Biometrics 000, 000–000 DOI: 000
000 0000

Web Appendix: *Supplemental Simulation Studies for* k-FWER control without p-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins

L. Finos*

University of Padova and Leiden University Medical Center

**email:* livio@stat.unipd.it

and

A. Farcomeni*

Sapienza - University of Rome

**email:* alessio.farcomeni@uniroma1.it

1. Simulation setting

In this Web Appendix we report the results of simulation studies under different settings, to complement the simulation study of Section 4 in the main paper.

We perform one-sample t-tests, with data generated from standard normals under the null hypothesis. We let $n = 5, 10, 20, 50$; $m = 1000, 10000, 100000$ ($m = 100, 1000, 10000$ in Section 4 and Section 5) and $\alpha = 0.05$. Unless stated otherwise, we fix the mean under the alternative hypotheses so that the single tests have a prescribed power of 70% (hence, the mean under the alternative decreases as the sample size n increases). The position of the false hypotheses are selected randomly at each iteration, and unless stated otherwise the number of false nulls is set so that 10% of the m hypotheses are false.

For each setting we generate the data, compute p-values, and apply the Lehmann and Romano (2005) (LR) and Guo and Romano (2007) (GR) step-down procedures; together with our procedure with data-driven order of the hypotheses (ORD) and its extension for general dependence (ORDdep). We perform 1000 Monte Carlo iterations for each setting. We report the fraction of true rejections to estimate power, and the fraction of iterations in which k or more false rejections occurred, as an estimate of the level of the multiple test. When provided, the latter is used to check that the level of the multiple test is below the prescribed level α.

2. Fraction of hypotheses under the alternative: 5%

Figure 1 reports the estimated power under the setting described in Section 1, when a proportion of 5% of the m hypotheses are false.

[Figure 1 about here.]

3. Distributed alternatives

In this section we generate data as described in Section 1. The difference with the setting of the main paper is that we use distributed alternatives. The effect size of each test corresponding to a false null hypothesis is randomly generated from $\chi_1^2\delta(n)_{70\%}$, where $\delta(n)_{70\%}$ is equal to 2.024, 1.325, 0.909 and 0.566 for $n = 5, 10, 20, 50$ respectively and χ_1^2 is a random draw from a chi-squared distribution with 1 degree of freedom. Figure 2 reports the estimated power in this setting.

[Figure 2 about here.]

4. k-FWER control under block dependence

In this Section we generate data as described in Section 1, but test statistics are now dependent. We use a block dependence assumption in order to mimick a genomic situation (it is frequently assumed that genes are dependent in blocks of size 20-30). We then generate data in blocks of size 25, in which observations within each block arise from a multivariate normal with a covariance matrix with diagonal elements equal to 1 and off-diagonal elements equal to $\rho = 0.3$. Figure 3 and Figure 4 report the estimated power and the estimated level of the multiple test, respectively.

[Figure 3 about here.]

[Figure 4 about here.]

5. k-FWER control under damped cosine dependence

The simulation setting is as described in Section 1, but now we let all test statistics be dependent, and generate data from a multivariate normal in $\mathcal{R}^m$.

In order to impose a meaningful dependence structure, we pretend data is scattered on a squared spatial grid. Random variables are then linked to an ideal map, equally scattered,

and indicized by a pair of coordinates (i, j), i indicating the "row" position and j indicating the "column" position. In each point (i, j) a sample of n normals is observed. A similar setting is used in Farcomeni (2006).

We determine the covariance matrix through a simplified version of a kernel commonly used in spatial statistics:

$$\mathrm{Cov}(X_{ij}, X_{i'j'}) = \cos(\frac{1}{\tau}\ d((i,j),(i',j')))e^{-\frac{1}{\tau}\ d((i,j),(i',j'))}, \tag{1}$$

where $d(\cdot,\cdot)$ is the euclidean distance function, (i, j) are the coordinates of the h-th test and (i', j') the coordinates of the z-th test. This function is used to generate both positive and negative correlations. The strenght of dependence is controlled by the tuning parameter τ, which is set as $\tau = 7.5$ in this section, leading to correlations as high as 0.9.

Figure 5 and Figure 6 report the estimated power and the estimated level of each multiple test, respectively.

[Figure 5 about here.]

[Figure 6 about here.]

