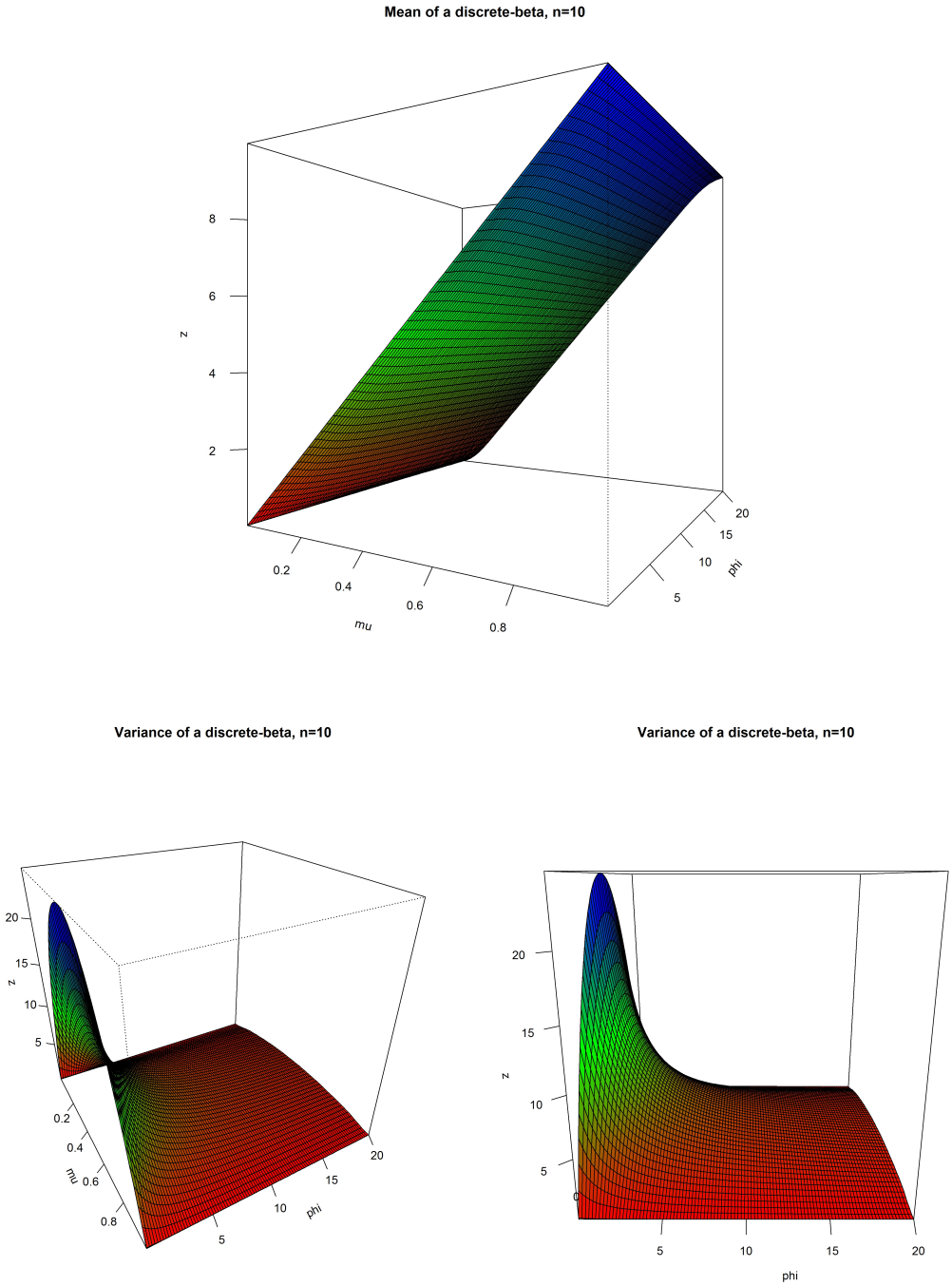

Figure 2.3: Top: expectation of a discrete-beta distribution with $n = 10$, plotted for $(\mu, \phi) \in (0, 1) \times (0, 20]$. Bottom: variance of a discrete-beta distribution with $n = 10$, plotted for $(\mu, \phi) \in (0, 1) \times (0, 20]$. Two different views of the same plot are shown.

2.2 Symmetric property

Using the relation

$$I_x(a,b) = 1 - I_{1-x}(b,a), \tag{2.10}$$

we can show that

$$\begin{aligned} P(Y=k) &= I_{\frac{k+1}{n+1}}(a,b) - I_{\frac{k}{n+1}}(a,b) \\ &= 1 - I_{1-\frac{k+1}{n+1}}(b,a) - \left[1 - I_{1-\frac{k}{n+1}}(b,a)\right] \\ &= I_{\frac{n-k+1}{n+1}}(b,a) - I_{\frac{n-k}{n+1}}(b,a) \end{aligned} \tag{2.11}$$

where $a = \mu\phi$ and $b = (1-\mu)\phi$. Eq. (2.11) is the probability that a discrete-beta distribution with parameters $(1-\mu, \phi)$ assumes the value $n-k$. Thus, if $\mu = \frac{1}{2}$, the pmf is symmetric with respect $x = \frac{n}{2}$ and, if n is even we have a single maximum in $x = \frac{n}{2}$, if n is odd, we have two equal maxima in $x = \lfloor \frac{n}{2} \rfloor$ and $x = \lfloor \frac{n}{2} \rfloor + 1$.

2.3 Identifiability of the model

We say that a model is identifiable if it is theoretically possible to learn the true value of the model's underlying parameter after obtaining an infinite number of observations from it. Mathematically, a parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability density functions or probability mass functions [13]. Identifiability is a property of the model, not of an estimator or estimation procedure. If the model is not identifiable, there is difficulty in doing inference. In the following, we prove that discrete-beta family is identifiable.

Let $n = 2$. Suppose that there exist $(\mu_1, \phi_1) \neq (\mu_2, \phi_2)$ such that

$$\begin{aligned} P(Y_1 = 0) &= P(Y_2 = 0); \\ P(Y_1 = 1) &= P(Y_2 = 1); \\ P(Y_1 = 2) &= P(Y_2 = 2); \end{aligned} \tag{2.12}$$

where $Y_1 \sim \text{Dbeta}(\mu_1, \phi_1)$ and $Y_2 \sim \text{Dbeta}(\mu_2, \phi_2)$. For the sake for clarity, let

$$a_1 = \mu_1\phi_1; \quad b_1 = (1-\mu_1)\phi_1; \quad a_2 = \mu_2\phi_2; \quad b_2 = (1-\mu_2)\phi_2. \tag{2.13}$$

From eq. (2.12) we obtain that

$$I_{\frac{1}{3}}(a_1, b_1) = I_{\frac{1}{3}}(a_2, b_2); \tag{2.14}$$

$$I_{\frac{2}{3}}(a_1, b_1) - I_{\frac{1}{3}}(a_1, b_1) = I_{\frac{2}{3}}(a_2, b_2) - I_{\frac{1}{3}}(a_2, b_2); \tag{2.15}$$

$$1 - I_{\frac{2}{3}}(a_1, b_1) = 1 - I_{\frac{2}{3}}(a_2, b_2) \tag{2.16}$$

whence we have

$$I_{\frac{1}{3}}(a_1, b_1) = I_{\frac{1}{3}}(a_2, b_2) \quad \text{and} \quad I_{\frac{2}{3}}(a_1, b_1) = I_{\frac{2}{3}}(a_2, b_2). \tag{2.17}$$

Eq. (2.17) implies that the latent variables have at least 3 intersections: at least one between $\left(0, \frac{1}{3}\right)$, at least one between $\left(\frac{1}{3}, \frac{2}{3}\right)$ and at least one between $\left(\frac{2}{3}, 1\right)$. However, it is not possible that two different beta distributions have more than two intersection between $(0, 1)$. Thus, we have that the distributions coincide and

$$a_1 = a_2 \quad \text{and} \quad b_1 = b_2, \tag{2.18}$$

which implies, from eq. (2.2), that

$$\mu_1 = \mu_2 \quad \text{and} \quad \phi_1 = \phi_2, \tag{2.19}$$

and discrete-beta model is identifiable. The result can be easily extended to $n > 2$ repeating the same argumentation on

$$I_{\frac{1}{n+1}}(a_1, b_1) = I_{\frac{1}{n+1}}(a_2, b_2) \quad \text{and} \quad I_{\frac{n}{n+1}}(a_1, b_1) = I_{\frac{n}{n+1}}(a_2, b_2). \tag{2.20}$$

Indeed, when $n = 2$, expected and observed relative frequencies coincide and we have a *saturated* model, since we have two parameters plus a constrain on the sum of the probabilities and just three relationships.

2.4 Comparison with binomial distribution and beta-binomial distribution

In this section, we compare our discrete-beta distribution with other two discrete distributions, the binomial and the beta-binomial defined on the same support $[0, 1, \ldots, n]$.

Let X be a random variable that follows a binomial distribution with parameters n and p. Its probability mass function (pmf) is defined as

$$P(X=k)=\binom{n}{k}p^k\left(1-p\right)^{n-k},\quad k=0,\cdots,n; \tag{2.21}$$

where $p\in[0,1]$ is called the probability of success. Let Y be a random variable with a beta-binomial density with parameters n, a and b, that is

$$P(Y=k)=\binom{n}{k}\frac{B\left(a+x,b+n-x\right)}{B\left(a,b\right)},\quad k=0,\cdots,n; \tag{2.22}$$

where $B(p,q)$ is the beta function (2.5), and a and b are two positive parameters. Thus we have that

$$\mathbb{E}[X]=np, \tag{2.23}$$

$$\mathbb{V}\mathrm{ar}[X]=np\left(1-p\right), \tag{2.24}$$

$$\mathbb{E}[Y]=n\frac{a}{a+b}, \tag{2.25}$$

$$\mathbb{V}\mathrm{ar}[Y]=n\frac{ab}{\left(a+b\right)^2}\frac{a+b+n}{a+b+1}. \tag{2.26}$$

It is known that, if X and Y have the same expected value, i.e. $p=\frac{a}{a+b}$, the variance of Y is greater than the variance of X for all values of a and b. Substituting $p=\frac{a}{a+b}$ in eq. (2.26), we have

$$\mathbb{V}\mathrm{ar}[Y]=np\left(1-p\right)\frac{a+b+n}{a+b+1}=\mathbb{V}\mathrm{ar}[X]\frac{a+b+n}{a+b+1}, \tag{2.27}$$

and with $n>1$ (in other words we exclude the Bernoulli case) the ratio in (2.27) is always greater than one. Due to this reason, the beta-binomial model is a popular and analytically tractable alternative to the binomial that captures *over-dispersion* with respect to the binomial model.

Let Z to be a random variable with discrete-beta density (2.3). Let $\mathbb{E}[Z]=\mu^{'}=np$. We can write $\mathbb{V}\mathrm{ar}[Z]$ (2.9) in the following way

$$\mathbb{V}\mathrm{ar}[Z]=(2n+1)\sum_{i=1}^{n}I_i-2\sum_{i=1}^{n}iI_i-\left(\sum_{i=1}^{n}I_i\right)^2$$

$$= \left(n - \mu^{'}\right)\left(n + \mu^{'} + 1\right) - 2\sum_{i=1}^{n} iI_i, \tag{2.28}$$

where $I_i = I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right)$ and (μ, ϕ) is a couple of parameters which satisfy $\mathbb{E}[Z] = \mu^{'}$. We have the minimum variance when the pmf is, in the limit case, concentrated in a single point of the set $[0, 1, \ldots, n]$ and the point is equal to $\mu^{'}$ (obviously it is a degenerated pmf). It is obtained by discretizing, for example, a beta distribution with $\mu = \frac{\mu^{'}}{n+1} + \frac{1}{2}\frac{1}{n+1}$ and $\phi \to +\infty$ (the limit case where the variance of the beta latent variable tends to zero). In this case, we have

$$\begin{aligned} 2\sum_{i=1}^{n} iI_i \to 2\sum_{i=\mu^{'}+1}^{n} i \cdot 1 &= 2\left(\sum_{i=1}^{n} i - \sum_{i=\mu^{'}}^{n} i\right) \\ &= 2\left(\frac{(n)(n+1)}{2} - \frac{\left(\mu^{'}\right)\left(\mu^{'}+1\right)}{2}\right) \\ &= n^2 + n - {\mu^{'}}^2 - \mu^{'} \end{aligned} \tag{2.29}$$

and eq. (2.28) tends to zero.

On the other hand, U-shapes are admitted by discrete-beta distribution. For a fixed $\mu^{'}$, every (μ, ϕ), such that $\mathbb{E}[Z] = \mu^{'}$ and $\phi < \min\left(\frac{1}{\mu}, \frac{1}{1-\mu}\right)$[1], generates a U-shape distribution (it follows from the properties of the beta latent distribution [41]). Thus, it can be over-dispersed. Since the variance is a continuous function of μ and ϕ, and $\mathbb{E}[Z] = \mu^{'}$ is also a continuous curve[2], we have that discrete-beta distribution can be under-dispersed, over-dispersed or it can have the same variance of a binomial with its same mean value.

Thus, if we set $\mathbb{E}[Z] = \mu^{'}$ and $\mathbb{V}\mathrm{ar}[Z] = \sigma^{2'}$, where $\mu^{'}$ and $\sigma^{2'}$ are admissible values,

[1]The existence of such points can be proved graphically intersecting the surface plotted in Figure 2.3 with the plane $z = \mu^{'}$ in the domain

$$D(\mu, \phi) = \begin{cases} \phi < \frac{1}{1-\mu}, & \text{if } 0 < \mu \leq \frac{1}{2} \\ \phi < \frac{1}{\mu}, & \text{if } \frac{1}{2} < \mu < 1 \end{cases}$$

[2]This curve is the intersection between the surface plotted in Figure 2.3 and the plane $z = \mu^{'}$.

in order to find μ and ϕ we have to solve the following system

$$\begin{cases} n - \sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) = \mu^{'} \\ \left(n - \mu^{'}\right)\left(n + \mu^{'} + 1\right) - 2\sum_{i=1}^{n} i I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) = \sigma^{2'}. \end{cases} \tag{2.30}$$

This is a non-linear system and we have not found a closed form solution. One way to solve it is using algorithms such as the Broyden Secant method, where the matrix of derivatives is updated after each major iteration using the Broyden rank 1 update, or full Newton method, where the Jacobian matrix of derivatives is recalculated at each iteration [21]. It is also possible to solve the minimization problem linked to this system

$$\min_{\mu,\phi}\left[\left(\mathbb{E}[Z] - \mu^{'}\right)^2 + \left(\mathbb{V}\mathrm{ar}[Z] - \sigma^{2'}\right)^2\right], \tag{2.31}$$

with, for example, a quasi-Newton method (also known as a variable metric algorithm), that is based on Newton's method to find the stationary point of a function, where the gradient is 0, but the Hessian matrix does not need to be computed, or a conjugate gradients method. Details of these methods, also with constrains, can be found in [57].

In Figures 2.4 and 2.5 comparisons, for different combinations of mean and variance, between discrete-beta distribution and beta-binomial distribution and binomial distribution, respectively, are shown.

2.5 Comparison with MUB distribution

In this section, we compare the variance of the discrete-beta distribution with that of the MUB distribution. Let X be a random variable with a MUB density (1.6) on the points $[1, \ldots, m]$. Then, the variance is

$$\mathbb{V}\mathrm{ar}[X] = (m-1)\left\{\pi\xi\left(1-\xi\right) + (1-\pi)\left[\frac{m+1}{12} + \pi\left(m-1\right)\left(\frac{1}{2} - \xi\right)^2\right]\right\}. \tag{2.32}$$

Note that X can be written as [60]

$$X = \pi\left(\mathrm{Bin}(m-1, 1-\xi) + 1\right) + (1-\pi)U_d(m), \tag{2.33}$$

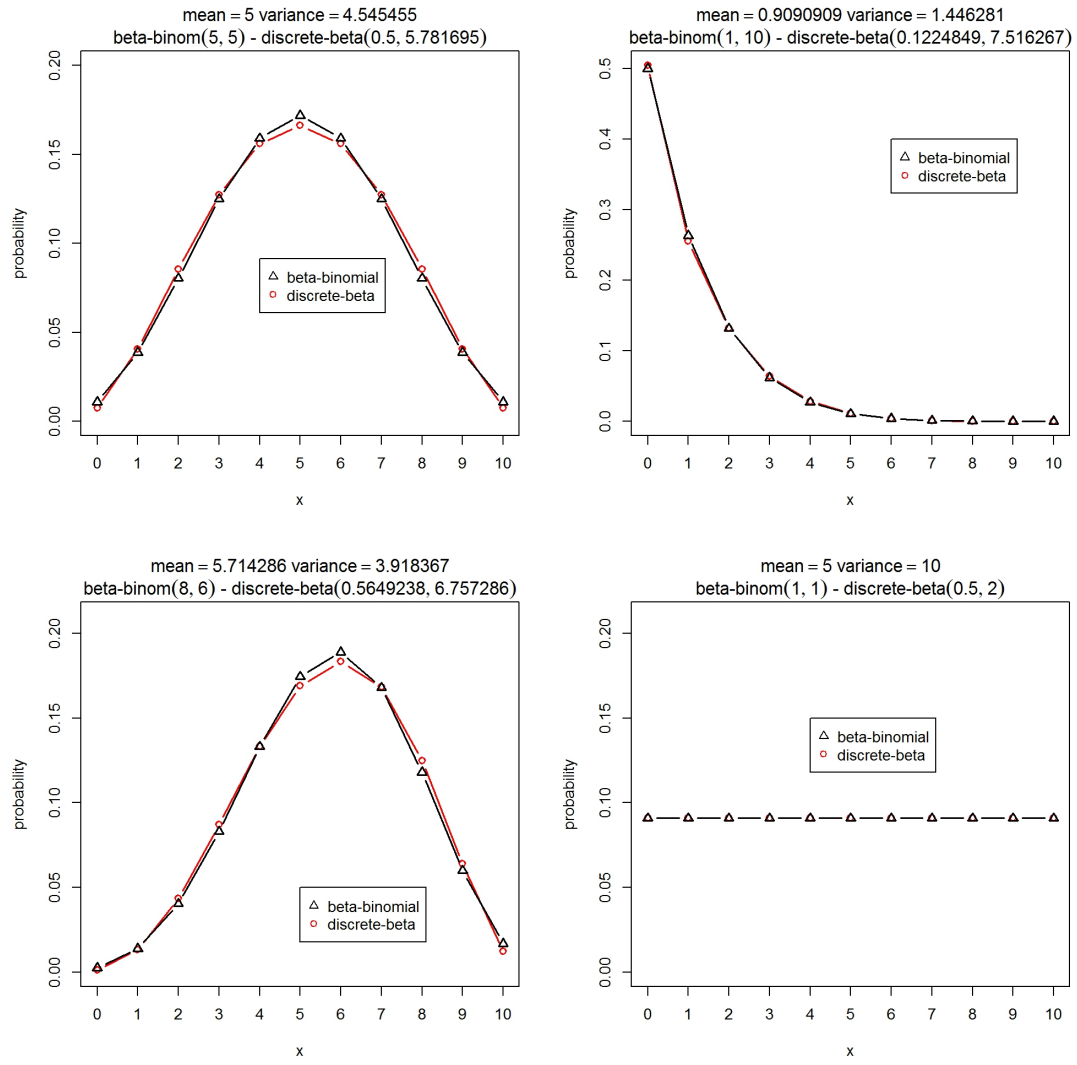

Figure 2.4: Comparison between discrete-beta distribution and beta-binomial distribution for different combinations of mean and variance. They coincide in the case of uniform distribution.

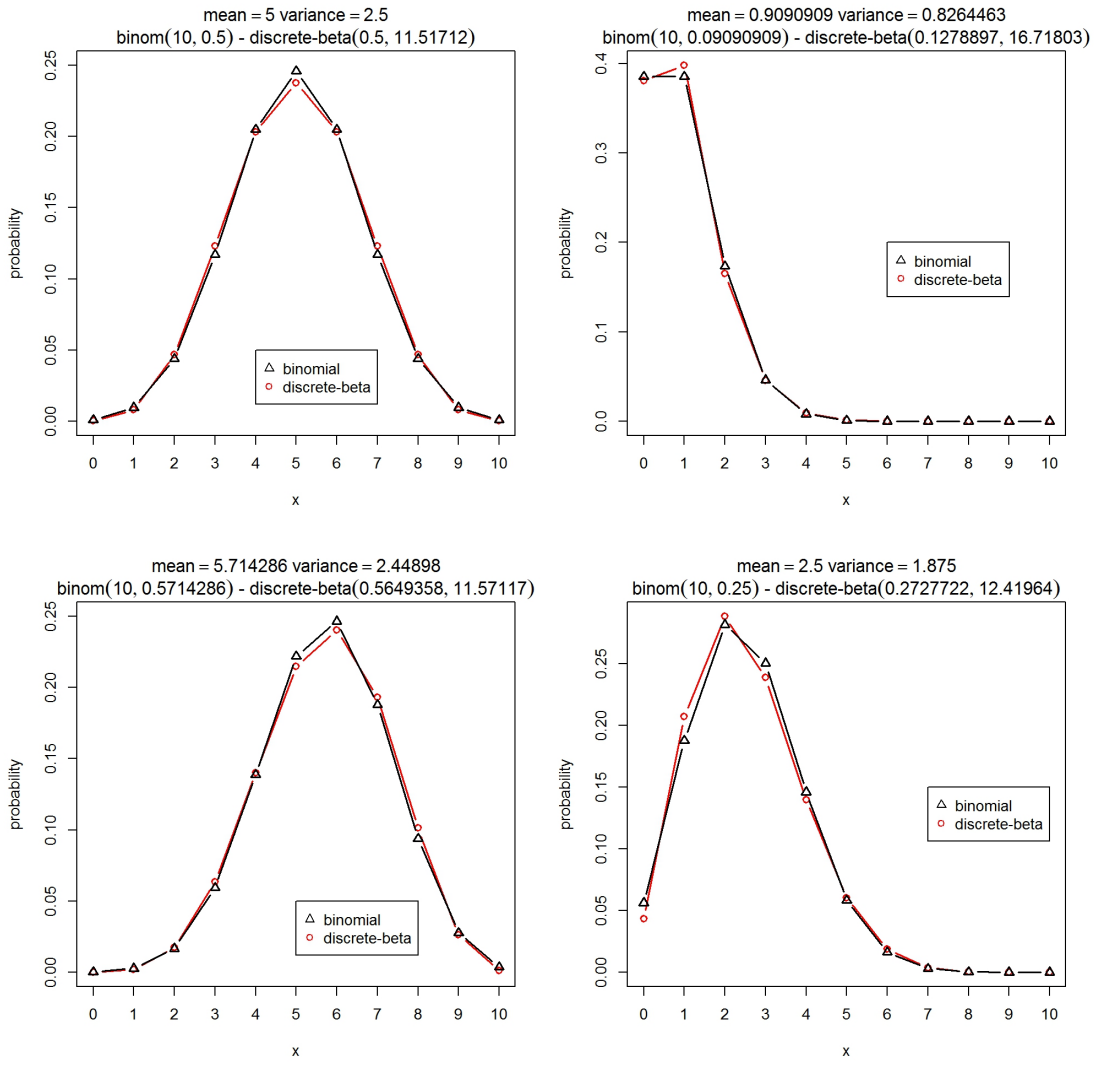

Figure 2.5: Discrete-beta distribution and binomial distribution plotted for different combinations of mean and variance.

