On the Shoulders of Gauss, Bessel, and Poisson: links, chunks, spheres — and conditional models

William D Heavlin
Google, Inc.
Mountain View, California

abstract

- For generalized linear models (GLMs), we define the *conditional likelihood,* which partitions (or chunks) data into subsets.
- To efficiently compute such a likelihood, which has combinatorial complexity, we introduce the *spherical approximation*.
- The resulting model estimates are highly linearized and computationally attractive.
- Further, these estimates refine our understanding of standard (unconditional) GLMs.

gloss

- GLMs (*links*)
- Conditional likelihood (*chunks*)

  Combinatorial complexity

- approximation (*spheres*)

  Computationally attractive

outline of what is to come

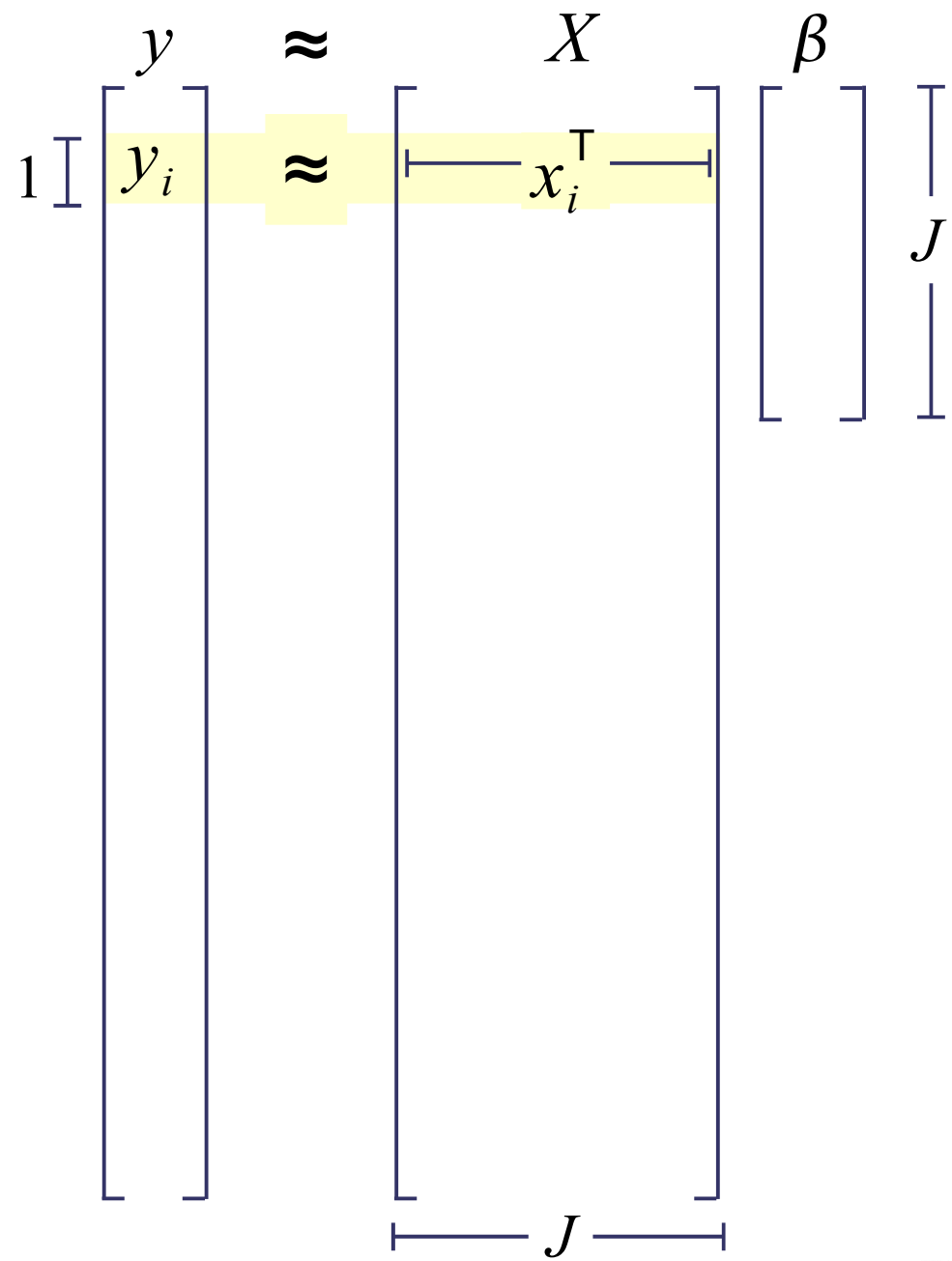
- notation and exponential models
- chunks, conditional models, and why
- the spherical approximation,
  its estimator and examples
- things similar++

Notation (I)

- (scalar) response y
- (vector) feature x
- β: J-vector of parameters

- Latent feature linearity:

$$\Pr\{y_i \mid x_i\} = \Pr\{y_i \mid x_i^\top \beta\}$$
$$= \Pr\{y_i \mid \eta_i\}$$

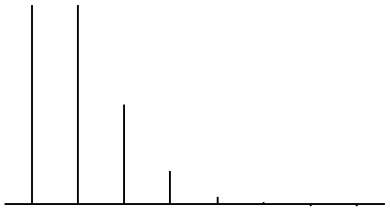


Three Models

- Poisson:

$$\Pr\{y_i|\eta_i\}=\frac{e^{-\eta_i}\eta_i^{y_i}}{y_i!},\ y_i=0,1,2,\ldots$$

Siméon-Denis Poisson

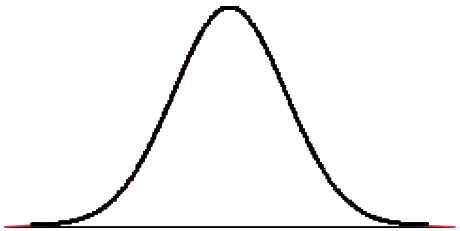


- Gaussian:

$$\Pr\{y_i|\eta_i\}\quad(-\infty<y_i<\infty)$$

$$=\frac{\exp\{-(y_i-\eta_i)^2/(2\sigma^2)\}}{\sqrt{2\pi\sigma}}$$

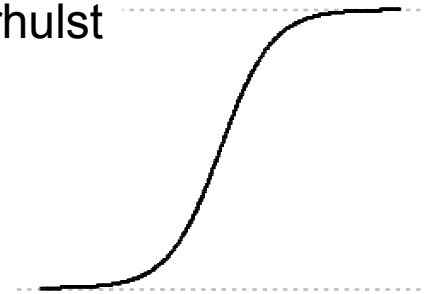
Carl Friedrich Gauss



- Logistic:

$$\Pr\{y_i|\eta_i\}=\frac{\exp\{y_i\eta_i\}}{1+\exp\{\eta_i\}},\ y_i=0,1$$

Pierre F Verhulst



Exponential family (EXP)

$$f(y_i|x_i) = \exp\{y_i h_0(\eta_i) + h_1(\eta_i) + h_2(y_i)\}$$

	$h_0(\eta_i)$	$+h_1(\eta_i)$	$+h_2(y_i)$
• Poisson: $\Pr\{y_i\|\eta_i\} = \frac{e^{-\eta_i}\eta_i^{y_i}}{y_i!}$	$\log(\eta_i)$	$-\eta_i$	$-\log(y_i!)$
• Gaussian: $\frac{\exp\{-(y_i-\eta_i)^2/(2\sigma^2)\}}{\sqrt{2\pi\sigma}}$	η_i/σ^2	$-\eta_i^2/(2\sigma^2)$	$-y_i^2/(2\sigma^2)$
• Logistic: $\Pr\{y_i\|\eta_i\} = \frac{\exp\{y_i\eta_i\}}{1+\exp\{\eta_i\}}$	η_i	$-\log(1+e^{\eta_i})$	0

