APPENDIX A. Proofs of the Theorems

APPENDIX A.1. Proof of Theorem 3.1

Proof. It is not difficult to observe that

$$\begin{gathered}\arg\min_{\boldsymbol{\mu}_k,\boldsymbol{\mu}_{mk}} L = (\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{x}_{ik}, k=1,...,K\},\\ \{\arg\min_{\boldsymbol{\mu}_{mk}} L_m, k=1,...,K\}),\end{gathered} \tag{A.1}$$

where

$$L_m := \frac{1}{2n_m}\sum_{k=1}^{K}\sum_{i\in C_m}\|\boldsymbol{x}^*_{ik}-\boldsymbol{\mu}_{mk}\|_k^2+\lambda\sum_{k=1}^{K}\omega_{mk}\|\boldsymbol{\mu}_{mk}\|_k. \tag{A.2}$$

The solution path to Equation (A.2) could be achieved by using Karush-Kuhn-Tucker conditions, presented as the following lemma:

Lemma A.1. *A necessary and sufficient condition for a vector $\boldsymbol{\mu}_m = (\boldsymbol{\mu}'_{m1},...,\boldsymbol{\mu}'_{mK})'$ with sparsity pattern $S_m := \{k : \boldsymbol{\mu}_{mk} \neq 0\}$ to be a solution to Equation* (A.2) *is:*

$$-(\boldsymbol{x}^*_{mk}/n_m-\boldsymbol{\mu}_{mk})+\lambda\cdot\frac{\omega_{mk}\boldsymbol{\mu}_{mk}}{\|\boldsymbol{\mu}_{mk}\|_k}=0, \qquad \forall \quad k\in S_m; \tag{A.3}$$

$$\|-\boldsymbol{x}^*_{mk}/n_m\|_k \le \lambda\omega_{mk}, \qquad \forall \quad k\in S_m^c, \tag{A.4}$$

where $\boldsymbol{x}^*_{mk} = \sum_{i\in C_m}\boldsymbol{x}^*_{ik}$.

It is not difficult to verify that the solution to Equations (A.3) and (A.4) can be presented as:

$$\widetilde{\boldsymbol{\mu}}_{mk} = (1-\frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+\widehat{\boldsymbol{\mu}}_{mk}. \tag{A.5}$$

This complete the proof. □

APPENDIX A.2. Proof of Theorem 3.2

Under assumption (A1) to (A3), we provide the following two Lemmas, which are necessary to prove the estimation consistency result.

Lemma A.2. *Under assumptions (A1), for all $m \in \{1,\dots,M\}$,*
for all $k \in S_m$, we define

$$A_{mk} := \{c_1 < \|\widehat{\boldsymbol{\mu}}_{mk}\|_k < c_2\}.$$

$\forall\ k \in S_m^c$, we define

$$B_{mk} := \{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k > \lambda\omega_{mk}\}.$$

Then for large enough n_m, we have

$$\begin{aligned}\mathbb{P}(A^c_{mk}) \le& \exp[-(\frac{-\sqrt{d_k+2v_{mk}}+\sqrt{-d_k+2n_mc_2^2}}{2})^2]\\ &+\exp[-\frac{(d_k+v_{mk}-n_mc_1^2)^2}{4(d_k+2v_{mk})}]\end{aligned}$$

and

$$\mathbb{P}(B_{mk}) \le \exp[-\frac{3}{16}d_k\cdot(\frac{n_m\lambda^2\omega_{mk}^2}{d_k}-1)^2] \tag{A.6}$$

where $v_{mk} = n_m\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk}$.

