Parameter estimation in biclassified blockmodels as mixture of contingency tables via the EM algorithm

Marianna Bolla, Fatma Abdelkhalek, József Mala
Institute of Mathematics
Budapest University of Technology and Economics
marib@math.bme.hu

CFE-CMStatistics, Pisa, 15 December, 2018.

Motivation

- We introduce a directed random contingency table model, where the entries are in $(0, 1)$, independent and beta-distributed with parameters depending on their row and column indices. We will find sufficient statistics, and based on them, give an algorithm to find ML estimates of the parameters, together with convergence proof.
- Then we extend the model to the multiclass scenario, where for fixed k and l, we are looking for k so-called out-clusters of the rows and l in-clusters of the columns so that the parameters of the beta-distributed entries also depend on their cluster memberships..
- To find the clusters and estimate the parameters, we use an EM iteration that alternately finds the cluster memberships, then fixing the memberships, estimates the parameters within the blocks separately.
- The algorithm is applied to generated and genetic data.

Motivation

- We introduce a directed random contingency table model, where the entries are in $(0, 1)$, independent and beta-distributed with parameters depending on their row and column indices. We will find sufficient statistics, and based on them, give an algorithm to find ML estimates of the parameters, together with convergence proof.
- Then we extend the model to the multiclass scenario, where for fixed k and l, we are looking for k so-called out-clusters of the rows and l in-clusters of the columns so that the parameters of the beta-distributed entries also depend on their cluster memberships..
- To find the clusters and estimate the parameters, we use an EM iteration that alternately finds the cluster memberships, then fixing the memberships, estimates the parameters within the blocks separately.
- The algorithm is applied to generated and genetic data.

Motivation

- We introduce a directed random contingency table model, where the entries are in $(0, 1)$, independent and beta-distributed with parameters depending on their row and column indices. We will find sufficient statistics, and based on them, give an algorithm to find ML estimates of the parameters, together with convergence proof.
- Then we extend the model to the multiclass scenario, where for fixed k and l, we are looking for k so-called out-clusters of the rows and l in-clusters of the columns so that the parameters of the beta-distributed entries also depend on their cluster memberships..
- To find the clusters and estimate the parameters, we use an EM iteration that alternately finds the cluster memberships, then fixing the memberships, estimates the parameters within the blocks separately.
- The algorithm is applied to generated and genetic data.

Motivation

- We introduce a directed random contingency table model, where the entries are in $(0, 1)$, independent and beta-distributed with parameters depending on their row and column indices. We will find sufficient statistics, and based on them, give an algorithm to find ML estimates of the parameters, together with convergence proof.
- Then we extend the model to the multiclass scenario, where for fixed k and l, we are looking for k so-called out-clusters of the rows and l in-clusters of the columns so that the parameters of the beta-distributed entries also depend on their cluster memberships..
- To find the clusters and estimate the parameters, we use an EM iteration that alternately finds the cluster memberships, then fixing the memberships, estimates the parameters within the blocks separately.
- The algorithm is applied to generated and genetic data.

The one-cluster random contingency table model

$\mathbf{W} = (w_{ij})$: $n \times m$ contingency table, $w_{ij} \in (0, 1)$.

Model: w_{ij} obeys a beta-distribution with parameters $a_i > 0$ and $b_j > 0$, independently of the other entries.

Notation: $\mathbf{a} := (a_1, \dots a_n)$ and $\mathbf{b} = (b_1, \dots b_m)$

a_i: the potential of row i to send messages out
b_j: the resistance of column j to receive messages in
w_{ij}: the weight of the $i \to j$ message.

The likelihood function in factorized form

$$L_{\mathbf{a},\mathbf{b}}(\mathbf{W}) = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{\Gamma(a_i + b_j)}{\Gamma(a_i)\Gamma(b_j)} w_{ij}^{a_i - 1}(1 - w_{ij})^{b_j - 1}$$

$$= C(\mathbf{a}, \mathbf{b}) \prod_{i=1}^{n}\prod_{j=1}^{m} \exp[(a_i - 1)\ln w_{ij} + (b_j - 1)\ln(1 - w_{ij})]$$

$C(\mathbf{a}, \mathbf{b})$: normalizing constant

$s_i := \sum_{j=1}^{m} \ln w_{ij} \quad (i = 1, \ldots, n)$: ith row-sum of the $n \times m$ matrix $\mathbf{U} = \mathbf{U}(\mathbf{W})$ of general entry $\ln w_{ij}$

