Drug response in cancer treatments in a Bayesian Multiple Task Multiple Kernel Learning framework using Variational Inference for updating

Master Thesis

to

Dr. Patrick Jähnichen

Humboldt-Universität zu Berlin
Institute for Informatics
Department of Knowledge Management in Bioinformatics

and

Prof. Dr. Bernd Droge

Humboldt-Universität zu Berlin
School of Business and Economics
Chair of Econometrics

by

Henry Webel

in partial fulfillment of the requirements
for the degree of
Master of Science
Berlin, January 30, 2018

Acknowledgement

I would like to thank the following people helping me to finish this thesis: Manuela Benary for explaining the data. Understanding and adding pathways would have not been possible without her help using biological databases. Bernd Droge for counselling on notation and statistical measures. James Costello for pointing to other interesting studies and explanations on the DREAM7 challenge. Patrick Jähnichen for explaining in depth Bayesian inference and directed acyclic graphs. He took many hours to give me hints leading to major progress. Mehmed Gönen for his hints on his model and hints on deriving it. Ammad-Uddin Muhammad for the explanations on data loading in the community study and hints on implementations. My parents for the support over all this years without I would have never been able to write a master thesis after all. And Christopher Lennan, Alexander Maywurm and Jan Reher for reading drafts of this thesis and providing insightful comments.

Abstract

Cancer is heterogeneous and therefore effectiveness of therapies is hard to predict. Personalized medical approaches envision to select drugs for treatment based on genetic dispositions. An approach modeling several gene informations and drugs simultaneously, called Bayesian multiple task multiple kernel learning (BMTMKL), is found to provide a flexible framework to analyse drug response. Both BMTMKL's ability to rank human cell samples according to their drug response and its ability in predicting the level of concentration needed in human tissue to obtain fifty percent growth inhibition is compared to a naive approach using random rankings or training means for prediction. It provides additional information for doctors which drugs to consider for patients in clinical trials. For building an understanding of the modeling approach, the training procedure using variational inference is derived.

Contents

List of Abbreviations v

List of Figures v

List of Tables v

1 Introduction 1

2 Model and Inference 3

2.1 The model . . . 3

2.2 Full conditional densities of latent variables . . . 8

2.2.1 Full conditional for the precision λ on kernel weights . . . 9

2.2.2 Full conditional for kernel weights a . . . 10

2.2.3 Full conditional for intermediate outputs G . . . 13

2.2.4 Full conditional for precision υ on intermediate outputs . . . 17

2.2.5 Full conditional on precision γ for biases . . . 18

2.2.6 Full conditional for precision ω on kernel coefficients . . . 18

2.2.7 Full Conditional on biases b and kernel coefficients e . . . 19

2.2.8 Full conditional for precision ϵ on outputs . . . 24

2.3 Log evidence as lower bound . . . 26

2.4 Lower bound of the model . . . 28

2.5 Parameter updating . . . 32

2.6 Parameter updates for the proposed model . . . 36

2.6.1 Updating the distributional parameters of the factors . . . 36

2.6.2 Updating hyperparameters . . . 38

2.7 Predictions or fitted values: . . . 43

3 Data 45

3.1 Drug data . . . 45

3.2 Six profiling datasets . . . 46

3.3 Additional data sets: pathway information and interactions . . . 48

3.3.1 Pathway information . . . 48

3.3.2 Interactions between profiling datasets . . . 49

3.4 Kernels 50
3.5 Data Limitations 51

4 Results **53**
4.1 Model evaluations - different models of grid search 53
4.2 Model evaluation in DREAM7 64
4.2.1 Challenge scoring method - wpc-index 64
4.2.2 Applying the wpc- index to models trained in Section 4.2.1 66

5 Conclusions **69**

References **70**

A Background Information **73**
A.1 Distributions 73
A.1.1 Gamma distribution 73
A.1.2 Normal distribution 74
A.1.3 Proportionality 75
A.2 Basic (Bayesian) Regression 75
A.3 Expectation operator: 77

B Definitions **77**

C Tables **79**

D Declaration of academic honesty **93**

List of Abbreviations

BMTMKL	Bayesian Multiple Task Multiple Kernel Learning
CNV	Copy Number Variation
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
GI_{50}	concentration of a drug at which growth is inhibited by 50 percent
HGNC_ID	HUGO Genome Nomenclature Committee Identifier
HUGO	Human Genome Organization
MSE	Mean Squared Error
NA	Not Available
Predict	comPREhensive Data Integration for Cancer Treatment
RPPA	Reverse Phase Protein Array
RNA	Ribonucleic Acid
r.v.	random variable
wpc-index	weighted probabilistic concordance index

List of Figures

1 Graphical model representation 6
2 Fitted growth inhibition curve for a hypothetical drug 47
3 Dimensionality reduction performed in BMTMKL for all cell lines with measured GI_{50} concentration for drug t 52
4 MSEs and ELBO for random model using standardized drug responses 61
5 MSEs and ELBO for ones model using standardized drug responses 61
6 MSEs and ELBO for default model using standardized drug responses 62

List of Tables

1 Unobserved real valued random variables of the model and their representations 4
2 Observed deterministic and real valued random variables and their representations 5
3 Full conditionals of *sets of random variables*, denoted factors 25
4 Distributions in joint 28

5 Six profiling data sets used in DREAM7 challenge (Costello et al., 2014) . . . 48
6 Pathways as reported in Costello et al. (2014) and in this thesis 50
7 Interacted profiling datasets 50
8 Five configurations of grid search on hyperparameters shape $\alpha \in \{10^{-10}, 1, 10\}$ and scale $\theta \in \{10^{-10}, 0.01, 1\}$ and their ELBO after 200 iterations using standardized drug response data 54
9 *Rescaled* fitted values of drug ten for five *training* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses and corresponding ranking 54
10 Fitted values of drug ten for five *training* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses and corresponding ranking 55
11 *Rescaled* predictions of drug ten for five *test* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses 56
12 Predictions of drug ten for five *test* cell lines given by models trained on standardized drug responses 56
13 Maximal and minimal ELBO configuration after 200 iterations of grid search of hyperparameters for shape $\alpha \in \{10^{-10}, 1, 10\}$ and scale $\theta \in \{10^{-10}, 0.01, 1\}$ using non-standardized drug response data 57
14 Configurations of Table 8 with different Mean Squared Error criteria adding a model using training means as a reference 58
15 Configurations of Table 8 replacing the maximal model as given in Table 13 trained with *non-standardized* drug responses 58
16 Fitted values for five *training* cell lines given by models trained on *non-standardized* $-\log_{10} GI_{50}$ drug responses and corresponding ranking 59
17 Predictions of drug ten for five *test* cell lines given by model trained on *non-standardized* drug responses 59
18 Fitted values of drug ten for five *training* cell lines, their variance and standard deviation in ones model trained on standardized $-\log_{10} GI_{50}$ drug responses . 60
19 Relative changes between ten iterations in the ones model of Figure 5 60
20 Number of iterations for minimal MSEs of default, ones and random model . 62
21 wpc- index for different models trained on standardized drug responses 67
22 Number of iterations and wpc-index for minimal MSEs of default, ones and random model 67
23 Variational densities - full conditionals with updated parameters 79

24 Parameter updates of variational distributions 80
25 Terms of ELBO expressed as variational parameters of random variables . . . 81
26 Terms of ELBO expressed in expectations of random variables 82
27 Expectation of gamma distributed random variable τ in terms of the variational parameters $\alpha^{\star}_{\tau_t}$ and $\beta^{\star}_{\tau_t} = \frac{1}{\theta^{\star}_{\tau_t}}$ 83
28 Derivatives of variational parameters w.r.t hyperparameters α_τ and β_τ of the gamma distributed random variable τ_t in rate notation 83
29 Derivatives of variational parameters w.r.t hyperparameters α_τ and θ_τ of the gamma distributed random variable τ_t in scale notation 84
30 Three combinations of hyperparameters obtaining minimal MSE on training and test data 85
31 Expected kernel weights and their precisions for models of Table 8 86
32 Expected intermediate outputs of drug ten for five *test* cell lines - default model 87
33 Expected intermediate outputs of drug ten for five *test* cell lines - ones model 87
34 Expected intermediate outputs of drug ten for five *test* cell lines of drug ten - random model 87
35 Summary statistics expected intermediate outputs of drug ten - default model 88
36 Summary statistics expected intermediate outputs of drug ten - ones model . 88
37 Summary statistics expected intermediate outputs of drug ten - random model 88
38 Expected biases and their precision for fitted values or predicitions for drug one to ten for five models of grid search trained on standardized drug responses 89
39 Drug number in data and its name. Drugs number 12, 26 and 27 have been excluded in Costello et al. (2014) 90
40 Summary statistics of $-\log_{10} GI_{50}$ drug responses for each drug for all observations (training and testing data). Costello et al. (2014) excluded drug 12, 26 and 27. 91
41 Summary statistics of $-\log_{10} GI_{50}$ drug responses for each drug for training data. Drugs 5, 24 and 26 have nearly no variation 92

1 Introduction

Cancer is a group of diseases caused by erratic, heterogeneous transformations in initially one human cell leading to uncontrolled growth. Cancers' heterogeneous nature makes it hard to predict drug response and to choose an appropriate treatment. Nowadays the human genome can be completely sequenced faster and at decreasing cost. This opens up the possibility to have gene information capturing the erratic, heterogeneous transformations available for clinical research. This thesis discusses one modeling approach that relates several high dimensional gene information of cancer cells to drug effectiveness.

Personalized medical approaches are envisioned to treat heterogeneous illnesses such as cancer. Cancer is classified by groups of altered cells. Cells related to cancer are called abnormal or malignant due to their malfunction leading to death. Since cancer is a non-deterministic malfunctioning of these cells, the aim is to determine the order of effectiveness of drugs given an individual incidence of cancer. Although cancer is heterogeneous in nature, it is classified by the functional type of a cell. A cancer class is divided further into subtypes. Many nowadays commonly applied therapies have been mainly established through trial and error, combining promising drugs in human trails. A detailed and comprehensive overview is provided by Mukherjee (2011).

Breast cancer has been divided into different subtypes (Daemen et al., 2013, 2015). Not every drug is effective for each cancer incidence due to specific genetic dispositions. Selecting drugs for trials on humans could thus be designed using predictions of drug effectiveness based on the human genome. Taking breast cancer as a common sort of cancer, Costello et al. (2014) engaged in a community study to determine the best models for ranking drugs according to their effectiveness of inhibiting breast cancer cell growth using different genomic, epigenomic and proteomic information of human breast cells. The most successful approach ranking cell lines according to their drug response was a *Bayesian multiple task multiple kernel learning* (BMTMKL) model. BMTMKL combines different datasets (kernels or views) and several outcome variables (tasks), which are measures of drug efficacy. It is Bayesian as it is based on a directed acyclical graph (DAG) model specifying distributional assumptions on all variables.

This thesis is a contribution to the Working Package 5 named *Model-Based Prediction of Drug Responsiveness* of the comPREhensive Data Integration for Cancer Treatment (PREDICT) project aiming at predicting the efficacy of drugs in cancer treatment on genomic information. The BMTMKL model, yielding the most accurate predictions for ranking drug

effectiveness in Costello et al. (2014), has been chosen as reference model to be discussed. This thesis reproduces the ranking obtained in the community study (ibid.). Since an overall ranking measure over all drugs has been used in Costello et al. (2014), additionally BMTMKL's ability to yield good fits and predictions in absolute drug response values is assessed.

The core code of the BMTMKL has been published on GitHub by Mehmet Gönen in R and Matlab. In terms of documentation, up to now there exists only a brief summary of the distributional assumptions of BMTMKL in the Supplementary Note 1 (Costello et al., 2014, S1, p. 3ff.) and the code itself. The model trained in Costello et al. (2014) has similarities to the models described in Gönen (2012a) or Gönen (2012b). However, understanding or assessing the code provided is infeasible without a detailed derivation of the implemented equations, which is not published according to my knowledge. Therefore, detailed mathematical descriptions of the used DAG model and the inference method *variational inference* (VI) for parameter optimization will be provided in Section 2 (Bishop, 2006; Blei et al., 2017; Hoffman et al., 2013; Murphy, 2012), leading to all implemented equations by Mehmet Gönen[1]. After deriving and understanding the model, extending the algorithm by a search on some previously fixed parameters, denoted as hyperparamters, is discussed in Section 2.6.2, as they have not been provided specifically for the challenge.

In Section 3 the used data is described. As preprocessing and selecting appropriate subsamples of the data is at least as important as the training procedure, the drug sensitivity measure and the available data including its transformations is analyzed in detail. For comparison, a complete reference dataset from the team around Mr. Gönen was solicited, but could not be obtained. However, Ammad-ud-din Muhammad provided many insights to the challenge data, including scripts.

In Section 4 a grid search on hyperparameters and in depth analysis of performance considering different hyperparameters is presented. The evaluation method used by Costello et al. (2014) is analyzed and compared to different Mean Squared Erros (MSE). As often, the choice of a model depends on the criteria chosen to evaluate it. Section 5 concludes.

[1]With one small caveat: The gamma distribution in the code and in the paper are differently parametrized. Here the rate notation was used, whereas Mehmet Gönen uses the scale notation.

2 Model and Inference

"The primary role of the latent variables is to allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler (typically exponential family) conditional distributions."(Bishop, 2006, p. 366)

2.1 The model

The model presented in this thesis is called BMTMKL, which is stressing the combination of several outcomes, denoted tasks, and multiple inputs, which are kernel matrices containing information of some kind of similarities in specific feature spaces between single observations. The inputs are named views and are shared between all tasks. Adding further hidden variables is a key to enable modeling the conditional distributions of outcomes given these inputs. The type of data this model can handle and the kernels constructed from it are described in Section 3. In this Subsection the model and necessary distributions for computing updates in a next step are presented.

