Longitudinal Probabilistic Clustering of Biomarker Trajectories for Scleroderma

Joint work with Jisoo Kim
Advisor: Scott Zeger

Motivation

- Scleroderma is a chronic rare disease
- Clinicians hope to
  - discover systematic heterogeneous patterns in patients' multivariate biomarker trajectories due to unobserved factors
  - have real-time updated information on which pattern a new patient may belong to
  - account for patient heterogeneity

Solution

- Latent mixture variable model
- Bayesian approach using a Kalman Filter way to recursively update BLUP for real-time clustering of trajectories
- Random effects to account for patient heterogeneity

Model Specification

$$Y(t) = \beta_0^{(\ell)} + b_{i0}^{(\ell)} + s(t)\beta_2^{(\ell)} + X\beta_1 + Xs(t)\beta_3 + \epsilon$$

$$\ell = 1,2, X = (Male, Ethnicity, Diffuse, Late\_onset)$$

$$logit[P(cluster\ 1)] = \alpha_0 + \alpha_1 Male + \alpha_2 Ethnicity + \alpha_3 Diffuse + \alpha_4 Late\_onset$$

Question being addressed: any unobserved factors are influencing Y over time besides X?

Real-time Update: BLUP in a Bayesian Paradigm

$$g_{it}(\alpha, \beta, G, R) = P\left(cluster\ 1 \middle| Y_{i,0:t}; \beta, G, R\right)$$

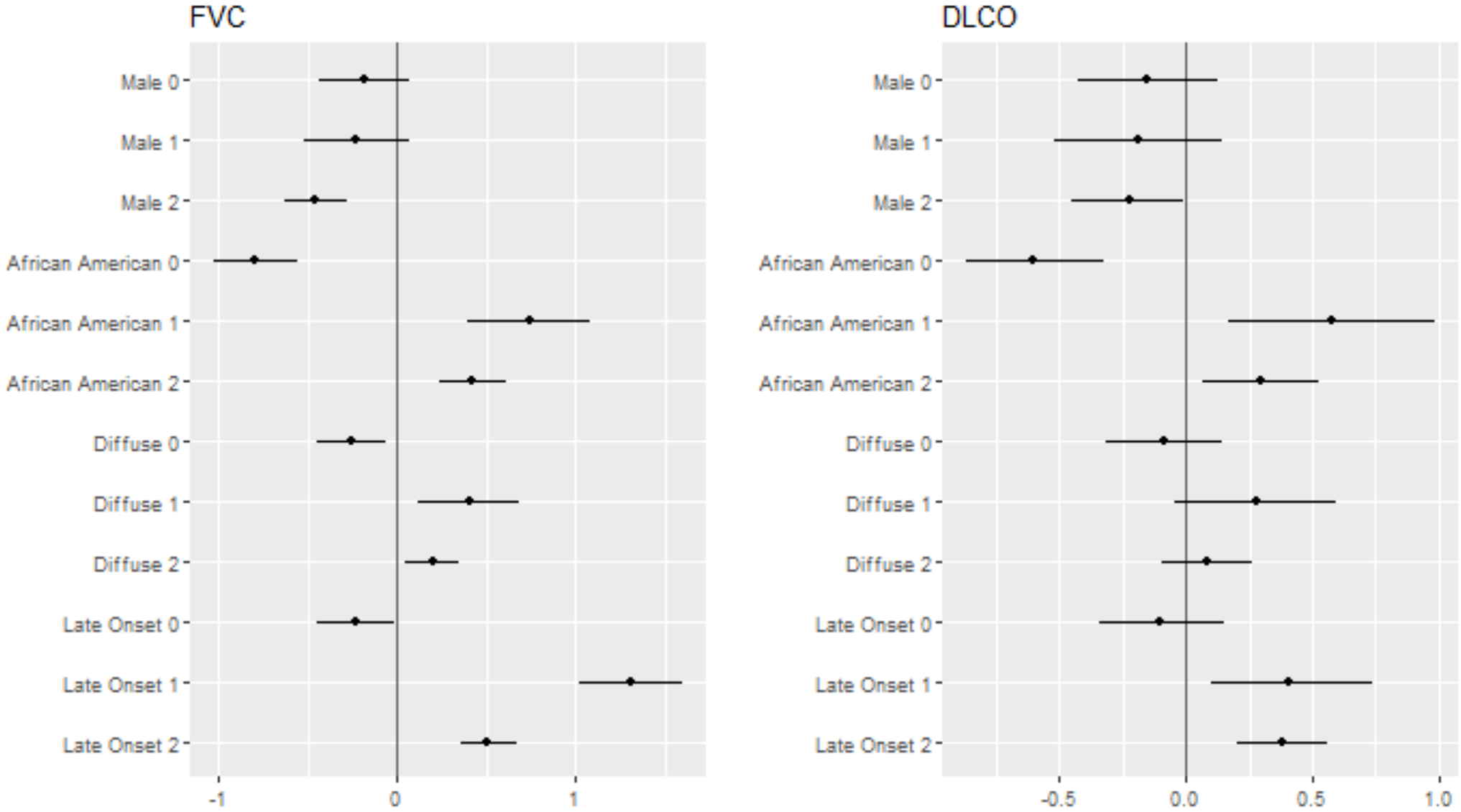
$$= \frac{P\left(Y_{i,0:t} \middle| cluster\ 1; \beta, G, R\right) \times P(cluster\ 1; \alpha)}{\sum_{c \in \{1,2\}} P\left(Y_{i,0:t} \middle| cluster\ c; \beta, G, R\right) \times P(cluster\ c; \alpha)}$$

