Download and installation of mixADA

Given that you have a recent version of R installed, mixADA can be installed by:

```
Installation from GitHub, using package devtools:
install.packages("devtools")
library("devtools")
install_github(repo="schaarschmidt/mixADA")
```

Data structure

The data should be a csv file (with decimal point, but comma-separating columns). It should contain

- "Sample ID": a column with unique sampleID (the same for spiked and unspiked),
- "Sample type": a column distinguishing different types of samples (spiked and unspiked samples, negative controls, positive controls, etc.),
- one or several columns for secondary factors, e.g. days, devices, plates (=runs), analysts, labs, technical replications, etc.,
- "Response variable": the endpoint (e.g. optical density)

Excel

	A	B	C	D	E	F	G
1	Sample ID	Sample type	Day	Plate / Run	Analyst	Repeat	Optical Density
2	sample 1	unspiked	1	1	AA	1	0.070
3	sample 1	unspiked	1	1	AA	2	0.090
4	sample 2	unspiked	1	1	AA	1	0.161
5	sample 2	unspiked	1	1	AA	2	0.122
6	sample 3	unspiked	1	1	AA	1	0.089
7	sample 3	unspiked	1	1	AA	2	0.064
8	sample 4	unspiked	1	1	AA	1	0.235
9	sample 4	unspiked	1	1	AA	2	0.250
10	sample 5	unspiked	1	1	AA	1	0.177
11	sample 5	unspiked	1	1	AA	2	0.181
12	sample 1	spiked	1	1	AA	1	0.087
13	sample 1	spiked	1	1	AA	2	0.092
14	sample 2	spiked	1	1	AA	1	0.078
15	sample 2	spiked	1	1	AA	2	0.088
16	sample 3	spiked	1	1	AA	1	0.088
17	sample 3	spiked	1	1	AA	2	0.053
18	sample 4	spiked	1	1	AA	1	0.077
19	sample 4	spiked	1	1	AA	2	0.062
20	sample 5	spiked	1	1	AA	1	0.074
21	sample 5	spiked	1	1	AA	2	0.071
22	NC	negative control	1	1	AA	1	0.069
23	NC	negative control	1	1	AA	2	0.066
24	HSP	high screen positive control	1	1	AA	1	1.574
25	HSP	high screen positive control	1	1	AA	2	1.439
26	LSP	low screen positive control	1	1	AA	1	0.323
27	LSP	low screen positive control	1	1	AA	2	0.297

CSV

```
Sample ID,Sample type,Day,Plate number,Analyst,Repeat,Result,
sample 1,unspiked,1,1,AA,1,0.070,
sample 1,unspiked,1,1,AA,2,0.090,
sample 2,unspiked,1,1,AA,1,0.161,
sample 2,unspiked,1,1,AA,2,0.122,
sample 3,unspiked,1,1,AA,1,0.089,
sample 3,unspiked,1,1,AA,2,0.064,
sample 4,unspiked,1,1,AA,1,0.235,
sample 4,unspiked,1,1,AA,2,0.250,
sample 5,unspiked,1,1,AA,1,0.177,
sample 5,unspiked,1,1,AA,2,0.181,
sample 1,spiked,1,1,AA,1,0.087,
sample 1,spiked,1,1,AA,2,0.092,
sample 2,spiked,1,1,AA,1,0.078,
sample 2,spiked,1,1,AA,2,0.088,
sample 3,spiked,1,1,AA,1,0.088,
sample 3,spiked,1,1,AA,2,0.053,
sample 4,spiked,1,1,AA,1,0.077,
sample 4,spiked,1,1,AA,2,0.062,
sample 5,spiked,1,1,AA,1,0.074,
sample 5,spiked,1,1,AA,2,0.071,
NC,negative control,1,1,AA,1,0.069,
NC,negative control,1,1,AA,2,0.066,
HSP,high screen positive control,1,1,AA,1,1.574,
HSP,high screen positive control,1,1,AA,2,1.439,
LSP,low screen positive control,1,1,AA,1,0.323,
LSP,low screen positive control,1,1,AA,2,0.297,
```

mixADA is available for two data structures:

i) un-normalized, raw data (with duplicated wells per run and ID, negative control groups for normalization must be in the data as proposed by Shankar et al 2008): `mixADAserver()`

ii) already normalized data: (expected to contain one value per run and ID, spiked/unspiked) as used by Kubiak et al. 2013: `mixADAsimple()`

```
# starting the mixADA application:
library("mixADA")
mixADAserver()                                          # or
mixADAsimple()
```

