AALBORG UNIVERSITY

Y-STR: Haplotype Frequency Estimation and Evidence Calculation

Mikkel Meyer Andersen, MSc Student

Supervised by associate professor Poul Svante Eriksen

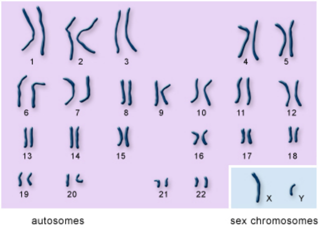
Department of Mathematical Sciences
Aalborg University, Denmark

June 16th 2010,
Presentation of Master of Science Thesis

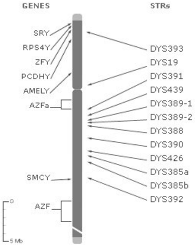
Outline

1. Short biological recap
2. Motivation and aims of using Y-STR
3. Frequency estimation
   - 3.1 Methods and comments to the methods
   - 3.2 Model control
4. Evidence calculation
5. Further work

Biological framework



U.S. National Library of Medicine



Example of (autosomal) STR DNA-type:

$$(\{14, 13\}, \{22\}, \{15, 16\}, \dots, \{22\})$$

Example of Y-STR DNA-type:

$$(17, 14, 22, \dots, 15)$$

LHS image is from http://ghr.nlm.nih.gov and RHS image is from http://history.earthsci.carleton.ca.

Motivation

- In some situations Y-STR is more sensible to use than A-STR (autosomal STR), e.g. to avoid noise in the trace from a rape victim
- Y-STR and A-STR differs in several areas, e.g. the number of alleles at each locus and statistical dependence between loci
- The statistical methods developed to handle A-STR cannot be applied on Y-STR directly, so reformulation is required (e.g. for calculating evidence) or new methods must be developed (to estimate Y-STR haplotype frequencies)

Aims

- Be able to calculate statistical evidence in trials
- Estimate frequencies for Y-STR haplotypes (also unobserved ones) is required to do this

First, methods for estimating frequencies for Y-STR haplotypes will be discussed and afterwards calculation of evidence will be introduced.

Estimating frequencies

Dimension reduction

Neither principal component analysis nor factor analysis yields good results, so that path has not been followed any further.

Existing methods

- Simple count estimates
  - Not precise enough
- One published method (used at `http://www.yhrd.org`): Frequency surveying introduced in ”A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males.“ from 2000 by L. Roewer *et al.*
  - Several problems exist; some will be presented in this presentation (some also presented in a talk at 7th International Y Chromosome User Workshop in Berlin, Germany, from April 22 to April 24, 2010)

New methods

- Graphical models would be an obvious choice
  - Structure based learning (e.g. PC algorithm) or score based learning (e.g. AIC and BIC)
  - Standard tests for conditional independence (e.g. $G^2 = 2nCE(A, B \mid \{S_i\}_{i \in I})$, where n is the sample size and CE is the cross entropy, which is χ^2_ν distributed when A and B are independent given $\{S_i\}_{i \in I}$) do not exploit the ordering in the data nor does it incorporate prior knowledge such as the single step mutation model
  - Better independence tests are required
- Ancestral awareness
- Classification models (e.g. classification trees, ordered logistic regression, and support vector machines)
- Kernel smoothing and model-based clustering

Frequency surveying

The idea: Bayesian approach

Notation:

- N: number of observations
- M: number of haplotypes (i.e. unique observations)
- N_i: the number of times the i'th haplotype has been observed
- $f_i = \frac{N_i - 1}{N - M}$: the frequency for the i'th haplotype

Model using Bayesian inference:

1. A priori: assume f_i is Beta distributed with parameters non-stochastic parameters u_i and v_i
2. Likelihood: Given f_i, then N_i is Binomial distributed
3. Posterior: Given N_i, then f_i is (still) Beta distributed (Beta distribution is a conjugate prior for the Binomial distribution)

Model expressed in densities using generic notation:

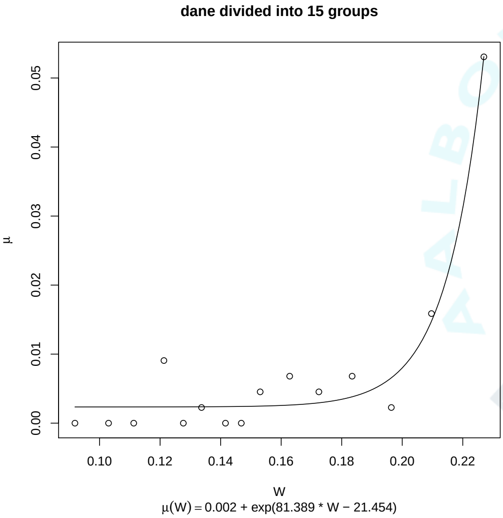
$$p\left(f_i \mid N_i\right) \propto p\left(N_i \mid f_i\right) p\left(f_i\right)$$