$z_j := \sum_{i=1}^{n} \ln(1 - w_{ij}) \quad (j = 1, \ldots, m)$: jth column-sum of the $n \times m$ matrix $\mathbf{V} = \mathbf{V}(\mathbf{W})$ of general entry $\ln(1 - w_{ij})$

The system of ML equations (MLE) in terms of the sufficient statistics

$$\frac{\partial \ln L_{\mathbf{a},\mathbf{b}}(\mathbf{W})}{\partial a_i} = \sum_{s=1}^{m} \frac{\Gamma'(a_i + b_s)}{\Gamma(a_i + b_s)} - m\frac{\Gamma'(a_i)}{\Gamma(a_i)} + s_i = 0, \quad i = 1, \ldots, n$$

$$\frac{\partial \ln L_{\mathbf{a},\mathbf{b}}(\mathbf{W})}{\partial b_i} = \sum_{s=1}^{n} \frac{\Gamma'(a_s + b_i)}{\Gamma(a_s + b_i)} - n\frac{\Gamma'(b_i)}{\Gamma(b_i)} + z_i = 0, \quad i = 1, \ldots, m$$

Fixed point iteration

$\boldsymbol{\theta} = f(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{b})$:

$$a_i = \psi^{-1}\left[\frac{1}{m}s_i + \frac{1}{m}\sum_{s=1}^{m}\psi(a_i + b_s)\right] =: g_i(\mathbf{a}, \mathbf{b}), \quad i = 1, \ldots, n$$

$$b_i = \psi^{-1}\left[\frac{1}{n}z_i + \frac{1}{n}\sum_{s=1}^{n}\psi(a_s + b_i)\right] =: h_i(\mathbf{a}, \mathbf{b}), \quad i = 1, \ldots, m$$

$\psi(x) = \frac{\partial \ln \Gamma(x)}{\partial x} = \frac{\Gamma'(x)}{\Gamma(x)}$ for $x > 0$ is the digamma function

g_is resp. h_js are the coordinate functions of $g : \mathbb{R}^{n+m} \to \mathbb{R}^n$ resp. $h : \mathbb{R}^{n+m} \to \mathbb{R}^m$

$f = (g, h)$

Starting with $\boldsymbol{\theta}^{(0)}$, we use the successive approximation $\boldsymbol{\theta}^{(it)} := f(\boldsymbol{\theta}^{(it-1)})$ for $it = 1, 2, \ldots$ until convergence.

Some facts needed for the convergence

As in minimally represented, regular exponential famlies, the MLE has a unique solution $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$, we have to prove the convergence to it.

Fact 1.

$$\sum_{i=1}^{n} s_i + \sum_{j=1}^{m} z_j = \sum_{i=1}^{n}\sum_{j=1}^{m} \ln w_{ij} + \sum_{j=1}^{m}\sum_{i=1}^{n} \ln(1 - w_{ij}) \leq -2 \ln 2 \, nm.$$

with equality if and only if $w_{ij} = \frac{1}{2}$ $(\forall i, j)$.

Fact 2. The function $\psi(2x) - \psi(x)$, $x \in (0, \infty)$ is decreasing and its range is $(\ln 2, \infty)$.

Facts continued

Fact 3.

$$M := \max\left\{ \max_{i\in\{1,\dots,n\}} \left(-\frac{s_i}{n}\right), \max_{i\in\{1,\dots,m\}} \left(-\frac{z_i}{m}\right) \right\}$$

and $\varepsilon > 0$ be the (only) solution of the equation $\psi(2x) - \psi(x) = M$. Then $(\hat{\mathbf{a}}, \hat{\mathbf{b}}) \geq \varepsilon\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{n+m}$ is the all 1's vector, and the inequality between vectors is understood componentwise.

Fact 4. With the solution ε of $\psi(2x) - \psi(x) = M$ we have $f(\varepsilon\mathbf{1}) \geq \varepsilon\mathbf{1}$.

Fact 5. If $(\mathbf{a}, \mathbf{b}) \geq (\mathbf{x}, \mathbf{y}) > \mathbf{0}$, then $f(\mathbf{a}, \mathbf{b}) \geq f(\mathbf{x}, \mathbf{y})$.