Kernels, the inputs, are indexed with subscript $k_{1:P}$[2] for kernels, whereas the drug data, the tasks or outcomes, will be indexed with subscript $t_{1:T}$. The interest is in predicting the distribution over the drug susceptibility, i.e. drug effectiveness, y_t. There are N_t measures of susceptibility for each drug t, which will be indexed with subscript $i, y_{t,i}$, where $i = 1, \ldots, N_t$: $y_t = y'_{t,1:N_t} = (y_{t,1}, \ldots, y_{t,N_t})'$. The expectation of a single $y_{t,i}$ is a linear combination of P real-valued intermediate variables[3], also indexed with $k_{1:P}$ and denoted $g_{t,i,k}$, and a real-valued coefficient e_k as well as an overall real-valued intercept for drug t, called bias, b_t: $\mathrm{E}\left[y_{t,i}|g_{t,i,1:P}, e_{1:P}, b_t\right] = \sum_{k=1}^{P} g_{t,i,k}e_k + b_t$. It can therefore be compared to a basic linear regression: $\mathrm{E}\left[y_t|G_t, e, b_t\right] = G_t \cdot e + \mathbb{1}_{N_t} b_t = (\mathbb{1}_{N_t}, G_t) \cdot (b_t, e')'$, where $\mathbb{1}_{N_t}$ is a column vector of N_t ones and $G_t = (g_{t,1}, \ldots, g_{t,P})$ a $(N_T \times P)$ matrix of intermediate output variables computed from each input k, and $e' = (e_1, \ldots, e_P)$. These regressions are task specific as the number of treated cell lines for each drug differ and a task specific intercept b_t is specified. It is thus possible to treat the model as a pooled regression as the kernel weights or kernel coefficients e are shared between drugs, and specifying for each drug its own intercept b_t.

The intermediate output vector $g_{t,k}$ of dimension $(N_t \times 1)$ can be interpreted as a variable originating from an input kernel matrix $K_{t,k} \in \mathbb{R}^{N_t \times N_t}$ by summing single input kernels, i.e. specific similarity measures, for a cell line with all other cell lines with the task-specific cell

[2] $k_{1:P} \Leftrightarrow k = 1, \ldots, P$, $t_{1:T} \Leftrightarrow t = 1, \ldots, T$ and $i_{1:N_t} \Leftrightarrow i = 1, \ldots, N_t$

[3] Assume for simplicity for the moment that all variables are observed.

Table 1: Unobserved real valued random variables of the model and their representations

factor	single r.v.	real valued random variables in several vector or matrix notations
λ	$\lambda_{t,i}$	$\lambda_t = (\lambda_{t,1}, \ldots, \lambda_{t,N_t})'_{(N_t \times 1)}$, $\lambda = (\lambda'_1, \ldots, \lambda'_T)'_{(N \times 1)}$, $N = \sum_{t=1}^{T} N_t$
a	$a_{t,i}$	$a_t = (a_{t,1}, \ldots, a_{t,N_t})'_{(N_t \times 1)}$, $a = (a'_1, \ldots, a'_T)'_{(N \times 1)}$, $N = \sum_{t=1}^{T} N_t$
G	$g_{t,k,i}$	$g_{t,k} = (g_{t,k,1}, \ldots, g_{t,k,N_t})'_{(N_t \times 1)}$, $G_t = (g_{t,1}, \ldots, g_{t,P})_{(N_T \times P)}$,
		$G = (G'_1, \ldots, G'_T)'_{(N \times P)}$ $g_t = (g'_{t,1}, \ldots, g'_{t,P})'_{(N_t \cdot P \times 1)}$,
		$G'_{t,i} = (g_{t,1,i}, \ldots, g_{t,P,i})'_{(P \times 1)}$, $G'_t G_t = \left(\sum_{i=1}^{N_t} G'_{t,i} G_{t,i}\right)_{(P \times P)}$,
		$\mathrm{tr}[G'_t G_t] = \sum_{i=1}^{N_t} G_{t,i} G'_{t,i} = \sum_{k=1}^{P} g'_{t,k} g_{t,k}$
υ	υ_t	$\upsilon = (\upsilon_1, \ldots, \upsilon_T)'_{(T \times 1)}$
γ	γ_t	$\gamma = (\gamma_1, \ldots, \gamma_T)'_{(T \times 1)}$
ω	ω_k	$\omega = (\omega_1, \ldots, \omega_P)'_{(P \times 1)}$, $\Omega = \mathrm{diag}(\omega)_{(P \times P)}$
b, e	b_t, e_k	$b = (b_1, \ldots, b_T)'_{(T \times 1)}$, $e = (e_1, \ldots, e_P)'_{(P \times 1)}$,
		$(b', e')' = (b_1, \ldots, b_t, e_1, \ldots, e_P)'_{(T+P \times 1)}$
ϵ	ϵ_t	$\epsilon = (\epsilon_1, \ldots, \epsilon_T)'_{(T \times 1)}$, $\mathrm{diag}(\epsilon)_{(T \times T)}$

line weights a_t of dimension $(N_t \times 1)$ for all input kernels. A single intermediate output for a specific cell line j is $g_{t,j,k} = \sum_{i=1}^{N_t} a_{t,i} \cdot k_k(j,i)$, where $k_k(j,i)$ is the corresponding entry of the kernel matrix. Due to the symmetry of the kernel matrix it holds that $k_k(j,i) = k_k(i,j)$.

The remaining variables $\lambda_t, \omega_k, e_k, \upsilon_t, \epsilon_t$ and γ_t (all $\in \mathbb{R}$) are the precisions, i.e. inverse covariances, for normal distributions governing the variables $a_t, e_k, g_{t,k}, y_t$ and b_t. All variables are specified as random variables in Bayesian statistics and most of them are not observed. A more precise and detailed link will be given in the following Section 2.2 when deriving the variational distribution.

In Table 1 all random variables and their different representations are listed. The observed data are P kernel matrices $K_{t,k} \in \mathbb{R}^{N_t \times N_t}$ and an outcome vector $y_t \in \mathbb{R}^{N_t}$ for each drug t (see Table 2). Different representations are needed in order to be able to express distributions of outcomes and random variables. Distributions, or more precisely, their corresponding densities of sets of random variables are stated in terms of vectors. For example, each $\lambda_{t,i}$ is independent of all others. Therefore the distribution of all $\lambda_{t,i} \in \{\lambda_{1:T,1:N_t}\}$ can be written in general notation as $p(\{\lambda_{1:T,1:N_t}\}) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} p(\lambda_{t,i})$, which is set for convenience equal to the vector representation of all $(\lambda_{1:T,1:N_t})' = \lambda$, which is $p(\lambda) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} p(\lambda_{t,i})$.

The set of latent random variables is defined as $\mathcal{Z} = \{\lambda, a, G, \upsilon, \gamma, \omega, b, e, \epsilon\}$, the input kernels for one drug t as $\mathcal{X}_t = \{K_{t,1}, \ldots, K_{t,P}\} = \{K_{t,k}\}_{k=1}^{P}$ with corresponding outcome

Table 2: Observed deterministic and real valued random variables and their representations

inital rep.	other summaries	transformations
$K_{t,k}$ $\in \mathbb{R}^{N_t \times N_t}$	$K_{t,k,i}$ is the row i of kernel matrix $K_{t,k}$, $K_{t,i} = (K_{t,1,i}, \ldots, K_{t,P,i})' \in \mathbb{R}^{P \times N_t}$ contains all rows i of all deterministic input matrices k, $K_t = (K_{t,1}, \ldots, K_{t,P})' \in \mathbb{R}^{P \cdot N_t \times N_t}$ is the stacked matrix of $K_{t,1:P}$	$\sum_{i=1}^{N_t} K'_{t,i} K_{t,i} = \sum_{k=1}^{P} K_{t,k} K_{t,k}$ $\sum_{k=1}^{P} g'_{t,k} K_{t,k} a_t = a'_t K_t g_t$, where g_t is the stacked vector of intermediate outcomes (see Table 1)
$y_t \in \mathbb{R}^{N_t \times 1}$	$y = (y'_1, \ldots, y'_T)' \in \mathbb{R}^{N \times 1}$ is a vector of stacked outcomes, where $\sum_{t=1}^{T} N_t = N$	

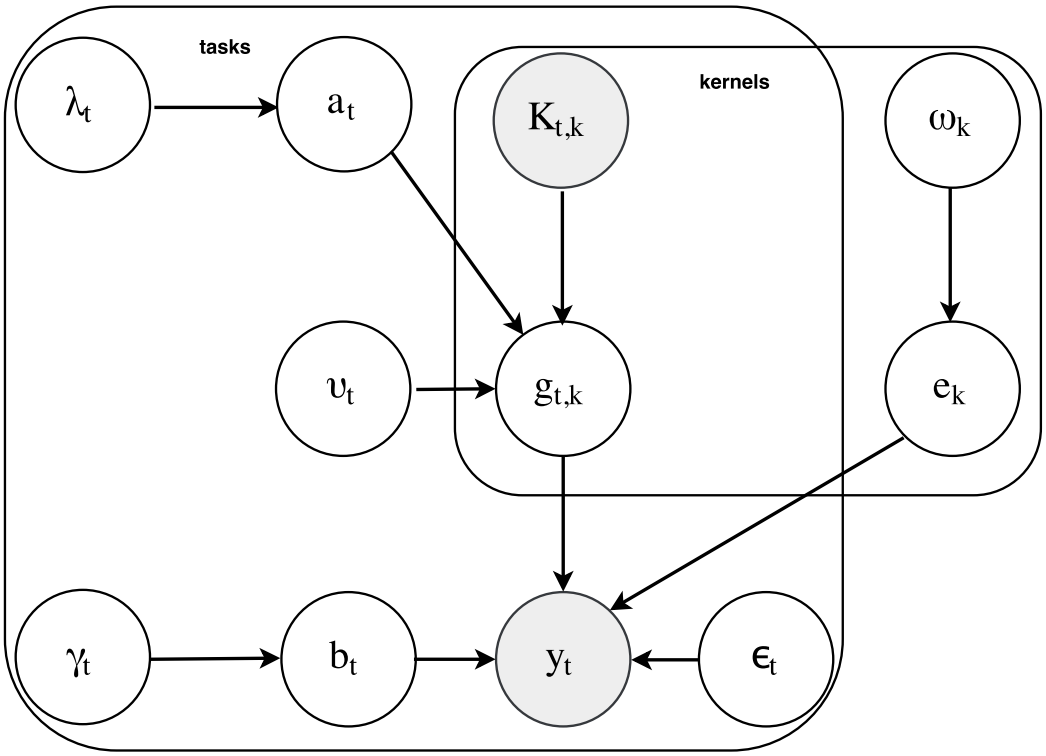
vector y_t, and for all T drugs as $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_T\}$ and $\mathcal{Y} = \{y_1, \ldots, y_T\} = \{y_t\}_{t=1}^T$. Although the kernels are shared between the drugs, only the cell lines for which the drug t is tested are used as task specific kernel matrices. The last two sets define the observed data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

The joint density of the model is therefore defined as a product of all (hidden) random variables conditional on the fixed inputs $\mathcal{X}$

$$
\begin{aligned}
p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) =& p(\lambda, a, G, \upsilon, \gamma, \omega, b, e, \epsilon, \mathcal{Y}|\mathcal{X}) \\
=& p(\lambda) p(a|\lambda) p(G|a, \upsilon, \mathcal{X}) p(\upsilon) p(\omega) p(e|\omega) p(b|\gamma) p(\gamma) p(\epsilon) p(\mathcal{Y}|b, e, G, \epsilon) \\
=& \prod_{k=1}^{P} p(\omega_k) p(e_k|\omega_k) \prod_{t=1}^{T} p(\lambda_t) p(a_t|\lambda_t) p(\upsilon_t) \left(\prod_{k=1}^{P} p(g_{t,k}|K_{t,k}, a_t, \upsilon_t) \right) \\
& \cdot p(\gamma_t) p(b_t|\gamma_t) p(\epsilon_t) p(y_t|e, G_t, b_t, \epsilon_t) \\
=& \prod_{k=1}^{P} p(\omega_k) p(e_k|\omega_k) \prod_{t=1}^{T} \left(\prod_{i=1}^{N_t} p(\lambda_{t,i}) p(a_{t,i}|\lambda_t) \right) p(\upsilon_t) \left(\prod_{k=1}^{P} \prod_{i=1}^{N_t} p(g_{t,k,i}|K_{t,k,i}, a_t, \upsilon_t) \right) \\
& \cdot p(\gamma_t) p(b_t|\gamma_t) p(\epsilon_t) p(y_t|e, G_t, b_t, \epsilon_t) \quad ,
\end{aligned}
\tag{1}
$$

where the dependence of random variables on different random variables is specified. It can be seen that each factor, i.e. set of latent variables, does only depend on its so called parent factors, c.p. Bishop (2006, ch. 8) or Wainwright and Jordan (2008, ch. 2.2). This assumption of conditional independence of a random variable given only the random variables pointing at it, can be easily checked comparing the graphical model representation in Figure 1 with the above joint distribution in equation (1). In Figure 1 the random variables are assumed

