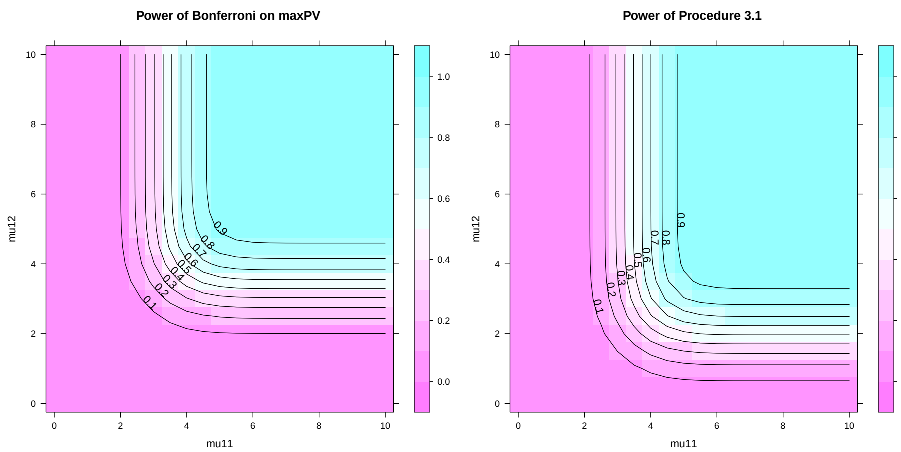

Figure 5: The power as function of the expectation in the first study (x-axis) and the expectation in the second study (y-axis), for the false no replicability null hypothesis, in a setting where one no replicability null hypothesis is false out of 100 no replicability null hypotheses. Left panel: Procedure that applies a Bonferroni correction on the maximum two study p-values for FWER control at level 0.05. Right panel: Procedure 3.1 with $(\alpha_1, \alpha) = (0.025, 0.05)$ and Bonferroni as the FWER controlling procedure.

dure. Moreover, for fixed $\mu_1 > \mu_2$, *the power of the two stage procedure is larger if* $(\mu_{11}, \mu_{21}) = (\mu_1, \mu_2)$ *than if* $(\mu_{11}, \mu_{21}) = (\mu_2, \mu_1)$.

Figure 6 shows the difference in power of Procedure 3.1 using Bonferroni with $c = \alpha_1/\alpha \in \{0.2, 0.5, 0.8\}$, *as well as the Bonferroni procedure on maximum p-values, from the power of Procedure 3.1 with optimal choice of c. Clearly, Procedure 3.1 with optimal choice of c can be much more powerful than the Bonferroni procedure on maximum p-values. Moreover, for the three choices* $c = 0.2$, $c = 0.5$ *and* $c = 0.8$, *the difference in power from the optimal power is fairly small, especially when the optimal power is above 0.9 (right panel). Figure 7 shows the power as a function of c for three configurations of* (μ_{11}, μ_{21}), *for which the power using the optimal c is 0.9. The power function is quite flat. The optimal c is below 0.5 in the top left panel, and above 0.5 in the top right and bottom panel. However, the difference in power between Procedure 3.1 with* $c = 0.5$ *and Procedure 3.1 with optimal c is small.*