University of Reading

Dankmar Böhning

Professor and Chair
in Applied Statistics

University of Reading

new and nice buildings

University of Reading

on a very nice and green campus

Postgraduate Studies

- MSc Biometry (1 year)
- 15-20 students per year

Postgraduate Studies

- PH programme in Applied Statistics

  (3 years)

  10 students

- MSc + PhD (3+1)

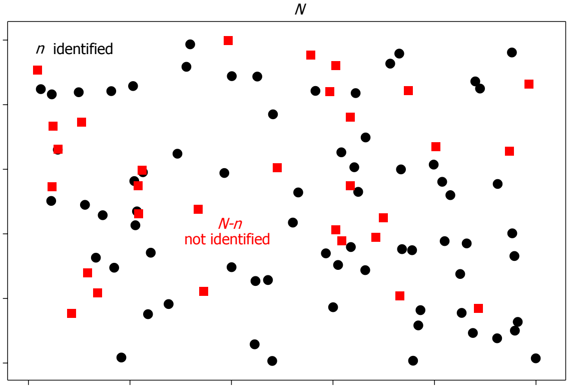
Capture-Recapture Estimation of Population Size by Means of Empirical Bayesian Smoothing

Dankmar Böhning

Professor and Chair in Applied Statistics, School of Biological Sciences
University of Reading

Pattaya, Thailand, 22 May 2009

a population has N units of which n are identified by some mechanism (trap, register, police database, ...)



Formulation of the Problem

- probability of identifying an unit is $(1 - p_0)$
- so that $N = \underbrace{(1 - p_0)N}_{observed} + \underbrace{p_0 N}_{hidden} = n + p_0 N$
- and the **Horvitz-Thompson** estimator follows:

$$\hat{N} = \frac{n}{1 - p_0}$$

- usually an **estimate** of p_0 is required

Formulation of the Problem as Frequencies of Frequencies

Frequencies of Frequencies

a common setting for estimating p_0 is the **Frequencies of Frequencies** setting:

Identifying Mechanism

the identifying mechanism provides a count Y of **repeated identifications** (w.r.t. to a reference period)

Illustration of the CR-Concept

Table: *Illustration with Case Data from Software Inspection (Wohlin et al. 1995)*

	Reviewers				
Error i	R1	R2	...	R22	Marginal Y_i
1	1	0	...	1	2
2	1	1	...	0	4
3	0	0	...	1	2
4	0	0	...	0	**0**
5	0	1	...	0	1
...	...	...	...	...	...
38	1	1	...	0	7

Formulation of the Problem as Frequencies of Frequencies

Marginal distribution

marginal distribution of Y is leading to frequencies $f_1, f_2, ..., f_m$ of the counts $1, 2, .., m$ (m is the largest observed count)

estimating f_0 on the basis of $f_1, f_2, ..., f_m$

zero counts are **not** observed: hence f_0 **is unknown**
Recall that $N = f_0 + n = f_0 + f_1 + f_2 + ... + f_m$, so that $\hat{f}_0$ leads to $\hat{N}$

Illustration of the frequencies of frequencies situation at hand of the software inspection data

Table: Zero-truncated count distribution of software errors

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	5	1	5	1	3	2	0	5	4	2

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}	n
3	1	0	2	0	1	0	0	0	1	36

Application Areas

- Epidemiology and Medicine
- Biology and Agriculture
- Social Science and Criminology
- Research on Terrorism
- Systems Engineering

Hser's Data on Estimating Hidden Intravenous Drug Users in Los Angeles 1989

- intravenous drug users in L.A. county were entered into the California Drug Abuse Data System (CAL-DADS)
- the data below refer to the frequency distribution of the episode count per drug user in 1989

the frequency distribution of the **episode count per drug user** for the year 1989:

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Del Rio Vilas's Data on Estimating Hidden Scrapie in Great Britain 2005

- sheep is kept in holdings in Great Britain (and elsewhere)
- the occurrence of scrapie is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) summarizing abbatoir survey, stock survey and the statutory reporting of clinical cases
- CSFS established since 2004

the frequency distribution of the **scrapie count within each holding** for the year 2005:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Drakos' Data on Estimating Hidden Transnational Terrorist Activity

- data on terrorism are provided by various databases including RAND terrorism chronology, the terrorism indictment and DFI International research on terrorist organizations on 153 countries and the period 1985-1998
- terrorism is violence or the threat of violence, calculated to create an atmosphere of fear and alarm