- linearized exponential "link:" $h_0(\eta_i) = \eta_i$

  Poisson: $\eta_i \leftarrow \exp(\eta_i)$

  Gaussian: $\eta_i/\sigma^2 \leftarrow \eta_i$ (known sigma)

  Logistic: $\eta_i \leftarrow \eta_i$

chunks: Bessel's (1826) personal equations

4 astronomers

- Bessel
- Walbeck
- Argelander
- Struve

in 1820-21	B — W	=	-1.041	secs
	S — W	=	-0.242	
hence	B — S	=	-0.799	
in 1823	B — A	=	-1.223	
	S — A	=	-0.202	
giving	B — S	=	-1.021	

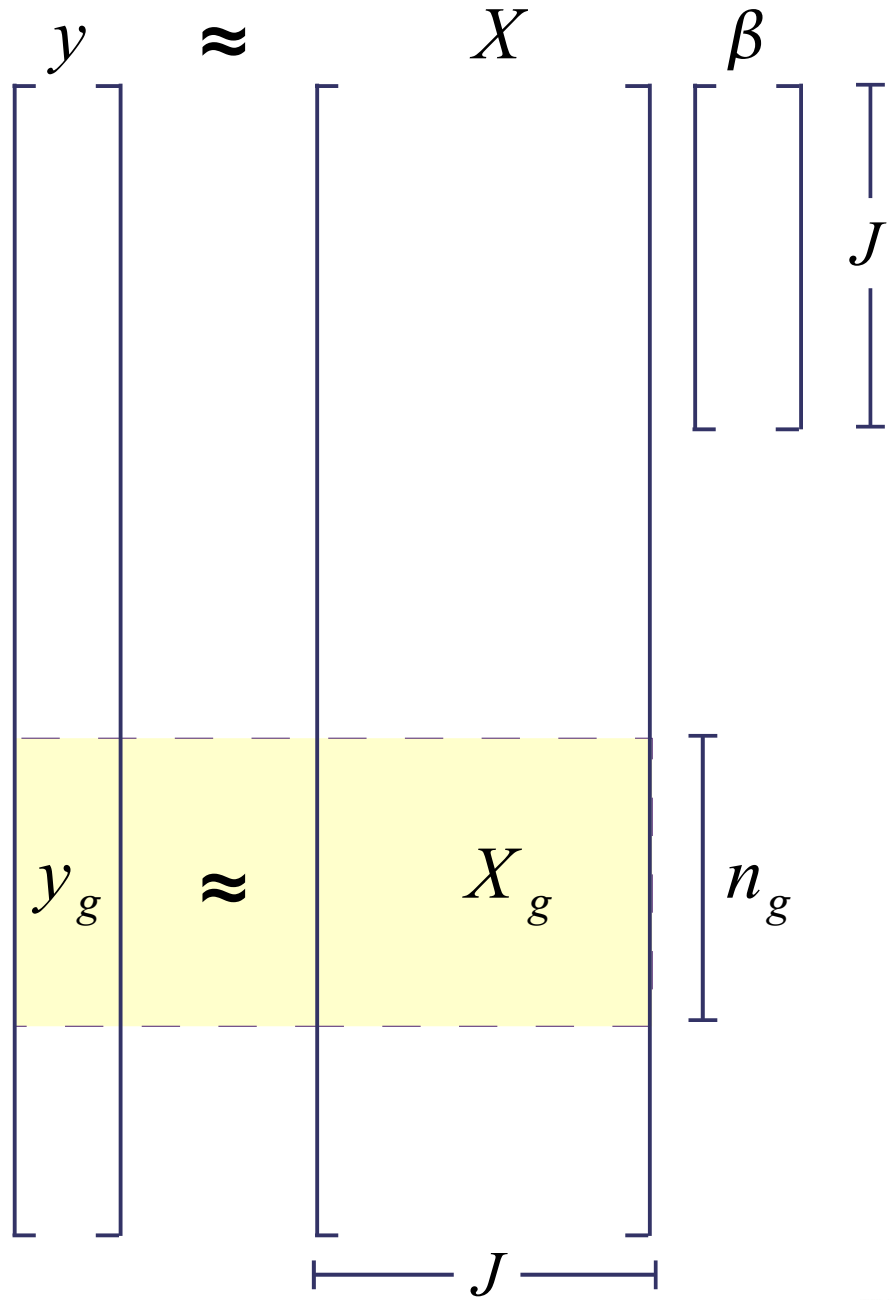
Notation (II) ("chunks")

- e.g. observers' "personal equations"

- chunked into groups

chunk $g \rightarrow$

- y_g, an n_g-vector
  X_g, an $n_g \times J$ feature matrix



Chunk effects

chunks have latent ("nuisance") effects:

$$y_{gi} \approx \mu_g + x_{gi}^{\mathsf{T}} \beta$$

chunk-specific

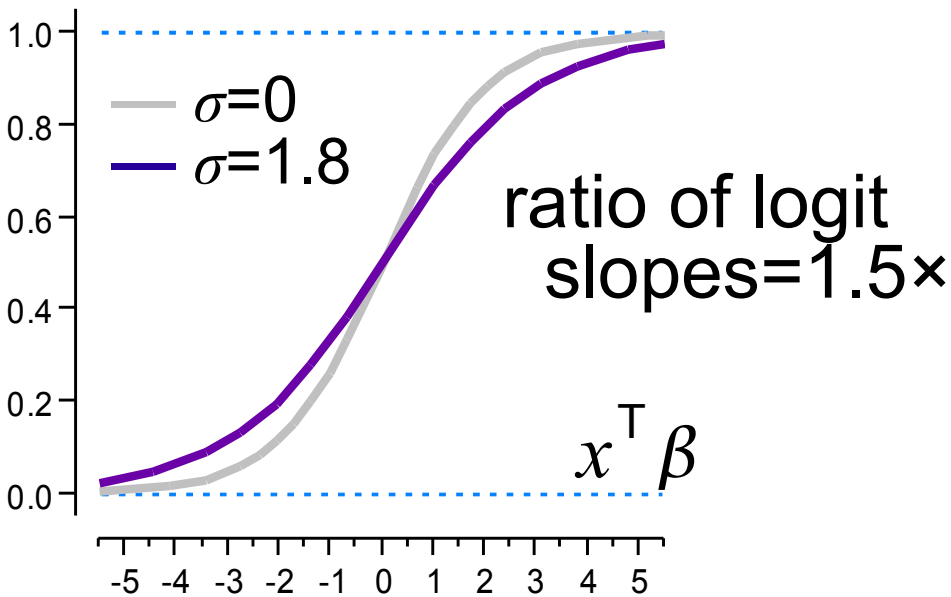
where chunks come from:

- single data or observer source/time or
- single entity obs'd at different times.

why chunks matter:

- confounding ...
- attenuation!

$$\int_{-\infty}^{\infty} \frac{e^{\sigma z + x^{\mathsf{T}} \beta}}{1 + e^{\sigma z + x^{\mathsf{T}} \beta}} f(z)\, dz$$



Conditioning ...

