Infinite mixtures of beta regression models for bounded-domain variables

Ioannis Kosmidis

ioannis.kosmidis@warwick.ac.uk

ikosmidis.com

github.com/ikosmidis

twitter.com/IKosmidis_

Reader in Data Science & Turing Fellow

University of Warwick & The Alan Turing Institute

14 April 2019

CRoNoS & MDA 2019, Limassol, Cyprus

Joint work with Achim Zeileis

Regression setting

Response

$y_i \in [0, 1]$

Explanatory variables

$x_i \in M \subset \Re^p$

Data

$(y_1, x_1^\top), \ldots, (y_n, x_n^\top)$

Aim

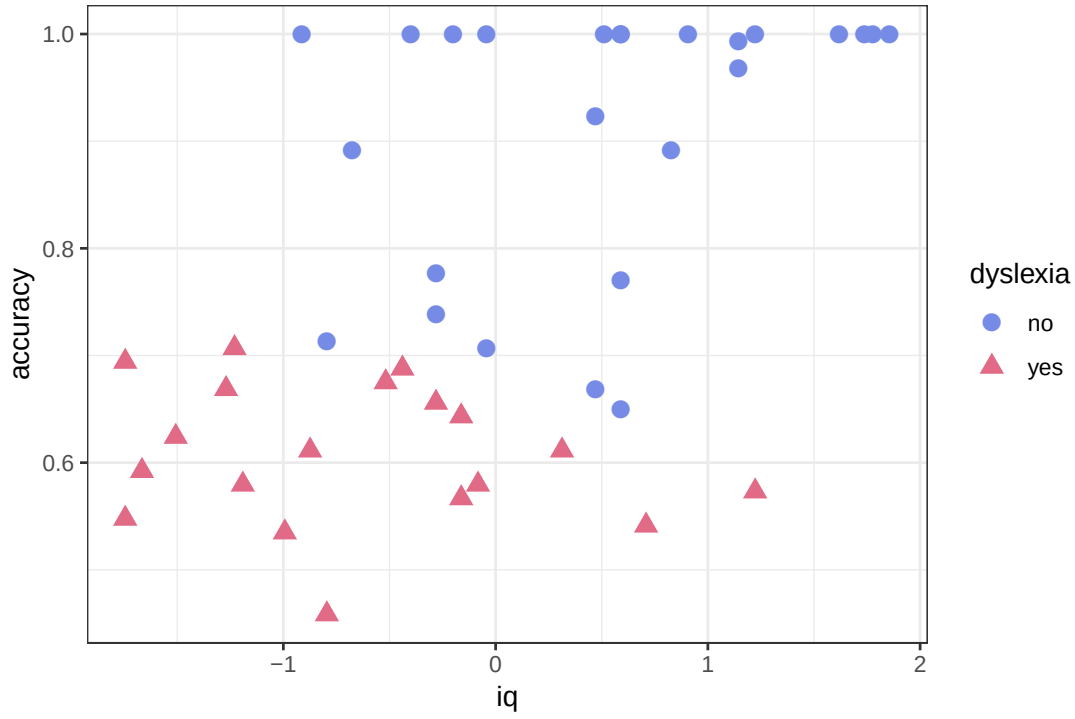
Use x to explain the variability in y

Common characteristics of bounded responses

Location and scale distributional properties are linked

Skewness and heteroscedasticity across explanatory settings

Reading accuracy



Reading accuracy for 44 nondyslexic and dyslexic Australian children[1]

Question

Is the different in the reading performance of dyslexic and non-dyslexic children the result of differences in general cognitive ability?

[1] data from Smithson and Verkuilen (2006)

Beta distribution

The density of the beta distribution has the form

$$f_{(b)}(y \mid \mu, \phi) = \frac{y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}}{B(\mu\phi, (1-\mu)\phi)} I(0 < y < 1)$$

where $0 < \mu < 1$, $\phi > 0$, $B(p, q)$ $(q, p > 0)$ is the beta function[2], and $I(A) = 1$ if A holds and 0, otherwise

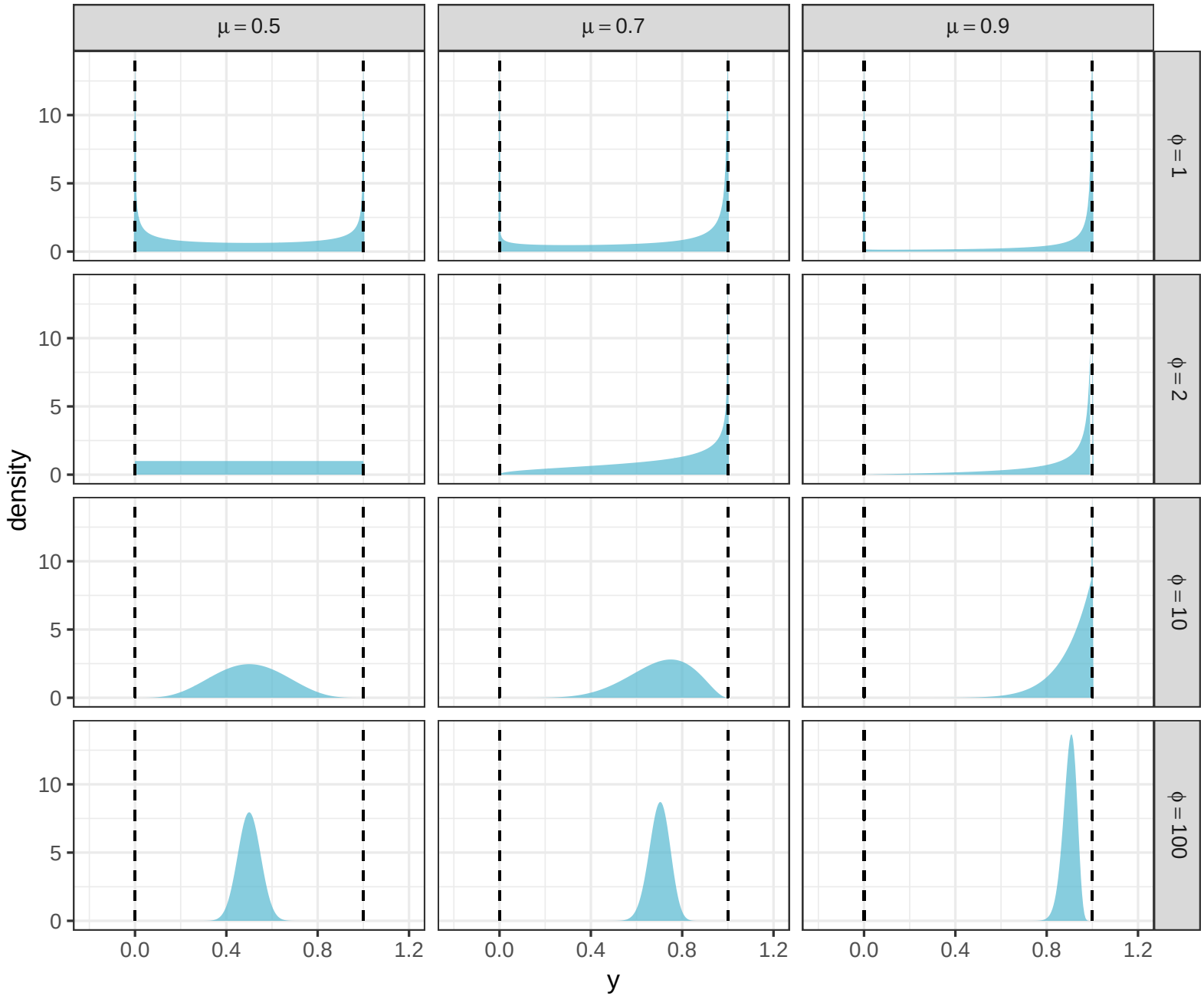
In this parameterization

$$E(Z \mid \mu, \phi) = \mu$$
$$\mathrm{Var}(Z \mid \mu, \phi) = \frac{\mu(1-\mu)}{1+\phi}$$

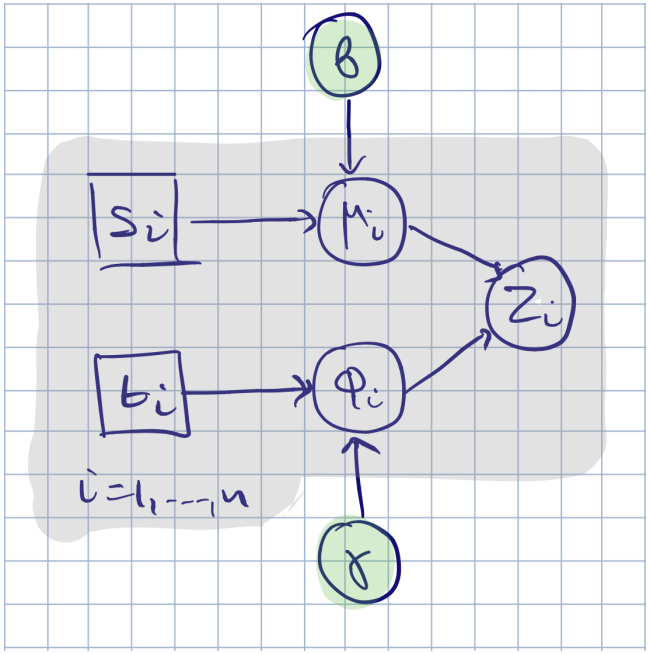
so ϕ is a *precision* parameter

[2] see, Abramowitz and Stegun (1964, §6.2.1)

Beta density



Beta regression



Responses are realizations of independent beta-distributed random variables $Y_1, \ldots, Y_n$ with means $\mu_1, \ldots, \mu_n$ and precisions $\phi_1, \ldots, \phi_n$, respectively

The mean and precision parameters are linked to explanatory variables as

$$g_1(\mu_i) = s_i^\top \beta$$
$$g_2(\phi_i) = t_i^\top \gamma$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ and $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$, $g_1(\cdot)$ and $g_2(\cdot)$ are monotone link functions, and s_i and t_i may share components

Beta regression likelihood

$$\prod_{i=1}^{n} f_{(b)}\left(y_i \mid g_1^{-1}(s_i^\top \beta), g_2^{-1}(t_i^\top \gamma)\right)$$

where $g_1^{-1}(\cdot)$ and $g_2^{-1}(\cdot)$ are the inverse link functions for the mean and the precision

R package `betareg` estimates β and γ using maximum likelihood (and bias-reducing adjusted equations)[3]

```
betareg(y ~ dyslexia * iq, data = ReadingSkills)

## Error in betareg(y ~ dyslexia * iq, data = ReadingSkills):  invalid
dependent variable, all observations must be in (0, 1)
```

If **at least one** of the observed responses is on the boundary, the Beta regression likelihood is zero, infinite or undetermined regardless of n

[3] see Grün, Kosmidis, and Zeileis (2012) for a range of modelling strategies and estimation methods based on beta regression

[4] using `betareg` version 3.1-0

Dealing with boundary observations

Response adjustment[5]

Hurdle models[6]

Two-limit Tobit regression[7]

[5] Smithson and Verkuilen (2006)

[6] see, for example, Cook et al. (2008), Calabrese (2012), Ospina and Ferrari (2012)

[7] see, for example, Maddala (1983, Section 6.7)

Extending the support of the beta distribution

Four-parameter beta distribution[8]

If Z has a beta distribution, then $Y = u_1 + (u_2 - u_1)Z$, $u_2 > u_1$ has the **four-parameter beta distribution** with density

$$f_{(b4)}(y \mid \mu, \phi, u_1, u_2) = \frac{f_{(b)}\left(\frac{y-u_1}{u_2-u_1} \mid \mu, \phi\right)}{u_2 - u_1}$$