$$P\left(Y_{i,0:t} \middle| cluster\ \ell; \beta, G, R\right)$$

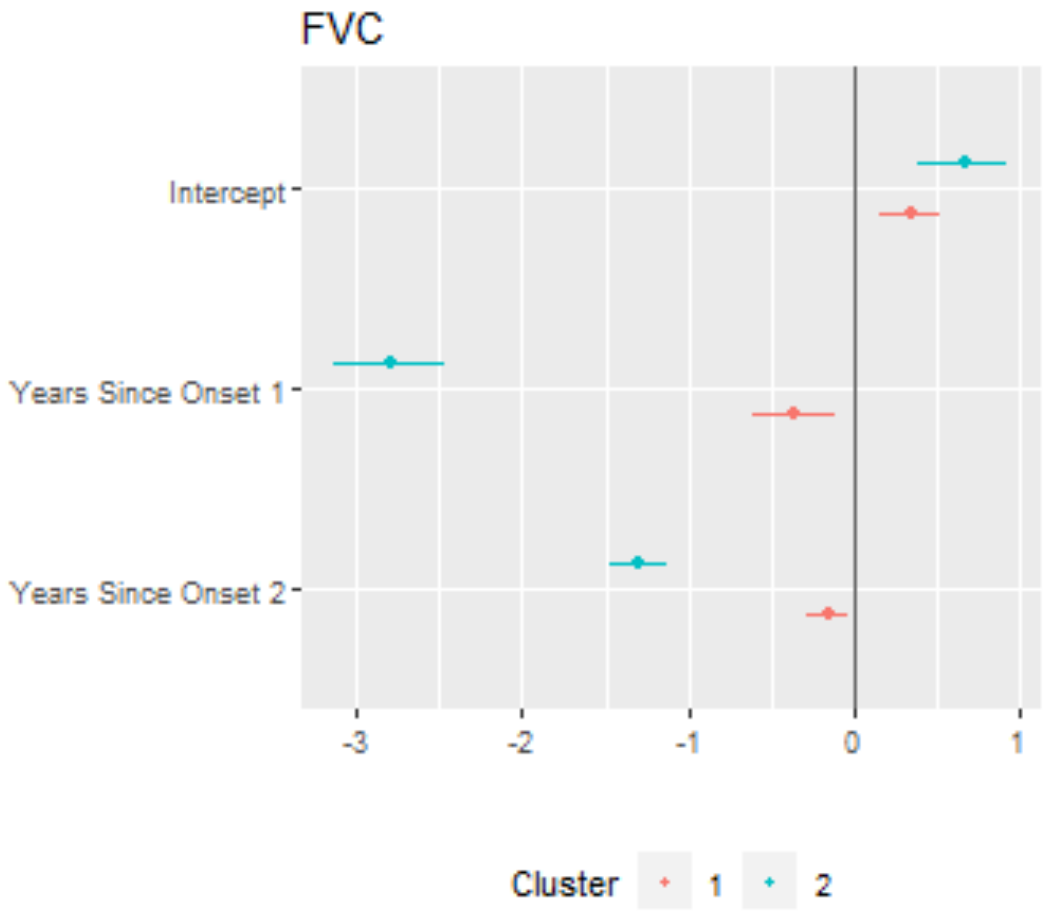
$$= \prod_{k \in 1:t} \int f\left(Y_{it} \middle| \beta_0^{(\ell)} + b_{i0}^{(\ell)} + s(t)\beta_2^{(\ell)} + X\beta_1 + Xs(t)\beta_3, R\right) f\left(b_{i0}^{(\ell)} \middle| 0, G, Y_{i,0:t}\right) db_{i0}^{(\ell)}$$