Convergence theorem

Theorem

Let $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{a}}, \hat{\mathbf{b}})$ be the unique solution of the MLE. Then the above mapping $f = (g, h)$ is a contraction in some closed neighborhood K of $\hat{\boldsymbol{\theta}}$, and so, starting at any $\boldsymbol{\theta}^{(0)} \in K$, the fixed point of the iteration $\boldsymbol{\theta}^{(it)} = f(\boldsymbol{\theta}^{(it-1)})$ exists and is $\hat{\boldsymbol{\theta}}$.

Since $f : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ is continuously differentiable in a neighborhood of $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{a}}, \hat{\mathbf{b}})$, existing theorems guarantee that there is a closed neighborhood K of $\hat{\boldsymbol{\theta}}$ such that f is a contraction on K. In particular, the fixed point iteration $f(\boldsymbol{\theta}^{(it-1)}) = \boldsymbol{\theta}^{(it)}$ $(it \to \infty)$ converges for every $\boldsymbol{\theta}^{(0)} \in K$ to $\hat{\boldsymbol{\theta}}$, which is the unique solution of the MLE.
Starting with $\varepsilon\mathbf{1}$, sooner or later we get into K.

The multiclass model

In the several clusters case, we are putting the blocks together. Here the statistics are sufficient only within the blocks.

Given the integers $1 \leq k \leq n$ and $1 \leq l \leq m$, we are looking for k-partition (clusters) $R_1, \ldots, R_k$ of the rows and $C_1, \ldots, C_l$ of the columns such that the row and column items are assigned to the clusters independently, and given the cluster memberships, the wight of the message sent by row $u \in R_i$ to $v \in C_j$ has weight $w_{uv} \sim \mathcal{B}eta(a_{uj}, b_{vi})$;
further, all these assignments are done independently.

Parameter estimation

Parameter matrices:
A: $n \times l$, where the jth column of **A** contains the parameters a_{uj} in the block $u \in R_i$, for $i = 1, \ldots, k$; $j = 1, \ldots, l$.
B: $m \times k$, where the ith column of **B** contains the parameters b_{vi} in the block $v \in C_j$, for $j = 1, \ldots, l$; $i = 1, \ldots, k$.

Here a_{uj} can be thought of as the potential of row-item u of cluster R_i to send messages out to C_j, and b_{vi} as the of column-item v of cluster C_j to receive messages in from R_i.

This is a mixture of exponential-family distributions, and as the mixing can be supervised by two multinomially distributed random variables (responsible for the memberships), the general theory of mixtures and the iteration of the EM algorithm can be used to estimate the parameters.

M-step: maximization

Starting with an initial clustering $R_1^{(0)}, \dots, R_k^{(0)}$ of the rows and $C_1^{(0)}, \dots, C_l^{(0)}$ of the columns, the t-th step of the iteration is as follows ($t = 1, 2, \dots$).

Maximization step within the blocks:
We update estimates of the parameters $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ within the kl blocks, separately. For the block $R_i^{(t)} \times C_j^{(t)}$, we use the one-cluster algorithm to find the estimates $a_{uj}^{(t)}$ for $u \in R_i^{(t)}$ and $b_{vi}^{(t)}$ for $v \in C_j^{(t)}$.

As each row u and column v uniquely corresponds to exactly one row- and column-cluster, respectively, in this way, the $R_i^{(t)} \times C_j^{(t)}$ parameter blocks for $i = 1, \dots, k$, $j = 1, \dots, l$ will fill in the $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ parameter matrices.

E-step: relocation between the blocks (Bayes rule)

Given the new estimates of the parameters $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$, we relocate u into the row-cluster R_{i*} and v into the column-cluster C_{j*} for which the contribution of w_{uv} to the overall likelihood is maximal. We do it separately for the rows and columns.

Relocation for the rows: For each u $(u = 1, \dots n)$ take the maximum of the following over i $(i = 1, \dots, k)$:

$$\sum_{v=1}^{m} \sum_{j=1}^{l} c_{vj} \left[\ln \frac{\Gamma(a_{uj}) + b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj} - 1) \ln w_{uv} + (b_{vi} - 1) \ln(1 - w_{uv}) \right].$$

If it is maximum for i^*, then we relocate u into the row-cluster R_{i*}.

$$\max_i \mathbb{E}(r_{ui} \mid \mathcal{M}^{(t-1)}) = \max_i \mathbb{P}(r_{ui} = 1 \mid \mathcal{M}^{(t-1)})$$

Bayes rule: $\max_i \mathbb{P}(\mathcal{M}^{(t-1)} \mid r_{ui} = 1) \times \mathbb{P}(r_{ui} = 1)$.