The shape properties of $f_{(b4)}(y \mid \mu, \phi, u_1, u_2)$ are those of the beta distribution but with support (u_1, u_2)

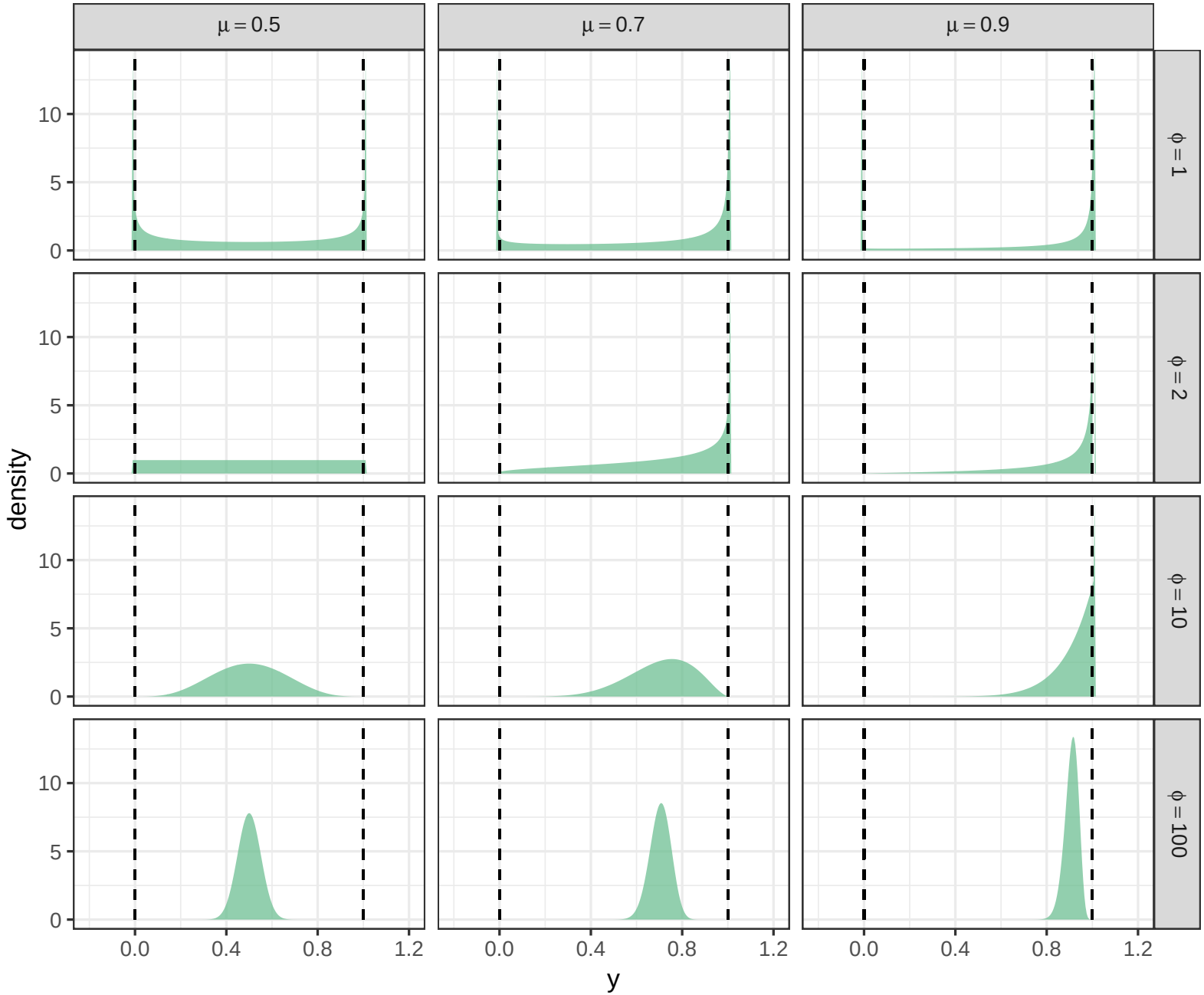
Restricted version

$$f_{(r)}(y \mid \mu, \phi, u) = f_{(b4)}(y \mid \mu, \phi, -u, 1 + u)$$

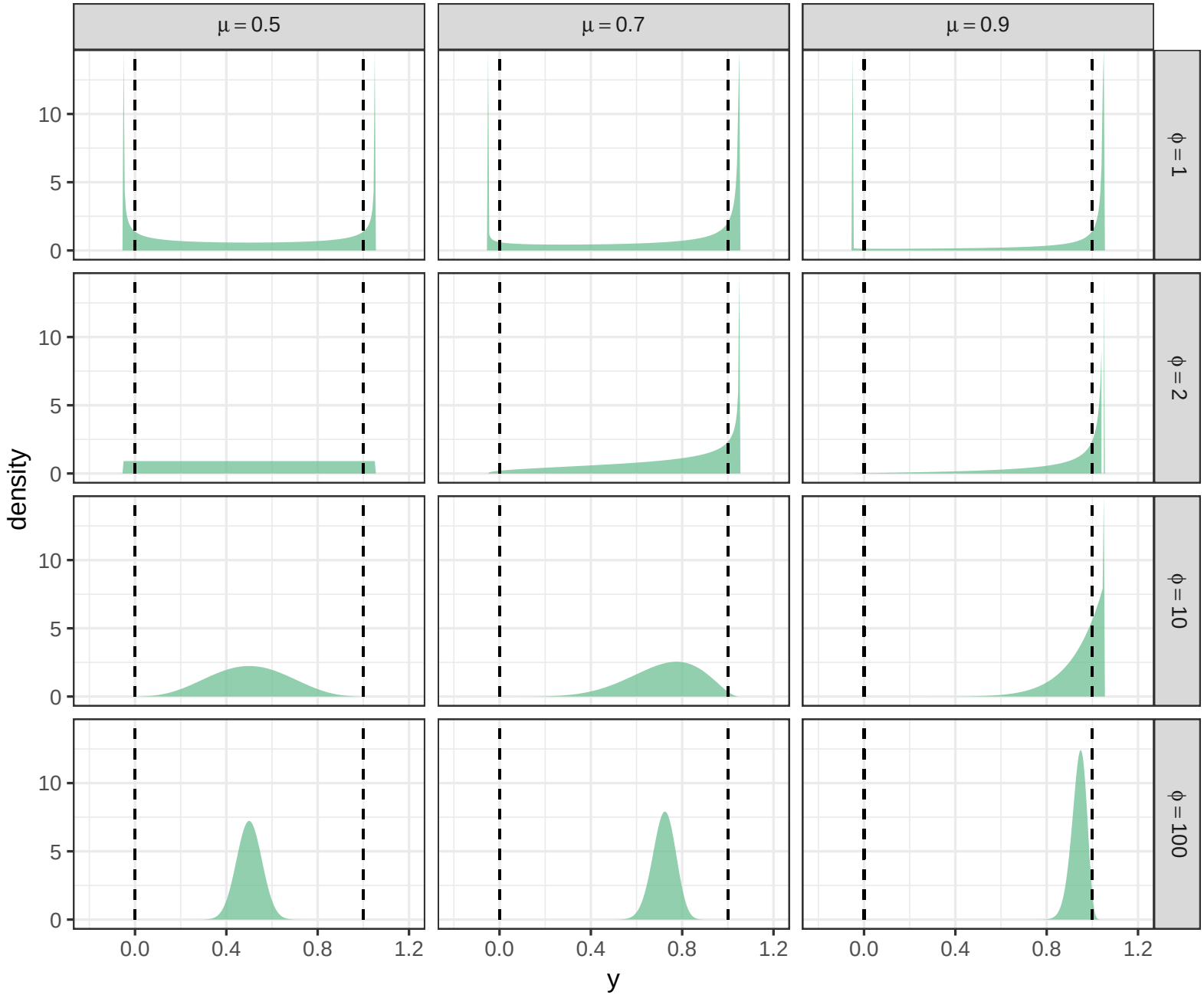
with $u > 0$

[8] see, Pearson (1895) where the unnormalized version of $F_{(b4)}$ is introduced

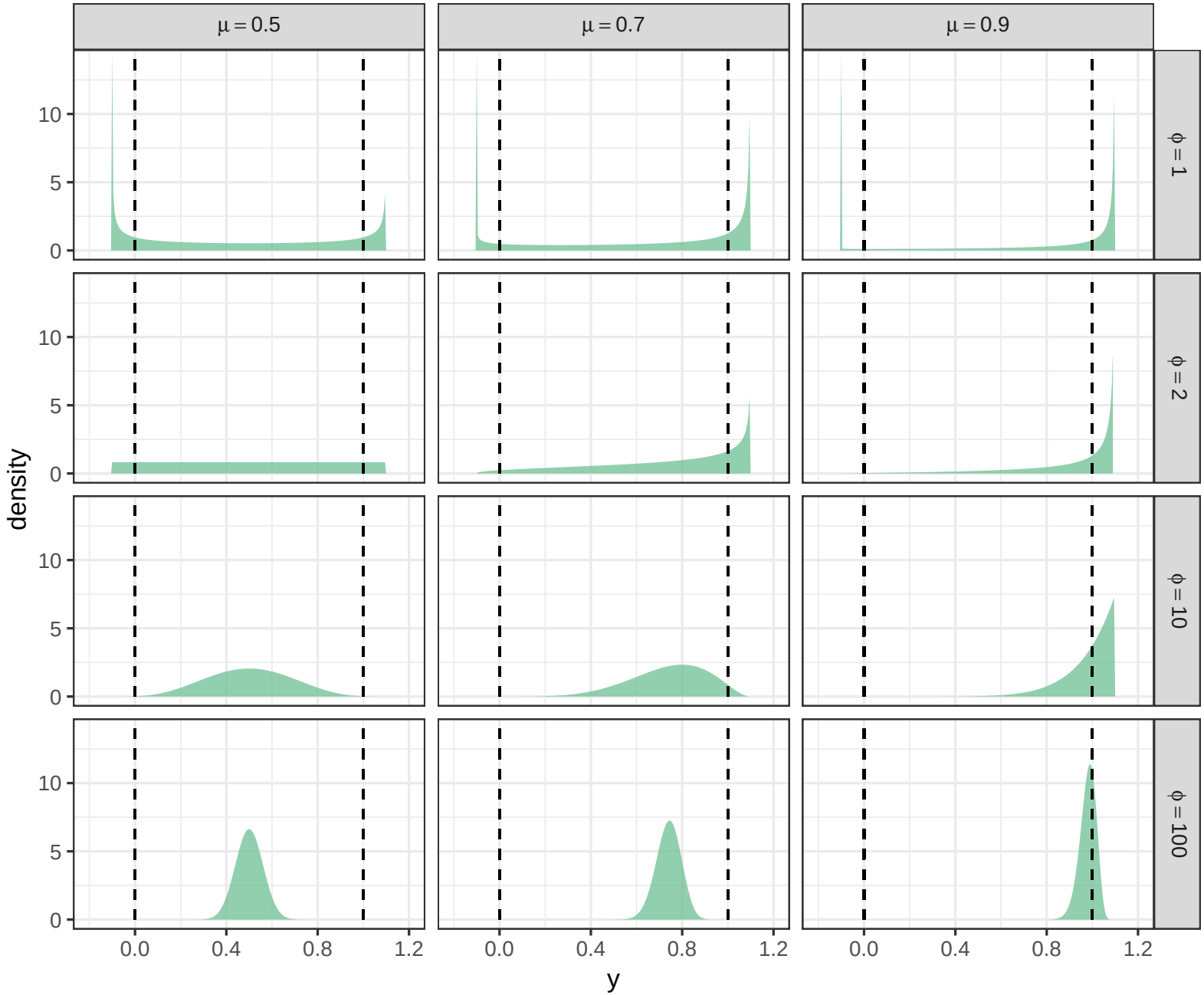
Restricted four parameter beta density ($u = 0.01$)