Lemma A.3. *Under assumptions (A1) and (A2), for all $m \in \{1,\dots,M\}$, for large enough n_m, and $\forall\ k \in S_m$, $\forall\ \epsilon > 0$, we have*

$$\begin{aligned}\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|A_{mk}) \le& \exp[-\tfrac{3}{16d_k}(\tfrac{c_3n_m\epsilon}{(1-\lambda_2)^2}-d_k)^2]\\ &+\phi(\|\boldsymbol{\mu}_{mk}\|^2 > \tfrac{\epsilon}{\lambda_1^2}),\end{aligned} \tag{A.7}$$

where $\lambda_2 = \frac{\lambda\omega_{mk}}{c_2}$, $\lambda_1 = \frac{\lambda\omega_{mk}}{c_1}$ and $\phi(\cdot)$ is the indicator function.

Combining Lemma A.2 and Lemma A.3, we have Theorem 3.2 which estimates the rate of convergence of $\widetilde{\boldsymbol{\mu}}_{mk}$ to $\boldsymbol{\mu}_{mk}$. We prove Theorem 3.2 as follows:

Proof. For any $m \in \{1,\dots,M\}$ and large enough n_m, $\forall\ k \in S_m$, $\forall\epsilon > 0$, we could use Lemma A.2 and Lemma A.3:

$$\begin{aligned}&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon)\\ =&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|A_{mk})\mathbb{P}(A_{mk})\\ &\quad+\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|A^c_{mk})\mathbb{P}(A^c_{mk})\\ \le&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|A_{mk})+\mathbb{P}(A^C_{mk})\\ \le&\exp(-\frac{3}{16d_k}(\frac{c_3n_m\epsilon}{(1-\lambda_2)^2}-d_k)^2)+\phi(\|\boldsymbol{\mu}_{mk}\|^2 > \frac{\epsilon}{\lambda_1^2})\\ &+\exp[-(\frac{-\sqrt{d_k+2v_{mk}}+\sqrt{-d_k+n_mc_2^2}}{2})^2]\\ &+\exp[-\frac{(d_k+v_{mk}-2n_mc_1^2)^2}{4(d_k+2v_{mk})}]\end{aligned}$$

$\forall\ k \in S_m^c$, $\forall\epsilon > 0$, similarly we have

$$\begin{aligned}&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon)\\ \le&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|B^C_{mk})+\mathbb{P}(B_{mk})\\ \le&\exp[-\frac{3}{64d_k}\cdot(n_m\lambda^2\omega_{mk}^2-2d_k)^2]\end{aligned}$$

The followings are straight forward calculations. □

APPENDIX A.3. Proof of Theorem 3.3

Proof. Using Equation (3.11) and assumption (A4), we have

$$[(\widehat{S}_1 = S_1)\cap(\widehat{S}_2 = S_2)] \subset [\widehat{S}_1\cup\widehat{S}_2 = S_1\cup S_2] \subset [\widehat{\mathcal{S}} = \mathcal{S}],$$

which implies that

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) &\le \mathbb{P}((\widehat{S}_1 \neq S_1)\cup(\widehat{S}_2 \neq S_2))\\ &\le \mathbb{P}(\widehat{S}_1 \neq S_1)+\mathbb{P}(\widehat{S}_2 \neq S_2).\end{aligned}$$

Using the same argument as in Theorem 3.2, we further have for $m \in \{1, 2\}$,

$$\mathbb{P}(\widehat{S}_m \neq S_m) \leq \sum_{k \in S_m} \mathbb{P}(A_{mk}^c) + \sum_{k \in S_m^c} \mathbb{P}(B_{mk}).$$

Choosing $\epsilon = \gamma\lambda^2$ and λ^2 sufficiently small, using Theorem 3.2, we have

$$\begin{aligned}\mathbb{P}(\widehat{S}_m \neq S_m) &\asymp \sum_{k=1}^{K} \mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 \geq 2\epsilon) \\ &\asymp K \cdot \max_k \mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 \geq 2\epsilon) \\ &\asymp K \cdot O(\exp(-c_4 n^2 \lambda^4)),\end{aligned}$$

where $c_4 = \min_k(\frac{3\gamma^2 c_3^2}{16 d_k}, \frac{3\omega_0^4}{64 d_k})$. Therefore