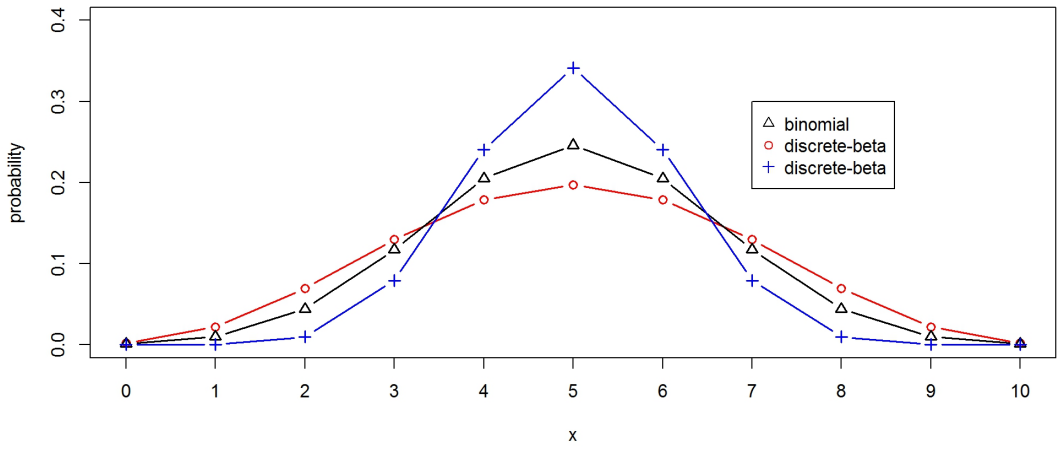

Figure 2.6: A binomial distribution with both under-dispersion and over-dispersion case of discrete-beta distribution with the same expected value.

where $\mathrm{Bin}(m-1, 1-\xi)$ is the binomial distribution and $U_d(m)$ is a uniform (rectangular) random variable defined over $[1, \ldots, m]$. The formula (2.32) can be decomposed in three parts (which can also be obtained by the classical variance decomposition [39])

$$\mathrm{Part}_1 = \pi\,(m-1)\,\xi\,(1-\xi) \tag{2.34}$$

$$\mathrm{Part}_2 = (1-\pi)\,\frac{m^2-1}{12} \tag{2.35}$$

$$\mathrm{Part}_3 = \pi(1-\pi)\,(m-1)^2\left(\frac{1}{2}-\xi\right)^2. \tag{2.36}$$

Part_3 is always non-negative. Part_1 and Part_2 are the variances respectively of $\mathrm{Bin}(m-1, 1-\xi)+1$ and $U_d(m)$ multiplied by their weights, respectively π and $1-\pi$. Moreover, we have that the variance of $\mathrm{Bin}(m-1, 1-\xi)+1$ is less or equal to the variance of $U_d(m)$, that is

$$\begin{aligned}(m-1)\,\xi\,(1-\xi) &\leq \frac{m^2-1}{12}\\ \xi\,(1-\xi) &\leq \frac{m+1}{12}\\ \xi^2 - \xi + \frac{m+1}{12} &\geq 0\end{aligned} \tag{2.37}$$

where eq. (2.37) is verified for all $\xi \in (0,1)$ and $m \geq 2$. Thus,

$$\begin{aligned}
\mathbb{V}\text{ar}[X] &= \text{Part}_1 + \text{Part}_2 + \text{Part}_3 \\
&\geq \text{Part}_1 + \text{Part}_2 = \pi\,(m-1)\,\xi\,(1-\xi) + (1-\pi)\,\frac{m^2-1}{12} \\
&\geq \pi\,(m-1)\,\xi\,(1-\xi) + (1-\pi)\,(m-1)\,\xi\,(1-\xi) \\
&= (m-1)\,\xi\,(1-\xi) \qquad (2.38)
\end{aligned}$$

where eq. (2.38) is the variance of $[\text{Bin}(m-1, 1-\xi)+1]$ that is equal to $\mathbb{V}\text{ar}[\text{Bin}(m-1, 1-\xi)]$. Then, a MUB distribution has a variance greater or equal to a binomial distribution and it can not be under-dispersed as a discrete-beta distribution could be (we showed this property in the previous section).

In Figure 2.7 the different shapes assumed by both distribution, with the same expectation and variance, are shown. Since their support are different, in order to compare the shapes, we choose first the expectation and variance for the CUB distribution on the set $[1, \ldots, m]$. Then, for the discrete-beta distribution we set

- the expectation equal to that of MUB distribution decreased by one unit;
- the same variance of the MUB distribution;
- $n = m - 1$.

Finally, we plot together on the same support $[1, \ldots, m]$ (but the real support of the discrete-beta distributions is $[0, \ldots, m-1]$). We can see that, in general, MUB distribution has tails heavier than those of discrete-beta.

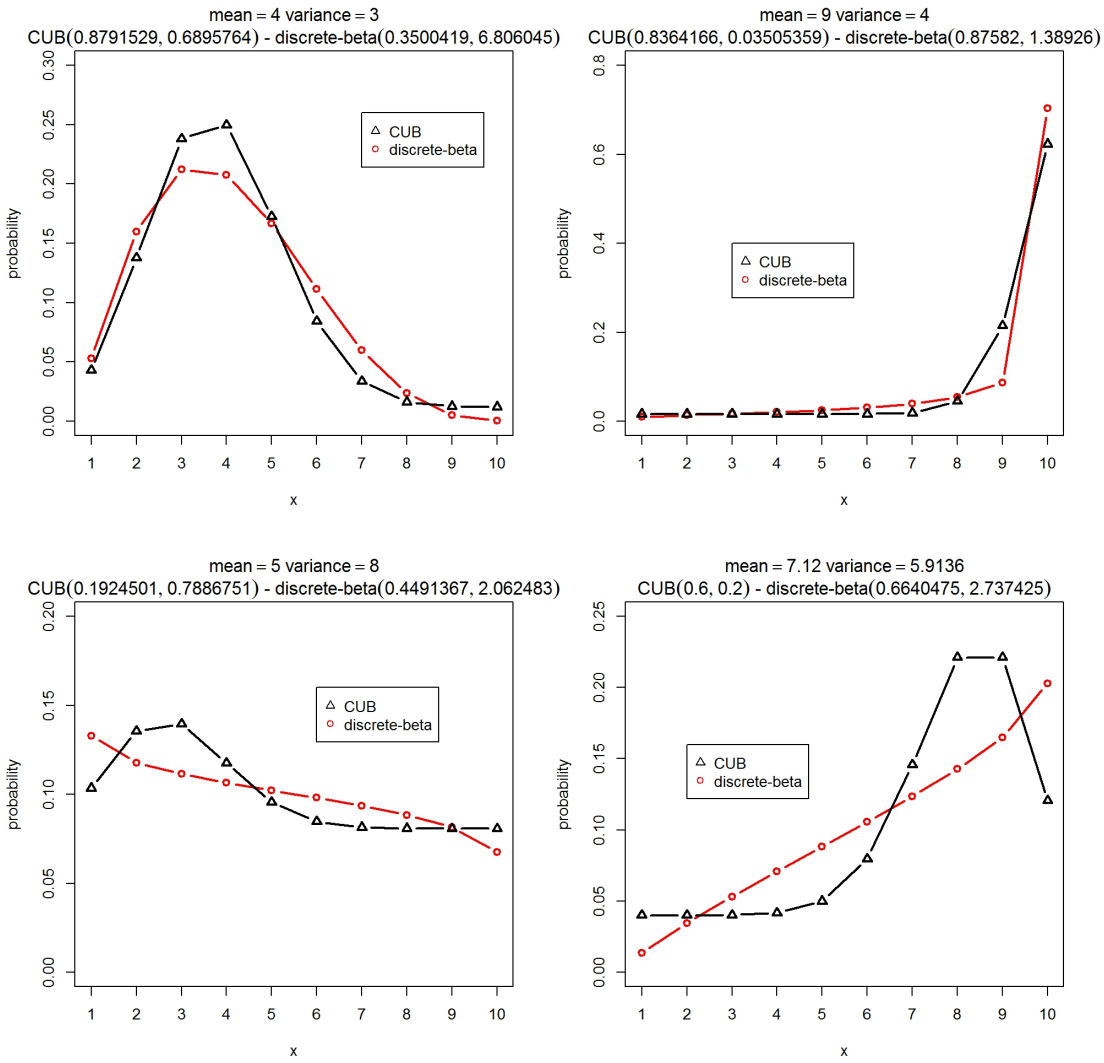

Figure 2.7: Comparison between discrete-beta distribution (shifted by 1) and MUB distribution for different combinations of mean and variance.

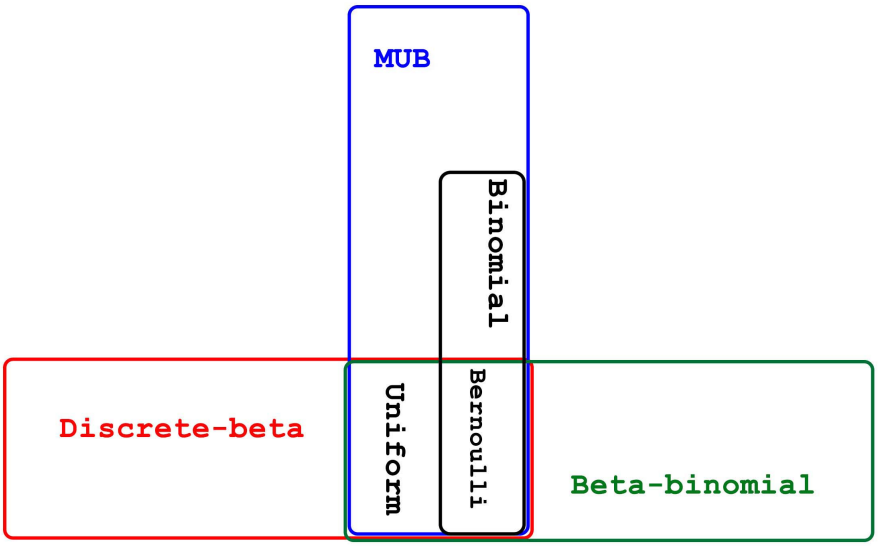

Figure 2.8: Intersection between discrete-beta, beta-binomial, CUB and binomial distribution.

Conclusion

In Figure 2.8 the results are summarised through a diagram. As we showed, only the discrete-beta distribution can be under-dispersed respect to the binomial case. If the outcomes reduced to a Bernoulli case, all the previous distributions tend to coincide. The uniform case is taken into account by all three discrete-beta distribution, CUB model and beta-binomial distribution. When the modalities are greater then three, MUB distribution and discrete-beta distribution can have very different behaviors since uncertainty is take into account in various ways. Instead, beta-binomial and binomial distributions can be well approximated by discrete-beta distribution, as we can see in the previous sections.

Chapter 3

Models with discrete-beta distribution

In this chapter, we investigate the problem of parameter estimation in both cases of a single population and in the presence of covariates. Let N be the sample size; in the first case we assume that individual responses follow the same distribution, with fixed parameters, in the latter case they can have different parameters. In both situations, we will use the maximum likelihood method for parameter estimation.

3.1 Single population

One of the primary uses of statistics is to estimate population parameters when the population is too large for a census to be practical. To accomplish this, a random sample of values from the population data set is drawn and the sample statistic is calculated to draw inferences to estimate the value of unknown population parameters.

Assuming that Y_i, the response random variable of the ith individual, follows a discrete-beta distribution, Dbeta(μ, ϕ, n), we have that

$$\begin{aligned} P\left(Y_i = y_i\right) &= \int_{\frac{y_i}{n+1}}^{\frac{y_i+1}{n+1}} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}\, dx \\ &= I_{\frac{y_i+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{y_i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right), \qquad (3.1) \\ & \qquad i = 1, 2, \cdots, N. \end{aligned}$$

In case of independent individuals, if $\mathbf{y} = (y_1, ..., y_N)$ is the vector of the observed values, we can compute the likelihood as

$$\begin{aligned}
\mathcal{L}(\mu, \phi|\ \mathbf{y}) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N|\ \mu, \phi) \\
&= \prod_{i=1}^{N} \int_{\frac{y_i}{n+1}}^{\frac{y_i+1}{n+1}} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}\, dx \\
&= \prod_{i=1}^{N} \left[I_{\frac{y_i+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{y_i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right] \\
&= \prod_{k=0}^{n} \left[I_{\frac{k+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right]^{n_k}, \qquad (3.2)
\end{aligned}$$

where n_k is the absolute frequency of the k value in the sample.

The log-likelihood function is given by

$$\begin{aligned}
l(\mu, \phi|\ \mathbf{y}) &= \log \mathcal{L}(\mu, \phi|\ \mathbf{y}) \\
&= \log \left\{ \prod_{k=0}^{n} \left[I_{\frac{k+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right]^{n_k} \right\} \\
&= \sum_{k=0}^{n} n_k \log \left[I_{\frac{k+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right]. \qquad (3.3)
\end{aligned}$$

Let $\boldsymbol{\theta} = [\mu, \phi]$. Following the maximum likelihood method, we have that $\hat{\boldsymbol{\theta}}(\mathbf{y}) = [\hat{\mu}(\mathbf{y}), \hat{\phi}(\mathbf{y})]$, the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, is computed as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}(\mathbf{y}) &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\mu, \phi|\ \mathbf{y}) = \arg\max_{\boldsymbol{\theta}} l(\mu, \phi|\ \mathbf{y}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{k=0}^{n} n_k \log \left[I_{\frac{k+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right]. \qquad (3.4)
\end{aligned}$$

Since the system of equations

$$\begin{cases} \dfrac{\partial}{\partial \mu} l(\mu, \phi|\ \mathbf{y}) = 0 \\ \dfrac{\partial}{\partial \phi} l(\mu, \phi|\ \mathbf{y}) = 0 \end{cases} \qquad (3.5)$$

has no closed form solutions, we prefer to maximize directly eq. (3.4) with a quasi-Newton method. We choose the algorithm implemented in `optim` function in the `R`

software. The choice of "L-BFGS-B" method in the `optim` function allows box constraints, that is each variable can be given a lower and/or upper bound; this method is described in detail in [12]. Good starting points for μ and ϕ is given by $(\mu_0, \phi_0) = \left(\frac{1}{n+1}\frac{\sum_i y_i}{N}, (n+1)\frac{N-1}{\sum_i (y_i-\mu_0)^2}\right)$, since μ is related to the mean end ϕ to the inverse of the variance.

In order to derive confidence intervals for μ and ϕ, we recall the asymptotic properties of maximum likelihood estimators. In particular, it is well known that the asymptotic variance-covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ of the ML estimators $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ is obtained by inverting the negative of the expectation of the second derivatives (the Hessian) of the log-likelihood function (eq. (3.3)). Since the *expected Fisher information matrix*, when the operations of integration with respect to x and differentiation with respect to θ can be interchanged in the expectation[26], can be written as

$$\left\{\mathcal{I}(\boldsymbol{\theta})\right\}_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta}|\ \mathbf{y})\right], \tag{3.6}$$

we have that

$$\mathbf{V}(\boldsymbol{\theta}) = \left[\mathcal{I}(\boldsymbol{\theta})\right]^{-1}. \tag{3.7}$$

An alternative method, which shares the same asymptotic properties, is based on the *observed information matrix* $\mathcal{J}(\boldsymbol{\theta})$, that is the negative of the second derivatives (the Hessian matrix) of the log-likelihood function

$$\left\{\mathcal{J}(\hat{\boldsymbol{\theta}})\right\}_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta}|\ \mathbf{y})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{3.8}$$

It is a sample-based version of the Fisher information. It is known that

$$\hat{\boldsymbol{\theta}}_N \approx N\left(\boldsymbol{\theta}, (\mathcal{J}(\hat{\boldsymbol{\theta}}_N)^{-1}\right) \tag{3.9}$$

where the subscript N is introduced to remind us that it is obtained by a sample of N individuals. Since what users actually do in multiparameter situation is to focus on confidence interval for single parameter, we obtain the following intervals

$$\left(\hat{\theta}_{Nj} - z_{\frac{\alpha}{2}}\sqrt{\left((\mathcal{J}(\hat{\boldsymbol{\theta}}_N)^{-1}\right)_{jj}}\ ,\ \hat{\theta}_{Nj} + z_{\frac{\alpha}{2}}\sqrt{\left((\mathcal{J}(\hat{\boldsymbol{\theta}}_N)^{-1}\right)_{jj}}\right) \tag{3.10}$$

where $z_{\frac{\alpha}{2}}$, as usual, is the $(1-\frac{\alpha}{2})$ quantile of a standard normal distribution.

In our case,

$$\mathcal{J}(\hat{\boldsymbol{\theta}}) = -\left[\begin{array}{cc} \frac{\partial^2}{\partial\mu^2} l(\boldsymbol{\theta}|\ \mathbf{y}) & \frac{\partial^2}{\partial\mu\partial\phi} l(\boldsymbol{\theta}|\ \mathbf{y}) \\ \frac{\partial^2}{\partial\phi\partial\mu} l(\boldsymbol{\theta}|\ \mathbf{y}) & \frac{\partial^2}{\partial\phi^2} l(\boldsymbol{\theta}|\ \mathbf{y}) \end{array}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_N} \tag{3.11}$$

and we have that

$$\begin{aligned} \frac{\partial}{\partial\theta_r} l(\boldsymbol{\theta}|\ \mathbf{y}) &= \frac{\partial}{\partial\theta_r}\left[\sum_{k=0}^{n} n_k \log P(Y=k)\right] \\ &= \sum_{k=0}^{n} n_k \frac{\partial}{\partial\theta_r} \log P(Y=k), \end{aligned} \tag{3.12}$$

$$\frac{\partial^2}{\partial\theta_r\partial\theta_s} l(\boldsymbol{\theta}|\ \mathbf{y}) = \sum_{k=0}^{n} n_k \frac{\partial^2}{\partial\theta_r\partial\theta_s} \log P(Y=k), \tag{3.13}$$

$$r = 1, 2; \quad s = 1, 2$$

with $\theta_1 = \mu$, $\theta_2 = \phi$ and

$$\begin{aligned} \log P(Y=k) &= \log \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}\ dx \\ &= \log\Gamma(\phi) - \log\Gamma(\mu\phi) - \log\Gamma((1-\mu)\phi) + \\ &\quad + \log \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}\ dx. \end{aligned} \tag{3.14}$$

For the sake of clarity, we will use the following notation

$$x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1} = h(x,\mu,\phi) \tag{3.15}$$

and we compute the first derivatives as

$$\begin{aligned} \frac{\partial}{\partial\mu}\log P(Y=k) &= -\phi\psi(\mu\phi) + \phi\psi((1-\mu)\phi) + \\ &\quad + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx} \end{aligned} \tag{3.16}$$

$$\frac{\partial}{\partial\phi}\log P(Y=k) = \psi(\phi) - \mu\psi(\mu\phi) - (1-\mu)\psi((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx} \tag{3.17}$$

where $\psi(t)$ is the *digamma* function, that is defined as the logarithmic derivative of the gamma function

$$\psi(t) = \frac{d}{dt}\log\Gamma(t). \tag{3.18}$$

Thus, the second derivatives can be written as

$$\frac{\partial^2}{\partial\mu^2}\log P(Y=k) = -\phi^2\psi_1(\mu\phi) + \phi^2\psi_1((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi^2\log^2\left(\frac{x}{1-x}\right)\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx} + \\ - \left(\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx}\right)^2 \tag{3.19}$$

$$\frac{\partial^2}{\partial\phi^2}\log P(Y=k) = \psi_1(\phi) - \mu^2\psi_1(\mu\phi) - (1-\mu)^2\psi_1((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]^2\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx} + \\ - \left(\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\ dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\ dx}\right)^2 \tag{3.20}$$

$$\frac{\partial^2}{\partial\phi\partial\mu}\log P(Y=k) = \psi_1(\mu\phi) - \mu\phi\psi_1(\mu\phi) + \psi_1((1-\mu)\phi) + (1-\mu)\phi\psi_1((1-\mu)\phi) +$$

$$+\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x+(1-\mu)\log(1-x)\right]\phi\log\left(\frac{x}{1-x}\right)\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx}+$$
$$-\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x+(1-\mu)\log(1-x)\right]\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx}*$$
$$*\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx} \tag{3.21}$$

where $\psi_1(t)$ is the *trigamma* function defined as

$$\psi_1(t)=\frac{d^2}{d^2t}\log\Gamma(t). \tag{3.22}$$

Substituting eq. (3.19), (3.20) and (3.21) in eq. (3.13), we can compute $\mathcal{J}(\hat{\boldsymbol{\theta}})$. Since the derivatives involve integrations, which have to be computed in a numerical way, it is also possible to estimate them using the quasi-Newton algorithm used to solve the problem (3.4).

3.1.1 Validation by simulation

In order to have a validation by simulation, we followed the steps showed below:

- we choose a couple of values for (μ^*,ϕ^*);
- we generated a random sample by a discrete-beta distribution with $\mu=\mu^*$ and $\phi=\phi^*$;
- we estimated (μ,ϕ) following the maximum likelihood method described in this section.

We used the R-script showed in Appendix A. Four cases are showed below. All intervals are calculated at 95% confidence level.

With $\mu=0.3$ and $\phi=6$, we obtain

```
$mu
     estimate lower bound upper bound Wald test
[1,]  0.29985     0.26541     0.33429  17.06363

$phi
     estimate lower bound upper bound Wald test
[1,]  5.92539     4.18542     7.66536   6.67457
```

With $\mu = 0.65$ and $\phi = 3.5$, we obtain

```
$mu
     estimate lower bound upper bound Wald test
[1,]  0.65491     0.61171      0.6981  29.71521

$phi
     estimate lower bound upper bound Wald test
[1,]  3.62504     2.62831     4.62176   7.12827
```

With $\mu = 0.5$ and $\phi = 2$, (discrete uniform distribution) we obtain

```
$mu
     estimate lower bound upper bound Wald test
[1,]  0.50868     0.45429     0.56307  18.32997

$phi
     estimate lower bound upper bound Wald test
[1,]  2.15492     1.58366     2.72618   7.39339
```

With $\mu = 0.1$ and $\phi = 20$, we obtain

```
$mu
     estimate lower bound upper bound Wald test
[1,]  0.10267     0.08809     0.11726  13.79562

$phi
     estimate lower bound upper bound Wald test
[1,] 19.37949    10.76797    27.99101   4.41073
```

We noted that μ is generally well estimated, while $\hat{\phi}$ has a larger uncertainty than $\hat{\mu}$.

3.2 Covariates

This section regards the introduction of covariates, i.e. concurrent variables of the outcome that improve both the results and the interpretation of the models. We do not follow the approach of generalized linear models, in which it is allowed for an arbitrary function (the link function) of the mean of response variable, $g(\mathbb{E}[Y])$, to vary linearly

with the predicted values (rather than assuming that the response itself must vary linearly). In fact, in our case where Y follows a discrete-beta distribution, for a given n, several pairs of different (μ, ϕ), generate the same expectation, as we have shown in the previous chapter. In a natural way, according to CUB model framework, we prefer to directly introduce covariates for μ and ϕ, parameters of the latent variable, without a direct reference to the expectation of this random variable. It is reasonable to assume that the parameters of the latent variable vary with the subjects characteristics as, for instance, gender, age, etc. In order to specify a correspondence among the values of covariates and the supports of μ and ϕ, we proposed the following mappings:

$$\text{logit}\,(\mu_i) = \beta_0 + \mathbf{x}_i^{\mu}\boldsymbol{\beta} \quad \text{or equivalently} \quad \mu_i = \frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i^{\mu}\boldsymbol{\beta}}} \tag{3.23}$$

$$\log\,(\phi_i) = \gamma_0 + \mathbf{x}_i^{\phi}\boldsymbol{\gamma} \quad \text{or equivalently} \quad \phi_i = e^{\gamma_0 + \mathbf{x}_i^{\phi}\boldsymbol{\gamma}} \tag{3.24}$$

$$i = 1, \cdots, N;$$

where $\mathbf{x}_i^{\mu}$ and $\mathbf{x}_i^{\phi}$ are the vectors of covariates of the ith individual for, respectively, μ_i and ϕ_i, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vector of the "regression" coefficient $(\beta_1, \beta_2, \cdots, \beta_{p-1})$ and $(\gamma_1, \gamma_2, \cdots, \gamma_{q-1})$ ($p-1$ and $q-1$ are the number of covariates for, respectively, μ_i and ϕ_i). $\mathbf{x}_i^{\mu}$ and $\mathbf{x}_i^{\phi}$ are allowed to be different vectors. For example, let Y_i to be the ith individual's response of a product satisfaction survey, with outcomes in the set $\{0, 1, 2, 3\}$, $\mathbf{x}_i^{\mu}$ to be the vector which has for components the values of age, sex, marital status and $\mathbf{x}_i^{\phi}$ to be a null vector, we have

$$\text{logit}\,(\mu_i) = \beta_0 + \beta_1 \text{age} + \beta_2\ \text{sex} + \beta_3 \text{m.status}; \tag{3.25}$$

$$\log\,(\phi_i) = \gamma_0. \tag{3.26}$$

Note that the greater $\mathbf{x}_i^{\mu}\boldsymbol{\beta}$ the greater is μ_i. Thus, in this situation, it is not convenient to change signs as in the cumulative logistic proportional odds models, but it is preferable the mapping we showed. Similarly, the greater $\mathbf{x}_i^{\phi}\boldsymbol{\gamma}$ the greater is ϕ_i. In a vector notation, we can write

$$\text{logit}\,(\boldsymbol{\mu}) = \mathbf{X}^{\mu}\boldsymbol{\beta} \quad \text{or equivalently} \quad \boldsymbol{\mu} = \frac{1}{1 + e^{-\mathbf{X}^{\mu}\boldsymbol{\beta}}} \tag{3.27}$$

$$\log\,(\boldsymbol{\phi}) = \mathbf{X}^{\phi}\boldsymbol{\gamma} \quad \text{or equivalently} \quad \boldsymbol{\phi} = e^{\mathbf{X}^{\phi}\boldsymbol{\gamma}} \tag{3.28}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ are the $N \times 1$ vectors that contain respectively the μ-values and the ϕ-values of all N individuals, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \cdots, \gamma_p)$, i.e. we add the intercept term to the previous vector, $\mathbf{X}^{\mu}$ and $\mathbf{X}^{\phi}$ are the design matrices, respectively, $(N \times p)$ and $(N \times q)$ matrices, in which the ith row is the vector of the ith individual covariates.