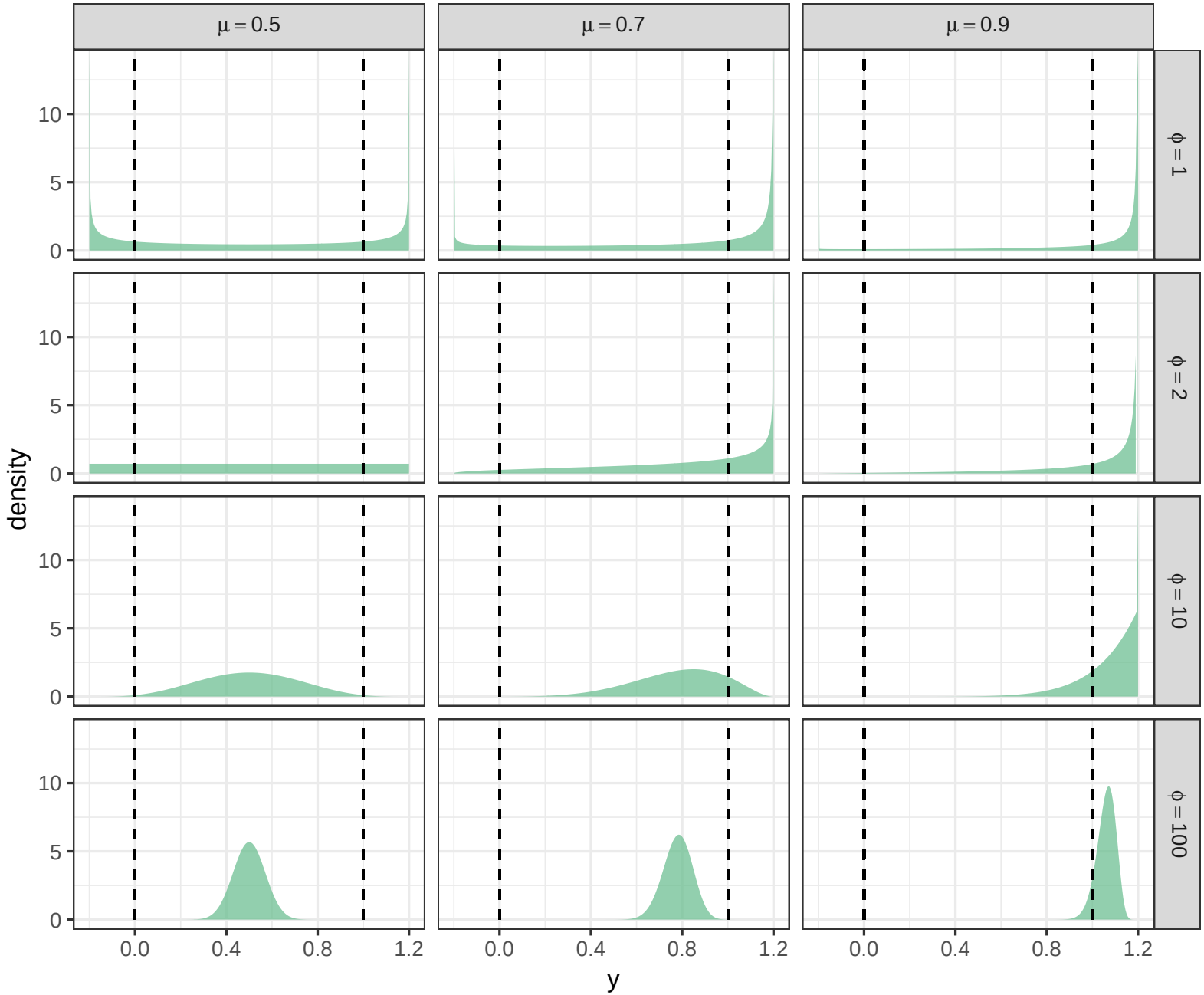
Restricted four parameter beta density ($u = 0.05$)



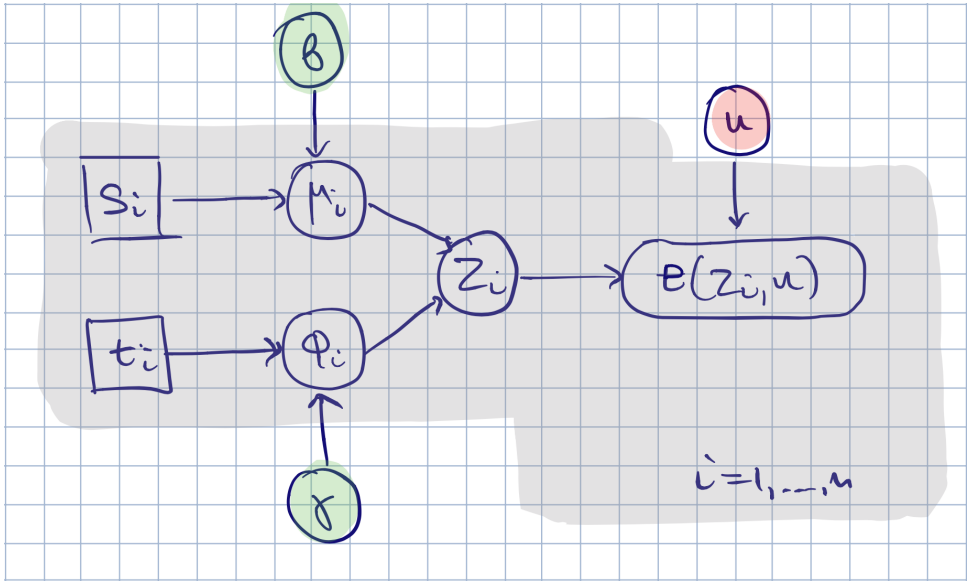
Restricted four parameter beta density ($u = 0.1$)



Restricted four parameter beta density ($u = 0.2$)



Response adjustment (Smithson and Verkuilen, 2006)



$$e(z, u) = -u + (1 + 2u)z$$

Estimation of u is hard

Smithson and Verkuilen (2006) propose to use an ad-hoc value for u

$$f_{(r)}\left(y_i \,|\, g_1^{-1}(s_i^\top \beta), g_2^{-1}(t_i^\top \gamma), \frac{1}{2(n-1)}\right)$$

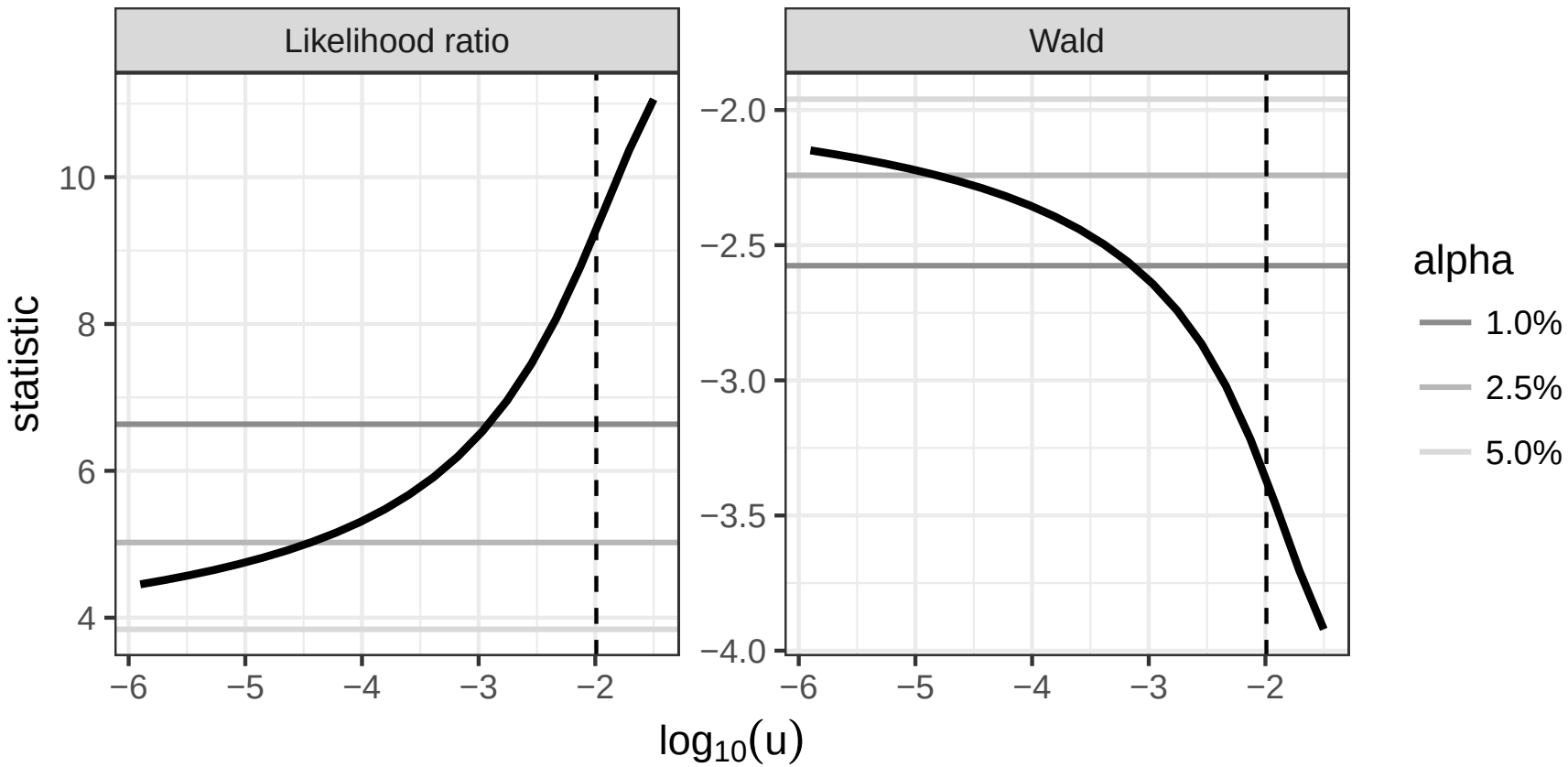
Pros: Parsimony, flexibility, estimation and inference are "borrowed" from beta regression; simply transform the responses to $(y + u)/(1 + 2u)$

Cons: Ad-hoc choice of u; not a generative model

[9] see, for example, Carnahan (1989) and Wang (2005) for extensive discussion on the issues with the estimation of u_1 and u_2 and resolutions in a Bayesian context, respectively.

