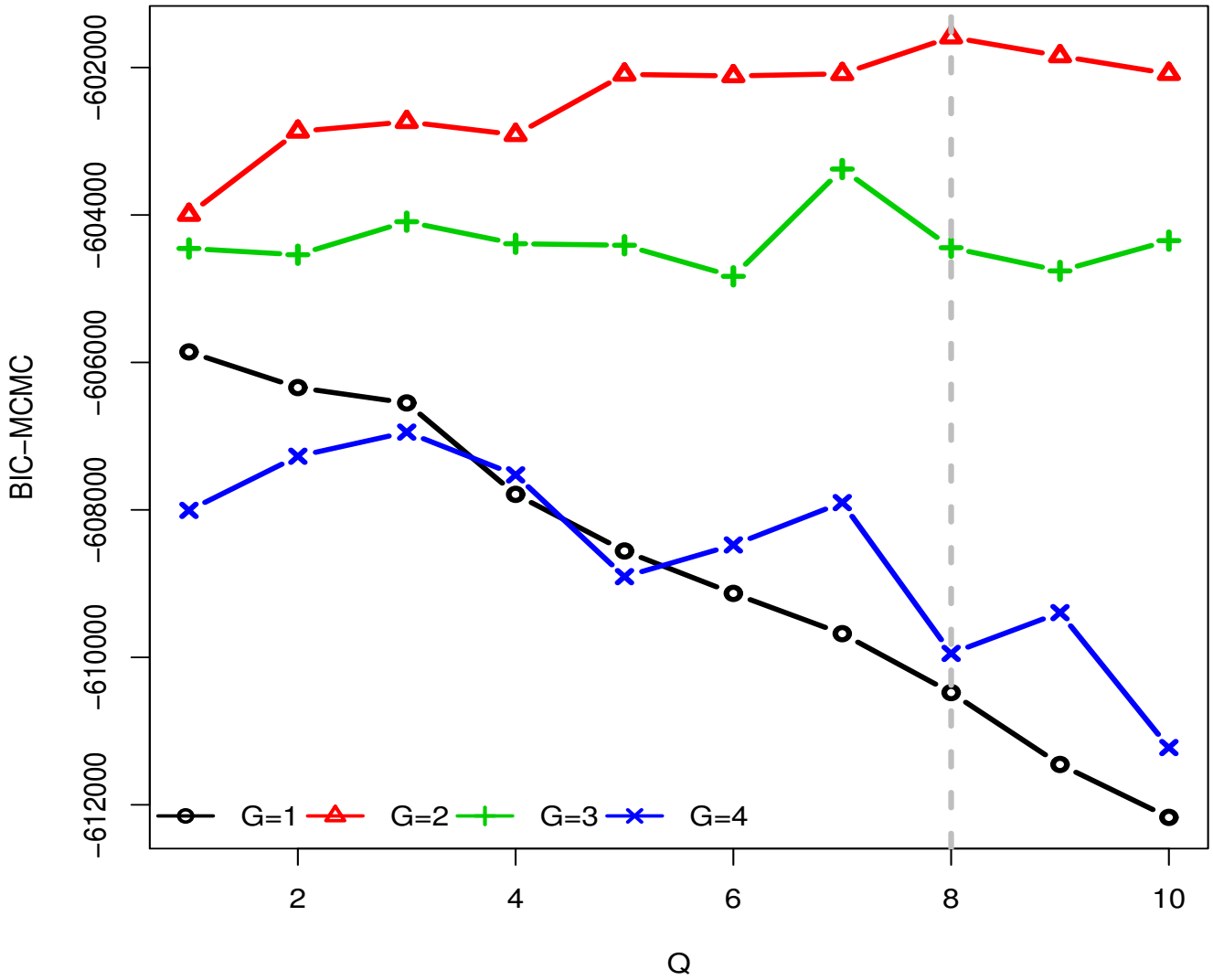
$$\text{BIC-MCMC} = 2 \times \log \tilde{\mathcal{L}} - \nu \times \log(N)$$

Application to the LIPGENE-SU.VI.MAX cohort.

The optimal model: $G = 2$ and $Q = 8$.