- conditioning eliminates these "nuisance effects" — as h_1-terms.
- unconditional likelihood:

$$\prod_g \prod_i f(y_{gi} | x_{gi})$$

- conditional likelihood:

$$\prod_g \frac{\prod_i f(y_{gi} | x_{gi})}{\sum_\tau \prod_i f(y_{g\tau(i)} | x_{gi})}$$

$$= \prod_g \frac{f(y_g | X_g)}{\sum_\tau f(P_\tau y_g | X_g)}$$

$n_g!$ terms

$P_\tau y_g$, the permutation of the n_g elements of y_g

- ...conditioning on $P_\tau y_g, P_\tau$ unknown, ≡the histogram of y_g

$$\overset{y_g}{\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}} \approx \overset{X_g}{\begin{bmatrix} \quad \end{bmatrix}} \begin{bmatrix} \beta \end{bmatrix}$$

condition on
#0's=7
#1's=2
#2's=1

...just 0's, 1's:
sum $= y_g^\top 1$

Conditioning, LFL, and EXP:

- conditional likelihood:

$$\prod_g \frac{f(y_g|\eta_g)}{\sum_\tau f(P_\tau y_g|\eta_g)}$$

- conditional+LFL:

$$\prod_g \frac{f(y_g|X_g\beta)}{\sum_\tau f(P_\tau y_g|X_g\beta)}$$

- conditional+LFL+EXP:

$$\prod_g \frac{\exp(y_g^\top h_0(X_g\beta))}{\sum_\tau \exp(y_g^\top P_\tau^\top h_0(X_g\beta))}$$

(free of h_1, h_2)

- …+linearized:

$$\prod_g \frac{\exp(y_g^\top X_g\beta)}{\sum_\tau \exp(y_g^\top P_\tau^\top X_g\beta)}$$

(h_0 is linear)

e.g. logistic regression:
...unconditionally,

$$\prod_{gi} \frac{e^{y_{gi}(\mu_g + x_{gi}^\top\beta)}}{1+e^{(\mu_g + x_{gi}^\top\beta)}}$$

paired logistic $(n_g = 2)$
conditional on $y_g^\top 1 = 1$,

$$\prod_g \frac{e^{y_{g1}x_{g1}^\top\beta + y_{g2}x_{g2}^\top\beta}}{e^{x_{g1}^\top\beta} + e^{x_{g2}^\top\beta}}$$

$$= \prod_g \frac{1}{1+e^{-(y_{g2}-y_{g1})(x_{g2}-x_{g1})^\top\beta}}$$

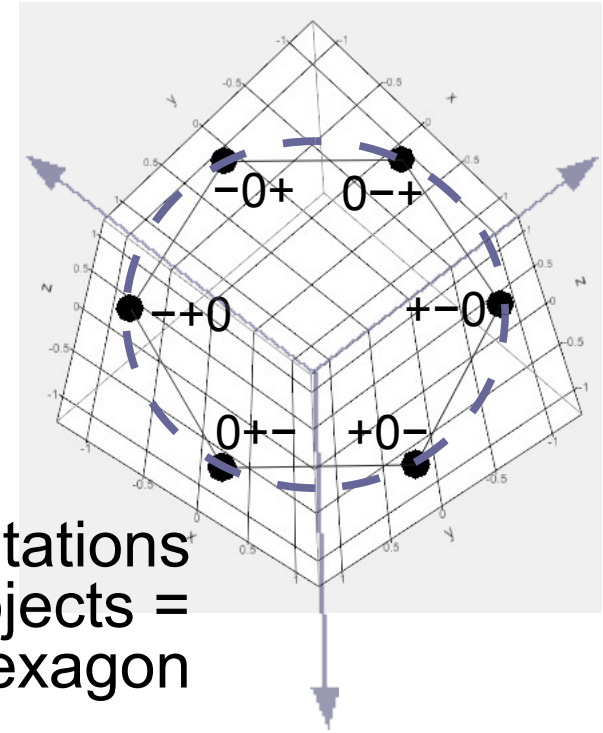
contrast of $g2$
vs $g1$=control

approximation:

- discrete permutation likelihood:

$$e^{L_g(\beta)} = \frac{\exp\left(y_g^{\mathsf{T}} \eta_g\right)}{\sum_\tau \exp\left(y_g^{\mathsf{T}} P_\tau^{\mathsf{T}} \eta_g\right)}$$

$$= \frac{\exp\{y_g^T \eta_g\}}{n_g!\ \text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\} : \tau \in \text{permutations}\}},$$



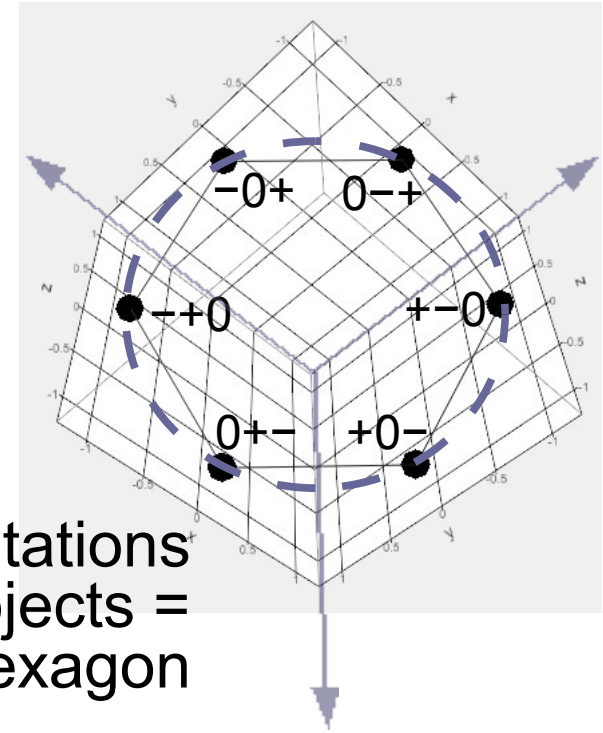
6 permutations of 3 objects = 2-D hexagon

approximation ("spheres"):

- discrete permutation likelihood:

$$e^{L_g(\beta)} = \frac{\exp(y_g^{\mathsf{T}} \eta_g)}{\sum_\tau \exp(y_g^{\mathsf{T}} P_\tau^{\mathsf{T}} \eta_g)}$$

$$= \frac{\exp\{y_g^T \eta_g\}}{n_g!\ \text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\} : \tau \in \text{permutations}\}},$$