Reading accuracy

Testing for interaction between `iq` and `dyslexia`



Hurdle models

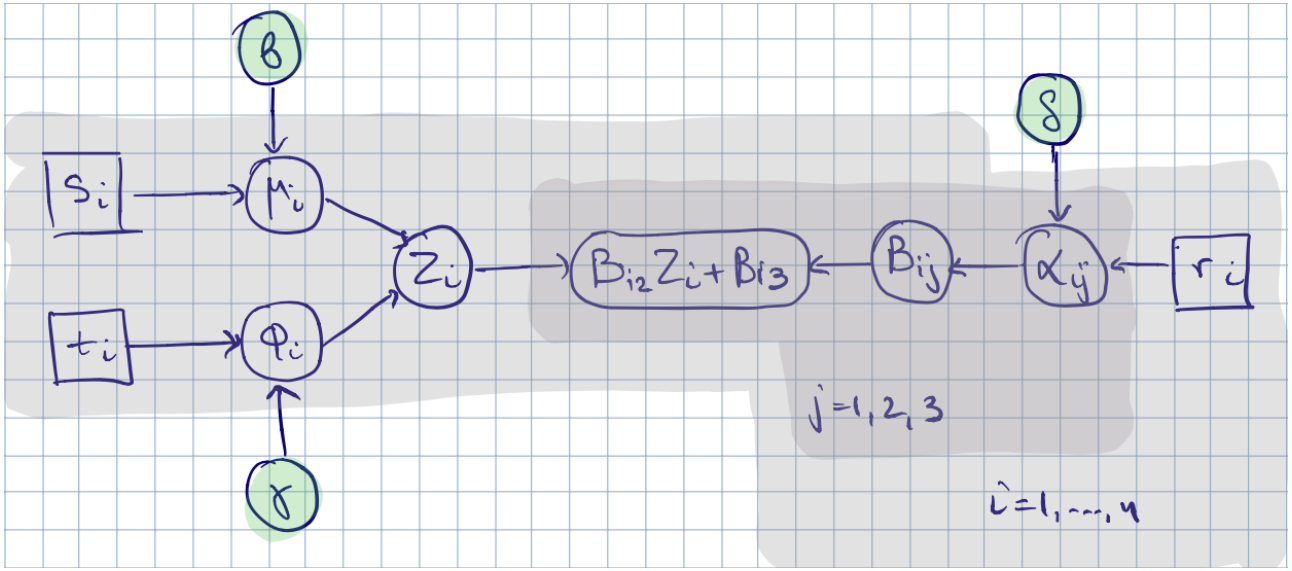
Hurdle specification (aka "Zero-or-one inflated Beta")[10]

$$f_{(zoib)}(y \mid \mu, \phi, \alpha_1, \alpha_2) = \begin{cases} \alpha_1\,, & y = 0 \\ \alpha_2 f_{(b)}(y \mid \mu, \phi)\,, & y \in (0,1) \\ \alpha_3\,, & y = 1 \end{cases}$$

with $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and $\alpha_j \geq 0$ $(j = 1, 2, 3)$

[10] see, Cook et al. (2008), Calabrese (2012), Ospina and Ferrari (2012)

Hurdle models



$B_{ij} \in \{0, 1\}$
$(j = 1, 2, 3)$
$\sum_j B_{ij} = 1$

Multinomial probabilities $\{\alpha_{ij}\}$ modelled in terms of r_i

Pros

Flexibility

Likelihood is **separable**; estimate δ using a multinomial regression model[11] and estimate β and γ with a beta regression on non-boundary observations

Cons

Not as parsimonious as Beta regression; model selection is hard

[11] e.g. using `multinom` or `polr` from the MASS **R** package

Two-limit tobit regression

Censored latent variable

$$Y_i^* \sim \mathsf{N}(x_i^\top \beta, \sigma^2)$$

$$Y_i = \max(\min(Y_i^*, 1), 0)$$

Pros

Parsimony; easy estimation and inference

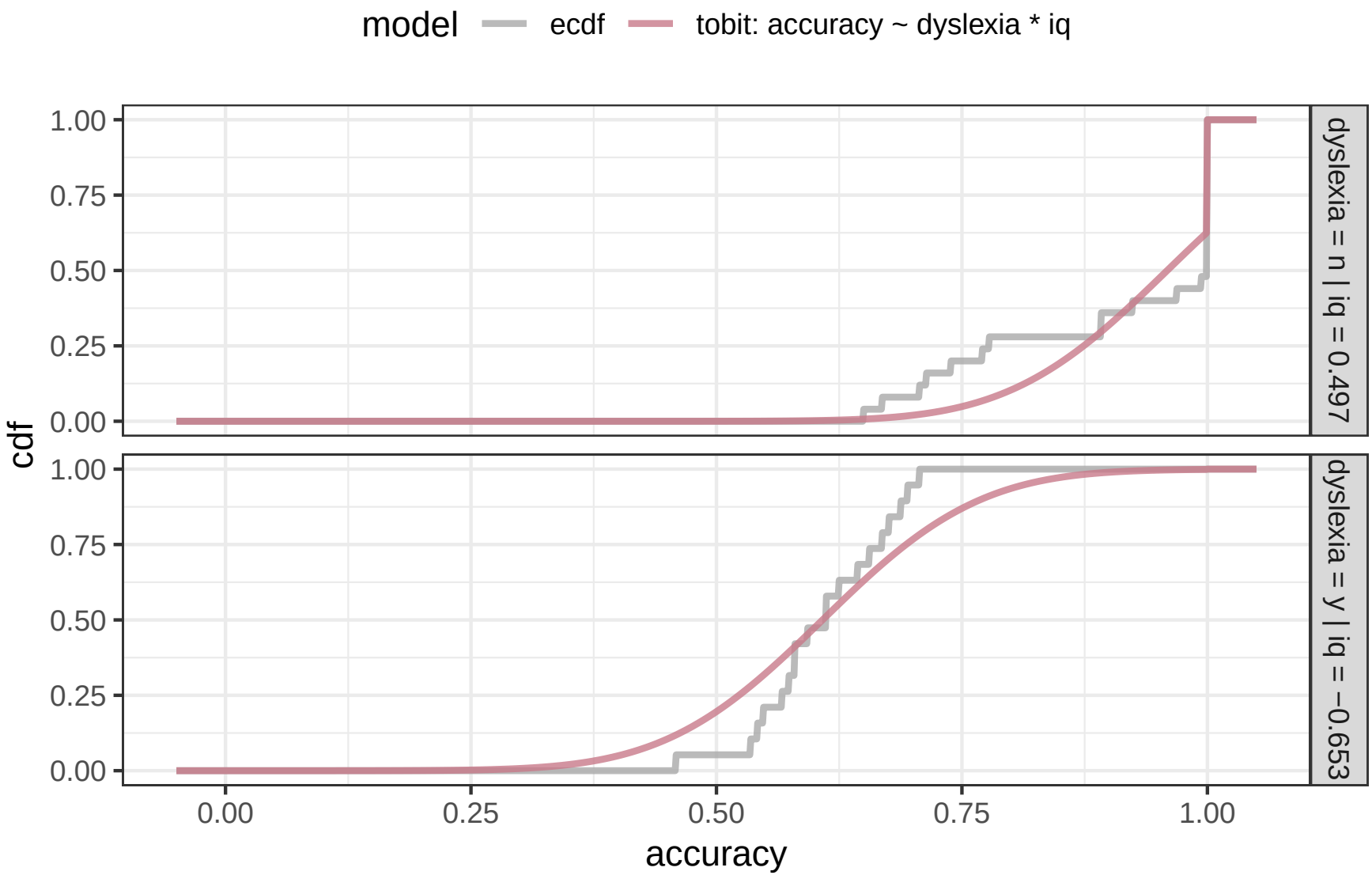
Cons

For non-boundary observations

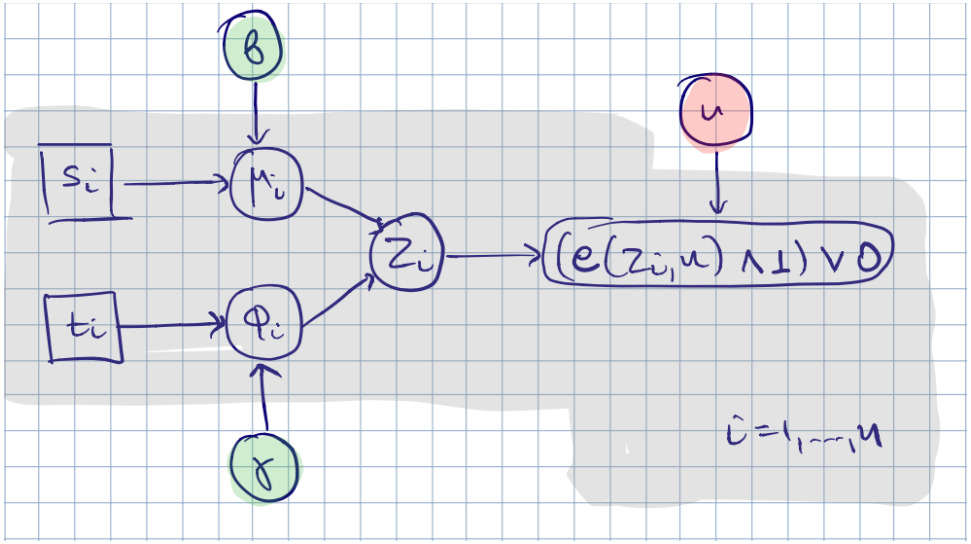
$$\text{two-limit tobit} \sim \text{Normal regression model}$$

[12] see, for example, Maddala (1983, Section 6.7)

Reading accuracy



Censored four-parameter beta (CB4) regression



Censored latent variable

$$Y_i^* \sim \text{RestrictedBeta4}(\mu_i, \phi_i, u)$$

$$Y_i = \max(\min(Y_i^*, 1), 0)$$

Key theoretical properties

$$\text{CB4 regression} \xrightarrow{u \to 0} \text{Beta regression}$$

$$\text{CB4 regression} \xrightarrow{u \to \infty} \text{Heteroscedastic tobit regression}$$

Likelihood inference

Log-likelihood

$$
\begin{aligned}
l_{(CB4)} = & \sum_{i:y_i \in (0,1)} \log f_{(b)}\left(\frac{y+u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma)\right) \\
& + \sum_{i:y_i = 0} \log F_{(b)}\left(\frac{u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma)\right) \\
& + \sum_{i:y_i = 1} \log \left\{1 - F_{(b)}\left(\frac{1+u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma)\right)\right\} \\
& - n_{01} \log(1+2u)
\end{aligned}
$$

n_{01} is the number of observations in $(0, 1)$ and $\sum_A(.) = 0$ is $A = \emptyset$

Implementation is straightforward; the key functions are the density and distribution function of the beta distribution

Likelihood inference

Log-likelihood derivatives with respect to β and γ

$$
\begin{aligned}
l_{(CB4)} = & \sum_{i:y_i \in (0,1)} \log f_{(b)} \left(\frac{y+u}{1+2u} \,|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \\
& + \sum_{i:y_i = 0} \log F_{(b)} \left(\frac{u}{1+2u} \,|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \\
& + \sum_{i:y_i = 1} \log \left\{ 1 - F_{(b)} \left(\frac{1+u}{1+2u} \,|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \right\} \\
& - n_{01} \log(1+2u)
\end{aligned}
$$

The derivatives of $\log f_{(b)}$ are obtained by the derivatives of the beta regression likelihood[13]

The derivatives of $\log F_{(b)}$ involve $\psi(.)$, $B(.,.)$ and ${}_3\tilde{F}_2$[14]

[13] see, Grün et al. (2012)