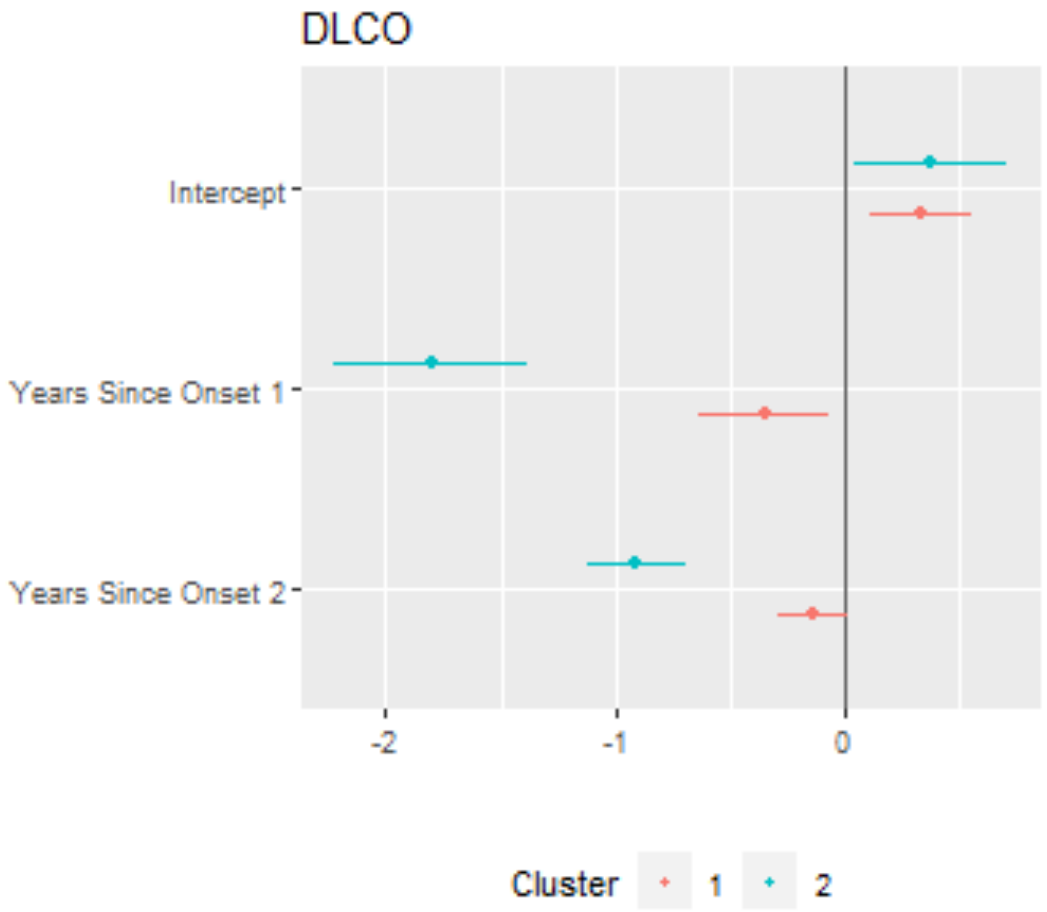
Both f's are gaussian distributed

Model Estimation – Shared Effects



Model Estimation – Cluster Specific Effects

