Advances in Visualizing Categorical Data
Using the vcd, gnm and vcdExtra Packages in R

Michael Friendly\(^1\)  Heather Turner\(^2\)  David Firth\(^2\)
Achim Zeileis\(^3\)

\(^1\)Psychology Department
York University

\(^2\)University of Warwick, UK

\(^3\)Department of Statistics
Universität Innsbruck

CARME 2011
Rennes, February 9–11, 2011
Slides: http://datavis.ca/papers/adv-vcd-4up.pdf
Co-conspirators

Heather Turner
University of Warwick

David Firth
University of Warwick

Achim Zeileis
Universität Innsbruck
Outline

1. Introduction
2. Generalized Mosaic Displays: vcd Package
4. 3D Mosaics: vcdExtra Package
5. Models and Visualization for Log Odds Ratios
Outline

1 Introduction
   • Brief History of VCD
   • Visual overview

2 Generalized Mosaic Displays: vcd Package
   • Extending mosaic-like displays
   • The strucplot framework

3 Generalized Nonlinear Models: gnm & vcdExtra Packages
   • Loglinear models and generalized linear models
   • Generalized nonlinear models: gnm package
   • Models for ordered categories

4 3D Mosaics: vcdExtra Package

5 Models and Visualization for Log Odds Ratios
   • Log odds ratios
   • Examples
Brief History of VCD

Hartigan and Kleiner (1981, 1984): representing an \( n \)-way contingency table by a “mosaic display,” showing a (recursive) decomposition of frequencies by “tiles”, area \( \sim \) cell frequency.

e.g., a 4-way table of viewing TV programs

Freq \( \sim \) Day + Week + Time + Network
Brief History of VCD

- Friendly (1994): developed the connection between mosaic displays and loglinear models
  - Showed how mosaic displays could be used to visualize both observed frequency (area) and residuals (shading) from some model.
  - 1\textsuperscript{st} presented at CARME 1995 (thx: Michael & Jörg!)
Brief History of VCD

- **Visualizing Categorical Data** (Friendly, 2000)
- But: mosaic-like displays have a long history (Friendly, 2002)!

von Mayr (1877)  Birch (1964)

2002: vcd project at TU & WU, Vienna (Kurt Hornik, David Meyer, Achim Zeileis) \(\rightarrow\) **vcd** package
Brief History of VCD

- **Visualizing Categorical Data** (Friendly, 2000)
- But: mosaic-like displays have a long history (Friendly, 2002)!

**von Mayr (1877)**

2002: vcd project at TU & WU, Vienna (Kurt Hornik, David Meyer, Achim Zeileis) \(\rightarrow\) vcd package

**Birch (1964)**
Brief History of VCD

- **Visualizing Categorical Data** (Friendly, 2000)
- But: mosaic-like displays have a long history (Friendly, 2002)!

  - von Mayr (1877)
  
  - Birch (1964)

- 2002: vcd project at TU & WU, Vienna (Kurt Hornik, David Meyer, Achim Zeileis) ➔ vcd package
Outline

1. **Introduction**
   - Brief History of VCD
   - Visual overview

2. **Generalized Mosaic Displays: vcd Package**
   - Extending mosaic-like displays
   - The strucplot framework

3. **Generalized Nonlinear Models: gnm & vcdExtra Packages**
   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. **3D Mosaics: vcdExtra Package**

5. **Models and Visualization for Log Odds Ratios**
   - Log odds ratios
   - Examples
Data, pictures, models & stories

Two paths to enlightenment

data

story
Data, pictures, models & stories

Two paths to enlightenment

Exploratory
Data, pictures, models & stories

Two paths to enlightenment

Exploratory

Model-based
Two paths to enlightenment

Data, pictures, models & stories

Exploratory

Model-based

Data → Visualization

Model → Summary

Story
Data, pictures, models & stories

Two paths to enlightenment

Exploratory vs. Model-based

Data → Visualization → Model → Summary → Inference

- Exploratory
- Model-based
Visual overview: Models for frequency tables

Generalized nonlinear models
\[ \text{gnm}(F \sim A+B+\text{Mult}(A,B), \text{family}=\text{poisson}) \]

Generalized linear models
\[ \text{glm}(F \sim A+B, \text{family}=\text{poisson}) \]

Loglinear models
\[ \text{loglm}(-A+B) \]

- Related models: logistic regression, polytomous regression, log odds models, ...
- Goals: Connect all with visualization methods
Visual overview: R packages

- vcd
  - visualization via strucplot framework
- gnm
  - generalized linear & nonlinear models
- vcdExtra
  - mosaic.glm()
  - mosaic3d()
  - glmlist() methods
  - LOR models
Outline

1 Introduction
   - Brief History of VCD
   - Visual overview

2 Generalized Mosaic Displays: vcd Package
   - Extending mosaic-like displays
   - The strucplot framework

3 Generalized Nonlinear Models: gnm & vcdExtra Packages
   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4 3D Mosaics: vcdExtra Package