Example 1: start mixADAsimple(), load 'Kubiak2013A1long.csv' [a)]

Data import & SCP	Select:	Remark
Response variable	value	
Sample type	study	*contains levels screening and confirmatory only*
Sample type level	screening	*defines which "sample type" is used*

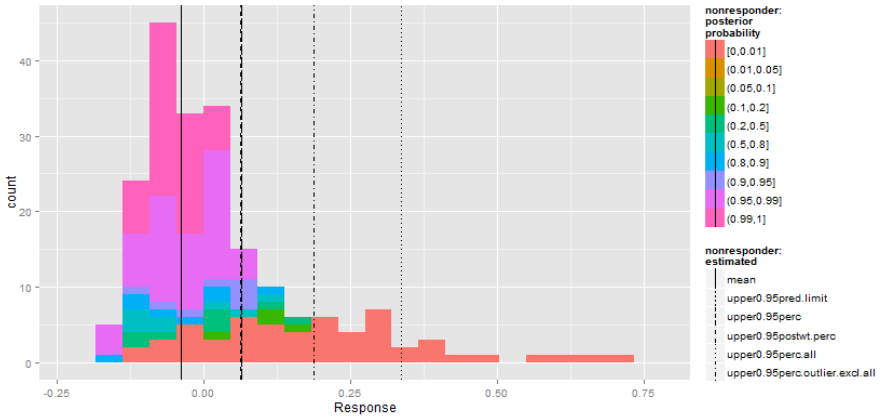
		for CP calculations (e.g. "unspiked" for the screening CP)
SampleID	ID	*Unique identifier of the 50 specimens*
Log-transform data	yes	*Calculate with and/or without*
random effects	Equal residual, different random effects	*Different choices are possible, with similar result* [b)]
Variable(s) def. technical replicates (runs)	plate	*Simple: one factor only* [c)]
Structure	Runs crossed with sampleIDs	*all sampleIDs once on each plate* [d)]

Start model fitting' by ticking the box (model fitting may take some time; if you want to refit the model after changing several options: remove tick in the box 'start', then apply all changes and finally tick the box 'Start model fitting)

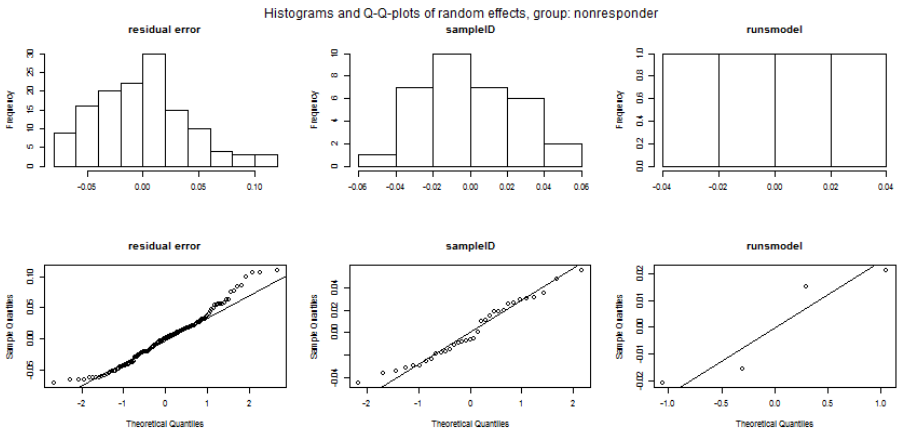
CCP estimation	**Select:**	**Remark**
Sample type level	confirmatory	
Compute CCP for	Percent inhibition	If negative inhibition is possible, take ratio spiked/unspiked values