6. Effect size $\delta = 1$

Figure 7 reports the estimated power under the setting described in Section 1, where we now fix the effect size $\delta(n) = 1$. Obviously, single inference power then increases with the sample size: the power of each individual test for $n = 5, 10, 20, 50$ is equal to 0.286, 0.562, 0.869 and 0.999 respectively.

[Figure 7 about here.]

7. Effect size $\delta = 1.5$

Figure 8 reports the estimated power under the setting described in Section 1, where we fix an effect size $\delta(n) = 1.5$. The power of each individual test, as a function of $n = 5, 10, 20, 50$ is 0.549, 0.887, 0.996 and 1.000 respectively.

[Figure 8 about here.]

8. Conclusions

In this Web Appendix we have showed results for simulations under different scenarios. The proposed simulations arise essentially the same comments as reported in the main paper.

We shall furthermore note that under dependence the Type I error rate is sometimes exceeded by GR and ORD, but for larger values of m all procedures yield a nominal error rate below α under block dependence. The damped cosine setting is designed for simulating a situation of strong dependence, and GR and ORD more often exceed the nominal error rate. This is a consequence of the high value chosen for the tuning parameter τ. Smaller values lead to procedures with a k-FWER below the nominal error level.

In terms of power, ORD is seen to dominate all other procedures for small sample size; while ORDdep seems to be often slightly better than the LR procedure.

Finally, we have simulated two settings in which the effect size is constant with respect to n. Power of the individual tests is then increasing with the sample size. In that case, our ORD procedure is not outperformed anymore for large n. With low sample size ORD outperforms the others, with large sample size sometimes ORD dominates the others, and sometimes GR dominates.

References

Farcomeni, A. (2006). More powerful control of the false discovery rate under dependence. *Statistical Methods & Applications* **15,** 43–73.

Guo, W. and Romano, J. (2007). A generalized Sidak-Holm procedure and control of genralized error rates under independence. *Statistical Applications in Genetics and Molecular Biology* **6,**.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33,** 1138–1154.

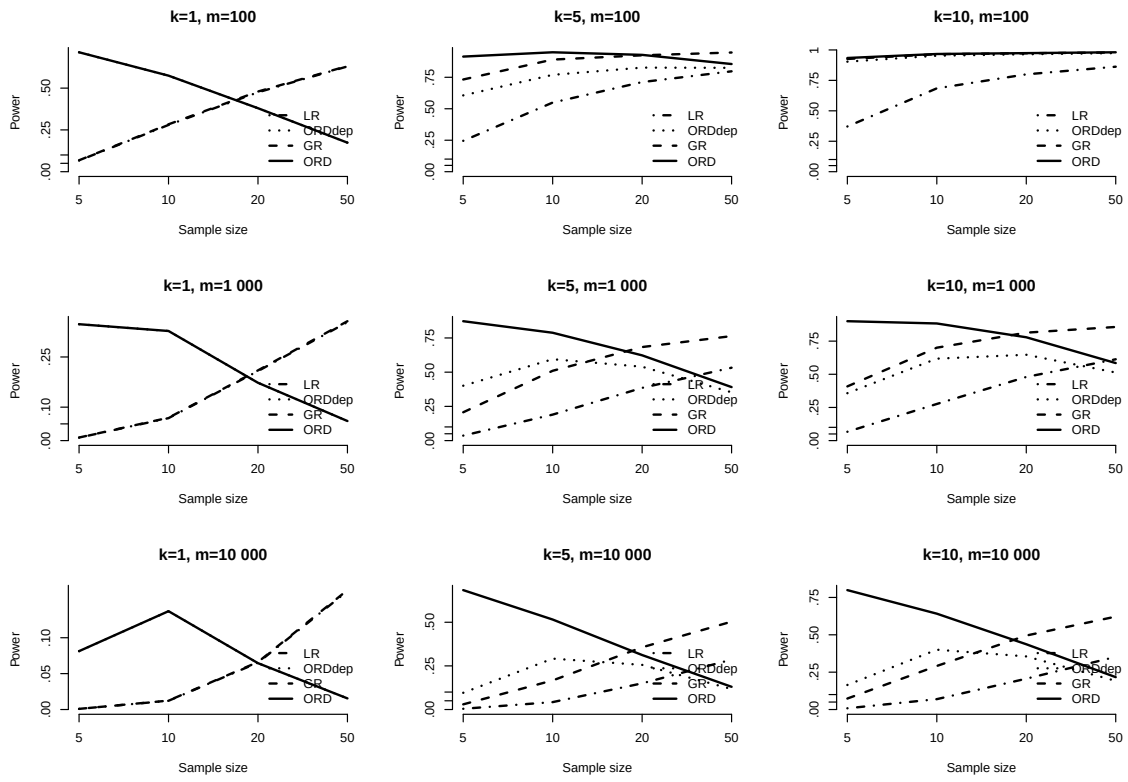

Figure 1. Proportion of correctly rejected hypotheses for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 5%, the power of each single test at 70% (i.e. effect size equal to 2.024, 1.325, 0.909 and 0.566 for $n = 5, 10, 20, 50$ respectively).