5 Models and Visualization for Log Odds Ratios
   - Log odds ratios
   - Examples
Extending mosaic displays

Initial ideas for mosaic displays were extended in a variety of ways:

- pairs plots and trellis-like layouts for marginal, conditional and partial views (Friendly 1999).
- varying the shape attributes of bar plots and mosaic displays
  - double-decker plots (Hofmann 2001),
  - spine plots and spinograms (Hofmann & Theus 2005)
- residual-based shadings to emphasize pattern of association in log-linear models or to visualize significance (Zeileis et al., 2007).
- dynamic interactive versions (ViSta, MANET, Mondrian):
  - linking of several graphs and models
  - selection and highlighting across graphs and models
  - interactive modification of the visualized models
Generalized mosaic displays

`vcd` package and the `strucplot` framework

- Various displays for \( n \)-way frequency tables
  - flat (two-way) tables of frequencies
  - fourfold displays
  - mosaic displays
  - sieve diagrams
  - association plots
  - doubledecker plots
  - spine plots and spinograms

- Commonalities
  - All have to deal with representing \( n \)-way tables in 2D
  - All graphical methods use area to represent frequency
  - Some are model-based — designed as a visual representation of an underlying statistical model
  - Graphical methods use visual attributes (color, shading, etc.) to highlight relevant statistical aspects
Generalized mosaic displays

vcd package and the strucplot framework

- Various displays for $n$-way frequency tables
  - flat (two-way) tables of frequencies
  - fourfold displays
  - mosaic displays
  - sieve diagrams
  - association plots
  - doubledecker plots
  - spine plots and spinograms

- Commonalities
  - All have to deal with representing $n$-way tables in 2D
  - All graphical methods use area to represent frequency
  - Some are model-based — designed as a visual representation of an underlying statistical model
  - Graphical methods use visual attributes (color, shading, etc.) to highlight relevant statistical aspects
Familiar example: UCB Admissions

Data on admission to graduate programs at UC Berkeley, by Dept, Gender and Admission

\[ \text{structable(Dept \sim Gender + Admit, UCBAdmissions)} \]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Dept</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Admitted</td>
<td>512</td>
<td>353</td>
<td>120</td>
<td>138</td>
<td>53</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rejected</td>
<td>313</td>
<td>207</td>
<td>205</td>
<td>279</td>
<td>138</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>Admitted</td>
<td>89</td>
<td>17</td>
<td>202</td>
<td>131</td>
<td>94</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rejected</td>
<td>19</td>
<td>8</td>
<td>391</td>
<td>244</td>
<td>299</td>
<td>317</td>
<td></td>
</tr>
</tbody>
</table>

or, as a two-way table (collapsed over Dept),

\[ \text{structable(~Gender + Admit, UCBAdmissions)} \]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Admitted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1198</td>
<td>1493</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>557</td>
<td>1278</td>
<td></td>
</tr>
</tbody>
</table>
Fourfold displays for $2 \times 2$ tables

**General ideas:**

- Model-based graphs can show *both data* and model *tests* (or other statistical features)
- Visual attributes tuned to support *perception* of relevant statistical comparisons

![Graph showing fourfold displays](image)

- **Quarter circles:** radius $\sim \sqrt{n_{ij}} \Rightarrow$ area $\sim$ frequency
- **Independence:** Adjoining quadrants $\approx$ align
- **Odds ratio:** ratio of areas of diagonally opposite cells
- **Confidence rings:** Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining rings overlap
Fourfold displays for $2 \times 2 \times k$ tables

- **Stratified analysis:** one fourfold display for each department
- Each $2 \times 2$ table **standardized** to equate marginal frequencies
- **Shading:** highlight departments for which $H_a : \theta_i \neq 1$
Mosaic displays

- **Tiles:** Area $\sim$ observed frequencies, $n_{ijk}$
- **Friendly shading** (highlight association pattern):
  - Residuals: $r_{ijk} = (n_{ijk} - \hat{m}_{ijk})/\sqrt{\hat{m}_{ijk}}$
  - Color— blue: $r > 0$, red: $r < 0$
  - Saturation: $|r| < 2$ (none), $> 4$ (max), else (middle)

(Other shadings highlight *significance*)

(Other color schemes: HSV, HCL, ...)

---

Model: $\sim$Dept+Gender+Admit

Model: $\sim$(Dept*Gender) + Admit

Model: $\sim$(Admit + Gender) * Dept
Mosaic displays: Fitting & visualizing models

Mutual independence model: \( \text{Dept} \perp \text{Gender} \perp \text{Admit} \)

```r
> berk.mod0 <- loglm(~Dept + Gender + Admit, data = UCB)
> mosaic(berk.mod0, gp = shading_Friendly, ...)
```

Model: \(~\text{Dept}+\text{Gender}+\text{Admit}\)
Joint independence model: Admit $\perp (\text{Gender}, \text{Dept})$

```r
> berk.mod1 <- loglm(~Admit + (Gender * Dept), data = UCB)
> mosaic(berk.mod1, gp = shading_Friendly, ...)
```

Model: $\sim\text{Admit} + (\text{Gender*Dept})$
Conditional independence model: $\text{Admit} \perp \text{Gender} \mid \text{Dept}$

> berk.mod2 <- loglm(~(Admit + Gender) * Dept, data = UCB)
> mosaic(berk.mod2, gp = shading_Friendly, ...)