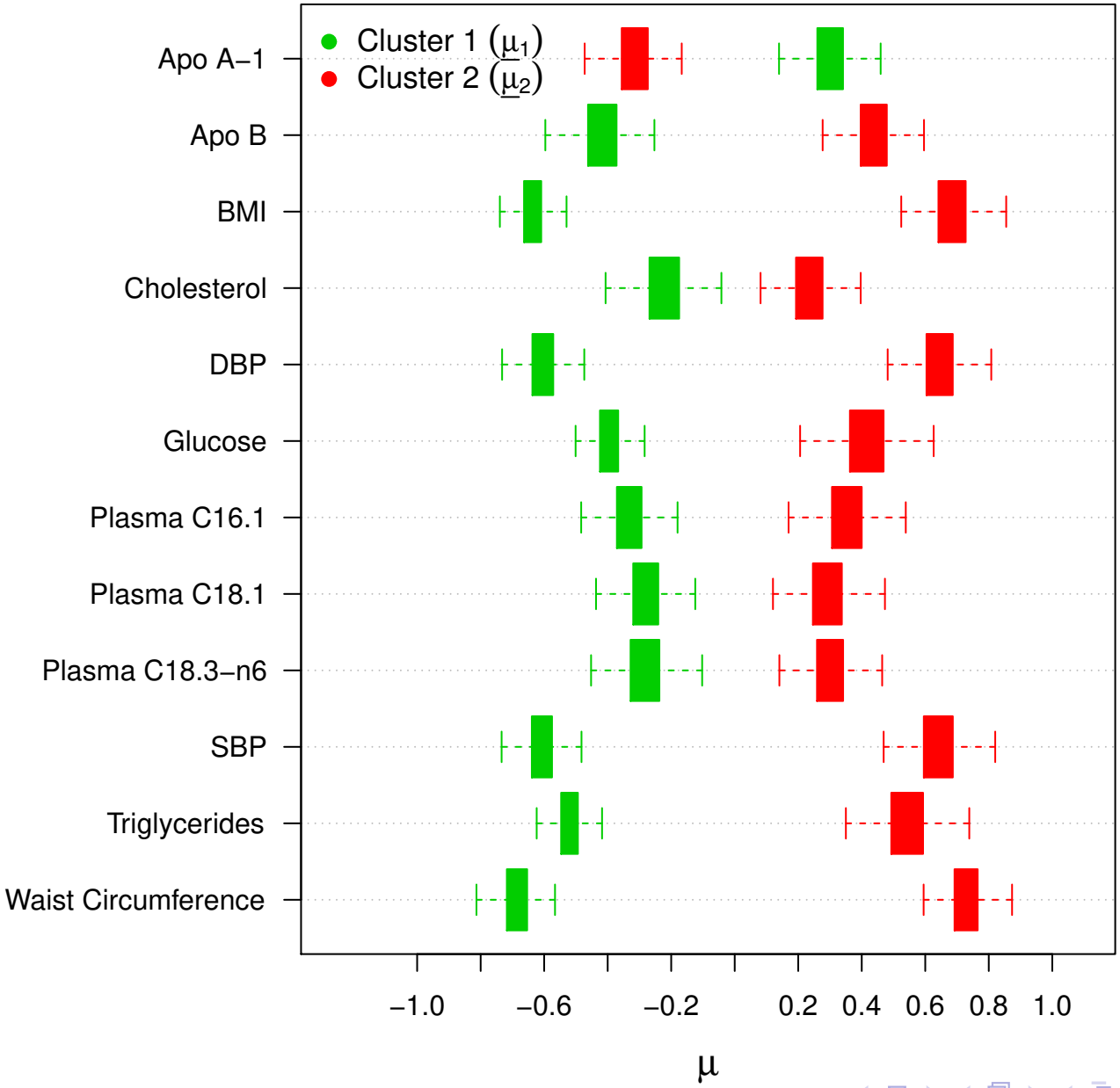


The optimal model: $G = 2$ and $Q = 8$.

