CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data
Authors	McParland, D.;Phillips, Catherine M.;Brennan, L.;Roche, H. M.;Gormley, I. C.
Publication date	2017-06-30
Original Citation	McParland, D., Phillips, C. M., Brennan, L., Roche, H. M. and Gormley, I. C. (2017) 'Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data', Statistics in Medicine, 36(28), pp. 4548-4569. doi:10.1002/sim.7371
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1002/sim.7371
Rights	© 2017, John Wiley & Sons, Ltd. This is the peer reviewed version of the following article: McParland, D., Phillips, C. M., Brennan, L., Roche, H. M. and Gormley, I. C. (2017) 'Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data', Statistics in Medicine, 36(28), pp. 4548-4569. doi:10.1002/sim.7371, which has been published in final form at https://doi.org/10.1002/sim.7371. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.
Download date	2026-11-03 03:23:46
Item downloaded from	https://hdl.handle.net/10468/6192

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data.

McParland, D., Phillips, C.M., Brennan, L., Roche, H.M. and Gormley, I.C.

Supplementary material.

1 Complete list of variables collected

Apo A-1 (g/L)	Apo B (g/L)	Body Mass Index (kg/m2)
Cholesterol (mmol/L)	Diastolic Blood Pressure (mm Hg)	Glucose (mmol/L)
plasma.C14.0 (*)	plasma.C16.0 (*)	plasma.C16.1 (*)
plasma.C18.0 (*)	plasma.C18.1 (*)	plasma.c18.2 (*)
plasma.C18.3.n.6 (*)	plasma.C18.3.n.3 (*)	plasma.C18.4 (*)
plasma.C20.1 (*)	plasma.C20.3 (*)	plasma.C20.4.n.6 (*)
plasma.C20.4.n.3 (*)	plasma.C20.5 (*)	plasma.C22.4 (*)
plasma.C22.5 (*)	plasma.C22.6 (*)	Systolic Blood Pressure (mm Hg)
Triglycerides (mmol/L)	Waist (cm)	

Table 1: Complete list of the 26 continuous phenotypic variables collected. Units are detailed in parentheses, where (*) denotes percentage of the total quantified fatty acids.

ABCA1 rs363717	*ABCA1* rs2066881	*ABCA1* rs2230808	*ABCA1* rs2297404
ABCA1 rs4149313	*ABCA1* rs1929842	*ABCA1* rs2515616	*ABCA1* rs2791952
ABCA1 rs2472510	*ABCA12* rs17430358	*ABCA12* rs10932587	*ABCA12* rs2970968
ABCC8 rs1799859	*ABCC8* rs2073583	*ABCC8* rs916829	*ABCC9* rs829060
ABCD1 rs5987140	*ABCD1* rs6643643	*ABCD2* rs7968837	*ABCD2* rs10877201
ABCD2 rs4072006	*ABCD2* rs6581228	*ABCG4* rs3802885	*ABCG5* rs6720173
ABCG5 rs4148189	*ABCG5* rs4245786	*ABCG5* rs11887534	*ABCG8* rs6709904
ABCG8 rs4148217	*ABCG8* rs6544718	*ACACA* rs725038	*ACACA* rs17138899
ACACB rs2300453	*ACACB* rs4766587	*ACACB* rs2284689	*ACACB* rs2075263
ACADL rs16844213	*ACADM* rs12032051	*ACADM* rs8763	*ACADS* rs3999408
ACADS rs695950	*ACAT1* rs2280332	*ACAT2* rs2146162	*ADIPOQ* rs266729
ADIPOQ rs17366568	*ADIPOQ* rs2241766	*ACSL1* rs4862417	*ACSL1* rs9997745
ACSM2A rs1133607	*ACSM2B* rs16970280	*ACSM2B* rs7499304	*ADD1* rs12503220
ADD1 rs6600769	*ADD1* rs4961	*ADD1* rs17777371	*ADD1* rs17834108

ADD1 rs4963
ADRA2A rs3750625
AGER rs2070600
AGT rs3889728
APOA2 rs5085
APOA4 rs5092
APOA5 rs3135506
APOBEC1 rs2302515
APOC3 rs4520
APOH rs8178847
APOL1 rs2012928
C3 rs344548
C3 rs2250656
GPR35 rs2975776
CEBPA rs12691
CPT2 rs1056438
CYP7A1 rs11786580
ENPP1 rs1044498
ESR1 rs9340954
FABP1 rs1530273
FABP3 rs951545
FABP7 rs1564900
FOXO1 rs7981045
GABRA6 rs6898571
GAD2 rs3781117
GCK rs12673242
GHRL rs696217
HK2 rs3771798
HNF4A rs2144908
IDE rs551266
IGF1 rs6219
IGF1 rs10860869
PDX1 rs1124607
PDX1 rs1124607
KCNIP2 rs2305189
LDLR rs6511720
LEPR rs1137100
LIPC rs1800588
LIPG rs4245232
LOC100507205 rs7480010
LPIN1 rs11524
LPL rs1800590
LPL rs1059611
LRP2 rs4668123
LRP5 rs587397
NCOA1 rs2083389
NFATC3 rs7205935
NOS3 rs743507
NPY rs16478
OLR1 rs1050289
PFDN1 rs9042
PPARGC1A rs3736265
PPARGC1A rs1878949
PPARGC1B rs7732671
PIK3R1 rs16897511
PLIN1 rs8179037