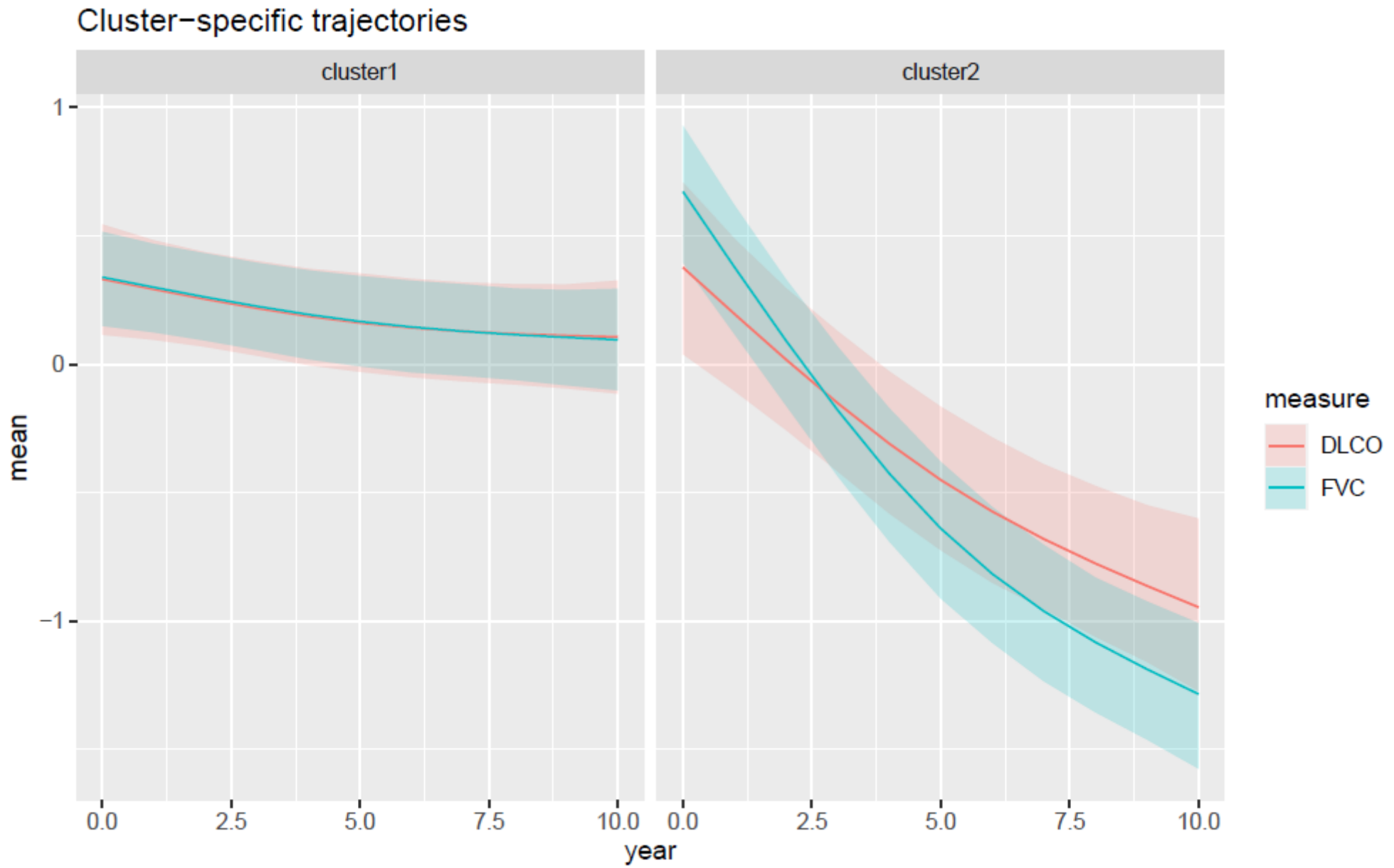


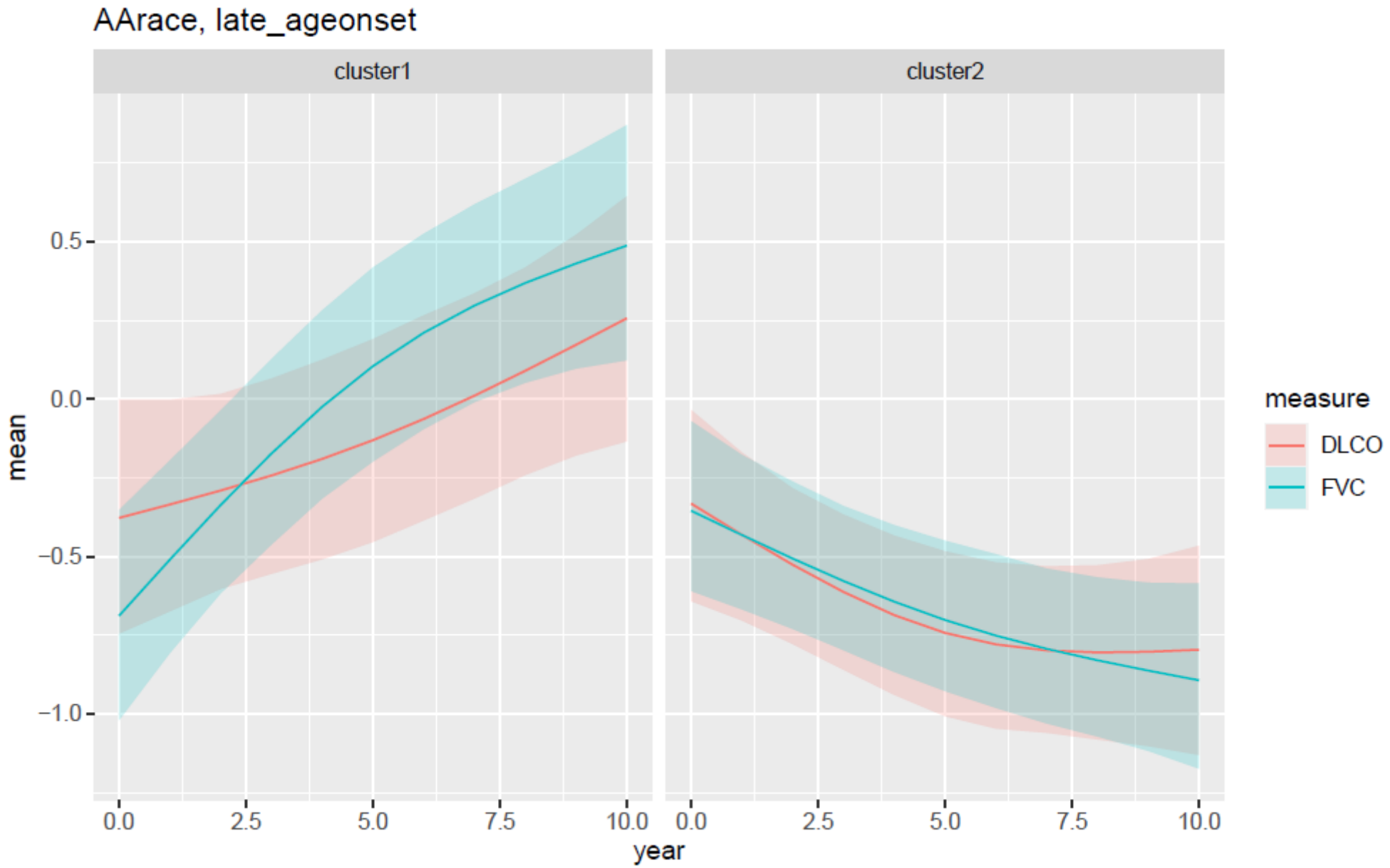


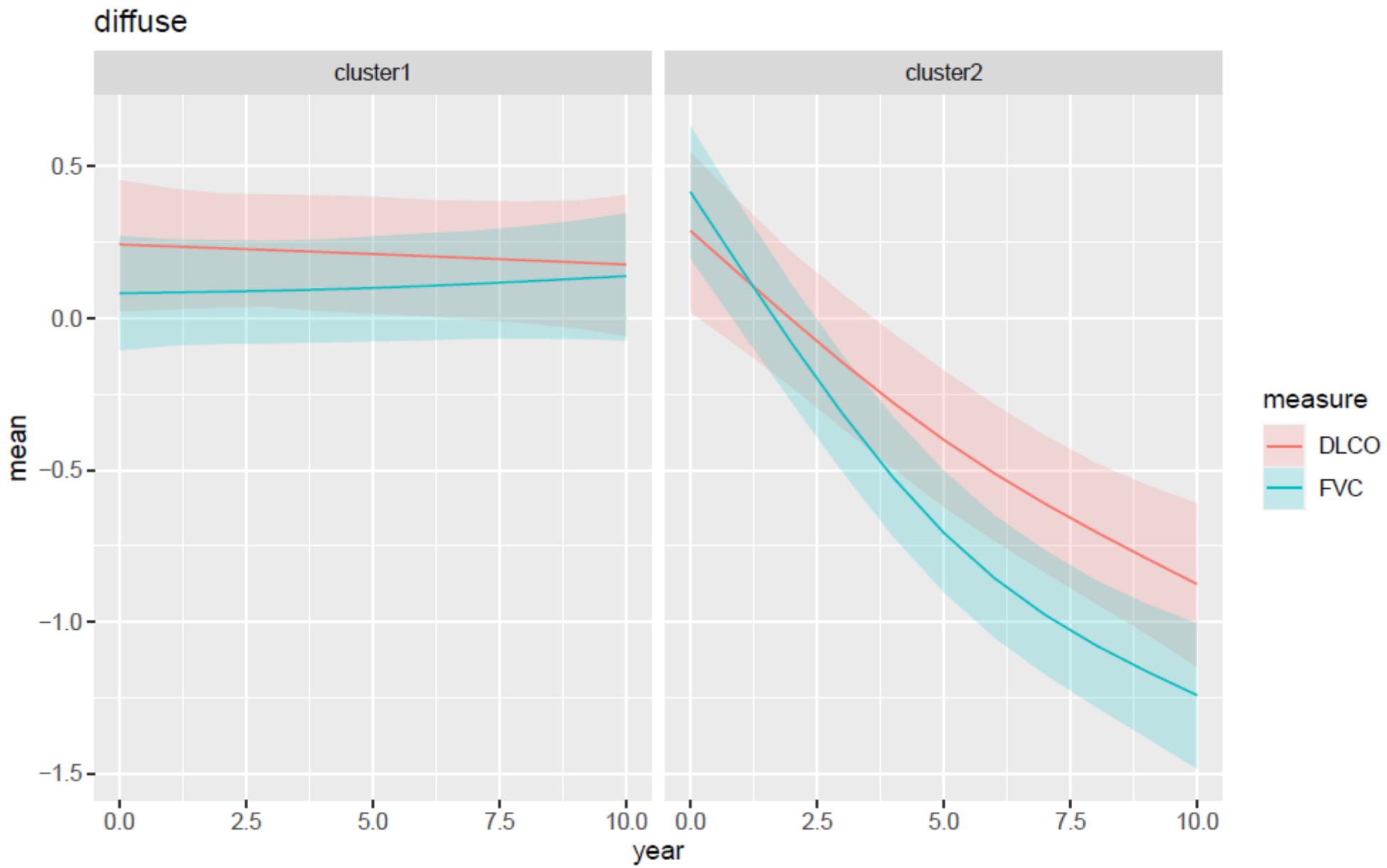
Model Estimation – Heterogeneity and Stochasticity