[14] implemented in the hypergeo R package

Likelihood inference

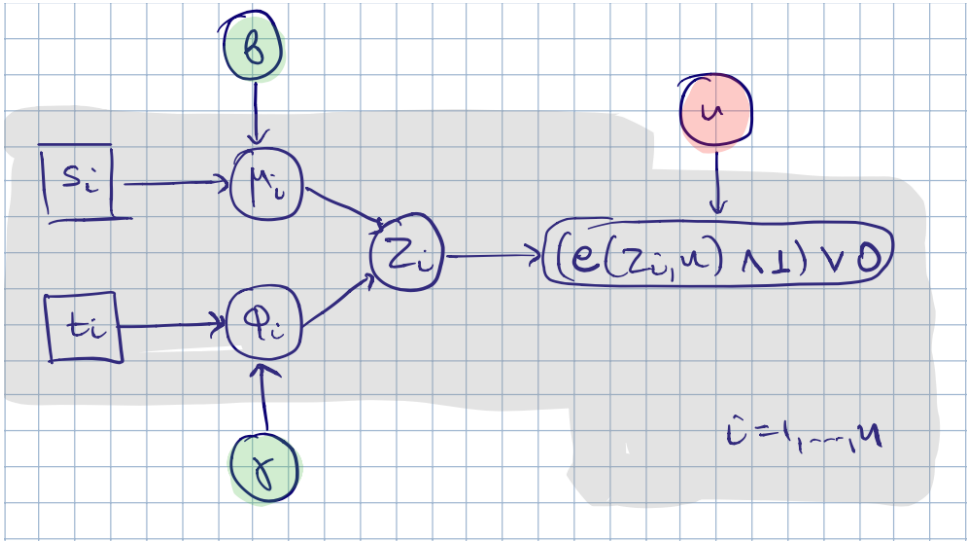
Log-likelihood derivatives with respect to u

$$
\begin{aligned}
l_{(CB4)} = & \sum_{i:y_i \in (0,1)} \log f_{(b)} \left(\frac{y+u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \\
& + \sum_{i:y_i = 0} \log F_{(b)} \left(\frac{u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \\
& + \sum_{i:y_i = 1} \log \left\{ 1 - F_{(b)} \left(\frac{1+u}{1+2u} \,\middle|\, g_1^{-1}(x_i^\top \beta), g_2^{-1}(w_i^\top \gamma) \right) \right\} \\
& - n_{01} \log(1+2u)
\end{aligned}
$$

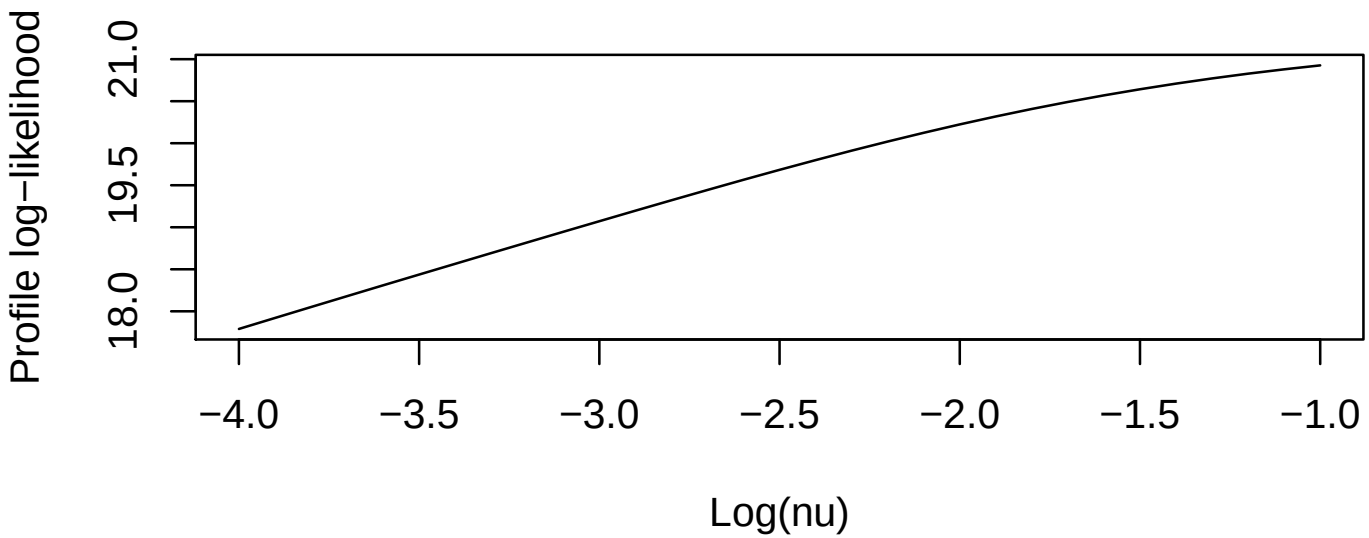
The derivatives of $logF_{(b)}$ and $log(1 - F_{(b)})$ can be written in terms of the beta density and distribution function

The derivatives of $logf_{(b)}$ are available in closed form

CB4 regression



Estimation of u is hard



Exponential-Beta mixture

$$u \sim Exp(\nu)$$
$$Y|u \sim \text{RestrictedBeta4}(\mu, \phi, u)$$

Marginal density

$$\frac{1}{\nu}\int_0^\infty f_{(r)}\left(y \,\middle|\, \mu, \phi, e\right) \exp\left\{-e/\nu\right\} de$$

The support is $\Re$ and Beta distribution is a boundary case (for $\nu \to 0$)

Marginal expectation: $-\nu + (1 + 2\nu)\mu$

Marginal variance: $\nu^2(1-2\mu)^2 + (1 + 4\nu + 8\nu^2)\dfrac{\mu(1-\mu)}{1+\phi}$

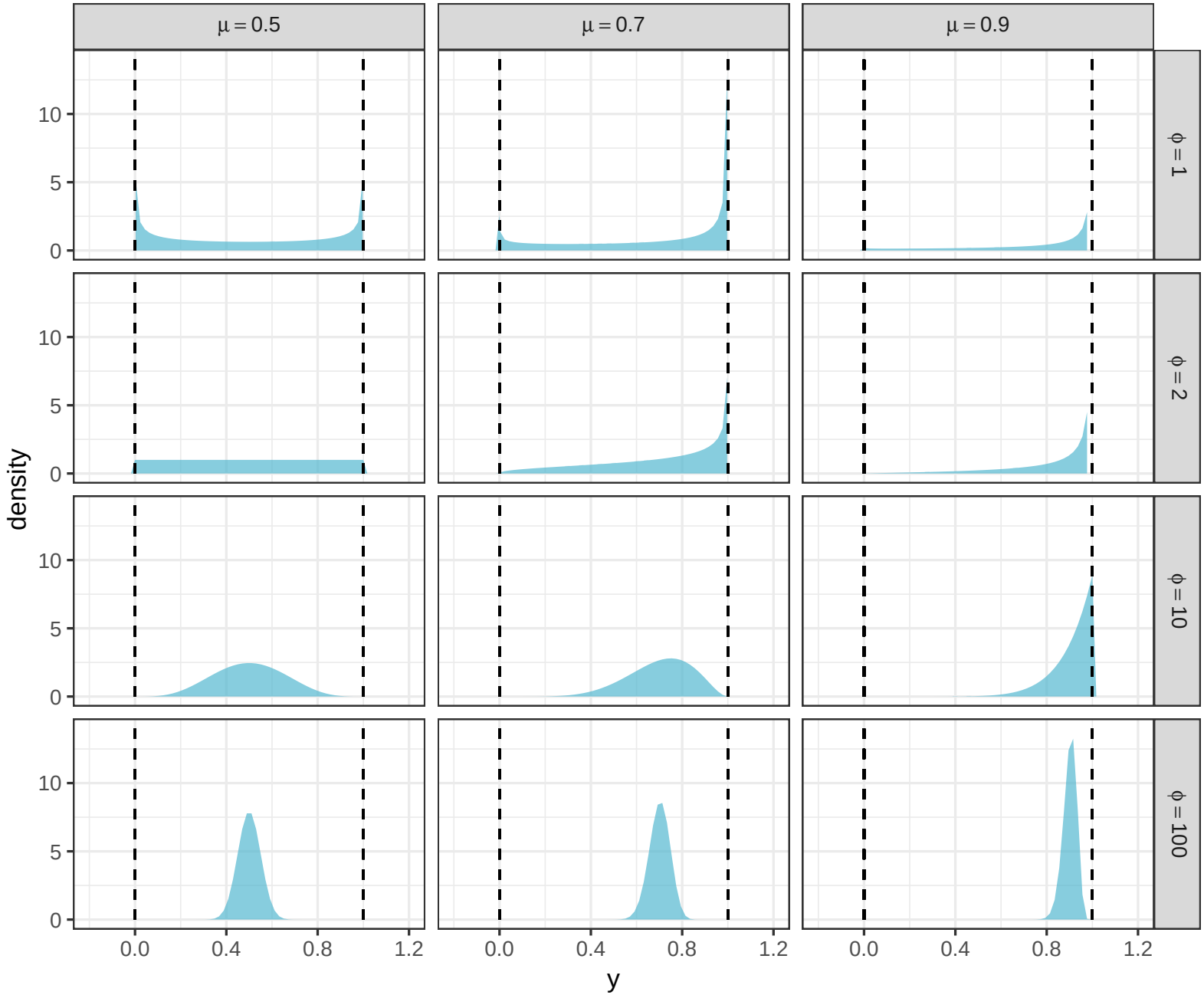
Gauss-Laguerre quadrature

$$\frac{1}{\nu}\int_0^{\infty} f_{(r)}\left(y \mid \mu, \phi, e\right) \exp\left(-\frac{e}{\nu}\right) de \simeq \sum_{t=1}^{T} W_t f_{(r)}\left(y \mid \mu, \phi, \nu Q_t\right)$$

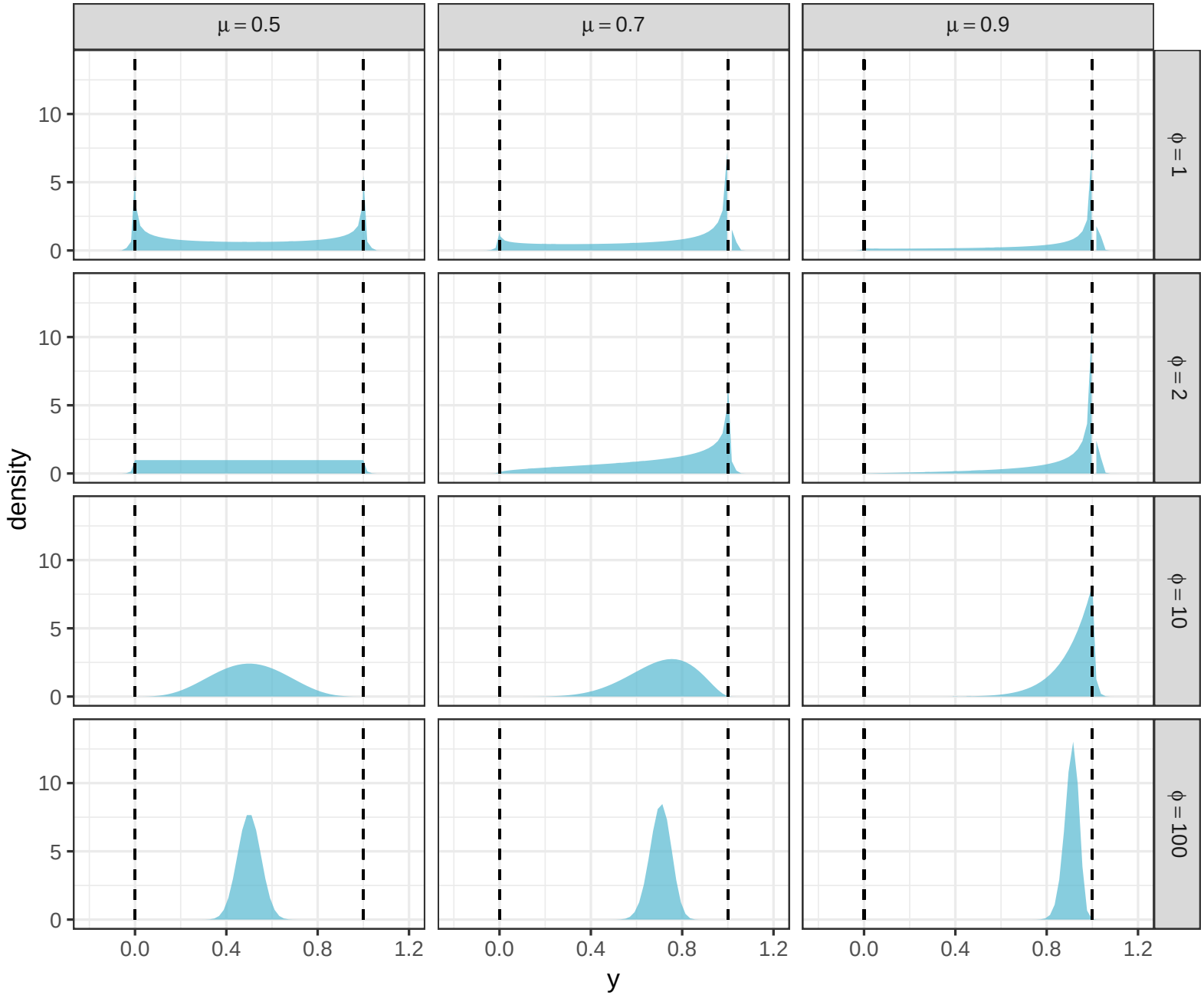
where W_t and Q_t $(t = 1, \ldots, T)$ are weights and nodes whose calculation and derivation are through Laguerre polynomials [15]