Figure 1: Graphical model representation



to have the following distributions:

$$\begin{aligned}
&\lambda_{t,i} \sim \mathcal{G}am(\alpha_\lambda, \beta_\lambda) \quad a_{t,i}|\lambda_{t,i} \sim \mathcal{N}(0, \lambda_{t,i}^{-1}) \qquad g_{t,k}|K_{t,k}, a_t, \upsilon_t \sim \mathcal{N}(K_{t,k}a_t, \upsilon_t^{-1} I_{N_t}) \\
&\upsilon_t \sim \mathcal{G}am(\alpha_\upsilon, \beta_\upsilon) \qquad \omega_k \sim \mathcal{G}am(\alpha_\omega, \beta_\omega) \qquad e_k|\omega_k \sim \mathcal{N}(0, \omega_k^{-1}) \\
&\gamma_t \sim \mathcal{G}am(\alpha_\gamma, \beta_\gamma) \qquad b_t|\gamma_t \sim \mathcal{N}(0, \gamma_t^{-1}) \qquad y_t|e, G_t, b_t, \epsilon_t \sim \mathcal{N}\left(G_t e + b_t \cdot \mathbb{1}_{N_t},\, \epsilon_t^{-1} I_{N_t}\right) \\
&\epsilon_t \sim \mathcal{G}am(\alpha_\epsilon, \beta_\epsilon)
\end{aligned} \tag{2}$$

The notation above gives a random variable with its distribution governed by its parameters, e.g. $\gamma_t \sim \mathcal{G}am(\alpha_\gamma, \beta_\gamma)$, where the random variables γ_t is gamma distributed with shape parameter α_λ and rate parameter β_λ. Its probability density function is in this field of research defined as $p(\gamma) = \mathcal{G}am(\gamma_t|\alpha_\gamma, \beta_\gamma)$, where a specific realization is denoted in the exact same way as the random variable itself. Equally, in the case of a normally distributed variable governed by its mean and variance, e.g. $b_t|\gamma_t^{-1} \sim \mathcal{N}(0, \gamma_t^{-1})$, the probability density function is referred to as $p(b_t|0, \gamma_t^{-1}) = \mathcal{N}(b_t\ |0, \gamma_t^{-1})$, where again a specific realization is denoted in the exact same way as the random variable itself. Note that I use as a distinction between dependence of a density on other random variables and dependence on parameters, including other random variables, the notation of $p(\cdot|\cdot)$, and for the two specific densities

$\mathcal{N}(\cdot|\cdot)$ and $\mathcal{G}am(\cdot|\cdot)$. The two densities of random variables used are described in Section A.1. The joint can then be given as

$$
\begin{aligned}
p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) = & \left(\prod_{k=1}^{P} \mathcal{G}am(\omega_k \mid \alpha_\omega, \beta_\omega) \mathcal{N}(e_k \mid 0, \omega_k^{-1}) \right) \prod_{t=1}^{T} \left(\prod_{i=1}^{N} \mathcal{G}am(\lambda_{t,i} \mid \alpha_\lambda, \beta_\lambda) \mathcal{N}(a_{t,i} \mid 0, \lambda_{t,i}^{-1}) \right) \\
& \cdot \mathcal{G}am(\upsilon_t \mid \alpha_\upsilon, \beta_\upsilon) \left(\prod_{k=1}^{P} \mathcal{N}(g_{t,k} | K_{t,k} a_t, \upsilon_t^{-1} I_{N_t}) \right) \mathcal{G}am(\gamma_t \mid \alpha_\gamma, \beta_\gamma) \mathcal{N}(b_t \mid 0, \gamma_t^{-1}) \\
& \cdot \mathcal{G}am(\epsilon_t \mid \alpha_\epsilon, \beta_\epsilon) \mathcal{N} \left(y_t \mid \sum_{k=1}^{P} e_k g_{t,k} + b_t \cdot \mathbb{1}_{N_t}, \epsilon_t^{-1} I_{N_t} \right)
\end{aligned}
$$

The gamma distributions have so called hyperparameters which are assumed to be fixed. The set of fixed hyperparameters is $\Theta = \{\alpha_\lambda, \beta_\lambda, \alpha_\upsilon, \beta_\upsilon, \alpha_\omega, \beta_\omega, \alpha_\gamma, \beta_\gamma, \alpha_\epsilon, \beta_\epsilon\}$. Setting the mean to zero for normal distributions is possible without loss of generality as it is corresponding to a linear transformation, implying that it does not depend on yet another hidden random variable. The mean of the weights for input kernel a_t, the bias b_t and the kernel coefficients e_k for each input kernel k are set to zero. All other parameters of distributions are random variables specified in the model.

In order to optimize the model one would, as a classic Bayesian, compute the posterior $p(\mathcal{Z}|\mathcal{Y}, \mathcal{X})$, which is the density of the distribution of all random variables given the data. As this involves marginalization over all latent variables in order to get the normalizing evidence, it is intractable. See Wainwright and Jordan (2008, ch. 2.3, ch. 5.1) for examples and details. A common approach is instead to formulate a variational approximation $q(\mathcal{Z}) \approxeq p(\mathcal{Z}|\mathcal{Y}, \mathcal{X})$, which factorizes into distributions for groups of latent variables. For the model in Costello et al. (2014) by Gönen, no explicit approximation for the posterior was mentioned. However the linked code and his previous papers (Gönen, 2012b; Gönen, 2012a) suggest the following approximated posterior distribution:

$$
q(\mathcal{Z}) = q(\lambda) q(a) q(G) q(\upsilon) q(\gamma) q(\omega) q(b, e) q(\epsilon) \tag{3}
$$

Except for the joint approximation of the biases and kernel coefficients, a classic mean field approach is performed, i.e. approximating each random variable independent of all other random variables, where some random variables are by construction of the data generating process univariate or multivariate. Each $q(\tau)$ for some (set of) random variables τ in the approximated posterior is sometimes referred to as factor - as these random variables are also grouped in the directed acyclic graph (DAG) of the model in Figure 1. The approximated posterior factorizes into the densities of the grouped random variables in $q(\mathcal{Z})$.

Having stated the assumed joint and the assumed approximated posterior, in the next Section the so called full conditionals for the factors or group of random variables are defined and derived. These are required to determine the optimal solution and updates for the factor densities $q(z), z \in \mathcal{Z}$, which constitute the approximate posterior. Deriving these before describing optimal solutions seems appropriate as they introduce the model in detail. In Gönen (2012a) the full conditionals are only stated without presenting important transformations and groupings of the random variables. In Costello et al. (2014, S1) only distributional assumptions of the prior are stated in their most basic form given in equations (2) along the published code. Beginning by deriving the full conditionals can further be justified as these serve also as a starting point for a Markov Chain Monte Carlo approach using Gibbs sampling (Bishop, 2006, ch. 11), which could also be applied as an alternative to variational inference.

Following the full conditionals, the objective for VI is derived in general in Subsection 2.3 before it is being stated for the chosen model in Subsection 2.4. This is followed in Subsection 2.5 by the derivation of the general updates on the basis of the previously introduced objective, before the specific updates in the model are calculated in Subsection 2.6. Finally, calculating predictions for new cell lines or fitted values for training cell lines given the learned distributional parameters of the described model is presented in Subsection 2.7. A brief introduction to the two continuous random variables used in BMTMKL, their densities and some of their properties or representations can be found in the appendix A.1.

2.2 Full conditional densities of latent variables

In the following each conditional distribution of a set of latent variables given all other latent variables and the data is derived. Distributions will be represented by their densities. The conditional densities will be denoted as *full conditional*, where the full prefix denotes the fact of conditioning one latent variable or a set of independent latent variables z_j on all other latent variables and the data $\{\mathcal{D}, \mathcal{Z}_{-z_j}\} = \{\mathcal{D}, \mathcal{Z} \setminus z_j\}$ (Blei et al., 2017, p. 863). The full conditional is also referred to as complete conditional. In the model presented only normal or gamma distributions are specified for random variables, of which all except the $\{y_t\}_{t=1}^T$ are hidden random variables. Some univariate distributions need to be reformulated into multivariate form. In a fully conjugated model a conditional distribution of a latent variable given all other latent variables and the data, i.e. the full or complete conditional, is again of the same kind as the prior distribution. It is thus the prior or model distribution of a latent variable that defines the distribution of its full conditional: A random variable with

gamma prior, i.e. distribution in the model, has a full conditional which is again a density of a gamma distributed random variable. The full conditional of a normally distributed random variable is again a density of a normally distributed random variable.

The assumed distributions for the latent variables in the model are introduced as they are needed for deriving the full conditionals of a latent variable by writing down their corresponding densities. Each derivation of a full conditional thus starts with referencing or stating the priors and is, if needed, followed by some comments on its form. See Table 1 for the list of hidden random variables and their representations. Note that Subsection 2.2 deals only with (conditional) prior distributions.

2.2.1 Full conditional for the precision λ on kernel weights

First, the density of the precision $\lambda_{t,i}$ on the weight [4] $a_{t,i}$ for the kernels and of the density of one input weight given this precision $\lambda_{t,i}$ are given. As all precisions are independent, the density of all $\lambda_{1:T,1:N_t} = \lambda$ is just the product of the single densities:

$$p(\lambda_{t,i}) = p(\lambda_{t,i}|\alpha_\lambda, \beta_\lambda) = \mathcal{G}am(\lambda_{t,i}\,|\alpha_\lambda, \beta_\lambda) \qquad \Rightarrow \quad p(\lambda) = \prod_{t=1}^{T}\prod_{i=1}^{N_t} \mathcal{G}am(\lambda_{t,i}\,|\alpha_\lambda, \beta_\lambda)$$
$$= \frac{1}{\Gamma(\alpha_\lambda)} \beta_\lambda^{\alpha_\lambda} \lambda_{t,i}^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda_{t,i}) \tag{4}$$

The same holds for the densities of $a_{t,i}$ summarized in $a = a_{1:T,1:N_t}$, whose joint prior density is as well the product of each single density.

$$p(a_{t,i}\,|\lambda_{t,i}) = \mathcal{N}(a_{t,i}\,|0, \lambda_{t,i}^{-1}) \quad \Rightarrow \quad p(a|\lambda) = \prod_{t=1}^{T}\prod_{i=1}^{N_t} \mathcal{N}(a_{t,i}\,|0, \lambda_{t,i}^{-1}) = \mathcal{N}(a|0, diag(\lambda)^{-1})$$
$$= \left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda_{t,i}}{2}\, a_{t,i}^2\right\} \tag{5}$$

The latent variable $\lambda_{t,i}$ is the precision for the weight $a_{t,i}$. Holding all other latent variables fixed, the full conditional depends on the two densities of latent variables $p(\lambda_{t,i})$ and $p(a_{t,i}|\lambda_{t,i})$, where in the ladder the random variable $\lambda_{t,i}$ is a parameter, namely the precision, of the random variable $a_{t,i}$. This is due to the conditional independence assumption of DAG models. Looking at the graph in Figure 1, the full conditional of a random variable depends on its children and itself. If one evaluates the full conditional for $\lambda_{t,i}$ given our joint in the nominator, one therefore only has to consider two distributions and evaluate the rest as fixed. For information on distributional proportionality w.r.t to a constant c, see appendix A.1.3. This conditioning w.r.t to all other latent variables except $\lambda_{t,i}$ is described using the set

[4] Many would say that weights have to sum to one. The factors $a_{t,i}$ do not have this restriction

$\mathcal{Z}_{-\lambda_{t,i}} = \{\mathcal{Z} \setminus \lambda_{t,i}\}$, which contains all latent variables except $\lambda_{t,i}$. One again obtains a conditional density of a gamma distributed random variable which will be denoted, as previously described, either *full conditional* or *complete conditional*:

$$\begin{aligned} p(\lambda_{t,i}|\mathcal{D}, \mathcal{Z}_{-\lambda_{t,i}}) &= \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{\int_{\lambda_{t,i}} p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) d\lambda_{t,i}} = c_1 \cdot p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) && \propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \\ &= c \cdot p(\lambda_{t,i}) p(a_{t,i}|\lambda_{t,i}) && \propto p(\lambda_{t,i}) p(a_{t,i}|\lambda_{t,i}) \end{aligned}$$

The integral $c_1 = \int_{\lambda_{t,i}} p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) d\lambda_{t,i} = p(\mathcal{Z}_{-z_j}, \mathcal{Y}|\mathcal{X})$ in the denominator of the first fraction is a density function without the random variable $\lambda_{t,i}$ and thus constant with respect to $\lambda_{t,i}$.

$$\begin{aligned} p(\lambda_{t,i}|\mathcal{D}, \mathcal{Z}_{-\lambda_{t,i}}) &\propto \frac{1}{\Gamma(\alpha_\lambda)} \beta_\lambda^{\alpha_\lambda} \lambda_{t,i}^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda_{t,i}) \cdot \left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda_{t,i}}{2}\, a_{t,i}^2\right\} \\ &\propto \lambda_{t,i}^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda_{t,i}) \cdot (\lambda_{t,i})^{\frac{1}{2}} \exp\left\{-\frac{\lambda_{t,i}}{2}\, a_{t,i}^2\right\} \\ &= \lambda_{t,i}^{\alpha_\lambda + \frac{1}{2} - 1} \cdot \exp\left\{-\left(\beta_\lambda + \frac{1}{2}\, a_{t,i}^2\right) \lambda_{t,i}\right\} \\ &\propto \mathcal{G}am\left(\lambda_{t,i} \middle|\, \alpha_\lambda + \frac{1}{2},\ \beta_\lambda + \frac{1}{2} a_{t,i}^2\right) \end{aligned}$$

The full conditional of the hidden random variable $\lambda_{t,i}$ is proportional to a gamma density with the scale parameter $\alpha(\lambda_{t,i}) = \alpha_{\lambda_{t,i}} = \alpha_\lambda + \frac{1}{2}$ and the rate parameter $\beta(\lambda_{t,i}) = \beta_{\lambda_{t,i}} = \beta_\lambda + \frac{1}{2} a_{t,i}^2$. The scale parameter is a constant for all $\lambda_{t,i}$ as it only involves fixed scalar values and no other random variables. Proportionality w.r.t. to a density of well defined distribution allows to state easily the normalizing constant as a function of the parameters defining the distribution. The normalization constant is specified in terms of shape and rate parameters as $\frac{1}{\Gamma(\alpha_{\lambda_{t,i}})} \beta_{\lambda_{t,i}}^{\alpha_{\lambda_{t,i}}}$. Would the gamma prior of $\lambda_{t,i}$ have been specified in scale notation instead of rate notation, i.e $\beta_\lambda = \frac{1}{\theta_\lambda}$, the full conditional in scale notation would be $p(\lambda_{t,i}|\mathcal{D}, \mathcal{Z}_{-\lambda_{t,i}}) = \mathcal{G}am\left(\lambda_{t,i}|\alpha_\lambda + \frac{1}{2}, \left(\frac{1}{\theta_\lambda} + \frac{1}{2} a_{t,i}^2\right)^{-1}\right)$ (c.p. appendix A.1.1). Gönen (2012a) uses the scale notation for gamma distributed variables and it will reappear in Section 2.6.2.