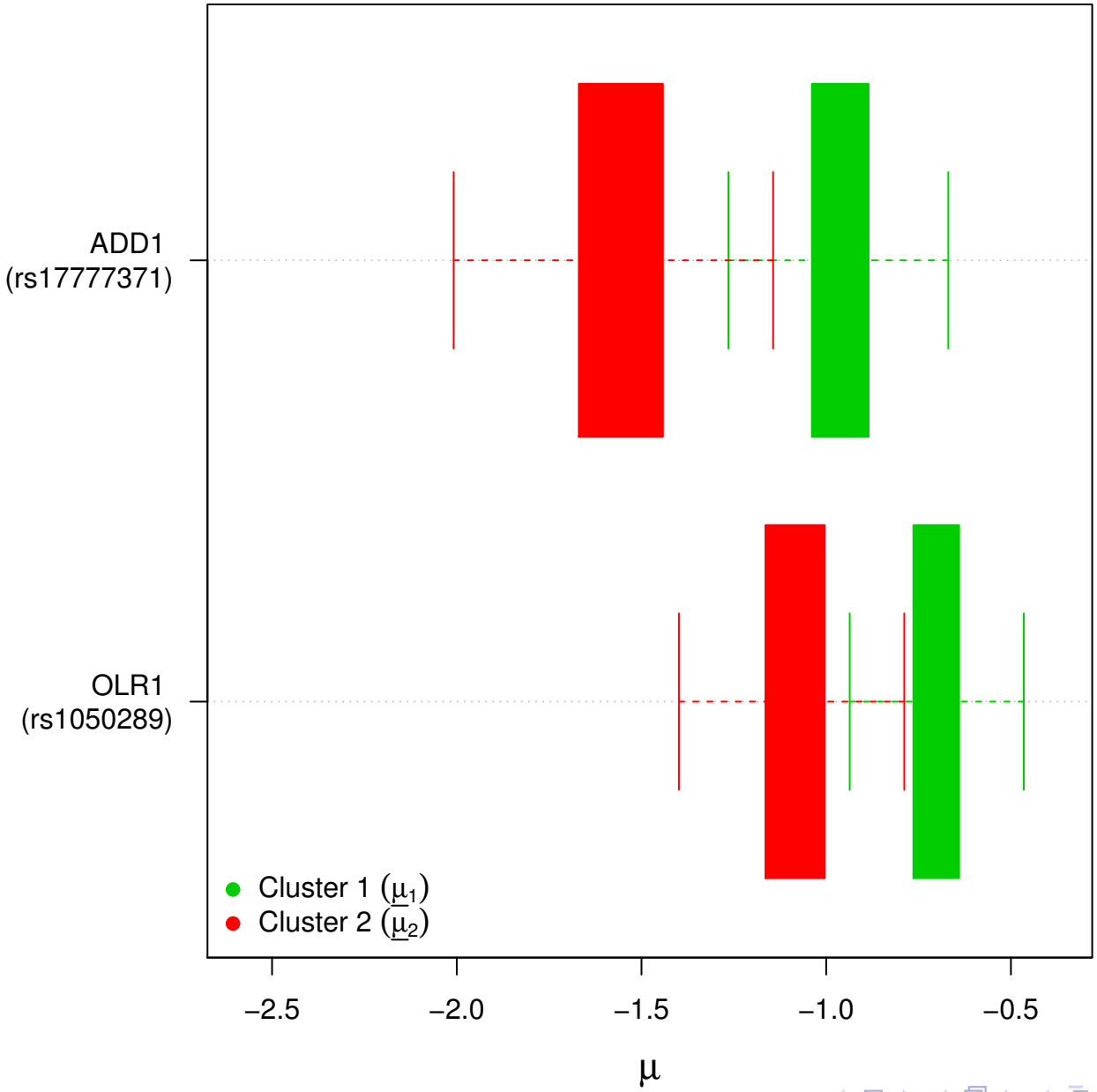


- Of the $J = 738$ original variables, 25 are retained: 12 phenotypic, 11 nominal SNPs and 2 binary SNPs.