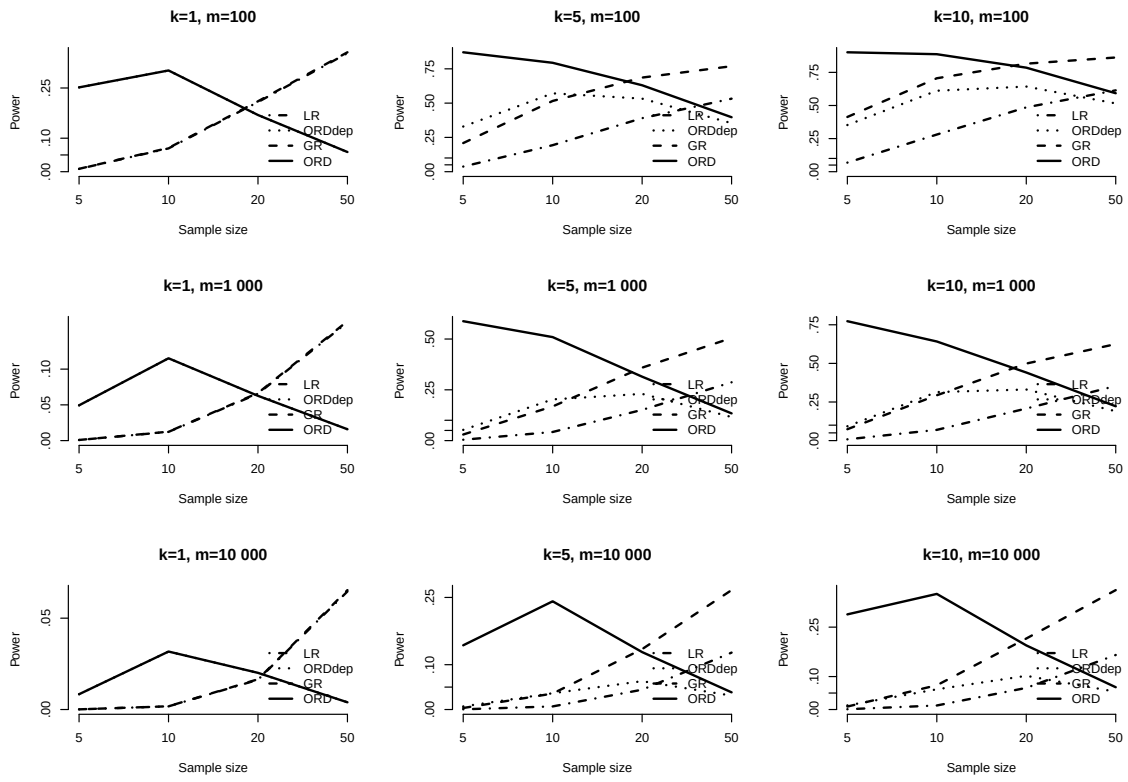

Figure 2. Proportion of correctly rejected hypotheses for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 10%, the effect size is randomly generated from a r.v. $\chi^2_1\delta(n)_{70\%}$.