ADIPOR1 rs10753929
ADRB3 rs4994
AGRP rs28937570
AKT2 rs7247515
APOA2 rs6413453
APOA5 rs651821
APOA5 rs662799
APOBEC1 rs1015102
APOC3 rs5128
APOL1 rs2239785
APOL6 rs5995133
C3 rs8107911
C3 rs163913
CARTPT rs3857384
CEBPA rs16967952
CREBBP rs130005
ENPP1 rs10457576
ESR1 rs11155814
ESR1 rs2474148
FABP2 rs1799883
FABP4 rs1486006
FOXC2 rs1035550
FXN rs1800651
GAD2 rs8190582
GAD2 rs11015008
GCK rs2908289
GIPR rs11671664
HK2 rs651071
HNF4A rs3212180
IDE rs1887922
IGF1 rs978458
IGF2 rs680
IGF2 rs3168310
IRS1 rs1801278
KCNIP2 rs550
LDLR rs8110695
LEPR rs10493380
LIPC rs2070895
LIPG rs3826577
LPIN1 rs2577262
LPIN1 rs4669781
LPL rs270
LPL rs268
LRP2 rs16823023
MKKS rs1547
NCOA1 rs11125744
NFATC3 rs12598
NPC1L1 rs3187907
NPY rs16475
OLR1 rs11053646
PDK4 rs12668651
PPARGC1A rs11734408
PPARGC1A rs2970865
PPARGC1B rs17572019
PIK3R1 rs895304
PLTP rs378114

ADRA2A rs521674
AGER rs184003
AGT AGT
AKT2 rs7254617
APOA4 rs5110
APOA5 rs1729411
APOB rs1042031
APOBEC1 rs7973596
APOE rs429358
APOL1 rs136175
ASIP rs819136
C3 rs1047286
CAPN10 rs2953171
CARTPT rs11575893
CPT1A rs4930610
CREBBP rs8046065
ENPP1 rs1044558
ESR1 rs1709183
EXT2 rs3740878
FABP2 rs10034661
FABP4 rs1054135
FOXO1 rs17446614
FXN rs2498429
GAD2 rs2236418
GAD2 rs928197
GCK rs1799884
GIPR rs1800437
HK2 rs3755452
HSD11B1 rs932335
IDE rs2249960
IGF1 rs2195239
IGF2 rs2230949
INSR rs1366234
IRS2 rs1865434
LCAT rs4986970
LDLR rs2738465
LEPR rs3790419
LIPC rs8034802
LMNA rs577492
LPIN1 rs2577261
LPIN2 rs607549
LPL rs328
LRP2 rs4667591
LRP2 rs16856843
MKKS rs221666
NEUROD1 rs8192556
NOS3 rs11771443
NPC1L1 rs217434
NR1H3 rs12221497
OLR1 rs3912640
MTTP rs982424
PPARGC1A rs2970869
PPARGC1B rs2010994
PPARGC1B rs26125
PLA2G2D rs10916711
PON1 rs854551

ADRA2A rs1800544
AGER rs1035798
AGT rs4762
APOA1 rs5069
APOA4 rs2239013
APOA5 rs2266788
APOB rs676210
APOC3 rs2070667
APOH rs3176975
APOL1 rs136176
ASIP rs819162
C3 rs2230199
CAPN10 rs2953166
CARTPT rs10515114
CETP rs4783962
CYP7A1 rs11786580
ENPP1 rs7754859
ESR1 rs1801132
FABP1 rs2197076
FABP2 rs10957056
FABP6 rs2277954
FOXO1 rs7986407
GABRA6 rs7704209
GAD2 rs7071922
GCK rs2268572
GHRL rs4684677
GIPR rs10423928
HMGCR rs3761738
MARK2P9 rs2209972
IDE rs11187025
IGF1 rs1019731
INSIG1 rs9768687
INSR rs7254060
KCNIP2 rs10883689
LCAT rs1109166
LEPR rs3790433
LEPR rs8179183
LIPC rs6078
LMNA rs4641
LPIN1 rs1050800
LPIN2 rs650543
LPL rs10099160
LRP2 rs2075252
LRP5 rs312015
MTTP rs3792683
NEUROD1 rs16867467
NOS3 rs3918227
NPC1L1 rs217420
NR1H3 rs7120118
PCK1 rs6070157
PPARGC1A rs3774923
PPARGC1A rs4383605
PPARGC1B rs741581
PIK3R1 rs706713
PLIN1 rs894160
PON2 rs7493

PON2 rs10487133
PPARA rs4253728
PPARA rs6008259
PPARG rs10865710
PRKAA1 rs3805492
PRKAG3 rs6436094
FAM65C rs914458
RETN rs3219177
RXRG rs17469611
SERPINE1 rs6090
SLC25A20 rs4974088
SLC2A2 rs5398
SLC2A2 rs5400
SLC39A6 rs1944319
SOS1 rs7577088
SREBF2 rs4822063
STAT3 rs2306580
TCF7L2 rs1885510
TNF rs1799724
UCP1 rs7688743
UCP3 rs15763
USF1 rs2073655
VLDLR rs1869592
ABCA1 rs2065412
ABCA1 rs1800977
ABCA12 rs6744831
ABCC8 rs916827
ABCC9 rs1283816
ABCC9 rs1517276
ABCG5 rs2278356
ABCG5 rs4131229
ABCG8 rs6752551
ACACB rs6606697
ACACB rs3742023
ACADL rs2286963
ACAT1 rs11212524
ADIPOQ rs1063538
ACE rs4344
ACSL1 rs6552828
ADIPOR1 rs2275737
ADRB2 rs1042713
AGT rs5051
APOB rs679899
APOE rs405509
APOL1 rs3886200
C3 rs2241393
CD36 rs1984112
CD36 rs1049673
CETP rs4783961
CETP rs5882
CYP7A1 rs3808607
ESR1 rs2234693
ESR1 rs1884051
FABP2 rs6857641
FABP7 rs7752838
GABRA6 rs6556558