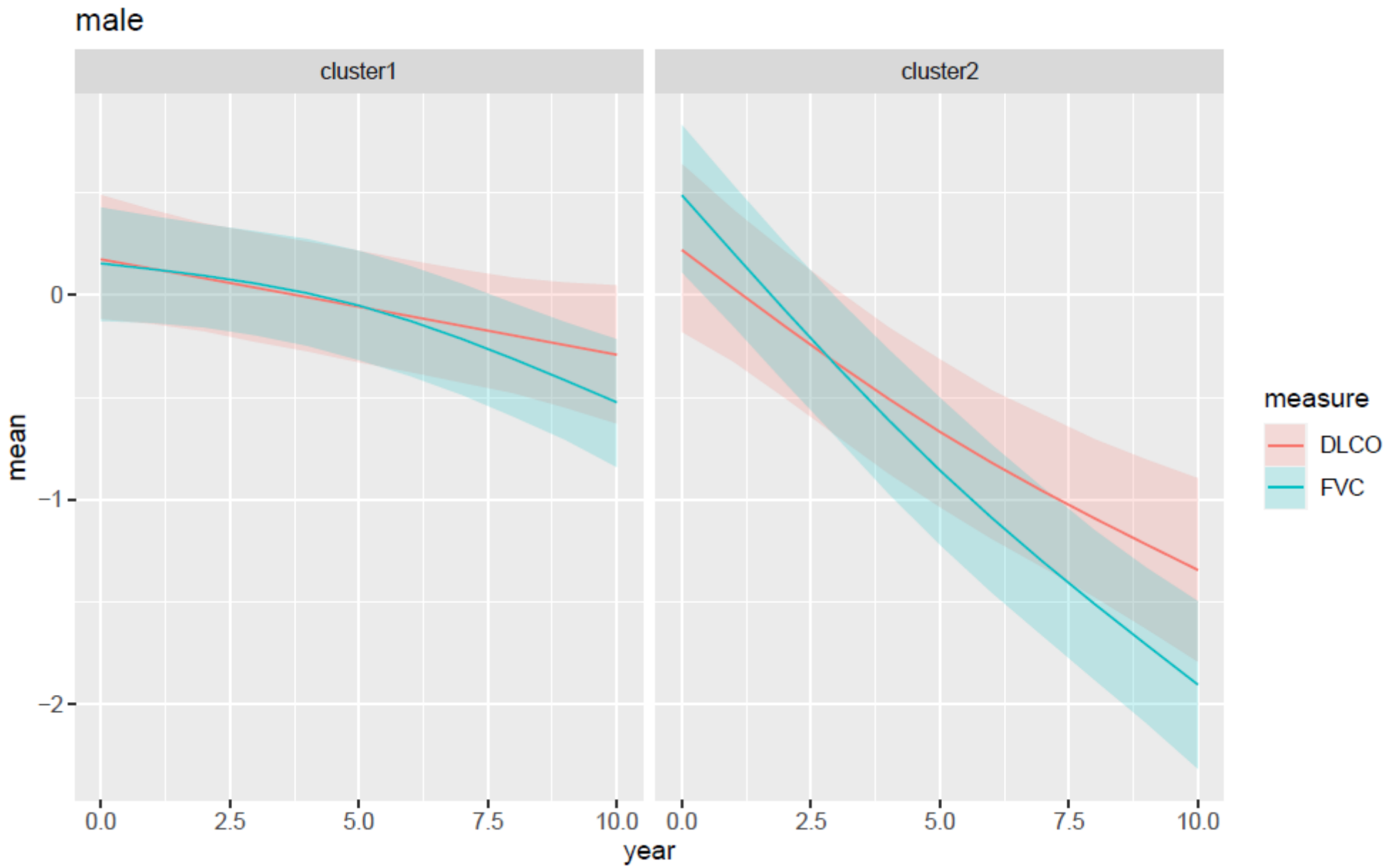
```
> summ_mat(G1, 2, varname=c("FVC b0", "DLCO b0"), frontword = "Cluster 1 G:"
[1] Cluster 1 G:
         FVC b0              DLCO b0
FVC b0  0.52 (0.42, 0.66) 0.33 (0.24, 0.45)
DLCO b0 0.33 (0.24, 0.45) 0.55 (0.43, 0.7)
> summ_mat(G2, 2, varname=c("FVC b0", "DLCO b0"), frontword = "Cluster 2 G:"
[1] Cluster 2 G:
         FVC b0              DLCO b0
FVC b0  0.43 (0.33, 0.55) 0.3 (0.21, 0.41)
DLCO b0 0.3 (0.21, 0.41)  0.52 (0.4, 0.66)
> summ_mat(R, 2, varname=c("FVC", "DLCO"), frontword = "Shared R:")
[1] Shared R:
      FVC                 DLCO
FVC   0.1 (0.09, 0.1)    0.05 (0.04, 0.05)
DLCO 0.05 (0.04, 0.05) 0.19 (0.17, 0.2)
```

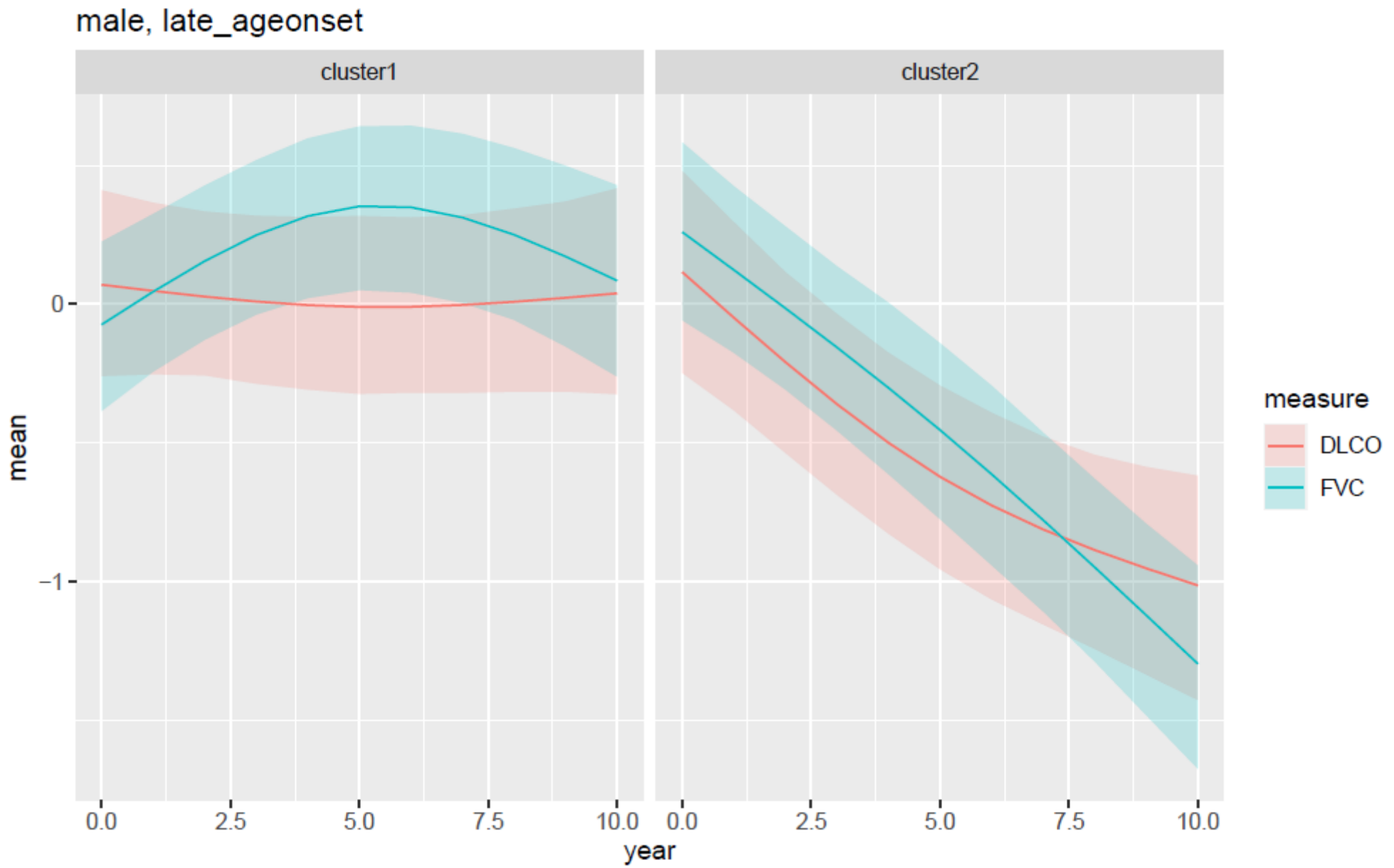
Cluster Patterns by Subgroup


Cluster−specific trajectories
cluster1
cluster2
1
0
−1
mean
0.0
2.5
5.0
7.5
10.0
year
measure
DLCO
FVC