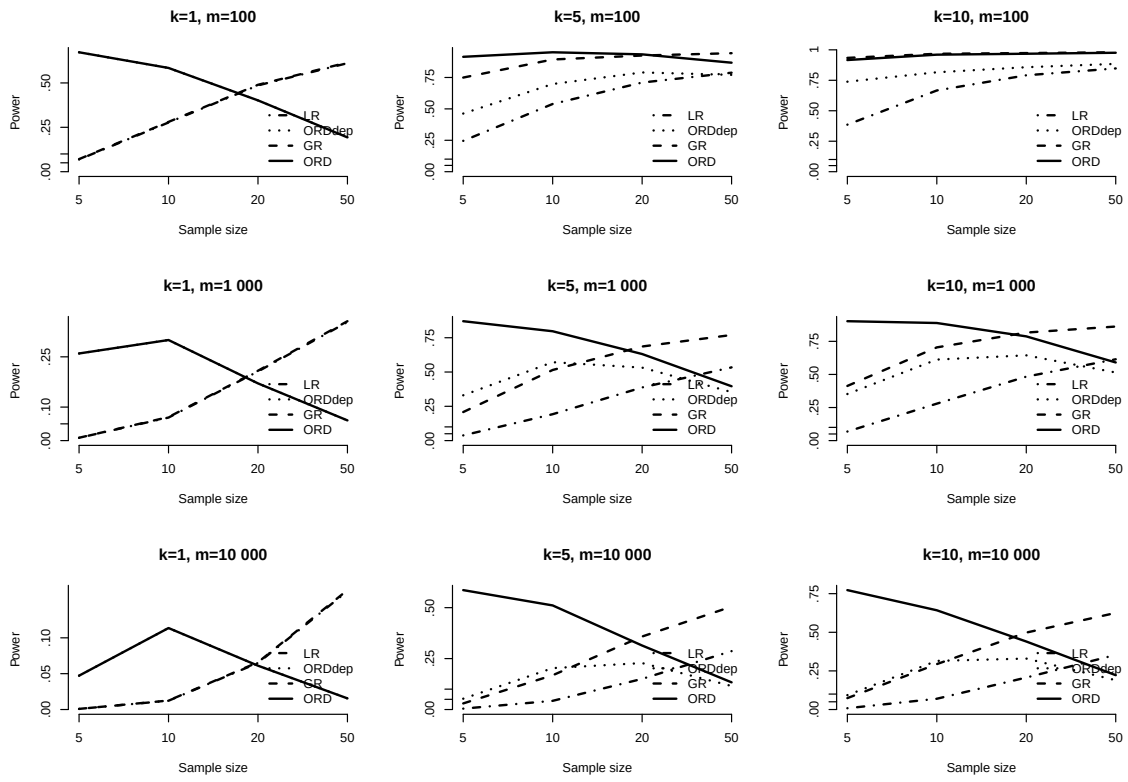

Figure 3. Proportion of correctly rejected hypotheses for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 10%, the power of each single test at 70%. Variables have block correlation structure (block size 25 and $\rho = .30$)