Drakos' Data on Estimating Hidden Transnational Terrorist Activity

the frequency distribution (Drakos 2007) of the **count of transnational terrorist activity** Y_{it} in country i and year t:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	...	f_{136}	n
-	286	114	101	59	33	21	20	19	11	...	1	785

- similar to other data sets there is an $f_0 = 1,357$
- however it is thought that there is a hidden number of terrorist activities which is of interest to be estimated
- estimate f_0 the frequency of periods and countries **with unrecorded terrorist activites**

Formulation of the Problem and the Idea for its Solution

Suppose we can find some model for the count probabilities

$$p_j = p_j(\lambda)$$

then estimate λ by some method (truncated likelihood) and then use the model for p_0:

$$\hat{N} = \frac{n}{1 - p_0(\hat{\lambda})}$$

Formulation of the Problem and the Idea for its Solution

Only to illustrate: Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

then estimate λ maximizing the zero-truncated Poisson likelihood

$$\prod_{j=1}^{m}\left(\frac{p_j}{1-p_0}\right)^{f_j} = \prod_{j=1}^{m}\left(\frac{1}{1-\exp(-\lambda)}\exp(-\lambda)\lambda^j/j!\right)^{f_j}$$

equivalently:

find solution $\hat{\lambda}$ of

$$\lambda = \frac{S}{n}(1 - \exp(-\lambda))$$

where $S = f_1 + 2f_2 + ... + mf_m$

in passing note:

$\lambda^{(j+1)} = \frac{S}{n}(1 - \exp(\lambda^{(j)}))$ is an **EM algorithm**

1. $\lambda^{(j)} = \frac{S}{N^{(j)}}$
2. $N^{(j+1)} = \frac{n}{1-\exp(-\lambda^{(j)})}$

estimate for N:

$$\hat{N} = \frac{n}{1 - \hat{p}_0} = \frac{n}{1 - \exp(-\hat{\lambda})}$$

What speaks against this simple solution?

However: using a simple Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

is **not** appropriate, since

- every unit is different
- there is population heterogeneity

so that more **realistic**

$$p_j = p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

where $\lambda(t)$ stands for the heterogeneity distribution of the Poisson parameter

Effects of Heterogeneity?

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
1	MLE-hom	**101.91**	12.98	13.12
	Chao	103.82	18.73	19.12
2	MLE-hom	**94.07**	7.02	9.19
	Chao	99.10	12.22	12.25
3	MLE-hom	**88.19**	4.96	12.81
	Chao	96.61	9.77	10.34
4	MLE-hom	**85.34**	4.30	15.30
	Chao	97.03	10.00	10.43
5	MLE-hom	**83.47**	3.71	16.94
	Chao	97.98	10.24	10.43

Effect of Heterogeneity on an estimator under homogeneity:

underestimation because of Jensen's inequality applied to $\exp(x)$:

$$\frac{n}{1-p_0} = \frac{n}{1-\int_0^\infty \exp(-t)\lambda(t)dt}$$
$$\geq \frac{n}{1-\exp\left(-\int_0^\infty t\lambda(t)dt\right)}$$
$$= \frac{n}{1-\exp(-\mu)},$$

where $\mu = \int_0^\infty t\lambda(t)dt$

Simple nonparametric estimates under heterogeneity

under heterogeneity

before we look at methods providing an estimate $\hat{\lambda}(t)$ in

$$p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

by means of

- **parametric** (Chao and Bunge 2002 *Biometrics*)
- or **nonparametric mixture models** (Böhning and Schön 2005, *JRSSC*)

interest is on the lower bound approach by **Chao** (1987, 1989, *Biometrics*) consider

$$p_j = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

with unknown $\lambda(t)$ for $t > 0$. Then, by the **Cauchy-Schwarz inequality** for random variables X and Y:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

it is

$$\left(\int_0^\infty \overbrace{\sqrt{\exp(-t)}}^{X} \overbrace{\sqrt{\exp(-t)}t}^{Y} \lambda(t)dt \right)^2$$

$$\leq \int_0^\infty \overbrace{\exp(-t)}^{X^2} \lambda(t)dt \int_0^\infty \overbrace{\exp(-t)t^2}^{Y^2} \lambda(t)dt$$