If we assume that the responses of individuals are independent of one another given the discrete-beta probabilities, i.e. $Y_i \sim \text{Dbeta}(\mu_i, \phi_i, n)$, the likelihood is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}) &= P(\mathbf{y} = \mathbf{y}|\ \boldsymbol{\beta}, \boldsymbol{\gamma}) \\
&= \prod_{i=1}^{N} \left[\frac{\Gamma\left(e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}\right)}{\Gamma\left(\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}\right) \Gamma\left(\left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}\right)} \right. \\
&\left. \cdot \int_{\frac{y_i}{n+1}}^{\frac{y_i+1}{n+1}} x^{\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} - 1} (1-x)^{\left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} - 1} dx \right] \\
&= \prod_{i=1}^{N} \left[I_{\frac{y_i+1}{n+1}} \left(\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}, \left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} \right) + \right. \\
&\left. - I_{\frac{y_i}{n+1}} \left(\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}, \left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} \right) \right],
\end{aligned} \tag{3.29}$$

and the log-likelihood is written as

$$\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}) &= \log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}) \\
&= \sum_{i=1}^{N} \log \left[I_{\frac{y_i+1}{n+1}} \left(\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}, \left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} \right) + \right. \\
&\left. - I_{\frac{y_i}{n+1}} \left(\frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}} e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}}, \left(1 - \frac{1}{1+e^{-\mathbf{x}_i^{\mu}\boldsymbol{\beta}}}\right) e^{\mathbf{x}_i^{\phi}\boldsymbol{\gamma}} \right) \right].
\end{aligned} \tag{3.30}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$. In order to maximize eq. (3.30), finding estimate of $\hat{\boldsymbol{\theta}}$, with $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}\ , \hat{\boldsymbol{\gamma}})$, and to compute the p-values associated to the parameters estimates, we follow the

same way showed for the model without covariates. In particular, we have that

$$\frac{\partial l}{\partial \beta_c} = \sum_{i=1}^{N} \frac{\partial}{\partial \beta_c} f(\mu, \phi, y_i); \quad c = 1, \cdots, p; \tag{3.31}$$

$$\frac{\partial^2 l}{\partial \beta_d \partial \beta_c} = \sum_{i=1}^{N} \frac{\partial^2}{\partial \beta_d \partial \beta_c} f(\mu, \phi, y_i); \quad c = 1, \cdots, p; \ d = 1, \cdots, p; \tag{3.32}$$

$$\frac{\partial l}{\partial \gamma_c} = \sum_{i=1}^{N} \frac{\partial}{\partial \gamma_c} f(\mu, \phi, y_i); \quad c = 1, \cdots, q; \tag{3.33}$$

$$\frac{\partial^2 l}{\partial \gamma_d \partial \gamma_c} = \sum_{i=1}^{N} \frac{\partial^2}{\partial \gamma_d \partial \gamma_c} f(\mu, \phi, y_i); \quad c = 1, \cdots, q; \ d = 1, \cdots, q; \tag{3.34}$$

$$\frac{\partial^2 l}{\partial \gamma_d \partial \beta_c} = \sum_{i=1}^{N} \frac{\partial^2}{\partial \gamma_d \partial \beta_c} f(\mu, \phi, y_i); \quad c = 1, \cdots, p; \ d = 1, \cdots, q; \tag{3.35}$$

where $f(\mu, \phi, y_i)$ denotes the function $f(\mu, \phi, y_i) = \log P(Y_i = y_i | \mu, \phi)$. We can compute the first derivatives using the chain rule derivation

$$\frac{\partial f}{\partial \beta_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c} + \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \beta_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c}, \quad c = 1, \cdots, p; \tag{3.36}$$

$$\frac{\partial f}{\partial \gamma_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \gamma_c} + \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \gamma_c} = \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \gamma_c}, \quad c = 1, \cdots, q. \tag{3.37}$$

and the second derivatives as

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_d \partial \beta_c} &= \frac{\partial}{\partial \beta_d}\left(\frac{\partial f}{\partial \beta_c}\right) = \frac{\partial}{\partial \beta_d}\left(\frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c}\right) \\ &= \frac{\partial^2 l}{\partial \mu^2}\frac{\partial \mu}{\partial \beta_d}\frac{\partial \mu}{\partial \beta_c} + \frac{\partial f}{\partial \mu}\frac{\partial^2 \mu}{\partial \beta_d \partial \beta_c}, \quad c = 1, \cdots, p; \ d = 1, \cdots, p; \end{aligned} \tag{3.38}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \gamma_d \partial \gamma_c} &= \frac{\partial}{\partial \gamma_d}\left(\frac{\partial f}{\partial \gamma_c}\right) = \frac{\partial}{\partial \gamma_d}\left(\frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \gamma_c}\right) \\ &= \frac{\partial^2 l}{\partial \phi^2}\frac{\partial \phi}{\partial \gamma_d}\frac{\partial \phi}{\partial \gamma_c} + \frac{\partial f}{\partial \phi}\frac{\partial^2 \phi}{\partial \gamma_d \partial \gamma_c}, \quad c = 1, \cdots, q; \ d = 1, \cdots, q; \end{aligned} \tag{3.39}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \gamma_d \partial \beta_c} &= \frac{\partial}{\partial \gamma_d}\left(\frac{\partial f}{\partial \beta_c}\right) = \frac{\partial}{\partial \gamma_d}\left(\frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c}\right) \\ &= \frac{\partial^2 l}{\partial \phi \partial \mu}\frac{\partial \phi}{\partial \gamma_d}\frac{\partial \mu}{\partial \beta_c} + \frac{\partial f}{\partial \mu}\frac{\partial^2 \mu}{\partial \gamma_d \partial \beta_c} \end{aligned}$$

$$= \frac{\partial^2 l}{\partial\phi\partial\mu}\frac{\partial\phi}{\partial\gamma_d}\frac{\partial\mu}{\partial\beta_c} \qquad c = 1, \cdots, p;\ d = 1, \cdots, q. \tag{3.40}$$

The derivatives with respect to μ and ϕ are the same written in the previous section. We also have that

$$\frac{\partial\mu}{\partial\beta_c} = \frac{x_c}{2\left(1+\cosh(\mathbf{x}\boldsymbol{\beta})\right)}, \quad c = 1, \cdots, p; \tag{3.41}$$

$$\frac{\partial^2\mu}{\partial\beta_d\partial\beta_c} = \frac{-x_d x_c \sinh(\mathbf{x}\boldsymbol{\beta})}{4\left(1+\cosh(\mathbf{x}\boldsymbol{\beta})\right)^2}, \quad c = 1, \cdots, p; \quad d = 1, \cdots, p; \tag{3.42}$$

$$\frac{\partial\phi}{\partial\gamma_c} = x_c e^{\mathbf{x}\boldsymbol{\gamma}}, \quad c = 1, \cdots, q; \tag{3.43}$$

$$\frac{\partial^2\phi}{\partial\gamma_d\partial\gamma_c} = x_d x_c e^{\mathbf{x}\boldsymbol{\gamma}}, \quad c = 1, \cdots, q; \quad d = 1, \cdots, q; \tag{3.44}$$

where x_l denote the value of the vector $\mathbf{x}$ at the position l.

As in the previous section, it is possible to estimate them directly during the quasi-Newton algorithm. We suggest as starting values for the algorithm a vector with all components equal to 0.1, since, due the mappings 3.23, we not expect large values of parameters. We also ran the algorithm with different initial values and we found that our proposal is a good compromise.

3.2.1 Model selection

A general way to compare models is by means of the Akaike information criterion (AIC), that is a measure of the relative quality of a statistical model, for a given set of data. In general case, AIC is defined as

$$\text{AIC} = 2k - 2l, \tag{3.45}$$

where k is the number of parameters in the statistical model, and l is the maximized value of the log-likelihood function for the estimated model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing

function of the number of estimated parameters. The AICc, is AIC with a correction for finite sample sizes

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{N-k-1}, \tag{3.46}$$

where N denotes the sample size. Thus, AICc is AIC with a greater penalty for extra parameters. In [11] it is strongly recommended using AICc, rather than AIC.

For nested models, it is also possible to use likelihood ratio statistic. It is a very useful tool for judging the usefulness of inserting a single or a group of variable in the estimated models. Consider two models, m_0 and m_1, where m_0 is a submodel of model m_1, that is, m_0 is simpler than m_1 and m_0 is nested in m_1. The likelihood ratio statistic for the comparison of m_0 and m_1 is

$$\text{LR} = -2(l_0 - l_1), \tag{3.47}$$

where l_0 is the log-likelihood of the simpler model and l_1 is the log-likelihood of the more complex model. The likelihood ratio statistic measures the evidence in the data for the extra complexity in m_1 relative to m_0 [14]. The likelihood ratio statistic asymptotically follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameter of m_0 and m_1 [81].

3.3 Case study

We study the responses of seventy-one patients with gastroesophageal reflux disease (GERD). The GERD patients group comprises forty GERD patients with grade A esophagitis (ERD) according to the Los Angeles classification, and thirty-one patients with GERD who had nonerosive reflux disease (NERD), but 2 day wireless Bravo pH system monitoring (Medtronic A/S, Skovlunde, Denmark) positive for pathological acid exposure. The Bravo probes were placed using standard techniques as recommended by the manufacturer. The pH-monitoring was intended to record data for 48 hours. Food intake, typical GERD symptoms and supine period data where recorded in a diary. The receiver and the diary were returned after a 48-hours recording period. The pH study data were uploaded to a computer and were analysed using the software provided by Medtronic. Patients were considered to have had episodes of reflux when pH was less than 4 for at least six seconds; episodes were considered to have ended when pH reached 5. The total number of reflux episodes, number of reflux episodes longer than

five minutes, and the mean duration of reflux episodes were also determined.

All patients that entered the study were evaluated (by water load test) before and after four weeks of standard therapy with proton pump inhibitors (esomeprazole, 40 mg per day). Endoscopy was performed to evaluate carefully the distal portion of the esophagus to determine the presence of any mucosal injury.

As a control group, 30 healthy volunteers with no abdominal symptoms or history of upper gastrointestinal disorders were recruited. Non of the patients and controls had previously undergone abdominal surgery, except appendectomy.

The study were carried out according to local ethical rules, after receiving patients and controls' informed consent, and in accordance with the recommendations of the Helsinki Declaration (Edinburgh revision, 2000).

Water load test (WLT) Before starting the test, the participants completed a symptom visual analogue scale (VAS, 0-10 cm; 0= absent, 10= maximal) to score the following: heartburn, postprandial fullness, vomiting, early satiety, nausea, bloating, epigastric pain, belching, epigastric burning.

After an overnight fast, WLT was performed by having subjects drink-room-temperature water for 5 minute or until they perceiving the "full" stomach sensation. Water was consumed from an unmasked flask that was taken from the subjects and refilled after each drink. The volume required to refill the flask to the initial level was recorded and the total volume consumed was determined by summing these volumes.

Recently, the water load test has been proposed as a non-invasive method to assess gastric sensation. The test is economic, easily performed, well tolerated and reproducible in healthy subjects and in patients with functional dyspepsia or gastroesophageal reflux disease. In [4] are showed that in GERD patients, with mild erosive esophagitis and non erosive reflux disease, the WLT is abnormal, similar but non identical to that reported in patients with functional dyspepsia.

3.3.1 Model comparison

In the following section, we analysed the data sets, presented in the previous section, adopting different models:

- the cumulative link model (1.5) estimated with the R-package `ordinal`; we used the function `clm` with the logistic link function and flexible cut-points;

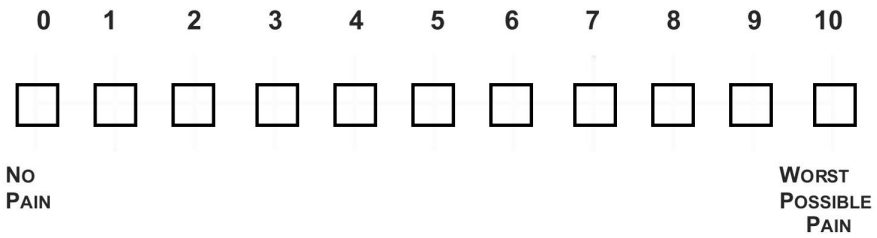

Figure 3.1: An example of VAS.

- the GAMLSS (Generalized Additive Models for Location, Scale and Shape) model [71]; we used the function `gamlss` from the R-package `gamlss`, with `family=BB`, i.e. we set the beta-binomial family;
- the CUB model, eq. (1.6); we use the R-script `CUB_3.0.R` kindly sent to us by Maria Iannario;
- the discrete-beta model, eq. (3.2); the script of the function `bdmod` is shown in the Appendix A.

We decided to analyse the score, VAS, of the patients before the therapy. In particular, the score of nausea, heartburn (pyrosis), vomiting, postprandial fullness, early satiety, bloating, epigastric pain and belching. We do not show the analysis of epigastric burning since covariates were proved to be insignificant. Due the predominance of 0 in the scores, we also do not include controls in the analysis (they have not abdominal symptoms or history of upper gastrointestinal disorder). Thus, our Y_i will be the score of the ith individual and the covariates should be BMI, age, sex, level of disease (ERD, NERD) and the volume of the water load test.

Nausea

Using the values of nausea as Y_i, we obtain with `clm`:

```
formula: nausea ~ BMI + Sesso
data:    gerd_pre_fac

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -115.31 252.61 8(1)  3.19e-11 2.4e+06

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
```

```
BMI     -0.5132     0.1686  -3.044  0.00233 **
Sessom  -0.6594     0.4569  -1.443  0.14892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
     Estimate Std. Error z value
1|2   -18.053      4.595  -3.929
2|3   -16.367      4.498  -3.639
3|4   -16.163      4.491  -3.599
4|5   -15.678      4.470  -3.507
5|6   -15.544      4.464  -3.482
6|7   -15.420      4.458  -3.459
7|8   -14.736      4.431  -3.326
8|9   -13.951      4.409  -3.164
9|10  -13.169      4.381  -3.006
```

with `gamlss`

```
Family:  c("BB", "Beta Binomial")

Call:  gamlss(formula = cbind(nausea, 10 - nausea) ~ BMI, sigma.formula = formulaphi,
    family = BB)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   11.7847      3.4801    3.386  0.001189
BMI           -0.3978      0.1327   -2.997  0.003824

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -0.5992      0.3757   -1.595   0.11546
Sessom        -0.9388      0.5383   -1.744   0.08576

------------------------------------------------------------------
No. of observations in the fit:  71
Degrees of Freedom for the fit:  4
      Residual Deg. of Freedom:  67
                      at cycle:  5

Global Deviance:     240.4093
            AIC:     248.4093
            SBC:     257.4601
```

```
**********************************************************************
```

with CUB:

```
======================================================================
============== CUB Program: version 3.0 (September 2013) ==============
======================================================================
=====>>> C U B (p,q) model <<<=====   ML-estimates via E-M algorithm
======================================================================
 Covariates for pai ==> p= 1      and     Covariates for csi ==> q= 2
======================================================================
*** m= 11   *** Sample size: n= 71    *** Iterations= 15  (maxiter=500)
======================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
======================================================================
beta_0       41.95214        16.19664         2.59018        0.009592576
beta_1       -1.54891        0.60845         -2.54567        0.01090682
gamma_0       -7.90875        2.91469         -2.71341        0.006659466
gamma_1       1.08501        0.39022         2.78051        0.005427359
gamma_2       -0.75267        0.37507         -2.00675        0.04477629
gamma_3       0.20597        0.11453         1.79839        0.07211523
======================================================================

Log-lik(beta^,gamma^)= -116.4404
Mean Log-likelihood  = -1.640006
----------------------------------------------------------------------
AIC-CUBpq          = 244.8809
BIC-CUBpq          = 258.4569
======================================================================
```

with discrete-beta model

```
bdmod(cbind(nausea,10-nausea),formulamu,formulaphi, gerd_pre)
$MU
              Estimate  Sd_error Wald_Test         Pval codemu
(Intercept)  9.4404285 2.8216381  3.345726 0.0008206741    ***
BMI         -0.3146795 0.1082288 -2.907539 0.0036428470     **

$PHI
              Estimate  Sd_error Wald_Test        Pval codephi
(Intercept)  0.9361496 0.3345929  2.797877 0.005143977      **
Sessom       1.0854145 0.4920932  2.205709 0.027404350       *
Esofagite1  -0.8391630 0.4866712 -1.724291 0.084655278       .

$AICc
[1] 249.2929

$AIC
[1] 248.3698
```

For each type of family model, we selected the one with the lower AIC and with the largest possible number of significant covariates. All comparisons are made with respect to the AIC index. When dropping the not significant covariates we obtain the same results with an AIC index larger, we prefer to present the model with also the not significant covariates but AIC index lower: this is the reason for which are also presented not significant covariates. `clm` tells us that BMI is significant and that the larger its value the larger the probability of the outcome to be in a lower rather than in a higher category. Discrete-beta and `gamlss` agree on this prediction. In addiction, our model states that males are more precise than females in the score (sex covariate is significant at 95%). In CUB model BMI, (β_1 value) plays an important role in π variable, the larger its value the larger the uncertainty. CUB also tells us that males have lower outcomes (γ_1 value) but ERD patients (significant at 95%) select a higher score than NERD patients.

Pyrosis

Using the values of nausea as Y_i, we obtain with `clm`:

```
formula: pirosis ~ Sesso + Max.vol..ml.
data:    gerd_pre_fac

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -112.20 244.40 7(2)  9.36e-13 8.5e+07

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Sessom       -1.83857    0.49357  -3.725 0.000195 ***
Max.vol..ml.  0.01121    0.00466   2.405 0.016152 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

with `gamlss`:

```
-------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
               Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)   -0.058899    1.300610  -0.04529  0.9640161
Sessom        -1.492741    0.340654  -4.38198  0.0000431
Max.vol..ml.   0.006076    0.003321   1.82935  0.0718640

-------------------------------------------------------------------
Sigma link function:  log
```

```
Sigma Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -6.51344    2.531444   -2.573   0.01234
Max.vol..ml.   0.01278    0.006194    2.064   0.04298

---------------------------------------------------------------------

Global Deviance:     235.3411
            AIC:     245.3411
            SBC:     256.6545
```

with CUB:

```
=============================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=============================================================================
beta_0       1.47632        0.55479         2.66104        0.00778997
beta_1       -1.47461        0.75106         -1.96337        0.0496032
gamma_0       39.23346        10.41457         3.76717        0.0001651086
gamma_1       1.55284        0.56187         2.7637        0.005715006
gamma_2       -7.14897        1.75627         -4.07054        4.690429e-05
=============================================================================
Log-lik(beta^,gamma^)= -123.9466
Mean Log-likelihood  = -1.745726
-----------------------------------------------------------------------------
AIC-CUBpq          = 257.8931
BIC-CUBpq          = 269.2065
=============================================================================
```

and with descrete-beta model:

```
$MU
                 Estimate    Sd_error  Wald_Test         Pval codemu
(Intercept)  -0.807761034 1.016133907 -0.7949356 4.266510e-01
Sessom       -1.324496171 0.272239596 -4.8651856 1.143494e-06    ***
Max.vol..ml.  0.007424267 0.002611504  2.8429091 4.470382e-03     **

$PHI
                Estimate    Sd_error Wald_Test         Pval codephi
(Intercept)   7.34821057 2.147437606  3.421851 0.0006219649     ***
Max.vol..ml. -0.01503162 0.005246766 -2.864930 0.0041710102      **

$AICc
[1] 246.2236

$AIC
[1] 245.3006
```

`clm` shows us that males are more likely to assign higher score than females and that patients who drink more water in WLT have a higher probability to score an higher value

of pyrosis. Discrete-beta and `gamlss` agree on these interpretations and, in addiction, tell us that the larger the value of the WLT the larger the variability (note that in `gamlss` the covariates are added on σ which is related to the reciprocal of the precision parameter: for this reason we have the opposite sign in covariate estimation respect to descrete-beta model). CUB agrees on the prediction of `clm` but states also that males have higher uncertainty than female.

Vomiting

Using the values of vomiting as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -148.75 321.50 5(1)  6.25e-07 2.1e+06

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
Sessom   0.7877     0.4332   1.818   0.0690 .
BMI     -0.3547     0.1452  -2.443   0.0146 *
```

with `gamlss`:

```
------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    7.3702      2.6544    2.777  0.007115
BMI           -0.3199      0.1025   -3.122  0.002652

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    8.9113        3.27    2.725  0.008190
BMI           -0.3868        0.13   -2.975  0.004075

------------------------------------------------------------------

Global Deviance:     297.6984
            AIC:     305.6984
            SBC:     314.7491
```

with CUB:

```
=====================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
```

```
=====================================================================
pai           0.43299        0.09164         4.7249        2.302286e-06
gamma_0      -37.24858         11.46304        -3.24945         0.001156284
gamma_1       1.62872        0.48567         3.35355         0.00079782
gamma_2      -3.4383         1.00161         -3.43277         0.0005974487
gamma_3      -2.25042         0.89331          -2.51919         0.01176252

=====================================================================
Log-lik(pai^,gamma^) = -154.2587
Mean Log-likelihood  = -2.172658
---------------------------------------------------------------------
AIC-CUB0q          = 318.5174
BIC-CUB0q          = 329.8308
```

with discrete-beta model:

```
$MU
              Estimate   Sd_error Wald_Test        Pval codemu
(Intercept)  6.9281974 2.45264541  2.824786 0.004731227     **
BMI         -0.2998144 0.09436039 -3.177333 0.001486360     **

$PHI
              Estimate  Sd_error Wald_Test        Pval codephi
(Intercept) -7.5106124 2.7592793 -2.721947 0.006489852      **
BMI          0.3302617 0.1081348  3.054166 0.002256875      **

$AICc
[1] 305.4879

$AIC
[1] 304.8819
```

CUB states that BMI (γ_1) is inversely proportional to the score and that males (γ_2) and ERD patients (γ_3) have higher probability to score higher outcomes than NERD patients. The other three models agree on BMI interpretation: the larger its value the larger the probability to be in a lower category of vomiting score. In addiction, `gamlss` and discrete-beta models state that a larger value of BMI also increases the precision, the accuracy of responses.