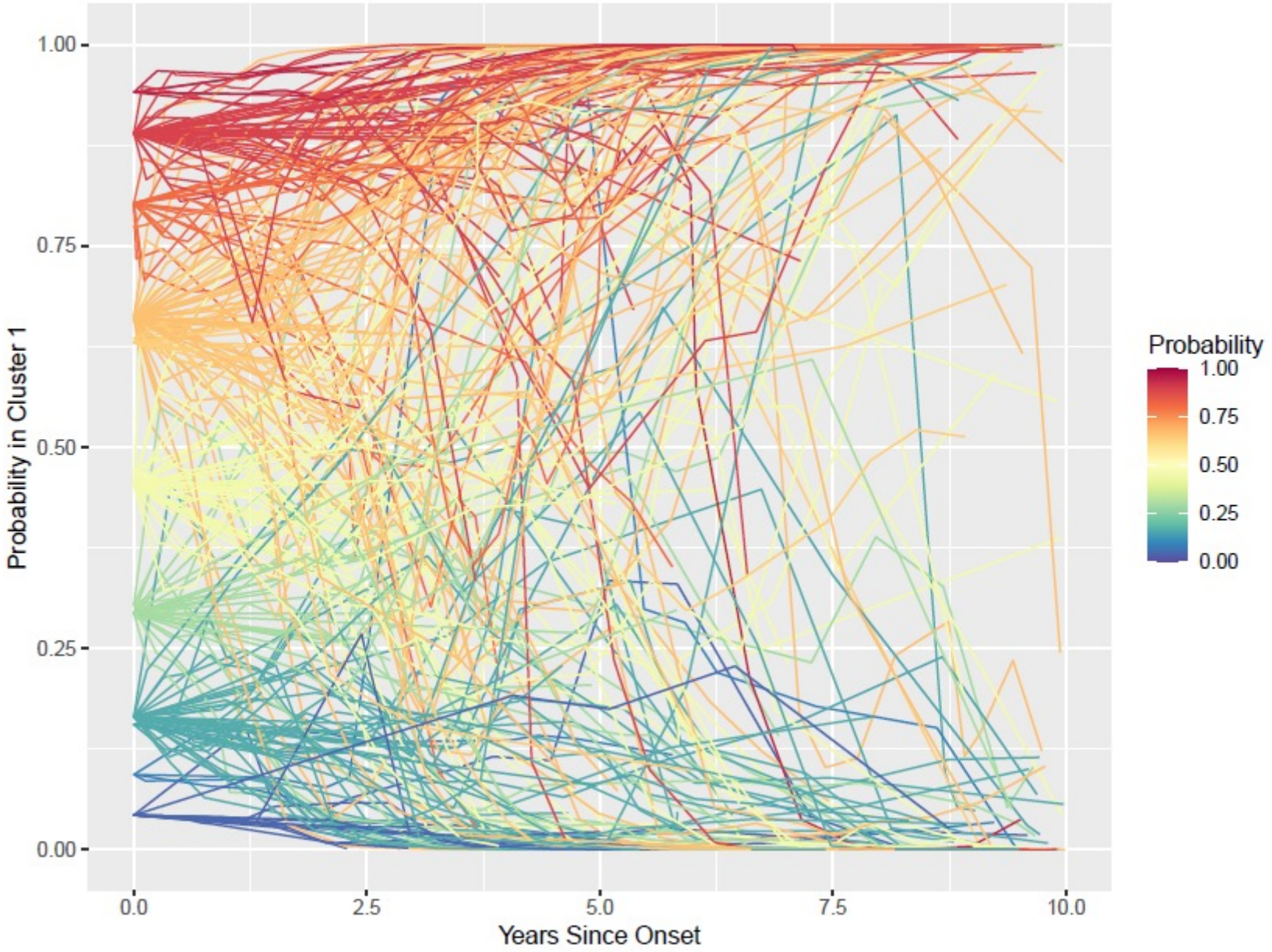

AArace, late_ageonset
cluster1
cluster2
mean
0.5
0.0
−0.5
−1.0
0.0
2.5
5.0
7.5
10.0
year
measure
DLCO
FVC