or,

$$p_1^2 \leq p_0 2 p_2$$

$$p_1^2 \leq p_0 2 p_2 \Leftrightarrow \frac{p_1^2}{2p_2} \leq p_0$$

leads to **Chao's lower bound estimate**

$$\hat{f}_0 = \frac{f_1^2}{2f_2}$$

or

$$\hat{N} = n + \hat{f}_0 = n + \frac{f_1^2}{2f_2}$$

Chao's estimator is good since **truely nonparametric** and conservative

Comparing the Estimators

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
1	MLE-hom	101.91	12.98	13.12
	Chao	**103.82**	18.73	19.12
2	MLE-hom	94.07	7.02	9.19
	Chao	**99.10**	12.22	12.25
3	MLE-hom	88.19	4.96	12.81
	Chao	**96.61**	9.77	10.34
4	MLE-hom	85.34	4.30	15.30
	Chao	**97.03**	10.00	10.43
5	MLE-hom	83.47	3.71	16.94
	Chao	**97.98**	10.24	10.43

Problems with the NPMLE

under heterogeneity:

$$p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

estimation under heterogeneity: the NPMLE

maximize zero-truncated Poisson mixture likelihood in Q

$$L(Q) = \prod_{j=1}^{m} \left(\frac{p_j}{1-p_0}\right)^{f_j} = \prod_{j=1}^{m} \left(\sum_{\ell=1}^{k} \frac{Po(j|t_\ell)\lambda_\ell}{1-\sum_i \exp(-t_i)\lambda_i}\right)^{f_j}$$

where

$$Q = \begin{pmatrix} t_1 & t_2 & \dots & t_k \\ \lambda_1 & \lambda_2 & \dots & \lambda_k \end{pmatrix}$$

Problems with the NPMLE

boundary problem:

$$f(0|\hat{Q}) \geq f_0/N$$

where

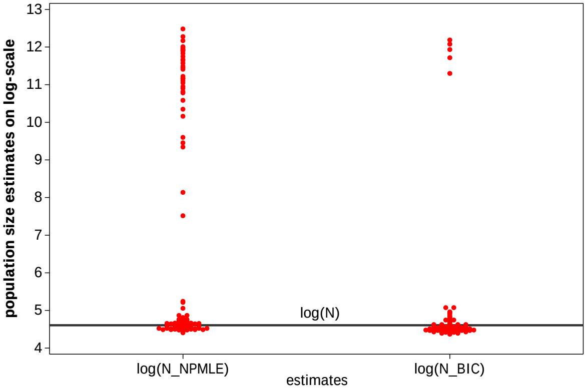
$$f(0|\hat{Q}) = \sum_{\ell} \exp(-t_\ell)\lambda_\ell$$

(Wang and Lindsay 2005, 2008; Harris 1991)

Illustration of Severity of Boundary Problem

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(t)$ *and* $N = 100$

t	estimator	mean	SD
1	Chao	102	17
	NPMLE	484	12,098
2	Chao	99	12
	NPMLE	4599	35,028
3	Chao	97	10
	NPMLE	12,517	52,425
4	Chao	97	9
	NPMLE	11,715	54,501
5	Chao	98	10
	NPMLE	4,657	33,069


population size estimates on log-scale
13
12
11
10
9
8
7
6
5
4
log(N)
log(N_NPMLE)
log(N_BIC)
estimates

The Idea for a robust approach of Zelterman (1988, *JSPI*)

empirical Bayes approach starts with an idea of Zelterman (1988)

- he noted that

$$\lambda = \frac{\lambda^{j+1}}{\lambda^j} = (j+1)\frac{\lambda^{j+1}/(j+1)!}{\lambda^j/j!}$$

$$\lambda = (j+1)\frac{Po(j+1;\lambda)}{Po(j;\lambda)}$$

- leading to the proposal

$$\hat{\lambda}_j = (j+1)\frac{f_{j+1}}{f_j}, \textbf{for } j = 1,\ \hat{\lambda} = \hat{\lambda}_1 = 2\frac{f_2}{f_1}$$

- Zelterman estimator

$$\hat{N}_Z = \frac{n}{1 - \exp(-2f_2/f1)}$$

The Idea of Zelterman

$\hat{\lambda} = 2\frac{f_2}{f_1}$ is **robust** in the sense that

- **positive:**
- it is **not affected** by any changes in counts larger than 2
- count distribution need only to behave **like** a Poisson for counts of 1 or 2
-
- **negative:**
- potential for large overestimation bias
- large variance

The Idea for an Empirical Bayes Approach

conventional Zelterman

$$\hat{N}_Z = \frac{n}{1 - \exp(-2f_2/f1)}$$

better:

$$\hat{N}_Z = \frac{f_1}{1 - \exp(-\lambda_1)} + \frac{f_2}{1 - \exp(-\lambda_2)} + \frac{f_3}{1 - \exp(-\lambda_3)} + \dots$$

but:

how to choose or estimate λ_x for $x = 1, 2, 3, \dots$?