Model: $\sim(\text{Admit} + \text{Gender}) \ast \text{Dept}$
Double decker plots

- Visualize dependence of one categorical (typically binary) variable on predictors
- Formally: mosaic plots with vertical splits for all predictor dimensions, highlighting response
Outline

1. Introduction
   - Brief History of VCD
   - Visual overview

2. Generalized Mosaic Displays: vcd Package
   - Extending mosaic-like displays
   - The strucplot framework

   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. 3D Mosaics: vcdExtra Package

5. Models and Visualization for Log Odds Ratios
   - Log odds ratios
   - Examples
The strucplot framework

A general, flexible system for visualizing $n$-way frequency tables:

- integrates tabular displays, mosaic displays, association plots, sieve plots, etc. in a common framework.

- $n$-way tables: variables partitioned into row and column variables in a “flat” 2D display using model formulae arguments allow for fitting *any* loglinear model via `loglm()` in the `MASS` package.

- high-level functions for all-pairwise views (`pairs()`), conditional views (`cotabplot()`).

- low-level functions control *all* aspects of labeling, shading, spacing, etc.
The strucplot framework

Components of the strucplot framework:

Level 4: Related
- `pairs()`, `cotabplot()`

Level 3: Convenience
- `mosaic()`, `sieve()`, `assoc()`, `doubledecker()`

Level 2: Coordinating
- `strucplot()`

Level 1: Low-level
- Graphical appearance control ("grapcon") functions / generators for `strucplot()` (Only the generators are shown below)

**Workhorse Functions**
- **Strucplot core**
  - `struc_mosaic()`, `struc_sieve()`, `struc_assoc()`

- **Labeling**
  - `labeling_border()`, `labeling_list()`, `labeling_cells()`

- **Legend**
  - `legend_resbased()`, `legend_fixed()`

**Parameter Functions**
- **Shading**
  - `shading_hsv()`, `shading_hcl()`, `shading_Friendly()`, `shading_max()`

- **Spacing**
  - `spacing_equal()`, `spacing_conditional()`, `spacing_highlighting()`, `spacing_increase()`
Pairwise bivariate plots

- Visualize all 2-way views of different independence models in \( n \)-way tables: \texttt{type=}
  - "pairwise": Burt matrix: bivariate, marginal views
  - "total": pairwise plots for mutual independence
  - "conditional": marginal independence, given all others
  - "joint": joint independence of all pairs from other variables

- Panel functions for upper, lower, diagonal panels
  - upper, lower: mosaic, assoc, sieve, ...
  - diagonal: barplot, text, mosaic, ...

```
Admit
Admitted Rejected
Gender
Male Female
Dept
A B C D E F
```

```
Admitted Rejected
Admit
0 500 1000 1500 2000 2500 3000
Male Female
Gender
0 200 400 600 800 1000
A B C D E F
Dept
0 500 1000 1500 2000 2500 3000
```
Pairwise bivariate plots

```r
> pairs(UCBAAdmissions, shade=TRUE, space=0.2,
+     diag_panel = pairs_diagonal_mosaic(offset_varnames=-3, ...))
```
Outline

1. Introduction
   - Brief History of VCD
   - Visual overview

2. Generalized Mosaic Displays: vcd Package
   - Extending mosaic-like displays
   - The strucplot framework

   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. 3D Mosaics: vcdExtra Package

5. Models and Visualization for Log Odds Ratios
   - Log odds ratios
   - Examples
Loglinear models and generalized linear models

- **Loglinear models**
  - Model fitting in the **vcd** package is based on loglinear models
    \[
    \log(m_{ij}) = \mu + \lambda^A_i + \lambda^B_j \equiv [A][B] \equiv \sim A + B
    \]
    \[
    \log(m_{ij}) = \mu + \lambda^A_i + \lambda^B_j + \lambda^{AB}_{ij} \equiv [AB] \equiv \sim A \ast B
    \]
  - Fit using iterative proportional fitting (**loglm()**) ➔ No standard errors, limited syntax for expressing models

- **Generalized linear models**
  - Link function:
    \[
    E(y \mid x) = g(\mu) = \eta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
    \]
  - Variance function: \( \text{Var}(y \mid x) = f(\mu) \)
  - Loglinear models as special cases with log link, Poisson dist
    \( n \mapsto \text{Var}(y \mid x) = \mu \)
Loglinear models and generalized linear models

- **Loglinear models**
  - Model fitting in the **vcd** package is based on loglinear models
    
    \[
    \log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B \equiv [A][B] \equiv A + B
    \]
    
    \[
    \log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \equiv [AB] \equiv A \ast B
    \]
  