Estimating u_i and v_i

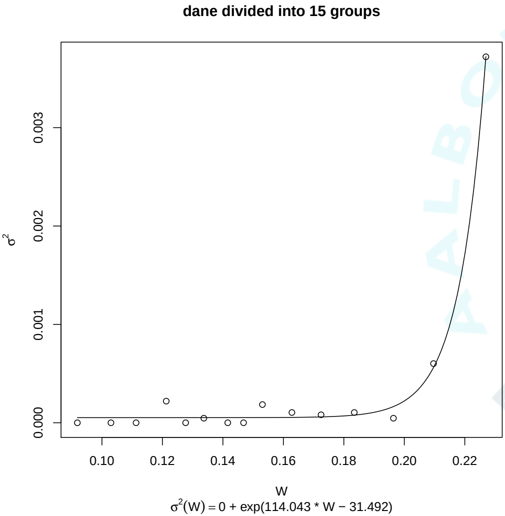
1. Calculate $W_i = \frac{1}{N-N_i} \sum_{i \neq j} \frac{N_j}{d_{ij}}$ for $i = 1, 2, \ldots, M$, where d_{ij} denotes the Manhattan distance/L^1 norm
2. Order the W_i's by size and divide into 15 (?) groups and calculate the mean and variance of the f_i's in each group
3. Fit regression models $\mu(W) = \beta_1 + \exp(\beta_2 W + \beta_3)$ and $\sigma^2(W) = \beta_4 + \exp(\beta_5 W + \beta_6)$ based on the 15 estimates
4. Calculate $\mu_i = \mu(W_i)$ and $\sigma_i^2 = \sigma^2(W_i)$ and use these to calculate the prior parameters u_i and v_i
5. Apply the Bayesian approach to obtain the posterier distribution, e.g. to estimate f_i using the posterior mean

Only `dane` could fit the full models – and the fit is not too comforting – the others resulted in $\mu_i < 0$ or $\sigma_i^2 < 0$ for some i's

Plot of the regression



Plot of the regression



Modified regression models

- Set $\beta_1 = \beta_4 = 0$ in the regression models yielding

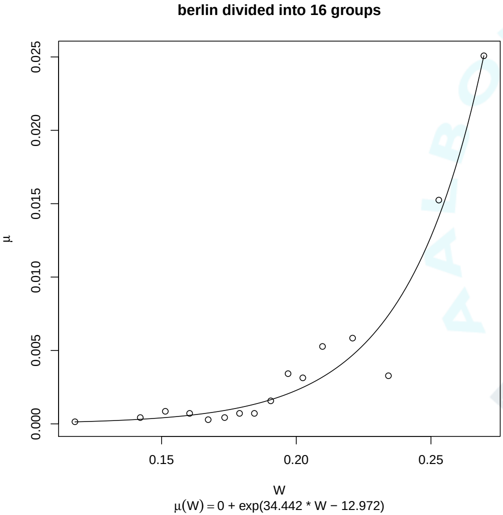
$$\mu(W) = \exp\left(\beta_2 W + \beta_3\right)$$

og

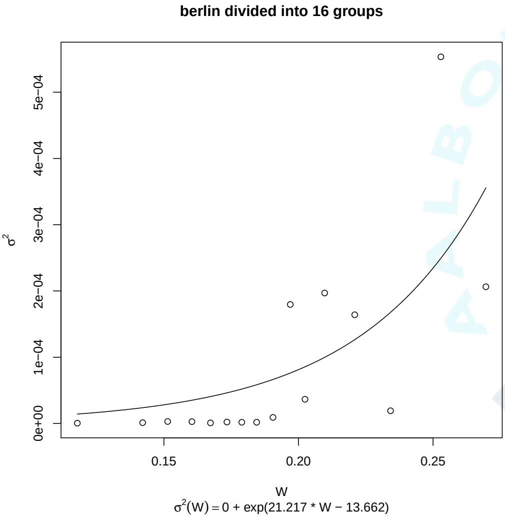
$$\sigma^2(W) = \exp\left(\beta_5 W + \beta_6\right)$$

- Now `berlin` makes the best fits, which seems quite reasonable for $\mu(W)$, but more doubtful for $\sigma^2(W)$
- `dane` fits almost as before