Since all $\lambda_{t,i}$ are assumed to be independent, the set of random variables $\{\lambda_{t,i}\}_{1:T,1:N_t}$ or equivalently the vector λ containing all $\lambda_{t,i}$ have as full conditional

$$p(\lambda|\mathcal{D}, \mathcal{Z}_{-\lambda}) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} \mathcal{G}am\left(\lambda_{t,i} \middle|\, \alpha_\lambda + \frac{1}{2},\ \beta_\lambda + \frac{1}{2} a_{t,i}^2\right) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} \mathcal{G}am\left(\lambda_{t,i}|\ \alpha_{\lambda_{t,i}},\ \beta_{\lambda_{t,i}}\right).$$

2.2.2 Full conditional for kernel weights a

The next latent variables of our data generating model are the normally distributed intermediate outputs $g_{t,k} \in \mathbb{R}^{N_t}$, where $v_t \in \mathbb{R}$ is their precision and $a_t \in \mathbb{R}^{N_t}$ the weights for

combining the P input kernels $K_{t,k} \in \mathbb{R}^{N_t \times N_t}\ \forall\ k = 1, \ldots, P$ to their mean. One single $g_{t,k,i}$ is the intermediate output for kernel k for cell line i. One intermediate output is therefore governed by a combination of an observed kernel matrix combined with task specific weights and the corresponding task specific precision.

All N_t cell line kernel values in the views are weighted for task t in the same way to obtain the mean of the intermediate outputs $g_{t,k}$ ($\forall k = 1, \ldots, P$). The random variable $g_{t,k} \in \mathbb{R}^{N_t}$ for each view or input kernel k has a distribution for a each cell line i, denoted, $g_{t,k,i}$, independent of all other cell lines and input kernels.

$$
\begin{aligned}
p(g_{t,k}|K_{t,k}, a_t, \upsilon_t) &= \mathcal{N}(g_{t,k}\ |K_{t,k}a_t, \upsilon_t^{-1} I_{N_t}) && (6)\\
&= \frac{(\det|\upsilon_t I_{N_t}|)^{1/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}(g_{t,k} - K_{t,k}a_t)'\upsilon_t I_{N_t}(g_{t,k} - K_{t,k}a_t)\right\} \\
&= \frac{(\upsilon_t)^{N_t/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\upsilon_t(g_{t,k} - K_{t,k}a_t)'(g_{t,k} - K_{t,k}a_t)\right\} && (7)\\
&= \frac{(\upsilon_t)^{N_t/2}}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\upsilon_t \sum_{i=1}^{N_t}(g_{t,k,i} - K_{t,k,i}a_t)^2\right\} \quad , i = 1, \ldots, N_t
\end{aligned}
$$

All intermediate outputs, which are random variables, can be combined in the matrix $G_t = \{g_{t,1}, \ldots, g_{t,P}\}$ of dimension $N_t \times P$.

$$
\begin{aligned}
\Rightarrow \quad p(G_t|\mathcal{X}_t, a_t, \upsilon_t) &= \prod_{k=1}^{P} \mathcal{N}(g_{t,k}\ |K_{t,k}a_t, \upsilon_t^{-1} I_{N_t}) && (8)\\
&= \left(\frac{(\upsilon_t)^{N/2}}{(2\pi)^{D/2}}\right)^P \exp\left\{-\frac{1}{2}\upsilon_t \sum_{k=1}^{P}\sum_{i=1}^{N_t}(g_{t,k,i} - K_{t,k,i}a_t)^2\right\} \\
&= \frac{(\upsilon_t)^{P\cdot N/2}}{(2\pi)^{P\cdot D/2}} \exp\left\{-\frac{\upsilon_t}{2} \sum_{k=1}^{P}(g_{t,k} - K_{t,k}a_t)'(g_{t,k} - K_{t,k}a_t)\right\} && (9)
\end{aligned}
$$

The prior distribution of the weight vector a_t for task t which is used in the distribution of intermediate outputs G_t is a multivariate normal with mean zero and a diagonal precision matrix $\Lambda_t = \text{diag}\{\lambda_{t,1}, \ldots, \lambda_{N_t}\}$. It can be given by jointly describing the N_t independent densities $p(a_{t,i}|\lambda_{t,i})$ given in equation (5):

$$
\begin{aligned}
p(a_t) = \prod_{i=1}^{N_t} \mathcal{N}(a_{t,i}\ |0, \lambda_{t,i}^{-1}) &= \left[\prod_{i=1}^{N_t}\left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}}\right] \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{N_t}\lambda_{t,i}a_{t,i}^2\right\} \\
&= \left[\prod_{i=1}^{N_t}\left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}}\right] \cdot \exp\left\{-\frac{1}{2}a_t'\Lambda_t a_t\right\}, \text{where } \Lambda_t = \text{diag}(\lambda_{t,1}, \ldots, \lambda_{t,N_t}) \qquad (10)
\end{aligned}
$$

The multivariate full conditional for weights a_t depends on all distributions dependent on a_t,

namely $p(a_t|\lambda_t)$ and $p(G_t|a_t, \upsilon_t)$, and is therefore task-specific:

$$
\begin{aligned}
p(a_t|\mathcal{D}, \mathcal{Z}_{-a_t}) &= \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{\int_{a_t} p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) da_t} = \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{p(\mathcal{Z}_{-a_t}, \mathcal{Y}|\mathcal{X})} \propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(a_t|\lambda_t) \cdot p(G_t|a_t, K_t, \upsilon_t^{-1}) \\
&= \prod_{i=1}^{N} p(a_{t,i}|\lambda_{t,i}) \cdot \prod_{k=1}^{P} p(g_{t,k}|a_t, K_{t,k}, \upsilon_t^{-1}) \\
&= \left[\prod_{i=1}^{N} \left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}}\right] \exp\left\{-\frac{1}{2} a_t' \Lambda_t a_t\right\} \cdot \frac{(v_t)^{P \cdot N/2}}{(2\pi)^{P \cdot D/2}} \exp\left\{-\frac{v_t}{2} \sum_{k=1}^{P} (g_{t,k} - K_{t,k} a_t)'(g_{t,k} - K_{t,k} a_t)\right\} \\
&\propto \exp\left\{-\frac{1}{2} a_t' \Lambda_t a_t\right\} \cdot \exp\left\{-\frac{v_t}{2} \sum_{k=1}^{P} (g_{t,k} - K_{t,k} a_t)'(g_{t,k} - K_{t,k} a_t)\right\} \\
&= \exp\left\{-\frac{1}{2} a_t' \Lambda_t a_t\right\} \cdot \exp\left\{-\frac{v_t}{2} \sum_{k=1}^{P} \left(g_{t,k}' g_{t,k} - g_{t,k}' K_{t,k} a_t - \underbrace{(K_{t,k} a_t)' g_{t,k}}_{g_{t,k}' K_{t,k} a_t} + (K_{t,k} a_t)' K_{t,k} a_t\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2} a_t' \Lambda_t a_t\right\} \cdot \exp\left\{-\frac{v_t}{2} \sum_{k=1}^{P} \left(-2 g_{t,k}' K_{t,k} a_t + a_t' K_{t,k} K_{t,k} a_t\right)\right\} \quad , K_{t,k} = K_{t,k}' \\
&= \exp\left\{-\frac{1}{2} a_t' \Lambda_t a_t + v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{v_t}{2} a_t' \left(\sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t\right\}
\end{aligned}
$$

The kernel matrices are symmetric, i.e. $K_{t,k} = K_{t,k}' \; \forall, k = 1, \ldots, P$. Next, the fact that a scalar can be written as the trace of a 1×1 matrix, as for example $a_t' \Lambda_t a_t = \text{tr}(a_t' \Lambda_t a_t) = \text{tr}(\Lambda_t a_t a_t')$, is used to be able to perform cyclic permutations within the trace operator of matrices and vectors. This is done to be able to formulate the full conditional in terms of the exponential family notation, which is one way to identify the parameters of interest in a normal distribution. See Section A.1.2 for details.

$$
\begin{aligned}
p(a_t|\mathcal{D}, \mathcal{Z}_{-a_t}) &\propto \exp\left\{-\frac{1}{2} \text{tr}[a_t' \Lambda_t a_t] + v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{v_t}{2} \text{tr}\left[a_t' \left(\sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t\right]\right\} \\
&= \exp\left\{-\frac{1}{2} \text{tr}[\Lambda_t a_t a_t'] + v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{v_t}{2} \text{tr}\left[\left(\sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t a_t'\right]\right\} \\
&= \exp\left\{v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{1}{2} \text{tr}\left[\Lambda_t a_t a_t' + \left(v_t \sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t a_t'\right]\right\} \\
&= \exp\left\{v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{1}{2} \text{tr}\left[\left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t a_t'\right]\right\} \\
&= \exp\left\{v_t \left(\sum_{k=1}^{P} g_{t,k}' K_{t,k}\right) a_t - \frac{1}{2} a_t' \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k}\right) a_t\right\}
\end{aligned}
$$

Looking at the natural parameters of the multivariate normal distribution in exponential family notation, it is possible to identify the precision and therefore the variance of the full

conditional as $\Sigma(a_t) = \Sigma_{a_t} = (\Lambda_{a_t})^{-1}$. :

$$\eta_1 = v_t \left(\sum_{k=1}^{P} g'_{t,k} K_{t,k} \right)' \qquad \stackrel{!}{=} \Lambda_{t,a} \mu_a$$

$$\text{and} \quad \eta'_2 = \eta_2 = -\frac{1}{2} \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right) \qquad \stackrel{!}{=} -\frac{1}{2} \Lambda_{t,a}$$

Thus the well defined and normalized full conditional $p(a_t|\mathcal{X}, \mathcal{Y}, \mathcal{Z}_{-a_t}) = p(a_t|\mathcal{D}, \mathcal{Z}_{-a_t})$ is

$$\Rightarrow \quad p(a_t|\mathcal{D}, \mathcal{Z}_{-a_t}) = \mathcal{N} \left(a_t \,\middle|\, (\Sigma_{a_t}) \left(\sum_{k=1}^{P} g'_{t,k} K_{t,k} \right)' v_t \,,\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \right)$$

$$= \mathcal{N} \left(a_t \,\middle|\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \left(\sum_{k=1}^{P} g'_{t,k} K_{t,k} \right)' v_t \,,\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \right)$$

The mean of the full conditional of a_t can be rewritten as $\left(\sum_{k=1}^{P} g'_{t,k} K_{t,k} \right)' = \sum_{k=1}^{P} K_{t,k} g_{t,k}$ due to symmetry of the kernel matrices $\{K_{t,k}\}_{k=1}^{P}$. Therefore the full conditional for all weight vectors $\{a_t\}_{t=1}^{T}$, again represented by the stacking them in a vector $a = (a'_1, \ldots, a'_T)'$, is

$$p(a|\mathcal{D}, \mathcal{Z}_{-a}) = \prod_{t=1}^{T} \mathcal{N} \left(a_t \,\middle|\, \Sigma_{a_t} \left(\sum_{k=1}^{P} K_{t,k} g_{t,k} \right) v_t \,,\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \right) \tag{11}$$

or without the placeholder for the variance in the mean

$$p(a|\mathcal{D}, \mathcal{Z}_{-a}) = \prod_{t=1}^{T} \mathcal{N} \left(a_t \,\middle|\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \left(\sum_{k=1}^{P} K_{t,k} g_{t,k} \right) v_t \,, \right.$$
$$\left. \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k} \right)^{-1} \right). \tag{12}$$

Note that calculating the mean involves the variance. Consequently, the variance has to be calculated before the mean.

2.2.3 Full conditional for intermediate outputs G

The derivation of the full conditional for the intermediate outputs $G_t = (g_{t,1}, \ldots, g_{t,P})$ for drug t involves two normal distributions. Both the multivariate normal distribution for the

outcomes y_t for drug t, which has the density

$$p(y_t|e, G_t, b_t, \epsilon_t) = \mathcal{N}\left(y_t \,\middle|\, \sum_{k=1}^{P} e_k g_{t,k} + b_t \cdot \mathbb{1}_{N_t},\, \epsilon_t^{-1} I_{N_t}\right) \qquad , G_t \cdot e = \sum_{k=1}^{P} e_k g_{t,k} \tag{13}$$

$$= \frac{(\det|\epsilon_t I_{N_t}|)^{1/2}}{(2\pi)^{N_t/2}} \cdot \exp\left\{-\frac{1}{2}\left(y_t - \sum_{k=1}^{P} e_k g_{t,k} - b_t \cdot \mathbb{1}_{N_t}\right)' \epsilon_t I_{N_t} \left(y_t - \sum_{k=1}^{P} e_k g_{t,k} - b_t \cdot \mathbb{1}_{N_t}\right)\right\}$$

$$= \left(\frac{\epsilon_t}{(2\pi)}\right)^{N_t/2} \cdot \exp\left\{-\frac{\epsilon_t}{2}(y_t - c_2)'(y_t - c_2)\right\} \qquad , c_2 = \sum_{k=1}^{P} e_k g_{t,k} + b_t \cdot \mathbb{1}_{N_t} \tag{14}$$

and again the prior distribution for the column of random intermediate outputs $g_{t,k}$. The density of a column $g_{t,k}$ is stated in equation (7).