  - Fit using iterative proportional fitting (**loglm()**)
  - \(\mapsto\) No standard errors, limited syntax for expressing models

- **Generalized linear models**
  - Link function:
    
    \[
    E(y \mid x) = g(\mu) = \eta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
    \]
  
  - Variance function: \(\text{Var}(y \mid x) = f(\mu)\)
  - Loglinear models as special cases with log link, Poisson dist
    \(\mapsto\) \(\text{Var}(y \mid x) = \mu\)
Outline

1. **Introduction**
   - Brief History of VCD
   - Visual overview

2. **Generalized Mosaic Displays: vcd Package**
   - Extending mosaic-like displays
   - The strucplot framework

3. **Generalized Nonlinear Models: gnm & vcdExtra Packages**
   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. **3D Mosaics: vcdExtra Package**

5. **Models and Visualization for Log Odds Ratios**
   - Log odds ratios
   - Examples
A generalized non-linear model (GNM) is the same as a GLM, except that we allow

\[ g(\mu) = \eta(x; \beta) \]

where \( \eta(x; \beta) \) is nonlinear in the parameters \( \beta \).

GNMs are very general, combining:

- classical nonlinear models
- standard link and variance functions for GLM families

In the context of models for categorical data, GNMs provide:

- parsimonious models for structured association
- models for multiplicative association (e.g., Goodman’s RC(1) model)
- multiple instances of multiplicative terms (RC(\( m \)) models)
- user-defined functions for custom models
Generalized nonlinear models: \texttt{gnm} package

Some models for structured associations in square tables

- quasi-independence (ignore diagonals)
  \[ \texttt{gnm(Freq} \sim \text{row + col + Diag(row, col), family = poisson)} \]

- symmetry \( \lambda_{ij}^{RC} = \lambda_{ji}^{RC} \)
  \[ \texttt{gnm(Freq} \sim \text{Symm(row, col), family = poisson)} \]

- quasi-symmetry \( = \text{quasi + symmetry} \)
  \[ \texttt{gnm(Freq} \sim \text{row + col + Symm(row, col), family = poisson)} \]

- fully-specified “topological” association patterns
  \[ \texttt{gnm(Freq} \sim \text{row + col + Topo(row, col, spec = RCmatrix), ...)} \]

All of these are actually GLMs, but the \texttt{gnm} package provides convenience functions \texttt{Diag}, \texttt{Symm}, and \texttt{Topo} to facilitate model specification.
Nonlinear models

- Nonlinear terms are specified in model formulae by functions of class "nonlin"

- Basic nonlinear functions: $\text{Exp}()$, $\text{Inv}()$, $\text{Mult}()$

- Nonlinear terms can be nested. e.g. for a UNIDIFF model:

  \[
  \log \mu_{ijk} = \alpha_{ik} + \beta_{jk} + \exp(\gamma_k)\delta_{ij}
  \]

  the exponentiated multiplier is specified as $\text{Mult}(\text{Exp}(C), A:B)$

- Multiple instances. e.g., Goodman’s RC(2) model:

  \[
  \log \mu_{rc} = \alpha_r + \beta_c + \gamma_{r1}\delta_{c1} + \gamma_{r2}\delta_{c2}
  \]

  specified using: $\text{instances}(\text{Mult}(A,B), 2)$

- User-defined functions of class "nonlin" allow further extensions

All of these are fully general, providing residuals, fitted values, etc.
Generalized nonlinear models: **vcdExtra package**

Provides glue, extending the **vcd** package visualization methods for glm and gnm models

- **mosaic.glm()** $\mapsto$ mosaic methods for class "glm" and class "gnm" objects
- **sieve.glm(), assoc.glm()** $\mapsto$ sieve diagrams and association plots
- Generalized residual types:
  - Pearson
  - deviance
  - standard (adjusted) — unit asymptotic variance
- Model lists:
  - **glmlist()** — methods for collecting, summarizing and visualizing a list of related models
  - **Kway()** — generate & fit models of form $\sim (A+B+\ldots)^k$.  


Outline

1. Introduction
   - Brief History of VCD
   - Visual overview

2. Generalized Mosaic Displays: vcd Package
   - Extending mosaic-like displays
   - The strucplot framework

   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. 3D Mosaics: vcdExtra Package

5. Models and Visualization for Log Odds Ratios
   - Log odds ratios
   - Examples
Models for ordered categories

Consider an $R \times C$ table having ordered categories

- In many cases, the $RC$ association may be described more simply by assigning numeric scores to the row & column categories.
- For simplicity, we consider only integer scores, 1, 2, \ldots here
- These models are easily extended to stratified tables