Postprandial fullness

Using the values of postprandial fullness as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -134.94 295.87 6(1)  2.80e-10 6.9e+05
```

```
Coefficients:
           Estimate Std. Error z value Pr(>|z|)
Sessom     -0.67983    0.45616  -1.490 0.136135
EtÃ         -0.07320    0.02105  -3.477 0.000507 ***
Esofagite1 -0.91591    0.45190  -2.027 0.042684 *
```

with `gamlss`:

```
Mu link function:  logit
Mu Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   4.18246     0.80470    5.198  2.187e-06
Sessom       -0.53376     0.32596   -1.637  1.064e-01
EtÃ           -0.05219     0.01491   -3.499  8.475e-04
Esofagite1    -0.72118     0.32775   -2.200  3.134e-02

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -4.79422    2.229721   -2.150   0.03527
Max.vol..ml.   0.00947    0.005376    1.762   0.08284

------------------------------------------------------------------
Global Deviance:     275.7001
            AIC:     287.7001
            SBC:     301.2762
```

with CUB:

```
 Covariates for pai ==> p= 2      and     Covariates for csi ==> q= 2
======================================================================
*** m= 11   *** Sample size: n= 71    *** Iterations= 20  (maxiter=500)
======================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
======================================================================
beta_0       1.10129        0.53259         2.0678        0.03865883
beta_1       -2.19267        0.73644         -2.97739        0.002907139
gamma_0       -1.24335        3.30461         -0.37625        0.706731
gamma_1       0.08638        0.02932         2.94611        0.003217979
gamma_2       -0.21406        0.11045         -1.93807        0.05261468
======================================================================
Log-lik(beta^,gamma^)= -140.596
Mean Log-likelihood  = -1.980225
----------------------------------------------------------------------
AIC-CUBpq          = 291.192
BIC-CUBpq          = 302.5054
======================================================================
```

with discrete-beta model:

```
$MU
               Estimate   Sd_error Wald_Test         Pval codemu
(Intercept)  3.74079080 0.69375592  5.392085 6.964487e-08    ***
Sessom      -0.45848689 0.27686134 -1.656016 9.771852e-02      .
EtÃ          -0.04602106 0.01282738 -3.587722 3.335799e-04    ***
Esofagite1  -0.67810256 0.28088377 -2.414175 1.577089e-02      *

$PHI
                Estimate    Sd_error Wald_Test        Pval codephi
(Intercept)   5.53554948 2.117477774  2.614218 0.008943183      **
Max.vol..ml. -0.01122309 0.004986512 -2.250689 0.024405229       *

$AICc
[1] 287.6956

$AIC
[1] 286.3831
```

According to all models, younger patients have higher probability to have higher values than older patients and, except the CUB model, that ERD patients have an higher probability to score lower category than NERD patients. CUB states that the higher the BMI value (γ_1) the higher the probability to have an higher outcome and that ERD patients have lower precision in choosing the score than NERD patients. Discrete-beta tells us that patients who drink more water in WLT have more uncertainty.

Early satiety

Using the values of early satiety as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -142.61 309.22 7(2)  4.92e-12 8.5e+06

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
EtÃ  -0.02958    0.01941  -1.524  0.12761
BMI -0.39867    0.13327  -2.991  0.00278 **
```

with `gamlss`:

```
Mu link function:  logit
Mu Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    9.7829     2.85941    3.421  0.001066
```

```
EtÃ            -0.0276     0.01439   -1.918  0.059329
BMI           -0.2992     0.10384   -2.881  0.005320

---------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
 -0.654221    0.228772   -2.859701    0.005648

---------------------------------------------------------------------
Global Deviance:     298.704
            AIC:     306.704
            SBC:     315.7547
```

with CUB:

```
parameters  ML-estimates  stand.errors    Wald-test      p-value
=========================================================================
pai           0.38807        0.07692         5.04511        4.532605e-07
gamma_0       -83.34086        19.24562         -4.33038        1.488522e-05
gamma_1       2.71526        0.62492         4.34497        1.392947e-05
gamma_2       0.16901        0.05503         3.07123        0.002131789
gamma_3       2.04906        0.70224         2.91789        0.003524086
=========================================================================
Log-lik(pai^,gamma^) = -149.832
Mean Log-likelihood  = -2.11031
-------------------------------------------------------------------------
AIC-CUB0q          = 309.664
BIC-CUB0q          = 320.9774
```

with discrete-beta model:

```
$MU
              Estimate   Sd_error Wald_Test         Pval codemu
(Intercept)  8.7735021 2.62505651  3.342215 0.0008311278    ***
EtÃ          -0.0259903 0.01319649 -1.969486 0.0488972688      *
BMI          -0.2649545 0.09518169 -2.783670 0.0053747641     **

$PHI
             Estimate  Sd_error Wald_Test         Pval codephi
(Intercept) 0.6680276 0.1875855  3.561189 0.0003691787     ***

$AICc
[1] 306.1333

$AIC
[1] 305.5272
```

The value of BMI is significant for all the models and the higher its value the higher the probability to have a lower category. CUB and discrete-beta models state that the

age is also significant and that the younger patients have an higher probability to score lower values. According to CUB and to expressed values, males declare to suffer less than females.

Bloating

Using the values of bloating as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -150.39 324.78 6(1)  6.32e-12 1.7e+08

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Max.vol..ml. -0.022057   0.005301  -4.161 3.17e-05 ***
Sessom        0.974014   0.451901   2.155   0.0311 *
```

with `gamlss`:

```
Mu link function:  logit
Mu Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)    5.82409    1.360187    4.282  6.047e-05
Max.vol..ml.  -0.01531    0.003493   -4.382  4.237e-05
Sessom         0.71046    0.307266    2.312  2.385e-02

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
-2.486e-03   5.698e-04  -4.363e+00   4.529e-05

------------------------------------------------------------------
Global Deviance:     318.3123
            AIC:     326.3123
            SBC:     335.363
```

with CUB:

```
=====================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=====================================================================
pai           0.38938        0.09113         4.2728        1.930336e-05
gamma_0       -152.6155        28.22168         -5.40774        6.382499e-08
gamma_1       -2.49755        0.75319         -3.31596        0.000913289
gamma_2       -2.25987        0.8711         -2.59427        0.009479201
gamma_3       25.75097        4.76509         5.40409        6.513823e-08
=====================================================================
```

```
Log-lik(pai^,gamma^) = -158.4807
Mean Log-likelihood  = -2.232122
------------------------------------------------------------------------
AIC-CUB0q          = 326.9614
BIC-CUB0q          = 338.2748
========================================================================
```

with discrete-beta model:

```
$MU
               Estimate    Sd_error Wald_Test         Pval codemu
(Intercept)   5.5703534 1.221067935  4.561870 5.069999e-06    ***
Max.vol..ml. -0.0145575 0.003143373 -4.631172 3.636018e-06    ***
Sessom        0.6562378 0.277251504  2.366941 1.793581e-02      *

$PHI
                Estimate     Sd_error Wald_Test        Pval codephi
Max.vol..ml. 0.002377821 0.0004109034  5.786812 7.17346e-09     ***

$AICc
[1] 326.6533

$AIC
[1] 326.0472
```

Males seem to have higher probability to have higher values than female and the higher the volume of water the higher the probability to have a lower score. According to CUB, ERD patients have higher values than NERD patients. Discrete-beta and `gamlss` state also that the higher the result of WLT the higher the precision of the score. Note that in CUB model we do not use WLT as covariate, but we prefer to add log(WLT) in order to avoid numerical problems (and because this transformation accelerates convergence).

Epigastric pain

Using the values of epigigastric pain as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -141.91 305.83 6(1)  8.57e-13 1.1e+08

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Max.vol..ml. -0.009392   0.004530  -2.073   0.0382 *
Sessom        0.705979   0.446095   1.583   0.1135
```

with `gamlss`:

```
Mu link function:  logit
Mu Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    0.880554    1.112603   0.7914   0.43148
Max.vol..ml.  -0.004654    0.002787  -1.6701   0.09956

---------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -4.6439     1.38886   -3.344  0.001357
EtÃ             0.0624     0.02664    2.342  0.022150

---------------------------------------------------------------------
Global Deviance:     288.1491
            AIC:     296.1491
            SBC:     305.1998
```

with CUB:

```
=====================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=====================================================================
beta_0       -282.161        583.5373         -0.48354        0.6287124
beta_1       14.24849        30.70544         0.46404         0.6426191
beta_2       39.18762        89.27738         0.43894         0.660705
beta_3       -1.97223        4.68276         -0.42117         0.6736309
gamma_0       -18.97485        6.6563         -2.85066         0.004362859
gamma_1       1.81832         1.04025         1.74796         0.08047095
gamma_2       0.36205         0.11137         3.25088         0.001150484
gamma_3       -0.47379         0.25938         -1.82663         0.06775542
=====================================================================
Log-lik(beta^,gamma^)= -137.2926
Mean Log-likelihood  = -1.933698
---------------------------------------------------------------------
AIC-CUBpq          = 290.5851
BIC-CUBpq          = 308.6866
=====================================================================
```

with discrete-beta model:

```
$MU
                 Estimate    Sd_error Wald_Test        Pval codemu
(Intercept)   2.122572117 0.970055583  2.188093 0.028662807      *
Max.vol..ml. -0.007460842 0.002464474 -3.027357 0.002467024     **

$PHI
                Estimate   Sd_error Wald_Test         Pval codephi
(Intercept)  3.29249069 0.78620792  4.187812 2.816571e-05     ***
```

```
EtÃ          -0.04079131 0.01672173 -2.439420 1.471087e-02       *

$AICc
[1] 300.5093

$AIC
[1] 299.9032
```

Discrete-beta and `clm` agree on the fact that WLT value is significant. In addition, CUB states also that males (γ_3) have higher probability to have a lower value than females. According to `gamlss` and our model, younger are more precise than older in giving the score.

Belching

Using the values of belching as Y_i, we obtain with `clm`:

```
 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  71   -146.96 315.91 7(2)  2.70e-12 1.2e+08

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Max.vol..ml.  0.011204   0.004815   2.327   0.0200 *
Esofagite1   -0.882794   0.475115  -1.858   0.0632 .
```

with `gamlss`:

```
Mu link function:  logit
Mu Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -2.056102    1.222433   -1.682   0.09730
Max.vol..ml.   0.006784    0.003267    2.076   0.04175
Esofagite1     -0.560593    0.305048   -1.838   0.07060

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -4.543903    2.061249   -2.204   0.03098
Max.vol..ml.   0.008863    0.005084    1.743   0.08595

------------------------------------------------------------------
Global Deviance:     323.3348
            AIC:     333.3348
            SBC:     344.6482
```

with CUB:

```
 Covariates for pai ==> p= 2      and     Covariates for csi ==> q= 4
=========================================================================
*** m= 11   *** Sample size: n= 71    *** Iterations= 22  (maxiter=500)
=========================================================================
parameters  ML-estimates  stand.errors    Wald-test      p-value
=========================================================================
beta_0       12.13277        24.53702         0.49447        0.6209743
beta_1       -1.56719        4.0673         -0.38531        0.7000078
beta_2       -0.0716        0.04198         -1.70557        0.08808817
gamma_0       64.71603        11.65825          5.55109        2.838939e-08
gamma_1       -13.45796        2.31666         -5.80921        6.276832e-09
gamma_2       1.99731        0.52652          3.79342        0.0001485865
gamma_3       0.06062        0.0234          2.5906        0.009580878
gamma_4       0.42286        0.17195          2.4592        0.0139247
=========================================================================
Log-lik(beta^,gamma^)= -156.5554
Mean Log-likelihood  = -2.205006
-------------------------------------------------------------------------
AIC-CUBpq          = 329.1109
BIC-CUBpq          = 347.2123
```

with discrete-beta model:

```
$MU
                 Estimate    Sd_error Wald_Test        Pval codemu
(Intercept)  -2.362550299 1.055355413 -2.238630 0.025180000      *
Max.vol..ml.  0.007512181 0.002862533  2.624313 0.008682399     **
Esofagite1   -0.550401594 0.270108595 -2.037705 0.041579459      *

$PHI
                 Estimate    Sd_error Wald_Test       Pval codephi
(Intercept)   4.803144878 1.629845778  2.946993 0.00320880      **
Max.vol..ml. -0.009633547 0.003985001 -2.417451 0.01562963       *

$AICc
[1] 334.6105

$AIC
[1] 333.6875
```

All the models agree on the significance of WLT values, but only CUB and our model state that ERD patients have higher probability to have lower values than NERD patients. In addition, CUB tells us that age (γ_3) and BMI (γ_4) are significant. On the other hand, for discrete-beta model patients who drink more water in WLT are less precise in giving the score.

3.3.2 Conclusion

Table 3.1: Summary of the result of the previous section. Models are presenting in non-decreasing order of the score.

Nausea	Model	AIC	Pirosis	Model	AIC	Vomiting	Model	AIC
	CUB	244.88		clm	244.40		D-beta	304.88
	D-beta	248.36		D-beta	245.30		gamlss	305.69
	gamlss	248.41		gamlss	245.34		CUB	318.51
	clm	252.61		CUB	257.89		clm	321.50
P.fullness	Model	AIC	E. satiety	Model	AIC	Bloating	Model	AIC
	D-beta	286.38		D-beta	305.52		clm	324.78
	gamlss	287.70		gamlss	306.70		D-beta	326.04
	CUB	291.19		clm	308.22		gamlss	326.31
	clm	295.87		CUB	309.66		CUB	326.96
Epig p.	Model	AIC	belching	Model	AIC			
	CUB	290.58		clm	315.91			
	gamlss	296.14		CUB	329.11			
	D-beta	299.90		gamlss	333.33			
	clm	305.83		D-beta	333.68			

In the table are summarized the result of AIC for each analysis. As we can note, discrete-beta models seems to be a good competitor with respect to the other standard methods. In the previous section we also note that the predictions of our model are very similar to those of `clm` and, above all, to those of `gamlss` model.

Chapter 4

Longitudinal ordinal data

One of the most common medical research designs is a pre-post study in which a single baseline health status measurement is obtained, an intervention is administered, and a single follow-up measurement is collected. The primary advantage of these studies is that they can investigate the *changes* in the outcomes. For example, if some subjects are given placebo while others are given an active drug, the two groups can be compared to see if the change in the outcome is different for those subjects who are actively treated as compared to control subjects. This design can be viewed as the simplest form of a prospective longitudinal study.

A *longitudinal study* refers to an investigation where participant (or object) outcomes are collected at multiple follow-up times.

Longitudinal studies generally yield multiple or repeated measurements on each subject. For example, HIV patients may be followed over time and monthly measures such as CD4 counts[1] or viral load are collected to characterize immune status and disease burden respectively. Such repeated measures data are correlated within subjects and thus require special statistical techniques for valid analysis and inference.

4.1 Discrete-beta model with random effects

In order to take into account this correlation between the repeated responses, we add *random effects* to the model (2.3) presented in the previous chapter. Mixed models have become very popular for the analysis of longitudinal data because they are flexible and

[1] CD4 cells are a type of white blood cell that fights infection. The CD4 count measures the number of CD4 cells in a sample of blood.

widely applicable. They assume that measurements from a single subject share a set of latent, unobserved, random effects which are used to generate an association structure between the repeated measurements [80].

Let y_{it}, $y_{it} \in \{0, \ldots, n\}$, be the outcome of the longitudinal ordinal variable Y_{it}, where $i \in \{0, 1, ..., N\}$ refers to individuals and $t \in \{0, 1, ..., T_i\}$ to the time; let $\mathbf{y}_i$ denote an $T_i \times 1$ vector of responses for the ith individual and assume that Y_{it}, given μ_{it} and ϕ_{it}, follows a discrete-beta distribution, i.e. $\left(Y_{it}|\mu_{it}, \phi_{it}\right) \sim \text{Dbeta}(\mu_{it}, \phi_{it}, n)$. These kinds of models are also called two-stage models. In the first stage, we assume that

$$\text{logit}(\boldsymbol{\mu}_i) = \mathbf{X}_i^{\mu}\boldsymbol{\beta} + \mathbf{Z}_i^{\mu}\mathbf{b}_{i\mu} \tag{4.1}$$

$$\log(\boldsymbol{\phi}_i) = \mathbf{X}_i^{\phi}\boldsymbol{\gamma} + \mathbf{Z}_i^{\phi}\mathbf{b}_{i\phi} \tag{4.2}$$

$$i = 1, \cdots, N;$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\phi}_i$ are the $T_i \times 1$ vectors, respectively of μ_{it} and ϕ_{it}. The population parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are treated as fixed effects, while $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ are random effects. Since we are in longitudinal case, random effects will be link to the time. Finally, $\mathbf{X}_i^{\mu}$ and $\mathbf{X}_i^{\phi}$ are design matrices, respectively of $T_i \times p$ and $T_i \times q$ dimensions, linking the fixed effects to $\boldsymbol{\mu}_i$ and $\boldsymbol{\phi}_i$; $\mathbf{Z}_i^{\mu}$ and $\mathbf{Z}_i^{\phi}$ are the matrices of between-subject covariates, respectively of $T_i \times k_{\mu}$ and $T_i \times k_{\phi}$ dimensions, associated to the random effects. If we suppose that the T_i values of $\mathbf{y}_i$, for the ith individual, are independent conditional on $\mathbf{b}_{i\mu}$, $\mathbf{b}_{i\phi}$ and the fixed-effects, the likelihood for the ith individual is

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) &= P(\mathbf{Y}_i = \mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) \\ &= \prod_{j=1}^{T_i}\left[I_{\frac{y_{ij}+1}{n+1}}\left(\mu_{ij}\phi_{ij}, (1-\mu_{ij})\phi_{ij}\right) - I_{\frac{y_{ij}}{n+1}}\left(\mu_{ij}\phi_{ij}, (1-\mu_{ij})\phi_{ij}\right)\right].\end{aligned} \tag{4.3}$$

At the second stage, we add assumptions on random effects. In particular, in our model $\mathbf{b}_{i\mu}$ is taken to be $\mathcal{N}(\mathbf{0}, \sigma_{\mu}^2\mathbf{D}_{\mu})$, and $\mathbf{b}_{i\phi}$ is taken to be $\mathcal{N}(0, \sigma_{\phi}^2\mathbf{D}_{\phi})$, independently

to each other and that satisfy the following condition

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i\mu} \\ \mathbf{b}_{i\phi} \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{D}_\mu & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_\phi \end{bmatrix} \right), \tag{4.4}$$

that is, its density function is

$$f_{\mathbf{b}_i}(\mathbf{b}_i) = \frac{1}{\sqrt{(2\pi)^{k_\mu + k_\phi}|\mathbf{D}|}} \exp\left(-\frac{1}{2} \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i \right). \tag{4.5}$$

Here $\mathbf{D}_\mu$ are $k_\mu \times k_\mu$ and $\mathbf{D}_\phi$ are $k_\phi \times k_\phi$ positive-definite covariance matrices, and $|\mathbf{D}|$ is the determinant of $\mathbf{D}$. In this case, we have a "conditional-independence model", since it implies that the T_i responses on individual i are independent conditional on $\mathbf{b}_i$ and the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Note that there is no requirement for balance in the data and that such "two-stage" model allows explicit modelling and analysis of between- and within- individual variation. Stage 1 allows modeling the within-subject variation (or occasion-to-occasion variation) separately for each subject. At Stage 2 we model the between-subject variation by postulating a distribution for the individual parameters $\mathbf{b}_i$.

Let $\boldsymbol{\theta}$ to be the vector of variance e covariance parameters found in $\mathbf{D}_\mu$ and $\mathbf{D}_\phi$, i.e. its length is $\left[k_\mu(k_\mu+1)/2 + k_\phi(k_\phi+1)/2\right]$, the classical approach is based on the maximum likelihood estimation of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ from the marginal distribution of $\mathbf{y}^T = (\mathbf{y}_1^T, ..., \mathbf{y}_N^T)$ [46, 79, 72]. Due to the chain rule[2] and the independence of $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ we have

$$\begin{aligned} f_{(\mathbf{Y}_i, \mathbf{b}_i)}(\mathbf{y}_i, \mathbf{b}_i) &= P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_i}(\mathbf{b}_i) \\ &= P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}). \end{aligned} \tag{4.6}$$

[2]In probability theory, the chain rule permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities.

$$P\left(\bigcap_{k=1}^{n} A_k \right) = \prod_{k=1}^{n} P\left(A_k \middle| \bigcap_{j=1}^{k-1} A_j \right), \quad k \geq 2.$$

Thus, the marginal distribution of $\mathbf{y}_i$ is given by

$$P(\mathbf{Y}_i = \mathbf{y}_i) = \int_{\mathbb{R}^\mu} \int_{\mathbb{R}^\phi} P(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})\ d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi}. \tag{4.7}$$

If we assume that observations on different individuals are independent, we obtain that the marginal likelihood of the data can be written in the following way

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{y}) &= \prod_{i=1}^{N} P(\mathbf{Y}_i = \mathbf{y}_i) \\
&= \prod_{i=1}^{N} \int_{\mathbb{R}^\mu} \int_{\mathbb{R}^\phi} P(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})\ d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi} \\
&= \prod_{i=1}^{N} \int_{\mathbb{R}^\mu} \int_{\mathbb{R}^\phi} \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})\ d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi},
\end{aligned} \tag{4.8}$$

and the log-likelihood is

$$\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{y}) &= \log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{y}) \\
&= \sum_{i=1}^{N} \log \int_{\mathbb{R}^\mu} \int_{\mathbb{R}^\phi} \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}|\ \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})\ d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi}.
\end{aligned} \tag{4.9}$$

Note that while the random effects $\mathbf{b}_i$ are not parameters, they are estimable quantities. The interest and focus on the unknown quantities $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\mathbf{b}_i$ depend on the objectives of the research. In a clinical trial the fixed treatment effects are usually of importance. In a biological application where the cluster of observations corresponds to a certain animal, or breed, the interest is in ranking the random effects and the best animal or breed. For genetic data the focus is on $\boldsymbol{\theta}$ which contains the components of genetic variability [79].

4.2 EM Algorithm

It is not trivial to maximize the likelihood (4.8): no closed-form expression is available and its computation involves multidimensional integrations. In this case, the estimators $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}})$ that maximize eq.(4.8) can be obtained using an EM algorithm. Its name stands for expectation-maximization, and it is so named because it alternates between calculating conditional expected values and maximized simplified likelihoods [68].

The EM algorithm [20] is a general-purpose optimization routine for computing maximum likelihood estimates. It was designed to be used for maximum likelihood estimation for situation in which augmenting the data set leads to a simpler problem. The actual data set is typically called the incomplete data in application of the EM algorithm. In the longitudinal data setting, the EM algorithm provides a convenient approach to computation, since the individual parameters $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ can be viewed as missing data.

If we were to observe $\mathbf{b}_{i\mu}$, $\mathbf{b}_{i\phi}$ in addition to $\mathbf{y}_i$, we could find simple closed-form maximum likelihood estimates of the component of $\boldsymbol{\theta}$:

$$\hat{\mathbf{D}}_\mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_{i\mu}\mathbf{b}_{i\mu}^T \qquad \hat{\mathbf{D}}_\phi = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_{i\phi}\mathbf{b}_{i\phi}^T, \tag{4.10}$$

or equivalently

$$\hat{\mathbf{D}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_i\mathbf{b}_i^T, \tag{4.11}$$

taking care to set $b_{ij}b_{ik}$ equal to zero for all $j = 1, \cdots k_\mu$ and $k = k_\mu + 1, \cdots, k_\mu + k_\phi$ (we have to remind condition (4.4)). For simplicity this condition is omitted from notation throughout this section, but when we write $\mathbf{b}_i\mathbf{b}_i^T$ we mean the matrix with this condition applied. Eq. (4.11) follows from the fact that, if $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ were observed, their likelihood would have the exponential-family form with sufficient statistic $\sum_{i=1}^{N} \mathbf{b}_i\mathbf{b}_i^T$, since they are known to have zero mean. If estimates of $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\gamma}}$, and $\tilde{\boldsymbol{\theta}}$, of, respectively, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are available, we can use them to compute "estimates" of the missing sufficient statistics, by setting them equal to their expectations, conditional on the observed data vector $\mathbf{y}$. For example, if we denote $\mathbf{t}_2 = \sum_{i=1}^{N} \mathbf{b}_i^T\mathbf{b}_i$, then we have

$$\tilde{\mathbf{t}}_2 = \mathbb{E}\left[\sum_{i=1}^{N} \mathbf{b}_i^T\mathbf{b}_i \middle| \mathbf{y}_i, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right]. \tag{4.12}$$

Substituting the numerator of (4.11) with the result of (4.12), we obtain a new $\tilde{\boldsymbol{\theta}}$, and, consequently $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$.

In a more formal way, at each iteration the expected of the augmented data log-likelihood is computed, conditional on the observed data and the current parameter values (E-step), than this pseudo-likelihood function is maximized to provide the next

parameter values (M-step). This gives an iterative scheme, as described in detail in the next two sections, that is used until it converges. Convergence is guaranteed under relatively unrestricted condition [83].