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) &\leq \mathbb{P}(\widehat{S}_1 \neq S_1) + \mathbb{P}(\widehat{S}_2 \neq S_2) \\ &\leq 2K \cdot O(\exp(-c_4 n^2 \lambda^2)) \to 0.\end{aligned}$$

This completes the proof. □

APPENDIX A.4. Proof of Theorem 3.4

Proof. Denote $\mathcal{A} = \{\widehat{\mathcal{S}} = \mathcal{S}\}$, then by Theorem 3.3,

$$\mathbb{P}(\mathcal{A}) = 1 - o(1),$$

and according to the proof in Theorem 3.3, we have for $j \in \{1, 2\}$,

$$\|\widetilde{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\|_2^2 = O_P(\epsilon) = O_P(\frac{\omega_0\sqrt{\log(n)\log(K)}}{n}). \tag{A.8}$$

Using the same techniques in (Bickel and Levina 2004)'s proof on theorem 2, we could prove that,

$$|\Psi_\Sigma(\widetilde{\Delta}, \widehat{\xi}) - \Psi_\Sigma(\Delta, \xi)| \leq C(\|\widetilde{\Delta} - \Delta\|_2^2 + \|\widehat{\xi} - \xi\|_2^2),$$

for all such that $\|\widetilde{\Delta} - \Delta\|_2 \leq \epsilon_1$, $\|\widehat{\xi} - \xi\|_2 \leq \epsilon_2$ with ϵ_1, ϵ_2 small enough. C is a constant only depending on the model assumptions.
And because of $\widetilde{\Delta} \to \Delta$ and $\widehat{\xi} \to \xi$, given n large enough such that both $\|\widetilde{\Delta} - \Delta\|_2 \leq \epsilon_1$ and $\|\widehat{\xi} - \xi\|_2 \leq \epsilon_2$ hold, we have

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq C\big(\|\widetilde{\Delta} - \Delta\|_2^2 + \|\widehat{\xi} - \xi\|_2^2\big).$$

For the first term, applying Equation (A.8), we have

$$\|\widetilde{\Delta} - \Delta\|_2^2 = O_P(s \cdot \frac{\omega_0\sqrt{\log(n)\log(K)}}{n}).$$

For the second term, given that $\mathcal{A}$ holds, we have

$$\begin{aligned}&\|\widehat{\xi} - \xi\|_2^2 \\ =&\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}}(\widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1)_{\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \\ &\quad\text{(for large enough } n\text{)} \\ =&\|((\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}} + \\ &(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} \cdot ((\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}} - (\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}})\|_2^2 \\ \leq&\|(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 + \\ &(\|(\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}}\|_2^2 + \|(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2) \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}}\|^2 \\ \leq&\|(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 + \\ &(\|(\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}}\|_2^2 + \|(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2) \cdot \\ &(\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\| + \|(\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|)^2.\end{aligned}$$

Because

$$\begin{aligned}&\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 \\ &\leq (\|(\widehat{\Sigma}_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma_{\mathcal{M}\mathcal{M}})^{-1}\| + \|(\Sigma_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|)^2 \\ &= (O_P(s\omega_0^2\sqrt{\frac{\log s \log \omega_0}{n}}) + a_{n,d})^2,\end{aligned}$$

where the last inequality is by applying Lemma (A.3) of (Bickel and Levina 2008), we have

$$\|\widehat{\xi} - \xi\|_2^2 = O_P(c_s s^2 \omega_0^4 \cdot \frac{\log s \log \omega_0}{n}) + c_s O(a_{n,d}^2). \tag{A.9}$$

This completes the proof. □

APPENDIX B. The General Trend by Changing λ

The minimum averaged misclassification errors for gNSC and NSC are highlighted in bold for both cross validation procedures. The results are shown in Table 3 and Table 4.

Table 3: Leave experiment out cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold.

