Probabilistic Boolean Tensor Decomposition – Supplementary Information

Tammo Rukat [1 2] **Chris C. Holmes** [1 3] **Christopher Yau** [4]

A. Derivation of the conditionals

Here we derive the full conditional distribution for a factor entry $f_{n_k l}$ as given in eq. (4) in the main paper. For constant priors, $p(f_{n_k l}) = \text{const.}$, the conditional is given by normalising the likelihood for $f_{n_k l} \in \{0, 1\}$. The likelihood has the form

$$p(x_{[n]}|.) = \sigma\left[\lambda \tilde{x}_{[n]}\left(1 - 2\prod_{l}\left[1 - \prod_{n\in[n]} f_{nl}\right]\right)\right] \tag{1}$$

and is factorial in the data-points $x_{[n]}$. Terms that do not depend on $f_{n_k l}$ cancel in the conditional and thus we take the product over all $[n]$ with n_k fixed. Then normalising gives

$$p(f_{n_k l} = 1|.) = \frac{\prod_{[n],\ n_k\text{fixed}} p(x_{[n]}|f_{n_k l} = 1, \text{rest})}{\prod_{[n],\ n_k\text{fixed}} p(x_{[n]}|f_{n_k l} = 1, \text{rest}) + \prod_{[n],\ n_k\text{fixed}} p(x_{[n]}|f_{n_k l} = 0, \text{rest})} \tag{2}$$

$$= \sigma\left[\sum_{\substack{[n]\\ n_k\text{fixed}}} \log \frac{p(x_{[n]}|f_{n_k l} = 1, \text{rest})}{p(x_{[n]}|f_{n_k l} = 0, \text{rest})}\right] . \tag{3}$$

Considering the term inside the logarithm in eq. (3) and using eq. (1) we find

$$\frac{p(x_{[n]}|f_{n_k l} = 1, \text{rest})}{p(x_{[n]}|f_{n_k l} = 0, \text{rest})} = \begin{cases} 1;\ \text{if } \left(\prod_{n\in[n]/n_k} f_{nl}\right)\prod_{l'\neq l}\left(1 - \prod_{n\in[n]} f_{nl'}\right) = 0 \\ \frac{1+\exp(\lambda\tilde{x}_{[n]})}{1+\exp(-\lambda\tilde{x}_{[n]})} = e^{\lambda\tilde{x}_{[n]}};\ \text{otherwise} . \end{cases} \tag{4}$$

The first equality describes all cases where the term inside the parenthesis in eq. (1) takes a value that is independent of the value of $f_{n_k l}$. The second term follows in all other cases and by expanding the logistic sigmoid. Hence, we can write eq. (3) as

$$p(f_{n_k l} = 1|.) = \sigma\left[\lambda \sum_{\substack{[n]\\ n_k\text{fixed}}} \tilde{x}_{[n]} \overbrace{\left(\prod_{n\in[n]/n_k} f_{nl}\right)\prod_{l'\neq l}\left(1 - \prod_{n\in[n]} f_{nl'}\right)}^{M_{(n_k,l)\to[n]}}\right] . \tag{5}$$

[1]Department of Statistics, University of Oxford, UK [2]The Alan Turing Institute, London, UK [3]Nuffield Department of Medicine, University of Oxford, UK [4]Centre for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, UK. Correspondence to: Tammo Rukat <tammo.rukat@stats.ox.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

B. Cancer-type legend

- BLCA: Bladder Urothelial Carcinoma
- BRCA: Breast invasive carcinoma
- CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma
- COAD: Colon adenocarcinoma
- ESCA: Esophageal carcinoma
- GBM: Glioblastoma multiforme
- HNSC: Head and Neck squamous cell carcinoma
- KIRC: Kidney renal clear cell carcinoma
- KIRP: Kidney renal papillary cell carcinoma
- LAML: Acute Myeloid Leukemia
- LGG: Brain Lower Grade Glioma
- LIHC: Liver hepatocellular carcinoma
- LUAD: Lung adenocarcinoma
- LUSC: Lung squamous cell carcinoma
- OV: Ovarian serous cystadenocarcinoma
- PAAD: Pancreatic adenocarcinoma
- PCPG: Pheochromocytoma and Paraganglioma
- PRAD: Prostate adenocarcinoma
- SARC: Sarcoma
- SKCM: Skin Cutaneous Melanoma
- STAD: Stomach adenocarcinoma
- TGCT: Testicular Germ Cell Tumours
- THCA: Thyroid carcinoma