Updates for the approximate distribution for the latent variable G_t can be either formulated in terms of the columns of G_t, i.e. views, or in terms of the rows of G_t, i.e. the cell lines. In order to state the updates in terms of the cell lines, the former equation is given in terms of a single cell line i, this is

$$p(g_{t,k,i}|K_{t,k,i}, a_t, \upsilon_t) = \mathcal{N}(g_{t,k,i}\,|K_{t,k,i}a_t, \upsilon_t^{-1}) = \left(\frac{\upsilon_t}{2\pi}\right)^{1/2} \exp\left\{-\frac{\upsilon_t}{2}(g_{t,k,i} - K_{t,k,i}a_t)^2\right\},$$

where $K_{t,k,i}$ stands for the row of the kernel matrix of view k for drug t. The column i of $K_{t,k}$ contains the same information as the row of the kernel matrix by construction as a symmetric matrix (see Section 3).

In the same way the prior distribution of a single cell line for drug susceptibiliy $y_{t,i}$ can be given, where the row vector $G_{t,i} = (g_{t,1,i}, \ldots, g_{t,P,i}) \in \mathbb{R}^P$ stands for the row i of G_t, giving all intermediate outputs [5] for cell line i for task t.

$$p(y_{t,i}|e, G_{t,i}, b_t, \epsilon_t) = \mathcal{N}\left(y_{t,i} \,\middle|\, \sum_{k=1}^{P} e_k g_{t,k,i} + b_t,\, \epsilon_t^{-1}\right) = \mathcal{N}\left(y_{t,i}\,|\,G_{t,i} \cdot e + b_t,\, \epsilon_t^{-1}\right)$$

$$= \left(\frac{\epsilon_t}{2\pi}\right)^{1/2} \exp\left\{-\frac{\epsilon_t}{2}\left(y_{t,i} - (G_{t,i}e + b_t)\right)^2\right\} \qquad \text{note: } G_{t,i}e = e'G'_{t,i}$$

Since it is common to state multivariate distributions in vector notation, the full conditional distribution for the intermediate outputs for a single cell line i will be stated in terms of the column vector $G'_{t,i} = (g_{t,1,i}, \ldots, g_{t,P,i})'$. In order to do so conveniently, the prior distribution for a vector $G'_{t,i}$ is stated in terms of all relevant kernel information for this vector of intermediate outputs. The different corresponding kernels for i are summerized in the matrix $K_{t,i} = (K_{t,1,i}, \ldots, K_{t,P,i})' \in \mathbb{R}^{P \times N_t}$ where a single $K_{t,k,i}$ is again the single row i of

[5] Recall: Speaking of drug susceptibility $y_{t,i}$ and intermediate outputs $G_{t,i}$ means the random variables of drug susceptibility and intermediate outputs.

the corresponding kernelmatrix for view k for drug t:

$$
\begin{aligned}
p(G'_{t,i}|K_{t,i}, a_t, \upsilon_t) &= \prod_{k=1}^{P} p(g_{t,k,i}|K_{t,k,i}, a_t, \upsilon_t) = \prod_{k=1}^{P} \left(\frac{\upsilon_t}{2\pi}\right)^{1/2} \cdot \exp\left\{-\frac{\upsilon_t}{2}(g_{t,k,i} - K_{t,k,i}a_t)^2\right\} \\
&= \left(\frac{\upsilon_t}{2\pi}\right)^{P/2} \cdot \exp\left\{-\frac{\upsilon_t}{2}\sum_{k=1}^{P}(g_{t,k,i} - K_{t,k,i}a_t)^2\right\} \qquad (15) \\
&= \left(\frac{\upsilon_t}{2\pi}\right)^{P/2} \cdot \exp\left\{-\frac{\upsilon_t}{2}(G'_{t,i} - K_{t,i} \cdot a_t)'(G'_{t,i} - K_{t,i} \cdot a_t)\right\}
\end{aligned}
$$

The full conditional for $G'_{t,i}$ can then be found as

$$
\begin{aligned}
&p(G'_{t,i}|\mathcal{D}, \mathcal{Z}_{-G'_{t,i}}) \propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(G'_{t,i}|K_{t,i}, a_t, \upsilon_t) \cdot p(y_{t,i}|e, G_{t,i}, b_t, \epsilon_t) \\
&\propto \exp\left\{-\frac{\upsilon_t}{2}(G'_{t,i} - K_{t,i} \cdot a_t)'(G'_{t,i} - K_{t,i} \cdot a_t) - \frac{\epsilon_t}{2}\left(y_{t,i} - (G_{t,i}e + b_t)\right)^2\right\} \\
&= \exp\left\{ -\frac{1}{2}\Big[\upsilon_t(G_{t,i}G'_{t,i} - 2a'_tK_{t,i}G'_{t,i} + a'_tK'_{t,i}K_{t,i}a_t) \right. \\
&\qquad \left. + \epsilon_t(y'_{t,i}y_{t,i} - 2y_{t,i}G_{t,i}e + (G_{t,i}e)^2 + 2b_tG_{t,i}e + b_t^2)\Big]\right\} \qquad \text{, note: } G_{t,i}e = e'G'_{t,i} \\
&\propto \exp\left\{-\frac{1}{2}\left(\upsilon_t\left[G_{t,i}G'_{t,i} - 2a'_tK_{t,i}G'_{t,i}\right] + \epsilon_t\left[-2y_{t,i}e'G'_{t,i} + e'G'_{t,i}G_{t,i}e + 2b_te'G'_{t,i}\right]\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(G_{t,i}\upsilon_tI_PG'_{t,i} + \epsilon_tG_{t,i}ee'G'_{t,i}\right) + \upsilon_ta'_tK'_{t,i}G'_{t,i} + \epsilon\left(y_{t,i}e'G'_{t,i} - b_te'G'_{t,i}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}G_{t,i}\left[\upsilon_tI_P + \epsilon_tee'\right]G'_{t,i} + \left[\upsilon_ta'_tK'_{t,i} + \epsilon\left(y_{t,i}e' - b_te'\right)\right]G'_{t,i}\right\}
\end{aligned}
$$

And from that one can again identify as before for the latent variable a_t the mean and precision of the full conditional (see Section A.1.2):

$$
\begin{aligned}
&\eta'_1 = \upsilon_ta'_tK'_{t,i} + \epsilon\left(y_{t,i}e' - b_te'\right) \quad \stackrel{!}{=} \Lambda_{G'_{t,i}}(\mu_{G'_{t,i}})' \quad \Rightarrow\ \mu_{G'_{t,i}} = \Sigma_{G'_{t,i}} \cdot [\upsilon_tK_{t,i}a_t + \epsilon_t(y_{t,i}e - b_te)] \\
&\eta'_2 = \eta_2 = -\frac{1}{2}\left(\upsilon_tI_P + \epsilon_tee'\right) \quad \stackrel{!}{=} -\frac{1}{2}\Lambda_{G'_{t,i}} \quad \Rightarrow\ \Sigma_{G'_{t,i}} = (\upsilon_tI_P + \epsilon_tee')^{-1} \\
&\Rightarrow\ p(G'_{t,i}|\mathcal{D}, \mathcal{Z}_{-G'_{t,i}}) = \mathcal{N}\left(G'_{t,i}\,|\,(\upsilon_tI_P + \epsilon_tee')^{-1} \cdot [\upsilon_tK_{t,i}a_t + \epsilon_t(y_{t,i}e - b_te)]\ ,\ (\upsilon_tI_P + \epsilon_tee')^{-1}\right)
\end{aligned}
$$

Therefore it is possible to state the full conditional distributions for intermediate outputs G_t for each task t in terms of N_t independent multivariate distributions of intermediate outputs of each cell line $G_{t,i}$:

$$
p(G_t|\mathcal{D}, \mathcal{Z}_{-G_t}) = \prod_{i=1}^{N_t}\mathcal{N}\left(G'_{t,i}\,|\,(\upsilon_tI_P + \epsilon_tee')^{-1} \cdot [\upsilon_tK_{t,i}a_t + \epsilon_t(y_{t,i}e - b_te)]\ ,\ (\upsilon_tI_P + \epsilon_tee')^{-1}\right)
$$

Then the full conditional for all intermediate outputs $\{G_t\}_{t=1}^{T}$ in terms of cell lines, and with writing the set as the stacked matrix $G = (G'_1, \ldots, G'_T)'$ of dimension $N \times P$, is

$$
p(G|\mathcal{D}, \mathcal{Z}_{-G}) \ \propto\ \prod_{t=1}^{T}\prod_{i=1}^{N_t}\mathcal{N}\left(G'_{t,i}\,|\,\Sigma_{G'_{t,i}} \cdot [\upsilon_tK_{t,i}a_t + \epsilon_t(y_{t,i}e - b_te)]\ ,\ (\upsilon_tI_P + \epsilon_tee')^{-1}\right), \qquad (16)
$$

where the variance term in the mean was abbreviated as before as $\Sigma_{G'_{t,i}} = (\upsilon_t I_P + \epsilon_t ee')^{-1}$.

Similarly, the full conditional distribution for G_t can be derived in terms of the approximation for the columns $g_{t,k}$ of G_t:

$$
\begin{aligned}
p(G_t|\mathcal{D}, \mathcal{Z}_{-G_t}) \propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto\ & p(G_t|a_t, \upsilon_t, \mathcal{X}) p(y_t|b_t, e, G_t, \epsilon_t) \\
=& \frac{(\upsilon_t)^{P \cdot N/2}}{(2\pi)^{P \cdot D/2}} \cdot \exp\left\{ -\frac{\upsilon_t}{2} \sum_{k=1}^{P} (g_{t,k} - K_{t,k} a_t)'(g_{t,k} - K_{t,k} a_t) \right\} \\
& \cdot \left(\frac{\epsilon_t}{(2\pi)} \right)^{N_t/2} \cdot \exp\left\{ -\frac{\epsilon_t}{2} (y_t - \hat{y}_t)'(y_t - \hat{y}_t) \right\} \quad , \hat{y}_t = \sum_{k=1}^{P} e_k g_{t,k} + b_t \cdot \mathbb{1}_{N_t} \\
\propto & \exp\left\{ -\frac{\upsilon_t}{2} \sum_{k=1}^{P} (g_{t,k} - K_{t,k} a_t)'(g_{t,k} - K_{t,k} a_t) - \frac{\epsilon_t}{2} (y_t - \hat{y}_t)'(y_t - \hat{y}_t) \right\} \\
\propto & \exp\left\{ -\frac{1}{2} \left(\upsilon_t \sum_{k=1}^{P} (g'_{t,k} g_{t,k} - 2(K_{t,k} a_t)' g_{t,k}) + \epsilon_t (c'_2 \hat{y}_t - 2 y'_t \hat{y}_t) \right) \right\}
\end{aligned}
$$

One can briefly calculate $\hat{y}'_t \hat{y}_t$ and $y'_t \hat{y}_t$ in order to be able to see which elements depend only on $g_{t,k}$:

$$
\begin{aligned}
\hat{y}'_t \hat{y}_t =& \sum_{k=1}^{P} e_k^2 g'_{t,k} g_{t,k} + 2 b_t \sum_{k=1}^{P} e_k \mathbb{1}'_{N_t} g_{t,k} + N_t b_t^2 \\
y'_t \hat{y}_t =& \sum_{k=1}^{P} e_k y'_t g_{t,k} + y'_t \mathbb{1}_{N_t} b_t
\end{aligned}
$$

Therefore the full conditional of G_t is proportional to:

$$
\begin{aligned}
& \propto \exp\left\{ -\frac{1}{2} \left(\upsilon_t \sum_{k=1}^{P} (g'_{t,k} g_{t,k} - 2(K_{t,k} a_t)' g_{t,k}) + \epsilon_t \sum_{k=1}^{P} \left(e_k^2 g'_{t,k} g_{t,k} + 2 b_t e_k \mathbb{1}'_{N_t} g_{t,k} - 2 e_k y'_t g_{t,k} \right) \right) \right\} \\
& = \exp\left\{ -\frac{1}{2} \sum_{k=1}^{P} \left(\upsilon_t g'_{t,k} g_{t,k} + \epsilon_t e_k^2 g'_{t,k} g_{t,k} \right) + \sum_{k=1}^{P} \left[\upsilon_t (K_{t,k} a_t)' g_{t,k} + \epsilon_t \left(e_k y'_t g_{t,k} - b_t \mathbb{1}'_{N_t} e_k g_{t,k} \right) \right] \right\} \\
& = \exp\left\{ \sum_{k=1}^{P} \left(-\frac{1}{2} \left(\upsilon_t g'_{t,k} g_{t,k} + \epsilon_t e_k^2 g'_{t,k} g_{t,k} \right) + \left[\upsilon_t \left(K_{t,k} a_t \right)' + \epsilon_t \left(e_k y'_t - b_t e_k \mathbb{1}'_{N_t} \right) \right] g_{t,k} \right) \right\}
\end{aligned}
$$

Since all terms in the exponent are a summation over the P views, one can again state the density function of G_t as a product of independent density functions over $g_{t,k}$ - as before in equation 8:

$$
\begin{aligned}
& p(g_{t,k}|\mathcal{D}, \mathcal{Z}_{-g_{t,k}}) \propto \exp\left\{ -\frac{1}{2} g'_{t,k} \left(\upsilon_t - \epsilon_t e_k^2 \right) g_{t,k} + \left[\upsilon_t \left(K_{t,k} a_t \right)' + \epsilon_t \left(e_k y'_t - b_t e_k \mathbb{1}'_{N_t} \right) \right] g_{t,k} \right\} \\
& \Rightarrow p(g_{t,k}|\mathcal{D}, \mathcal{Z}_{-g_{t,k}}) \propto \mathcal{N}(g_{t,k}|\, \Sigma_{g_{t,k}} \cdot [\upsilon_t (K_{t,k} a_t) + \epsilon_t e_k (y_t - b_t \mathbb{1}_{N_t})] \ , \left(\upsilon_t - \epsilon_t e_k^2 \right)^{-1} I_{N_T})
\end{aligned}
$$

This shows that a formulation of a full conditional for intermediate outputs an be either obtained by specifying the distribution over the columns or rows of the matrix of random variables in G_t. Computationally both formulations are driven by the inversion of the precision

matrix in order to obtain the variance. Thus, the choice should depend on the dimensionality of used inputs, which are both views and the number of cell lines (Gönen, 2017). As there are 35 cell lines for training and a maximum of 22 views in the data used by Costello et al. (2014), his leads to the choice of the definition in terms of the rows of G_t. This choice is common in statistics where the number of observations is generally higher than the number of inputs.