[15] see, Abramowitz and Stegun (1964, §25.4.45) for the approximation of integrals of the form $\int_0^{\infty} f(x)e^{-x}dx$

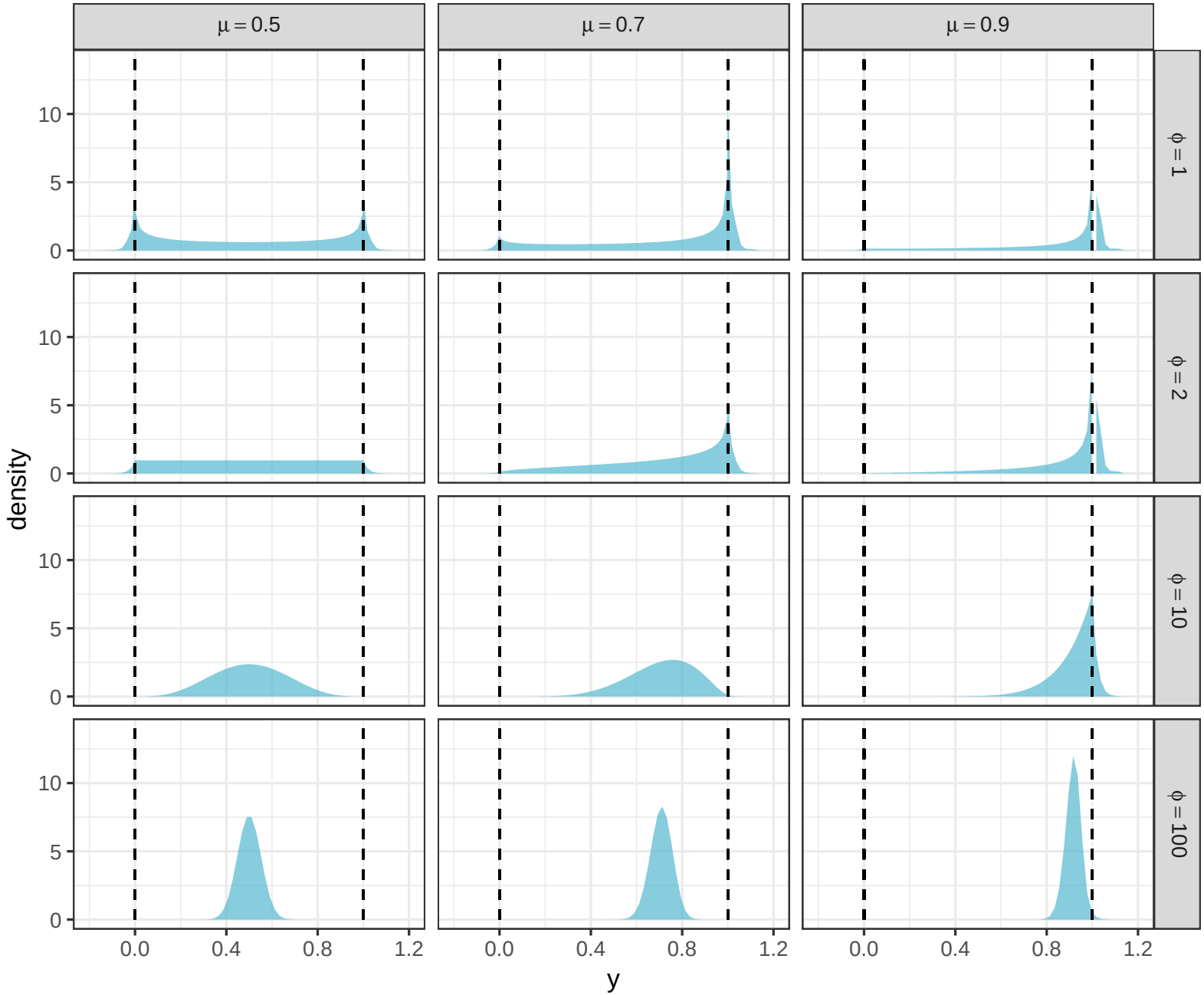
Exponential-Beta mixture ($\nu = 0.001$)



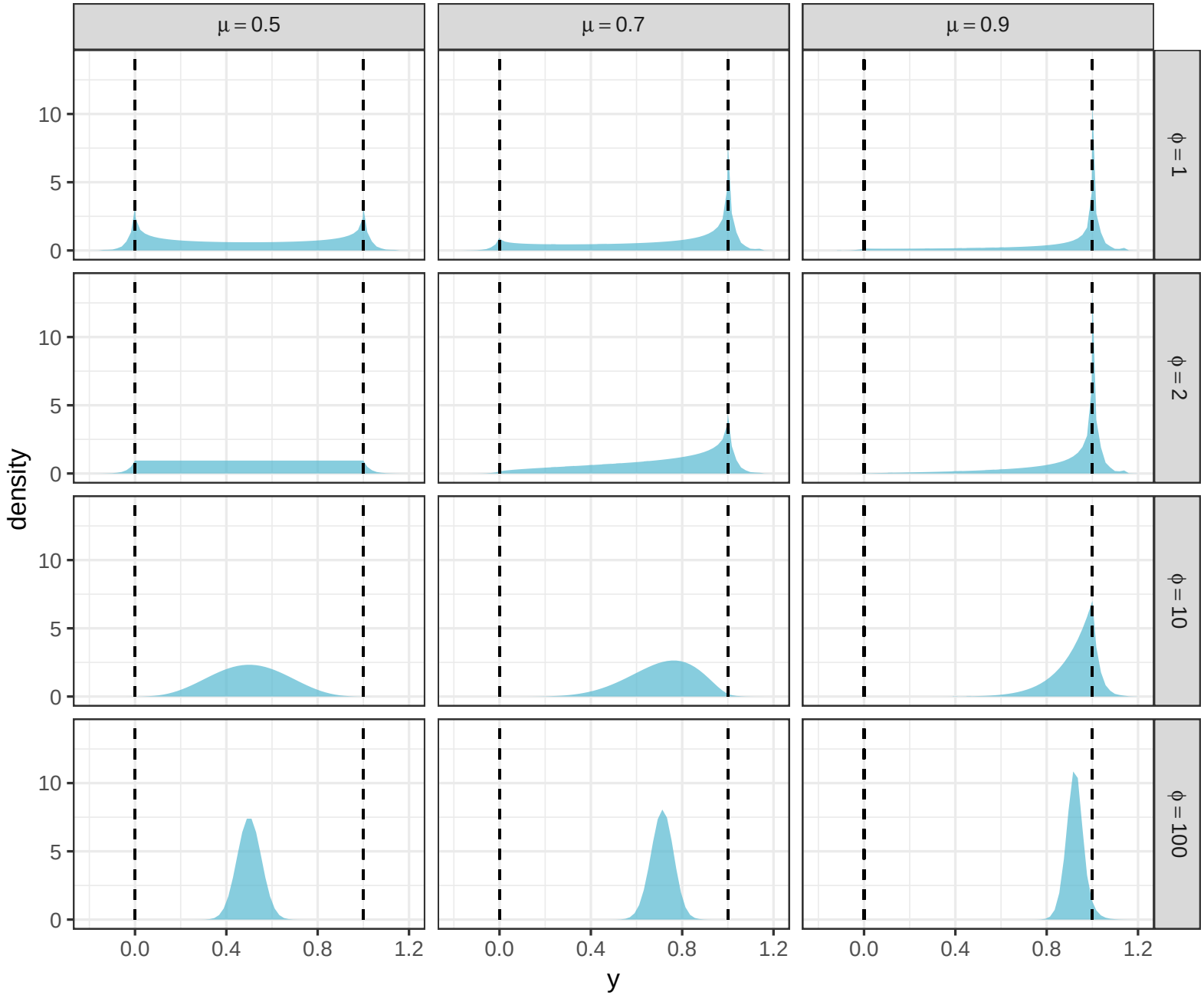
Exponential-Beta mixture ($\nu = 0.01$)



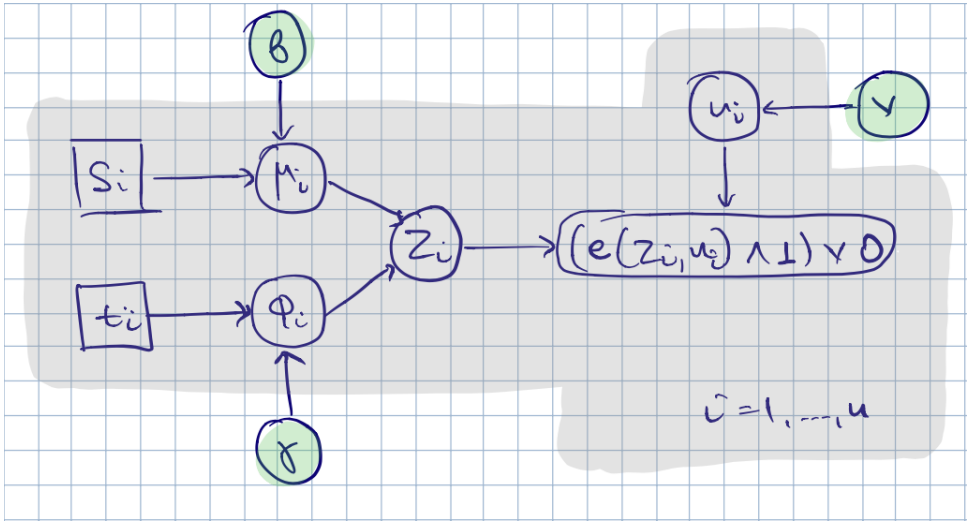
Exponential-Beta mixture ($\nu = 0.02$)



Exponential-Beta mixture ($\nu = 0.03$)