PON2 rs4729189
PPARA rs1800206
PPARD rs2267665
PPARG rs1801282
PRKAA1 rs10074991
PTPN1 rs6126033
PYY rs162430
RETN rs3745367
SCARB1 rs701106
SERPINE1 rs2227657
SLC27A4 rs7030121
SLC2A2 rs5406
SLC2A4 rs5412
SLC6A14 rs12720074
SREBF1 rs11868035
SREBF2 rs5996080
STAT3 rs8069645
TCF7L2 rs290481
TNF rs1800750
UCP1 rs1800592
UCP3 rs1800849
USF1 rs2073653
VLDLR rs2290465
ABCA1 rs2230806
ABCA12 rs2225063
ABCA12 rs1523718
ABCC8 rs1799854
ABCC9 rs864360
ABCC9 rs2138723
ABCG5 rs10205816
ABCG5 rs13396273
ABCG8 rs4952689
ACACB rs2268387
THYN1 rs570113
ACADL rs3764913
ACAT1 rs10890819
ACE rs4291
ACE rs4359
ACSL1 rs13120078
ADIPOR1 rs10920533
AGT rs2067853
APOA1 rs5070
APOB rs1367117
APOE rs440446
APOL6 rs9610329
C3 rs7257062
CD36 rs1761667
CPT1A rs2305508
CETP rs820299
CPT2 rs1799821
CYP7A1 rs3808607
ESR1 rs1514348
FABP1 rs2241883
FABP2 rs2282688
FOXC2 FOXC2
GABRA6 rs3219151

PDZK1 rs1284300
PPARA rs6007662
PPARD rs2076169
PPARG rs3856806
PRKAA1 rs466108
PTPN1 rs6020608
PYY rs231461
RXRB rs6531
SCARB1 rs10846744
SERPINE1 rs2227672
SLC27A4 rs6478827
SLC2A2 rs10513684
SLC2A4 rs5415
SLC27A6 rs174006
SREBF2 rs4501042
SREBF2 rs4822062
HNF1A rs3999413
TGFBI rs4572
TNF rs1800629
UCP2 rs17132534
TSTD1 rs2073658
USF1 rs1556259
VLDLR rs8210
ABCA1 rs4743764
ABCA12 rs4533467
ABCA12 rs12464205
ABCC8 rs2040653
ABCC9 rs829080
ABCD2 rs11172721
ABCG5 rs4148187
ABCG5 rs3806471
ABCG8 rs4953028
ACACB rs3742026
THYN1 rs1048761
ACADM rs1251079
ACAT2 rs927450
ACE rs4295
ACE rs4461142
ACSL1 rs12503643
ADIPOR2 rs6489323
AGT rs2478523
APOA2 rs5082
APOB rs512535
APOH rs6933
C3 rs11569562
GPR35 rs2953161
CD36 rs3211816
CPT1A rs4930248
CETP rs7205804
CREBBP rs129968
ENPP1 rs858344
ESR1 rs6557171
FABP1 rs2970901
FABP6 rs12523547
FXN rs2309393
GCK rs2268574

PPARA rs12330015
PPARA rs4253778
PPARG rs6809631
PRKAA1 rs17239241
PRKAA2 rs11206887
PTPN1 rs2230604
RETN rs1862513
RXRB rs2076310
SCD rs3870747
SERPINE1 rs2227692
SLC27A5 rs4801275
SLC2A2 rs5404
SLC39A5 rs17118403
SOS1 rs2168043
SREBF2 rs11702960
STAT3 rs1053005
TCF7L2 rs17685538
TNF rs1799964
UCP1 rs11932232
UCP3 rs7930460
USF1 rs2774276
VLDLR rs1454626
ABCA1 rs2482432
ABCA1 rs2740487
ABCA12 rs4673937
ABCC8 rs757110
ABCC8 rs2237967
ABCC9 rs7301876
ABCG4 rs668033
ABCG5 rs4148182
ABCG5 rs3806470
ACACA rs1266182
ACACB rs2284685
ACAD8 rs473041
ACADM rs1250876
ADIPOQ rs822395
ACE rs4343
ACE rs8066276
ADD1 rs3775068
ADIPOR2 rs1058322
AGT rs699
APOB rs693
APOBEC1 rs9651863
APOL1 rs136147
C3 rs344550
CARTPT rs6453132
CD36 rs3211931
CETP rs17231506
CETP rs4784744
CREBBP rs886528
ESR1 rs532010
ESR1 rs6927072
FABP2 rs1511025
FABP7 rs2243372
FXN rs7861997
GCK rs2908290