2.2.4 Full conditional for precision υ on intermediate outputs

The derivation of the full conditional of the random variables υ, the precisions on the intermediate outcomes, is similar to the derivation of the full conditional for the precisions λ on the weights a.

$$p(\upsilon_t) = \mathcal{G}am(\upsilon_t \,|\alpha_\upsilon, \beta_\upsilon) = \frac{1}{\Gamma(\alpha_\upsilon)} \beta_\upsilon^{\alpha_\upsilon} \upsilon_t^{\alpha_\upsilon - 1} \exp(-\beta_\upsilon \upsilon_t) \tag{17}$$

$$\Rightarrow \quad p(\upsilon) = \prod_{t=1}^{T} \mathcal{G}am(\upsilon_t \,|\alpha_\upsilon, \beta_\upsilon) = \left(\frac{\beta_\upsilon^{\alpha_\upsilon}}{\Gamma(\alpha_\upsilon)}\right)^T \left(\prod_{t=1}^{T} \upsilon_t\right)^{\alpha_\upsilon - 1} \exp\left(-\beta_\upsilon \sum_{t=1}^{T} \upsilon_t\right) \tag{18}$$

The density for all precision parameters of the intermediate outputs $\{\upsilon_t\}_{t=1}^T$ is the product of T independent densities of the random variables υ_t. The density of this set is again represented as a vector in υ (see Table 1) and its representation is chosen to facilitate later calculations of its logarithm (see Section 2.4). The same holds for the random variables in γ, ω and ϵ.

The full conditional for the gamma prior υ_t can therefore be derived as before the full conditional for $\lambda_{t,i}$:

$$\begin{aligned}
p(\upsilon_t|\mathcal{D}, \mathcal{Z}_{-\upsilon_t}) = \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{\sum_{X,y,Z_{-\upsilon}} p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})} \quad &\propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \quad \propto p(\upsilon_t) \cdot p(g_t|a_t, K_t, \upsilon_t^{-1}) \\
&= \mathcal{G}am(\upsilon_t \,|\alpha_\upsilon, \beta_\upsilon) \cdot \prod_{k=1}^{P} \mathcal{N}(g_{t,k} \,|K_{t,k} a_t, \upsilon_t^{-1}) \\
&\propto \upsilon_t^{\alpha_\upsilon - 1} \exp(-\beta_\upsilon \upsilon_t) \cdot (\upsilon_t)^{P \cdot N_t/2} \exp\left\{-\frac{1}{2} \upsilon_t c_{\upsilon_t}\right\} \;,\; c_{\upsilon_t} = \sum_{k=1}^{P} \sum_{i=1}^{N_t} (g_{t,k,i} - K_{t,k,i} a_t)^2 \\
&= \upsilon_t^{\alpha_\upsilon + \frac{P \cdot N_t}{2} - 1} \cdot \exp\left\{-\left(\beta_\upsilon + \frac{1}{2} c_{\upsilon_t}\right) \upsilon_t\right\} \\
&\propto \mathcal{G}am\left(\upsilon_t \,\middle|\, \alpha_\upsilon + \frac{P \cdot N_t}{2} \,,\, \beta_\upsilon + \frac{c_{\upsilon_t}}{2}\right) \quad , c_{\upsilon_t} = \sum_{k=1}^{P} (g_{t,k} - K_{t,k} a_t)' (g_{t,k} - K_{t,k} a_t)
\end{aligned}$$

The full conditional for the vector υ is

$$p(\upsilon|\mathcal{D}, \mathcal{Z}_{-\upsilon}) \propto \prod_{t=1}^{T} \mathcal{G}am\left(\upsilon_t \,\middle|\, \alpha_\upsilon + \frac{P \cdot N_t}{2} \,,\, \beta_\upsilon + \frac{c_{\upsilon_t}}{2}\right) \,, c_{\upsilon_t} = \sum_{k=1}^{P} \|(g_{t,k} - K_{t,k} a_t)\|^2, \tag{19}$$

where the L2 norm, or euclidean norm, is defined as $\|x\| = \sqrt{x'x}$.

2.2.5 Full conditional on precision γ for biases

The full conditional for the precisions of the bias (intercept) γ can be obtained likewise. The random variables $\gamma = (\gamma_1, \ldots, \gamma_T)'$ have a gamma distribution as prior and are used as the precision in the normal distribution of biases $b = (b_1, \ldots, b_T)'$. Their distributions in the model are:

$$p(\gamma_t|\alpha_\gamma, \beta_\gamma) = \quad \mathcal{G}am(\gamma_t \,|\alpha_\gamma, \beta_\gamma) \; = \frac{1}{\Gamma(\alpha_\gamma)} \beta_\gamma^{\alpha_\gamma} \gamma_t^{\alpha_\gamma - 1} \exp(-\beta_\gamma \gamma_t) \tag{20}$$

$$\Rightarrow \quad p(\gamma|\alpha_\gamma, \beta_\gamma) = \prod_{t=1}^{T} \mathcal{G}am(\gamma_t \,|\alpha_\gamma, \beta_\gamma) = \left(\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\right)^T \left(\prod_{t=1}^{T} \gamma_t^{\alpha_\gamma - 1}\right) \exp\left(-\beta_\gamma \sum_{t=1}^{T} \gamma_t\right) \tag{21}$$

$$p(b_t|\gamma_t) = \qquad \mathcal{N}(b_t \,|0, \gamma_t^{-1}) = \left(\frac{\gamma_t}{2\pi}\right)^{1/2} \exp\left\{-\frac{\gamma_t}{2} b_t^2\right\} \tag{22}$$

$$\Rightarrow \quad p(b|\,\gamma) = \prod_{t=1}^{T} \mathcal{N}(b_t \,|0, \gamma_t^{-1}) = \mathcal{N}(b \,|\, 0, \mathrm{diag}(\gamma)) = \frac{\prod_{t=1}^{T} \gamma_t}{(2\pi)^{-T/2}} \underbrace{\exp\left\{-\frac{1}{2} \sum_{t=1}^{T} \gamma_t b_t^2\right\}}_{\exp\left\{-\frac{1}{2} b' \,\mathrm{diag}(\gamma) b\right\}} \tag{23}$$

The full conditional for one random variable γ_t is:

$$\begin{aligned}
p(\gamma_t|\mathcal{D}, \mathcal{Z}_{-\gamma_t}) &\propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(\gamma_t) \cdot p(b_t|\gamma_t) \\
&= \mathcal{G}am(\gamma_t \,|\alpha_\gamma, \beta_\gamma) \cdot \mathcal{N}(b_t \,|0, \gamma_t^{-1}) \\
&= \frac{1}{\Gamma(\alpha_\gamma)} \beta_\gamma^{\alpha_\gamma} \gamma_t^{\alpha_\gamma - 1} \exp(-\beta_\gamma \gamma_t) \cdot \left(\frac{\gamma_t}{2\pi}\right)^{1/2} \exp\left\{-\frac{\gamma_t}{2} b_t^2\right\} \\
&\propto \gamma_t^{\alpha_\gamma - 1} \exp(-\beta_\gamma \gamma_t) \cdot (\gamma_t)^{1/2} \exp\left\{-\frac{\gamma_t}{2} b_t^2\right\} \\
&= \gamma_t^{\alpha_\gamma + \frac{1}{2} - 1} \exp\{-(\beta_\gamma + \tfrac{1}{2} b_t^2)\gamma_t\} \\
&\propto \mathcal{G}am\left(\gamma_t \,\middle|\, \alpha_\gamma + \frac{1}{2}\,,\, \beta_\gamma + \frac{1}{2} b_t^2\right)
\end{aligned}$$

The full conditional for the vector γ is the product of all γ_t over all drugs t:

$$p(\gamma|\mathcal{D}, \mathcal{Z}_{-\gamma}) \propto \prod_{t=1}^{T} \mathcal{G}am\left(\gamma_t \,\middle|\, \alpha_\gamma + \frac{1}{2}\,,\, \beta_\gamma + \frac{1}{2} b_t^2\right). \tag{24}$$

2.2.6 Full conditional for precision ω on kernel coefficients

The derivation of the full conditional of the precisions ω for the kernel coefficients e for the views is done as before for the full conditional of the factors λ, υ and ω. The prior of the

random variables $\omega = (\omega_1, \ldots, \omega_P)'$ and the normal prior density of the random variables $e = (e_1, \ldots, e_P)'$ are needed:

$$p(\omega_k|\alpha_\omega, \beta_\omega) = \quad \mathcal{G}am(\omega_k\,|\alpha_\omega, \beta_\omega) \;= \frac{1}{\Gamma(\alpha_\omega)} \beta_\omega^{\alpha_\omega} \omega_k^{\alpha_\omega - 1} \exp(-\beta_\omega \omega_k) \tag{25}$$

$$\Rightarrow\; p(\omega) = \prod_{k=1}^{P} \mathcal{G}am(\omega_k\,|\alpha_\omega, \beta_\omega) = \left(\frac{\beta_\omega^{\alpha_\omega}}{\Gamma(\alpha_\omega)}\right)^P \left(\prod_{k=1}^{P} \omega_k\right)^{\alpha_\omega - 1} \exp\left(-\beta_\omega \sum_{k=1}^{P} \omega_k\right) \tag{26}$$

$$p(e_k|\omega_k) = \quad \mathcal{N}(e_k\,|0, \omega_k^{-1}) \;= \left(\frac{\omega_k}{2\pi}\right)^{1/2} \exp\left\{-\frac{\omega_k}{2} b_t^2\right\} \tag{27}$$

$$\Rightarrow \quad p(e|\,\omega) = \prod_{k=1}^{P} \mathcal{N}(e_k\,|0, \omega_k^{-1}) = \mathcal{N}(e\,|\,0, \mathrm{diag}(\omega)) = \mathcal{N}(e\,|\,0, \Omega) \quad , \Omega = \mathrm{diag}(\omega) \tag{28}$$

$$= (2\pi)^{-P/2} |\Omega|^{1/2} \exp\left\{-\frac{1}{2} e' \Omega e\right\}$$

The full conditional for the latent variable ω_k is then:

$$\begin{aligned}
p(\omega_k|\mathcal{D}, \mathcal{Z}_{-\omega_k}) &\propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(\omega_k) \cdot p(e_k|\,\omega_k) \\
&= \mathcal{G}am(\omega_k\,|\alpha_\omega, \beta_\omega) \cdot \mathcal{N}(e_k\,|0, \omega_k^{-1}) \\
&= \frac{1}{\Gamma(\alpha_\omega)} \beta_\omega^{\alpha_\omega} \omega_k^{\alpha_\omega - 1} \exp(-\beta_\omega \omega_k) \cdot \left(\frac{\omega_k}{2\pi}\right)^{1/2} \exp\left\{-\frac{\omega_k}{2} e_k^2\right\} \\
&\propto \omega_k^{\alpha_\omega - 1} \exp(-\beta_\omega \omega_k) \cdot (\omega_k)^{1/2} \exp\left\{-\frac{\omega_k}{2} e_k^2\right\} \\
&= \omega_k^{\alpha_\omega + \frac{1}{2} - 1} \exp\left\{-(\beta_\omega + \tfrac{1}{2} e_k^2)\, \omega_k\right\} \\
&\propto \mathcal{G}am\left(\omega_k|\; \alpha_\omega + \frac{1}{2}\,,\; \beta_\omega + \frac{1}{2} e_k^2\right)
\end{aligned}$$

The full conditional of the random variables of all ω_k contained in ω is

$$p(\omega|\mathcal{D}, \mathcal{Z}_{-\omega}) \propto \prod_{k=1}^{P} \mathcal{G}am\left(\omega_k \middle|\; \alpha_\omega + \frac{1}{2}\,,\; \beta_\omega + \frac{1}{2} e_k^2\right) \tag{29}$$

2.2.7 Full Conditional on biases b and kernel coefficients e

Next the full conditional for the random variables b_t and e are derived separately. As these two are highly related to each other as being the coefficients for the outcomes, however statistically independent, it will be computationally advantageous to derive an approximating distribution jointly for these random variables in a second step (Gönen, 2017).

Furthermore it can be interesting to model approximate the biases apart of the kernel coefficients in order to add prior knowledge of drug similarities, e.g. by modeling a variance-covariance for the biases capturing drug relations.