a) provided data set resembles that presented by Kubiak et al. 2013, Tab. A1, w.r.t. structure and summary statistics
b) trying different options and comparing BIC values shows: a model allowing different variances between sampleIDs and runs fits the data best.
c) if several factors are available: several factors can be chosen, and their levels will be combined to **one single factor** defining runs to be used in the model fitting
d) each plate contains all sampleIDs: a simple 2 factorial design (check the 1st figure). In other cases there may is no clearly best choice: you may adapt the variables def. technical replications such that a simple design results or choose 'simplified pool over runs' to fit a mixture model for only one value per sampleID.

Selected results



Interpretation: Blue to green colors indicate that for a number of specimens the classification in this 2-component mixture model is unclear, they cannot be assigned clearly to either of the two groups.



Interpretation: Histograms or QQ-plots can be used. There may be slight violations of the normality assumptions in the non-responder group (the group with lower mean). For the residual error (left column) there a slight indications of a left-skewed distribution (some observations have more extreme positive residual errors than expected for a normal distribution). The distribution of specimen-means (mid column) is symmetric (histogram, upper row) and does not clearly deviate from a normality assumption (lower row). Not enough technical runs (i.e., plates) to say anything about their distribution (right column).

Mixture model fit: parameter estimates and size of groups (a posteriori)

	labels	mean	V.ID	V.runs	V.res	no.ID	no.obs
Comp.1	responder	0.1517	0.0164	0.0177	0.0014	17	68
Comp.2	nonresponder	-0.0368	0.0010	0.0011	0.0014	33	132

17 out of 50 specimens are assigned to the group with higher mean (labeled ‘responder’, the putative ADA-positive subgroup). 33 specimens are assigned to the group with low mean response, the putative ADA-negative group (labeled ‘nonresponder’). Note that the high mean in the putative ADA-positive group goes along with high variances between specimens and between runs.

Box-Cox-Lambda and likelihood ratio test (LRT) for normality and lognormality in mixed effects model

	lambda	LogLikelihood
Assumption: Normal distribution	1.0000	203.9867
Assumption: Lognormal distribution	0.0000	205.3459
Estimate*	-1.7000	206.2030

	H0	HA	statLRT	Pr(> chi(df=1))
1	Normal (lambda=1)	Dev. from Normal	4.4325	0.0353
2	Lognormal (lambda=0)	Dev. from Lognormal	1.7142	0.1904

The estimate of the Box-Cox-Parameter is -1.7 (indicating skewness). For this lack-of-fit tests a larger p-value argues for its underlying distribution, i.e. lognormal is more likely (p=0.1904).

Cutpoint estimation

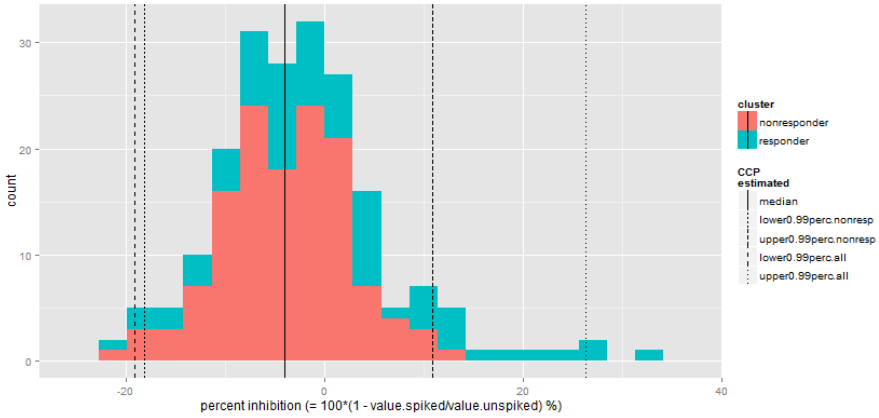
	value	backtransformed	group	Estimated
2	0.0641	**1.0662**	nonresponder	**upper0.95pred.limit**