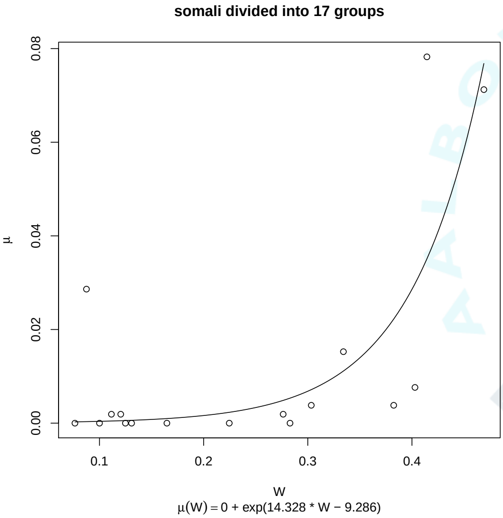
Plot of the regression



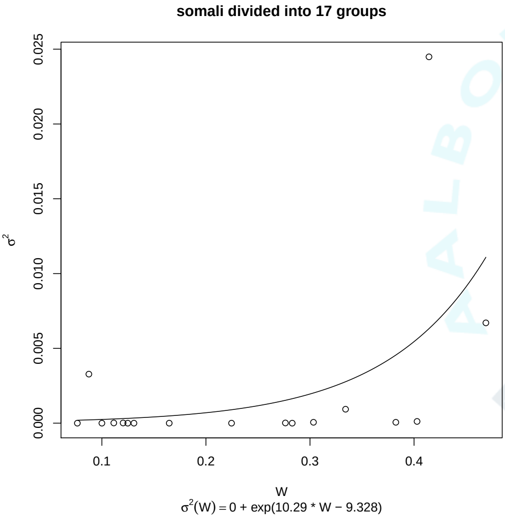
Plot of the regression



Plot of the regression



Plot of the regression



Changes made on http://www.yhrd.org

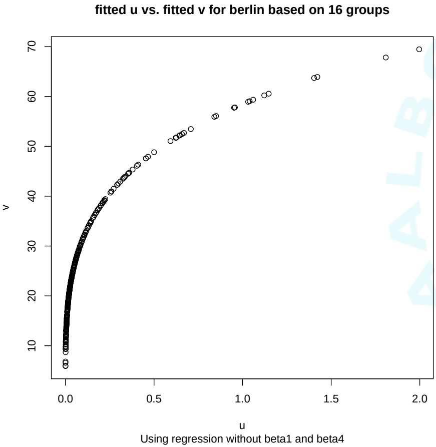
At the 7th International Y Chromosome User Workshop in Berlin, 2010, Sascha Willuweit (one of the persons behind `http://www.yhrd.org`) mentioned a couple of changes between their implementation at `http://www.yhrd.org` and the original article:

- Using the reduced regression models, i.e. without intercepts β_1 and β_4
- The number of groups are determined by fitting several regressions and choosing the best one (details for selecting the minimum number of groups was not mentioned)

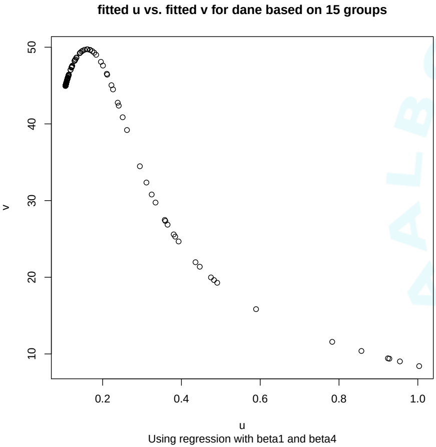
Comments and proposals for changes

- Not a statistical model, more an ad-hoc method
- $\mu_i = \mu(W_i) = \exp(aW_i + b)$ is not bounded above such that $\mu(W) \geq 1$ for $W \geq -\frac{b}{a}$: for `berlin` $a = 34.44$ and $b = -12.97$ so $\mu_i \geq 1$ for $W_i \geq 0.377$ ($0 \leq W_i \leq 1$ and $0 < \mu_i < 1$ by definition)
- Fitting u_i and v_i: only W_i to fit *two* exponential regression models
- The model is not consistent: generalisation to a Dirichlet prior and a multinomial likelihood might solve this

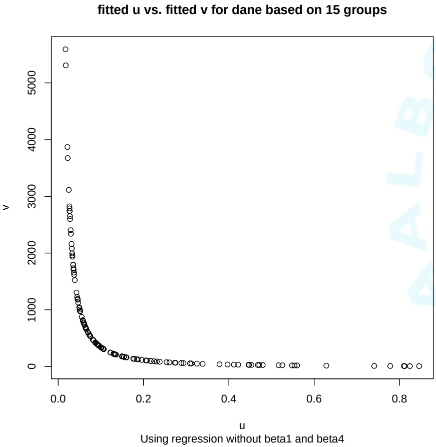
Fitting two parameters (u_i and v_i) using only one (W_i)



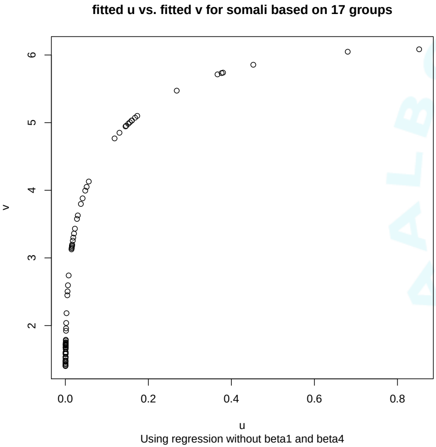
Fitting two parameters (u_i and v_i) using only one (W_i)



Fitting two parameters (u_i and v_i) using only one (W_i)



Fitting two parameters (u_i and v_i) using only one (W_i)



Idea: use second moment

- $W_i = \frac{1}{N-N_i} \sum_{i \neq j} \frac{N_j}{d_{ij}} = \frac{1}{N-N_i} \sum_{i \neq j} \frac{N_j}{\sum_{k=1}^r d_{ijk}}$ uses first moment of the allele differences on each locus of r loci
- Maybe also use the second moment to introduce

$$Z_i = \frac{1}{N - N_i} \sum_{i \neq j} \frac{N_j}{\sum_{k=1}^r \left(d_{ijk} - \frac{d_{ij}}{r} \right)^2}$$