6 permutations of 3 objects = 2-D hexagon

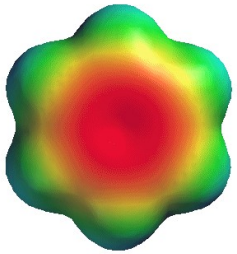
- spherical approximation:

$$\text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\} : \tau \in \text{permutations}\}$$

$$\approx \text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\} : \tau \in \text{rotations}\}$$

$$= \frac{\int_{y^T y = y_g^T y_g} \exp\{y^T \eta_g\}\, dy}{\int_{y^T y = y_g^T y_g} dy}$$

integration over n_g-sphere's surface



benzene: electrostatic potential

graphite STM

von Mises-Fisher distribution

- discrete likelihood:

$$e^{L_g(\beta)} = \frac{\exp\{y_g^T \eta_g\}}{n_g!\ \text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\}:\ \tau \in \text{permutations}\}}$$

wolog $y_g^\mathsf{T} 1 = 0$

$\eta_g^\mathsf{T} 1 = 0$

- spherical:

$$\approx \frac{\exp\{y_g^T \eta_g\}}{n_g!\ \text{ave}\{\exp\{y_g^T P_\tau^T \eta_g\}:\ \tau \in \text{rotations}\}}$$

von Mises-Fisher (Watson & Williams 1956)

$$= \frac{\int_{y^T y = y_g^T y_g} \exp\{y^T \eta_g\}\, dy}{\int_{y^T y = y_g^T y_g} dy} = \frac{\int_{u^T u = 1} \exp\{\kappa_g u^T \mu_g\}\, dy}{\left(\frac{2\pi}{\kappa_g}\right)^{n_g/2} \kappa_g I_{n_g/2-1}(\kappa_g)}$$

$I_\nu(\kappa) \equiv$ modified Bessel function of order ν.

$\kappa_g \equiv \kappa_{g1} \kappa_{g2}$,

unit vector $\mu_g \equiv \eta_g / \kappa_{g2}$,

chunk g standard deviation $\kappa_{g1} \equiv \sqrt{y_g^\mathsf{T} y_g}$

chunk g prediction s.d. $\kappa_{g2} \equiv \sqrt{\eta_g^\mathsf{T} \eta_g}$

likelihood gradient (I)

- vM-F directional mean
  $=\rho_\nu(\kappa)\mu$, μ s.t. $\mu^T\mu=1$ (unit vector)

  $$\rho_\nu(\kappa)=\frac{I_\nu(\kappa)}{I_{\nu-1}(\kappa)} \quad \text{(a scalar)}, \quad \nu=\frac{n_g}{2}$$

- spherical likelihood $$e^{L_g(\beta)}=\frac{\exp\{y_g^T X_g \beta\}}{n_g!\int \exp\{y^T X_g \beta\}\,dy}$$

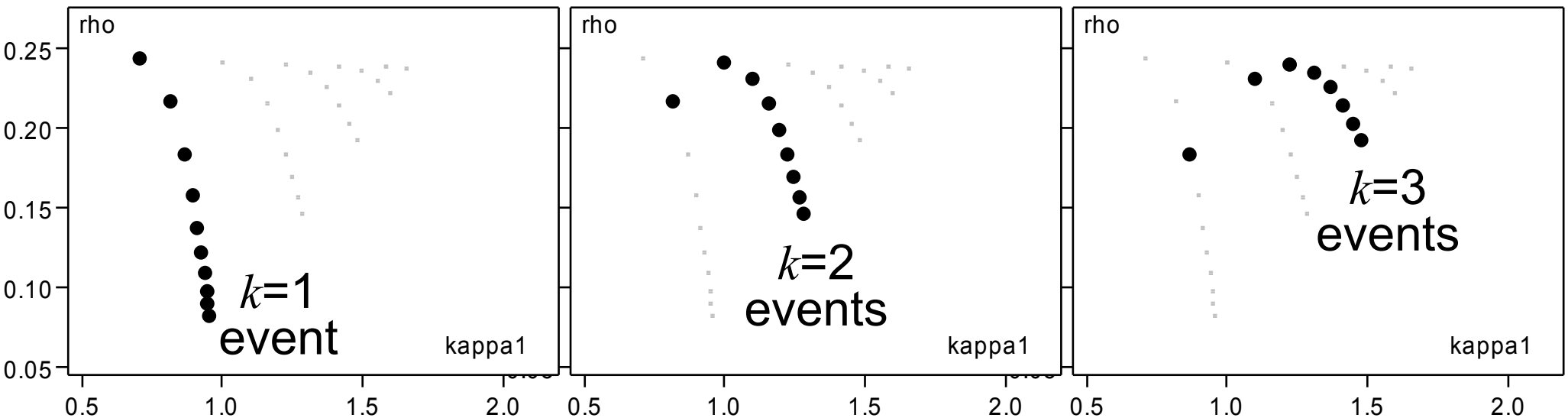
- gradient of log-likelihood:

  $$\nabla_\beta L_g(\beta)=X_g^T\left(y_g-\frac{\kappa_{g1}}{\kappa_{g2}}\rho(\kappa_{g1}\kappa_{g2})X_g\beta\right),$$

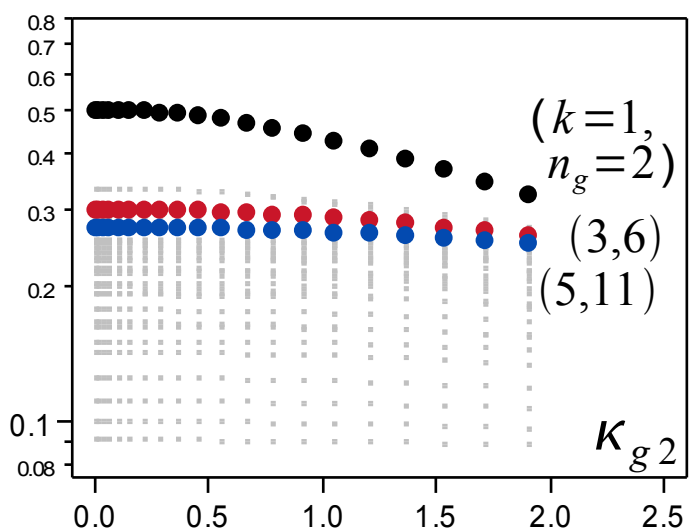
what is $\frac{\kappa_{g1}}{\kappa_{g2}}\rho(\kappa_{g1}\kappa_{g2})$ like?