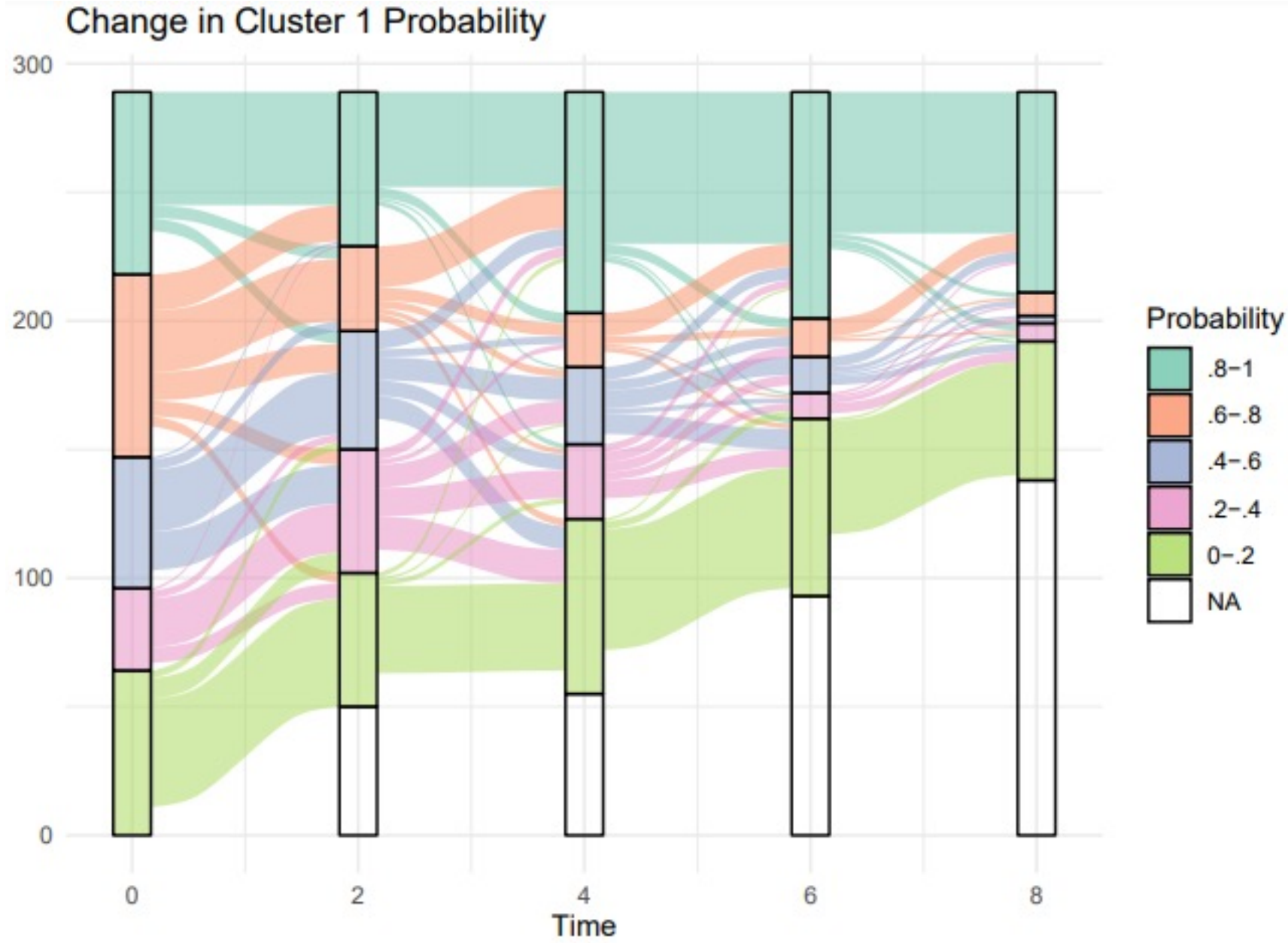

diffuse
cluster1
cluster2
0.5
0.0
−0.5
−1.0
−1.5
mean
0.0
2.5
5.0
7.5
10.0
year
measure
DLCO
FVC


male
cluster1
cluster2
mean
0
−1
−2
0.0
2.5
5.0
7.5
10.0
year
measure
DLCO
FVC


male, late_ageonset
cluster1
cluster2
mean
0
−1
0.0
2.5
5.0
7.5
10.0
year
measure
DLCO
FVC

Time-Varying Prediction of Cluster Membership




Change in Cluster 1 Probability
300
200
100
0
0
2
4
6
8
Time
Probability
.8−1
.6−.8
.4−.6
.2−.4
0−.2
NA

Model Choice

BIC: smaller the better

L: number of clusters

L=1	L=2	L=3
4478.411	4124.750	4136.595

Other Ways of Updating Trajectory Prediction

$$P(Y_{0:t}|X, splines) = \int P(Y_{0:t}|X, splines, b_i)P(b_i|X, splines)db_i$$
$$= \int P(Y_{0:t}|X, splines, b_i)P(b_i)db_i$$
$$= \int \left[P(Y_0|X, splines) \prod_{k\in\{1,\dots,t\}} P(Y_k|\bar{Y}_{k-1}, X, splines, b_i)\right] P(b_i)db_i$$

$$P(Y_{0:t}|X, splines) = P(Y_0|X, splines) \prod_{k\in\{1,\dots,t\}} P(Y_k|\bar{Y}_{k-1}, X, splines\ )$$
$$= \prod_{k\in\{0,\dots,t\}} \left[\int P(Y_k|\bar{Y}_{k-1}, X, splines, b_i\ )P(b_i|\bar{Y}_{k-1}, X, splines)db_i\right]$$
$$= \prod_{k\in\{0,\dots,t\}} \left[\int P(Y_k|\bar{Y}_{k-1}, X, splines, b_i\ )P(b_i|\bar{Y}_{k-1})db_i\right]$$

By model assumption, $P(b_i|X, splines) = P(b_i) = N(0, G)$,
While $P(b_i|\bar{Y}_{k-1}, X, splines) = P(b_i|\bar{Y}_{k-1}, X, splines) \neq N(0, G)$

Different Specifications of a Clustering Model

$$Y = \beta_0^{\ell} + X\beta_1^{\ell} + Xs(t)\beta_2^{\ell} + s(t)\beta_3^{\ell} + b_{i0}^{\ell} + \epsilon$$
$$\ell = 1,2$$

Question being addressed: any unobserved factors influencing the effect of X on Y over time? AND any unobserved factors STILL influencing the progression of Y after accounting for all of those effects?

$$Y(t) = \beta_0^{(\ell)} + b_{i0}^{(\ell)} + s(t)\beta_2^{(\ell)} + X\beta_1 + Xs(t)\beta_3 + \epsilon$$
$$\ell = 1,2$$

Question being addressed: any unobserved factors are influencing Y over time besides X?

Future extension

$$Y = \beta_0^{\ell} + \phi_X * X\left[\beta_1^{\ell} + s(t)\beta_2^{\ell}\right] + s(t)\beta_3^{\ell} + b_{i0}^{\ell} + \epsilon$$

