Semi-algebraic geometry of Poisson regression

Thomas Kahle
Otto-von-Guericke Universität Magdeburg

joint work with Kai Oelbermann and Rainer Schwabe

Psychometrics is the field of objective measurement of skill, knowledge, ability, attitudes, personality, ….

Measuring Intelligence

The Berlin intelligence structure model (Jäger et al. 1984–) consists of 12 components of intelligence. Four "operational facets":

- Processing capacity (How many cores?)
- Processing speed (CPU frequency)
- Creativity (Hardware bugs?)
- Short-term memory (Size of CPU Cache)

are combined with "content categories": symbolic, numerical, verbal.

Measuring mental speed

- Give many simple tasks and measure processing speed.
- Historically test items from hand-crafted databases
  - labor intensive creation
  - subjects learn them
  - bias is hard to control

Measuring mental speed

- Give many simple tasks and measure processing speed.
- Historically test items from hand-crafted databases
  - labor intensive creation
  - subjects learn them
  - bias is hard to control
- Better: Rule-based item generation
  - Define rules with fixed influence on difficulty.
  - Trivial to generate more items by combining rules.
- Example: **MS^{2}T**: Münster mental speed test, Doebler/Holling in *Learning and individual differences* (2015).

Example of rule based item generation

= red phone

Example of rule based item generation

= red phone

Rule 1: Give the opposite of the correct answer

Example of rule based item generation

= red phone

Rule 1: Give the opposite of the correct answer

Rule 2: Apply Rule 1 only if the item in the picture is green.

Rules! on your phone

...
36. Even monsters
35. Red animals
34. Multiples of three
33. Primes
32. Third column
31. Ascending except Whales
30. Shake if Whales
29. Bipeds
28. Foxes
27. Fives
26. 5s-9s
...

Task: Model number of correct answers as a function of rules.

Regression

- Influences (Rules) are binary $\mathbf{x} \in \{0, 1\}^k$.
- Response is a count whose mean depends deterministically on $\mathbf{x}$.

Task: Model number of correct answers as a function of rules.

Regression

- Influences (Rules) are binary $\mathbf{x} \in \{0,1\}^k$.
- Response is a count whose mean depends deterministically on $\mathbf{x}$.

General principle of statistical regression

The expected value of the dependent variable Y is a deterministic function of the influences X:

$$\mathbb{E}(Y|X = x) = r(x)$$

The Rasch Poisson counts model

- The number of correct answers is Poisson distributed:

$$\text{Prob}(\#\text{correct answers} = m) = \frac{\lambda^m e^{-\lambda}}{m!}$$

- Intensity $\lambda = \theta\sigma$ depends on ability θ of subject and easiness σ.

Calibration of rule influence

- Assume ability θ of a subject is known (or at least fixed).
- Want to calibrate the influence of rules on σ.

Poisson regression: Influence on exponential scale – log-linear model

$$\lambda(\mathbf{x}) = \theta\sigma(\mathbf{x}) = \theta \exp(f(\mathbf{x}) \cdot \beta)$$

Calibration of rule influence

- Assume ability θ of a subject is known (or at least fixed).
- Want to calibrate the influence of rules on σ.

Poisson regression: Influence on exponential scale – log-linear model

$$\lambda(\mathbf{x}) = \theta\sigma(\mathbf{x}) = \theta \exp(f(\mathbf{x}) \cdot \beta)$$

- Binary rules: $\mathbf{x} \in \{0, 1\}^k$
- Regression functions f translate settings into numbers.

No interaction $f(\mathbf{x}) = (1, x_1, x_2, \ldots, x_k)$

Pairwise interaction $f(\mathbf{x}) = (1, x_1, \ldots, x_k, x_1x_2, \ldots, x_{k-1}x_k)$

$\ldots$

Saturated model $f(\mathbf{x}) = (\prod_{i \in A} x_i : A \subseteq \{1, \ldots, k\})$

Multiplicative structure

$$\lambda(\mathbf{x}) = \theta \exp(f(\mathbf{x}) \cdot \beta) = \prod_{A \subseteq \mathbf{x}} e^{\beta_A}$$

- Convenient: Rules determine which factors appear.
  - Will often choose $\beta_A < 0$
- Implicit equations in $\lambda(\mathbf{x})$:
  - Independence: (2×2)-minors

$$\lambda(00, \beta)\lambda(11, \beta) = \lambda(10, \beta)\lambda(01, \beta)$$

  - All terms up to order $k - 1$: One generator

$$\prod_{|\mathbf{x}| \text{ odd}} \lambda(\mathbf{x}, \beta) = \prod_{|\mathbf{x}| \text{ even}} \lambda(\mathbf{x}, \beta)$$

  - In between: Query MBDB, 4ti2, or give up.

General framework

In a generalized linear model, the expectation varies as

$$\mathbb{E}(Y|X=x) = g^{-1}(f(x)\cdot\beta)$$

- f is a vector of regression functions
- β is a vector of parameters
- A link function g (e.g. id, $\log$) couples the expectation value and the linear predictor.
- Distributions around the mean from exponential family (e.g Gauss, Poisson, Binomial, Gamma, ...).

