Appendix

Here we want to show that the sample correlation coefficients r_i and r_j converge in distribution to independent normally distributed random variables, with $n \to \infty$. Then taking these independent asymptotic distributions in our approximate formula for p (Equation 2), we obtain independence of Z_i, Z_j (as functions of independent variables generate independent variables).

We assume that the vectors X_i, X_j and Y come from normal distributions with zero mean. We define the matrix $M_{n\times 3} = \left[X_i^T, X_j^T, Y^T\right]$. M can be seen as n independent samples from p-variate ($p = 3$) normal distribution with zero mean and the covariance matrix $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$ (since X_i is independent of X_j and Y, and X_j and Y are correlated).

Then the matrix $S = \frac{1}{n} M^T M$ is the sample covariance matrix, where the elements $s_{1,3}$ and $s_{2,3}$ of this matrix represent the sample correlation coefficients r_i and r_j, respectively.

The matrix nS has the Wishart distribution $W_{p=3}(\Sigma, n)$ with n degrees of freedom. Asymptotic behaviour of this distribution was studied in many papers; here we use the result given in [5] which states that for $n \to \infty$ and fixed p, the matrix $\frac{1}{\sqrt{n}}(S - nI)$ (where I is the $p \times p$ identity matrix) converges in distribution to the random matrix Z whose elements are normally and independently distributed. Hence r_i and r_j converge in distribution to independent normally distributed random variables, for $n \to \infty$.

References

1. Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set?. Bioinformatics 21(2):171–178
2. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome of cancer. Proceedings of the National Academy of Sciences 103(15):5923–5928
3. Fisher RA (1915) Frequency distribution of the values of correlation coefficient in samples from an indefinitely large population. Biometrica 10(4):507–521
4. Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:3–32
5. Jonsson D (1982) Some Limit Theorems for the Eigenvalues of a Sample Covariance Matrix. Journal of Multivariate Analysis 12(1):1–38
6. Maciejewski H (2013) Predictive Modelling in High-Dimensional Data: Prior Domain Knowledge-Based Approaches. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław
7. Wu MC, Lin X (2009) Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. Statistical Methods in Medical Research 18(6):577–593