The Idea for an Empirical Bayes Approach

how to choose or estimate λ_x for $x = 1, 2, 3, ...$?

a poor choice:

$$\lambda_{\mathbf{x}} = \mathbf{x}$$

so that

$$\hat{N}_Z = \frac{f_1}{1-\exp(-1)} + \frac{f_2}{1-\exp(-2)} + \frac{f_3}{1-\exp(-3)} + ...$$

The Idea for an Empirical Bayes Approach

We think of the mixing distribution $\lambda(t)$ as a prior distribution on t so that

$$\lambda_x = E(t|x) = \int_0^\infty t \frac{Po(x|t)\lambda(t)}{\int_0^\infty Po(x|\theta)\lambda(\theta)d\theta} dt \qquad (1)$$

is the *posterior mean* w.r.t the prior $\lambda(t)$ and Poisson likelihood for observation x.

Note that (1) can be further simplified to

$$\lambda_x = E(t|x) = \frac{\int_0^\infty tPo(x|t)\lambda(t)dt}{\int_0^\infty Po(x|t)\lambda(t)dt}$$

$$= \frac{\int_0^\infty te^{-t}t^x/x!\lambda(t)dt}{\int_0^\infty e^{-t}t^x/x!\lambda(t)dt}$$

$$(x+1)\frac{\int_0^\infty Po(x+1|t)\lambda(t)dt}{\int_0^\infty Po(x|t)\lambda(t)dt}$$

$$= (x+1)\frac{p_{x+1}}{p_x},$$

where $p_x = \int_0^\infty Po(x|t)\lambda(t)d(t)$ is the **marginal density** of X

An empirical Bayes version of Zelterman

choice of λ_x:

$$\lambda_x = E(t|x) = (x+1)\frac{p_{x+1}}{p_x}$$

to achieve

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1-\exp[-\lambda_x]} = \sum_{x=1}^{m} \frac{f_x}{1-\exp[-(x+1)p_{x+1}/p_x]}$$

with $p_x = \int_0^\infty Po(x|t)\lambda(t)dt$

empirical Bayes:

p_x can be estimated by the relative, empirical frequency f_x/N so that

$$\widehat{E(t|x)} = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

provides an estimate of the posterior mean $E(t|x) = \lambda_x$

important:

- the unknown denominators N cancel out
- idea is a special case of the nonparametric, empirical Bayes estimator (Robbins 1955, Carlin and Louis 1996).

Robbins approach:

hence, using

$$\widehat{E(\lambda|x)} = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

the empirical Bayes approach (Robbins) leads to

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)\frac{f_{x+1}}{f_x}]}$$

Empirical Bayesian Smoothing

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)\frac{p_{x+1}}{p_x}]}$$

with

$$p_x = \int_0^\infty Po(x|t)\lambda(t)dt$$

offers **options:**

1. Robbins
2. nonparametric smoothing with discrete mixture model
3. parametric smoothing with Gamma-mixing distribution
4. nonparametric smoothing with empirical distribution function

$$\hat{p}_x = \sum_{y=1}^{m} Po(x|y)\frac{f_y}{n}$$

Empirical Bayesian Smoothing

1. Robbins (no need for estimating $\lambda(t)$!!!)
2. nonparametric smoothing with discrete mixture model (computational expensive!)
3. parametric smoothing with Gamma-mixing distribution (computational instable)
4. nonparametric smoothing with empirical distribution function (not a good estimate of the mixing distribution)

Software Inspection

Table: Zero-truncated count distribution of software errors

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	5	1	5	1	3	2	0	5	4	2

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robbins	$\Gamma(t)$	EDF
49	1	36	244.1	36	50	37	37
	2	38	**211.4**	37			
	3	124,279	215.2	40			
	4	84,946	219.7	40			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Drug Use in California 1989

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Drug Use in California 1989

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robb.	$\Gamma(t)$	EDF
38,637	1	26,426	57,944	26,426	34,776	35,572	26,434
	2	39,183	52,262	33,757			
	3	58,224	**52,083**	34,756			
	4	424,168	52,085	34,766			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Hidden Scrapie in Great Britain