$$\ell = 1,2$$

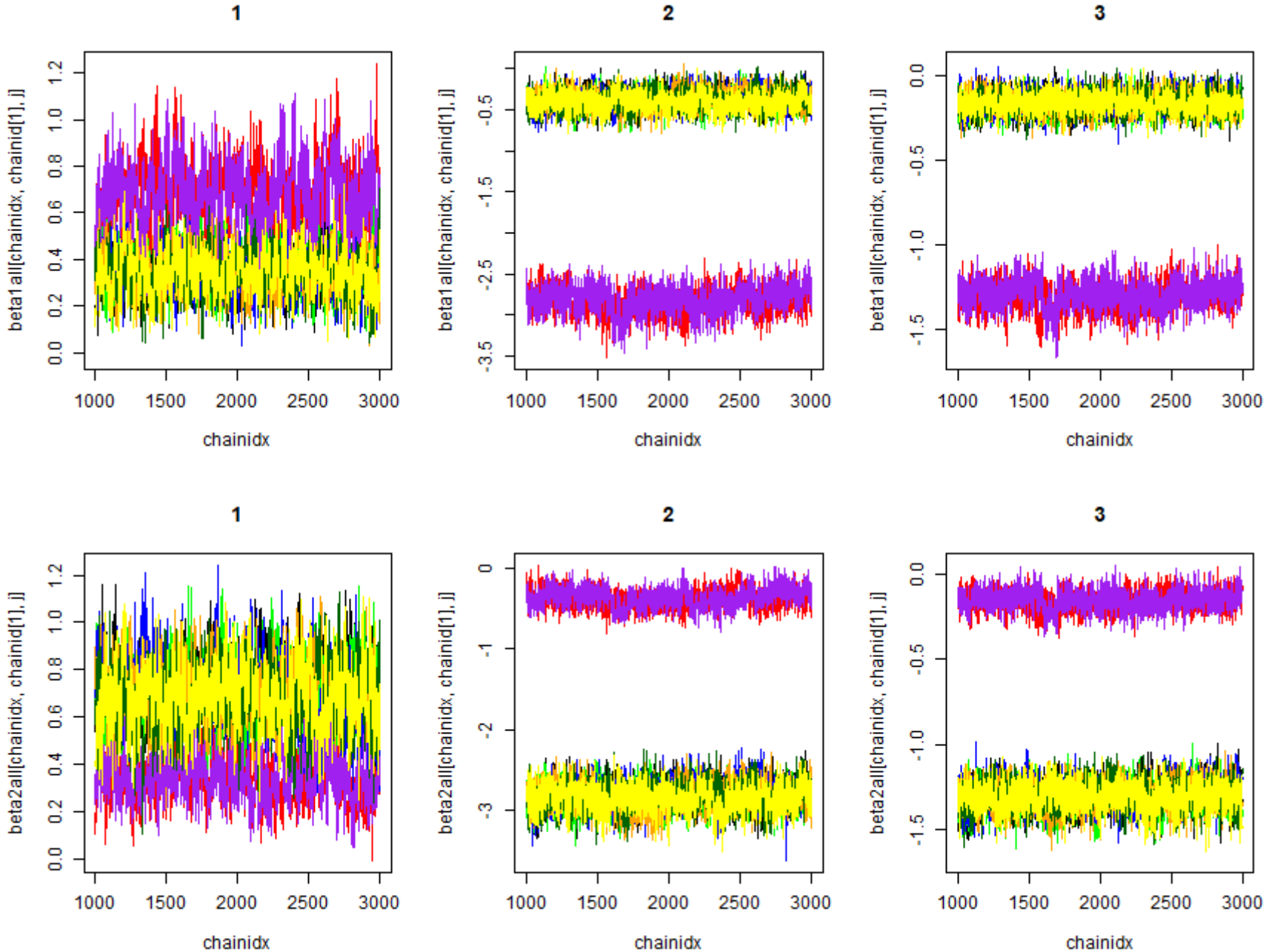
$$\phi_X = 1\{is\ X\ cluster\ specific\}$$

$$\phi_X \sim binary$$

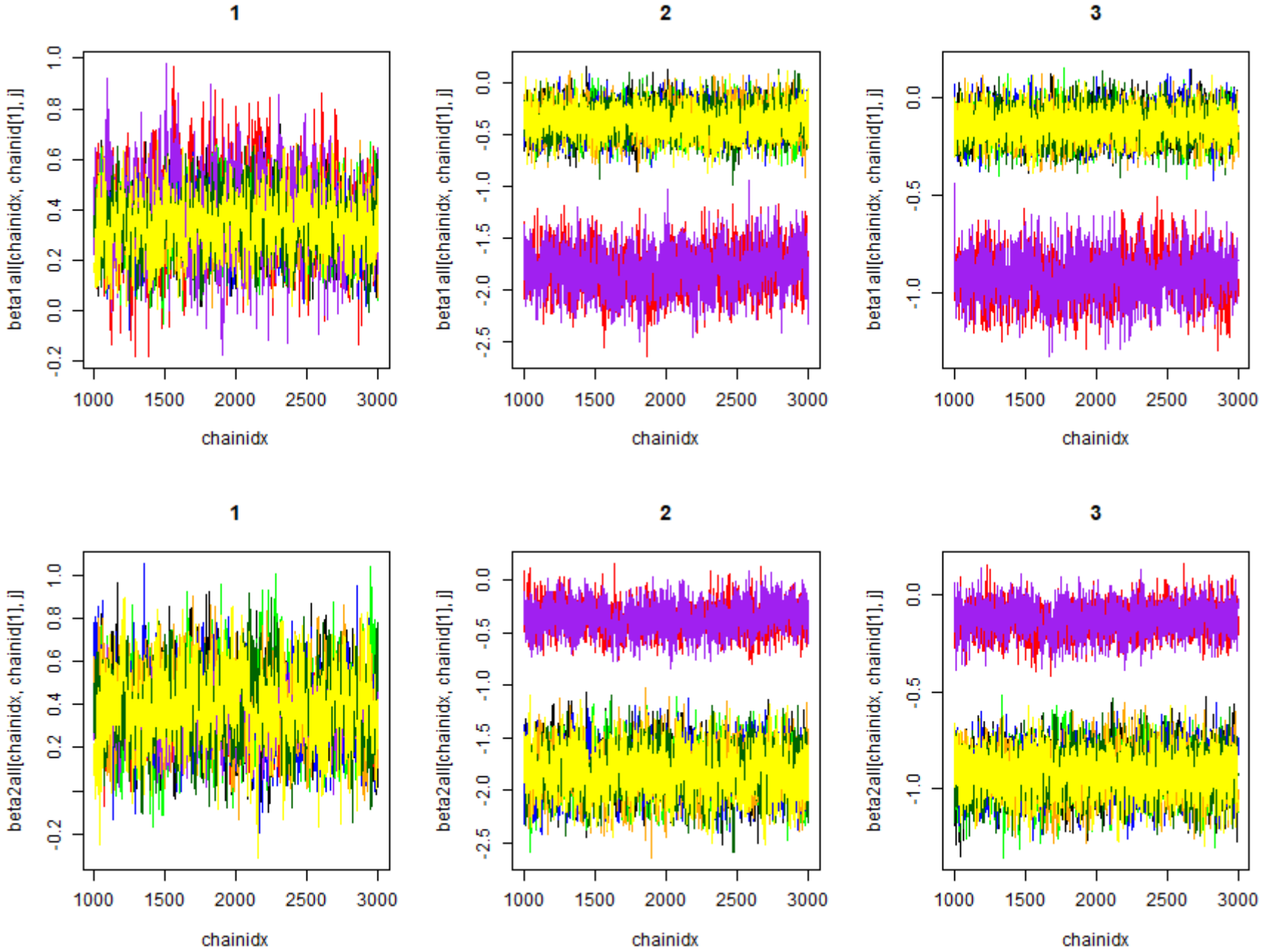
Spike and slab prior on $\beta^{(2)}$ in sets (main effects and spline interactions)

Label Switching

FVC betas


1
2
3
beta1all[chainidx, chainid[1], j]
beta2all[chainidx, chainid[1], j]
chainidx
1000
1500
2000
2500
3000

DLCO betas


1
2
3
beta1.all[chainidx, chainid[1], j]
beta2.all[chainidx, chainid[1], j]
chainidx
1000
1500
2000
2500
3000

Alphas of Intercept male AArace diffuse late_ageonset