Formal derivations for the paper (please cite):

Kliger, L., & Yovel, G. (2019). The functional organization of high-level visual cortex determines the representation of complex visual stimuli. (*bioRxiv*). https://doi.org/10.1101/2019.12.22.852293

According to the normalization model (Reynolds & Heeger, 2009), the measured neuronal response of a specific neuron (i.e., neuron j) to a pair of stimuli, for example, a face and a body (denoted by F and B respectively) is given by the following equation:

$$R_j(F+B) = \gamma \frac{F_j + B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k},$$

where the measured response of a specific neuron (i.e. neuron j) to a pair of stimuli presented together, $R_j(F+B)$, equals to the response of the neuron to the sum of the two stimuli, $F + B_j$, divided by the sum of the responses of the surrounding neurons (the normalization pool) to the two stimuli, $\Sigma_k F_k + \Sigma_k B_k$, and a constant, σ. The constants γ and σ are free parameters that are fitted to the data (Carandini & Heeger, 2012).

A face-selective neuron that resides in a face-selective area responds more to faces than bodies, i.e.,

$$F_j \gg B_j,$$

and is surrounded by neurons that are mostly face-selective, i.e.,

$$\sum_k F_k \gg \sum_k B_k.$$

Based on the normalization equation we can predict that the response to the face and body presented together will be dominated by the response to the face:

$$R_j(F+B) \approx \gamma \frac{F_j}{\sigma + \sum_k F_k} \approx R_j(F)$$

Similarly, in a body-selective area, the response to a face and a body of a body-selective neuron, i.e.,

$$F_j \ll B_j,$$

surrounded mostly by body-selective neurons, i.e.,

$$\Sigma_k F_k \ll \Sigma_k B_k,$$

will be dominated by the response to the body:

$$R_j(F+B) \approx \gamma \frac{B_j}{\sigma + \sum_k B_k} \approx R_j(B)$$

In addition, face- and body-selective areas usually reside in neighboring locations and the border between them contains two populations of neighboring neurons that are selective to either a face or a body. In an area that has a similar proportion of neurons that are selective to faces and bodies,

$$\Sigma_k F_k \approx \Sigma_k B_k,$$

the face and the body will contribute equally to the representation of the face and body presented together. Thus, the response to the face and body when presented together will be the mean of the responses to each of them when presented alone:

$$R_j(F+B) = \gamma \frac{F_j + B_j}{\sigma + \sum_k F_k + \sum_k B_k}$$

$$\approx \frac{\gamma F_j}{\sigma + 2\sum_k F_k} + \frac{\gamma B_j}{\sigma + 2\sum_k B_k}.$$

$$\approx \frac{1}{2} R_j(F_j) + \frac{1}{2} R_j(B_j).$$

More generally, we can reform the normalization equation to express the response to a pair of stimuli as a linear combination of the responses to each of the isolated stimulus composing the pair.

Since

$$R_j(F+B) = \gamma \frac{F_j + B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k},$$

we can separate the right side of the equation to two parts, yielding

$$R_j(F+B) = \frac{\gamma F_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k} + \frac{\gamma B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \ .$$

Next, we multiply each part by a term equal to 1:

$$R_j(F+B) = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k} \cdot \frac{\gamma F_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k} + \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k B} \cdot \frac{\gamma B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \ .$$

Rewriting the equation, it becomes:

$$R_j(F+B) = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \cdot \frac{\gamma F_j}{\sigma + \Sigma_k F_k} + \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B} \cdot \frac{\gamma B_j}{\sigma + \Sigma_k B_k} \ .$$

Since the response to the isolated face and body according to the normalization model is given by:

$$R_j(F) = \frac{\gamma F_j}{\sigma + \Sigma_k F_k}$$

and

$$R_j(B) = \frac{\gamma B_j}{\sigma + \Sigma_k B_k} \ ,$$

the response to the face and body presented together becomes:

$$R_j(F+B) = \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \cdot R_j(F) + \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B} \cdot R_j(B).$$

We can write this equation as

$$R_j(F+B) = \beta_F \cdot R_j(F) + \beta_B \cdot R_j(B),$$

where

$$\beta_F = \frac{\sigma + \sum_k F_k}{\sigma + \sum_k F_k + \sum_k B_k},$$

$$\beta_B = \frac{\sigma + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

are the coefficients of the isolated face and the body in this linear combination. Note that the coefficients depend only on the selectivity of the surrounding neurons (i.e., the normalization pool) to the stimuli presented, and not on the selectivity of neuron j itself.