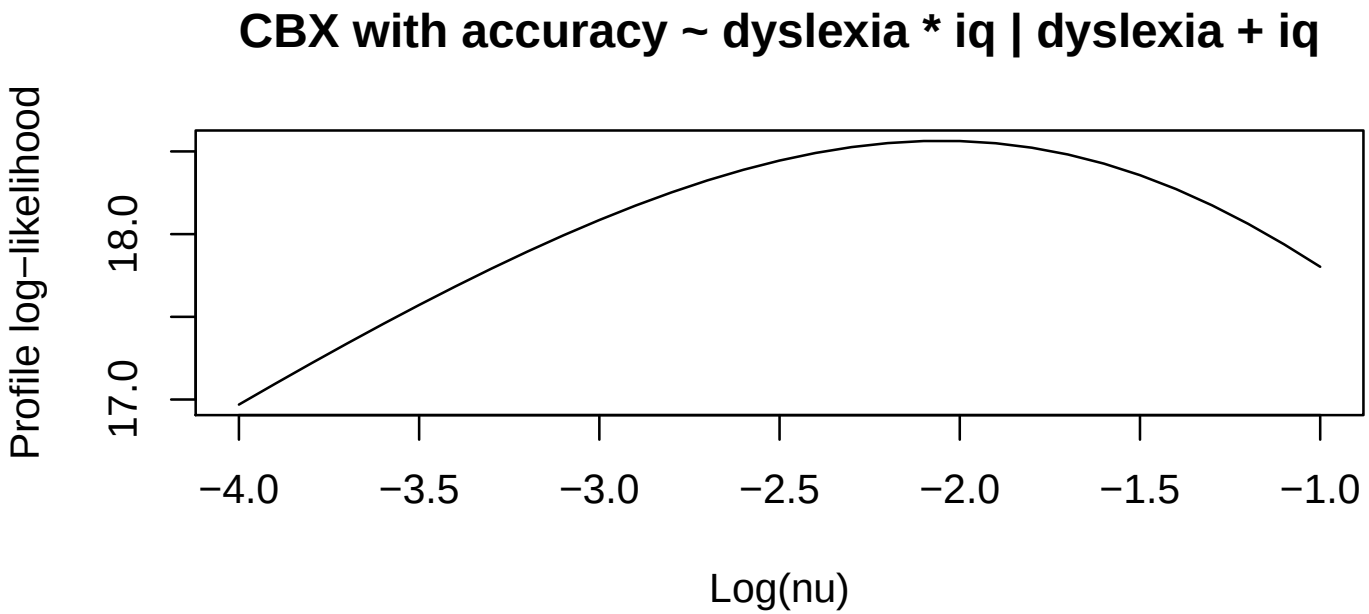
Censored Exponetial-Beta regression (CBX regression)



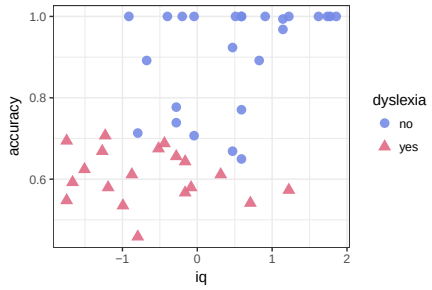
Parsimony and flexibility

Fast approximate log-likelihood evaluation through Gauss-Laguerre quadrature

ν can be estimated from data



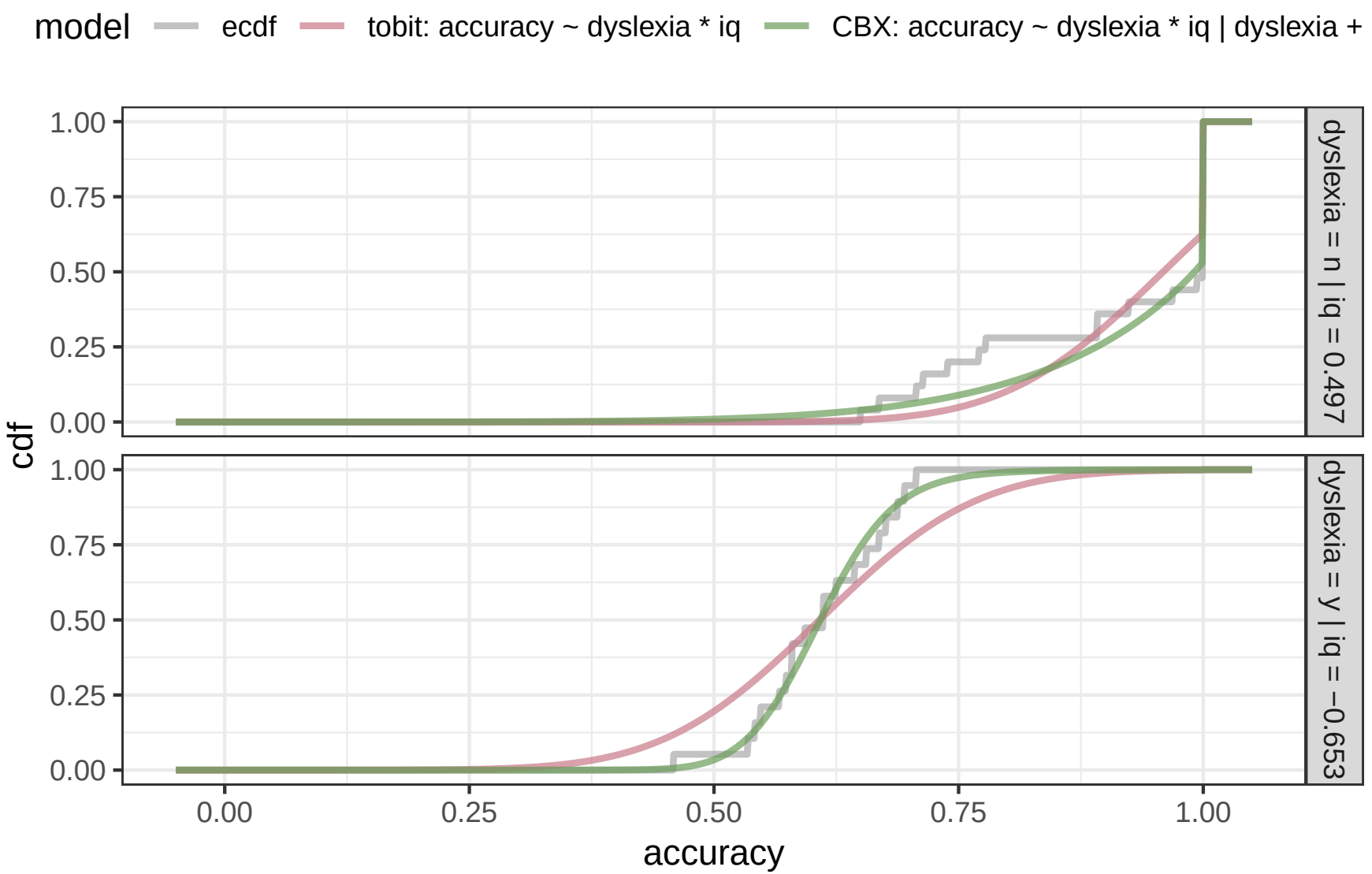
Reading accuracy



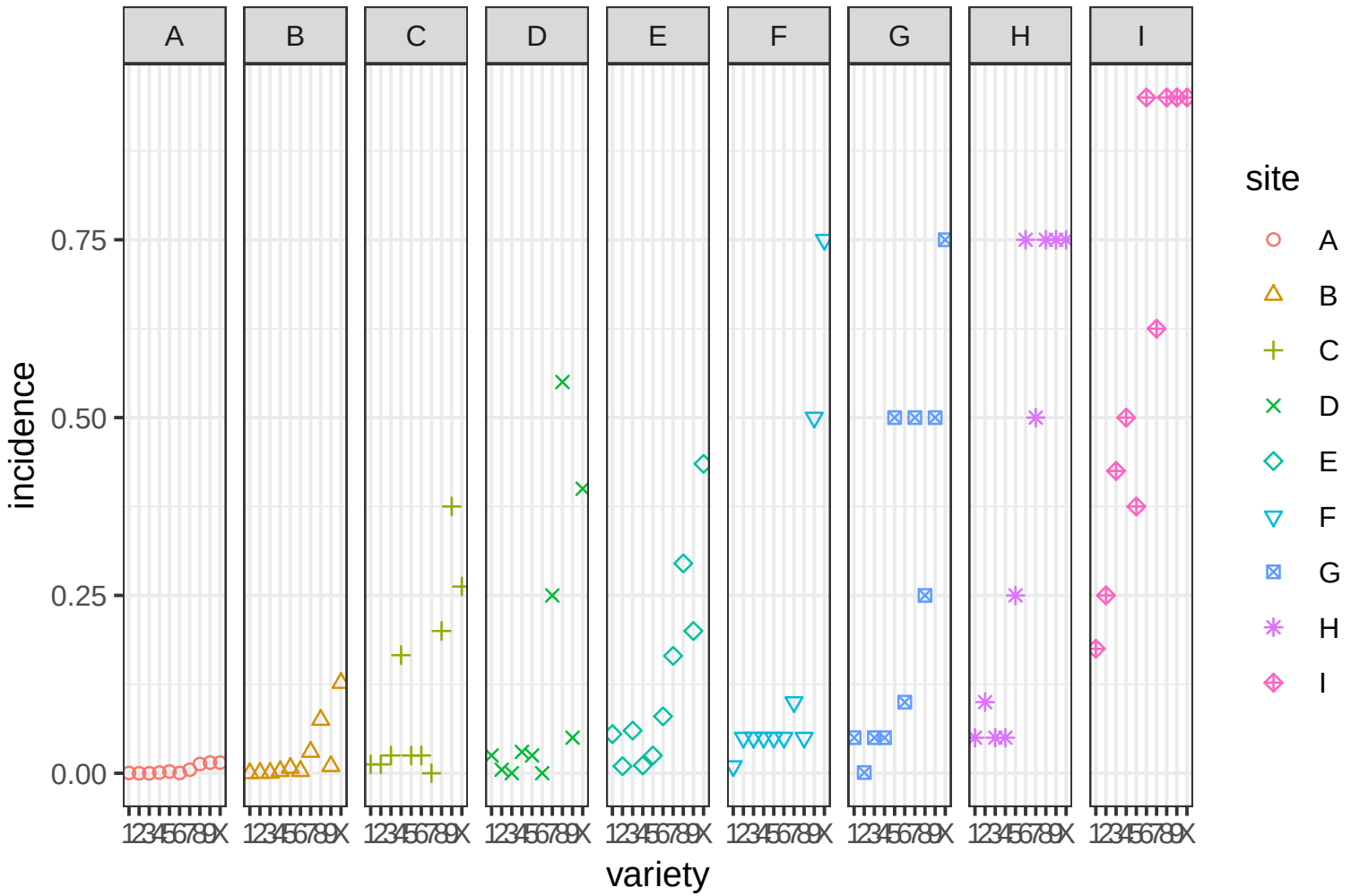
Model	μ	ϕ	α
Tobit	dyslexia $*$ iq		
Hurdle 1	dyslexia $*$ iq	dyslexia $+$ iq	1
Hurdle 2	dyslexia $*$ iq	dyslexia $+$ iq	dyslexia
Hurdle 3	dyslexia $*$ iq	dyslexia $+$ iq	dyslexia $+$ iq
CBX	dyslexia $*$ iq	dyslexia $+$ iq	

	Number of parameters	AIC	BIC
Tobit	5	-10.74	-1.82
Hurdle 1	8	-5.00	9.27
Hurdle 2	9	-21.80	-5.74
Hurdle 3	10	-21.80	-3.96
CBX	8	-21.13	-6.86