- Fit μ_i's and σ_i^2's by multiple regression using a grid of W_i and Z_i values
- 15 groups only correspond to a 4×4-grid, which is way too coarse – requiring 15 groups of W_i's and Z_i's, the grid would have size $15 \cdot 15 = 225$: requires a lot of observations!
- Too few observations in `berlin`, `dane`, and `somali`, but it would be interesting to see how it would perform compared to just using the W_i's

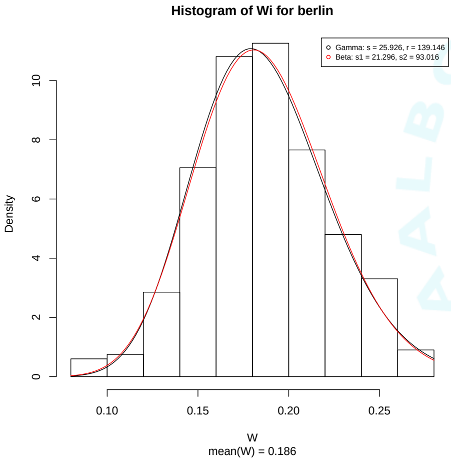
Generalised frequency surveying

- Assume a priori that $\boldsymbol{f} \sim \text{Dirichlet}\,(\boldsymbol{\alpha})$ where $\boldsymbol{f} = (f_1, f_2, \dots, f_K)$ is the vector of frequencies for all possible haplotypes
- Use the likelihood $\boldsymbol{N}|\boldsymbol{f} \sim \text{Multinomial}\,(N_+, \boldsymbol{f})$
- The posterior becomes $\boldsymbol{f}|\boldsymbol{N} \sim \text{Dirichlet}\,(\alpha_1 + N_1, \dots, \alpha_K + N_K) = \text{Dirichlet}\,(\alpha_1 + N_1, \dots, \alpha_n + N_n, \alpha_{n+1}, \dots, \alpha_K)$ where $f_1, f_2, \dots, f_n$ are the frequencies for the observed haplotypes and $f_{n+1}, f_{n+2}, \dots, f_K$ are for the unobserved haplotypes

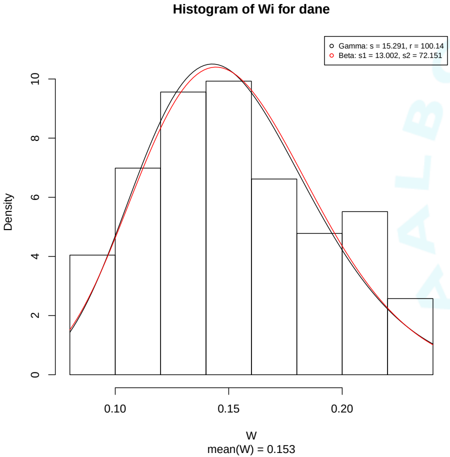
Marginal distribution

- Let $\alpha_+ = \sum_{i=1}^{K} \alpha_i$
- The marginal posterior distribution for the i'th haplotype is $f_i | N_i \sim \text{Beta}\left(\alpha_i + N_i, \sum_{j=1}^{K} (\alpha_j + N_j) - (\alpha_i + N_i)\right) = \text{Beta}\left(\alpha_i + N_i, \alpha_+ - \alpha_i + N_+ - N_i\right)$
- $\mathbf{E}\left[f_i | N_i\right] = \frac{\alpha_i + N_i}{\sum_{j=1}^{K}\left(\alpha_j + N_j\right) - (\alpha_i + N_i) + (\alpha_i + N_i)} = \frac{\alpha_i + N_i}{\alpha_+ + N_+}$
- $\sum_{i=1}^{K} \mathbf{E}\left[f_i | N_i\right] = (\alpha_+ + N_+)^{-1} \sum_{i=1}^{K} (\alpha_i + N_i) = 1$
- Incorporate prior knowledge can be done by specifying the prior parameter α_i for all possible haplotypes, but with this approach α_+ might be problematic to calculate for large K

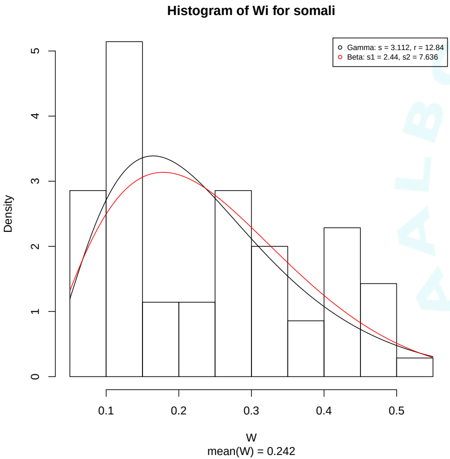
Approximating α_+



Approximating α_+



Approximating α_+



Approximating α_+

- As earlier stated, $0 \leq W_i \leq 1$ so the Beta distribution is the right choice theoretically
- Assume that $\alpha_i = h(W_i)$ for some function h such that $\alpha_+ = \sum_{i=1}^{K} h(W_i)$
- Denote by f_β the density of a fitted Beta-distribution, then