C. Principal component analysis of gene expression data

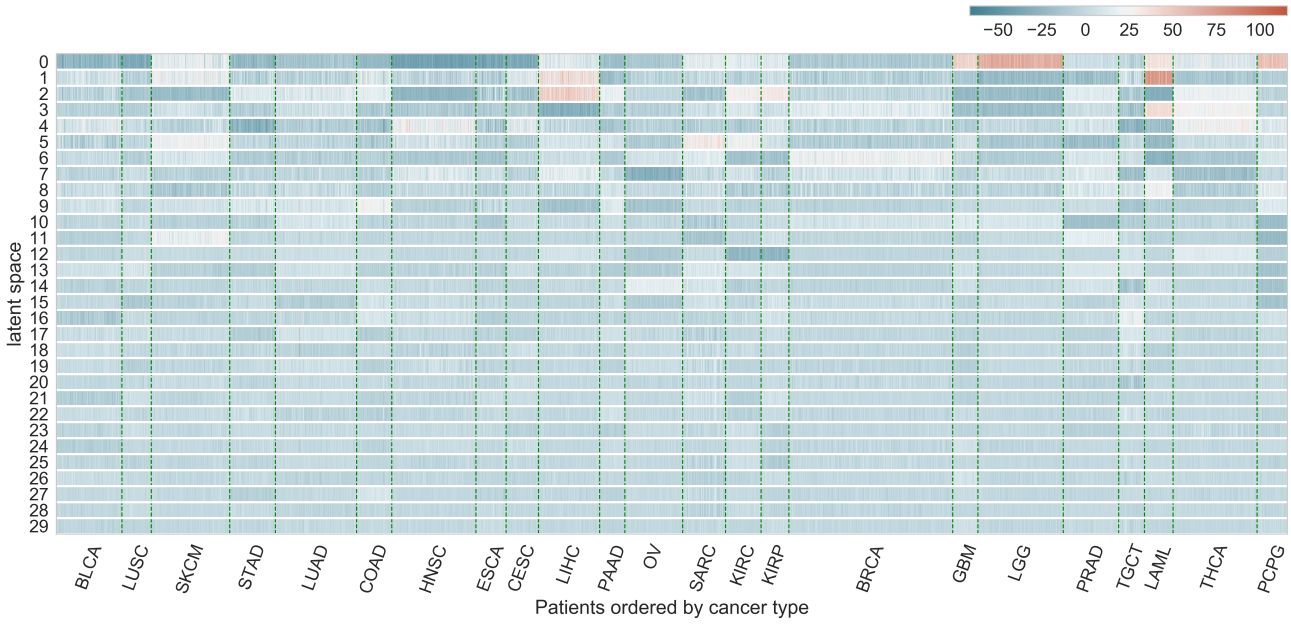

Figure 1: Latent representation of the patient×gene-expression under principal component analysis and arranged in analogy to Fig. 4(c) in the main paper.

D. DBTF on Gene Expression Data

Fig. 2 shows patient representation for relative gene expression networks computed by dbtf. The analysis is limited to 350 genes, since our 32 core, 128 GB machine runs out of memory for larger analyses using the implementation provided by Park et al. [2017]. Data that was treated as missing in our original analysis is treated as 0 here. Comparing to Fig. 4(c) we observe similar but noisier disease-specific patterns.

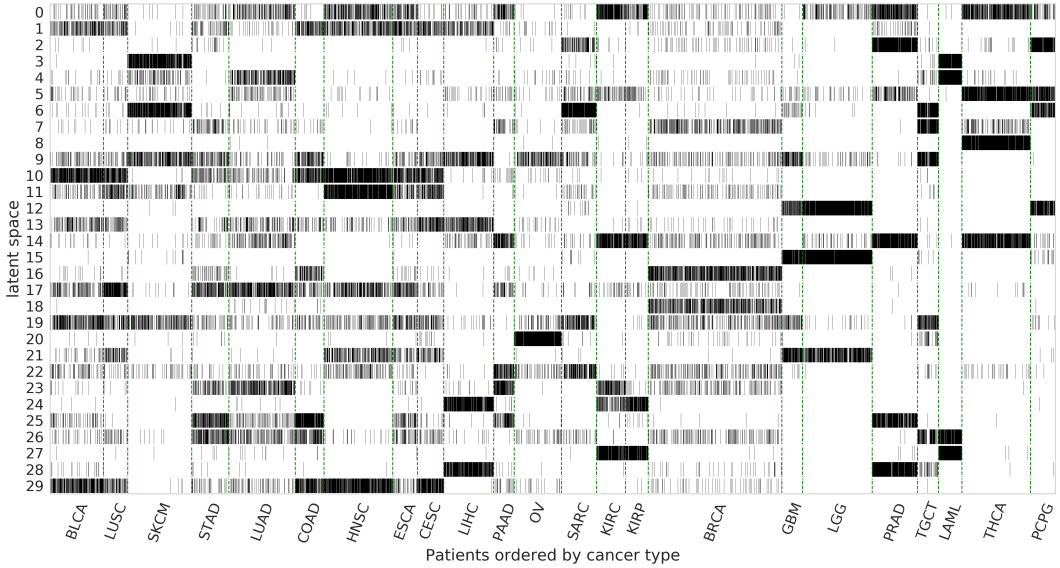

Figure 2: Latent representation of the patient×gene-expression under dbtf arranged in analogy to Fig. 4(c) in the main paper.