Uniform law: $\mathbb{P}(r_{ui} = 1) = \frac{1}{k}$ or else, $\frac{n_i^{(t-1)}}{n}$.

Relocation for the columns

For each v ($v = 1, \dots m$) take the maximum of the following over j ($j = 1, \dots, l$):

$$\sum_{u=1}^{n} \sum_{i=1}^{k} r_{ui} \left[\ln \frac{\Gamma(a_{uj}) + b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj} - 1) \ln w_{uv} + (b_{vi} - 1) \ln(1 - w_{uv}) \right].$$

If it is maximum for j^*, then we relocate v into the column-cluster C_{j^*}.

It is also equivalent to

$$\max_j \mathbb{E}(c_{vj} \mid \mathcal{M}^{(t-1)}) = \max_j \mathbb{P}(c_{vj} = 1 \mid \mathcal{M}^{(t-1)}).$$

Practical considerations

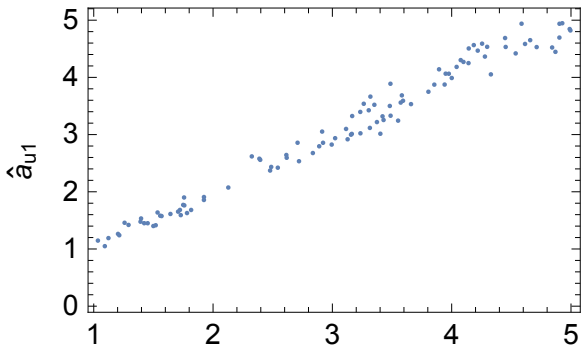
Relocation is a discrete maximization. Break ties arbitrarily.

In the E-step, we get a new clustering $R_1^{(t)}, \ldots, R_k^{(t)}$ of the rows and $C_1^{(t)}, \ldots, C_l^{(t)}$ of the columns, with which we go back to the M-step.

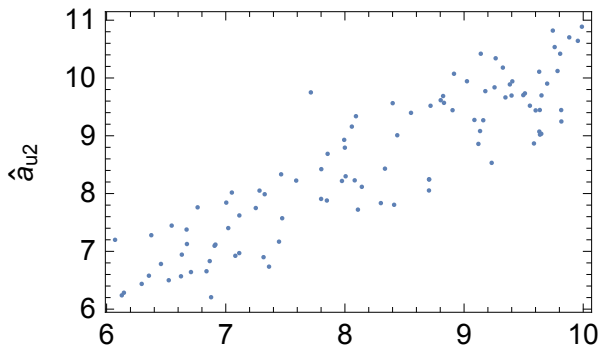
As in both steps we increase the likelihood, and the likelihood function is bounded from above with the sum of the existing maxima over the blocks, the iteration must converge to a local maximum of it.
A good starting, for example, with spectral biclustering helps a lot.
The iteration also resembles the collaborative filtering algorithm.

Application for simulated data

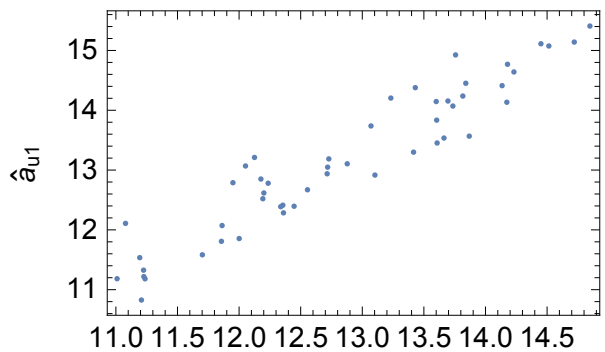
Figure: The original versus the estimated parameters a_{uj}'s