$$\alpha_+ \approx K \int_0^1 f_\beta(W)\, h(W) dW$$

Approximating α_+

- To get equality between the first prior parameter in surveying and the generalised surveying, let $\alpha_i = u_i = \frac{\mu_i^2(1-\mu_i)}{\sigma_i^2}$
- Because $\mu_i = \mu(W_i) = \exp(aW_i + b)$ can result in $\mu_i \geq 1$ then $1 - \mu_i \leq 0$ such that $\alpha_i \leq 0$ which is now allowed
- For `berlin`, $\mu_i \geq 1$ for $W_i \geq 0.377$ so that all contributions to the integral in the α_+ approximation is negative for $W_i \geq 0.377$
- `berlin`: $\alpha_+ = 50205.04$ and the uncovered probability mass is estimated to 0.986
- `dane`: $-b/a = 0.271$ and $\alpha_+ = -7897176$
- `somali`: $-b/a = 0.648$ and $\alpha_+ = -1535169$
- $\alpha_i = u_i$ and the exponential regression is unusable, but the distribution of the W_i's might be helpful for other choices

Problem

- Maybe getting too much attention because it is the only published method for estimating haplotype frequencies
- At the 7th International Y Chromosome User Workshop in Berlin, 2010, Michael Krawczak (one of the authors of the original articles) gave a talk where the associated slides included the statement ”[frequency surveying has] never [been] thoroughly studied and validated“

Ancestral awareness

The idea: approximate when identified a common (partial) ancestor

The basic idea is to find $I = \{i_1, i_2, \ldots, i_q\} \subseteq \{1, 2, \ldots, r\}$ such that

$$P\left(\bigcap_{j \notin I} L_j = a_j \;\middle|\; \bigcap_{i \in I} L_i = a_i \right) \approx \prod_{j \notin I} P\left(L_j = a_j \;\middle|\; \bigcap_{i \in I} L_i = a_i \right)$$

is a good approximation.

Example

- Assume that we have loci L_1, L_2, L_3, L_4 and $I = \{1, 2\}$
- Then

$$P\left(L_1, L_2, L_3, L_4\right) = P\left(L_1\right) P\left(L_2 | L_1\right) P\left(L_3, L_4 | L_1, L_2\right) \quad (1)$$

$$\approx P\left(L_1\right) P\left(L_2 | L_1\right) \prod_{j=3}^{4} P\left(L_j | L_1, L_2\right) \quad (2)$$

How to chose I

- I is called an ancestral set, because it can be interpreted as a set of alleles that is common with one's ancestors
- I can be found using a greedy approach adding the j to I that maximises $P\left(L_j = a_j \mid \cap_{i \in I} L_i = a_i\right)$
- Stop adding elements to I, e.g. when only a percentage of the observations is left to use for calculating the marginal probabilities conditional on I

Drawbacks

- The approach is simple, but like it is not a statistical model

Classification models

Classification models

The idea: alternately perceive different loci as a response variable

Let $L_1, L_2, \ldots, L_r$ be the r different loci available in the haplotype. Then fit

$$L_{i_1} \sim \sum_{k \notin \{i_1\}} L_k \tag{3}$$

$$L_{i_2} \sim \sum_{k \notin \{i_1, i_2\}} L_k \tag{4}$$

$$\vdots \tag{5}$$

$$L_{i_{r-2}} \sim \sum_{k \notin \{i_1, i_2, \ldots, i_{r-2}\}} L_k \tag{6}$$

$$L_{i_{r-1}} \sim \sum_{k \notin \{i_1, i_2, \ldots, i_{r-1}\}} L_k = L_{i_r} \tag{7}$$

and use the empirical distribution for L_{i_r}.

Properties

- A class of statistical models
- The classifications can be done with some of the several available classifications methods such as classification trees, ordered logistic regression, or support vector machines
- Selection of i_j should be done using standard model selection criteria depending on the classification model used
- Does not incorporate prior knowledge

Kernel smoothing and model based clustering

The idea: create a density around each haplotype

- Put a scaled density/mass (called a kernel) around each haplotype with mass equal to its relative frequency $\frac{N_i}{N_+}$
- In this way unobserved haplotypes get probability mass from the (near) neighbours

Choice of kernel

- A straightforward approach is the Gaussian kernel $K\left(\boldsymbol{z}|\boldsymbol{x}_i,\lambda\right) = \left(2\pi\lambda^2\right)^{-\frac{r}{2}}\det\left(\Sigma\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\lambda^2}(\boldsymbol{x}_i-\boldsymbol{z})\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{z})^\top\right)$ where λ is called a smoothing parameter that has to be chosen
- A frequency estimate for any given haplotype $\boldsymbol{z}$ is $g(\boldsymbol{z}) = \frac{1}{N_+}\sum_{i=1}^{n} N_i K\left(\boldsymbol{z}|\boldsymbol{x}_i,\lambda\right)$
- To incorporate prior knowledge, $K\left(\boldsymbol{z}|\boldsymbol{x}_i,N_i,\lambda\right) = \left(2\pi\frac{\lambda^2}{N_i}\right)^{-\frac{r}{2}}\det\left(\Sigma\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\frac{\lambda^2}{N_i}}(\boldsymbol{x}_i-\boldsymbol{z})\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{z})^\top\right)$ could be used