4.2.1 E-step

In the E-step we compute the expectation of the log-likelihood of the augmented data $(\mathbf{y}_i, \mathbf{b}_i)$ conditional on the observed data and the current parameter values $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\theta}}$. Let

$$\mathcal{Q}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}\right) = \mathbb{E}\left[l\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}; \mathbf{y}, \mathbf{b}\right) | \mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right]. \tag{4.13}$$

Under the hypothesis of independence, we have that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{b}) &= \prod_{i=1}^{N} f_{(\mathbf{Y}_i, \mathbf{b}_i)}(\mathbf{y}_i, \mathbf{b}_i) \\
&= \prod_{i=1}^{N} P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}) \\
&= \prod_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma} |\ \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})
\end{aligned} \tag{4.14}$$

and consequently

$$\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{b}) &= \log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{b}) \\
&= \log \left[\prod_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma} |\ \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}) \right] \\
&= \sum_{i=1}^{N} \log \left[P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}) \right] \\
&= \sum_{i=1}^{N} \left[\log P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) + \log f_{\mathbf{b}_i}(\mathbf{b}_i) \right] \\
&= \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \log P(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{b}_i) + \log f_{\mathbf{b}_i}(\mathbf{b}_i) \right] \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P(\mathbf{Y}_{it} = \mathbf{y}_{it} | \mathbf{b}_i) + \sum_{i=1}^{N} \log f_{\mathbf{b}_i}(\mathbf{b}_i).
\end{aligned} \tag{4.15}$$

Thus, $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ can be written as:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) &= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{T_i} \log P(\mathbf{Y}_{it} = \mathbf{y}_{it}|\mathbf{b}_i) + \sum_{i=1}^{N} \log f_{\mathbf{b}_i}(\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{T_i} \log P(\mathbf{Y}_{it} = \mathbf{y}_{it}|\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] + \mathbb{E}\left[\sum_{i=1}^{N} \log f_{\mathbf{b}_i}(\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{T_i} \mathbb{E}\left[\log P(\mathbf{Y}_{it} = \mathbf{y}_{it}|\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] + \sum_{i=1}^{N} \mathbb{E}\left[\log f_{\mathbf{b}_i}(\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] \\
&= \mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathcal{Q}_2(\boldsymbol{\theta}), \qquad (4.16)
\end{aligned}$$

where the two terms $\mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathcal{Q}_2(\boldsymbol{\theta})$ are

$$\mathcal{Q}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{N}\sum_{j=1}^{T_i} \mathbb{E}\left[\log P(\mathbf{Y}_{it} = \mathbf{y}_{it}|\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] \qquad (4.17)$$

$$\mathcal{Q}_2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbb{E}\left[\log f_{\mathbf{b}_i}(\mathbf{b}_i)\middle|\mathbf{y}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right]. \qquad (4.18)$$

In order to compute the conditional expectations, we need the density of $\mathbf{b}_i|\mathbf{y}_i$. It can be obtained by the Bayes' formula

$$f_{\mathbf{b}_i|\mathbf{y}_i}(\mathbf{b}_i|\mathbf{y}_i) = \frac{P(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{b}_i) f_{\mathbf{b}_i}(\mathbf{b}_i)}{P(\mathbf{Y}_i = \mathbf{y}_i)}. \qquad (4.19)$$

Substituting eq. (4.7) and (4.6) in eq. (4.19), we achieve the result.

The computation of the conditional expectations in eq. (4.17) and (4.18) is not trivial because they involve multidimensional integrals. These expectations are not available in closed form, but it possible to approximate these integrals by numerical methods or Monte Carlo simulations. For dimensions less or equal to 2, Gaussian quadrature may be used. We used `cubature` package in `R` software. This package allows to compute multiple integrals in an adaptive way where integrations are based on the algorithms described in [27, 6]. These algorithms are best suited for a moderate number of dimensions (say, < 7), and is superseded for high-dimensional integrals by other methods (e.g. Monte Carlo variants or sparse grids)[1]. An example of sparse grid method can be found in [35].

4.2.2 M-step

The M-step conveniently separates the estimation of the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from the variance components $\boldsymbol{\theta}$. It has different interpretations for $\mathcal{Q}_1$ and $\mathcal{Q}_2$ from the E-step.

To maximize $\mathcal{Q}_2$, note that this is the log-likelihood corresponding to N independent observations from the "prior" random effects distribution $f_{\mathbf{b}_i}(\mathbf{b}_i)$ where the standard sufficient statistics are replaced with their conditional expectations. In the case of a diagonal $\mathbf{D}$ the estimates are

$$\hat{\sigma}_g^2 = \frac{1}{N}\sum_{i=1}^{N} E\left[b_{ig}^2 | \mathbf{y}_i, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right] \qquad g = 1, \cdots, k_\mu + k_\phi \tag{4.20}$$

where $\hat{\sigma}_g^2$ denotes the g-element of the diagonal of $\mathbf{D}$. In general case, when the variance-covariance matrices $\mathbf{D}_\mu$ and $\mathbf{D}_\phi$ are unconstrained, $\mathcal{Q}_2$ is maximized by

$$\hat{\mathbf{D}} = \frac{1}{N}\sum_{i=1}^{N} E\left[\mathbf{b}_i \mathbf{b}_i^T | \mathbf{y}_i, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}\right], \tag{4.21}$$

that is the same formula that we achieve combining eq. (4.12) and (4.11).

About the first term, $\mathcal{Q}_1$, we have not an analytic solution and we must use numerical algorithms. As in the previous chapter, it is possible to maximize it with the quasi-Newton algorithm.

The EM algorithm starts with a set of initial values for the parameters. A good starting points for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is given by a regression with no random effect (described in the previous chapter). The initial value for $\mathbf{D}$ can be taken as the identity matrix. The algorithm then iterates between these two steps until convergence. Note that, due to the approximation of the E-step, the convergence of the log-likelihood is only approximately monotone. The algorithm is considered to converge when all parameter estimates become stable and no further improvements can be made to the likelihood value. However, to reduce computational time, since the rate of convergence is very slow [53], it is often common practice for the algorithm to be stopped before complete convergence using heuristic approaches. For example, in [79] the algorithm is stopped when the relative variation of all parameter values is less than 1 per cent (or after a fixed number of steps), then convergence is assessed by visual inspection and the final estimates are the average of the EM sequence over the convergent portion of the chain.

4.3 Variance-covariance matrix estimation

As in the previous section, we estimate the variance-covariance matrix of the vector $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}$ trough the inverse of observed information matrix. The EM algorithm only generates estimates and does not give the variance estimates as a byproduct, as do the quasi-Newton method. To obtain variance estimate, extra computations must be performed.

In [51], Louis derived a procedure for extracting the information matrix when the EM is used. The technique requires computation of a complete-data gradient vector or second derivatives matrix, but not those associated with the incomplete data likelihood. Let $\boldsymbol{\nu} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$, then

$$I(\boldsymbol{\nu}) = E\left[-\ddot{l}(\boldsymbol{\nu}; \mathbf{y}, \mathbf{b})|\mathbf{y}, \boldsymbol{\nu}\right] - E\left[\mathbf{s}(\boldsymbol{\nu}; \mathbf{y}, \mathbf{b})\mathbf{s}(\boldsymbol{\nu}; \mathbf{y}, \mathbf{b})^T|\mathbf{y}, \boldsymbol{\nu}\right], \tag{4.22}$$

where $\ddot{l}$ and s denote the matrix of second and vector of first derivatives of l with respect to $\boldsymbol{\nu}$. However, expectations must be computed. Thus, we preferred to estimate the second derivatives directly from the marginal log-likelihood (4.8) in a numerical way. Richardson's extrapolation joint to finite differences is a method for calculating (usually) accurate numerical first and second order derivatives. Simple difference method is also an option: it is usually less accurate but is much quicker than Richardson's extrapolation and provides a useful cross-check. `R`-package `numDeriv` implements these methods.

4.4 Prediction of random effects

In some applications the magnitude of random effects is of interest. Estimation, or prediction of random individual effects $\mathbf{b}_i$ is obtained by using an empirical Bayes strategy [72]

$$\hat{\mathbf{b}}_i = E\left[\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\nu}}\right], \tag{4.23}$$

with variance

$$\hat{\mathbf{v}}_i = \mathbb{V}\mathrm{ar}\left[\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\nu}}\right]. \tag{4.24}$$

Conditional on $\hat{\boldsymbol{\nu}}$, the random effects $\hat{\mathbf{b}}_i$ depend only on $\mathbf{y}_i$, the data from cluster i. Note that the asymptotic of $\hat{\boldsymbol{\nu}}$ is governed by the number of individuals N, whereas the asymptotic of $\hat{\mathbf{b}}_i$ is governed by the cluster size T_i.

4.5 Case Study

Returning to the case study presented in the previous subsection 3.3, after the therapy and the positive endoscopy, thirty-one patients (eighteen ERD patients and thirteen NERD patients) undergone surgery. Nissen fundoplication[3] (NF) is the most common and effective surgical treatment for gastroesophageal reflux disease. Previous studies have shown that after NF gastric emptying of solids and liquids is accelerated, the resting pressure of the lower esophageal sphincter (LES) is increased and transient LES relaxations are reduced.

After the surgery, the participants completed a symptom visual analogue scale (VAS, 0-10 cm; 0= absent, 10= maximal) to score the following: heartburn, postprandial fullness, vomiting, early satiety, nausea, bloating, epigastric pain, belching, epigastric burning.

Our Y_i will be the score of the ith individual and the covariates should be BMI, age, sex, level of disease (ERD, NERD) and the time (pre-post surgery).

4.5.1 Model comparison

In the following section, we analysed the data set, presented in the previous section, adopting our model and a generalized linear model with binomial family. We present only results on nausea symptom. With `binomial` model we obtain:

```
Generalized linear mixed model fit by the Laplace approximation
Formula: cbind(nausea, 10 - nausea) ~ BMI + esofagite + tempo + (1 | Paziente)
+ (0 + tempo | Paziente)
   Data: Nissen_t
   AIC   BIC logLik deviance
 145.4 162.4 -64.68    129.4
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Paziente (Intercept) 0.030433 0.17445
 Paziente tempo0      1.653684 1.28596
          tempo1      1.887979 1.37404  0.444
Number of obs: 62, groups: Paziente, 31

Fixed effects:
```

[3] In a fundoplication, the gastric fundus (upper part) of the stomach is wrapped, or plicated, around the lower end of the esophagus and stitched in place, reinforcing the closing function of the lower esophageal sphincter. The esophageal hiatus is also narrowed down by sutures to prevent or treat concurrent hiatal hernia, in which the fundus slides up through the enlarged esophageal hiatus of the diaphragm. In a Nissen fundoplication, also called a complete fundoplication, the fundus is wrapped all the way 360 degrees around the esophagus.

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.7241     3.3869   1.395   0.1631
BMI          -0.2339     0.1334  -1.753   0.0796 .
esofagiteS    0.9790     0.5028   1.947   0.0515 .
tempo1       -2.1056     0.3727  -5.650  1.6e-08 ***
---
```

and with discrete-beta model

```
$MU
               Estimate  Sd_error       Pval
(Intercept)  1.65428385 0.9632638 0.08591093      .
Sessom       0.44350422 0.2489684 0.07485239      .
BMI         -0.11338419 1.8939167 0.95226111
eta          0.01173522 0.9964852 0.99060384
tempo1      -1.83604170 0.9205365 0.04609369      *

$PHI
               Estimate  Sd_error         Pval
(Intercept) -2.05315175 0.7548059 0.0065261101      **
eta          0.07459756 0.3748485 0.8422570700
tempo1       2.88745270 0.8230477 0.0004510739     ***

$VarcovMU
             (Intercept)     tempo1
(Intercept)  0.64283059 0.04961232
tempo1       0.04961232 0.70120039

$AIC
[1] 242.1153
```

Both models say that patients, after the surgery, have an higher probability to score lower values than before the surgery. Our model also state that patients become more precise after NF. As we note from the AIC, the discrete-beta models could be probably improved eliminating some covariates.[4]

[4]Since the EM algorithm has a slow convergence, when we noted that the changes in $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ have a little influence on $\mathcal{Q}_1$, we stopped it. The we set $\hat{\boldsymbol{\theta}}_F$ equal to the $\hat{\boldsymbol{\theta}}$ obtained from the last step of the EM and we maximized directly the marginal log-likelihood in order to obtain the final estimate of $\hat{\boldsymbol{\beta}}_F$ and $\hat{\boldsymbol{\gamma}}_F$ with $\hat{\boldsymbol{\theta}}_F$. We also use the correction, proposed in [67], of second derivatives if it is necessary. Other solutions to this numerical problem are under investigation.

Conclusions

In this work we proposed a new discrete probability distribution useful when we work with ordered categorical data. The discrete-beta distribution, Dbeta(μ, ϕ, n), has a highly flexible shape and it can be either over-dispersed or under-dispersed with respect to the binomial distribution. It has only two parameters, μ and ϕ, which have a very clear interpretation: μ is the mean of the beta latent variable and the greater its value the higher the probability to obtain higher score; ϕ is a precision parameter (of the latent variable) and the higher its value the lower the variance. For $n > 2$ the model is identifiable and, adding directly covariates on parameters μ and ϕ (according to CUB model framework), it is very suitable to regression.

The assumption of the equispaced cut-points, combined with the finite support of the beta distribution, allows us to decrease the number of parameters in the model and to obtain the flexible shapes described previously in this work. On the other hand, except the U shapes, the model can not take into account bimodal shapes and so on. Thus, in these cases our model is not recommended. Future improvements should regard the mixture with an other distribution, such as an uniform distribution or another beta distribution, in order to take into account more uncertainty and bimodal shapes. Another idea is to change the set of cut-points, without increase exponentially the number of parameters.

We also introduced random effects in order to take into account the correlations in longitudinal data. An EM algorithm is proposed to find estimation of parameters. Further investigations need to be done to increase the speed of convergence of the algorithm and to achieve a better approximation of parameter estimates.

Other possible investigations involve the introduction of the Bayesian information criterion (BIC), which is a criterion for model selection, based, in part, on the likelihood function and closely related to the Akaike information criterion (AIC), and the

comparisons with CUBE models. In this thesis, we restricted the comparisons between the most used model, the cumulative logistic regression, and the other models with one or two parameters (CUB, beta-binomial and binomial models).

Appendix A

In this appendix, R-script for maximum likelihood estimation, for single population and in presence of covariates are shown.

For single population:

```
x <- rbetadiscr(100,size,mu,phi)
t1 <- proc.time()
stima=optim(par=c(mean(x),1/var(x)), fn=verosim, x=x, size=size,  method = "L-BFGS-B", hessian=T,
lower = c(2^{-1074}, 0+.Machine$double.eps),
upper = c(1-.Machine$double.eps, Inf))
t2 <- proc.time()
t2-t1

hess1 <- stima$hessian

hess3 <- solve(hess1)
sdval2 <- diag(hess3)
sdval <- sqrt(sdval2)

z<- qnorm(0.975)
interval_mu <- c( stima$par[1] - z*sdval[1], stima$par[1] + z*sdval[1])
interval_phi <- c( stima$par[2] - z*sdval[2], stima$par[2] + z*sdval[2])
sol_mu <- matrix(round(c(stima$par[1], interval_mu, stima$par[1]/sdval[1]),5), nrow=1)
dimnames(sol_mu)[[2]] <- c("estimate","lower bound", "upper bound", "Wald test")

sol_phi <- matrix(round(c(stima$par[2], interval_phi, stima$par[2]/sdval[2]),5), nrow=1)
dimnames(sol_phi)[[2]] <- c("estimate","lower bound", "upper bound", "Wald test")

list( mu= sol_mu, phi= sol_phi)
```

In presence of covariates:

```
bdmod <- function(y,formulamu, formulaphi, data)
{
#creo la matrice di design per mu
mat_mu <- model.matrix(formulamu, data)
mat_mup <<- mat_mu
```

```
#creo la matrice di design per phi
mat_phi <- model.matrix(formulaphi, data)
mat_phip <<- mat_phi

n <- dim(data)[1]

size <- sum(y[1,])

# funzione di log verosim da massimizzare

logverosim <- function(par,x,size,mat_mu,mat_phi){
beta_mu <- par[1:dim(mat_mu)[2]]
beta_phi <- par[(dim(mat_mu)[2]+1) : (dim(mat_mu)[2]+dim(mat_phi)[2])]


mu <- 1/(1+exp(- (mat_mu %*% beta_mu))) #vettore dei mu
phi <- exp(mat_phi %*% beta_phi) #vettore dei phi
mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))

alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))

a <- pbeta(x/(size +1),alfa1,alfa2)
b <- pbeta((x+1)/(size +1),alfa1,alfa2)
#if ( sum(is.na(a))!= 0) cat("problema con mu=", mu, " e phi=", phi, "\n ")
#if ( sum(is.na(b))!= 0) cat("problema con mu=", mu, " e phi=", phi, "\n ")
l <- sum(log(b-a))
-l
}

stima = optim(par=rep(0.1, dim(mat_mu)[2] + dim(mat_phi)[2]) , fn=logverosim,
x=y[,1], size=size, mat_mu=mat_mu, mat_phi=mat_phi,
method = "BFGS", hessian=TRUE )

per_mu = matrix(stima$par[1:dim(mat_mu)[2]], ncol=1)
per_mu = data.frame(per_mu)
dimnames(per_mu)[[1]] <- dimnames(mat_mu)[[2]]
dimnames(per_mu)[[2]] <- c("Estimate")

per_phi = matrix(stima$par[(dim(mat_mu)[2]+1) : (dim(mat_mu)[2]+dim(mat_phi)[2])], ncol=1)
per_phi = data.frame(per_phi)
```

```
dimnames(per_phi)[[1]] <- dimnames(mat_phi)[[2]]
dimnames(per_phi)[[2]] <- c("Estimate")


simm <- stima$hessian

simm <- solve(simm)
sdval <- diag(simm)
sdval <- sqrt(sdval)

pvalmu <- 2*(1-pnorm(abs(per_mu[,1]/sdval[1:dim(per_mu)[1]])))
pvalphi <- 2*(1-pnorm(abs(per_phi[,1]/sdval[-(1:dim(per_mu)[1])])))

codemu <- rep(" ", dim(per_mu)[1])
codemu[which(pvalmu<=0.001)] <- rep("***",length(pvalmu[pvalmu<=0.001]))
codemu[which(pvalmu>0.001 & pvalmu<=0.01)] <-
rep("**",length(which(pvalmu>0.001 & pvalmu<=0.01)))
codemu[which(pvalmu>0.01 & pvalmu<=0.05)] <-
rep("*",length(which(pvalmu>0.01 & pvalmu<=0.05)))
codemu[which(pvalmu>0.05 & pvalmu<=0.1)] <-
rep(".",length(which(pvalmu>0.05 & pvalmu<=0.1)))

codephi <- rep(" ", dim(per_phi)[1])
codephi[which(pvalphi<=0.001)] <- rep("***",length(pvalphi[pvalphi<=0.001]))
codephi[which(pvalphi>0.001 & pvalphi<=0.01)] <-
rep("**",length(which(pvalphi>0.001 & pvalphi<=0.01)))
codephi[which(pvalphi>0.01 & pvalphi<=0.05)] <-
rep("*",length(which(pvalphi>0.01 & pvalphi<=0.05)))
codephi[which(pvalphi>0.05 & pvalphi<=0.1)] <-
rep(".",length(which(pvalphi>0.05 & pvalphi<=0.1)))


per_mu <- cbind(per_mu, Sd_error= sdval[1:dim(per_mu)[1]],
Wald_Test=per_mu[,1]/sdval[1:dim(per_mu)[1]], Pval= pvalmu, codemu)
per_phi <- cbind(per_phi, Sd_error=sdval[-(1:dim(per_mu)[1])],
Wald_Test=per_phi[,1]/sdval[-(1:dim(per_mu)[1])], Pval=pvalphi, codephi)

k= length(stima$par)
N <- dim(data)[1]
AIC = 2*( k + stima$value)
AICc = AIC + + 2*k*(k+1)/(N -k -1)

list( MU = per_mu, PHI = per_phi, AICc = AICc, AIC= AIC)
}
```

Probability mass function, cumulative probability function, quantile function and random function of a discrete-beta distribution.

```
#####################################################################
```

```
#
# dbetadiscr(k,n,mu,sigma)
#
# k = outcome (between 0 and n)
# size = number of possibility
# mu = mean of the latent beta distribution
# phi = precision parameter (alpha + beta) of the lantent
#   beta distribution
#########################################################################

dbetadiscr <<- function(x,size,mu,phi,log=FALSE){

if (sum(x<0)!=0 | sum(x>size)!=0 | sum(round(x))!= sum(x))
cat ("ERROR: x must be integer between 0 and n \n")
else
if (size<=0 | round(size)!= size)
cat ("ERROR: size must be integer greater than 0 \n")
else
if (sum(mu<0)!=0 | sum(mu>1)!=0)
cat ("ERROR: mu must be between 0 and 1 \n")
else
if (sum(phi<0)!=0 )
cat ("ERROR: phi must be positive \n")

else
pbeta((x+1)/(size+1),mu*phi, (1-mu)*phi) -
pbeta((x)/(size+1),mu*phi, (1-mu)*phi)
}


#########################################################################
#
# pbetadiscr(x,size,mu,phi)
#
# x = real number or vector of real
# size = number of possibility
# mu = mean of the latent beta distribution
# phi = precision parameter (alpha + beta) of the lantent
#   beta distribution
#########################################################################

pbetadiscr <<- function(x,size,mu,phi,lower.tail = TRUE, log.p = FALSE){

if (size<=0 | round(size)!= size)
cat ("ERROR: size must be integer greater than 0 \n")
else
if (sum(mu<0)!=0 | sum(mu>1)!=0)
cat ("ERROR: mu must be between 0 and 1 \n")
```

```
else
if (sum(phi<0)!=0 )
cat ("ERROR: phi must be positive \n")
else {
k = floor(x)
pbeta((k+1)/(size+1),mu*phi, (1-mu)*phi,lower.tail = lower.tail,
        log.p = log.p)
  }
}
```

```
#########################################################################
#
# qbetadiscr(u,n,mu,phi)
#
# u = quantile (between 0 and 1)
# size = number of possibility
# mu = mean of the latent beta distribution
# phi = precision parameter (alpha + beta) of the lantent
#   beta distribution
#########################################################################

qbetadiscr <<- function(p,size,mu,phi,lower.tail = TRUE, log.p = FALSE){

if (sum(p<0)!=0 | sum(p>1)!=0)
cat ("ERROR: p must be between 0 and 1 \n")
else
if (size<=0 | round(size)!= size)
cat ("ERROR: size must be integer greater than 0 \n")
else
if (sum(mu<0)!=0 | sum(mu>1)!=0)
cat ("ERROR: mu must be between 0 and 1 \n")
else
if (sum(phi<0)!=0 )
cat ("ERROR: phi must be positive \n")
else
{
ris = ceiling(qbeta(p, mu*phi, (1-mu)*phi,lower.tail = lower.tail,
        log.p = log.p) * (size+1))-1
ris[ris==-1] = rep(0,length(ris[ris==-1]))
ris
}
}
```

```
#########################################################################
#
# rbetadiscr(n,size,mu,phi)
#
# n = number of observations. If length(n) > 1, the length is taken to be
# the number required.
# size = number of possibility
# mu = mean of the latent beta distribution
# phi = precision parameter (alpha + beta) of the lantent
#   beta distribution
#########################################################################


rbetadiscr <<- function(n,size,mu,phi){

if (n<=0 | round(n)!= n)
cat ("ERROR: n must be integer greater than 0 \n")
else
if (size<=0 | round(size)!= size)
cat ("ERROR: size must be integer greater than 0 \n")
else
if (mu<0 | mu>1)
cat ("ERROR: mu must be between 0 and 1 \n")
else
if (phi<0)
cat ("ERROR: phi must be positive \n")
else
{
a <- rbeta(n, mu*phi, (1-mu)*phi)
sol <- floor(a*(size+1))
sol[sol==size+1] = rep(size,length(sol[sol==size+1]))
sol
}
}
```

Appendix B

R-script for estimation in case of longitudinal data. In this script, random effects are added only for μ.