Phenotypic cluster means



Binary SNP cluster means



SNP interpretations

Gene	SNP	Associated biological pathway
ADD1	`rs17777371`	Blood pressure regulation
APOB	`rs512535`	Lipid metabolism
APOL1	`rs136147`	Lipid metabolism
CETP	`rs4784744`	Lipid metabolism
GYS1	`rs2270938`	Glucose homeostasis
SLC6A14	`rs2071877`	Amino acid transporter
⋮	⋮	⋮

Correspondence between sub-phenotypes and 7-year follow-up diagnosis

Correspondence between sub-phenotypes and 7-year follow-up diagnosis

		Follow up data	
		Healthy	MetS
Initial data	Cluster 1 (‘Healthy’)	220	42
	Cluster 2 (‘At risk’)	39	204

Correspondence between sub-phenotypes and 7-year follow-up diagnosis

		Follow up data	
		Healthy	MetS
Initial data	Cluster 1 (‘Healthy’)	220	42
	Cluster 2 (‘At risk’)	39	204

- 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?

		Follow up data	
		Healthy	MetS
Initial data	Healthy	194	31
	MetS	65	215

- Rand index: 0.69 (adjusted Rand: 0.38).

Better than just using the phenotypic abnormality criterion?

		Follow up data	
		Healthy	MetS
Initial data	Healthy	194	31
	MetS	65	215

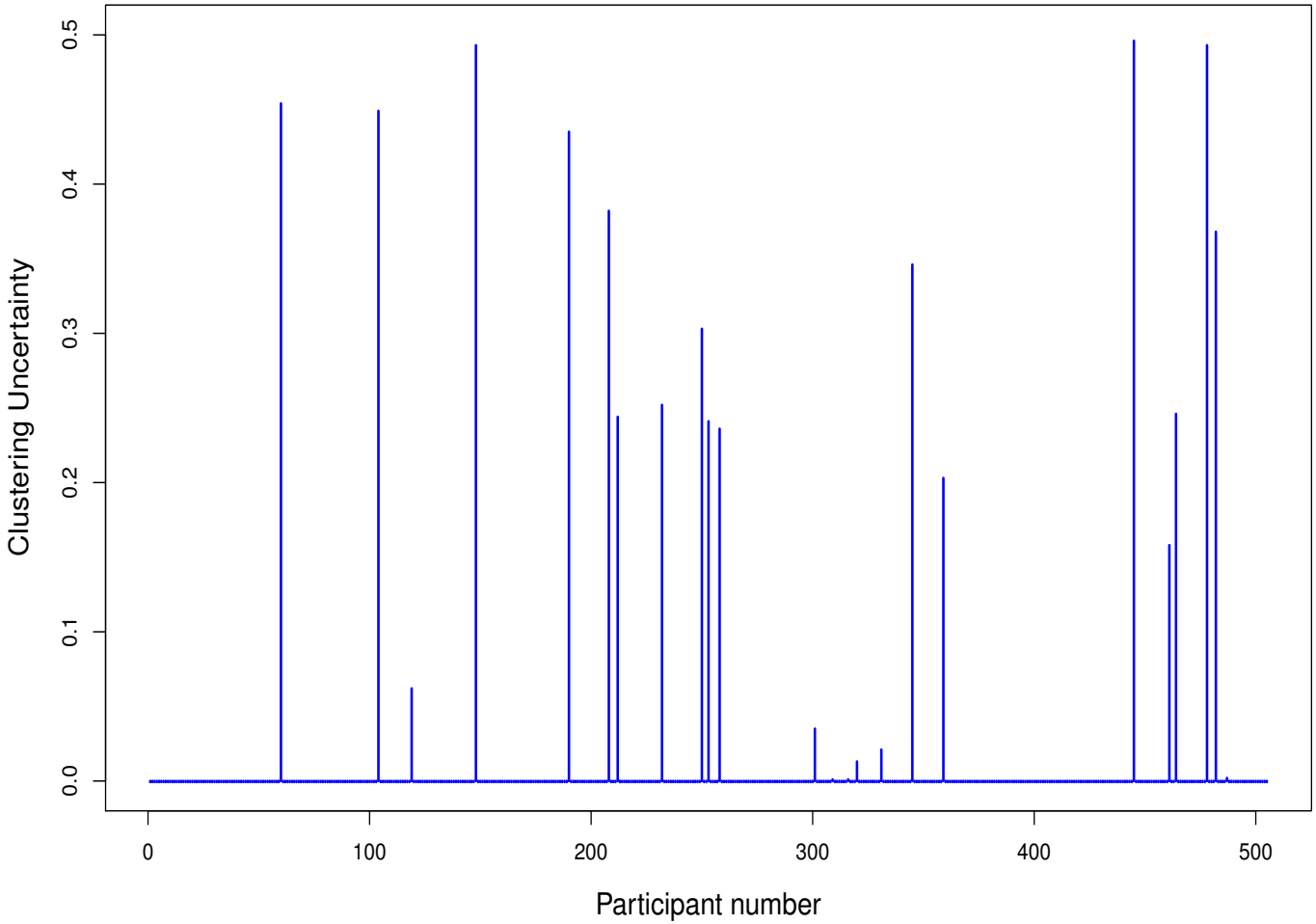
- Rand index: 0.69 (adjusted Rand: 0.38).
- Highlights the importance of utilising *both* phenotypic and genotypic factors.
- Suggests potential utility of early screening.