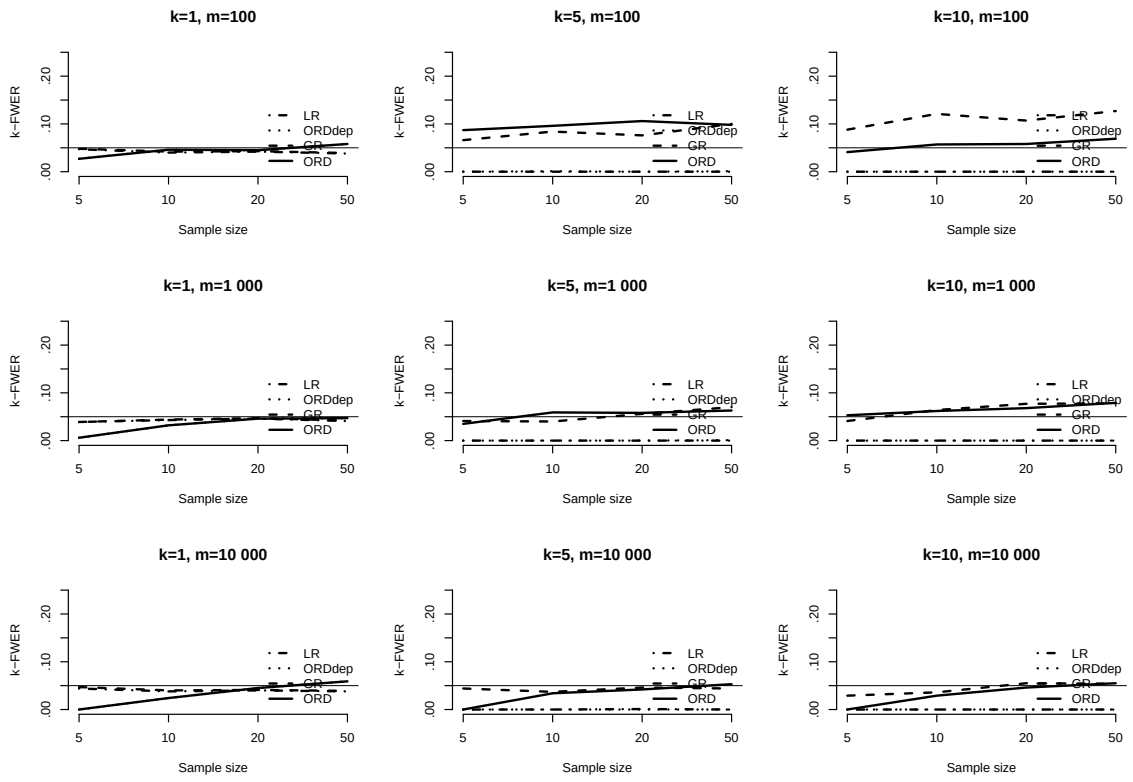

Figure 4. Estimated k-FWER for different values of n and m. α is set to .05, the proportion of false nulls is 10%, the power of each single test at 70%. Variables have block correlation structure (block size 25 and $\rho = .30$)

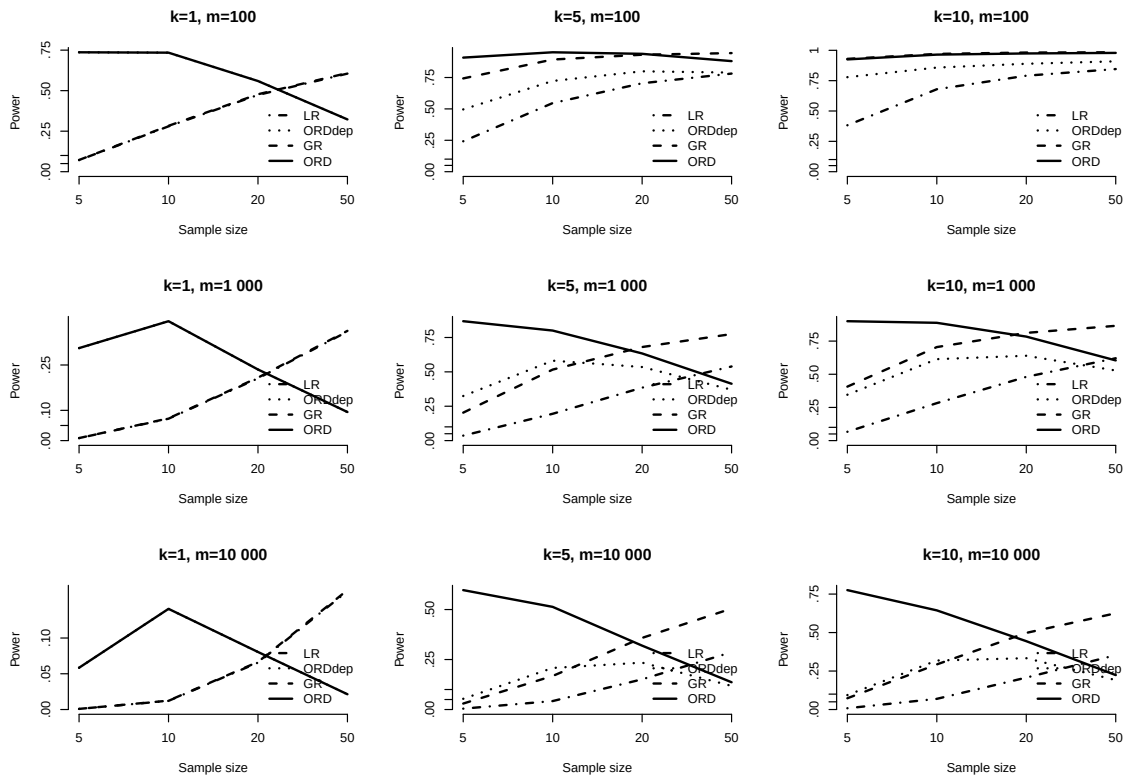

Figure 5. Estimated power for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 10%, the power of each single test at 70%. Variables have damped cosine dependence with $\tau = 7.5$.