```
bdmod <- function(y,formulamu, formulaphi, randommu, data)
{
library(Matrix)
library(mnormt)
library(cubature)
library(MCMCpack)

attach(data)

size <- sum(y[1,])

#creo la matrice totale di design per mu
mat_mutot <- model.matrix(formulamu, data)

#creo la matrice totale di design per phi
mat_phitot <- model.matrix(formulaphi, data)

#########################################################################
# matrice sparsa dei random effect per mu e matrici parziali

a <- all.vars(randommu[[2]])
eval(parse(text=paste("f <- ~ ", a[1], sep=" ")))
datasplit <- split(data, eval(parse(text=a[2])))
mat_mu <- NULL
r_mu <- NULL
T <- dim(datasplit[[1]])[1]
for (i in (1:length(datasplit)) ) {
r_mu[[i]] <- model.matrix(f,datasplit[[i]])
mat_mu[[i]] <- model.matrix(formulamu, datasplit[[i]])
}
nbmu <- dim(r_mu[[1]])[2]
r_mutot <- bdiag(r_mu)

# matrice sparsa dei random effect per phi e matrici parziali

mat_phi <- NULL
for (i in (1:length(datasplit)) ) {
mat_phi[[i]] <- model.matrix(formulaphi, datasplit[[i]])
}
#########################################################################

detach(data)

lrmu <- dim(r_mu[[1]])[1]*(dim(r_mu[[1]])[1] +1)/2
theta <- rep(0, lrmu)
theta[ cumsum(rep(1,  dim(r_mu[[1]])[1]) + c(0, (dim(r_mu[[1]])[1]-1):1))] =  rep(1,  dim(r_mu[[1]])[1])
```

```
varcovmu <- xpnd( theta[ 1: lrmu])


verosimp <- function(par,x,size,mat_mu,mat_phi,r_mu,varcovmu){

beta_mu <- par[1:dim(mat_mu)[2]]
beta_phi <- par[-(1:dim(mat_mu)[2])]

fyi <- function(bstim,x,size,mat_mu,mat_phi,r_mu,varcovmu,beta_mu, beta_phi){

b_mu <- bstim
mu <- 1/(1+exp(- (mat_mu %*% beta_mu) - (r_mu %*% (b_mu))))
phi <- exp(mat_phi %*% beta_phi )

# controls
mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))

alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))


a <- pbeta(x/(size +1),alfa1,alfa2)
b <- pbeta((x+1)/(size +1),alfa1,alfa2)

l <- prod(abs(b-a))* dmnorm(b_mu, rep(0,dim(r_mu)[1]), varcovmu)
l
}

adaptIntegrate(fyi, lower=-10*diag(varcovmu), upper=10*diag(varcovmu), x=x,size=size,
mat_mu=mat_mu,mat_phi=mat_phi,r_mu=r_mu,varcovmu=varcovmu,beta_mu=beta_mu,
beta_phi=beta_phi, maxEval=15000 )$integral
}

### log-verosimiglianza totale

denominatore_fcondizionata <- function(par, y , T, size, datasplit, mat_mu, mat_phi, r_mu, varcovmu ){

lv <<- NULL
for (i in (1:length(datasplit)) ) {
lv <<- c(lv, verosimp(par=par,x=y[(i*T -T +1):(i*T),1],size=size,mat_mu=mat_mu[[i]],mat_phi=mat_phi[[i]],
r_mu=r_mu[[i]],varcovmu=varcovmu) )
}
lv
}



densityb2 <- function(bstim,b2,x,size,mat_mu,mat_phi,r_mu,varcovmu,beta_mu, beta_phi)
{
b_mu <- bstim
mu <- 1/(1+exp(- (mat_mu %*% beta_mu) - (r_mu %*% (b_mu))))
phi <- exp(mat_phi %*% beta_phi )

mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))
```

```
alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))

a <- pbeta(x/(size +1),alfa1,alfa2)
b <- pbeta((x+1)/(size +1),alfa1,alfa2)

l <- prod(bstim[b2])* prod(abs(b-a))* dmnorm(b_mu, rep(0,dim(r_mu)[1]), varcovmu)
l
}

logverosim_cond <- function(par,x,size,mat_mu,mat_phi,r_mu,varcovmu,beta_muv, beta_phiv){

beta_mu <- par[1:dim(mat_mu)[2]]
beta_phi <- par[(dim(mat_mu)[2]+1) : (dim(mat_mu)[2]+dim(mat_phi)[2])]

finteg <- function(bstim,x,size,mat_mu,mat_phi,r_mu,varcovmu,beta_mu,beta_phi,beta_muv, beta_phiv){

b_mu <- bstim[1:dim(r_mu)[1]]

mu <- 1/(1+exp(- (mat_mu %*% beta_mu) - (r_mu %*% b_mu)))
phi <- exp(mat_phi %*% beta_phi)


mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))

alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))


a <<- pbeta(x/(size +1),alfa1,alfa2) E
b <<- pbeta((x+1)/(size +1),alfa1,alfa2)
c <- b-a

muv <- 1/(1+exp(- (mat_mu %*% beta_muv) - (r_mu %*% b_mu)))
phiv <- exp(mat_phi %*% beta_phiv)

muv[muv==1] <- rep(1-.Machine$double.eps, length(muv[muv==1]))
muv[muv==0] <- rep(2^(-1074), length(muv[muv==0]))
phiv[phiv==0] <- rep(2^(-1074), length(phiv[phiv==0]))
phiv[phiv==Inf] <- rep(2^(1023), length(phiv[phiv==Inf]))

alfa1v <- muv*phiv
alfa2v <- (1-muv)*phiv
alfa1v[alfa1v==0] <- rep(2^(-1074), length(alfa1v[alfa1v==0]))
alfa1v[alfa1v==Inf] <- rep(2^(1023), length(alfa1v[alfa1v==Inf]))
alfa2v[alfa2v==0] <- rep(2^(-1074), length(alfa2v[alfa2v==0]))
alfa2v[alfa2v==Inf] <- rep(2^(1023), length(alfa2v[alfa2v==Inf]))


av <<- pbeta(x/(size +1),alfa1v,alfa2v)
bv <<- pbeta((x+1)/(size +1),alfa1v,alfa2v)
cv <- bv-av
lik <- prod(abs(c))
```

```
if (lik==0) lik=2^(-1074)
l <- log(lik)* prod(abs(bv-av))*dmnorm(b_mu, rep(0,dim(r_mu)[1]), varcovmu)
-l
}
adaptIntegrate(finteg, lower=-10*diag(varcovmu),
upper=+10*diag(varcovmu),
x=x,size=size,mat_mu=mat_mu,mat_phi=mat_phi,r_mu=r_mu,
varcovmu=varcovmu, beta_mu=beta_mu, beta_phi=beta_phi,beta_muv=beta_muv,beta_phiv=beta_phiv,
maxEval=15000 )$integral
}

fminim <- function(par, y , T, size, datasplit, mat_mu, mat_phi, r_mu,  varcovmu, varcovphi,
beta_muv, beta_phiv ){

lg <<- NULL
for (i in (1:length(datasplit)) ) {
lg <<- c(lg, logverosim_cond(par=par,x=y[(i*T -T +1):(i*T),1],size=size,mat_mu=mat_mu[[i]],
mat_phi=mat_phi[[i]],r_mu=r_mu[[i]],varcovmu=varcovmu,beta_muv=beta_muv, beta_phiv=beta_phiv) )
}
sum(lg/lv)
}

log_ver0 <- -Inf
log_ver1 <- Inf
n_iter = 1
parbeta <- NULL
pargamma <- NULL

inizio <- function(par,x,size,mat_mu,mat_phi){
beta_mu <- par[1:dim(mat_mu)[2]]
beta_phi <- par[(dim(mat_mu)[2]+1) : (dim(mat_mu)[2]+dim(mat_phi)[2])]

mu <- 1/(1+exp(- (mat_mu %*% beta_mu)))
phi <- exp(mat_phi %*% beta_phi)
mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))

alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))


a <- pbeta(x/(size +1),alfa1,alfa2)
b <- pbeta((x+1)/(size +1),alfa1,alfa2)
l <- sum(log(b-a))
-l
}


stima = optim(par=rep(0.1, dim(mat_mu[[1]])[2] + dim(mat_phi[[1]])[2]) , fn=inizio, x=y[,1], size=size,
 mat_mu=mat_mutot, mat_phi=mat_phitot,  method = "BFGS" )
beta_mu <- stima$par[1:dim(mat_mu[[1]])[2]]
beta_phi <- stima$par[-(1:dim(mat_mu[[1]])[2])]
log_ver <- NULL
thetatot <- theta

while ( abs(log_ver1 - log_ver0) > 0.001 & n_iter<=30 ){

cat("E \n")
```

```
theta1 <- matrix(rep(0, length(datasplit)*length(theta)), nrow= length(datasplit))

for (i in (1:length(datasplit)) ) {
k=1 #inizializzo il contatore della posizione di theta
for (j in (1: dim(varcovmu)[1]) ) { #ciclo su tutte le righe per theta-mu
for (s in (j:dim(varcovmu)[2])) {
theta1[i,k] <- adaptIntegrate(densityb2, lower=-10*diag(varcovmu),
upper=10*diag(varcovmu), b2= c(j,s),
x=y[(i*T -T +1):(i*T),1],size=size,mat_mu=mat_mu[[i]],mat_phi=mat_phi[[i]],r_mu=r_mu[[i]],
varcovmu=varcovmu,beta_mu=beta_mu, beta_phi=beta_phi, maxEval=15000 )$integral
k = k+1
}
}
}

lv <<- denominatore_fcondizionata(par=c(beta_mu,beta_phi), y=y , T=T, size=size, datasplit=datasplit,
 mat_mu =mat_mu, mat_phi=mat_phi, r_mu=r_mu, varcovmu=varcovmu)

if( any(lv==0) ) cat("problema con il denominatore")

theta1 <- theta1 / (matrix(rep(lv, dim(theta1)[2]), ncol= dim(theta1)[2]))
theta <- apply(theta1, 2, sum)/length(lv)

cat("M \n")
t1=proc.time()
stima = optim(par=c(beta_mu,beta_phi), fn=fminim, y=y, size=size,
datasplit= datasplit, T=T, r_mu=r_mu, varcovmu=varcovmu,
 mat_mu=mat_mu, mat_phi=mat_phi, beta_muv=beta_mu, beta_phiv=beta_phi,  method = "BFGS" )

beta_mu <- stima$par[1:dim(mat_mu[[1]])[2]]
beta_phi <- stima$par[-(1:dim(mat_mu[[1]])[2])]


varcovmu <- xpnd( theta[ 1: lrmu])

log_ver0 <- log_ver1
log_ver1 <- stima$value
n_iter <- n_iter+1

parbeta <- rbind(parbeta,beta_mu)
pargamma <- rbind(pargamma,beta_phi)
log_ver <- c(log_ver, log_ver1)
thetatot <- rbind(thetatot, theta)
}

verosim_der2 <- function(par,theta, y,mat_mu,mat_phi,size,r_mu,low,up){

nbeta <- dim(mat_mu)[2]
ngamma <- dim(mat_phi)[2]
beta_mu <- par[1:nbeta]
beta_phi <- par[(nbeta+1) : (nbeta+ngamma)]

varcovmu <- xpnd(theta)


finteg <- function(bstim,x,size,mat_mu,mat_phi,r_mu,varcovmu,low,up,beta_mu,beta_phi){

b_mu <- bstim[1:dim(r_mu)[1]]

mu <- 1/(1+exp(- (mat_mu %*% beta_mu) - (r_mu %*% b_mu)))
phi <- exp(mat_phi %*% beta_phi)

mu[mu==1] <- rep(1-.Machine$double.eps, length(mu[mu==1]))
mu[mu==0] <- rep(2^(-1074), length(mu[mu==0]))
phi[phi==0] <- rep(2^(-1074), length(phi[phi==0]))
```

```
phi[phi==Inf] <- rep(2^(1023), length(phi[phi==Inf]))

alfa1 <- mu*phi
alfa2 <- (1-mu)*phi
alfa1[alfa1==0] <- rep(2^(-1074), length(alfa1[alfa1==0]))
alfa1[alfa1==Inf] <- rep(2^(1023), length(alfa1[alfa1==Inf]))
alfa2[alfa2==0] <- rep(2^(-1074), length(alfa2[alfa2==0]))
alfa2[alfa2==Inf] <- rep(2^(1023), length(alfa2[alfa2==Inf]))


a <- pbeta(x/(size +1),alfa1,alfa2)
b <- pbeta((x+1)/(size +1),alfa1,alfa2)
c <- b-a

lik <- prod(abs(c))
if (lik==0) lik=2^(-1074)
l <- lik*dmnorm(b_mu, rep(0,dim(r_mu)[1]), varcovmu)
l
}
adaptIntegrate(finteg, lower=low, upper=up,
x=y,size=size,mat_mu=mat_mu,mat_phi=mat_phi,r_mu=r_mu,
varcovmu=varcovmu, beta_mu=beta_mu, beta_phi=beta_phi, maxEval=15000 )$integral
}

f <- function(par, theta, y , T, size, datasplit, mat_mu, mat_phi, r_mu, low, up){

ver <<- NULL
for (i in (1:length(datasplit)) ) {
#attach(datasplit[[i]]) #forse non mi serve
ver <<- c(ver, verosim_der2(par=par,theta=theta,y=y[(i*T -T +1):(i*T),1],size=size,
mat_mu=mat_mu[[i]],mat_phi=mat_phi[[i]],r_mu=r_mu[[i]], low=low,up=up) )
}
-sum(log(ver))
}


results <- optim(par=c(beta_mu,beta_phi) , fn=f,theta=theta, y=y, T=T, size=size, datasplit=datasplit,
mat_mu=mat_mu, mat_phi=mat_phi, r_mu=r_mu,low=low, up=up,  method = "BFGS", hessian=T)


per_mu = matrix(results$par[1:dim(mat_mu[[1]])[2]], ncol=1)
per_mu = data.frame(per_mu)
dimnames(per_mu)[[1]] <- dimnames(mat_mu[[1]])[[2]]
dimnames(per_mu)[[2]] <- c("Estimate")

per_phi = matrix(results$par[-(1:dim(mat_mu[[1]])[2])], ncol=1)
per_phi = data.frame(per_phi)
dimnames(per_phi)[[1]] <- dimnames(mat_phi[[1]])[[2]]
dimnames(per_phi)[[2]] <- c("Estimate")


simm <- risults$hessian
simm <- PDFORCE(simm)

simm <- solve(simm)
sdval <- diag(simm)
sdval <- sqrt(sdval)

pvalmu <- 2*(1-pnorm(abs(per_mu[,1]/sdval[1:dim(per_mu)[1]])))
pvalphi <- 2*(1-pnorm(abs(per_phi[,1]/sdval[-(1:dim(per_mu)[1])])))

codemu <- rep(" ", dim(per_mu)[1])
codemu[which(pvalmu<=0.001)] <- rep("***",length(pvalmu[pvalmu<=0.001]))
codemu[which(pvalmu>0.001 & pvalmu<=0.01)] <- rep("**",length(which(pvalmu>0.001 & pvalmu<=0.01)))
codemu[which(pvalmu>0.01 & pvalmu<=0.05)] <- rep("*",length(which(pvalmu>0.01 & pvalmu<=0.05)))
```

```
codemu[which(pvalmu>0.05 & pvalmu<=0.1)] <- rep(".",length(which(pvalmu>0.05 & pvalmu<=0.1)))

codephi <- rep(" ", dim(per_phi)[1])
codephi[which(pvalphi<=0.001)] <- rep("***",length(pvalphi[pvalphi<=0.001]))
codephi[which(pvalphi>0.001 & pvalphi<=0.01)] <- rep("**",length(which(pvalphi>0.001 & pvalphi<=0.01)))
codephi[which(pvalphi>0.01 & pvalphi<=0.05)] <- rep("*",length(which(pvalphi>0.01 & pvalphi<=0.05)))
codephi[which(pvalphi>0.05 & pvalphi<=0.1)] <- rep(".",length(which(pvalphi>0.05 & pvalphi<=0.1)))


per_mu <- cbind(per_mu, Sd_error= sdval[1:dim(per_mu)[1]], Pval= pvalmu, codemu)
per_phi <- cbind(per_phi, Sd_error=sdval[-(1:dim(per_mu)[1])], Pval=pvalphi, codephi)

k= length(c(risults$par,theta))
N <- dim(data)[1]
AIC = 2*( k + risults$value)

dimnames(varcovmu)[[1]] <- dimnames(r_mu[[1]])[[2]]
dimnames(varcovmu)[[2]] <- dimnames(r_mu[[1]])[[2]]

list( MU = per_mu, PHI = per_phi, VarcovMU = varcovmu, AIC= AIC)