Problem

- The model can be inaccurate if the kernel has small variance, because then the actual mass when evaluated over the discrete grid can differ greatly from the relative frequencies
- Discrete kernels could be tried instead, e.g. the multinomial

Model based clustering

- Estimating a frequency for a haplotype using kernel smoothing require evaluating as many densities as the number of haplotypes in the database
- Model based clustering can be used to perform clustering first to minimise the required number of density evaluations
- Same problem as with kernel smoothing if the variances are too small

Comparing models

Comparing models

- Model verification is as always crucial
- One important feature of a model is to be able to efficiently obtain further samples of haplotypes according to their probability under a model

Different ways of comparing models

- Estimated unobserved probability mass
- Marginal deviances (for a model's single and pairwise compared to observed)
- *Several more should be definitely considered*

Unobserved probability mass: point estimate

- K_0: the number of singletons
- N: the number of observations
- In 1968, Robbins showed that

$$V = \frac{K_0}{N+1}$$

is an unbiased estimate of the unobserved probability mass.

Unobserved probability mass: limiting consistent variance estimate

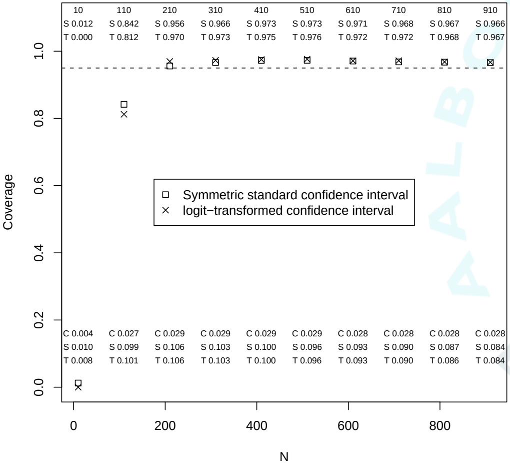
- K_1: the number of doubletons
- In 1986, Bickel and Yahav showed that under some regularity conditions,

$$\hat{\sigma}^2 = \frac{K_0}{N^2} - \frac{(K_0 - 2K_1)^2}{N^3}$$

  is limiting consistent estimate of the variance of the unobserved probability mass
- Both V and the variance estimate can be verified by simulation

Unobserved probability mass: simulation study

Confidence interval coverage for true population size Q = 10000



Based on 10000 simulations and 10−90 probabilities. Line at 0.95.

Unobserved probability mass

- The estimate seems like a good and simply way of perform model verification, but it cannot stand alone as we shall soon see
- It can also be used to fit model parameters, e.g. the smoothing parameter in the kernel smoothing model

Unobserved probability mass: approximate confidence intervals

	berlin	*dane*	*somali*
V	0.364	0.602	0.277
$\frac{\hat{\sigma}}{V}$	0.061	0.078	0.120
95% conf. int.	$[0.321; 0.408]$	$[0.510; 0.694]$	$[0.212; 0.342]$
S: $\beta_1/\beta_4 \neq 0$	NA	0.643	NA
S: $\beta_1/\beta_4 = 0$	0.479	0.703	0.335
`rpart`	0.478	0.71	0.42
`svm`	0.526	0.792	0.43
`polr`	0.639	0.886	NA
Ancestor: 10%	0.39	0.589	0.182
Ancestor: 15%	0.454	0.668	0.22
Ancestor: 20%	0.466	0.713	0.246

Marginal deviances

- Depending on the model, exact marginals can be difficult to obtain
- If haplotypes can be sampled according to their probability under a model, then marginals can be approximated by simulating a huge number of haplotypes under that model
- Using only the observed marginals should correspond to this, but – at least for small databases – this is not the case according to simulations studies performed with the classification models

Deviance for pairwise marginals

- Let $\{u\}_{ij}$ be the two-way table with the observation counts, $\{\tilde{p}\}_{ij}$ the table of predicted probabilities under a model $\mathcal{M}_0$, $\{\hat{p}\}_{ij}$ the relative probabilities such that $\hat{p}_{ij} = \frac{u_{ij}}{u_{++}}$
- For the pairwise marginal tables, the deviance is $d = -2 \log \left(\frac{L\left(\{\tilde{p}\}_{ij}\right)}{L\left(\{\hat{p}\}_{ij}\right)} \right)$ where $L\left(\{p\}_{ij}\right) = \prod_{i,j} p_{ij}^{u_{ij}}$ is proportional to the likelihood (the constant $\frac{u_{++}!}{\prod_{i,j} u_{ij}}$ is cancelled out in the fraction)
- Then $d = -2 \sum_{i,j} u_{ij} \log \left(\frac{\tilde{p}_{ij}}{\hat{p}_{ij}} \right) \sim \chi^2_\nu$

Comparing models

- A deviance is calculated for each pair of loci
- To compare models these deviances can summed and used for relative comparisons (the sum is not χ^2 distributed)
- The deviance is calculated similar for single marginals

Deviance sums for single marginals

	berlin	*dane*	*somali*
rpart	1.236	1.449	0.268
svm	13434.632	2363.861	5664.264
polr	3.114	4.623	NA

Sum of deviances for observed single marginals vs. simulated single marginals for the classification method specified in each row.