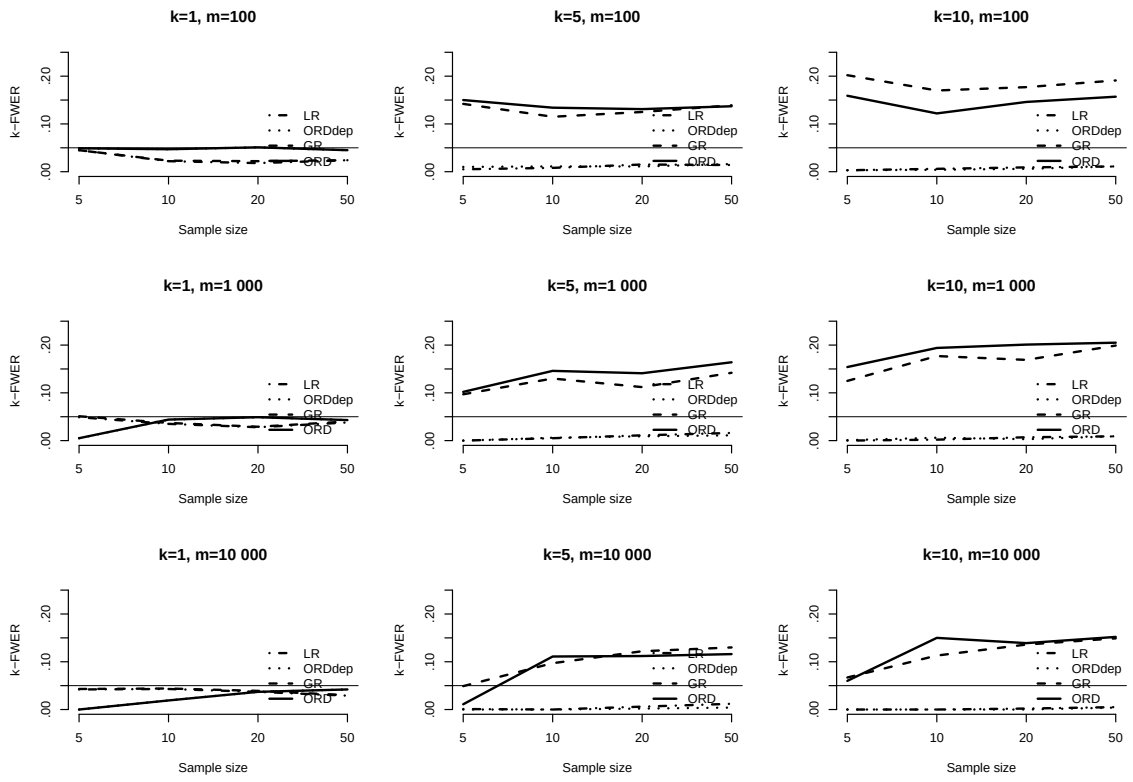

Figure 6. Estimated k-FWER different values of n and m. α is set to .05, the proportion of false nulls is 10%, the power of each single test at 70%. Variables have damped cosine dependence with $\tau = 7.5$.

