

On a new approach to regression with ordinal explanatory variables

Kaspar Rufibach
Department of Statistics
Stanford University (until August 31, 2007)
Supported by Swiss National Science Foundation

ROeS – Seminar Bern
September 11, 2007

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

Goal: model $y \in \mathbb{R}$ based on feature vector $\mathbf{x} \in \mathbb{R}^p$

Given: Training set $(y_i, (x_{ij})_{j=1}^p), i = 1, \dots, n$

Maximize **concave** criterion function

$$\ell(\boldsymbol{\beta}) = \ell(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) : \mathbb{R}^n \times \mathbb{R}^{n \times p} \times \mathbb{R}^p \rightarrow \mathbb{R}$$

over $\boldsymbol{\beta} \in \mathbb{R}^p$:

$$\hat{\boldsymbol{\beta}} := \arg \max_{\boldsymbol{\beta} \in \mathcal{P}} \ell(\boldsymbol{\beta})$$

for some suitable $\mathcal{P} \subseteq \mathbb{R}^p$.

Examples:

- $\mathbf{y} \in \mathbb{R}^n$: **ordinary least squares**,

$$l_1(\boldsymbol{\beta}) = - \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.$$

- $\mathbf{y} \in \{0, 1\}^n$: **logistic regression**,

$$l_2(\boldsymbol{\beta}) = - \sum_{i=1}^n \left(-y_i \mathbf{x}_i^\top \boldsymbol{\beta} + \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})) \right).$$

- Survival times: **prop. hazard model** (Cox-Regression),

$$l_3(\boldsymbol{\beta}) = \sum_{s=1}^D \alpha_{(s)} - \sum_{s=1}^D \log \left(\sum_{k \in R_s} \alpha_k \right)$$

for $\alpha_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, $\alpha_{(s)}$ for s -th failure time, $s = 1, \dots, D$

3 types of explanatory variables $\mathbf{x}_{.j} = (x_{ij})_{i=1}^n$:

- **quantitative**: $x_{ij} \in \mathbb{R}$, e.g. age, body weight, hormone levels
- **grouped ("factors")**: $x_{ij} \in \{1, \dots, g\}$, e.g. sex, country, types of treatments
- **ordered factors**: $x_{ij} \in \{1, \dots, g\}$, e.g. performance status/stage in oncology, answers in surveys, e.g. "never", "sometimes", "always"

Proportional hazard regression:

$$\begin{array}{ccccccccccc}
 (T_1, \delta_1) & x_{11} & \cdots & x_{1p_1} & f_{11} & \cdots & f_{1p_2} & o_{11} & \cdots & o_{1p_3} \\
 \vdots & \sim & \vdots & & & & & & & \vdots \\
 (T_n, \delta_n) & x_{n1} & \cdots & x_{np_1} & f_{n1} & \cdots & f_{np_2} & o_{n1} & \cdots & o_{np_3}
 \end{array}$$

⇓

y	x₁	x₂	f₁	o₁	o₂
7.2+	42	1.1	2	1	3
6.5+ ~	34	2.1	1	3	1
⋮	⋮				⋮
8.3	23	0.7	2	3	5

No ordering for **f₁**: 1, 2

Ordered levels for **o₁, o₂**: 1 < ... < 5

Approaches to incorporate \mathbf{o}_j :

- Ignore grouping, \mathbf{o}_j as quantitative \Rightarrow implicitly assume **levels spaced according to coding** \Rightarrow estimated regression parameter $\hat{\beta}_j$ difficult to interpret
- Treat as unordered factor \Rightarrow dummy variables $\mathbf{o}_{j_2}, \dots, \mathbf{o}_{j_{p_j}} \Rightarrow \hat{\beta}_2 \leq \dots \leq \hat{\beta}_{j_{p_j}}$ **not guaranteed**. Many levels \Rightarrow many predictors. Regularization!
- Get simple models (interactions!) \Rightarrow **pool levels**, even dichotomization of \mathbf{o}_j . Cutoff(s)? Kind of model selection!
- Polynomial contrasts: dummies $\mathbf{o}_{j_i} = i^2 \{\mathbf{o}_j = i\}$, $i = 2, \dots, p_j$. To avoid correlated estimators/tests \Rightarrow modify design matrix, **get orthogonal contrasts** \Rightarrow `as.ordered()` in R

