APPENDIX: ASYMPTOTIC CONSIDERATIONS

In this study, we have mainly focused on methodological development. In this Appendix, we provide some heuristic theoretical justification for our proposed approaches.

As the first step of the proposed study, the PCs of the original and expanded sets need to be consistently estimated. In our study, we compute the PCs from the sample variance-covariance matrix. Denote m as the pathway size and n as the sample size. It has been shown that if $m/n \to 0$, then under mild conditions, estimation of the first PC is consistent. Since we focus on a fixed number of PCs, consistency of the first c^* PCs requires the same $m/n \to 0$. With the expanded set, its dimension is $O(m^2)$. Thus we will require $m^2/n \to 0$ for analysis using nonlinear effects. In previous studies, regularization methods have been suggested to improve estimation of the variance matrix and hence also of the PCs. We refer to Bickel and Levina (2008) for more details. If certain assumptions on the variance matrix can be made, then less strict requirements on m can be assumed. For example in Bickel and Levina (2008), the "bandable" assumption has been made, and the requirement of $m/n \to 0$ can be replaced with the much looser $\log(m)/n \to 0$. However, we note that in Bickel and Levina and references therein, it is assumed that genes are only weakly correlated with most other genes. This can be a reasonable assumption for genes belonging to different pathways. In our study, we consider genes within the same pathways only. It is not clear whether such a bandable assumption is realistic in this context.

As the second step of the study, p-values from multiple pathways (and/or multiple sets of representative features per pathway) need to be combined and analyzed. Denote M as the total number of pathways and consider the special case where we are simply computing t-tests of gene effects, as in the setting of Kosorok and Ma (2007). If we can assume that the PCs are consistent as argued in the previous paragraphs and if we assume $\log(M)/n \to 0$, then uniform consistency of the p-values follows from arguments in Kosorok and Ma because the estimated PCs correspond to normalized data of dimension M and sample size n. It appears that these arguments can be extended to the regression model setting of the current paper, but we do not pursue the details further here.

REFERENCES

Bickel, P.J. and Levina, E. (2008) Covariance regularization by thresholding. *Annals of Statistics*. To appear.

Kosorok, M.R. and Ma, S. (2007) Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Annals of Statistics*, 35 (4), 1456-1486.