λ	gNSC	gene number	NSC	gene number
1.0	0.158(0.3644)	10982(1.3)	0.301(0.0823)	8611(23.39)
2.5	0.150(0.3682)	10189(74.07)	0.240(0.0834)	5778(21.62)
3.5	0.175(0.3715)	8376(16.51)	**0.149(0.0479)**	4495(17.26)
4.0	0.168(0.3488)	7380(15.94)	0.150(0.0487)	3975(14.49)
5.5	0.085(0.1054)	5179(13.48)	0.251(0.0455)	2888(9.89)
6.5	**0.089(0.0505)**	3901(22.50)	0.275(0.0925)	2251(9.53)
7.0	0.120(0.0505)	3515(26.12)	0.275(0.0925)	2081(10.43)
8.5	0.120(0.0505)	2265(40.39)	0.245(0.0950)	1594(10.72)
10.0	0.120(0.0505)	1182(38.05)	0.222(0.0962)	1196(10.38)

Table 4: 10-fold cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold.

λ	gNSC	gene number	NSC	gene number
1.0	0.141(0.0945)	10982(0.52)	0.145(0.0892)	8538(8.36)
2.5	0.143(0.0892)	10054(19.97)	**0.136(0.0906)**	5670(11.14)
3.0	0.144(0.0843)	9188(26.95)	0.140(0.0900)	4975(9.88)
4.0	0.146(0.0806)	7154(26.06)	0.137(0.0899)	3871(9.17)
5.0	**0.139(0.0866)**	5502(21.33)	0.139(0.0852)	3057(7.86)
5.5	0.181(0.0866)	4874(17.54)	0.144(0.0882)	2735(7.77)
7.0	0.182(0.0923)	3356(15.22)	0.156(0.0852)	2009(5.86)
8.5	0.184(0.0911)	2021(14.62)	0.164(0.0827)	1503(5.53)
10.0	0.189(0.0920)	1052(11.19)	0.170(0.0858)	1143(4.97)

APPENDIX C. QQ Plots of the Data

For each gene in each sample class we present the Quantile-to-Quantile plot (QQ plot) to visualize the normality. Three of them are shown in Figure 4.

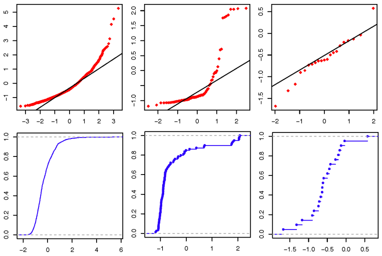
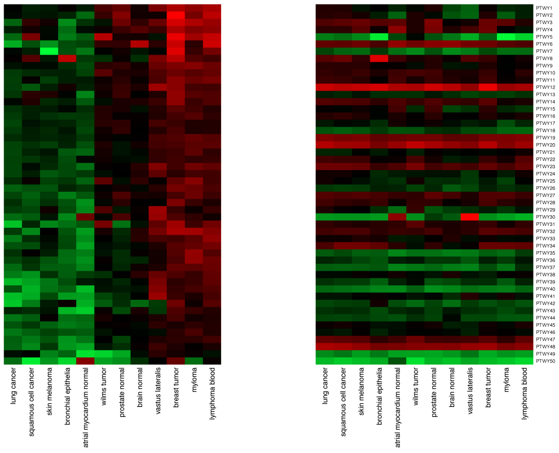


Figure 4: Non-Gaussian Data. The two rows are QQ-plot and empirical cdf plot of three different genes.

APPENDIX D. Heatplots of gNSC on GPL96

To illustrate our result more clearly, we randomly pick 12 tissue types and 100 gene pathways for visualization. Figure 5(a) presents the negative "shrinkage amount," i.e. $(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})$ shown in Equation (3.5), of the combination of one certain gene pathway indexed by k and one certain tissue type indexed by m. closer the color to green, the lower the negtive shrinkage amount is. The Figure 1 presents the significant associations, i.e. the threshold term $(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+$, between gene pathways and tissue types: red color suggests that the corresponding pathway and tissue type are estimated to be associated, while green suggests not. Moreover, the expression levels in one block are summarized to be the mean of all gene expression values confined in this block and the result can be observed in Figure 5(b). Figures 5(a) and 1 illustrate the statistical significance levels and Figure 5(b) illustrates the biological expression levels. A detailed index for all gene pathways are presented in the Appendix F.