GHRL rs35683	*P3H3* rs3759348	*P3H3* rs4963517	*GNB3* rs5440
GNB3 rs5443	*GYS1* rs2270938	*GYS1* rs5464	*HHEX* rs1111875
HK2 rs656489	*HK2* rs10496195	*HK2* rs4241264	*HMGCR* rs10038095
HNF4A rs3212183	*HNF4A* rs6017340	*HNF4A* rs6031596	*HNF4A* rs3818247
LIPE rs1206034	*IDE* rs2149632	*IDE* rs1544210	*IGF1* rs6214
IGF1 rs7136446	*IGF1R* rs4966014	*IGF1R* rs6598541	*IGF1R* rs2715428
IGF1R rs2229765	*IGF1R* rs7166565	*IGF1R* rs2593053	*IGF2* rs734351
IGF2 rs3213216	*IGF2* rs3741211	*IKBKB* rs3747811	*IKBKB* rs5029748
IKBKB rs10958713	*IKBKB* rs3747811	*IKBKB* rs5029748	*IKBKB* rs10958713
IL6 rs1800797	*IL6* rs1800795	*IL6* rs2069832	*INSIG1* rs9769506
INSIG1 rs9770068	*PDX1* rs4581569	*MTTP* rs10516445	*INSR* rs7248104
INSR rs17254521	*INSR* rs1346490	*INSR* rs919275	*PDX1* rs4581569
IRS2 rs2289046	*IRS2* rs4771646	*KCNIP2* rs874885	*KCNJ11* rs2285676
KCNJ11 rs5210	*KCNJ11* rs5215	*KCNJ11* rs5219	*LDLR* rs5930
LDLR rs688	*LEP* rs13228377	*LEP* rs11763517	*LEP* rs11761556
LEPR rs2025805	*LEPR* rs6673324	*LEPR* rs1137101	*LEPR* rs12067936
LEPR rs1805096	*LIPC* rs11629736	*LIPC* rs8028759	*LIPC* rs6083
LIPG rs2000813	*LIPG* rs2097055	*LIPG* rs2276269	*LIPG* rs6507931
LIPH rs6788865	*LIPH* rs9790230	*LPIN1* rs4315495	*LPIN1* rs7595221
LPIN1 rs10209969	*LPIN2* rs641287	*LPIN2* rs1785282	*LPL* rs1534649
LRP2 rs2161039	*LRP2* rs831022	*LRP2* rs700550	*LRP2* rs2544377
LRP2 rs6716834	*LRP2* rs4668136	*LRP2* rs3755166	*LRP5* rs4988300
LRP5 rs7944040	*LRP5* rs638051	*LTA* rs915654	*LTA* rs2239704
LTA rs909253	*LTA* rs1041981	*LTA* rs915654	*LTA* rs2239704
LTA rs909253	*LTA* rs1041981	*MKKS* rs6077783	*SLX4IP* rs6039932
MTTP rs3816873	*MTTP* rs1057613	*NCOA1* rs920418	*NCOA1* rs2033861
NEUROD1 rs1801262	*NFATC3* rs7190307	*NOS3.1* NOS3.1	*NOS3* rs1800783
NPY rs16129	*NR1H2* rs17373080	*NR1H2* rs1405655	*OLR1* rs10505755
OLR1 rs1050283	*TMEM52B* rs2742113	*PDK4* rs3807891	*MTTP* rs3811800
PPARGC1A rs8192678	*PPARGC1A* rs2970848	*PPARGC1A* rs4235308	*PPARGC1A* rs2946385
PPARGC1A rs2970870	*PPARGC1B* rs10060424	*PPARGC1B* rs10747516	*PPARGC1B* rs9324628
PIK3R1 rs34303	*PIK3R1* rs10940160	*PLA2G2D* rs584367	*PLIN1* rs1052700
PLTP rs394643	*PON1* rs854549	*PON1* rs662	*PON1* rs854560
PON1 rs2299261	*PON1* rs2299262	*PPARA* rs135539	*PPARG* rs2972164
PPARG rs709154	*PRKAA2* rs2746342	*PTPN1* rs6012953	*PTPN1* rs3787334
PTPN1 rs2143511	*PTPN1* rs2038526	*PTPN1* rs2426159	*PYY* rs1058046
RETN rs3745369	*RXRA* rs1805352	*RXRA* rs4240705	*RXRA* rs1805343
RXRG rs283690	*RXRG* rs2134095	*RXRG* rs157865	*RXRG* rs2281985
SCAP rs12636851	*SCAP* rs6800271	*SCARB1* rs838891	*SCARB1* rs5888
SCARB1 rs989892	*SCARB1* rs4765616	*SCARB1* rs10846748	*SCARB1* rs10773109
SCARB1 rs3924313	*SCARB1* rs10773111	*SCARB1* rs11615630	*SCD* rs3071
SCD rs2234970	*SCD* rs3978768	*SERPINE1* rs2227631	*SLC25A14* rs2235800
SLC25A20 rs4974085	*SLC27A1* rs4808650	*SLC27A1* rs10423221	*SLC27A1* rs2278280
SLC27A2 rs7176501	*SLC27A2* rs2278167	*SLC27A2* rs1365508	*SLC27A3* rs10158450
SLC2A4 rs5435	*SLC39A6* rs8092264	*SLC39A6* rs2236718	*SLC39A6* rs948419
SLC6A14 rs7059155	*SLC6A14* rs2071877	*SLC6A14* rs2011162	*SLC27A6* rs2577531
SLC27A6 rs185411	*SLC27A6* rs1181965	*SOS1* rs2888586	*SREBF1* rs9902941
TOM1L2 rs6502618	*SREBF2* rs2267443	*STAT3* rs744166	*HNF1A* rs1169288
HNF1A rs1169302	*HNF1A* rs2464196	*HNF1A* rs1169307	*HNF1A* rs735396
TCF7L2 rs12255372	*TCF7L2* rs3814573	*TCF7L2* rs7903146	*TGFB1* rs4803455
TGFB1 rs2241715	*UCP1* rs12502572	*UCP2* rs660339	*UCP2* rs659366
USF1 rs2516840	*VLDLR* rs2242103	*VLDLR* rs1454627	*VLDLR* rs2242104

Table 2: Complete list of the 712 categorical genotypic SNP variables analysed (and the associated genes). The first 371 SNP variables (by row) are binary variables and the remaining 341 SNP variables are nominal with 3 levels.

2 Mixture of Factor Analysers for Mixed Data, with Variable Selection – a Simulation Study

2.1 Motivation

In order to assess the performance of the mixture of factor analysers for mixed data (MFA-MD) model with incorporated variable selection algorithm, and to assess the likelihood approximation and associated BIC-MCMC model selection criterion, a simulation study was conducted. Mixed type data were simulated from an MFA-MD model as detailed below. A range of MFA-MD models with varying values of G and Q was then fitted to the simulated data and the results analysed.

2.2 Data Simulation

In order to simulate mixed type data, continuous data were first simulated from a mixture of $G = 2$ Gaussian distributions:

$$f(\underline{z}_i) \;=\; \sum_{g=1}^{G} \pi_g \mathrm{MVN}_D\left(\underline{\mu}_g,\ \Lambda_g\Lambda_g^T + \Psi\right)$$

The matrices Λ_g were filled in with values simulated from a standard Gaussian distribution. The diagonal entries of Ψ corresponding to continuous variables were specified by taking the absolute value of Gaussian deviates with mean 1 and variance 1. The continuous data, which acts as the 'latent' continuous data of the MFA-MD model, were then transformed into 'observed' mixed type data using the criteria described in Section 3. A data set consisting of $N = 500$ observations on $J = 738$ variables was simulated. The variable types were divided as follows; 26 were continuous, 371 were binary and 341 were nominal. Thus the simulated data are similar in nature to the LIPGENE data set.