2.2.7.1 Full conditional only for biases b:

The full conditional for the bias is a combination of two normal densities. The first normal is the prior of the bias used in the mean of the normal distribution of drug outcomes. Their densities can be found in equation (14) and (22):

$$
\begin{aligned}
p(b_t|\mathcal{D}, \mathcal{Z}_{-b_t}) &\propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(b_t\,|\gamma_t) \cdot p(y_t\,|b_t, e, G_t, \epsilon_t) \\
&= \mathcal{N}(b_t\,|0, \gamma_t^{-1}) \cdot \mathcal{N}\left(y_t \,\middle|\, \sum_{k=1}^{P} e_k g_{t,k} + b_t \cdot \mathbb{1}_{N_t},\, \epsilon_t^{-1} I_{N_t}\right) \\
&= \left(\frac{\gamma_t}{2\pi}\right)^{1/2} \exp\left\{-\frac{\gamma_t}{2} b_t^2\right\} \cdot \left(\frac{\det|\epsilon_t I_{N_t}|}{2\pi}\right)^{1/2} \cdot \exp\left\{-\frac{\epsilon_t}{2}\left\|y_t - \sum_{k=1}^{P} e_k g_{t,k} - b_t \cdot \mathbb{1}_{N_t}\right\|\right\} \\
&\propto \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2}\left\|y_t - \sum_{k=1}^{P} e_k g_{t,k} - b_t \cdot \mathbb{1}_{N_t}\right\|\right\} \qquad , \sum_{k=1}^{P} e_k g_{t,k} = G_t e \\
&= \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2}(y_t - g_t - b_t \cdot \mathbb{1}_{N_t})'(y_t - g_t - b_t \cdot \mathbb{1}_{N_t})\right\} \\
&= \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2}(y_t' y_t - 2y_t'(G_t e + b_t \mathbb{1}_{N_t}) + e' G_t' G_t e + 2e' G_t' \mathbb{1}_{N_t} b_t + b_t^2 \mathbb{1}_{N_t}' \mathbb{1}_{N_t})\right\} \\
&= \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2}\left(y_t' y_t - 2y_t' G_t e - 2y_t' \mathbb{1}_{N_t} b_t + e' G_t' G_t e + 2e' G_t' \mathbb{1}_{N_t} b_t + b_t^2 \mathbb{1}_{N_t}' \mathbb{1}_{N_t}\right)\right\} \\
&= \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2}\left(\underbrace{y_t' y_t - 2y_t' G_t e}_{=c_1} - 2\underbrace{y_t' \mathbb{1}_{N_t}}_{\sum_i y_{it}} b_t + \underbrace{e' G_t' G_t e}_{=c_2} + 2\underbrace{e' G_t' \mathbb{1}_{N_t}}_{=\sum_i g_{it}} b_t + b_t^2 \underbrace{\mathbb{1}_{N_t}' \mathbb{1}_{N_t}}_{=N_t}\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\left\{-\frac{\gamma_t}{2} b_t^2 - \frac{\epsilon_t}{2} N_t b_t^2 + \epsilon_t y_t' \mathbb{1}_{N_t} b_t - \epsilon_t e' G_t' \mathbb{1}_{N_t} b_t\right\} \cdot \exp\{-\frac{\epsilon_t}{2}(y_t' y_t - 2y_t' G_t e + e' G_t' G_t e)\} \\
&\propto \exp\left\{-\frac{1}{2}\left((\gamma_t + \epsilon_t N_t) b_t^2 + 2\epsilon_t(-y_t' \mathbb{1}_{N_t} + e' G_t' \mathbb{1}_{N_t}) b_t\right)\right\} \\
&= \exp\left\{\left((\epsilon_t(y_t' \mathbb{1}_{N_t} - e' G_t' \mathbb{1}_{N_t}), -\frac{1}{2}(\gamma_t + \epsilon_t N_t)\right)\begin{pmatrix} b_t \\ b_t^2 \end{pmatrix}\right\}
\end{aligned}
$$

This equation can now be rewritten as a sufficient statistic vector $t(b_t) = (b_t, b_t^2)'$ and natural parameters $\eta' = (\eta_1, \eta_2)$.

$$
\eta_1 = \eta_1' = \Sigma_{b_t} \mu_{b_t} = \epsilon_t(y_t' \mathbb{1}_{N_t} - e' G_t' \mathbb{1}_{N_t}) \qquad , \quad \eta_2 = -\frac{1}{2}\Lambda_{b_t} = -\tfrac{1}{2}(\gamma_t + \epsilon_t N_t)
$$

$$
\mu_{b_t} = \Sigma_{b_t} \epsilon_t (y_t' \mathbb{1}_{N_t} - e' G_t' \mathbb{1}_{N_t}) \qquad \text{and} \quad \Sigma_{b_t} = (\Lambda_{b_t})^{-1} = (\gamma_t + \epsilon_t N_t)^{-1}
$$

The full conditional for a single bias b_t reveals that its mean is a weighted difference between all outcomes for drug t minus the prediction on the basis of intermediate outputs. It thus captures the overall mean for drug t of unexplained differences.

2.2.7.2 Full conditional only for kernel coefficients e

The full conditional only for the kernel coefficients can be derived with the densities given in equations (13), (27) and (28):

$$
\begin{aligned}
p(e|\mathcal{D}, \mathcal{Z}_{-e}) &\propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \propto p(e|\omega) \cdot p(\mathcal{Y}|e, G, b, \epsilon) \\
&= \left(\prod_{k=1}^{P} \mathcal{N}(e_k \,|0, \omega_k^{-1}) \right) \cdot \prod_{t=1}^{T} \mathcal{N}\left(y_t \,|\, G_t e + b_t \cdot \mathbb{1}_{N_t},\, \epsilon_t^{-1} I_{N_t}\right) \\
&= \left(\prod_{k=1}^{P} \left(\frac{\omega_k}{2\pi}\right)^{1/2} \exp\left\{-\frac{\omega_k}{2} e_k^2\right\} \right) \sqrt{\prod_{t=1}^{T} \frac{\det|\epsilon_t I_{N_t}|}{(2\pi)^{N_t}}} \cdot \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \epsilon_t \|y_t - G_t e - b_t \mathbb{1}_{N_t}\| \right\} \\
&\propto \left(\prod_{k=1}^{P} \exp\left\{-\frac{\omega_k}{2} e_k^2\right\} \right) \cdot \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \epsilon_t \left(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\right)' \left(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\right) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\sum_{k=1}^{P} \omega_k e_k^2 + \sum_{t=1}^{T} \epsilon_t \left(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\right)' \left(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\right) \right] \right\} \\
&= \exp\Bigg\{ -\frac{1}{2} \Bigg[e' \operatorname{diag}(\omega) e + \sum_{t=1}^{T} \epsilon_t \big(y_t' y_t - y_t' G_t e - y_t' b_t \mathbb{1}_{N_t} - e' G_t' y_t \\
&\qquad\qquad + e' G_t' G_t e + e' G_t' b_t \mathbb{1}_{N_t} - b_t \mathbb{1}_{N_t}' y_t + b_t \mathbb{1}_{N_t}' G_t e + b_t^2 N_t \big) \Bigg] \Bigg\} \\
&= \exp\left\{ -\frac{1}{2} \left[e' \operatorname{diag}(\omega) e + \sum_{t=1}^{T} \epsilon_t \big(-2 y_t' G_t e + e' G_t' G_t e + 2 b_t \mathbb{1}_{N_t}' G_t e + c \big) \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2} e' \left(\operatorname{diag}(\omega) + G_t' \epsilon_t G_t \right) e + \left(\sum_{t=1}^{T} \epsilon_t \big(y_t' - b_t \mathbb{1}_{N_t}' \big) G_t \right) e \right\}
\end{aligned}
$$

One can again recognise the sufficient statistic $t(e) = (e, ee')'$ and the natural parameters $\eta' = (\eta_1', \eta_2')$:

$$
\eta_1' = \mu_e^{\star\prime} \Lambda_e = \sum_{t=1}^{T} \epsilon_t \left(y_t' - b_t \mathbb{1}_{N_t}'\right) G_t \quad \text{and} \quad \eta_2 = -\frac{1}{2} \Lambda_e^{\star} = -\frac{1}{2} \left(\operatorname{diag}(\omega) + \sum_{t=1}^{T} G_t' \epsilon_t G_t \right)
$$

Having these natural parameters, one identifies a normal with mean and covariance of :

$$
\mu_e^{\star} = \Sigma_e^{\star} \sum_{t=1}^{T} \epsilon_t G_t' (y_t - b_t \mathbb{1}_{N_t}) \quad \text{and} \quad \Sigma_e^{\star} = (\Lambda_e^{\star})^{-1} = \left(\operatorname{diag}(\omega) + \sum_{t=1}^{T} G_t' \epsilon_t G_t \right)^{-1}
$$

For the mean of the full conditional for kernel coefficients e the relation with linear regression becomes apparent. The inverse, build of a shifted sum of weighted squared intermediate outputs, is multiplied with intermediate outputs as well as outcomes demeaned by their bias terms. This estimator resembles a mix of L2 regularized regression, sometimes called Ridge regression, and weighted ordinary least squares estimators.

2.2.7.3 Joint full conditional for biases b and kernel coefficients e

Now the full conditional for the vector of random variables $(b', e')'$ is derived jointly as it is assumed in the approximate posterior $q(\mathcal{Z})$ given in equation 3. In order to give the joint full conditional, the joint distribution of b and e is constructed in multivariate form:

$$\begin{aligned} p(b, e|\gamma, \omega) &= \mathcal{N}\left(\begin{pmatrix} b \\ e \end{pmatrix} \middle| 0, \begin{bmatrix} \text{diag}(\gamma) & 0 \\ 0 & \Omega \end{bmatrix}\right) = \mathcal{N}\left(\begin{pmatrix} b \\ e \end{pmatrix} \middle| 0, \text{diag}(\gamma', \omega')\right) \\ &= (2\pi)^{-(T+P)/2} |\text{diag}(\gamma', \omega')| \cdot \exp\left\{-\frac{1}{2}(b', e')\,\text{diag}(\gamma', \omega')(b', e')'\right\} \end{aligned} \tag{30}$$

Since the kernel coefficients are equally determined by all outputs y_t, but the biases (intercepts) are drug specific, the prior distribution on the basis of equations 13 and 14 are given for all outcomes $\{y_t\}_{t=1}^T = \mathcal{Y}$, where G_t defines the matrix of intermediate outputs for drug t as introduced before ($G_t e = \sum_k g_{t,k} e_k$):

$$\begin{aligned} p(\mathcal{Y}|e, G, b, \epsilon) &= \prod_{t=1}^{T} \mathcal{N}(y_t|G_t \cdot e + b_t \mathbb{1}_{N_t}, \epsilon_t I_{N_t}) \quad \text{, set: } G_t e + b_t \mathbb{1}_{N_t} = (\mathbb{1}_{N_T}, G_t)(b_t, e')' = \hat{y}_t \\ &= \frac{\prod_{t=1}^{T} \epsilon_t^{N_t/2}}{(2\pi)^{N/2}} \cdot \exp\left\{-\frac{\epsilon_t}{2} \sum_{t=1}^{T} (y_t - \hat{y}_t)'(y_t - \hat{y}_t)\right\} \qquad , N = \sum_{t=1}^{T} = N_t \\ &\propto \exp\left\{-\frac{1}{2}(y_1' - \hat{y}_1', \ldots, y_T' - \hat{y}_T')\,\text{diag}(\epsilon)(y_1' - \hat{y}_1', \ldots, y_T' - \hat{y}_T')'\right\} \end{aligned}$$

Next the single substraction $y_t - \hat{y}_t$ is reformulated as a matrix multiplication where the combined vector of the random variables of interest b and e appears:

$$\begin{aligned} &\left(y_1' - \left[(\mathbb{1}_{N_{t_1}}, G_1)(b_1, e')'\right]', \ldots, \; y_T' - \left[(\mathbb{1}_{N_T}, G_T)(b_T, e')'\right]'\right)' \\ =&\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}}_{N \times 1} - \underbrace{\begin{pmatrix} \mathbb{1}_{N_1} & & & G_1 \\ & \ddots & & \vdots \\ & & \mathbb{1}_{N_T} & G_T \end{pmatrix}}_{N \times (T+P)} \underbrace{\begin{pmatrix} b \\ e \end{pmatrix}}_{(T+P) \times 1} \qquad , N = \sum_{t=1}^{T} N_t \\ =&y - \left(\text{diag}(\mathbb{1}_{N_1}, \ldots, \mathbb{1}_{N_T}), G\right)(b', e')' \quad , G = \left(G_1', \ldots, G_T'\right)', y = \left(y_1', \ldots, y_T'\right)' \\ =&y - B \cdot (b', e')' \qquad , B = \left(\text{diag}(\mathbb{1}_{N_1}, \ldots, \mathbb{1}_{N_T}), G\right) = \left(\text{diag}(\{\mathbb{1}_{N_t}\}_{t=1}^T), G\right) \end{aligned}$$

The matrix B is a $N \times (T + P)$ matrix. It combines a block-diagonal matrix of ones, representing in each block the number of cell lines for drug t, and a stacked matrix of intermediate outputs for all drugs t. Thus the joint density for all y_t can be reformulated being proportional

to:

$$p(\mathcal{Y}|e,G,b,\epsilon) \propto \exp\Bigg\{ -\frac{1}{2}\Bigg(\Bigg[y' - (b',e') \begin{pmatrix} \mathbb{1}'_{N_1} & & \\ & \ddots & \\ & & \mathbb{1}'_{N_T} \\ G'_1 & \dots & G'_T \end{pmatrix} \Bigg] \begin{pmatrix} I_{N_1}\epsilon_1 & & \\ & \ddots & \\ & & I_{N_T}\epsilon_T \end{pmatrix} \Bigg[y - \begin{pmatrix} \mathbb{1}_{N_1} & & & G_1 \\ & \ddots & & \vdots \\ & & \mathbb{1}_{N_T} & G_T \end{pmatrix} \begin{pmatrix} b \\ e \end{pmatrix} \Bigg] \Bigg) \Bigg\}$$