- as a function of data norm κ_1

~sample variance



- as a function of prediction norm κ_2



~flat, weak decline

likelihood gradient (II)

- gradient of log-likelihood:

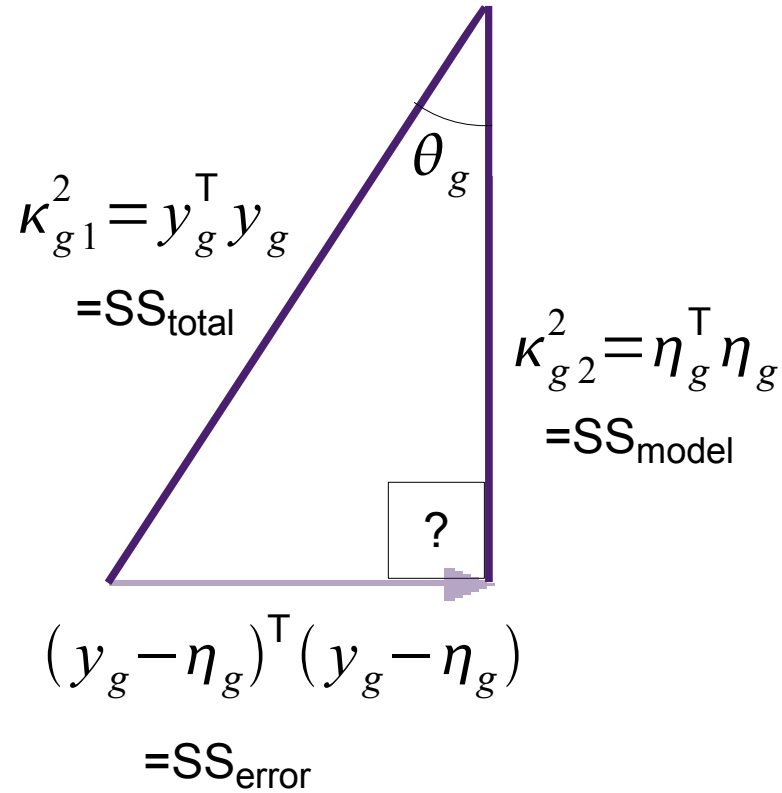
$$\nabla_\beta L_g(\beta) = X_g^T\left(y_g - \frac{\kappa_{g1}}{\kappa_{g2}}\rho(\kappa_{g1}\kappa_{g2})X_g\beta\right)$$

$$= X_g^T\left(y_g - \frac{1}{r_g}\rho(r_g\kappa_{g1}^2)X_g\beta\right),$$

pseudo-correlation $r_g \equiv r_g(\beta)$

s.t. $r_g^2\kappa_{g1}^2(\beta) = \kappa_{g2}^2$.

BTW, $r_g = \cos(\theta_g)$



likelihood gradient (III)

- gradient of log-likelihood:

$$\nabla_\beta L_g(\beta) = X_g^T\left(y_g - \frac{\kappa_{g1}}{\kappa_{g2}}\rho(\kappa_{g1}\kappa_{g2})X_g\beta\right)$$

$$= X_g^T\left(y_g - \frac{1}{r_g}\rho(r_g\kappa_{g1}^2)X_g\beta\right),$$

pseudo-correlation $r_g \equiv r_g(\beta)$

s.t. $r_g^2\kappa_{g1}^2(\beta) = \kappa_{g2}^2$.

Define variance $s_g^2 \equiv y_g^\mathsf{T} y_g / n_g$. Then this holds:

$$\frac{s_g^2}{1+r_g^2 s_g^4 t}\Bigg|_{t=1} \le \frac{\rho(r_g\kappa_{g1}^2)}{r_g} = \frac{\rho(r_g n_g s_g^2)}{r_g} \le \frac{s_g^2}{1+r_g^2 s_g^4 t}\Bigg|_{t=\nu/(\nu+1)}$$

"Poisson" "binomial"

fixed point estimator

$$\nabla_\beta L_g(\beta) = X_g^T\left(y_g - \frac{1}{r_g}\rho(r_g n_g s_g^2) X_g \beta\right),$$

Summing over g, setting equal to zero:

$$\left[\sum_g \frac{\rho(r_g n_g s_g^2)}{r_g(\beta)} X_g^T X_g\right]\beta = \sum_g X_g^T y_g$$

(iteration till $r_g(\beta)$ converges)

(1) Recognizable as Gauss's normal equations.

(2) Weak dependence on $r_g \Rightarrow$ fast convergence.

(3) $\rho(r_g n_g s_g^2)/r_g \approx s_g^2$, downweights low-info chunks.

(4) r_g measures prediction strength within chunk g.

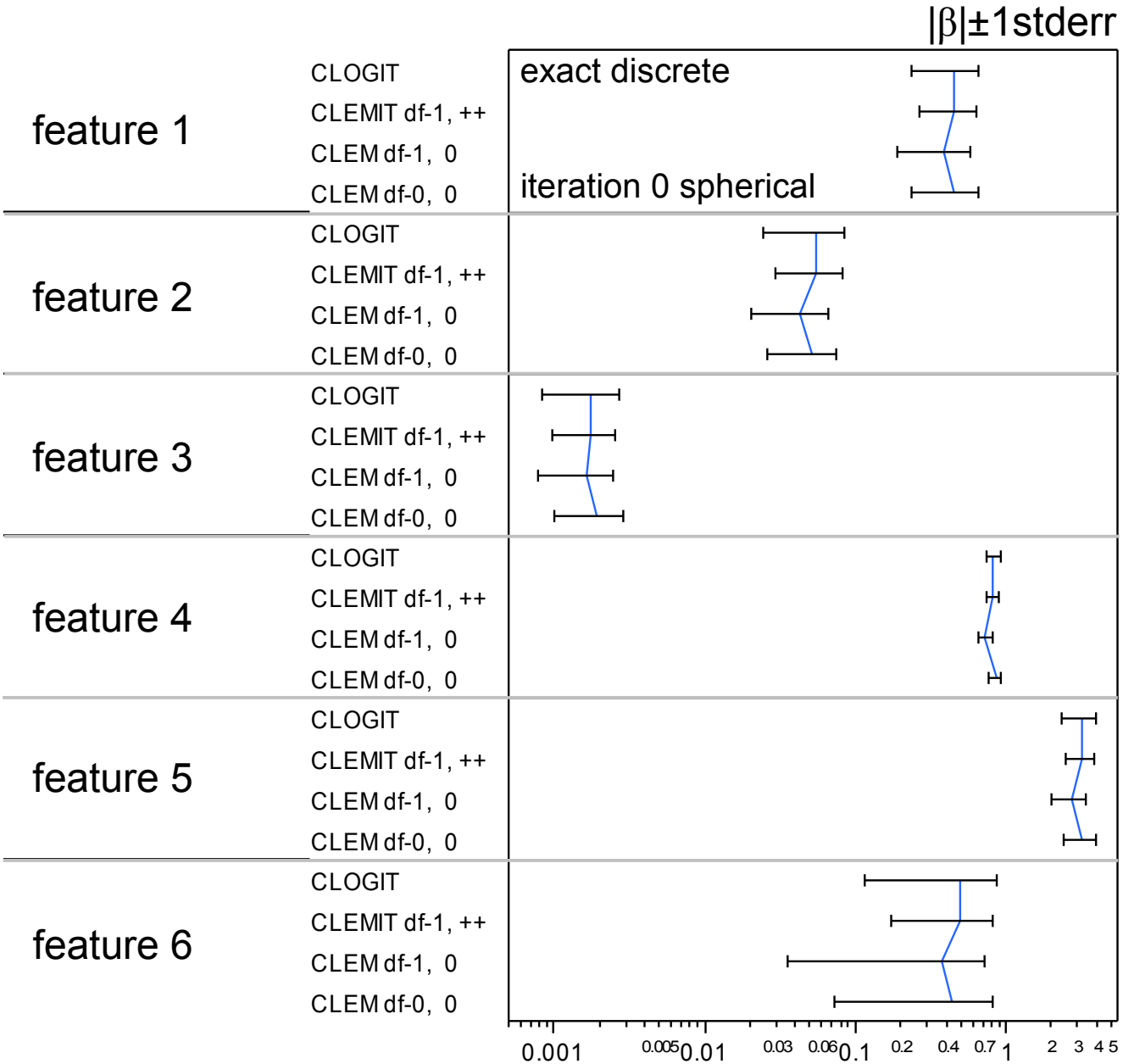
example 1, paired data

data

- 504 preference tasks
- n_g=2: choose "A" or "B"
- features: various quality signals