Two simulation settings were considered, both involving a mixture of two Gaussian distributions. The mixture components were well separated in the first simulation setting, but not well separated in the second. The degree of separation was achieved by specifying values for $\underline{\mu}_1$ and setting $\underline{\mu}_2$ equal to the negative of $\underline{\mu}_1$. Values for each entry of $\underline{\mu}_1$ were simulated from a Gaussian distribution with mean 3 and variance 1 for the well separated case but with mean 1.5 and variance 1 for the less well separated case. The mixing weights π_g were set equal to 0.5 so that clusters were approximately equal in size. The dimension of the latent trait Q was specified to be 2.

In order to test the variable selection algorithm, 600 of the $J = 738$ variables were selected at random and the columns of the data corresponding to these variables were replaced with standard Gaussian deviates so that they did not contain clustering information. The variable selection algorithm should identify these noise variables and remove them from the model. Further the BIC-MCMC criterion, involving the approximated likelihood function, should select the generating $G = Q = 2$ model.

2.3 Results

A range of MFA-MD models, with $G = 1, \ldots, 4$ and $Q = 1, \ldots, 3$, were fitted to the simulated data sets. Each MCMC chain burned in for 5000 iterations, with a further

50,000 iterations, thinned every 50th. The thresholds for the variable selection algorithm ε were set at 0.9 for continuous variables and 0.95 for categorical variables. The BIC-MCMC criterion was evaluated for each model and the model with the largest BIC-MCMC selected as optimal. Further, the set of retained variables in the optimal model and the set of known clustering variables were compared.

2.3.1 Simulation setting 1: Well Separated Mixture Components

Table 3 details the resulting BIC-MCMC values from fitting the MFA-MD models to the well separated simulated data. The true generating $G = 2$ and $Q = 2$ MFA-MD model has the largest BIC-MCMC value. The cluster memberships returned by the optimal MFA-MD model correspond exactly with the true component labels (Table 4). The variable selection algorithm correctly retained the 138 clustering variables, along with an additional 135 of the noise variables. Examination of the cluster means of the retained variables suggests that a less conservative threshold ε would have removed the retained non-clustering variables.

2.3.2 Simulation setting 2: Less Well Separated Mixture Components

In the case of the less well separated components, the similar behaviour was observed. Table 5 details the BIC-MCMC values, and the true generating $G = Q = 2$ model is selected as optimal. The resulting MFA-MD model clustered all the observations correctly, and retained 271 variables, of which 134 were non-clustering variables. One true clustering variable was also removed during the variable selection process.

The simulation study suggests that the approaches taken to variable selection and likelihood approximation, along with the BIC-MCMC model selection tool, perform well with the context of the MFA-MD model. Conservative values of the variable selection threshold ε are recommended to ensure all clustering variables are retained.

Table 3: The BIC-MCMC values for each of the MFA-MD models fitted to the well separated simulated data. The optimal model is shown in bold.

		Q		
		1	2	3
G	1	-661361	-662344	-663317
	2	-592824	**-592371**	-605088
	3	-599299	-599256	-600558
	4	-605993	-608576	-608571

Table 4: A cross tabulation of cluster membership versus the true cluster labels for the simulated data set with well separated clusters.

	True Component 1	True Component 2
Cluster 1	242	0
Cluster 2	0	258

Table 5: The BIC-MCMC values for each of the MFA-MD models fitted to the less well separated simulated data. The optimal model is shown in bold.

		Q		
		1	2	3
G	1	-663809	-664818	-665927
	2	-595978	**-595751**	-596256
	3	-598852	-600665	-602721
	4	-604278	-605234	-606663

3 Derivation of the Full Conditional Posterior Distributions

3.1 Mixing Weights

A Dirichlet prior distribution is used for the mixing weights $\underline{\pi}$.

$$\underline{\pi} \sim \text{Dirichlet}(\underline{\alpha})$$

This is a conjugate prior which leads to a Dirichlet full conditional posterior.

$$h(\underline{\pi}|\ldots) \propto \prod_{g=1}^{G} \pi_g^{(n_g+\alpha_g)-1}$$

$$\Rightarrow \underline{\pi}|\ldots \sim \text{Dirichlet}(\underline{\delta}_\pi)$$

where $\underline{\delta}_\pi = (n_1 + \alpha_1, \ldots, n_G + \alpha_G)$ and $n_g = \sum_{i=1}^{N} \ell_{ig}$.

3.2 Allocation Vectors

A priori the allocation vectors, $\underline{\ell}_i$ are assumed to be Multinomial$(1, \underline{\pi})$ distributed. Thus the full conditional posterior is also multinomial.

$$\begin{aligned} h(\underline{\ell}_i|\ldots) \propto \prod_{g=1}^{G} \Bigg\{ & \pi_g \left[\prod_{j=1}^{A} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, \psi_{jj})\right] \left[\prod_{j=A+1}^{B} \prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}\{z_{ij}<0|y_{ij}\}}\right] \\ & \times \left[\prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T} \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}(y_{ij}=s)}\right] \Bigg\}^{\ell_{ig}} \end{aligned}$$

$$\Rightarrow \underline{\ell}_i|\ldots \sim \text{Multinomial}(\underline{p})$$

where

$$\begin{aligned} p_g = & \left\{ \pi_g \left[\prod_{j=1}^{A} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, \psi_{jj})\right] \left[\prod_{j=A+1}^{B} \prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}\{z_{ij}<0|y_{ij}\}}\right] \right. \\ & \left. \times \left[\prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T} \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}(y_{ij}=s)}\right] \right\} \\ \times & \left\{ \sum_{g=1}^{G} \pi_g \left[\prod_{j=1}^{A} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, \psi_{jj})\right] \left[\prod_{j=A+1}^{B} \prod_{k=1}^{K_j} N(z_{ij}|\tilde{\underline{\lambda}}_{gj}^T \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}\{z_{ij}<0|y_{ij}\}}\right] \right. \\ & \left. \times \left[\prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1}|\tilde{\underline{\lambda}}_{gj}^{k-1^T} \tilde{\underline{\theta}}_i, 1)^{\mathbb{I}(y_{ij}=s)}\right] \right\}^{-1} \end{aligned}$$