Example from **oncology** (IMSV consulting case Fleischmann):

Dependent: **OS** of 102 metastatic prostate cancer patients

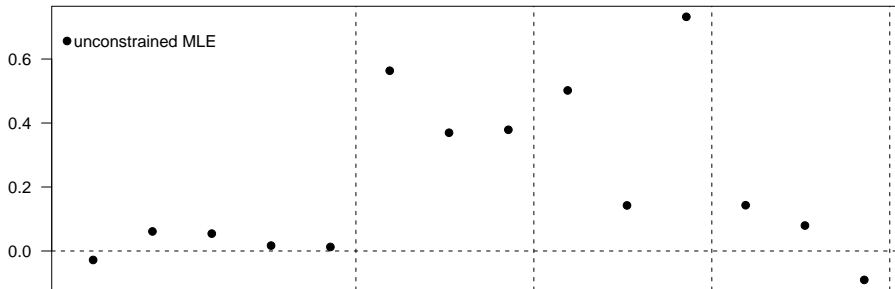
Predictors:

- age (years)
- Extracapsular extension of lymph node metastases (yes/no)
- Maximal diameter biggest metastasis (mm)
- Cancer volume in prostate (cm³)
- #lymph nodes with metastasis (count)
- **Tumor staging**: pT2 < pT3a < pT3b < pT4
- **Primary tumor Gleason score**: 6 < 7 < 8 < 9, 10
- **Lymph nodes Gleason score**: 6, 7 < 8 < 9 < 10

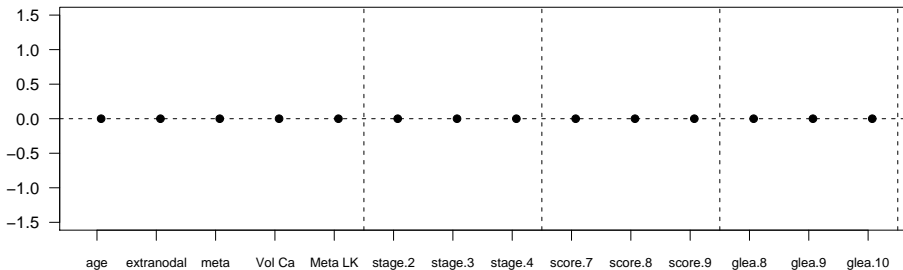
Higher stage/score corresponds to higher risk \Rightarrow **increasing coefficients for dummy variables desirable**

PROSTATE CANCER EXAMPLE

Estimated coefficients: beta hat



gradient at beta hat



- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

Generate dummy variables for the (ordered) factors:

\mathbf{y}		\mathbf{x}_1	\mathbf{x}_2	\mathbf{f}_1	\mathbf{o}_1	\mathbf{o}_2
7.2+		42	1.1	2	1	3
6.5+	~	34	2.1	1	3	1
\vdots		\vdots				\vdots
8.3		23	0.7	2	3	5

↓

\mathbf{y}		\mathbf{x}_1	\mathbf{x}_2	$\mathbf{f}_{1.2}$	$\mathbf{o}_{1.2}$	$\mathbf{o}_{1.3}$	$\mathbf{o}_{2.2}$	$\mathbf{o}_{2.3}$	$\mathbf{o}_{2.4}$	$\mathbf{o}_{2.5}$
7.2+		42	1.1	1	0	0	0	1	0	0
6.5+	~	34	2.1	0	0	1	0	0	0	0
\vdots		\vdots								
8.3		23	0.7	1	0	1	0	0	0	1