(a) (b)

Figure 5: Heatplots for gNSC. The 100 pathways in this figure are randomly chosen.

APPENDIX E. Heatplot of the Keyword-Gene Relevance

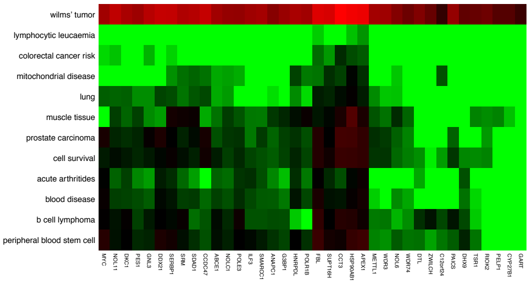


Figure 6: NSC Results of Keywords v.s. Genes.

APPENDIX F. Gene Pathways

Table 5: The index of all gene pathways IDs.

ID	Pathway name
PTWY1	landis erbb2 breast preneoplastic up
PTWY2	waesch anaphase promoting complex
PTWY3	sanchez mdm2 targets
PTWY4	naderi breast cancer prognosis dn
PTWY5	begum targets of pax3 foxo1 fusion and pax3
PTWY6	der ifn alpha response dn
PTWY7	negative regulation of cytokine biosynthetic process
PTWY8	aldo keto reductase activity
PTWY9	xu hgf signaling not via akt1 6hr
PTWY10	valk aml cluster 15
PTWY11	nam fxyd5 targets dn
PTWY12	creighton akt1 signaling via mtor dn
PTWY13	gargalovic response to oxidized phospholipids green up
PTWY14	appierto response to fenretinide up
PTWY15	chiaradonna neoplastic transformation kras cdc25 dn
PTWY16	yao temporal response to progesterone cluster 8
PTWY17	tonks targets of runx1 runx1t1 fusion granulocyte up
PTWY18	ebauer targets of pax3 foxo1 fusion dn
PTWY19	zhan v2 late differentiation genes
PTWY20	hu angiogenesis dn
PTWY21	shipp dlbcl vs follicular lymphoma dn
PTWY22	mori mature b lymphocyte dn
PTWY23	caffarel response to thc dn
PTWY24	nucleotide metabolic process
PTWY25	yang breast cancer esr1 bulk dn
PTWY26	hoegerkorp cd44 targets temporal dn
PTWY27	vesicle membrane
PTWY28	nikolsky breast cancer 1q21 amplicon
PTWY29	naderi breast cancer prognosis up
PTWY30	rickman head and neck cancer f
PTWY31	myllykangas amplification hot spot 9
PTWY32	scheidereit ikk targets
PTWY33	radaeva response to ifna1 dn
PTWY34	zhang interferon response
PTWY35	regulation of t cell activation
PTWY36	finetti breast cancers kinome gray
PTWY37	positive regulation of lymphocyte activation
PTWY38	cortical cytoskeleton
PTWY39	rna polymerase ii transcription mediator activity
PTWY40	meiotic cell cycle
PTWY41	kim response to tsa and decitabine dn
PTWY42	vitamin transport
PTWY43	regulation of rho gtpase activity
PTWY44	brain development
PTWY45	regulation of binding
PTWY46	ginestier breast cancer znf217 amplified up
PTWY47	zhan multiple myeloma subgroups
PTWY48	proteasome complex
PTWY49	organic anion transmembrane transporter activity
PTWY50	krishnan furin targets dn

APPENDIX G. Context Analysis

Microarray data often have text documents of samples (e.g. sample description, experiment description, etc.). We propose a new "index-gene relevance" approach to applying NSC or gNSC to context analysis.