$\Rightarrow$ general theory for estimation, testing, fit, etc.

Experimental design

- Can observe n times: generate $(Y_i|\mathbf{x}_i)$ for *chosen* $\mathbf{x}_i$.
- How to pick $\mathbf{x}_i$ so that our experiment is most informative about the parameters?
- A design is a choice of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \{0,1\}^k$.
- An approximate design is a choice of real weights $w_{\mathbf{x}} \geq 0, \mathbf{x} \in \{0,1\}^k$ with $\sum_{\mathbf{x}} w_{\mathbf{x}} = 1$.

Optimal experimental design

A design is good if the variance of unbiased estimators is low.

Fisher Information

- Information gained from observing a single experiment (one value of the Poisson variable, given a setting $\mathbf{x}$) is measured with the Fisher Information

$$M(\mathbf{x}, \beta) = \lambda(\mathbf{x}, \beta) f(\mathbf{x}) f(\mathbf{x})^T$$

- Information of an approximate design w

$$M(w, \beta) = \sum_{\mathbf{x}} w_{\mathbf{x}} \lambda(\mathbf{x}, \beta) f(\mathbf{x}) f(\mathbf{x})^T$$

- Connection to estimator variance: Cramer-Rao inequality.

Experimental design as an optimization problem

Optimality

A design is locally D-optimal at β if it maximizes the determinant of the information matrix.

Optimal experimental design

- Chicken and Egg Problem: Optimal design depends on β.
- BUT: "Regions of optimality" are often semi-algebraic.

Remarks

- Person with highest ability provides most information!
- Optimization can be carried out with $\theta = 1, \beta_0 = 0$.

Two independent rules (Graßhoff/Holling/Schwabe)

- Settings $\mathbf{x} \in \{00, 01, 10, 11\}$, $\quad \lambda(\mathbf{x}, \beta) =: \lambda_{\mathbf{x}} = \prod_i e^{x_i \beta_i}$
- Weights $w_{00} + w_{01} + w_{10} + w_{11} = 1$.

$$f(00)^T = (1,0,0) \quad f(10)^T = (1,1,0)$$
$$f(01)^T = (1,0,1) \quad f(11)^T = (1,1,1)$$

$$f(00)f(00)^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad f(10)f(10)^T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$f(01)f(01)^T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} \quad f(11)f(11)^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Two independent rules (Graßhoff/Holling/Schwabe)

- Settings $\mathbf{x} \in \{00, 01, 10, 11\}$, $\lambda(\mathbf{x}, \beta) =: \lambda_{\mathbf{x}} = \prod_i e^{x_i \beta_i}$
- Weights $w_{00} + w_{01} + w_{10} + w_{11} = 1$.

Information of the design w:

$$M(w, \beta) = \begin{pmatrix} \sum_{\mathbf{x}} w_{\mathbf{x}} \lambda_{\mathbf{x}} & w_{11}\lambda_{11} + w_{10}\lambda_{10} & w_{11}\lambda_{11} + w_{01}\lambda_{01} \\ w_{11}\lambda_{11} + w_{10}\lambda_{10} & w_{11}\lambda_{11} + w_{10}\lambda_{10} & w_{11}\lambda_{11} \\ w_{11}\lambda_{11} + w_{01}\lambda_{01} & w_{11}\lambda_{11} & w_{11}\lambda_{11} + w_{01}\lambda_{01} \end{pmatrix}$$

with determinant

$$\det(M(w, \beta)) = w_{11}w_{10}w_{01}\lambda_{11}\lambda_{10}\lambda_{01} + w_{11}w_{10}w_{00}\lambda_{11}\lambda_{10}\lambda_{00} + \\ w_{11}w_{01}w_{00}\lambda_{11}\lambda_{01}\lambda_{00} + w_{01}w_{10}w_{00}\lambda_{01}\lambda_{10}\lambda_{00}$$

Maximize as a function of parameters β_1, β_2.

Two independent rules (Graßhoff/Holling/Schwabe)

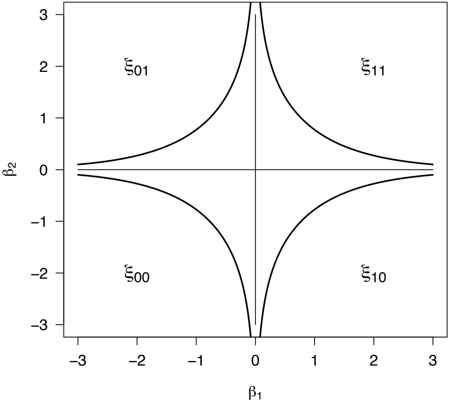
$\xi_{00} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$

$\vdots$

$\xi_{11} = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Origin: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

Diamond: Full support



Curve in lower right quadrant:

$$\lambda_{10} + \lambda_{01} + \lambda_{11} = 1 \Leftrightarrow e^{\beta_1} + e^{\beta_2} + \mathrm{e}^{\beta_1+\beta_2} = 1 \Leftrightarrow \beta_2 = \log \frac{1 - e^{\beta_1}}{1 + e^{\beta_1}}$$

If rules make problem hard, then 11 is not very informative.