Deviance sums for pairwise marginals

	berlin	*dane*	*somali*
rpart	Inf	768.444	772.936
svm	Inf	24971.264	53797.852
polr	1761.153	Inf	NA

Sum of deviances for observed pairwise marginals vs. simulated pairwise marginals for the classification method specified in each row.

Calculating evidence

Evidence: motivation of usage

- The purpose is to get an unbiased opinion in a trial
- Often formulated as the hypothesis

$$\begin{aligned} H_p &: \text{The suspect left the crime stain} \\ &\quad \text{together with } n \text{ additional contributors.} \\ H_d &: \text{Some other person left the crime stain} \\ &\quad \text{together with } n \text{ additional contributors.} \end{aligned}$$

- Then the likelihood ratio given by $LR = \frac{P(E|H_p)}{P(E|H_d)}$ is calculated
- In a courtroom it can then be stated that the evidence E is LR times more likely to have arisen under H_p than under H_d (formulation is from "Interpreting DNA Mixtures“ by Weir *et al.*, 1997)

Computational challenge

- The actual evaluation of LR can be a computation heavy task: the number of combinations giving rise to the same trace grows with the number of contributors

Two contributors

Assume

$$
\begin{aligned}
H_p &: \text{The suspect left the crime stain} \\
& \text{together with one additional contributors.} \\
H_d &: \text{Some other person left the crime stain} \\
& \text{together with one additional contributors.}
\end{aligned}
$$

Two contributors: notation

Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ and define

$$T = \boldsymbol{a} \oplus \boldsymbol{b} = (\{a_1, b_1\}, \{a_2, b_2\}, \ldots, \{a_1, b_n\}) \tag{8}$$

$$T \ominus \boldsymbol{a} = (\{b_1\}, \{b_2\}, \ldots, \{b_n\}) \tag{9}$$

where the sets are multisets.

Two contributors

Let T be the trace, $\boldsymbol{h}_s$ the suspect's haplotype, and $\boldsymbol{h}_1$ the additional contributor's haplotype. Then $T = \boldsymbol{h}_s \oplus \boldsymbol{h}_1$ and $\boldsymbol{h}_1 = T \ominus \boldsymbol{h}_s$. Say that $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ is consistent with the trace T if $\boldsymbol{h}_1 \oplus \boldsymbol{h}_2 = T$, which is denoted $(\boldsymbol{h}_1, \boldsymbol{h}_2) \equiv T$. This makes

$$
\begin{aligned}
LR &= \frac{P\left(E|H_p\right)}{P\left(E|H_d\right)} && (10)\\
&= \frac{P\left(\boldsymbol{h}_s, T \ominus \boldsymbol{h}_s\right)}{\sum_{(\boldsymbol{h}_1,\boldsymbol{h}_2)\equiv T} P\left(\boldsymbol{h}_s, \boldsymbol{h}_1, \boldsymbol{h}_2\right)} && (11)\\
&= \frac{P\left(\boldsymbol{h}_s\right) P\left(T \ominus \boldsymbol{h}_s\right)}{P\left(\boldsymbol{h}_s\right) \sum_{(\boldsymbol{h}_1,\boldsymbol{h}_2)\equiv T} P\left(\boldsymbol{h}_1\right) P\left(\boldsymbol{h}_2\right)} && (12)\\
&= \frac{P\left(T \ominus \boldsymbol{h}_s\right)}{\sum_{(\boldsymbol{h}_1,\boldsymbol{h}_2)\equiv T} P\left(\boldsymbol{h}_1\right) P\left(\boldsymbol{h}_2\right)} && (13)
\end{aligned}
$$

by assuming that haplotypes are independent.

Two contributors: number of terms in the denominator

- Let $T = (T_1, T_2, \dots, T_r)$, T_i is a set of alleles such that $|T_i| \in \{1, 2\}$
- Let $\mathcal{H}_T = T_1 \times T_2 \times \cdots \times T_r$
- In the non-trivial case a $j \in \{1, 2, \dots, r\}$ exists such that $T_j = \{a_1, a_2\}$ with $a_1 \neq a_2$
- Let $T'_j = \{a_1\}$ (such that one of the alleles is removed) and

$$\mathcal{H}'_T = T_1 \times \cdots \times T_{j-1} \times T'_j \times T_{j+1} \times \cdots \times T_r$$