}
```

Bibliography

[1] Cubature (multi-dimensional integration). http://ab-initio.mit.edu/wiki/index.php/Cubature, 11/12/2013.

[2] A. Agresti. *Analysis of ordinal categorical data (Second Edition)*. Wiley, 2010.

[3] J. A. Anderson and R. P. Philips. Regression, discrimination and measurement models for ordered categorical variables", journal = "Applied Statistics. 30:22–31, 1981.

[4] E. Battaglia, M. Grassini, M. Navino, P. Niola, C. Verna, A. Mazzocchi, C. Clerici, A. Morelli, and G. Bassotti. Water load test before and after ppi therapy in patients with gastro-oesophageal reflux disease. *Digestive and Liver Disease*, 39:1052–1056, 2007.

[5] M.P. Becker. *Encyclopedia of biostatistics*, volume 6, chapter Ordered categorical data, pages 3869–3876. John Wiley, 2005.

[6] J. Berntsen, T. O. Espelid, and A. Genz. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software (TOMS)*, 17:437–451, 1991.

[7] R.D. Bock and L.V. Jones. *The measurement and prediction of judgment and choice*. Holden-Day, San Francisco, 1968.

[8] G.E. Bonney. Logistic regression for dependent binary observations. *Biometrics*, 43:951–973, 1987.

[9] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society. Series B*, 61:265–285, 1999.

[10] N. E. Breslow and D. G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[11] K.P. Burnham and D.R. Anderson. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach (2nd ed.)*. Springer, 2002.

[12] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, September 1995.

[13] G. Casella and R.L. Berger. *Statistical inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[14] R. H. B. Christensen. Analysis of ordinal data with cumulative link models - estimation with the r-package ordinal. Technical report, 2013.

[15] B. A. Coull and A. Agresti. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, 56:73–80, 2000.

[16] C. Cox. Location-scale cumulative odds models for ordinal data: A generalized nonlinear model approach. *Statistics in Medicine*, 14:1191–1203, 1995.

[17] R. Crouchley. A Random-Effects Model for Ordered Categorical Data. *Journal of the American Statistical Association*, 90:489–498, 1995.

[18] R. W. Mee D. A. Harville. A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, 40:393–408, 1984.

[19] A. D'Elia and D. Piccolo. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49:917 – 934, 2005.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[21] J. E. Dennis, Jr. and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16)*. Soc for Industrial & Applied Math, 1996.

[22] Bas Engel. A simple illustration of the failure of pql, irreml and aphl as approximate ml methods for mixed models for binary data. *Biometrical Journal*, 40, 1998.

[23] F. Ezzet and J. Whitehead. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, 10:901–906, 1991. (Comment and Reply: 12 (1993) 2147-2151).

[24] S. Ferrari and F.o Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31:799–815, 2004.

[25] M. Ganjali. Fitting transitional models to longitudinal ordinal response data using availabel software. 2010. ICOTS8 Invited Paper.

[26] M. Gasparini. *Modelli probabilistici e statistici*. CLUT, 2006.

[27] A. C. Genz and A.A. Malik. Remarks on algorithm 006: An adaptive algorithm for numerical integration over an n-dimensional rectangular region. *Journal of Computational and Applied Mathematics*, 6:295–302, 1980.

[28] L. A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74:537–552, September 1979.

[29] L. A. Goodman. The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics*, 39:149–160, 1983.

[30] J Hartzel, A Agresti, and B Caffo. Multinomial logit random effects models. *Statistical Modelling*, 1:81–102, 2001.

[31] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

[32] P. J. Heagerty and S. L. Zeger. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91:1024–1036, 1996.

[33] D. Hedeker and R. D. Gibbons. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944, 1994.

[34] D. Hedeker and R. D. Gibbons. Mixor: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.

[35] F. Heiss and V. Winschel. Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144:62–80, 2008.

[36] M. Iannario. Hierarchical cub models for ordinal variables. *Communications in Statistics-Theory and Methods*, 41(16-17):3110–3125, 2012.

[37] M. Iannario. Modelling shelter choices in a class of mixture models for ordinal responses. *Statistical Methods & Applications*, 21:1–22, 2012.

[38] M. Iannario. Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics - Theory and Methods*, 43(4):771–786, 2014.

[39] M. Iannario and D. Piccolo. A new statistical model for the analysis of customer satisfaction. *Quality Technology & Quantitative Management*, 7, 2010.

[40] Maria Iannario. Cube models for interpreting ordered categorical data. *Quaderni di Statistica*, 14:137–140, 2012.

[41] N.L. Johnson and S. Kotz. *Distributions in statistics.* Wiley series in probability and mathematical statistics. Wiley, 1970.

[42] V. E. Johnson and J. H. Albert. *Ordinal data modeling.* Statistics for Social Science and Public Policy. New York, NY: Springer. 258 p. , 1999.

[43] J. D. Kalbfleisch and J. F. Lawless. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80:863–871, 1985.

[44] G. Kauermann. Modeling longitudinal data with ordinal response by varying coefficients, 1999.

[45] MG Kenward, E Lesaffre, and G Molenberghs. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, 50:945–953, 1994.

[46] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, December 1982.

[47] K. Lee and M. J. Daniels. Marginalized models for longitudinal ordinal data with application to quality of life studies. *Statistics in Medicine*, 27:4359–4380, 2008.

[48] S. R. Lipsitz, K. Kim, and L. Zhao. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13:1149–1163, 1994.

[49] I. Liu and A. Agresti. The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 14:1–73, 2005.

[50] Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81:624–629, 1994.

[51] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[52] P. Mc Cullagh. Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society*, 42:109–142, 1980.

[53] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[54] Donald G Morrison. Purchase intentions and purchase behavior. *The Journal of Marketing*, pages 65–74, 1979.

[55] L. R. Muenz and L. V. Rubinstein. Markov models for covariate dependence of binary sequences. *Biometrics*, 41:91–101, 1985.

[56] G. Muniz-Terrera, A. van den Hout, R.A. Rigby, and D.M. Stasinopoulos. Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical Methods in Medical Research*, 2012.

[57] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer series in operations research and financial engineering. Springer, 1999.

[58] M. Palta and C. Lin. Latent variables, measurement error and methods for analysing longitudinal binary and ordinal data. *Statistics in Medicine*, 18:385–396, 1999.

[59] B. Peterson and F. Harrell. Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39:205–217, 1990.

[60] D. Piccolo. On the Moments of a Mixture of Uniform and Shifted Binomial random variables. *Quaderni di Statistica.*, 5, 2003.

[61] D. Piccolo. Observed information matrix in MUB models. *Quaderni di Statistica.*, 8:33–78, 2008.

[62] D. Piccolo. Inferential issues in cube models with covariates. *Communications in Statistics - Theory and Methods*, 43:forthcoming, 2014.

[63] D. Piccolo and A. D'Elia. A new approach for modelling consumers' preferences. *Food Quality and Preference*, 19:247 – 259, 2008.

[64] J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.

[65] A. Punzo. Discrete beta-type models. In Hermann Locarek-Junge and Claus Weihs, editors, *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 253–261. Springer Berlin Heidelberg, 2010.

[66] H. J. Ribaudo, M. Bacchi, J. Bernhard, and S. G. Thompson. A multilevel analysis of longitudinal ordinal data: evaluation of the level of physical performance of women receiving adjuvant therapy for breast cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162:349–360, 1999.

[67] L. R. Schaeffer. Modification of negative eigenvalues to create positive definite matrices and approximation of standard errors of correlation estimates. http://lirpa.aps.uoguelph.ca/elares/sites/default/files/PDforce.pdf, 18/12/2013.

[68] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance components*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1992.

[69] G. Simon. Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69:971–976, December 1974.

[70] E. J. Snell. A scaling procedure for ordered categorical data. *Biometrics*, 20:592–607, 1964.

[71] D. Mikis Stasinopoulos and Robert A. Rigby. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23:1–46, 12 2007.

[72] R. Stiratelli, N. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971, 1984.

[73] T.R. Ten Have. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics*, 52:473–491, 1996.

[74] A. Y. Toledano and C. Gatsonis. Ordinal regression methodology for roc curves derived from correlated data. *Statistics in Medicine*, 15:1807–1826, 1996.

[75] G. Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43:39–55, 1990.

[76] G. Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11:275–295, May 1991.

[77] G. Tutz. Generalized semiparametrically structured ordinal models. *Biometrics*, 59:263–273, 2003.

[78] G. Tutz and W. Hennevogl. Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, 22:537–557, September 1996.

[79] F. Vaida and R. Xu. Proportional hazards model with random effects. *Statistics in medicine*, 19:3309–3324, 2000.

[80] G. Verbeke, G. Molenberghs, and D. Rizopoulos. *Longitudinal Research with Latent Variables*, chapter 2: Random Effects Models for Longitudinal Data, pages 37–96. Springer, 2010.

[81] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

[82] O. D. Williams and J. E. Grizzle. Analysis of Contingency Tables Having Ordered Response Categories. *Journal of the American Statistical Association*, 67:55–63, March 1972.

[83] C.F. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11:95–103, 1983.

Part II
A Proximity-Based Method to Identify Genomic Regions Correlated with a Continuously Varying Environmental Variable

Evolutionary Bioinformatics

METHODOLOGY

A Proximity-Based Method to Identify Genomic Regions Correlated with a Continuously Varying Environmental Variable

Cornelia Di Gaetano[1,2], Giuseppe Matullo[1,2], Alberto Piazza[1,2], Moreno Ursino[3] and Mauro Gasparini[3]

[1]Department of Genetics, Biology and Biochemistry, University of Turin, Turin, Italy. [2]HuGeF, Human Genetics Foundation, Turin, Italy. [3]Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy.
Corresponding author email: cornelia.digaetano@unito.it

Abstract: Knowledge of markers in the human genome which show spatial patterns and display extreme correlation with different environmental determinants play an important role in understanding the factors which affect the biological evolution of our species. We used the genotype data of more than half a million single nucleotide polymorphisms (SNPs) from the data set Human Genome Diversity Panel (HGDP-CEPH -CEPH) and we calculated Spearman's correlation between absolute latitude and one of the two allele frequencies of each SNP. We selected SNPs with a correlation coefficient within the upper 1% tail of the distribution. We then used a criterion of proximity between significant variants to focus on DNA regions showing a continuous signal over a portion of the genome. Based on external information and genome annotations, we demonstrated that most regions with the strongest signals also have biological relevance. We believe this proximity requirement adds an edge to our novel method compared to the existing literature, highlighting several genes (for example *DTNB, DOT1L, TPCN2, RELN, MSRA, NRG3*) related to body size or shape, human height, hair color, and schizophrenia. Our approach can be applied generally to any measure of association between polymorphic frequencies and continuously varying environmental variables.

Keywords: adaptations, spatial patterns, latitude, point processes, outlier approach

Evolutionary Bioinformatics 2013:9 29–42

doi: 10.4137/EBO.S10211

This article is available from http://www.la-press.com.

Introduction

From an evolutionary point of view, human biological variation can result from natural selection, genetic drift and demographic processes. In human population genetics, several ways have been found to highlight genes that may be subject to selective pressures, and in recent years whole genome scanning techniques have made it possible to find signatures of selection.[1–4] The Human Genome Diversity Project (HGDP-CEPH) database[5] has been repeatedly investigated in order to identify markers in the human genome which show geographical patterns and to explain how different selective forces can shape human genetic variations across continents. One strategy for the detection of spatial selection signatures is the outlier approach.[2,6,7] Using genome-wide data sets genotyped in different human populations, genetic variables—such as single nucleotide polymorphisms (SNPs)—that exhibit extreme correlations with latitude or with other environmental determinants are identified as candidate targets for selective pressure. By "extreme correlation" we mean that the value of a certain statistic, measuring the strength of the relationship between allele frequencies and latitude or other environmental variables, falls in the tails of the distribution of the same statistic over the whole genome. Many choices are possible for the relevant statistic, ranging from a simple (either Pearson or Spearman) correlation coefficient between the latitude and the frequencies of either one or two alleles of a SNP to a Bayes factor comparing two models that do and do not, respectively, take into account the effect of a dichotomous environmental variable on the distribution of a genetic variant. From a technical point of view, the outlier approach is just a reformulation of the concept of *statistical significance*, ie, variation with respect to a reference distribution.

The outlier approach has been used to study sodium homeostasis balance as an example of adaptation. In hot and dry climates, genes influencing salt and water retention are favored by selection, explaining in this way large inter-ethnic differences in the prevalence of salt-sensitive hypertension.[8,9] Other important research has been conducted to assess the correlation between four variables that summarize climate and the frequencies of 873 tag SNPs in 82 genes related to energy metabolic pathways.[6] The outlier approach has also been used to demonstrate that allele frequencies of a subset of genes coding for blood group antigens vary with levels of pathogen richness, supporting the idea that these loci affect susceptibility to infectious diseases.[10] This finding, which is compatible with previous evidences on the correlation between HLA class I diversity and pathogen richness,[11] is important for stressing the role of diseases and pathogens, like virus protozoa fungi, in shaping human variations.[12] Finally, a very comprehensive article on the HGDP-CEPH database (enriched with the Hap Map and other human populations databases) has recently been published, in which the outlier approach is used to highlight polymorphisms and pathways correlated with ecoregion membership and diet.[13]

Our idea is to reinforce the outlier approach by considering a criterion of *proximity* between significant variants. In the search for targets of selective pressure, we believe it is important to focus on those DNA regions which *repeatedly* contain values which are labeled as significant by the outlier approach. In other words, we look for evidence of a continuous signal over a portion of the genome which can strengthen the significance of a cluster of markers labeled as significant by the outlier approach alone and we built statistical tools.

In this paper we therefore adopt a search-and-confirm approach which integrates the outlier approach by identifying regions of the genome where not just one, but a significant number of SNPs are located in the tails of the distribution of the relevant statistic, when compared to the number of SNPs originally genotyped in the same region. This is done in the following three steps, which are further illustrated in the complete workflow process diagram in Figure 1:

1. The outlier method: We identify 1% significant SNPs as having an absolute value of the Spearman correlation coefficient with latitude above its 99th percentile;
2. The proximity-based algorithm: Using the methods described in detail in the Materials and Methods section, we select candidate regions in the genome which exhibit the strongest signals, ie, the regions where the significant SNPs identified above are present at a significantly higher rate when compared to the number of originally genotyped SNPs;
3. Biological relevance: We investigate the biological relevance of the strongest signals by comparing our

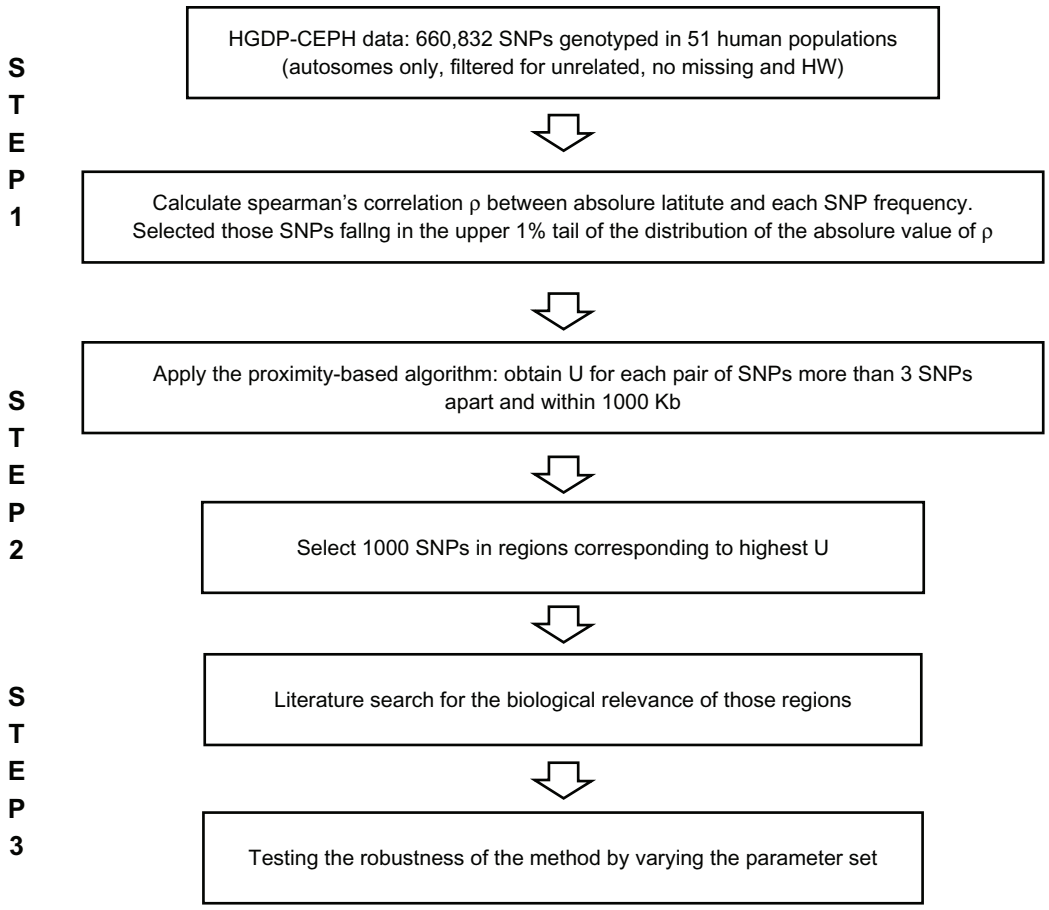

Figure 1. Graphical workflow process for the study.

data with results from Genome Wide Association studies (GWAs),[14] by studying the canonical pathway processes through gene-annotation enrichment analysis[15] and by comparing our analysis with previously published genomic scans for selective sweep.[3,16]

Materials and Methods

We describe here our methods with reference to the three-step process described in the Introduction.

Step 1: Our data and the outlier method

We used a data set of 660,832 SNPs genotyped in 51 human populations distributed worldwide from the HGDP-CEPH panel.[1] As underlined by a previous article,[17] within the HGDP-CEPH panel there are some closely-related individuals; in order to overcome this possible source of bias we excluded one member of each relative pair and we used 938 HGDP-CEPH individuals. Information about sample sizes and latitudes of the populations can be found on the CEPH homepage http://www.cephb.fr/en/hgdp/table.php.[5] Only 22 autosomes are included in our analysis; we also removed SNPs with more than 10% of missing genotypes and the ones that failed the Hardy-Weinberg *equilibrium* test in at least one population. After filtering, we use 545,209 SNPs.

Statistical analysis is performed using R.[18] We calculated Spearman's correlation (the correlation coefficient between the ranks of two variables) between absolute latitude and one of the two alleles of each SNP and, using the outlier approach, we identified those SNPs which have an absolute Spearman's correlation coefficient falling in the upper 1% tail of the distribution (Fig. 2).

Step 2: The proximity-based algorithm

For each chromosome, we now have two sequences of serial positions: one for all genotyped SNPs and one for the significant SNPs, the latter of which are included in the former. Each chromosome is indexed by the sequence of base pairs: as an approximation, we can view a chromosome as a linear segment and

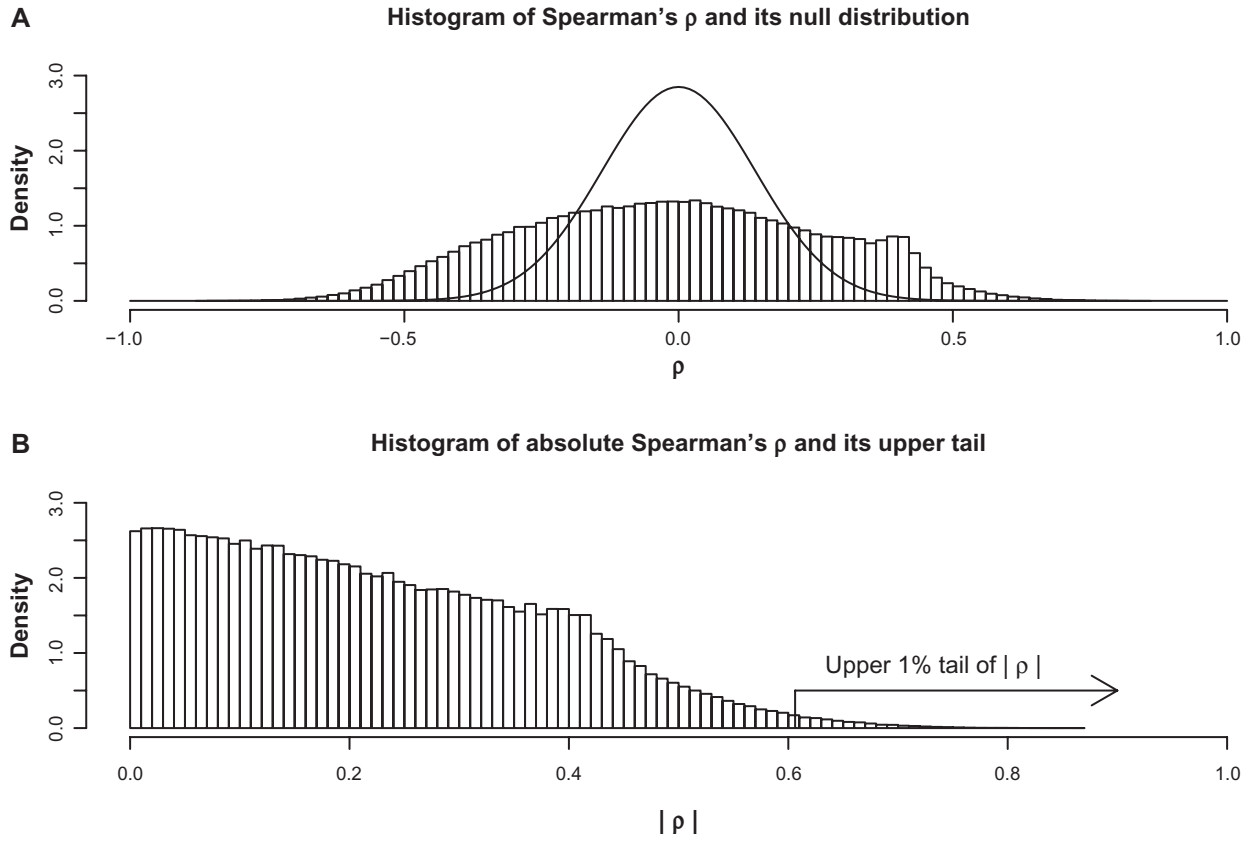

Figure 2. (**Panel A**) Histogram of the values of Spearman's correlation coefficient over all the SNPs and theoretical approximate density of the Spearman's correlation coefficient under the hypothesis of population null correlation. (**Panel B**) Histogram of the absolute values of Spearman's correlation coefficient over all the SNPs. Using the outlier approach, we identify significant SNPs in the 1% upper tail of this distribution.

the position of a SNP as a point of that linear segment. Based on the two sequences of points, we can define two cumulative counts depending on a generic point l, known in statistics as *counting processes*:

$S(l)$ = number of SNPs with a position smaller than or equal to l

$S_{.01}(l)$ = number of significant SNPs with a position smaller than or equal to l

with l varying from 1 (the first bp in the chromosome) to the position of the last bp of the chromosome. As an example, the two counting processes are plotted for chromosome 1 in Figure 3. Cumulative counts are a convenient way to compare the incidences of the different kinds of SNPs over different genomic regions (a simple dot plot would not do it, due to the sheer number of SNPs involved). If, over a certain segment of the chromosome, there is a greater-than-usual incidence of significant SNPs, then the relative increment of $S_{.01}(l)$ over that segment will be greater than the relative increment of $S(l)$ over the same segment. In other words, the graph of the $S_{.01}(l)$ counting process will be steeper than $S(l)$, up to a proportionality factor. Our proposal is to identify those genome regions which exhibit extreme concentrations of outlying SNPs.

We could formalize this search as a change-point problem for counting processes: in certain intervals to be estimated, the intensity of the $S(l)$ point process—a function modelling the instantaneous rate of incidence of the process—would be higher than in other regions. Due to the size of the problem and to the approximate nature of our search- and -confirm approach, we prefer a simpler *proximity-based algorithm* as follows.

For each pair of significant SNPs located at points l_1 and l_2, with $l_1 < l_2$ on the chromosome, we define

$$U(l_1, l_2) = \frac{S_{.01}(l_2) - S_{.01}(l_1)}{S(l_2) - S(l_1)}$$

ie, the observed incidence rate of significant SNPs per original SNP. This statistic over the sliding window

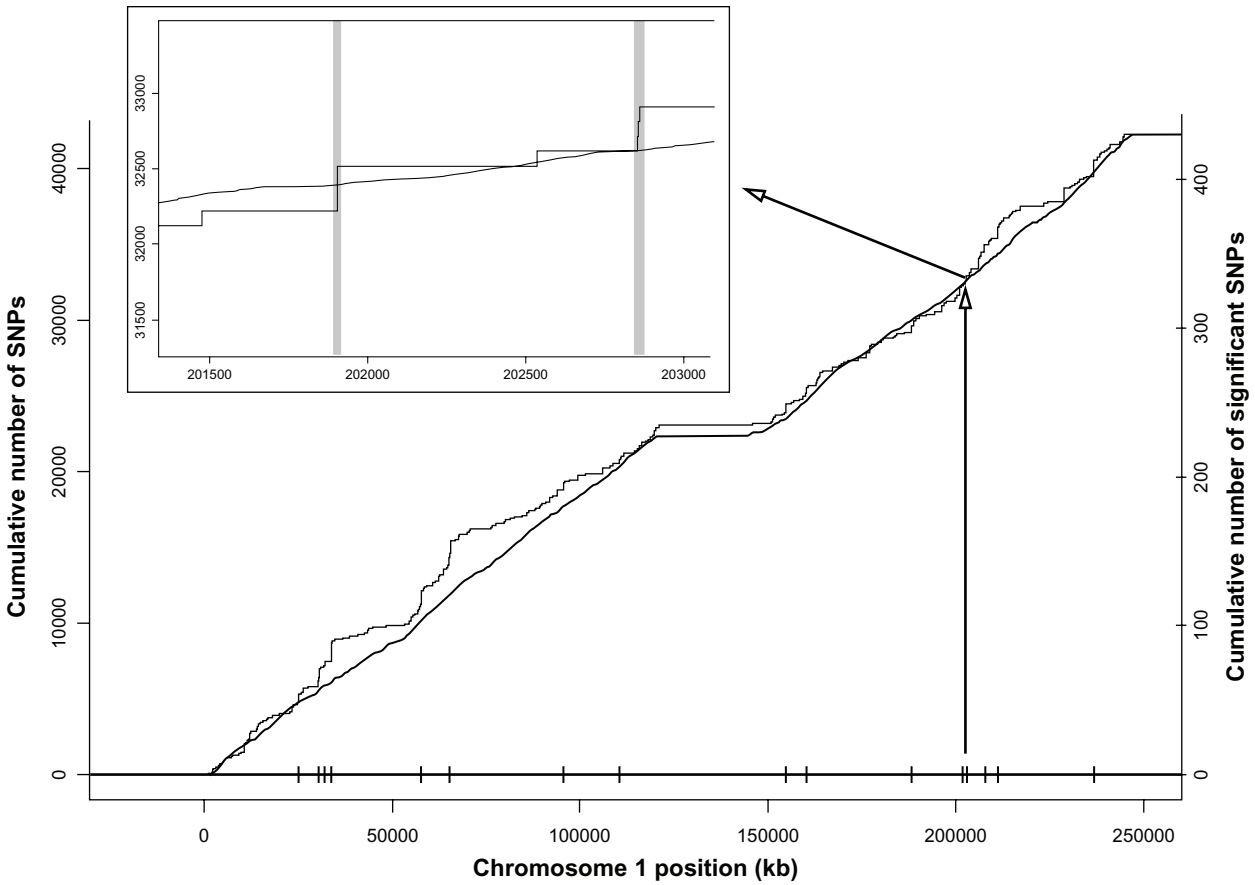

Figure 3. Counting process representation of the location of the candidate regions of chromosome 1.
Notes: The thicker step function represents cumulative counts of all originally genotyped SNPS and refers to the main ordinate scale, on the left. The thinner step function represents cumulative counts of significant SNPs and refers to the ordinate scale on the right. Sixteen regions identified by our method are shown as small vertical segments on the abscissa axis. The zooming box on the upper left part of the graph shows two of them (gray bands) located around position 202500 kb, as guided by the arrows.

(l_1, l_2) plays a central role in our proximity-based algorithm.

As a technical note, it would probably be a good idea to penalize large windows, for example by dividing the $U(l_1, l_2)$ statistic above by a penalty term $(l_2–l_1)^g$ with g equal to some number between 0 and 1. The final results would not change a lot (results not shown) and it would be difficult to commit to a specific g; therefore we decide to use the $U(l_1, l_2)$ statistic without a penalty term.

For each chromosome and for each significant SNP in position l_1, we computer $U(l_1, l_2)$ for each of the other significant SNPs in position l_2 within a distance of 1000 Kb from the original one. This is done to reduce the problem to a manageable size, under the assumption that relevant proximities are smaller than 1000 Kb.

We built the new reference distribution of all $U(l_1, l_2)$ values over all chromosomes, excluding from the analysis all $U(l_1, l_2)$ values relative to intervals (l_1, l_2) which included fewer SNPs than a threshold s, which has been chosen to be equal to 3 in this work. This is done to avoid very high automatic values of $U(l_1, l_2)$ when two significant SNPs happen to be adjacent. We selected the first 1000 SNPs contained in regions corresponding to the highest $U(l_1, l_2)$ values. A fixed number, rather than a fix tail area, was chosen to facilitate the discussion of the robustness of our method to varying parameters (see end of section Results).

Step 3: Biological relevance of the strongest signals

To accomplish step 3 as outlined in the Introduction, we proceeded to the biological cross-validation of our findings, which insofar had been based mainly on statistical grounds. We focused on the genes tagged by the SNPs we found, since our goal was to detect continuous signals coming from proximal groups of

SNPs belonging to the same gene. To link our findings to the results of genome wide data, we first compared our gene list with the June 2012 update of the Catalog of Publish ed GWAs.[14] Next, we scanned our gene list using a bioinformatic enrichment tool named Genecodis 2.0[15] to obtain a summary of the most enriched biological processes or pathways. Finally, we compared our analysis with previously published genomic scans for selective sweep in order to find possible overlaps in signals.

Results

We calculated Spearman's correlation between absolute latitude and one of the two alleles of the SNPs found in the HGDP-CEPH panel and, following the outlier approach, we identified those SNPs which have an absolute Spearman's correlation coefficient falling in the upper 1% tail of the distribution. The histogram of Spearman's correlations ρ's is plotted in Figure 2A. Its null distribution for 51 pairs of numbers has been overlaid on the same graph (Fig. 2A). It is a normal distribution with variance 1/50 due to a well-known result.[19] The discrepancy between the two distributions is due to SNPs which are correlated with latitude for reasons other than chance alone, for example due to environmental selection factors. Following the outlier approach, the upper 1% of the distribution of the absolute value of ρ, corresponding to $|\rho| > 0.606$, is identified in the histogram of the absolute value of ρ (Fig. 2B). It corresponds to 5452 outlying SNPs in the tails of the ρ distribution.

The candidate regions and the annotations emerging from the application of Step 2 described in the Introduction are contained in Additional 1 in the online supporting information. As an example, candidate regions which were identified in chromosome 1 are shown in Figure 3. The 1000 top SNPs emerging from the proximity-based algorithm enabled us to identify 467 intergenic and 533 genic SNPs, harboring 146 genes. We found 23 coding non synonymous (NS) changes and 6 coding synonymous changes. 372 were intronic and 107 were on the mRNA 3′UTR.

Finally, we gathered the biological knowledge of the strongest signals by comparing them to the Catalog of Published Genome-Wide Association Studies updated to June 2012. The genes which appear on this Catalog and additionally appear in candidate regions according to our proximity-based algorithm, are shown in Additional file 2. A short list of the most interesting signals are shown in Table 1. Several genes shown in that table are associated with metabolism-related phenotypes (like celiac disease for *IL21* interleukin 21, Gene id 59067)[20] and adiposity (*MSRA* Gene id4482) or variants associated with hair color in Europeans, like *TPCN2* gene (two pore segment channel 2, gene ID 219931)[21] and several with schizophrenia. At the same time, we compared our gene list with genes reported in OMIM. Several of our genes which show a correlation with latitude also implied some traits. For example, *DOT1L* gene (DOT1-like, histone H3 methyltransferase *Saccharomyces cerevisiae*) gene ID 84444 is associated with height[22] or *DTNB* gene dystrobrevin, beta ID 1838 which is affecting adult human height.[23] A complete table with the genes reported also in OMIM Disease database is in Additional file 3.

We analyzed Kyoto Encyclopedia of Genes and Genomes pathways (KEGG) using as reference set all genes in the Entrez-gene database and, as a statistical test, the hypergeometric one with a Benjamini-Hochberg correction for multiple testing at significance level equal to 0.05. Several KEGG pathways reached significance. The first was the extracellular matrix (ECM) receptor interaction (KEGG number: hsa04512) for the following genes: *RELN* reelin gene ID 5649; *ITGB6* integrin beta 6 gene ID 3694; *COL6A3* collagen, type VI, alpha 3 gene ID 1293. This pathway reaches a raw *P*-value of the hypergeometric test equal to 0.0011 and a *P*-value adjusted for multiplicity around 0.01. In order to look for overlaps with scans of the human genome for signals of positive natural selection, we compared our results with SNPs with significant composite of multiple signals (CMS) but only one intersection was found between the two gene lists concerning rs2256670 and rs2711853 both on *RELN* reelin, gene ID 5649.[16] A variety of choices were made in the actual implementation of the proximity-based algorithm described in Step 2 in the previous section. The two most important parameters set to reasonable values are (a) the maximum distance over which we search, which is set to 1000 Kb in Step 2, and (b) the minimum number of consecutive SNPs required, which is set to 3 in Step 2. In order to study the robustness of our method with respect to different values of these parameters, we varied the maximum distance and

Table 1. List of several genes reported in previously published GWAs and showing continuous correlation signals with our proximity based method.

Reported gene(s)	Trait	Region	NCBI ID	Gene description	Reference
C9orf3	Erectile dysfunction and prostate cancer treatment	9q22.32	84909	Chromosome 9 open reading frame 3	41
ABL1	Response to amphetamine	9q34.12	25	v-abl Abelson murine leukemia viral oncogene homolog 1	42
DTNB	Adult human height	2p23.3	1838	Dystrobrevin, beta	23
DTNB	Coronary heart disease	2p23.3	1838	Dystrobrevin, beta	43
TPCN2	Hair pigmentation in Europeans	11q13.3	219931	Two pore segment channel 2	21
DOT1L	Associated with height	19p13.3	84444	DOT1-like, histone H3 methyltransferase (*S. cerevisiae*)	22,35
RELN	Susceptibility and clinical phenotype in multiple sclerosis	7q22.1	5649	Reelin	44
RELN	Increases the risk of schizophrenia only in women	7q22.1	5649	Reelin	34
IL21	Celiac disease	4q27	59067	Interleukin 21	38,45
DOCK2	Protein quantitative trait loci	5q35.1	1794	Dedicator of cytokinesis 2	46
FRMD4B	Celiac disease	3p14.1	23150	FERM domain containing 4B	38
MAGI2	Hippocampal atrophy	7q21.11	9863	Membrane associated guanylate kinase, WW and PDZ domain containing 2	47
NCALD	Cognitive performance	8q22.3	83988	Neurocalcin delta	48
NRG3	Response to iloperidone treatment (QT prolongation)	10q23.1	10718	Neuregulin 3	49
RUNX3	Celiac disease	1p36.11	864	Runt-related transcription factor 3	38
SDK1	Quantitative traits	7p22.2	221935	Sidekick homolog 1 (chicken)	50
MSRA	Adiposity	8p23.1	4482	Methionine sulfoxide reductase A	27
MSRA	Hypertension	8p23.1	4482	Methionine sulfoxide reductase A	28
MSRA	Schizophrenia	8p23.1	4482	Methionine sulfoxide reductase A	51
MSRA	Bipolar disorder and schizophrenia	8p23.1	4482	Methionine sulfoxide reductase A	26

noticed (not shown) that the results where unchanged for distances down to 100 Kb.

The algorithm is instead sensitive to the minimum number of consecutive SNPs required: if we increase it from 3 to 5, for example (it would not make sense to consider a minimum much higher than 5), different SNPs and regions turn out to be significant, as shown in Table 2. For example, the number of selected SNPs shared when applying a minimum of 5 and when applying a minimum of 3 is 64%. This made us consider what would happen for varying this threshold. The changes are not dramatic (Table 2) but some interesting genes, like *AGT, ADCY9* and *WWOX* would come out from the analysis with a threshold equal to 5.

Discussion

In this paper we examined the HGDP-CEPH data again by integrating the outlier approach with a novel proximity-based algorithm.

Only latitude was used for ecological conditions, rather than using a multiplicity of variables as in Hancock et al[6] for example. We made this choice for the sake of simplicity, since latitude is correlated with different variables like short wave radiation flux, mean winter and summer temperatures, rainfall and pathogen richness. It should therefore provide a good proxy for the selective pressures that shaped variation in our genome. Even though we use a simple correlation measure such as Spearman's ρ with latitude only, we emphasize that the resulting signal should be a continuous and persistent proportion of background information, represented by all originally genotyped SNPs. We believe this proximity requirement adds an edge to our novel method when compared to existing literature. Our approach is applicable to any measure of association between polymorphic frequencies and environmental variables. It could be applied, for example, to complex statistics such as the minimum rank statistic, based on Bayes Factors and on rank transformations, of Hancock et al.[13]

Table 2. Percentage of common SNPs when varying the minimum number of consecutive SNPs required.

% concordance	3 SNPs	4 SNPs	5 SNPs
3 SNPs	100.00%	74.70%	64.00%
4 SNPs		100.00%	80.80%
5 SNPs			100.00%

With our method we identifed different genes, some of them already reported in the literature, dealing with different traits or diseases. GWAs include the scanning of all or most of the genes of different individuals aimed at finding susceptibility loci for traits or diseases. GWAs, so far, have allowed the identification of more than 7688 associated SNPs in humans. We compared our list of genes with GWAs results. Some interesting signals can be pointed out, for instance the correlation between skin pigmentation and latitude. It is well known that two coding variants in *TPCN2* are associated with hair color in Europeans.[21] At the same time *MSRA* (methionine sulfoxide reductase A gene) is related to the melanin formation in the hair follicle melanocyte.[24] Remarkably, *MSRA* gene is also related to schizophrenia[25,26] but also with adiposity[27] and hypertension.[28]

Several other genes in our list (see Additional file 1) can be associated with vitamin D related genes, known to show a latitude driven cline.[7] An example is *SMARCA2*, (SWI/SNF related, matrix associated, act in dependent regulator of chromatin, subfamily a, member 2), described as a component of a human multiprotein complex that that interacts directly with the vitamin D receptor. Schizophrenia genes are correlated with latitude and in our list several schizophrenia genes appear, like *GRID1*,[29,30] *MAGI2,*[31] *NRG3,*[32] *NRXN3,*[33] *RARB* and *RELN*.[34]

Region *CYP19A1* in our list is known from GWAs to exhib it association with adult height[35,36] whose distribution is related to latitude. Two more genes in our list, DOT1-like, histone H3 methyltransferase (*S. cerevisiae*)[22,35] and dystrobrevin, beta[23] are reported in OMIM to be related with height.

Several others genes are related to Celiac Disease (CD) which strongly correlates with latitude. Infectious agents are implicated in the pathogenesis of many autoimmune diseases like CD. This observation may imply that there is a relationship between one or more infectious agents, latitude related environmental exposure to gluten and others genetic susceptibility loci, and the development of this disease. For a complete review see Plot and Amital, 2009.[37] The *RUNX3* gene and *IL21*, in our list, are implicated with CD.[38] In the same paper, another gene *FRMD4B* previously known as *GRSP1*, appearing in our Table 1 is also associated with CD.[38] *RUNX3* gene is also required for CD8 T cell development during thymopoiesis.[39]

One of the most interesting genes highlighted by our work is *ANK2* (ankyrin 2, neuronal) which is implicated in cardiac arrhythmias due to abnormal variations in QT interval.[40]

Finally, the enrichment of genes in the KEGG pathway called extracellular matrix (ECM) receptor interaction (KEGG number: hsa04512) is note worth because these molecules are exploited by a number of pathogenic micro-organisms as receptors for cell entry. This can be interpreted as a signal of different forces played by pathogens on living cells in different environments.

Conclusions

Our study complements the growing body of knowledge surrounding scans for natural selection in humans using a method that uses the proximity criterion in addition to the outlier approach. Our findings support the hypothesis that latitudinal genetic diversity gradients are present in humans and reflect genetic adaptations to different environmental pressures that have shaped the human genome.

Acknowledgements

The authors thank the Human Genetic Foundation (HuGeF) Laboratory of Genomic Variation in Human Populations and Complex Disease group for biological and bioinformatics discussion.

Authors' Contributions

CDG designed the biological rationale for the study and MG provided the statistical tools to implement it, both authors wrote and revised the manuscript. MU and AP performed the data analysis. GM revised the manuscript. All authors read and approved the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. Feb 22, 2008; 319(5866):1100–4.
2. Coop G, Pickrell JK, Novembre J, et al. The role of geography in human adaptation. *PloS Genet.* 2009;5(6):e1000500.
3. Pickrell JK, Coop G, Novembre J, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research.* 2009; 19(5):826–37.
4. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 2010;20(4): R208–15.
5. Cann HM, de Toma C, Cazes L, et al. A human genome diversity cell line panel. *Science.* 2002;296(5566):261–2.
6. Hancock AM, Witonsky DB, Gordon AS, et al. Adaptations to climate in candidate genes for common metabolic disorders. *PloS Genet.* 2008; 4(2):e32.
7. Amato R, Pinelli M, Monticelli A, Miele G, Cocozza S. Schizophrenia and vitamin D related genes could have been subject to latitude-driven adaptation. *BMC Evol Biol.* 2010;10:351.
8. Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A. CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet*. Dec 2004;75(6):1059–69. Epub Oct 18, 2004.
9. Young JH, Chang YP, Kim JD, et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PloS Genet.* 2005;1(6):e82.
10. Young JH, Chang YP, Kim JD, et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PloS Genet.* 2005;1(6):e82.
11. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 2005;15(11):1022–7.
12. Pozzoli U, Fumagalli M, Cagliani R, et al. The role of protozoa-driven selection in shaping human genetic variability. *Trends Genet*. 2010;26(3):95–9.
13. Hancock AM, Witonsky DB, Ehler E, et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A*. 2010;107 Suppl 2:8924–30.
14. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106(23):9362–7.
15. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*. 2007;8(1):R3.
16. Grossman SR, Shylakhter I, Karlsson EK, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science.* 2010;327(5967):883–6.
17. Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet.* 2006;70(Pt 6):841–7.
18. Development Core Team R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

19. Hotelling HPR. Rank correlation and tests of significance involving no assumption of normality. *Ann Math Statist.* 1936;7(14).
20. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet.* 2008; 40(4):395–402.
21. Sulem P, Gudbjartsson DF, Stacey SN, et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet*. 2008;40(7): 835–7.
22. Lettre G, Jackson AU, Gieger C, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet.* 2008;40(5):584–91.
23. Gudbjartsson DF, Walters GB, Thorleifsson G, et al. Many sequence variants affecting diversity of adult human height. *Nat Genet*. 2008;40(5): 609–15.
24. Schallreuter KU, Salem MM, Hasse S, Rokos H. The redox—biochemistry of human hair pigmentation. *Pigment Cell Melanoma Res.* 2011;24(1): 51–62.
25. Walss-Bass C, Soto-Bernardini MC, Johnson-Pais T, et al. Methionine sulfoxide reductase: a novel schizophrenia candidate gene. *Am J Med Genet B Neuropsychiatr Genet*. 2009;150B(2):219–25.
26. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry.* 2012;17(9):880–6.
27. Lindgren CM, Heid IM, Randall JC, et al. Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PloS Genet.* 2009;5(6):e1000508.
28. Levy D, Larson MG, Benjamin EJ, et al. Framingham Heart Study 100 K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med Genet.* 2007;8 Suppl 1:S3.
29. Treutlein J, Muhleisen TW, Frank J, et al. Dissection of phenotype reveals possible association between schizophrenia and Glutamate Receptor Delta 1 (GRID1) gene promoter. *Schizophr Res.* 2009;111(1–3):123–30.
30. Chen X, Lee G, Maher BS, et al. GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia. *Mol Psychiatry.* 2011;16(11):1117–29.
31. Koide T, Banno M, Aleksic B, et al. Common variants in MAGI2 gene are associated with increased risk for cognitive impairment in schizophrenic patients. *PLoS One.* 2012;7(5):e36836.
32. Kao WT, Wang Y, Kleinman JE, et al. Common genetic variation in Neuregulin 3 (NRG3) influences risk for schizophrenia and impacts NRG3 expression in human brain. *Proc Natl Acad Sci U S A.* 2010;107(35): 15619–24.
33. Gauthier J, Siddiqui TJ, Huashan P, et al. Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. *Hum Genet.* 2011;130(4):563–73.
34. Shifman S, Johannesson M, Bronstein M, et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PloS Genet.* 2008;4(2):e28.
35. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature.* 2010;467(7317):832–8.
36. Okada Y, Kamatani Y, Takahashi A, et al. A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. *Hum Mol Genet.* 2010;19(11):2303–12.
37. Plot L, Amital H. Infectious associations of Celiac disease. *Autoimmun Rev.* 2009;8(4):316–9.
38. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet.* 2010;42(4): 295–302.
39. Woolf E, Xiao C, Fainaru O, et al. Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc Natl Acad Sci U S A.* 2003;100(13):7731–6.
40. Sedlacek K, Stark K, Cunha SR, et al. Common genetic variants in ANK2 modulate QT interval: results from the KORA study. *Circ Cardiovasc Genet.* 2008;1(2):93–9.
41. Kerns SL, Ostrer H, Stock R, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys.* 2010;78(5):1292–300.
42. Hart AB, Engelhardt BE, Wardle MC, et al. Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (CDH13). *PLoS One.* 2012;7(8):e42646.
43. Lettre G, Palmer CD, Young T, et al. Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *Plos Genet.* 2011;7(2):e1001300.
44. Baranzini SE, Wang J, Gibson RA, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet.* 2009;18(4):767–78.
45. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet.* 2008; 40(4):395–402.
46. Melzer D, Perry JR, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PloS Genet.* 2008; 4(5):e1000072.
47. Potkin SG, Guffanti G, Lakatos A, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One.* 2009;4(8):e6501.
48. Cirulli ET, Kasperaviciute D, Attix DK, et al. Common genetic variation and performance on standardized cognitive tests. *Eur J Hum Genet.* 2010;18(7):815–20.
49. Volpi S, Heaton C, Mack K, et al. Whole genome association study identifies polymorphisms associated with QT prolongation during iloperidone treatment of schizophrenia. *Mol Psychiatry.* 2009;14(11):1024–31.
50. Lowe JK, Maller JB, Pe'er I, et al. Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PloS Genet.* 2009;5(2):e1000365.
51. Ma X, Deng W, Liu X, et al. A genome-wide association study for quantitative traits in schizophrenia in China. *Genes Brain Behav.* 2011;10(7):734–9.