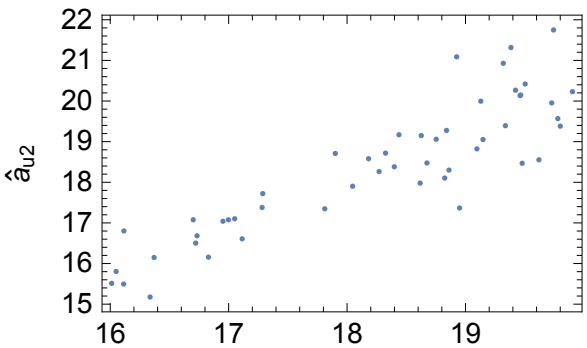
a_{u1} , u$\in R_1$, MSE = 0.029557



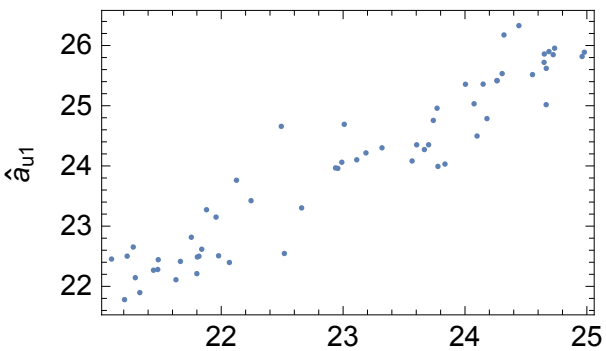
a_{u2} , u$\in R_1$, MSE = 0.415935



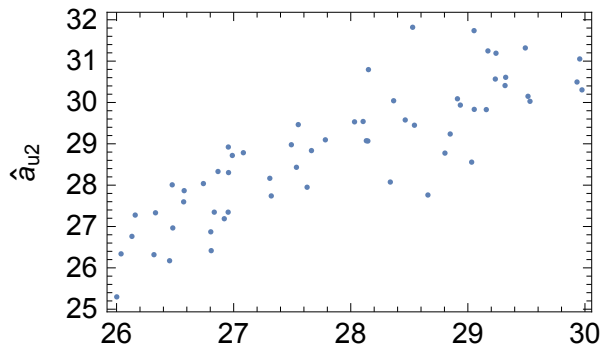
a_{u1} , u$\in R_2$, MSE = 0.2547340



a_{u2} , u$\in R_2$, MSE = 0.6126018

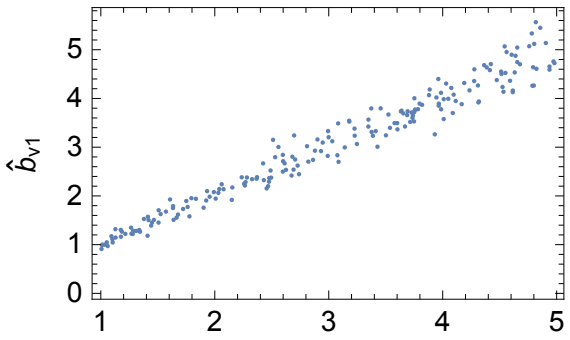


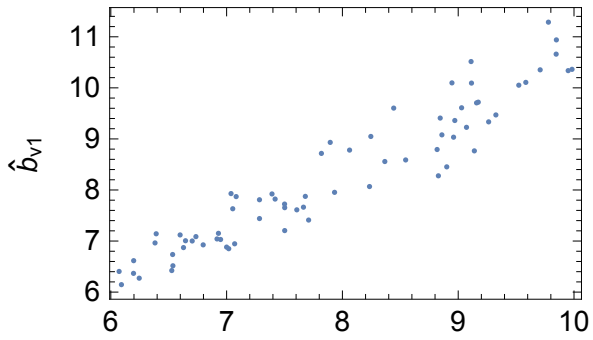
a_{u1} , u$\in R_3$, MSE = 1.0826113

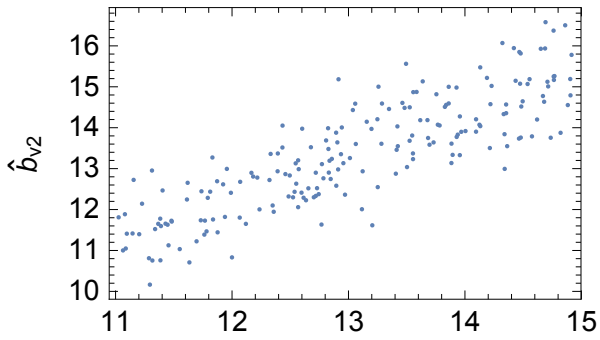


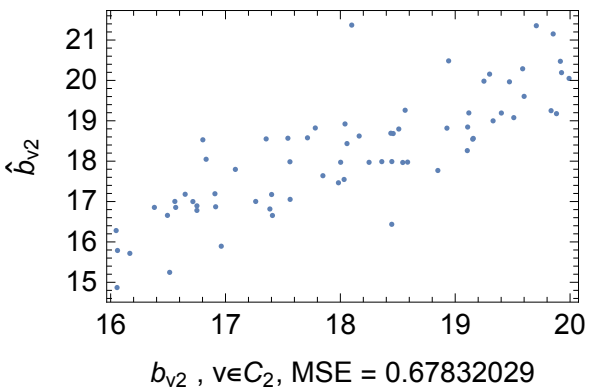
a_{u2} , u$\in R_3$, MSE= 1.57340291

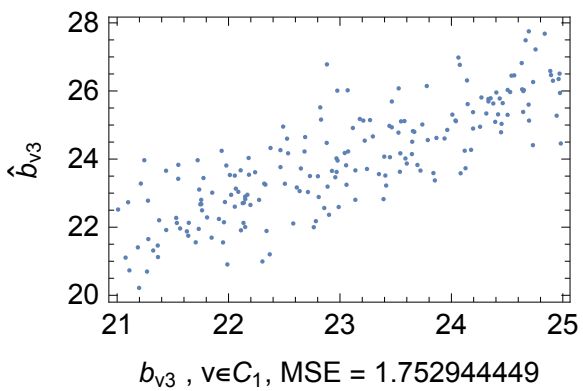
Figure: The original versus the estimated parameters b_{vi}'s