<table>
<thead>
<tr>
<th>R:C model</th>
<th>$\mu_{ij}^{RC}$</th>
<th>df</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform association</td>
<td>$i \times j \times \gamma$</td>
<td>1</td>
<td>$i:j$</td>
</tr>
<tr>
<td>Row effects</td>
<td>$\alpha_i \times j$</td>
<td>$(I - 1)$</td>
<td>R:j</td>
</tr>
<tr>
<td>Col effects</td>
<td>$i \times \beta_j$</td>
<td>$(J - 1)$</td>
<td>i:C</td>
</tr>
<tr>
<td>Row+Col eff</td>
<td>$j \alpha_i + i\beta_j$</td>
<td>$I + J - 3$</td>
<td>R:j + i:C</td>
</tr>
<tr>
<td>RC(1)</td>
<td>$\phi_i \psi_j \times \gamma$</td>
<td>$I + J - 3$</td>
<td>Mult(R, C)</td>
</tr>
<tr>
<td>Unstructured (R:C)</td>
<td>$\mu_{ij}^{RC}$</td>
<td>$(I - 1)(J - 1)$</td>
<td>R:C</td>
</tr>
</tbody>
</table>

```r
> Yama.tab <- xtabs(Freq ~ Father + Son + Country, data = Yamaguchi87)
> structable(Country + Son ~ Father, Yama.tab[, , 1:2])
```

<table>
<thead>
<tr>
<th>Country</th>
<th>US</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UpNM</td>
<td>LoNM</td>
<td>UpM</td>
<td>LoM</td>
<td>Farm</td>
<td></td>
</tr>
<tr>
<td>Father</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UpNM</td>
<td></td>
<td>1275</td>
<td>364</td>
<td>274</td>
<td>272</td>
<td>17</td>
<td>474</td>
</tr>
<tr>
<td>LoNM</td>
<td></td>
<td>1055</td>
<td>597</td>
<td>394</td>
<td>443</td>
<td>31</td>
<td>300</td>
</tr>
<tr>
<td>UpM</td>
<td></td>
<td>1043</td>
<td>587</td>
<td>1045</td>
<td>951</td>
<td>47</td>
<td>438</td>
</tr>
<tr>
<td>LoM</td>
<td></td>
<td>1159</td>
<td>791</td>
<td>1323</td>
<td>2046</td>
<td>52</td>
<td>601</td>
</tr>
<tr>
<td>Farm</td>
<td></td>
<td>666</td>
<td>496</td>
<td>1031</td>
<td>1632</td>
<td>646</td>
<td>76</td>
</tr>
</tbody>
</table>

See: `demo("yamaguchi-xie", package="vcdExtra")`
First thought: try MCA

```r
> library(ca)
> Yama.dft <- expand.dft(Yamaguchi87)
> yama.mjca <- mjca(Yama.dft)
> plot(yama.mjca, what = c("none", "all"))
```

Yamaguchi data: Mobility in US, UK and Japan, MCA

- Dim 1: Farm vs. Other (52.6%)
- Dim 2: Occ. Status (28.0%)

- Dimensions seem to have reasonable interpretations
- 2nd glance: do they?
- How do they relate to theories of social mobility?
- How to understand Country effects?
Models for stratified mobility tables

Baseline models:

- **Perfect mobility**: \( \text{Freq} \sim (R+C) \times L \)
- **Quasi-perfect mobility**: \( \text{Freq} \sim (R+C) \times L + \text{Diag}(R, C) \)

Layer models:

- **Homogeneous**: no layer effects
- **Heterogeneous**: e.g., \( \mu_{i,j,k}^{RCL} = \delta_{i,j}^{RC} \exp(\gamma_k^L) \)

Extended models: Baseline \( \oplus \) Layer model (R:C model)

<table>
<thead>
<tr>
<th>R:C model</th>
<th>Homogeneous</th>
<th>Layer model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row effects</td>
<td>( \sim + R:j )</td>
<td>( \sim + \text{Mult}(R:j, \exp(L)) )</td>
</tr>
<tr>
<td>Col effects</td>
<td>( \sim + i:C )</td>
<td>( \sim + \text{Mult}(i:C, \exp(L)) )</td>
</tr>
<tr>
<td>Row+Col eff</td>
<td>( \sim + R:j + i:C )</td>
<td>( \sim + \text{Mult}(R:j + i:C, \exp(L)) )</td>
</tr>
<tr>
<td>RC(1)</td>
<td>( \sim + \text{Mult}(R, C) )</td>
<td>( \sim + \text{Mult}(R, C, \exp(L)) )</td>
</tr>
<tr>
<td>Full R:C</td>
<td>( \sim + R:C )</td>
<td>( \sim + \text{Mult}(R:C, \exp(L)) )</td>
</tr>
</tbody>
</table>
Yamaguchi data: Baseline models

Minimal, null model asserts Father $\perp$ Son | Country

> yamaNull <- gnm(Freq ~ (Father + Son) * Country, data = Yamaguchi87, + family = poisson)
> mosaic(yamaNull, ~Country + Son + Father, condvars = "Country", ...)