Denote this new design matrix by $\mathbf{X} \in \mathbb{R}^{n \times d}$

Define estimators:

$$\hat{\beta} := \underset{\beta \in \mathcal{B}}{\text{maximize}} \ell_3(\beta) \quad \hat{\eta} := \underset{\beta \in \mathbb{R}^d}{\text{maximize}} \ell_3(\beta)$$

where

$$\mathcal{B} = \{\beta \in \mathbb{R}^d : \beta_{o_{1.2}} \leq \beta_{o_{1.3}}, \beta_{o_{2.2}} \leq \beta_{o_{2.3}} \leq \beta_{o_{2.4}} \leq \beta_{o_{2.5}}\}.$$

Remember model:

$$\mathbf{y} \sim \mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{f}_{1.2} \quad \mathbf{o}_{1.2} \quad \mathbf{o}_{1.3} \quad \mathbf{o}_{2.2} \quad \mathbf{o}_{2.3} \quad \mathbf{o}_{2.4} \quad \mathbf{o}_{2.5}$$

Accordingly for OLS, logistic regression

Only relevant if #levels ≥ 3 (1 level “lost” as a reference)

Fixed response and $\mathbf{X} \Rightarrow$ the functions $l_i : \mathbb{R}^d \rightarrow [-\infty, \infty)$ are

- strictly concave
- coercive:

$$l_i(\boldsymbol{\beta}) \rightarrow -\infty \quad \text{if} \quad \|\boldsymbol{\beta}\| \rightarrow \infty,$$

- continuously differentiable on

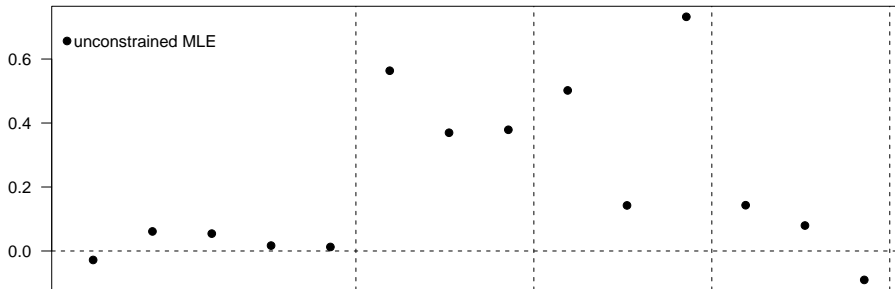
$$\text{dom}(l_i) := \{\boldsymbol{\beta} \in \mathbb{R}^d : l_i(\boldsymbol{\beta}) > -\infty\}$$

- The set \mathcal{B} is closed and convex.

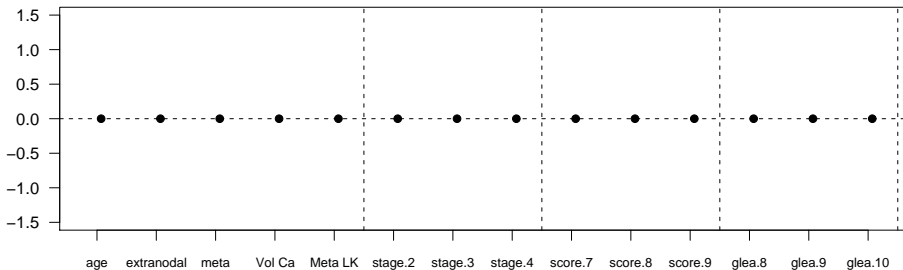
\Rightarrow the vectors $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\gamma}}_i$ **exist and are unique** for $i = 1, 2, 3$.