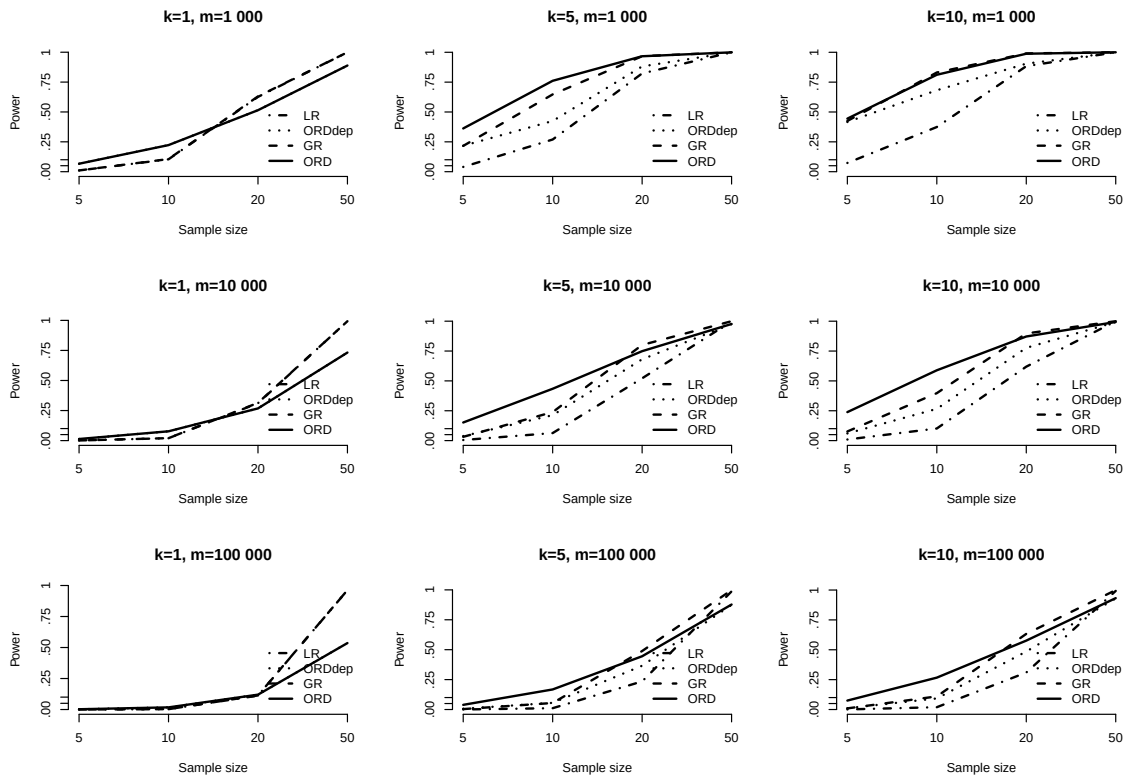

Figure 7. Estimated power for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 10% and the effect size $\delta = 1$.

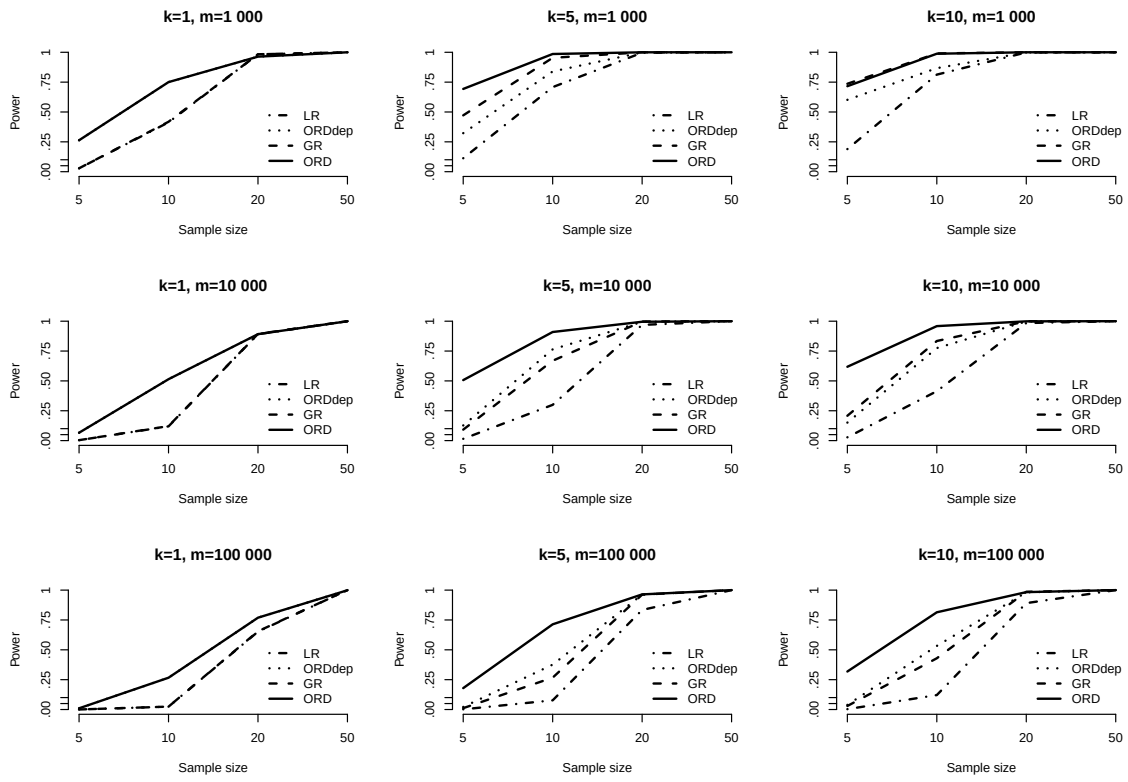

Figure 8. Estimated power for different values of k, n and m. α is set to .05, the proportion of false nulls is set at 10% and the effect size $\delta = 1.5$.