3.3 Marginal Variance Parameters

An inverse gamma prior distributions is specified for the diagonal elements of Ψ which correspond to continuous items j.

$$\psi_{jj} \sim \mathcal{G}^{-1}(\beta_1, \beta_2)$$

This is a conjugate prior and leads to an inverse gamma posterior distribution.

$$\begin{aligned}
h(\psi_d|\ldots) &\propto \prod_{i=1}^{N}\left[\mathrm{N}\left(z_{id}|\underline{\lambda}_{gd}^T\tilde{\underline{\theta}}_i, \psi_d\right)\right]^{\ell_{ig}} \mathcal{G}^{-1}(\beta_1, \beta_2) \\
&\propto \psi_d^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{g=1}^{G}\sum_{i:\ell_{ig}=1}\psi_d^{-1}\left(z_{id} - \underline{\lambda}_{gd}^T\tilde{\underline{\theta}}_i\right)^2\right\}\psi_d^{-\beta_1-1}\exp\left(-\frac{\beta_2}{\psi_d}\right) \\
&\propto \psi_d^{-\beta_1-\frac{N}{2}-1}\exp\left\{-\frac{1}{2\psi_d}\sum_{g=1}^{G}\sum_{i:\ell_{ig}=1}\left(z_{id} - \underline{\lambda}_{gd}^T\tilde{\underline{\theta}}_i\right)^2 - \frac{\beta_2}{\psi_d}\right\} \\
&\propto \psi_d^{-\beta_1-\frac{N}{2}-1}\exp\left\{\psi_d^{-1}\left[\frac{1}{2}\sum_{g=1}^{G}\left(\underline{z}_{gd} - \tilde{\theta}_g\underline{\lambda}_{gd}\right)^T\left(\underline{z}_{gd} - \tilde{\theta}_g\underline{\lambda}_{gd}\right) + \beta_2\right]\right\}
\end{aligned}$$

$$\Rightarrow \psi_d|\ldots \sim \mathcal{G}^{-1}\{b_{1j}, b_{2j}\}$$

where

$$b_{1j} = \beta_1 + \frac{N}{2}$$

and

$$b_{2j} = \left[\frac{1}{2}\sum_{g=1}^{G}\left(\underline{z}_{gd} - \tilde{\theta}_g\underline{\lambda}_{gd}\right)^T\left(\underline{z}_{gd} - \tilde{\theta}_g\underline{\lambda}_{gd}\right) + \beta_2\right]$$

3.4 Latent Traits

A priori the latent traits, $\underline{\theta}_i$, are assumed to be standard multivariate Gaussian distributed.

$$\underline{\theta}_i \sim \mathrm{MVN}_q(\underline{\theta}_i|\mathbf{0}, \mathbf{I})$$

The resulting full conditional posterior is also multivariate Gaussian.

$$\begin{aligned}
h(\underline{\theta}_i|\ldots) &\propto \prod_{d=1}^{D}\mathrm{N}\left(z_{id}|\tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i, 1\right)\mathrm{MVN}_q(\underline{0}, \mathbf{I}_q) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left(z_{id} - \tilde{\underline{\lambda}}_{gd}^T\tilde{\underline{\theta}}_i\right)^2\right\}\exp\left\{-\frac{1}{2}\underline{\theta}_i^T\underline{\theta}_i\right\} \\
&\propto \exp\left\{-\frac{1}{2}\underline{\theta}_i^T\left[\Lambda_g^T\Lambda_g + \mathbf{I}_q\right]\underline{\theta}_i + \left[\left(\underline{z}_i - \underline{\mu}_g\right)^T\Lambda_g\right]\underline{\theta}_i\right\}
\end{aligned}$$

$$\Rightarrow \underline{\theta}_i|\ldots \sim \mathrm{MVN}_q\left\{\underline{\mu}_\theta, \Sigma_\theta\right\}$$

where

$$\underline{\mu_\theta} = \left[\Lambda_g^T \Lambda_g + \mathbf{I}_q\right]^{-1} \left[\Lambda_g^T \left(\underline{z}_i - \underline{\mu}_g\right)\right]$$

and

$$\Sigma_\theta = \left[\Lambda_g^T \Lambda_g + \mathbf{I}_q\right]^{-1}$$

3.5 Loadings Matrix and Mean

A multivariate Gaussian prior is specified for $\underline{\tilde{\lambda}}_{gd}$.

$$\underline{\tilde{\lambda}}_{gd} \sim MVN(q+1)\left(\underline{\tilde{\lambda}}_{gd} | \underline{\mu}_\lambda, \Sigma_\lambda\right)$$

This is a conjugate prior and so leads to a multivariate Gaussian full conditional posterior.