Two contributors: number of terms in the denominator

- Now the denominator of LR can be written as $2 \sum_{\boldsymbol{h}_1 \in \mathcal{H}'_T} P(\boldsymbol{h}_1) P(T \ominus \boldsymbol{h}_1)$
- If k denotes the number of loci in the trace with only one allele, and we assume that we have the non-trivial case with $0 \leq k < r$, we have that $|\mathcal{H}_T| = \prod_{i=1}^{r} |T_i| = 2^{r-k}$ such that $|\mathcal{H}'_T| = \frac{|\mathcal{H}_T|}{2} = 2^{r-k-1} \leq 2^{r-1}$
- This means that for r loci, a maximum of $2 \cdot 2^{r-1} = 2^r$ haplotype frequencies have to be calculated, e.g. $2^{10} = 1024$
- If a trace has two contributors with no known suspects, the two most likely contributors can be chosen to be $\arg\max_{\boldsymbol{h}_1 \in \mathcal{H}'_T} P(\boldsymbol{h}_1) P(T \ominus \boldsymbol{h}_1)$

n contributors: formulation of the LR for one locus

- LR defined generally in ”Forensic interpretation of Y-chromosomal DNA mixtures“ by Wolf *et al.*, 2005
- At a given locus, let E_t, E_s, and E_k be the set of alleles from the trace, the suspect, and the known contributors, respectively
- Assume n unknown contributors and let A_n denote the set of alleles carried by the these n unknown contributors
- Let $P_n(V; W) = P(W \subseteq A_n \subseteq V)$
- Then

$$LR = \frac{P_n(E_t; E_t \setminus (E_s \cup E_k))}{P_{n+1}(E_t; E_t \setminus E_k)} \tag{14}$$

$$= \frac{P(E_t \setminus (E_s \cup E_k) \subseteq A_n \subseteq E_t)}{P(E_t \setminus E_k \subseteq A_{n+1} \subseteq E_t)} \tag{15}$$

n contributors: formulation of the LR for m loci

For m loci we have

$$LR = \frac{P_n\left(\bigcap_{i=1}^{m}\{E_{t,i}; E_{t,i} \setminus \left(E_{s,i} \cup E_{k,i}\right)\}\right)}{P_{n+1}\left(\bigcap_{i=1}^{m}\{E_{t,i}; E_{t,i} \setminus E_{k,i}\}\right)}$$

n contributors: example from thesis, page 18

Let $E_{t,1} = \{1,2\}$, $E_{t,2} = \{2\}$, $E_{s,1} = \{1\}$, $E_{s,2} = \{2\}$, $E_{k,1} = E_{k,2} = \varnothing$. Then

$$\begin{aligned}
LR &= \frac{P_1\left(\bigcap_{i=1}^{2}\{E_{t,i}; E_{t,i} \setminus (E_{s,i} \cup E_{k,i})\}\right)}{P_2\left(\bigcap_{i=1}^{2}\{E_{t,i}; E_{t,i} \setminus E_{k,i}\}\right)} \\
&= \frac{P\left(\bigcap_{i=1}^{2}\{E_{t,i} \setminus (E_{s,i} \cup E_{k,i}) \subseteq A_1^i \subseteq E_{t,i}\right)}{P\left(\bigcap_{i=1}^{2}\{E_{t,i} \setminus E_{k,i} \subseteq A_2^i \subseteq E_{t,i}\}\right)} \\
&= \frac{P\left(\{E_{t,1} \setminus E_{s,1} \subseteq A_1^1 \subseteq E_{t,1}\} \cap \{E_{t,2} \setminus E_{s,2} \subseteq A_1^2 \subseteq E_{t,2}\right)}{P\left(\{E_{t,1} \setminus E_{k,1} \subseteq A_2^1 \subseteq E_{t,1}\} \cap \{E_{t,2} \setminus E_{k,2} \subseteq A_2^2 \subseteq E_{t,2}\}\right)} \\
&= \frac{P\left(\{\{2\}^1 \subseteq A_1^1 \subseteq \{1,2\}^1\} \cap \{A_1^2 \subseteq \{2\}^2\}\right)}{P\left(\{\{1,2\}^1 \subseteq A_2^1 \subseteq \{1,2\}^1\} \cap \{\{2\}^2 \subseteq A_2^2 \subseteq \{2\}^2\}\right)} \\
&= \frac{P\left(\{2\}^1 \cap \{2\}^2\right)}{P\left(\{1,2\}^1 \cap \{2,2\}^2\right)} \\
&= \frac{P\left(\boldsymbol{h}_1 = (2,2)\right)}{P\left(\boldsymbol{h}_1 = (1,2), \boldsymbol{h}_2 = (2,2)\right) + P\left(\boldsymbol{h}_1 = (2,2), \boldsymbol{h}_2 = (1,2)\right)}
\end{aligned}$$

n contributors: challenges

- It is complicated
- One key ingredient in calculating the *LR* is to be able to estimate frequencies for unobserved haplotypes
- The next step is to be able to calculate the *LR* efficiently even for a large number of contributors

Further work

Further work

- Estimating Y-STR haplotype frequencies
  - Better incorporation of prior knowledge in a statistical model, e.g. graphical models with other test statistics (ordinal data and incorporating prior knowledge such as the single step mutation model)
  - More and better ways to verify models
  - Larger datasets (`http://www.yhrd.org` has gathered a lot of data, both publicly available in journals and directly from laboratories, but none is available for others, yet)
- Y-STR Mixtures
  - Efficient calculation of *LR*
  - Use quantitative information (the amount of DNA material which can be seen in the EPG) instead of only the qualitative
- Model the signal in the EPG (electropherogram)

Questions?

Questions?