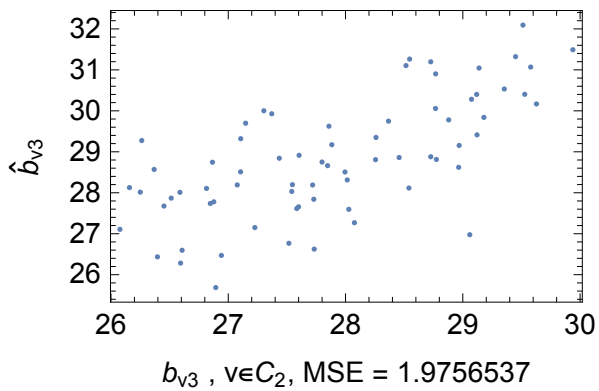












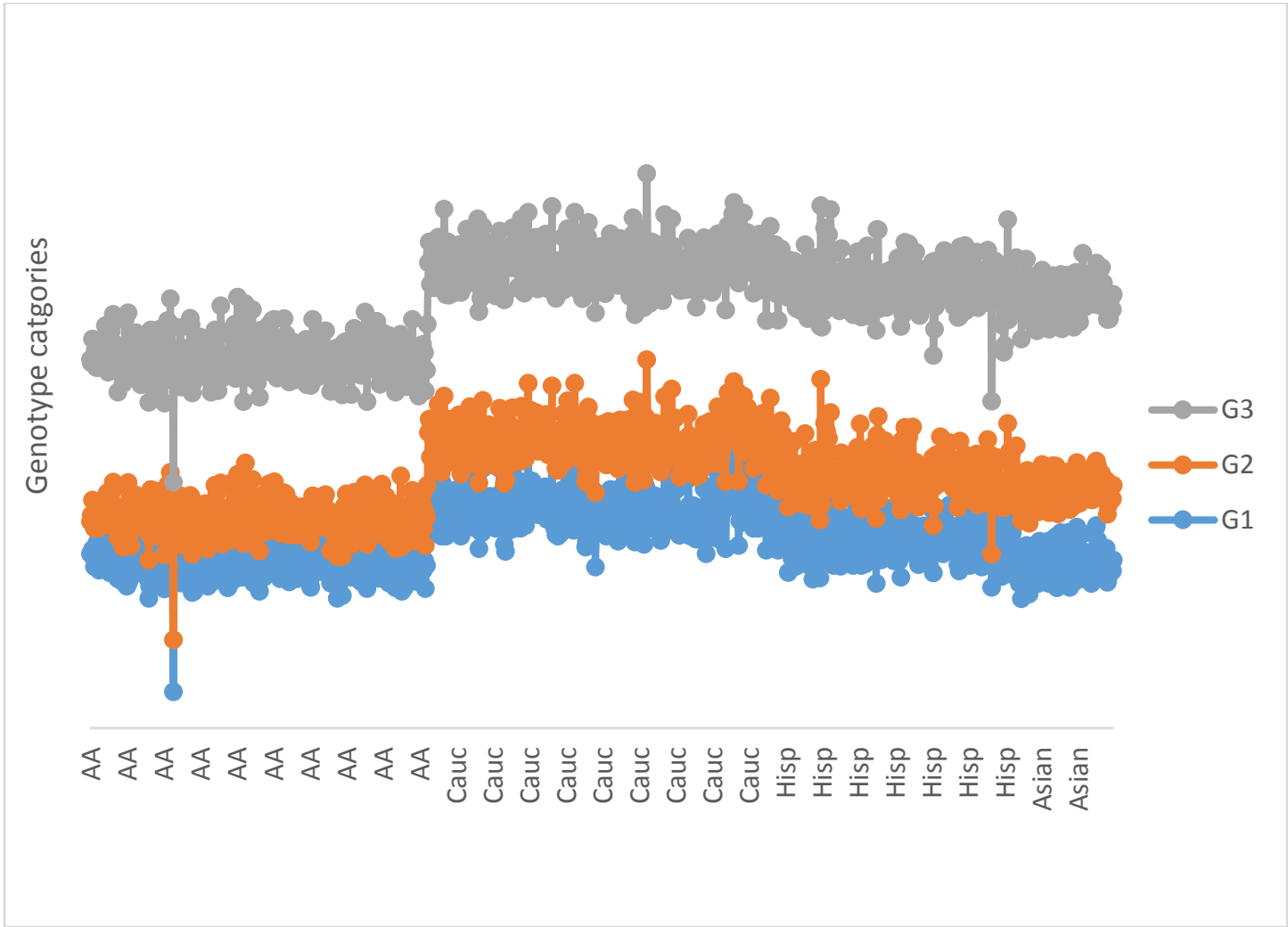
Application for genetic data

$n = 1036$ unrelated samples in the U.S. population, divided into $k = 4$ population groups:
R_1: African American (AA)