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robb.	$\Gamma(t)$	EDF
353	1	170	313.9	170	320	313	164
	2	274	**260.0**	310			
	3	1,111	263.2	320			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Hidden International Terroristic Activity

the frequency distribution (Drakos 2007) of the **count of transnational terrorist activity** Y_{it} in country i and year t:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	...	f_{136}	n
-	286	114	101	59	33	21	20	19	11	...	1	785

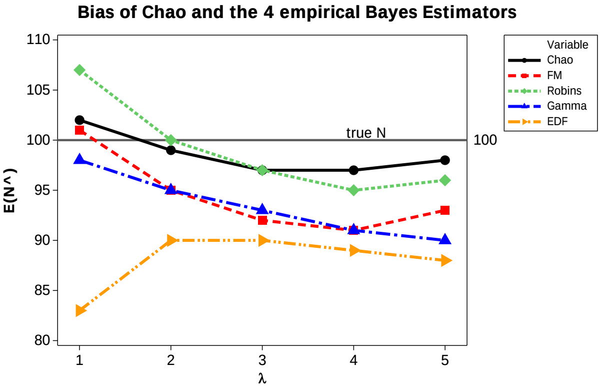
Hidden International Terroristic Activity

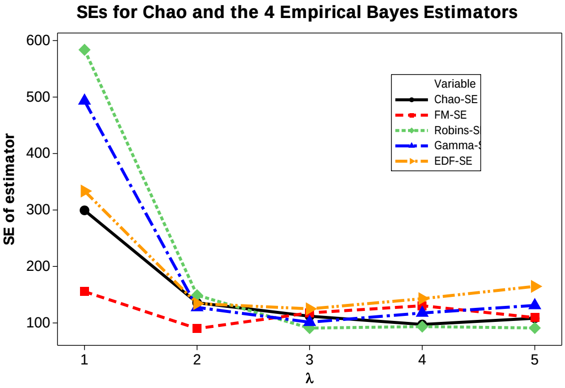
Table: *Estimate* $\hat{N}$

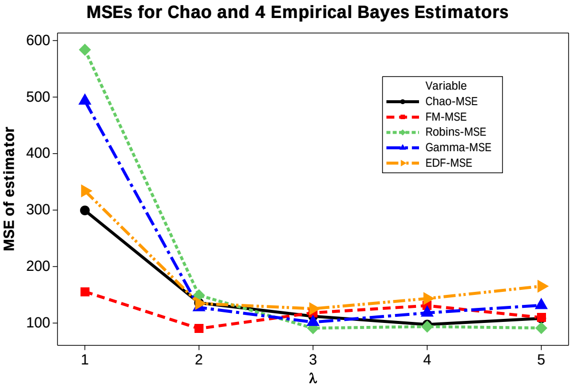
conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robb.	$\Gamma(t)$	EDF
1,144	1	787	9,699	787	1,044	847	912
	2	839	5,131	837			
	3	892	4,250	885			
	4	940	3960	918			
	5	964	3,909	930			
	6	965	3,893	930			
	7	966	3,899	931			
	8	1,386,688	**3,864**	1,035			
	9	963,940	3,874	1,035			
	10	1,474,006	3,883	1,033			

Simulation

- $N = 100$
- $Y \sim 0.5Po(1) + 0.5Po(\lambda)$
- count data $Y_i = 0$ removed from data set
- all estimators of interest computed delivering a value of $\hat{N}$


Bias of Chao and the 4 empirical Bayes Estimators
110
105
100
95
90
85
80
E(N^)
1
2
3
4
5
λ
true N
100
Variable
Chao
FM
Robins
Gamma
EDF


SEs for Chao and the 4 Empirical Bayes Estimators
SE of estimator
600
500
400
300
200
100
1
2
3
4
5
λ
Variable
Chao-SE
FM-SE
Robins-S
Gamma-S
EDF-SE


MSEs for Chao and 4 Empirical Bayes Estimators
MSE of estimator
600
500
400
300
200
100
1
2
3
4
5
λ
Variable
Chao-MSE
FM-MSE
Robins-MSE
Gamma-MSE
EDF-MSE

Conclusions

- direct application of nonparametric mixture models leads to **dramatic overestimation bias**
- Zelterman's estimator can be corrected and **generalized** to a valid estimator
- count specific parameters can be estimated via **posterior means**
- using as priors **estimated mixture models**
- **excellent properties** of the NPMLE can be **regained**

where to find things:

- Software by Kuhnert (2009): CR_Smooth
- references, publications, talks:
- www.reading.ac.uk/∼sns05dab