More precisely, we encode relationships between representative medical terms, and pertaining genes in matrix form, to which we apply NSC or gNSC. The procedure of creating the index-gene relevance matrix includes four steps: preparing documents for each microarray sample; creating a tf-idf (defined later) based term-document matrix; creating an index-document matrix; and lastly combining text information with the microarray data. We describe the details of these steps below.

Step 1. Document Preparation. In order to use the text information efficiently, we extract the biologically meaningful words and phrases from the text description files and map them into existing knowledge sources. For each sample, we produce a text file with sample specific information, in which "meaningful phrases" and related words are organized into rows. Rather than describing the semantic criteria by which we extract "meaningful phrases," we simply provide an example.

Table 6: Document Preparation

Procedure
1: Download the file GPL96_family.soft.gz (`www.affymetrix.com`) with text description of the samples and the experiments information of the samples.
2: Retrieve sample information (GSM files, description of the specific sample information involved in one entire experiment) for all experiments from the GPL96 family.soft file.
3: Extract the biologically meaningful sample and experiment information from GSM files. Sample title, sample source, sample organism, sample characteristics and sample description are extracted.
4: Use MetaMap to map all the information extracted from previous step to several knowledge sources, including GO, MSH, HUGO, OMIM and NCI. MetaMap can break the input text into several phrases by its lexical/syntactic analysis and then map those phrases to the knowledge source.
5: We denote each phrase together with its all related words (all the words are lower-cased) as a block (one row in the file) in the text document file, which has the same name as the original GSM files.

To utilize the text information of GPL96 data, we prepare the sample documents in five steps as shown in Table 6.

Step 2. TF-IDF based Term-doc Matrix. "Term frequency-inverse document frequency" (tf-idf) (Wu, Luk, Wong and Kwok 2008) is one of the most commonly used relevance weighting factors in today's information retrieval and text mining systems, and is our preferred relevance metric. The tf-idf value increases proportionally to the number of times a word appears in a specific document, but is offset by the frequency of the word in the corpus. This provides a good measure of relevance which controls for the fact that some words are generally more common than others.

Let doc_i be the text document of sample i and $Doc = \{doc_i : i = 1, 2, \cdots, n\}$ be the set of all documents. Each document is represented as a list of words:

$$doc = (w_1, ..., w_{N_{doc}}),$$

where w_i $(i = 1, 2, \cdots, N_{doc})$ are the words in the doc document, including repetition. N_{doc} is the total number of words in doc. We extract all *distinct* words from all documents and use W to represent the set of all words:

$$W = \cup_{doc \in Doc} \cup_{j=1}^{N_{doc}} \{doc(j)\}, \qquad (G.1)$$

where $doc(j)$ is the j-th word of doc. Note that W here is a set of word where each elements is unique. For each extracted term w and the document doc, we define the term-count $tc(w, doc)$ to be the number of times that the term w appears in doc.

To prevent the bias towards longer documents, the term-frequency(tf) is defined as:

$$tf(w, doc) = \frac{tc(w, doc)}{|doc|}, \tag{G.2}$$

where $|doc| = N_{doc}$ is the length (total number of words) of the document doc.

In addition, we introduce the inverse document frequency for each word w which is a measure of the general importance of w:

$$idf(w) = \log \frac{|Doc|}{|\{doc : w \in doc\}|}, \tag{G.3}$$

where $|Doc|$ is the total number of documents which equals the number of samples and $|\{doc : w \in doc\}|$ is the number of documents in which the word w appears. Based on tf-idf score we build the term-doc matrix tdM:

$$tdM(w, doc) = tf(w, doc)idf(w). \tag{G.4}$$

As we can see, tf-idf score is

1. highest when the term w appears many times within a small number of documents;
2. lower when the term w appears fewer in a document, or appears in too many documents;
3. lowest when the term appears in nearly all documents.