R_2: Caucasian (Cauc)
R_3: Hispanic (Hisp)
R_4: Asian (Asian)

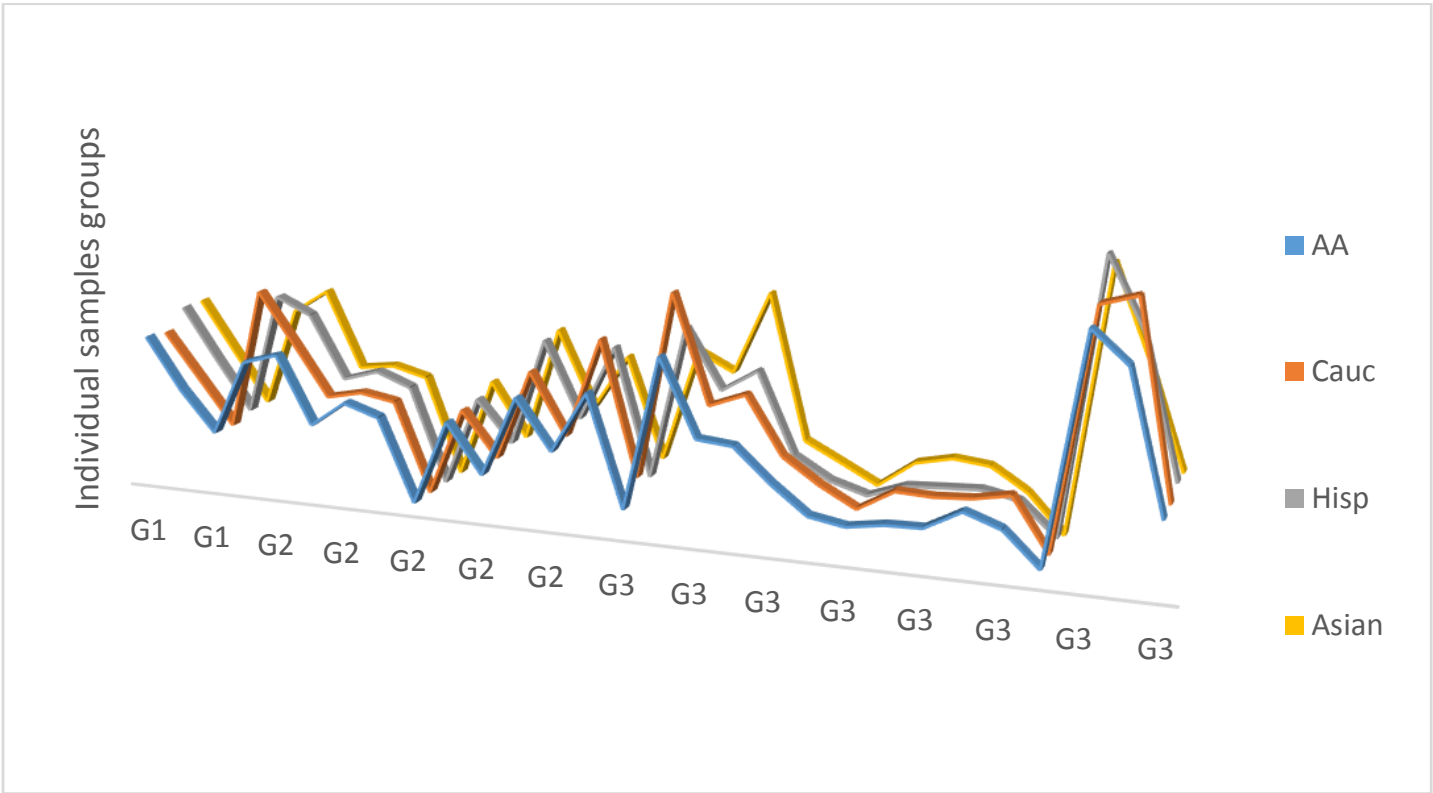
$m = 29$ loci, divided into $l = 3$ genotypes by the experts:
C_1: U.S. core genes (G1)
C_2: non-core U.S. genes (G2)
C_3: general genes (G3)