3	0.0624	1.0644	nonresponder	upper0.95postwt.perc
4	0.0634	1.0655	nonresponder	upper0.95perc
5	0.3356	1.3988	all	upper0.95perc.all
6	0.1869	1.2055	all	upper0.95perc.outlier.excl.all (Shankar-style)

A model with 1 subgroup (BIC = -262.944) DOES NOT provide a better fit to this data set than the model with 2 subgroups shown here (BIC = -334.064). In group 'nonresponder': 'pred.int': prediction limit (for 1 future observation) based on fitting a random effects model to those observations that were classified as 'nonresponder' in the 2-component mixture model; 'postwt.perc' percentile of a sample of the original observations, weighted by the posterior probability to be member of group 'nonresponder'; 'perc': percentile of those observations that were classified as 'nonresponder' in the 2-component mixture model; 'perc.all' in group 'all': percentile of all observations (irrespective of classification in responders or nonresponders); 'perc.outlier.excl.all' in group 'all': percentile of all observations after exclusion of potential outliers:18 observations excluded because they exceed Q75+1.5*IQR= 0.2749 namely from samples 14:1, 18:1, 19:4, 20:1, 25:1, 28:3, 41:1, 46:4, 48:2.

The SCP is 1.07 (log-normal model can be assumed, see Box-Cox, above), mixing distribution was selected, for ADA- classified samples only, prediction interval takes plate effects into account)

CCP



Estimated median and empirical percentiles for percent inhibition (= 100*(1 - value.spiked/value.unspiked) %)

	group	estimated	value
2	nonresponder	lower0.99perc.nonresp	-18.2026
3	nonresponder	**upper0.99perc.nonresp**	**10.8000**
4	all	lower0.99perc.all	-19.1483
5	all	upper0.99perc.all	26.3016

The CCP is 10.8%

Example 2: start mixADAserver() and load ‘dps.csv’ (Shankar-design)

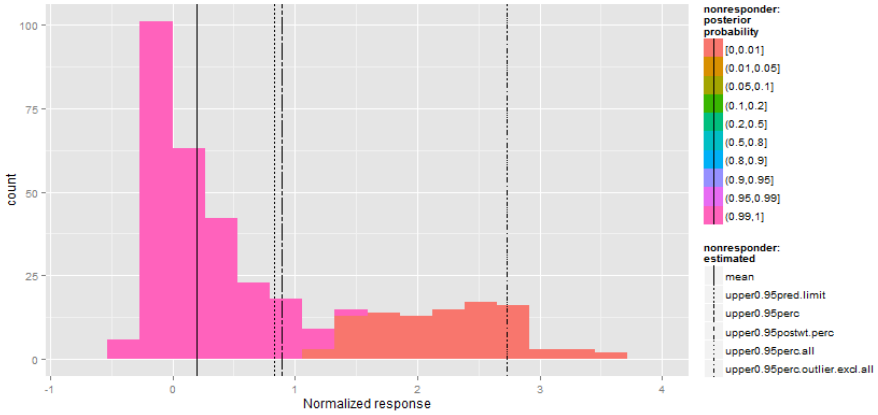
Normalization & SCP	Select:	Remark
Response variable	result	
Sample type	sampletype	*contains levels: NC, LowQC, HighQC, treated, untreated*
Level for normalization	NC	*here, NC is a neg. control suitable for normalization*
Sample type level	untreated	*Biological samples (uninhibited)*
SampleID	sampleID	*Unique identifier of specimens*
Normalization:	log-transform and subtract	
	median	*mean leads to similar results*
Runs for normalization	plate, day	*plate is a unique ID for the technical unit only within each day* [a)]
random effects	Equal residual, different random effects	*Different choices are possible, with similar result*
Runs for model fitting	day	*Each sampleID is analysed on each day but not on each plate per day* [b)]
Structure	Runs crossed with sampleIDs	*A simple crossed design sampleIDs once on each plate*

a) within each day, there are several plates, numbered 1, …, 5 within each day. Choosing plate only as ‘Run for normalization’ would lead to pooling plates with the same number but from different days. Choosing day only, would lead to pooling all plates within a given day for the normalization step.