Reading accuracy



Incidence of leaf blotch on barley



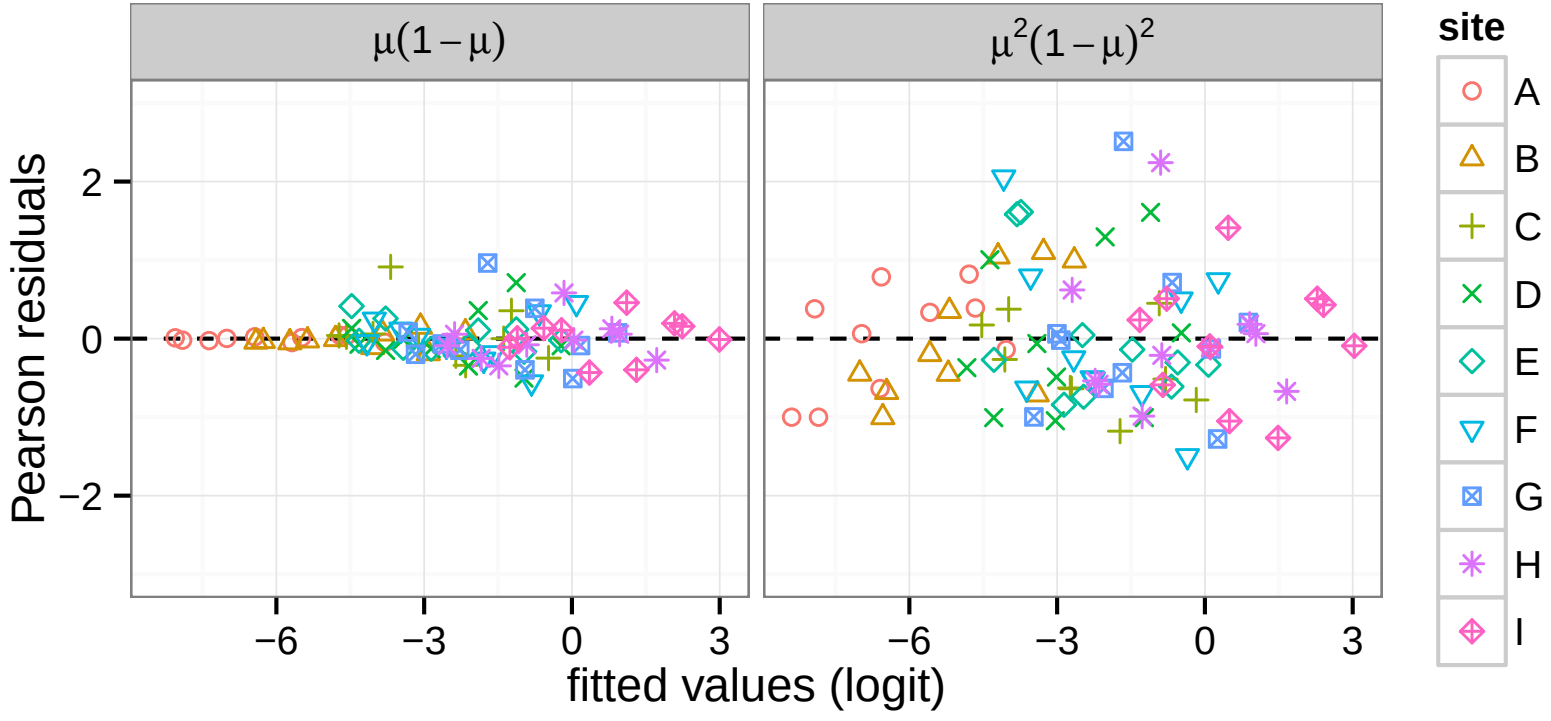
Incidence of *Rhynchosporium secalis* (percentage leaf area affected) on 10 varieties of barley tested at 9 sites in a variety trial in 1965[16]

[16] from the example that motivated quasi-likelihoods in Wedderburn (1974)

Quasi-likelihood

Wedderburn (1974) modelled incidence as pseudo-binomial response

Variance functions: $\mu(1-\mu)$ and $\mu^2(1-\mu)^2$



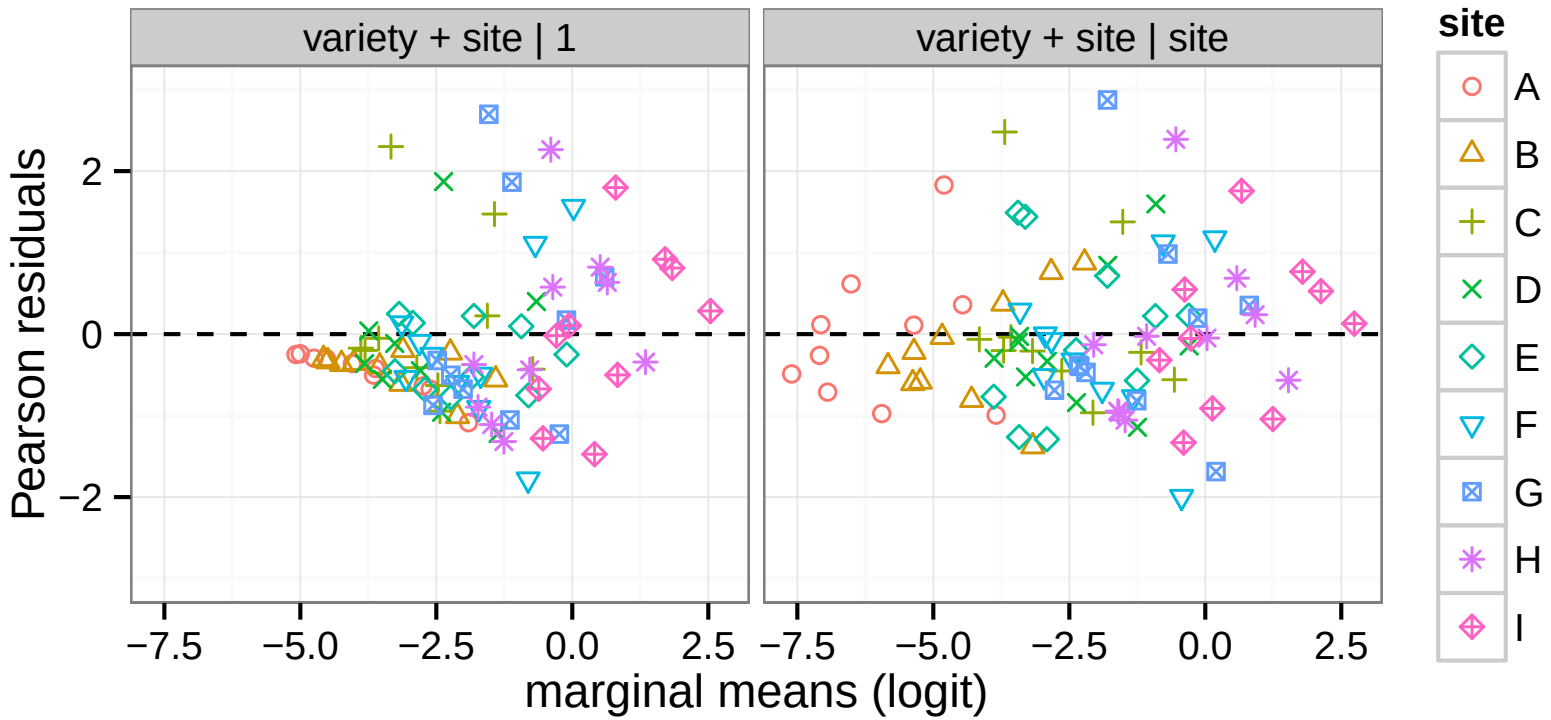
Deviance function **is not defined** when there are boundary observations[17]

[17] see, McCullagh and Nelder (1989, §9.2.4) for details

CBX regression

`variety + site | 1` and `variety + site | site`

50 Gauss-Laguerre quadrature points

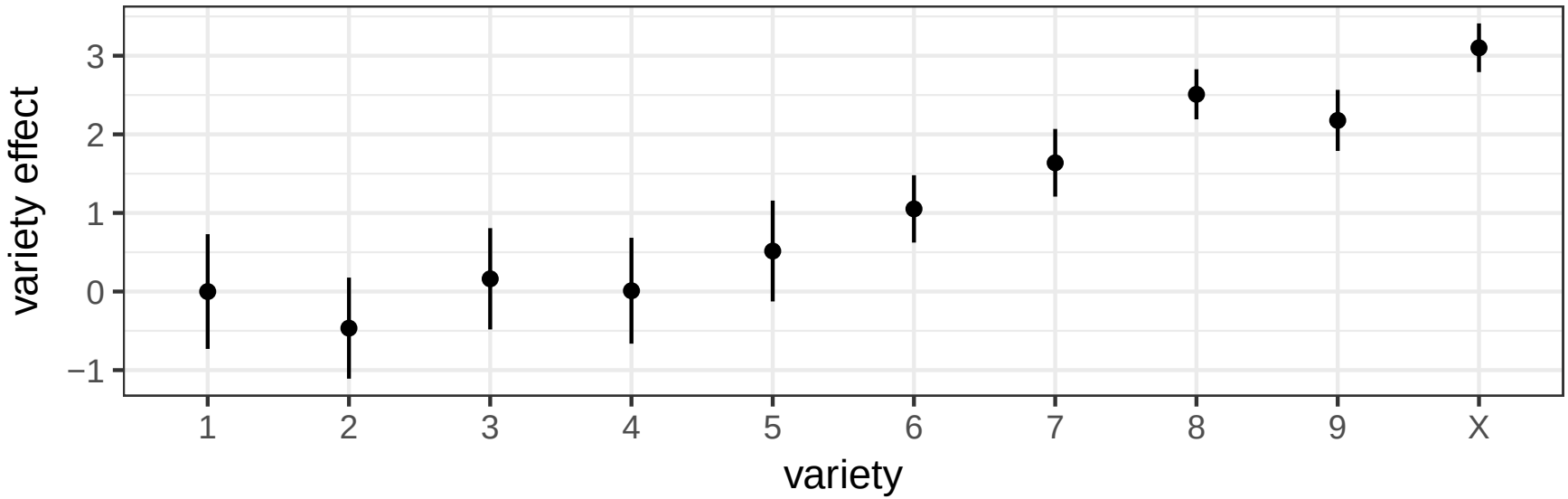


Testing for heteroscedasticity

Model	Parameters	Approximate log-likelihood	Df	LR statistic
`site + variety | 1`	20	126.81		
`site + variety | site`	28	147.56	8	41.50

Strong evidence against the model without site-dependent precision (0.999 quantile of χ^2_8 is 26.12)

95% point-wise comparison intervals[18]



[18] computed using the `qvcalc` R package (Firth and De Menezes, 2004)

Summary and discussion

Discussion

CB4 regression bridges the gap between heteroscedastic tobit and beta regression **with only 1 extra parameter**

CBX regression is an "exponential - CB4 regression mixture" that alleviates the identifiability issues of CB4 with the same number of parameters

CBX covers "less ground" between tobit and beta regression than CB4

Computationally efficient estimation and ready inferential procedures

Current and future work

Estimation and inference from CB4 and CBX regression in **betareg**

CBX regression mixtures and trees

References I

Abramowitz, M. and I. A. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (ninth Dover printing ed.). New York: Dover.

Calabrese, R. (2012). Regression model for proportions with probability masses at zero and one. Working Papers 201209, Geary Institute, University College Dublin.

Carnahan, J. V. (1989). Maximum likelihood estimation for the 4-parameter beta distribution. *Communications in Statistics - Simulation and Computation 18*(2), 513–536.

Cook, D. O., R. Kieschnick, and B. McCullough (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance 15*(5), 860–867.

Firth, D. and R. X. De Menezes (2004). Quasi-variances. *Biometrika 91*(1), 65–80.

Grün, B., I. Kosmidis, and A. Zeileis (2012). Extended beta regression in r: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software 48*(1), 1–25.

References II

Maddala, G. S. (1983, Mar). *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Ospina, R. and S. L. Ferrari (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis 56*(6), 1609–1623.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 186*, 343–414.

Smithson, M. and J. Verkuilen (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods 11*(1), 54–71.

Wang, J. Z. (2005). A note on estimation in the four-parameter beta distribution. *Communications in Statistics - Simulation and Computation 34*(3), 495–501.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika 61*, 439–447.