results

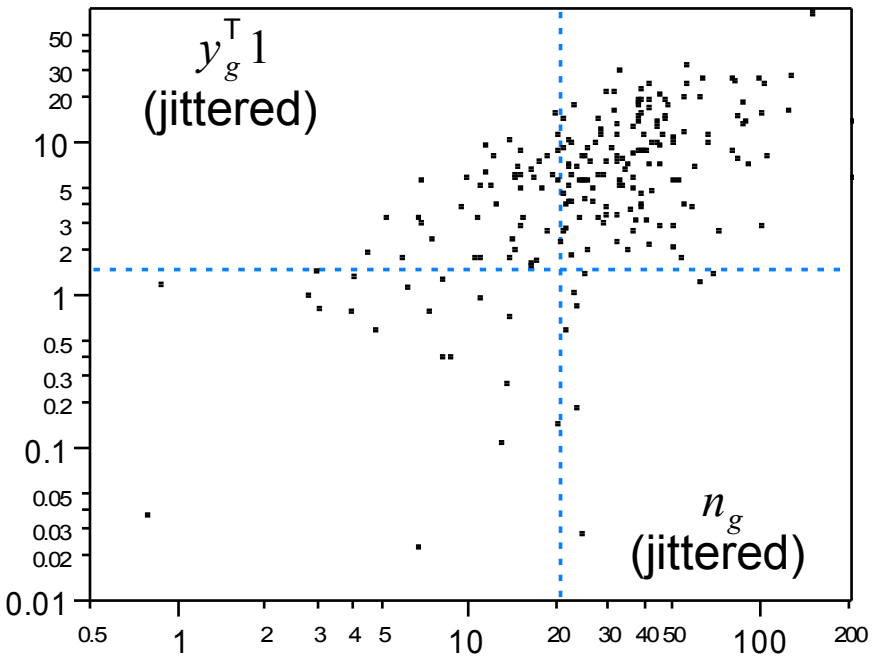
- 1st iteration gets close
- dimension-corrected form converges to same estimate

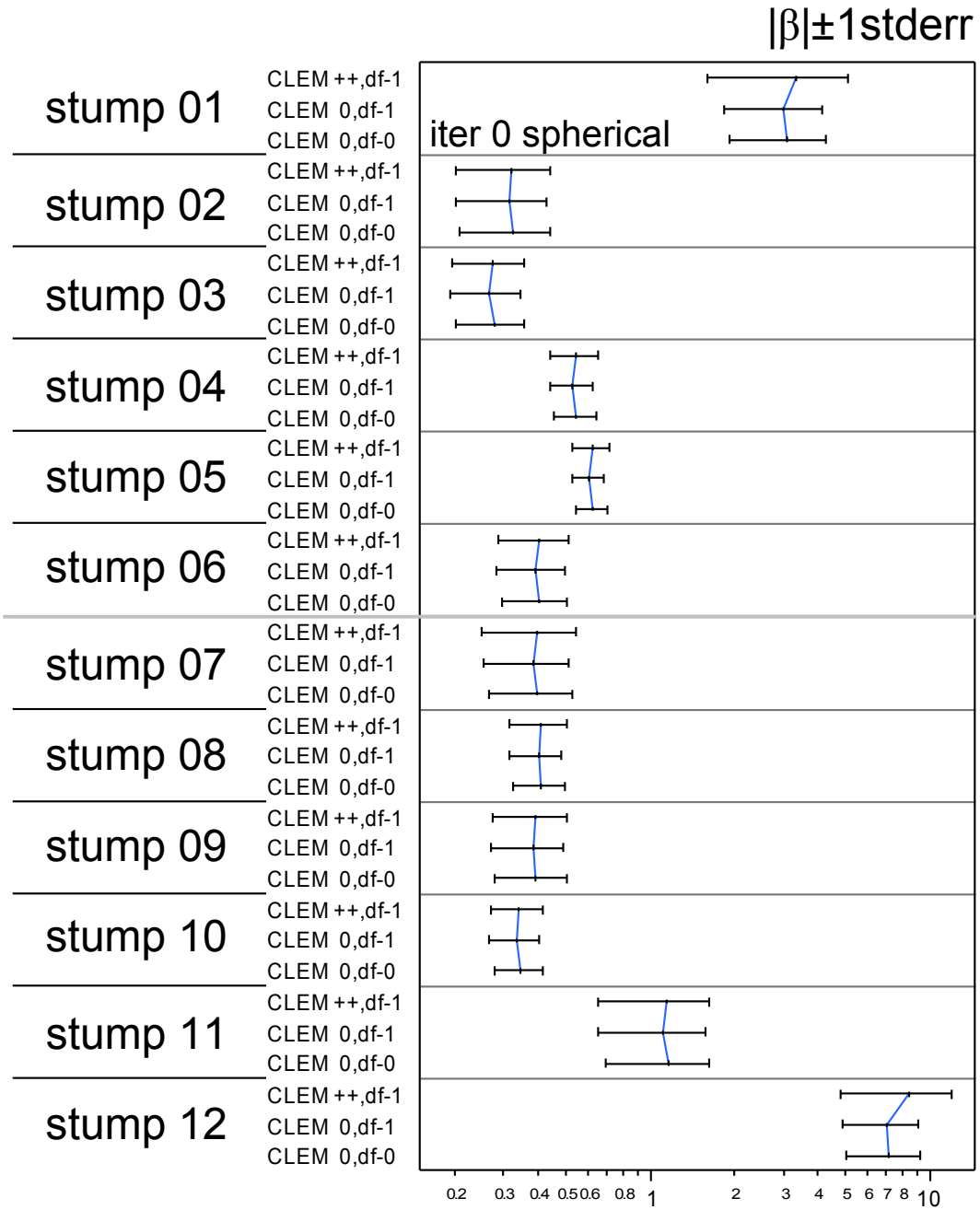


example 2

user panel study

- response: 0-1 self-report
- features: user task statistics,
- … split into 0-1 features ("stumps")
- 7753 tasks for 223 users (chunks),
- median=27 tasks/user





something similar (I): IRLS

iteratively re-weighted least squares

- e.g unchunked logistic $\Pr\{y_i=1|x_i\}=\frac{\exp\{x_i^\mathsf{T}\beta\}}{1+\exp\{x_i^\mathsf{T}\beta\}}\equiv\mu_i$
- estimate for iteration k