```
#########################################################################
#
# proximity.R – by MU and MG, November 2012
#               R programs to analyze the HGDP-CEPH data according to
#               the proximity-based method in Di Gaetano et al.
#
#########################################################################

# the following instructions assume that data have been read from the
# text files from the web page http://hagsc.org/hgdp/files.html
#  via,  for  example,  read.table("HGDP_Map.txt")  or  read.csv2("id2.csv",
sep = ",")
# into a dataframe called "data"

### function to compute allele's frequencies

freqfun<- function(data){
k<-0
freq <- namefreq <- NULL

for (i in dimnames(data)[[2]][-(1:3)]){

### do the frequency table
tav <- table(data[,3],data[,i])
### we exclude the tables with dim 1 or 2, in which
### there is no variation

### heterozygous, all homozygous and missing
if (dim(tav)[[2]]==4) {
# search of allele with greatest frequency
allmagg <- c(sum(tav[,2]),sum(tav[,4]))
if (allmagg[1] > allmagg[2]) magg <- 2 else magg <-4
freq <- cbind(freq, round((tav[,magg]+tav[,3]/2)/(tav[,2]+tav[,3]+tav[,4]),3))
# use the variable name for the table
namefreq <- c(namefreq,i)
}

### heterozygous, all homozygous and no missing
if (dim(tav)[[2]]==3 & (dimnames(tav)[[2]][1]!="--")) {
# search of allele with greatest frequency
allmagg <- c(sum(tav[,1]),sum(tav[,3]))
if (allmagg[1] > allmagg[2]) magg <- 1 else magg <-3

freq <- cbind(freq, round((tav[,magg]+tav[,2]/2)/(tav[,1]+tav[,2]+tav[,3]),3))
# use the variable name for the table
namefreq <- c(namefreq,i)
}
```

```
dimnames(freq)[[2]] <- namefreq
cat(k <- k+1,"\n")

}
freq
}

n_min = 5
finestra = 1000000

### Computation of U for chrom 1

ovr1_sig <- ovr1[ovr1[,"Lsig01"] == "sig01",]
matrice_rappovr1 <- NULL

# loop over all significant SNPs

for (i in 1:(dim(ovr1_sig)[1]-1) ) { #last SNP is automatically processed

# p = number of subsequent snps to that processed
p <- n_min-1
while ( (i+p)<= dim(ovr1_sig)[1] & (ovr1_sig[i+p,4]-ovr1_sig[i,4])<= finestra )
{
   iniz <- ovr1_sig[i,4]
   fin <- ovr1_sig[i+p,4]
   u <- (ovr1_sig[i+p,9]-ovr1_sig[i,9]+1)/(ovr1_sig[i+p,7]-ovr1_sig[i,7]+1)
   x <- c(i,p+1,1,iniz,fin,u)
   matrice_rappovr1 <- rbind(matrice_rappovr1,x)
   p <- p+1
}

}
dimnames(matrice_rappovr1)[[2]] <- c("SNP","n SNP", "chrom", "reg in", "reg
fin","U")

### Selection of SNP
N = 1000

# union of results of all chromosomes
matrice_totale = rbind(matrice_rappovr1,matrice_rappovr2,matrice_rappovr3,
               matrice_rappovr4,matrice_rappovr5,matrice_rappovr6,
matrice_rappovr7,

     matrice_rappovr8,matrice_rappovr9,matrice_rappovr10,matrice_rappovr11,

     matrice_rappovr12,matrice_rappovr13,matrice_rappovr14,matrice_rappovr15,
```

```
     matrice_rappovr16,matrice_rappovr17,matrice_rappovr18,matrice_rappovr19,
        matrice_rappovr20,matrice_rappovr21,matrice_rappovr22)

matrice_totale <- data.frame(matrice_totale)
# decreasing order
matrice_totale <- matrice_totale[order(matrice_totale$U,decreasing = TRUE),]

# loop to extract result
ovr_sig <- list(ovr1_sig,ovr2_sig,ovr3_sig,ovr4_sig,ovr5_sig,ovr6_sig,ovr7_sig,

     ovr8_sig,ovr9_sig,ovr10_sig,ovr11_sig,ovr12_sig,ovr13_sig,ovr14_sig,
             ovr15_sig,ovr16_sig,ovr17_sig,ovr18_sig,ovr19_sig,ovr20_sig,
             ovr21_sig,ovr22_sig)
z = 1
selezione <- NULL
n_SNP = 0

while (n_SNP < N )
       {
       a = matrice_totale[z,1]
       b = matrice_totale[z,2]
       c = matrice_totale[z,3]

       sel <- cbind( ovr_sig[[c]][a:(a+b-1),1], rep(c,b) )
       selezione <- rbind(selezione, sel)

       z <- z+1

       # test to obtain unique solutions
       selezione <- unique(selezione)

       n_SNP <- dim(selezione)[1]
}

dimnames(selezione)[[2]] <- c("name_SNP" , "chrom" )
selezione <- selezione[1:N,]
```

Additional Files

Additional file 1: SNPs and regions from the proximity-based algorithm.

Additional file 1 in the online supporting information contains all regions selected by the proximity-based method, duly annotated.

Additional file 2: The complete list of genes reported in previously published GWAs and showing continuous correlation signals with our proximity based method.

Additional file 3: The complete list of genes reported in OMIM and showing continuous correlation signals with our proximity based method.

Additional file 4: R scripts.

Additional file 4 in the online supporting information contains R scripts to perform the necessary calculations.

Genes reported in previous GWAS and showing continous correlation signals with our proximity based method

MSRA	4482	ENSG00000175806	MSRA	methionine sulfoxide reductase A
C14orf143	90141	ENSG00000140025	C14orf143	chromosome 14 open reading frame 143
CPNE8	144402	ENSG00000139117	CPNE8	copine VIII
DAB1	1600	ENSG00000173406	DAB1	disabled homolog 1 (Drosophila)
ATF7IP	55729	ENSG00000171681	ATF7IP	activating transcription factor 7 interacting protein
NCALD	83988	ENSG00000104490	NCALD	neurocalcin delta
RUNX3	864	ENSG00000020633	RUNX3	runt-related transcription factor 3
DOCK4	9732	ENSG00000128512	DOCK4	dedicator of cytokinesis 4
SCN3A	6328	ENSG00000153253	SCN3A	sodium channel, voltage-gated, type III, alpha subunit
RELN	5649	ENSG00000189056	RELN	reelin
NRXN3	9369	ENSG00000021645	NRXN3	neurexin 3
PPA2	27068	ENSG00000138777	PPA2	pyrophosphatase (inorganic) 2
NRG3	10718	ENSG00000185737	NRG3	neuregulin 3
ITPR2	3709	ENSG00000123104	ITPR2	inositol 1,4,5-triphosphate receptor, type 2
GRID1	2894	ENSG00000182771	GRID1	glutamate receptor, ionotropic, delta 1
DOCK2	1794	ENSG00000134516	DOCK2	dedicator of cytokinesis 2
NOX4	50507	ENSG00000086991	NOX4	NADPH oxidase 4
ANK2	287	ENSG00000145362	ANK2	ankyrin 2, neuronal
RGNEF	64283	ENSG00000214944	RGNEF	Rho-guanine nucleotide exchange factor
NRXN1	9378	ENSG00000179915	NRXN1	neurexin 1
COL6A3	1293	ENSG00000163359	COL6A3	collagen, type VI, alpha 3
C21orf33	8209	ENSG00000160221	C21orf33	chromosome 21 open reading frame 33
IL21	59067	ENSG00000138684	IL21	interleukin 21
TPCN2	219931	ENSG00000162341	TPCN2	two pore segment channel 2
RHPN2	85415	ENSG00000131941	RHPN2	rhophilin, Rho GTPase binding protein 2
CYP19A1	1588	ENSG00000137869	CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1
ATP10B	23120	ENSG00000118322	ATP10B	ATPase, class V, type 10B
FRMD4B	23150	ENSG00000114541	FRMD4B	FERM domain containing 4B
SMARCA2	6595	ENSG00000080503	SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
CDH10	1008	ENSG00000040731	CDH10	cadherin 10, type 2 (T2-cadherin)
ABL1	25	ENSG00000097007	ABL1	c-abl oncogene 1, receptor tyrosine kinase
RARB	5915	ENSG00000077092	RARB	retinoic acid receptor, beta
FLJ43663	378805	NA	FLJ43663	hypothetical LOC378805
POLN	353497	ENSG00000130997	POLN	polymerase (DNA directed) nu
PLA2R1	22925	ENSG00000153246	PLA2R1	phospholipase A2 receptor 1, 180kDa
MAGI2	9863	ENSG00000187391	MAGI2	membrane associated guanylate kinase, WW and PDZ domain containing 2
DSCAML1	57453	ENSG00000177103	DSCAML1	Down syndrome cell adhesion molecule like 1
TAF1B	9014	ENSG00000115750	TAF1B	TATA box binding protein (TBP)-associated factor, RNA polymerase I, B, 63kDa

C9orf3	84909	ENSG00000148120	C9orf3	chromosome 9 open reading frame 3
DOT1L	84444	ENSG00000104885	DOT1L	DOT1-like, histone H3 methyltransferase (S. cerevisiae)
FLJ45139	400867	NA	FLJ45139	FLJ45139 protein
LRIG3	121227	ENSG00000139263	LRIG3	leucine-rich repeats and immunoglobulin-like domains 3
CNTN6	27255	ENSG00000134115	CNTN6	contactin 6
DTNB	1838	ENSG00000138101	DTNB	dystrobrevin, beta
KCNH7	90134	ENSG00000184611	KCNH7	potassium voltage-gated channel, subfamily H (eag-related), member 7
MCC	4163	ENSG00000171444	MCC	mutated in colorectal cancers
KIRREL3	84623	ENSG00000149571	KIRREL3	kin of IRRE like 3 (Drosophila)
ATP2B4	493	ENSG00000058668	ATP2B4	ATPase, Ca++ transporting, plasma membrane 4
SNTB1	6641	ENSG00000172164	SNTB1	syntrophin, beta 1 (dystrophin-associated protein A1, 59kDa, basic component 1)
SDK1	221935	ENSG00000146555	SDK1	sidekick homolog 1, cell adhesion molecule (chicken)
CSMD2	114784	ENSG00000121904	CSMD2	CUB and Sushi multiple domains 2
PPIAP14	5486	NA	PPIAP14	peptidylprolyl isomerase A (cyclophilin A) pseudogene 14
TSC22D2	9819	ENSG00000196428	TSC22D2	TSC22 domain family, member 2