E. Real World Data Analysed with dbtf

Fig. 3 shows the results of dbtf, corresponding to the analysis in Fig. 4(b) of the main paper. While this solution looks similar to a possible point estimate of the TensOrM analysis, dbtf lack the ability to characterise posterior uncertainty which is crucial in this example. For the hospital data, dbtf infers only 3 latent dimensions. These correspond to dimensions 2, 5, and 7 of the TensOrM analysis. Note, that we can not assess the predictive performance since dbtf is unable to deal with held-out data

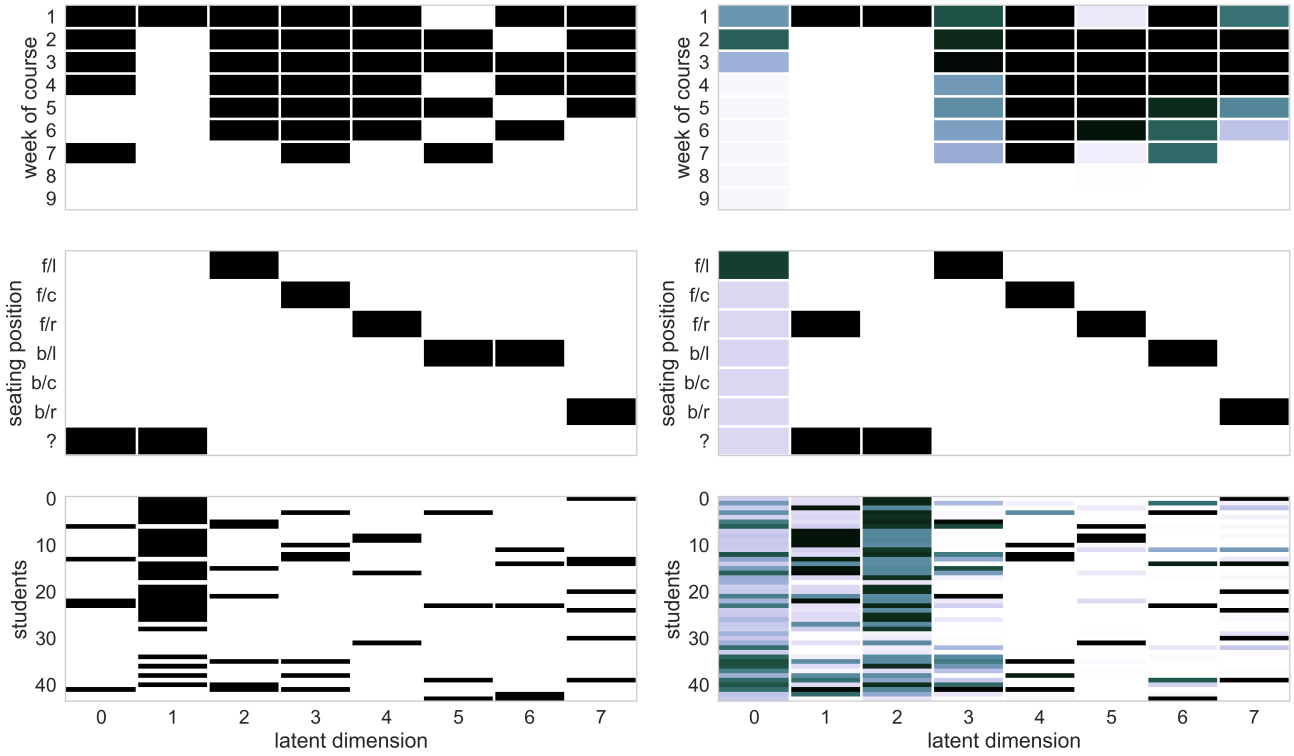

Figure 3: Student seating throughout course, analysed with dbtf (left) The latent representation are ordered to emphasis similarity to Fig. 4(b) in the main paper, a copy of which is shown on the right-hand-side for easier reference. The correspondence is as follows [dbtf latent dimensions $\rightarrow$ TensOrM latent dimensions]: [0 $\rightarrow$ (0,1)], [1 $\rightarrow$ 2], [2 $\rightarrow$ 3], [3 $\rightarrow$ 4], [4 $\rightarrow$ 5], [(5,6), $\rightarrow$ 6], [7 $\rightarrow$ 7].