$X^\mathsf{T}D_{w[k]}X\hat{\beta}[k]=X^\mathsf{T}D_{w[k]}y$, where $D_{w[k]}=\mathrm{diag}(w[k])$

s.t. $w_i[k]=\mu_i[k](1-\mu_i[k])$, and $\mu_i[k]\equiv\frac{\exp\{x_i^\mathsf{T}\hat{\beta}[k-1]\}}{1+\exp\{x_i^\mathsf{T}\hat{\beta}[k-1]\}}$

weights	IRLS	spherical estimator
X-weight	$\sqrt{w_i[k]}$	$\sqrt{\rho_g/r_g}$
y-weight	$\sqrt{w_i[k]}$	$\sqrt{r_g/\rho_g}$
meaning	obs i theoretical variance	within chunk g sample variance
iteration	w_i strong $\mathrm{F}^\mathrm{n}(\beta)$	$\frac{\rho_g}{r_g}$ weak $\mathrm{F}^\mathrm{n}(\beta)$ $\Rightarrow$ faster convergence

something similar (II): analysis of R^2_{within}

N-D plot:

deconstructs global R^2

sort by $y_g^T y_g$ ascending

plot model-info cumsum:

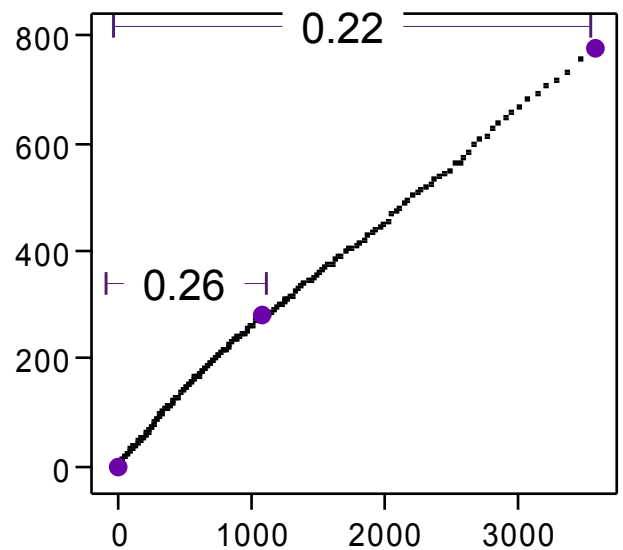
$$N[G] \equiv N[G-1] + \eta_G^T \eta_G$$

vs response-info cumsum:

$$D[G] \equiv D[G-1] + y_G^T y_G$$

cumsum numerators of r_g^2
=model-info

$$N[G] = N[G-1] + \sum_i (\eta_{gi} - \bar{\eta}_g)^2$$



cumsum denominators of r_g^2
=response-info

$$D[G] = D[G-1] + \sum_i (y_{gi} - \bar{y}_g)^2$$

properties:

- slope-from-origin
- local slopes
- reversible sums

- better fit within lower-info chunks

something similar (III): marginal regression

marginal regression: $\hat{\alpha}_{\text{MR}} = [\text{diag}(X^{\mathsf{T}} X)^{-1}] X^{\mathsf{T}} y$ so $\propto X^{\mathsf{T}} y$

- better scaling in J $\hat{\beta}_{\text{MR}} = \hat{\alpha} \times 1\{|(\hat{\alpha})| \geq \theta\}$, for threshold θ

- spherical version:

$$\hat{\alpha}_{\text{sphere}} = [\sum_g \frac{\rho_g}{r_g} \text{diag}(X_g^{\mathsf{T}} X_g)]^{-1} \sum_g X_g^{\mathsf{T}} y_g$$

$$\hat{\beta}_{\text{sphere}} = \hat{\alpha}_{\text{sphere}} \times 1\{|\hat{\alpha}_{\text{sphere}}| \geq \theta\}$$

	marginal regression	spherical estimator
response y	continuous	continuous (Gaussian); discrete (Poisson, logistic)
conditioning X	centered $1^{\mathsf{T}} y = 0;\ 1^{\mathsf{T}} X = 0$	centered within chunk $1^{\mathsf{T}} y_g = 0;\ 1^{\mathsf{T}} X_g = 0$
iteration	not obvious	(optional) update r_g

… marginal regression++

initialization:

$$z_{g0} \equiv 1$$
$$r_{g0} \leftarrow \text{some } r_0, 0 < r_0 < 1$$

iterate k=0, 1, 2, ...

$$u_{gk} \equiv z_{gk} / \| z_{gk} \|$$
$$M_{gk} \equiv [I - u_{gk} u_{gk}^{\mathsf{T}}]$$
$$y_g \leftarrow M_{gk} y_g$$
$$X_g \leftarrow M_{gk} X_g$$

residuals

marginal regression

$$\hat{\alpha}_k = [\sum_g \frac{\rho_g}{r_{gk}} \text{diag}(X_g^{\mathsf{T}} X_g)]^{-1} \sum_g X_g^{\mathsf{T}} y_g$$
$$\hat{\beta}_k = \hat{\alpha}_k \times 1\{|\hat{\alpha}_k| \geq \theta\}$$

$$z_{gk} \equiv X_g \hat{\beta}_k$$
$$r_{g(k+1)}^2 \leftarrow \frac{z_{gk}^{\mathsf{T}} z_{gk}}{y_{gk}^{\mathsf{T}} y_{gk}}$$