PROSTATE CANCER EXAMPLE

Estimated coefficients: beta hat

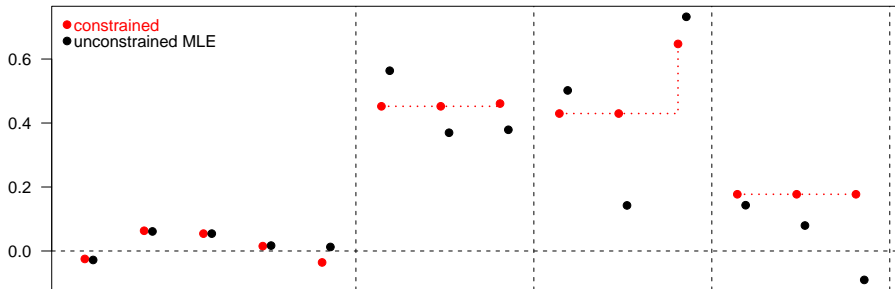


gradient at beta hat

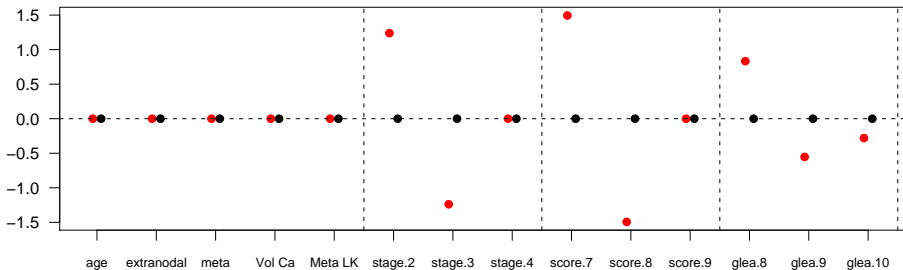


PROSTATE CANCER EXAMPLE

Estimated coefficients: beta hat



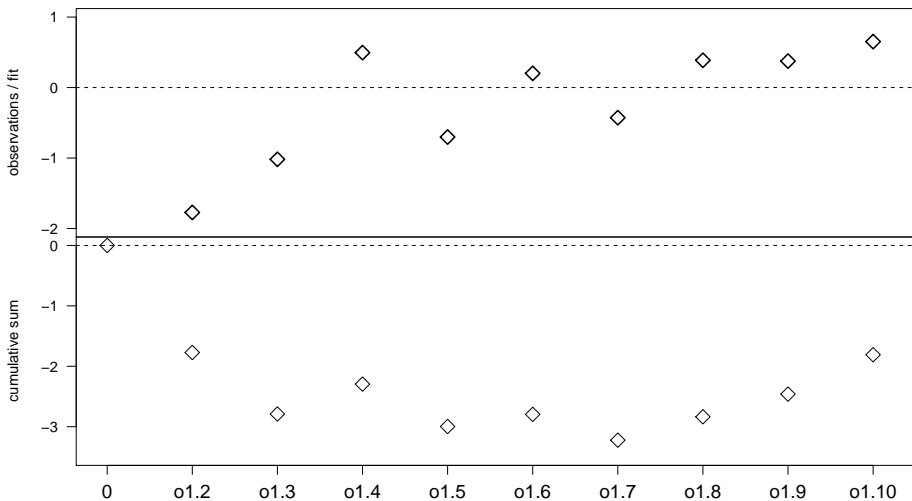
gradient at beta hat



Features of this new estimator:

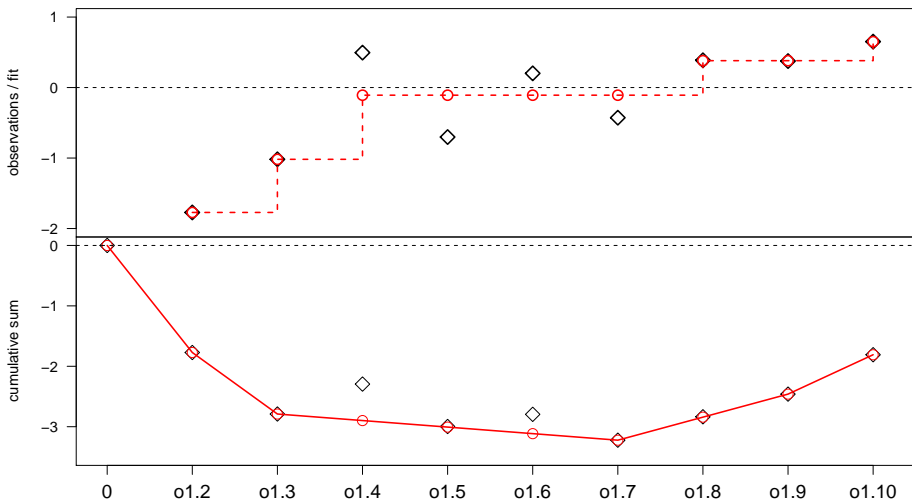
- **intuitive** approach, exploits available knowledge **precisely**
- get an **“estimate” for levels** \Rightarrow codings are often chosen arbitrarily
- **interpretable** coefficients for dummy variables
- some sort of **regularization (without specifying any tuning parameters!)**, especially relevant for **small data sets**