w_{uv}'s: measured allele frequencies.

The estimated parameters in the A matrix, toward to the genotypes



The estimated parameters in the B matrix, toward to the population groups



Conclusions

The starting clusters were those given by the experts, and they did not change much during the iteration until the fifth step, when the algorithm converged. The estimated parameters in **A** showed that all the four groups of U.S. individuals have low core U.S. genotype representations and higher representations of both non-core U.S. and general genotypes. At the same time, the estimated parameters in **B** indicated that all the three genotype categories have the lowest representation in the African American samples and the highest representation in the Asian group samples.
`http://dx.doi.org/10.1016/j.fsigen.2017.08.011`

- C. R. Hill, D. L. Duewer, M. C. Kline, M. D. Cober, J. M. Dutler, U. S. population data for 29 autosomal STR loci, *Forensic Sci. Int. Genet.* **7** (2013), e82-e83.
- C. R. Steffen, M. D. Coble, K. B. Gettings, P. M. Vallone, Corrigendum to 'U. S. population data for 29 autosomal STR loci', *Forensic Sci. Int. Genet.* **31** (2017), e36-e40.

One-cluster estimates for emigration–immigration data (2011)

i	Country	a_i	b_i	i	Country	a_i	b_i
1	Australia	0.26931	1475.7	18	Japan	0.23211	9926.9
2	Austria	0.27403	632.8	19	Korea	0.22310	4199.2
3	Belgium	0.33380	46.1	20	Luxembourg	0.17543	107.9
4	Canada	0.27383	2363.2	21	Mexico	0.26706	4655.9
5	Chile	0.21236	28940.5	22	Netherlands	0.37754	39.5
6	Czech Rep.	0.31188	470.2	23	New Zealand	0.20542	2568.1
7	Denmark	0.26514	847.3	24	Norway	0.22646	519.1
8	Estonia	0.23235	25602.3	25	Poland	0.62846	1106.5
9	Finland	0.29357	1100.1	26	Portugal	0.31011	1606.6
10	France	0.52721	37.9	27	Slovak Rep.	0.27871	42451.2
11	Germany	0.62020	1.6	28	Slovenia	0.19720	6824.5
12	Greece	0.29708	6319.1	29	Spain	0.39732	182.4
13	Hungary	0.31443	32750.8	30	Sweden	0.39627	57.3
14	Iceland	0.18051	2950.7	31	Switzerland	0.33611	4524.6
15	Ireland	0.27555	364.5	32	Turkey	0.25900	146175.8
16	Israel	0.25854	1926.1	33	United Kingdom	0.49301	48.6
17	Italy	0.50522	135.1	34	United States	0.38019	2433.7