Geometry of fixed parameter optimization problem

- Maximize log-concave function $\det$ over
- Polytope of design matrices

$$P_\beta = \mathrm{conv}\{\lambda(\mathbf{x}, \beta) f(\mathbf{x}) f(\mathbf{x})^T : \mathbf{x} \in \{0,1\}^k\}$$

Note: Both target function and geometry of P_β depend on β.

Three Independent rules

- $\beta = 0$: Cyclic polytope
- $\beta \neq 0$: Simplex

Candidates for optimal designs

Full support

- For $\beta = 0$, equal weights on all design points $\mathbf{x} \in \{0,1\}^k$.
- Numerical optimization in region with full support
  - Need to round before realization
- Caratheodory's theorem: Solution in w not unique.

Restricted support

- A design is saturated if the support of w has the same size as the number of parameters.
  - This is the minimal number (otherwise $\det = 0$)
  - Can be expensive to change setting $\mathbf{x}$ (not here)
  - All weights must be equal $\rightarrow$ Optimal weights rational
  - Model validation (test for for higher interaction) is impossible.

The corner design

If rules make the problem hard

Fix an interaction order d. The corner design w^* consists of equal weights on the points

$$\left\{\mathbf{x} \in \{0,1\}^k : |\mathbf{x}|_1 \leq d\right\}$$

Optimality of the corner design

Theorem

Consider the Rasch Poisson counts model with interaction order d and k binary predictors. Denote $\mu_A = e^{\beta_A}$, $|A| \leq d$. The design w^* is D-optimal if and only if for all $C \subseteq [k]$ with $|C| = d + 1$

$$\prod_{A \subseteq C} \mu_A + \sum_{B \subseteq C} \prod_{\substack{A \subseteq C, \\ A \neq B}} \mu_A \leq 1$$

Note: inequalities are imposed in parameter space.

Optimality of the corner design

Theorem

Consider the Rasch Poisson counts model with interaction order d and k binary predictors. Denote $\mu_A = e^{\beta_A}$, $|A| \leq d$. The design w^* is D-optimal if and only if for all $C \subseteq [k]$ with $|C| = d+1$

$$\prod_{A \subseteq C} \mu_A + \sum_{B \subseteq C} \prod_{\substack{A \subseteq C, \\ A \neq B}} \mu_A \leq 1$$

Example: k independent rules (Graßhoff/Holling/Schwabe)

Design w^* is optimal if for all pairs i, j

$$\mu_i \mu_j + \mu_i + \mu_j \leq 1.$$

Technology: the Kiefer-Wolfowitz Theorem

For saturated designs, the optimization problem is solved in general by

Kiefer-Wolfowitz general equivalence theorem

Let w be a saturated design. $\Psi = \mathrm{diag}(1, (\mu_A)_{|A|\leq d})$, and F the matrix with rows $\{f(\mathbf{x}) : \mathbf{x} \in \mathrm{supp}(w)\}$. Then w is locally D-optimal if and only if for all $\mathbf{x} \in \{0,1\}^k$

$$\lambda(\mathbf{x})(F^{-T} f(\mathbf{x}))^T \Psi^{-1} (F^{-T} f(\mathbf{x})) \leq 1$$

- For corner design w^* can determine F^{-T} explicitly.
- Equality holds on the design points $\mathbf{x} \in \mathrm{supp}(w)$
- For $|\mathbf{x}|_1 = d + 1$ we get inequalities in the theorem
- Remaining inequalities redundant by monotonicity arguments.

Other saturated designs

Conjecture

If $\beta_A < 0$ then no saturated design except w^* is ever optimal.

Kiefer-Wolfowitz

- For each saturated design get (rational) inequality system
  - Don't know how to invert F by hand.
- Need to show that inequality system is infeasible.
  - Best software comes from optimization community
  - Positivstellensatz

Evidence in easy cases

- Grasshoff/Holling/Schwabe did $d = 1$, $k = 3$ by hand:
  - Up to symmetry there are 4 inequality systems to be checked.
  - Could find two inequalities that contradict each other.
- Magma, Maxima, Maple: DNF
- Numerics: For $d = 1, k = 4$
  - used moment relaxations with Sage/Matlab/Yalmip/MOSEK
  - Challenge: Conditioning of the resulting SDP

Goal: Explicit Positivstellensatz certificates.

Outlook

- Interpretation: Optimal design wants many combinations, but avoid low intensity.
- Geometry of the information matrix polytope?
- Inequalities in $\lambda(\mathbf{x})$?

Related work

- Russel et al (2009): Similar results for (independent) continuous predictors.
- Yang et al. (2012): successful application of quantifier elimination in a similar setting (binary response).

Outlook

- Interpretation: Optimal design wants many combinations, but avoid low intensity.
- Geometry of the information matrix polytope?
- Inequalities in $\lambda(\mathbf{x})$?

Related work

- Russel et al (2009): Similar results for (independent) continuous predictors.
- Yang et al. (2012): successful application of quantifier elimination in a similar setting (binary response).

Thanks!