[FC][SC] Null [FS] association (perfect mobility)
Yamaguchi data: Baseline models

But, theory $\mapsto$ ignore diagonal cells

> yamaDiag <- update(yamaNull, ~. + Diag(Father, Son):Country)
> mosaic(yamaDiag, ~Country + Son + Father, condvars = "Country", ...)

[FC][SC] Quasi perfect mobility, +Diag(F,S)
Yamaguchi data: Fit models for homogeneous association

gnm package makes it easy to fit collections of models, with simple update() methods

```r
> Rscore <- as.numeric(Yamaguchi87$Father)
> Cscore <- as.numeric(Yamaguchi87$Son)
> yamaRo <- update(yamaDiag, ~. + Father:Cscore)
> yamaCo <- update(yamaDiag, ~. + Rscore:Son)
> yamaRpCo <- update(yamaDiag, ~. + Father:Cscore + Rscore:Son)
> yamaRCo <- update(yamaDiag, ~. + Mult(Father, Son))
> yamaFIo <- update(yamaDiag, ~. + Father:Son)
```
Yamaguchi data: Models for heterogeneous association

Log-multiplicative (UNIDIFF) models:

```r
> yamaRx <- update(yamaDiag, ~ . + Mult(Father:Cscore, Exp(Country)))
> yamaCx <- update(yamaDiag, ~ . + Mult(Rscore:Son, Exp(Country)))
> yamaRpCx <- update(yamaDiag, ~ . + Mult(Father:Cscore +
+ Rscore:Son, Exp(Country)))
> yamaRCx <- update(yamaDiag, ~ . + Mult(Father,Son, Exp(Country)))
> yamaFIx <- update(yamaDiag, ~ . + Mult(Father:Son, Exp(Country)))
```

GNM model methods:

- Summary methods: `print(model), summary(model), ...`
- Extractor methods: `coef(model), residuals(model), ...`

Visualization:

- Diagnostics: `plot(model)`
- Mosaics, etc: `mosaic(model)`
Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

```r
> models <- glmlist(yamaNull, yamaDiag,
+                    yamaRo, yamaRx, yamaCo, yamaCx, yamaRpCo,
+                    yamaRpCx, yamaRCo, yamaRCx, yamaFIo, yamaFIx)
> summarise(models)

Model Summary:

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>Chisq</th>
<th>Df</th>
<th>Pr(&gt;Chisq)</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>yamaNull</td>
<td>5592</td>
<td>48</td>
<td>0.0000</td>
<td>5496</td>
<td>5099</td>
<td></td>
</tr>
<tr>
<td>yamaDiag</td>
<td>1336</td>
<td>33</td>
<td>0.0000</td>
<td>1270</td>
<td>997</td>
<td></td>
</tr>
<tr>
<td>yamaRo</td>
<td>156</td>
<td>29</td>
<td>0.0000</td>
<td>98</td>
<td>-142</td>
<td></td>
</tr>
<tr>
<td>yamaRx</td>
<td>148</td>
<td>27</td>
<td>0.0000</td>
<td>94</td>
<td>-130</td>
<td></td>
</tr>
<tr>
<td>yamaCo</td>
<td>68</td>
<td>29</td>
<td>0.0001</td>
<td>10</td>
<td>-230</td>
<td></td>
</tr>
<tr>
<td>yamaCx</td>
<td>59</td>
<td>27</td>
<td>0.0004</td>
<td>5</td>
<td>-219</td>
<td></td>
</tr>
<tr>
<td>yamaRpCo</td>
<td>39</td>
<td>26</td>
<td>0.0509</td>
<td>-13</td>
<td>-228</td>
<td></td>
</tr>
<tr>
<td>yamaRpCx</td>
<td>33</td>
<td>24</td>
<td>0.1034</td>
<td>-15</td>
<td>-213</td>
<td></td>
</tr>
<tr>
<td>yamaRCo</td>
<td>38</td>
<td>26</td>
<td>0.0642</td>
<td>-14</td>
<td>-229</td>
<td></td>
</tr>
<tr>
<td>yamaRCx</td>
<td>32</td>
<td>24</td>
<td>0.1240</td>
<td>-16</td>
<td>-214</td>
<td></td>
</tr>
<tr>
<td>yamaFIo</td>
<td>36</td>
<td>22</td>
<td>0.0288</td>
<td>-8</td>
<td>-190</td>
<td></td>
</tr>
<tr>
<td>yamaFIx</td>
<td>31</td>
<td>20</td>
<td>0.0560</td>
<td>-9</td>
<td>-174</td>
<td></td>
</tr>
</tbody>
</table>
```
Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

```r
> BIC <- matrix(summarise(models)$BIC[-(1:2)], 5, 2, byrow = TRUE)
```

- Homogeneous models all preferred by BIC
- (Xie preferred heterogeneous models)
- Little difference among Col, Row+Col and RC(1) models
- $\leftrightarrow$ R:C association $\sim$ Row scores (Father’s status)
Yamaguchi data: Comparing models

`glmlist()` and related methods facilitate model comparison

```r
> AIC <- matrix(summarise(models)$AIC[-(1:2)], 5, 2, byrow = TRUE)
```

- AIC prefers heterogeneous models
- Row+Col and RC(1) model fit best
- `R:C` association $\sim$ Father’s status, not just scores
- Model summary plots provide sensitive comparisons!
3D mosaic displays

- Loglinear models rely on $\log(n_{ijk}) \sim$ linear model
  - $\mapsto n_{ijk} \sim$ multiplicative model
- Mosaic displays rely on (nested) use of Area = Height $\times$ Width to represent frequencies in $n$-way tables
- How to take this to 3D?