$$\begin{aligned}
h(\underline{\tilde{\lambda}}_{gd} | \ldots) &\propto \prod_{i=1}^{N} \left[\mathrm{N}\left(z_{id} | \underline{\tilde{\lambda}}_{gd}^T \underline{\tilde{\theta}}_i, 1\right)\right]^{\ell_{ig}} \left[\mathrm{MVN}_{(q+1)}\left(\underline{\tilde{\lambda}}_{gd} | \underline{\mu}_\lambda, \Sigma_\lambda\right)\right] \\
&\propto \exp\left\{-\frac{1}{2} \sum_{i:\ell_{ig}=1} \left(z_{id} - \underline{\tilde{\lambda}}_{gd}^T \underline{\tilde{\theta}}_i\right)^2\right\} \exp\left\{-\frac{1}{2}\left(\underline{\tilde{\lambda}}_{gd} - \underline{\mu}_\lambda\right)^T \Sigma_\lambda^{-1} \left(\underline{\tilde{\lambda}}_{gd} - \underline{\mu}_\lambda\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2} \underline{\tilde{\lambda}}_{gd}^T \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T \tilde{\Theta}_g\right] \underline{\tilde{\lambda}}_{gd} + \underline{\tilde{\lambda}}_{gj}^T \left[\tilde{\Theta}_g^T \underline{z}_{gd} + \Sigma_\lambda^{-1} \underline{\mu}_\lambda\right]\right\}
\end{aligned}$$

$$\Rightarrow \qquad \underline{\tilde{\lambda}}_{gd} | \ldots \sim \mathrm{MVN}_{(q+1)}\left\{\underline{\zeta}_\lambda, \Omega_\lambda\right\}$$

where $\underline{z}_{gd} = \{z_{id}\}$ for all respondents i in cluster g, $\tilde{\Theta}_g$ is a matrix, the rows of which are $\underline{\tilde{\theta}}_i$ for members of cluster g

$$\underline{\zeta}_\lambda = \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T \tilde{\Theta}_g\right]^{-1} \left[\tilde{\Theta}_g^T \underline{z}_{gd} + \Sigma_\lambda^{-1} \underline{\mu}_\lambda\right]$$

and

$$\Omega_\lambda = \left[\Sigma_\lambda^{-1} + \tilde{\Theta}_g^T \tilde{\Theta}_g\right]^{-1}.$$

4 Evolution of the *variable selection phase*

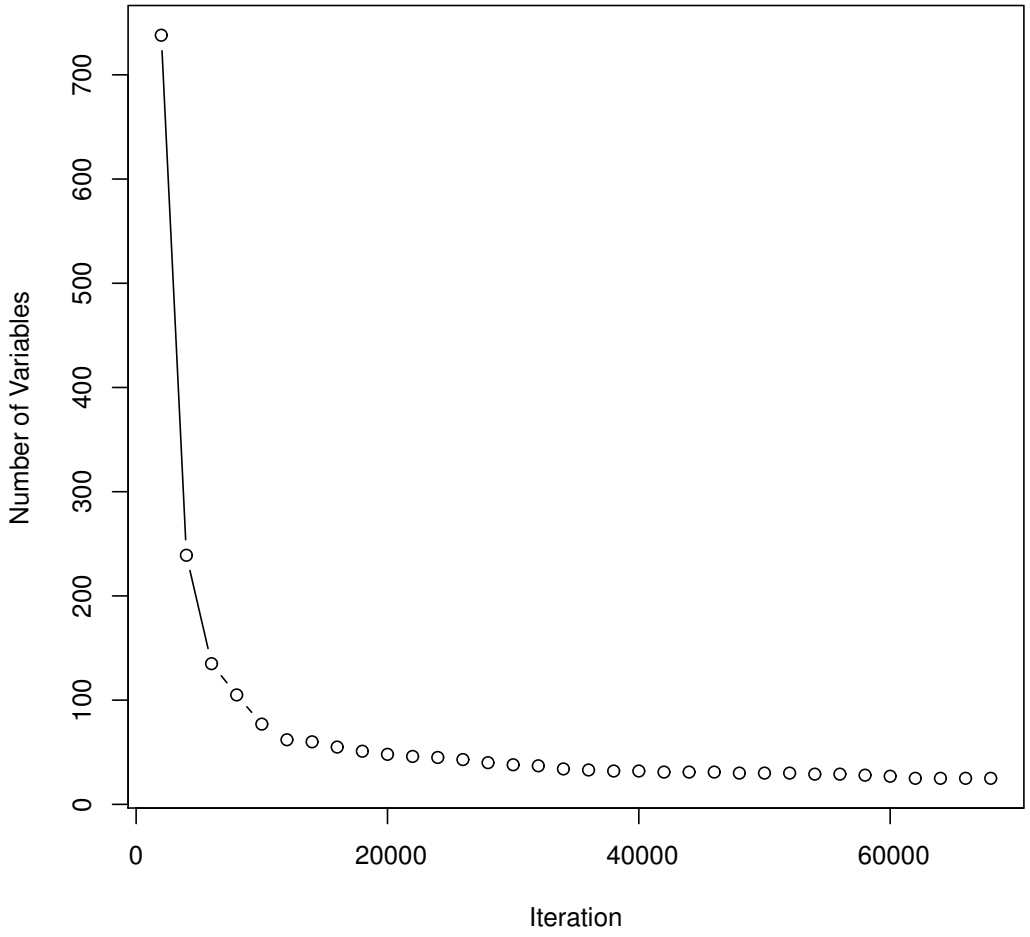


Figure 1: The evolution of the number of variables retained by the model over the duration of the variable selection phase of the model fitting algorithm.