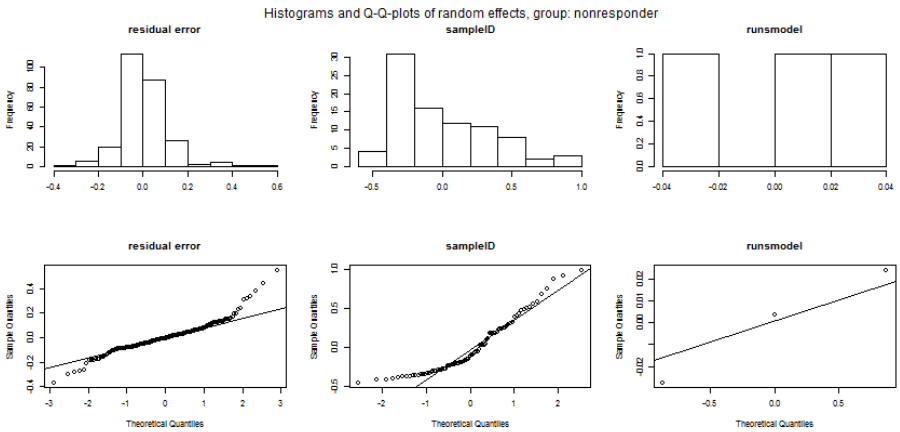
b) within each day, each sampleID is present. There are too many sampleIDs in total to be assessed on each plate. Accounting for plates nested within days would require a more complicated model than implemented in mixADA. Further, thorough investigation of the data reveals that sampleIDs may have been assigned to plates ordered by response (not shown). If so, including plate effects in the model would errroneously assign biological variance (sampleID) to between plate variance.

Start model fitting’ by ticking the box

Selected results and interpretation



Interpretation: colors indicate the posterior probability to belong to the subgroup with the lower mean (putative ADA-negatives), i.e., high probabilities (violet) indicate that the observations are classified to the group with the low mean response.



Interpretation: Normality assumptions are not met. Left column: residual errors have a symmetric histogram (upper row), but the QQ-plot (lower row) shows variance heterogeneity; mid column: variance of sample ID shows a clearly right skewed distribution (histogram and QQ-plot); right column: ignore, 3 runs (provide no basis for any interpretation of their distribution.)

Mixture model fit: parameter estimates and size of groups (a posteriori)

	labels	mean	V.ID	V.runs	V.res	no.ID	No.obs
Comp.1	responder	2.2313	0.2853	0.0156	0.0096	33	99
Comp.2	nonresponder	0.1966	0.1254	0.0083	0.0096	87	261

87 specimens have been classified to the group with low mean (putative ADA-negative), 33 specimens are classified to the group with high mean response (putative ADA-positive). The Group with high mean also exhibits higher variances between IDs and between runs (i.e. days), residual variance was assumed equal.

Box-Cox-Lambda and LRT for normality and lognormality in mixed effects model

	lambda	LogLikelihood
Assumption: Normal distribution	1.0000	-126.8141
Assumption: Lognormal distribution	0.0000	-28.6209
Estimate*	-1.1000	5.4074

	H0	HA	statLRT	Pr(> chi(df=1))
1	Normal (lambda=1)	Dev. from Normal	264.4430	0.0000
2	Lognormal (lambda=0)	Dev. from Lognormal	68.0568	0.0000

The data deviate from both normal distribution and from log-normal distribution (p<0.0001). Therefore, a nonparametric percentile interval should be used (see below)

Estimated mean, prediction limit and quantiles for 'nonresponder'

	value	backtransformed	group	estimated
2	0.8300	2.2985	nonresponder	upper0.95pred.limit
3	0.9222	2.4356	nonresponder	upper0.95postwt.perc
4	0.9223	**2.4358**	nonresponder	**upper0.95perc**
5	2.7346	15.3474	all	upper0.95perc.all
6	2.7346	15.3474	all	upper0.95perc.outlier.excl.all

The SCP is 2.44 (95% percentile of the 'non-responder' subgroup)

For technical questions send an e-mail to schaarschmidt(at)biostat.uni-hannover.de