Mutual independence: ~Hair+Eye+Sex

Mutual independence: Expected frequencies
3D mosaic displays

- `mosaic3d()` in the `vcdExtra` package
- partition unit cube $\rightarrow$ nested set of 3D tiles, Volume $\sim$ frequency
- uses `rgl` package: interactive, 3D graphs

```
> mosaic3d(HEC)
```

```
> mosaic3d(HEC, type="expected")
```
Outline

1. **Introduction**
   - Brief History of VCD
   - Visual overview

2. **Generalized Mosaic Displays: vcd Package**
   - Extending mosaic-like displays
   - The strucplot framework

3. **Generalized Nonlinear Models: gnm & vcdExtra Packages**
   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. **3D Mosaics: vcdExtra Package**

5. **Models and Visualization for Log Odds Ratios**
   - Log odds ratios
   - Examples
Log odds ratios

In any two-way, $R \times C$ table, all associations can be represented by a set of $(R - 1) \times (C - 1)$ odds ratios,

$$
\theta_{ij} = \frac{n_{ij}/n_{i+1,j}}{n_{i,j+1}/n_{i+1,j+1}} = \frac{n_{ij} \times n_{i+1,j+1}}{n_{i+1,j} \times n_{i,j+1}}
$$

$$
\ln(\theta_{ij}) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \ln \begin{pmatrix} n_{ij} & n_{i+1,j} & n_{i,j+1} & n_{i+1,j+1} \end{pmatrix}^T
$$

\begin{array}{|c|c|}
\hline
& j & j+1 \\
\hline
i & 1 & -1 \\
i+1 & -1 & 1 \\
\hline
\end{array}
In any two-way, $R \times C$ table, all associations can be represented by a set of $(R - 1) \times (C - 1)$ odds ratios,

$$\theta_{ij} = \frac{n_{ij}/n_{i+1,j}}{n_{i,j+1}/n_{i+1,j+1}} = \frac{n_{ij} \times n_{i+1,j+1}}{n_{i+1,j} \times n_{i,j+1}}$$

$$\ln(\theta_{ij}) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \ln \begin{pmatrix} n_{ij} & n_{i+1,j} & n_{i,j+1} & n_{i+1,j+1} \end{pmatrix}^T$$
Log odds ratios

- \( \ln \theta_{ij} \sim N(0, \sigma^2) \), with estimated asymptotic standard error:

\[
\hat{\sigma}(\ln \theta_{ij}) = \left( n_{ij}^{-1} + n_{i+1,j}^{-1} + n_{i,j+1}^{-1} + n_{i+1,j+1}^{-1} \right)^{1/2}
\]

- This extends naturally to \( \theta_{ij} \mid k \) in higher-way tables, stratified by one or more “control” variables.

- Many models have a simpler form expressed in terms of \( \ln(\theta_{ij}) \).
  - e.g., Uniform association model

\[
\ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \gamma a_i b_j \equiv \ln(\theta_{ij}) = \gamma
\]

- Direct visualization of log odds ratios permits more sensitive comparisons than area-based displays.
Models for log odds ratios: Computation

- Consider an $R \times C \times K_1 \times K_2 \times \ldots$ frequency table $n_{ij...}$, with factors $K_1, K_2 \ldots$ considered as strata.
- Let $\mathbf{n} = \text{vec}(n_{ij...})$ be the $N \times 1$ vectorization of the table.
- Then, all log odds ratios and their asymptotic covariance matrix can be calculated as:
  
  \[
  \ln(\hat{\theta}) = \mathbf{C} \ln(\mathbf{n})
  \]
  
  \[
  S = \text{Var}[\ln(\theta)] = \mathbf{C} \text{ diag}(\mathbf{n})^{-1} \mathbf{C}^T
  \]

  where $\mathbf{C}$ is an $N$-column matrix containing all zeros, except for two $+1$ elements and two $-1$ elements in each row.
- e.g., for a $2 \times 2$ table, $\mathbf{C} = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}$
- With strata, $\mathbf{C}$ can be calculated as $\mathbf{C} = \mathbf{C}_{RC} \otimes \mathbf{I}_{K_1} \otimes \mathbf{I}_{K_2} \otimes \cdots$
- \text{loddsratio()} in \text{vcdExtra} package provides generic methods (\text{coef()}, \text{vcov()}, \text{confint()}, \ldots)
A log odds ratio linear model for the $\ln(\theta)$ is

$$\ln(\theta) = X\beta$$

where $X$ is the design matrix of covariates.