Simple toy example: LS - regression with $n = 10$ ordered groups, each containing only 1 observation: $\mathbf{y} \in \mathbb{R}^{10}$, $\mathbf{o}_1 = (1, \dots, 10)$



Define: $C\mathbf{v}(k) = \sum_{i=1}^k v_i$. $C(0, y_2, \dots, y_{10})(k)$.

Simple toy example: LS - regression with $n = 10$ ordered groups, each containing only 1 observation: $\mathbf{y} \in \mathbb{R}^{10}$, $\mathbf{o}_1 = (1, \dots, 10)$



Define: $C\mathbf{v}(k) = \sum_{i=1}^k v_i$. $C(0, y_2, \dots, y_{10})(k)$.

$C(0, \hat{\beta}_{o_{1.2}}, \dots, \hat{\beta}_{o_{1.10}})(k), k = 1, \dots, 10$.

Initial questions:

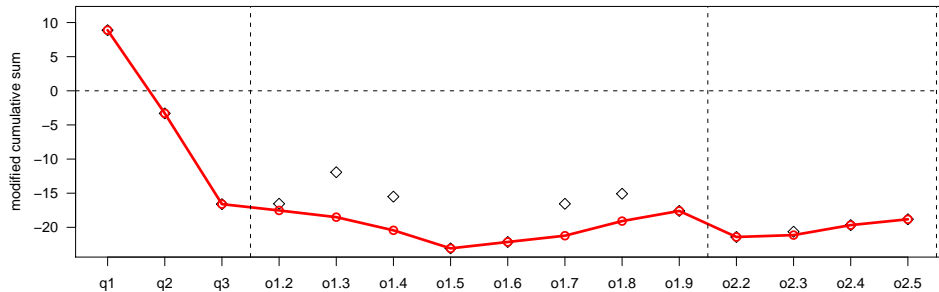
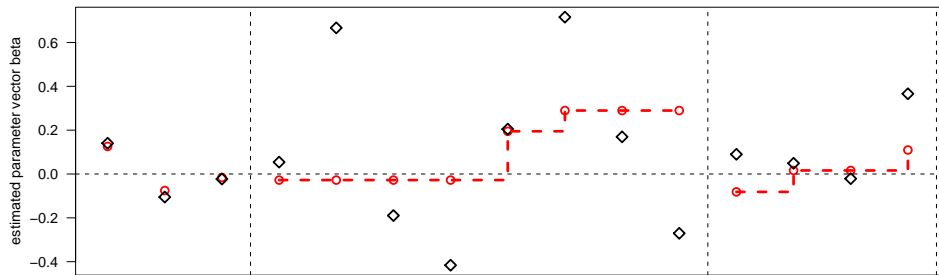
- Does characterization **generalize** to our setting?
- If yes, how?
- If no, why not?
- “Piecewise convexity”?
- What about logistic and Cox?

