

Introduction to Bayesian record linkage

Brenda Betancourt and Andee Kaplan

Duke University
Department of Statistical Science

February 8, 2018

Slides available at <http://bit.ly/cimat-bayes>

Why Bayes?

Common approaches

Felligi-Sutner
Clustering
Beka's

Empirical Bayes graphical model for record linkage

blink package

R package that removes duplicate entries from multiple databses using the empirical Bayes graphical method:

```
install.packages("blink")
```

- Formatting data for use with blink
- Tuning parameters
- Running the Gibbs sampler (estimate model parameters)
- Output

RLdata500 data

We will continue with the RLdata500 dataset in the RecordLinkage package consisting of 500 records with 10% duplication.

```
library(blink) # load blink library
library(RecordLinkage) # load data library
data("RLdata500") # load data
head(RLdata500) # take a look
```

##	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd
## 1	CARSTEN	<NA>	MEIER	<NA>	1949	7	22
## 2	GERD	<NA>	BAUER	<NA>	1968	7	27
## 3	ROBERT	<NA>	HARTMANN	<NA>	1930	4	30
## 4	STEFAN	<NA>	WOLFF	<NA>	1957	9	2
## 5	RALF	<NA>	KRUEGER	<NA>	1966	1	13
## 6	JUERGEN	<NA>	FRANKE	<NA>	1929	7	4

Formatting the data

```
# categorical variables
```

```
X.c <- as.matrix(RLdata500[, c("by", "bm", "bd")])
```

```
p.c <- ncol(X.c)
```

```
# string variables
```

```
X.s <- as.matrix(RLdata500[, c("fname_c1", "lname_c1")])
```

```
p.s <- ncol(X.s)
```

X.c and X.s include all files stacked on top of each other, for categorical and string variables respectively

```
# keep track of which rows of are in which files
```

```
file.num <- rep(c(1, 2, 3), c(200, 150, 150))
```

Tuning parameters

Hyperparameters

```
# Subjective choices for distortion probability prior  
a <- 1  
b <- 999
```

Distortion

```
# string distance function example  
d <- function(s1, s2) {  
  adist(s1, s2) # approximate string distance  
}  
  
# steepness parameter  
c <- 1
```