The (asymptotic) ML estimates $\hat{\beta}$ are obtained by GLS via

$$\hat{\beta} = \left( X^T S^{-1} X \right)^{-1} X^T S^{-1} \ln \hat{\theta}$$

where $S = \text{Var}[\ln(\theta)]$ is the estimated covariance matrix.

Standard diagnostic and graphical methods can be adapted to this case.

- diagnostics: influence plots, added-variable plots, 
- visualization: effect plots, 

Outline

1. Introduction
   - Brief History of VCD
   - Visual overview

2. Generalized Mosaic Displays: vcd Package
   - Extending mosaic-like displays
   - The strucplot framework

   - Loglinear models and generalized linear models
   - Generalized nonlinear models: gnm package
   - Models for ordered categories

4. 3D Mosaics: vcdExtra Package

5. Models and Visualization for Log Odds Ratios
   - Log odds ratios
   - Examples
Example: Breathlessness & Wheeze in Coal Miners

> fourfold(CoalMiners, mfcol = c(2, 4), fontsize = 18)

- There is a strong + association at all ages
- But can you see the trend?
Example: Breathlessness & Wheeze in Coal Miners

> (lor.CM <- loddsratio(CoalMiners))

log odds ratios for Wheeze and Breathlessness by Age
25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64
3.695  3.398  3.141  3.015  2.782  2.926  2.441  2.638

Fit linear and quadratic models in Age using WLS:

> lor.CM.df <- as.data.frame(lor.CM)
> age <- seq(25, 60, by = 5)
> CM.mod1 <- lm(LOR ~ age, weights=1/ASE^2, data=lor.CM.df)
> CM.mod2 <- lm(LOR ~ poly(age, 2), weights=1/ASE^2, data=lor.CM.df)
> anova(CM.mod1, CM.mod2)

Analysis of Variance Table
Model 1: LOR ~ age
Model 2: LOR ~ poly(age, 2)

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>6.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5.60</td>
<td>1</td>
<td>0.742</td>
<td>0.66</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Example: Breathlessness & Wheeze in Coal Miners

Plot log odds ratios and fitted regressions: The trend is now clear!
Attitudes toward corporal punishment

A four-way table, classifying 1,456 persons in Denmark (Punishment data in vcd package).

- **Attitude**: approves moderate punishment of children (moderate), or refuses any punishment (no)
- **Memory**: Person recalls having been punished as a child?
- **Education**: highest level (elementary, secondary, high)
- **Age group**: (15–24, 25–39, 40+)

<table>
<thead>
<tr>
<th>Education</th>
<th>Attitude</th>
<th>Age Memory</th>
<th>15–24</th>
<th>25–39</th>
<th>40+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>21</td>
<td>93</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>Yes</td>
<td>2</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>Secondary</td>
<td>No</td>
<td>Yes</td>
<td>5</td>
<td>45</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>No</td>
<td>2</td>
<td>26</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>No</td>
<td>1</td>
<td>19</td>
<td>4</td>
</tr>
</tbody>
</table>
Attitudes toward corporal punishment

Fourfold plots: Association of Attitude with Memory

> cotabplot(punish, panel = cotab_fourfold)
Log odds ratio plot

```r
> (lor.pun <- loddsratio(punish))
```

Log odds ratios for memory and attitude by age, education

<table>
<thead>
<tr>
<th>age</th>
<th>elementary</th>
<th>secondary</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-24</td>
<td>-1.7700</td>
<td>-0.2451</td>
<td>0.3795</td>
</tr>
<tr>
<td>25-39</td>
<td>-1.6645</td>
<td>-0.4367</td>
<td>0.4855</td>
</tr>
<tr>
<td>40+</td>
<td>-0.8777</td>
<td>-1.3683</td>
<td>-1.8112</td>
</tr>
</tbody>
</table>

- Structure now completely clear
- Little difference between younger groups
- Opposite pattern for the 40+
- Need to fit an LOR model to confirm appearances (SEs large)
- (These methods are under development)
Effective data analysis for categorical data depends on:

- Flexible models, with syntax to specify possibly complex models — *easily*
- Flexible visualization tools to help understand data, models, lack of fit, etc. — *easily*

The **vcd** package provides very general visualization methods via the strucplot framework

The **gnm** package extends the class of applicable models for contingency tables considerably

- Parsimonious models for structured associations
- Multiplicative and other nonlinear terms

The **vcdExtra** package provides glue, and a testbed for new visualization methods
Further information

http://www.jstatsoft.org/v17/i03/vignette("strucplot", package="vcd").

http://CRAN.R-project.org/package=gnm/vignette("gnmOverview", package="gnm").

**vcdExtra**  Friendly M & others (2010). **vcdExtra**: vcd additions.  
http://CRAN.R-project.org/package=vcdExtra/vignette("vcd-tutorial").
References