update fit & info score

something necessary (IV): m-stage models

predictors

0: conditional estimate: β

...from spherical approximation

$X_{g\,(n_g \times J)}$

$\Downarrow$

1: scale estimate: σ

- for Gaussian,
- for attenuation correction
- for scaling in J

$z_{g\,(n_g \times 1)} \equiv X_g \beta$

$\Downarrow$

2: intercept estimate: μ

- for Poisson,
- for logistic

$z_g \equiv X_g \beta$ or $X_{g\,(n_g \times 1)} (\beta / \sigma^2)$

two 1-D calculations

$\Downarrow$

predictive model $f(\mu + x_i^{\mathsf{T}} \beta)$

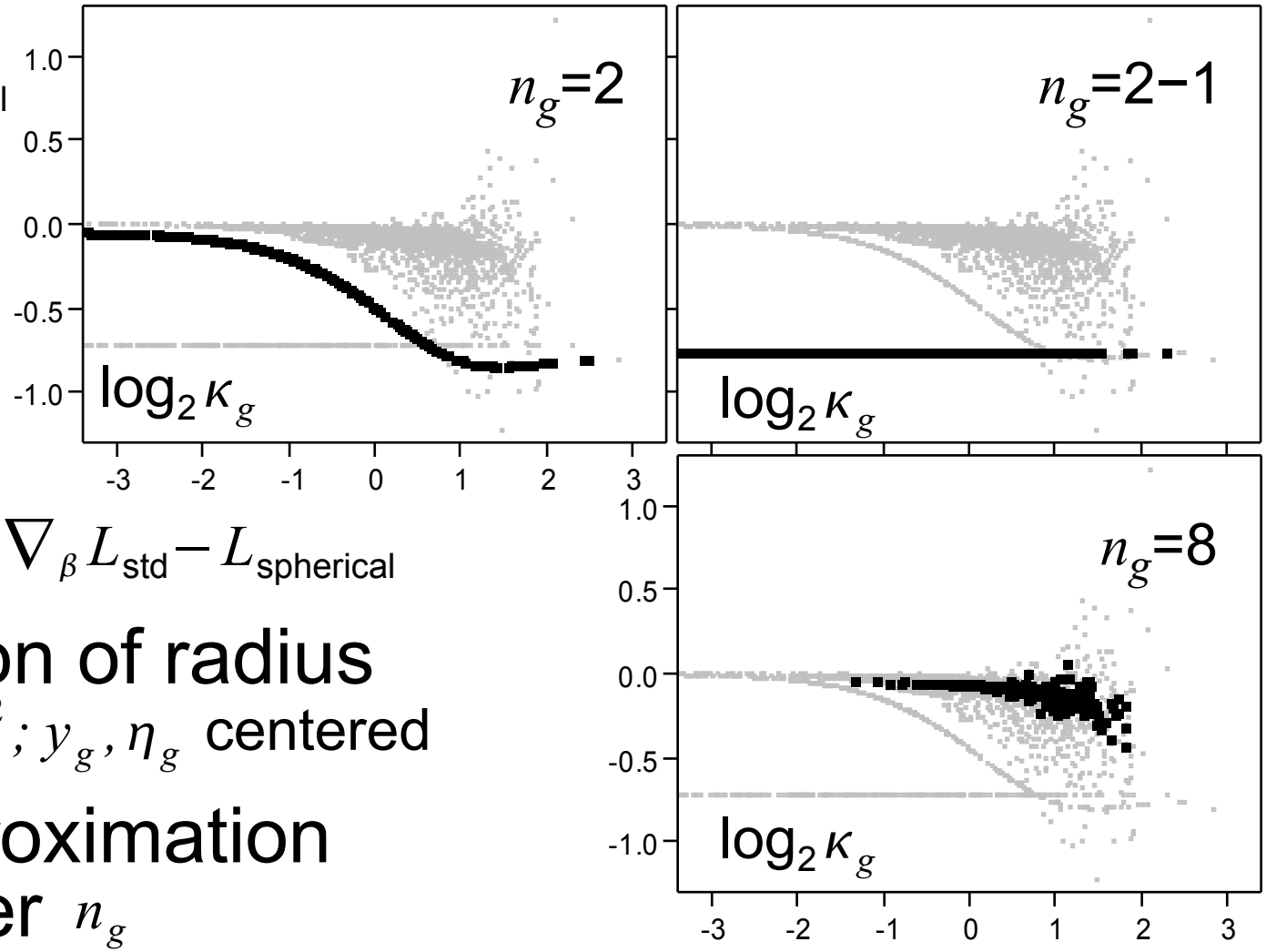
summary

- framework:
  (1) linearized predictor space (LFL)
  (2) generalized linear models ("links", EXP)
  (3) stratified samples ("chunks")
- problem $\Rightarrow$ approach:
  (1) nuisance effects, attenuation $\Rightarrow$ conditional models
  (2) combinatorial denominator $\Rightarrow$ spherical approx ("spheres")
  (3) need precise algorithm $\Rightarrow$ fixed point iteration
  (4) need parallel algorithm $\Rightarrow$ marginal model++ approach
- some interesting consequences:
  (1) fast algorithm, especially as $n_g \rightarrow$ big
  (2) s_g measures a chunk's potential information
  (3) r_g measures a chunk's prediction satisfaction, a model diagnostic.

- some things not talked about enough:
  (1) scaling for large, large J
  (2) attenuation, intercept effects, and the n_g=everything case
  (3) models in series ("stages," "layers")
  (4) model deconstruction and inference

numerical analysis (I)

$$\frac{\partial}{\partial \|\eta_g\|} L_{\text{std}} - L_{\text{spherical}}$$



<u>critical issue</u>:

- non-constant $\nabla_\beta L_{\text{std}} - L_{\text{spherical}}$
- esp. as function of radius
  $\kappa_g = [y_g^\mathsf{T} y_g \eta_g^\mathsf{T} \eta_g]^{1/2}$; y_g, η_g centered
- spherical approximation better for larger n_g
- … for smaller κ_g
  e.g. rare events: $y_g^\mathsf{T} y_g$ small
  e.g. weak learners: $\eta_g^\mathsf{T} \eta_g = \|\eta_g\|^2$ small

numerical analysis (II)

$$\frac{1}{r}\rho_\nu(rns^2) = \frac{1}{r}\frac{I_\nu(rns^2)}{I_{\nu-1}(rns^2)} = \frac{1}{r}\frac{I_\nu(2r\nu s^2)}{I_{\nu-1}(2r\nu s^2)} = \frac{1}{r}\frac{\sum_k \frac{(x/2)^{\nu+2k}}{k!\,\Gamma(\nu+k+1)}}{\sum_k \frac{(x/2)^{\nu-1+2k}}{k!\,\Gamma(\nu+k)}}\Bigg|_{x=2r\nu s^2}$$

"~a Bessel distribution"

$$= \frac{x}{2r}\frac{\sum_k \frac{(x/2)^{2k}}{k!\,\Gamma(\nu+k+1)}}{\sum_k \frac{(x/2)^{2k}}{k!\,\Gamma(\nu+k)}} = s^2\frac{\sum_k \frac{(x/2)^{2k}\nu}{k!\,\Gamma(\nu+k+1)}}{\sum_k \frac{(x/2)^{2k}(\nu+k)}{k!\,\Gamma(\nu+k)(\nu+k)}} = s^2\frac{\sum_k \frac{(x/2)^{2k}\nu}{k!\,\Gamma(\nu+k+1)}}{\sum_k \frac{(x/2)^{2k}(\nu+k)}{k!\,\Gamma(\nu+k+1)}} = s^2\frac{\nu\sum_k p(k|x,\nu)}{\sum_k (\nu+k)\,p(k|x,\nu)}$$

$$= s^2\frac{\nu}{\nu+E\{k|x,\nu\}}$$

bounds on $E\{k|x,\nu\}$

lower (binomial): $\nu r^2 s^4 t$,
where $t=\nu/(1+\nu)$

upper (Poisson): $\nu r^2 s^4$

bounds on ρ_ν/r:

upper: $\frac{s^2}{1+r^2 s^4 t}$

lower: $\frac{s^2}{1+r^2 s^4}$ ← useful for 1st iteration

so ρ_ν/r re-weights strata,
mainly through s^2,
weakly, secondarily via r^2

a few references

- K. V. Mardia and P. E. Jupp, *Directional Statistics*, 2nd ed. John Wiley & Sons, 2000.
- P. McCullough and J. A. Nelder, *Generalized Linear Models,* 2nd edition, John Wiley & Sons, 1999.
- D.W. Hosmer, Jr. and S. Lemeshow, *Applied Logistic Regression,* 2nd edition, John Wiley & Sons, 2000.
- S. M. Stigler, *The History of Statistics: the measurement of uncertainty before 1900.* Harvard University Press, 1986.
- S. M. Plis, T. Lane, V. D. Calhoun, "Permutations as Angular Data: Efficient Inference in Factorial Spaces," ICDM, pp.403-410, 2010 IEEE International Conference on Data Mining, 2010.