Quantifying sub-phenotype membership uncertainty

- Synonymous with concepts of precision medicine & nutrition.

Quantifying sub-phenotype membership uncertainty

- Synonymous with concepts of precision medicine & nutrition.

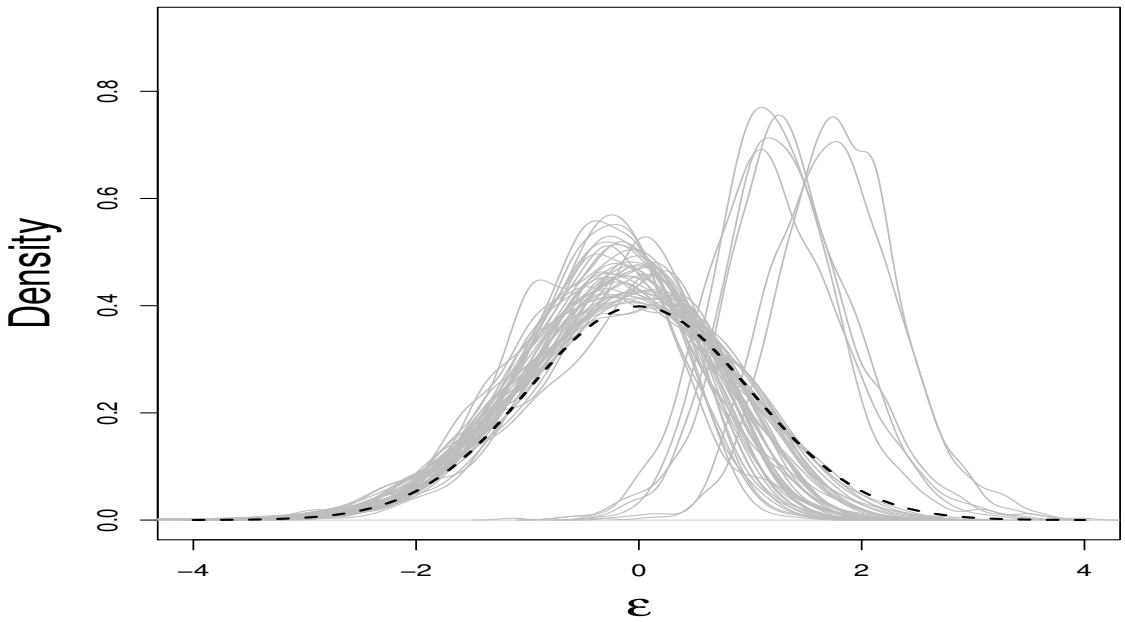


Assessing model fit

- 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.



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.

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.

Bibliography

- McParland, D., Gormley, I.C. et al. (2016)
  "Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic & genotypic data."
  Under Review.

- McParland, D. and Gormley, I.C. (2016)
  "Model based clustering for mixed data: `clustMD`"
  Advances in Data Analysis and Classification.

- McParland, D., Gormley, I. C. et al. (2014).
  "Clustering South African households based on their asset status using latent variable models."
  The Annals of Applied Statistics.

(Grant no: 09/RFP/MTH2367)