5 Bayesian residual and Bayesian latent residual plots

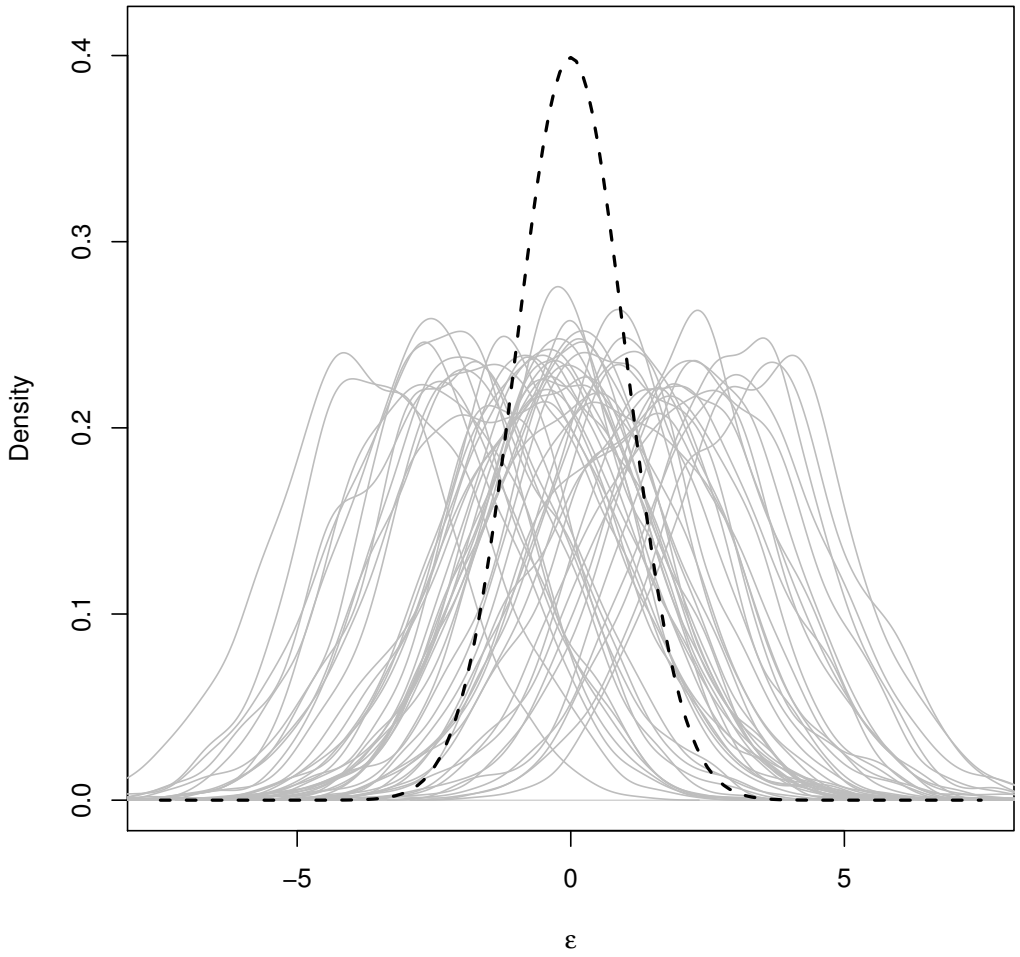
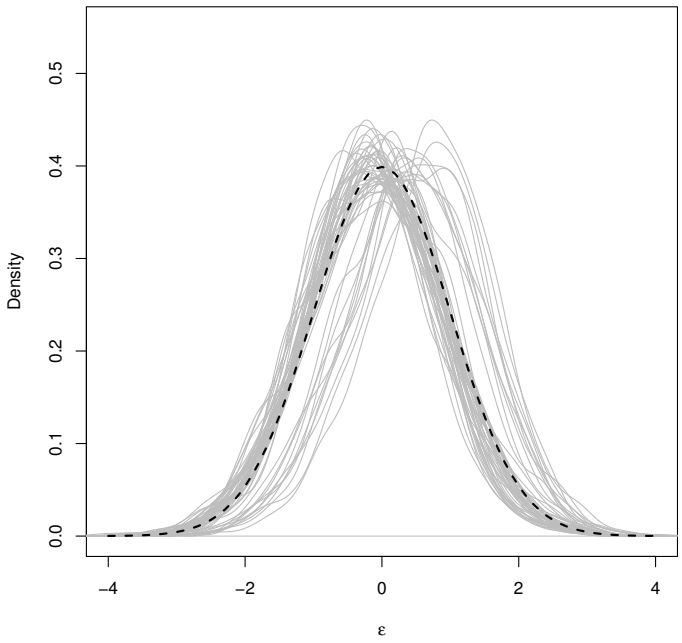
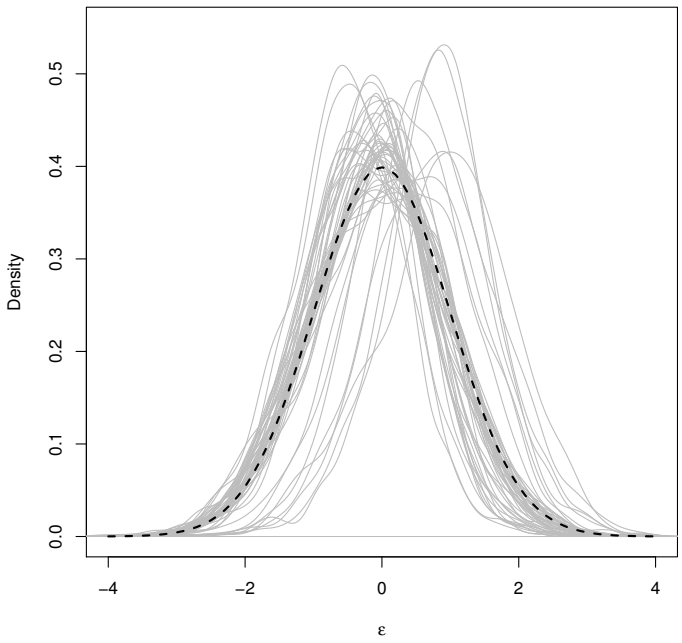


Figure 2: Density estimates of the Bayesian residuals for the waist circumference variable for 50 randomly selected volunteers. The standard Gaussian density curve is shown by the black dashed line.



(a) First latent dimension.



(b) Second latent dimension.

Figure 3: Density estimates of the Bayesian latent residuals for the `rsrs2071877` SNP on the $SLC6A14$ gene, for 50 randomly selected volunteers. The standard Gaussian density curve is shown by the black dashed line.