Step 3. Index-doc Matrix. The tf-idf based term-doc matrix reflects how important a word is to a document in the corpus. However, we are not only interested in words but also biologically pertinent phrases with multiple terms (e.g. “breast cancer,” “brain tumor,” etc.). We measure the relevance of such phrases as follows:

Let $p_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,|p_i|})$ be any word ($|p_i| = 1$) or phrase ($|p_i| \geq 2$) in the dictionary, where $w_{i,j}$ ($j = 1, 2, \cdots, |p_i|$) are the single terms in p_i and $|p_i|$ is the number of terms. Let $P = (p_1, p_2, \cdots, p_N)$ be the list of all words and phrases. Let ind_i be the index of p_i such that words or phrases with same meaning (e.g., “brain” and “brains”) in the dictionary have the same index. Therefore, each index can represent a synonym word group. Let $I = \{ind_1^*, ind_2^*, \cdots, ind_M^*\}$ be the set of indices, where each element is unique. Note that since the indices of the words and phrases in the word list P are not unique, M need not equal N. We build a index-doc matrix, idM, based on the tf-idf score:

$$idM(ind^*, doc) = \max_{i:ind_i = ind^*} \left\{ \frac{\sum_j tdM(w_{i,j}, doc)}{N_{p_i}} \right\}, \tag{G.5}$$

where each element is the maximum of the mean tf-idf score for words and phrases with same meaning.

Step 4. Gene-Index Relevance Matrix. In order to apply NSC to context analysis, we combine the text information with our microarray data. We generate the relevance matrix of words/phrases with genes based on the gene expression levels in microarray data and the index-doc matrix generated above. Let the m-th synonym group be the words or phrases with their index $= ind_m^*$. We can use ind_m^* can represent this group.

The index-doc matrix we get from Step 3 can be seen as a measure of relevance of synonym word groups and samples, while the gene expression levels in microarray data can be seen as a measure of gene-sample relevance. We measure the connection of gene and synonym groups by multiplying the elements of the above two relevance matrix.

In mathematics, we measure the connection of the m_{th} synonym group and the j_{th} gene for a given sample i as:

$$R_m(g_j, doc_i) = idM(ind_m^*, doc_i) \times x_{ij}, \tag{G.6}$$

where x_{ij} is the expression level of j_{th} gene in sample i and idM is the index-doc matrix in Step 3.

To respect the structure of the NSC input data, we collect the R matrices to form the gene-index relevance matrix:

$$R = (R_1, R_2, \cdots, R_M). \tag{G.7}$$

Here each synonym group is regarded as a sample class, i.e. C_m, in NSC.

By applying NSC to the gene-index relevance matrix, we can select the most relevant genes with all words and phrases in the dictionary. Moreover, we can trivially generalize the NSC approach to context analysis to the group version by first combining the pathway information with the microarray data before using gNSC to the corresponding gene-index relevance matrix.

Remark G.1. *As alluded to previously (cf. Remark 4.1), although the dimension of the gene-index relevance matrix R is $d \times (n \times M)$, which can grow prohibitively large, our process can still be completed in minutes. This efficiency is due to the use of sufficient statistics.*

Table 7: NSC(/gNSC) for Context Analysis

Algorithm
1: Extract the biological meaningful words and phrases from the text description files of each microarray sample and map them into existing knowledge sources. Prepare a text document file for each sample.
2: Calculate the term-frequency(tf) and the inverse document frequency(idf) for each word w, then multiply them to get the tf-idf score for each word and each document. Build the term-doc matrix(tdM) based on the score.
3: Divide the words into synonym groups and calculate the index-doc matrix(idM), based on equation G.5, for each synonym group and each document.
4: For gNSC, sort the genes in the microarray data by pathways.
5: Calculate the R matrixes and bind them together to get the gene-index relevance matrix.
6: Apply NSC or gNSC on the gene-index relevance matrix to selection the significant genes or pathways for each word or phrase in the dictionary.