When using fMRI to measure the response to multi-category stimulus, we measure the BOLD signal, which is an approximae measure of the response of thousands of neurons in a small patch of cortex (e.g., a 2x2x2 mm^3 voxel). The response of all of the neurons in a voxel can be written as:

$$\Sigma_j R_j(F+B) = \Sigma_j\left(\gamma \frac{F_j + B_j}{\sigma + \Sigma_{k(j)} F_{k(j)} + \Sigma_{k(j)} B_{k(j)}}\right),$$

where k(j) indicates the k'th neuron in the normalization pool of neuron j.

Assuming that all neurons in a given voxel has a similar normalization pool, i.e., a similar surrounding, we can rewrite the equation such that k is no longer a function of j:

$$\Sigma_j R_j(F+B) = \Sigma_j\left(\gamma \frac{F_j + B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k}\right)$$

$$= \Sigma_j\left(\gamma \frac{F_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k}\right) + \Sigma_j\left(\gamma \frac{B_j}{\sigma + \Sigma_k F_k + \Sigma_k B_k}\right).$$

Now, following the exact same derivation as before, we can rewrite the equation as:

$$\Sigma_j R_j(F+B) = \Sigma_j\left(\frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \cdot \frac{\gamma F_j}{\sigma + \Sigma_k F_k}\right) + \Sigma_j\left(\frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B} \cdot \frac{\gamma B_j}{\sigma + \Sigma_k B_k}\right)$$

$$= \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \cdot \Sigma_j\left(\frac{\gamma F_j}{\sigma + \Sigma_k F_k}\right) + \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B} \cdot \Sigma_j\left(\frac{\gamma B_j}{\sigma + \Sigma_k B_k}\right)$$

$$= \frac{\sigma + \Sigma_k F_k}{\sigma + \Sigma_k F_k + \Sigma_k B_k} \cdot \Sigma_j R_j(F) + \frac{\sigma + \Sigma_k B_k}{\sigma + \Sigma_k F_k + \Sigma_k B} \cdot \Sigma_j R_j(B)$$

Similar to the case of a single neuron, we can write this equation as

$$\Sigma_j R_j(F+B) = \beta_F \cdot \Sigma_j R_j(F) + \beta_B \cdot \Sigma_j R_j(B),$$

where

$$\beta_F = \frac{\sigma + \sum_k F_k}{\sigma + \sum_k F_k + \sum_k B_k},$$

$$\beta_B = \frac{\sigma + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

are the coefficients of the isolated face and the body in this linear combination and are the same as the coefficients for the single neuron response. These coefficients are dependent on the local selectivity to each of the isolated categories, which can be measured effectively by fMRI.

We can further see that the difference between the coefficients is given by:

$$\beta_F - \beta_B = \frac{\sigma + \sum_k F_k}{\sigma + \sum_k F_k + \sum_k B_k} - \frac{\sigma + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

$$= \frac{\sum_k F_k - \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

i.e., the difference between the coefficients is determined by the difference in the selectivity of the normalization pool to the two categories of the multi-category stimulus.

The sum of the coefficients is given by:

$$\beta_F + \beta_B = \frac{\sigma + \sum_k F_k}{\sigma + \sum_k F_k + \sum_k B_k} + \frac{\sigma + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

$$= \frac{2\sigma + \sum_k F_k + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k}$$

$$= \frac{\sigma + \sum_k F_k + \sum_k B_k}{\sigma + \sum_k F_k + \sum_k B_k} + \frac{\sigma}{\sigma + \sum_k F_k + \sum_k B_k}$$

$$= 1 + \frac{\sigma}{\sigma + \sum_k F_k + \sum_k B_k}$$

i.e., the sum of the coefficients is slightly higher than 1, being equal to 1 plus a small positive term (σ is usually a small positive number, see Reynolds & Heeger, 2009).

Taking it all together, the multi-category response can be described as a weighted mean of the responses to the isolated parts (with sum of weights slightly higher than 1). Moreover, the weights are determined by the proportion of neurons in the surrounding selective to either of the categories.