Initial questions:

- Does characterization **generalize** to our setting? **Somehow.**
- If yes, how? **Only analytically, not geometrically.**
- If no, why not? **More involved situation.**
- “Piecewise convexity”? **No!**
- What about logistic and Cox? **Similar to least squares.**

Complicated toy example:

- $n = 100$ observations
- 3 quantitative variables
- 2 factors with 9 and 5 levels
- least squares



Lemma (Characterization of solution)

An arbitrary admissible vector $\hat{\gamma}$ maximizes l *iff*:

$$\left(\nabla l(\hat{\gamma})\right)_{s=1}^c = 0$$

and for every ordered factor j we have

$$\sum_{s \in S_{j,k}} \left(\nabla l(\hat{\gamma})\right)_s \geq 0$$

$$\sum_{s \in T_{j,k}} \left(\nabla l(\hat{\gamma})\right)_s \leq 0$$

where $S_{j,k}, T_{j,k} \subseteq \{1, \dots, p\}$ are specific sets and k depends on the "kinks" of $\hat{\gamma}$ for a given j .

Back to complicated toy example:

Variable	unconstrained		constrained estimator			
	$\hat{\eta}_1$	$\nabla \ell(\hat{\eta}_1)$	$\hat{\beta}_1$	$\nabla \ell(\hat{\beta}_1)$	$\sum \downarrow \nabla \ell(\hat{\beta}_1)$	$\sum \uparrow \nabla \ell(\hat{\beta}_1)$
q1	0.140	0	0.126	0	0	0
q2	-0.105	0	-0.075	0	0	0
q3	-0.023	0	-0.016	0	0	0
f1.2	0.054	0	-0.028	1.959	1.959	0
f1.3	0.667	0	-0.028	11.201	13.161	-1.959
f1.4	-0.190	0	-0.028	-3.294	9.867	-13.161
f1.5	-0.416	0	-0.028	-9.867	0	-9.867
f1.6	0.205	0	0.195	0	0	0
f1.7	0.716	0	0.290	9.342	9.342	0
f1.8	0.170	0	0.290	-1.336	8.005	-9.342
f1.9	-0.271	0	0.290	-8.005	0	-8.005
f2.2	0.090	0	-0.082	0	0	0
f2.3	0.049	0	0.016	0.988	0.988	0
f2.4	-0.021	0	0.016	-0.988	0	-0.988
f2.5	0.366	0	0.110	0	0	0

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm**
- ④ Statistical inference
- ⑤ Further problems

General problem: Find the vector

$$\hat{\beta} := \arg \max_{\beta \in \mathcal{B}} \ell(\beta).$$

where

$$\mathcal{B} = \{\beta \in \mathbb{R}^d : \mathbf{v}_i^\top \beta \leq 0, i = 1, \dots, q\}$$

for given (linearly independent) vectors $\mathbf{v}_i \in \mathbb{R}^d$.

⇒ active set algorithm

Basic active set strategy consists of **two alternating** steps:

- **Unconstrained optimization** where “violating” dummy variables are pooled (i.e. added).
- **Vary** set of newly “violating” dummy variables in very specific way.

Finitely many iteration loops: once correctly identified “set of jumps” of $\hat{\beta} \Rightarrow$ unconstrained optimization!

No tuning parameters in algorithm.

Very efficient, easily deals with large datasets.

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- **Confidence interval** for or **test** whether $\beta_i = 0$ vs. $\beta_i \neq 0$ for some $i = 1, \dots, d$?

Since **fast algorithm** available

\Rightarrow **Bootstrap pairs**, i.e. draw M samples (y_m^*, \mathbf{x}_m^*) , $m = 1, \dots, M$ (with replacement) from all observations (y_i, \mathbf{x}_i) , $i = 1, \dots, n$, compute standard errors, CI, tests, ...

Efron & Tibshirani (1993) for more on bootstrapping regression

- Likelihood ratio tests ?

- ① Problem, (de-)motivating example, existing approaches
- ② A new estimator
- ③ Computation via active set algorithm
- ④ Statistical inference
- ⑤ Further problems

- Striking application ?!?! Dose-response models?
- Combine with **penalization methods** such as Lasso, elastic net - what properties does $\hat{\beta}$ then have ? Computation ?
- **Asymptotics**: Rate of convergence?

Software for R available upon request

Thank you.