The prior density in compact notation for all $\mathcal{Y}$ is:

$$\begin{aligned} p\big(\mathcal{Y}|e,G,b,\epsilon\big) =& \frac{\prod_{t=1}^{T} \epsilon_t^{N_t/2}}{(2\pi)^{N/2}} \qquad , \; B = \big(\operatorname{diag}(\mathbb{1}_{N_1},\dots,\mathbb{1}_{N_T}), G\big) \\ & \cdot \exp\Bigg\{ -\frac{1}{2}\Bigg(\big[y' - (b',e')B'\big] \cdot \operatorname{diag}\big(\{I_{N_t}\epsilon_t\}_{t=1}^{T}\big) \cdot \big[y - B(b',e')'\big] \Bigg) \Bigg\} \end{aligned} \tag{31}$$

Therefore the full conditional for b and e combined can be written as

$$\begin{aligned} & p(b,e|\mathcal{D},\mathcal{Z}_{-b,-e}) \\ & \propto p(\mathcal{Z},\mathcal{Y}|\mathcal{X}) \propto p(b,e|\gamma,\omega) \cdot p(\mathcal{Y}|e,G,b,\epsilon) \\ & \propto \exp\Bigg\{ -\frac{1}{2}(b',e') \operatorname{diag}(\gamma',\omega')(b',e')' \Bigg\} \cdot \exp\Bigg\{ -\frac{1}{2}\Bigg(\big[y' - (b',e')B'\big] \\ & \cdot \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T) \cdot \big[y - B\cdot(b',e')'\big] \Bigg) \Bigg\} \\ & = \exp\Bigg\{ -\frac{1}{2}\Bigg((b',e') \operatorname{diag}(\gamma',\omega')(b',e')' + y' \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T) y \\ & - 2(y'_1,\dots,y'_T) \cdot \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T) \cdot B(b',e')' \\ & + (b',e')(\operatorname{diag}(\{\mathbb{1}_{N_t}\}_{t=1}^{T}) \vdots G)' \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T)(\operatorname{diag}(\{\mathbb{1}_{N_t}\}_{t=1}^{T}) \vdots G)(b',e')' \Bigg) \Bigg\} \\ \propto & \exp\Bigg\{ (y'_1,\dots,y'_T) \cdot \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T) \cdot (\operatorname{diag}(\{I_t\epsilon_t\}_{T=1}^{T}) \vdots G)(b',e')' \\ & - \frac{1}{2}\Bigg((b',e') \operatorname{diag}(\gamma,\omega)(b',e')' + (b',e')B' \operatorname{diag}(I_{N_1}\epsilon_1,\dots,I_{N_T}\epsilon_T) B(b',e')' \Bigg) \Bigg\} \\ & = \exp\Bigg\{ \Big(y'_1\epsilon_1\mathbb{1}_{N_1}, \quad \dots, \quad y'_T\epsilon_T\mathbb{1}_{N_T}, \quad \textstyle\sum_t y'_t\epsilon_t G_t \Big) \begin{pmatrix} b \\ e \end{pmatrix} \\ & \qquad -\frac{1}{2}(b',e') \begin{pmatrix} \gamma_1 + N_1\epsilon_1 & & & \epsilon_1\mathbb{1}'_{N_1}G_1 \\ & \ddots & & \vdots \\ & & \gamma_T + N_T\epsilon_T & \epsilon_T\mathbb{1}'_{N_T}G_T \\ \epsilon_1 G'_1\mathbb{1}_{N_1} & \dots & \epsilon_T G'_T\mathbb{1}_{N_T} & \operatorname{diag}(\omega) + \sum_t G'_t\epsilon_t G_t \end{pmatrix} \begin{pmatrix} b \\ e \end{pmatrix} \Bigg\} \end{aligned}$$

The variance of the full conditional is therefore the inverse of a matrix which consists of four building blocks. The upper left is a diagonal $T \times T$ matrix which contains the precisions

of outputs and bias. The upper right block is the transposed of the lower left block. The matrix contains in its rows the sums of all intermediate outputs over all observations and weights it by the precision of the outputs. It has dimension $T \times P$. The lower right matrix has dimensions $P \times P$ and is a sum of the diagonal matrix of the precision of the kernel coefficients and a sum over all drugs of the sandwich-matrices of the intermediate output of drug t as the bread with the corresponding precision of the drug sensitivity as the butter. The full conditional is thus proportional to a normal distribution parametrized as

$$p(b, e|\mathcal{D}, \mathcal{Z}_{-b,-e}) \propto \mathcal{N}\left(\begin{pmatrix} b \\ e \end{pmatrix} \middle| \Sigma_{(b',e')'} \cdot \begin{pmatrix} y_1' \epsilon_1 \mathbb{1}_{N_1} \\ \vdots \\ y_T' \epsilon_T \mathbb{1}_{N_T} \\ \sum_t G_t' \epsilon_t y_t \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} \gamma_1 + N_1 \epsilon_1 & & & \epsilon_1 \mathbb{1}_{N_1}' G_1 \\ & \ddots & & \vdots \\ & & \gamma_T + N_T \epsilon_T & \epsilon_T \mathbb{1}_{N_T}' G_T \\ \epsilon_1 G_1' \mathbb{1}_{N_1} & \dots & \epsilon_T G_T' \mathbb{1}_{N_T} & \text{diag}(\omega) + \sum_t G_t' \epsilon_t G_t \end{pmatrix}^{-1} \right) \tag{32}$$

The mean of the joint density function of b and e is harder to interpret as it involves the inversion of the variance in block form. Since it is computationally more efficient to derive the full conditional for this ensemble of random variables (Gönen, 2017), it is implemented although it is harder to interpret. For an interpretation of the means of the random variables, please see the explanations for the separate full conditionals of b_t and e. Note that it will be necessary to extract distributions for both anyways as b_t and e appear in other full conditionals separately.

2.2.8 Full conditional for precision ϵ on outputs

Finally the full conditional for the precisions ϵ for the observed outputs $\mathcal{Y} = \{y_t\}_{t=1}^T$ has to be derived. This can be done for the outcome and bias of each drug t separately. The prior of ϵ_t has as prior density:

$$\begin{aligned} p(\epsilon_t) = \mathcal{G}am(\epsilon_t \,|\alpha_\epsilon, \beta_\epsilon) \qquad &= \frac{1}{\Gamma(\alpha_\epsilon)} \beta_\epsilon^{\alpha_\epsilon} \epsilon_t^{\alpha_\epsilon - 1} \exp(-\beta_\epsilon \epsilon_t) \\ \Rightarrow \quad p(\epsilon) = \prod_{t=1}^T \mathcal{G}am(\epsilon_t \,|\alpha_\epsilon, \beta_\epsilon) \quad &= \Gamma(\alpha_\epsilon)^{-T} \beta_\epsilon^{T\alpha_\epsilon} \left(\prod_{t=1}^T \epsilon_t^{\alpha_\epsilon - 1} \right) \cdot \exp\left(-\beta_\epsilon \sum_{t=1}^T \epsilon_t \right) \end{aligned} \tag{33}$$

Therefore its full conditional can be derived in conjunction with the normal distribution for outputs as stated in equation (14), where the scalar product is replaced by the notation in

terms of the euclidean norm:

$$
\begin{aligned}
p(\epsilon_t|X,y,Z_{-\epsilon_t}) &\propto p(\mathcal{Z},\mathcal{Y}|\mathcal{X}) \propto p(\epsilon_t)p(y_t\,|b_t,e,G_t,\epsilon_t) \\
&= \mathcal{G}am(\epsilon_t\,|\alpha_\epsilon,\beta_\epsilon)\cdot\mathcal{N}\left(y_t\,\middle|\,\sum_{k=1}^{P}e_k g_{t,k}+b_t\cdot\mathbb{1}_{N_t},\,\epsilon_t^{-1}I_{N_t}\right) \\
&= \frac{1}{\Gamma(\alpha_\epsilon)}\beta_\epsilon^{\alpha_\epsilon}\epsilon_t^{\alpha_\epsilon-1}\exp(-\beta_\epsilon\epsilon_t)\cdot\left(\frac{\det|\epsilon_t I_{N_t}|}{(2\pi)^{N_t}}\right)^{\frac{1}{2}}\exp\left\{-\frac{\epsilon_t}{2}\|y_t-G_te-b_t\cdot\mathbb{1}_{N_t}\|^2\right\} \\
&\propto \epsilon_t^{\alpha_\epsilon-1}\exp(-\beta_\epsilon\epsilon_t)\cdot\epsilon_t^{\frac{N_t}{2}}\cdot\exp\left\{-\frac{c_{\epsilon_t}}{2}\epsilon_t\right\} \qquad ,\, c_{\epsilon_t}=\|y_t-G_te-b_t\cdot\mathbb{1}_{N_t}\|^2 \\
&= \epsilon_t^{(\alpha_\epsilon+\frac{N_t}{2})-1}\cdot\exp\left\{-(\beta_\epsilon+\frac{c_{\epsilon_t}}{2})\,\epsilon_t\right\} \\
&\propto \mathcal{G}am\left\{\epsilon_t\,|\,\alpha_\epsilon+\frac{N_t}{2}\,,\,\beta_\epsilon+\frac{c_{\epsilon_t}}{2}\right\}
\end{aligned}
$$

Then the full conditional of the vector ϵ containing all $\{\epsilon_t\}_{t=1}^T$ is

$$
p(\epsilon|\mathcal{D},\mathcal{Z}_{-\epsilon})\propto\prod_{t=1}^{T}\mathcal{G}am\left\{\epsilon_t\,\middle|\,\alpha_\epsilon+\frac{N_t}{2}\,,\,\beta_\epsilon+\frac{c_{\epsilon_t}}{2}\right\}\quad ,c_{\epsilon_t}=\|y_t-G_te-b_t\cdot\mathbb{1}_{N_t}\|^2\quad . \tag{34}
$$

All full conditionals are summarized in Table 3.

Table 3: Full conditionals of *sets of random variables*, denoted factors

r.v.	full conditional for factor
λ	$p(\lambda\|\mathcal{D},\mathcal{Z}_{-\lambda})=\prod_{t=1}^{T}\prod_{i=1}^{N_t}\mathcal{G}am\left(\lambda_{t,i}\|\,\alpha_\lambda+\frac{1}{2},\,\beta_\lambda+\frac{1}{2}a_{t,i}^2\right)$
a	$p(a\|\mathcal{D},\mathcal{Z}_{-a})=\prod_{t=1}^{T}\mathcal{N}\left(a_t\,\middle\|\,\Sigma_{a_t}\left(\sum_{k=1}^{P}K_{t,k}g_{t,k}\right)v_t\,,\,\left(\Lambda_t+v_t\sum_{k=1}^{P}K_{t,k}K_{t,k}\right)^{-1}\right)$
G	$p(G\|\mathcal{D},\mathcal{Z}_{-G})=\prod_{t=1}^{T}\prod_{i=1}^{N_t}\mathcal{N}(G'_{t,i}\,\|\,\Sigma_{G'_{t,i}}\cdot[v_tK_{t,i}a_t+\epsilon_t(y_{t,i}e-b_te)]\,,\,(v_tI_P+\epsilon_tee')^{-1})$
υ	$p(\upsilon\|\mathcal{D},\mathcal{Z}_{-\upsilon})=\prod_{t=1}^{T}\mathcal{G}am\left(\upsilon_t\,\middle\|\,\alpha_\upsilon+\frac{P\cdot N_t}{2}\,,\,\beta_\upsilon+\frac{c_{\upsilon,t}}{2}\right),c_{\upsilon,t}=\sum_{k=1}^{P}(\|(g_{t,k}-K_{t,k}a_t)\|_2)^2$
γ	$p(\upsilon\|\mathcal{D},\mathcal{Z}_{-\upsilon})=\prod_{t=1}^{T}\mathcal{G}am\left(\gamma_t\,\middle\|\,\alpha_\gamma+\frac{1}{2}\,,\,\beta_\gamma+\frac{1}{2}b_t^2\right)$
ω	$p(\gamma\|\mathcal{D},\mathcal{Z}_{-\gamma})=\prod_{k=1}^{P}\mathcal{G}am\left(\omega_k\|\,\alpha_\omega+\frac{1}{2}\,,\,\beta_\omega+\frac{1}{2}e_k^2\right)$
b,e	$p(b,e\|\mathcal{D},\mathcal{Z}_{-(b,e)})=\mathcal{N}\left(\begin{matrix}b\\e\end{matrix}\,\middle\|\,\Sigma_{(b',e')'}\cdot\left(y'_1\epsilon_1\mathbb{1}_{N_1},\;\ldots,\;y'_T\epsilon_T\mathbb{1}_{N_T},\;\sum_t y'_t\epsilon_tG_t\right)'\right.,$ $\left.\begin{pmatrix}\gamma_1+N_1\epsilon_1 & & & \epsilon_1\mathbb{1}'_{N_1}G_1\\ & \ddots & & \vdots\\ & & \gamma_T+N_T\epsilon_T & \epsilon_T\mathbb{1}'_{N_T}G_T\\ \epsilon_1G'_1\mathbb{1}_{N_1} & \ldots & \epsilon_TG'_T\mathbb{1}_{N_T} & \mathrm{diag}(\omega)+\sum_t G'_t\epsilon_tG_t\end{pmatrix}^{-1}\right)$
ϵ	$p(\epsilon\|\mathcal{D},\mathcal{Z}_{-\epsilon})=\prod_{t=1}^{T}\mathcal{G}am\left\{\epsilon_t\,\middle\|\,\alpha_\epsilon+\frac{N_t}{2}\,,\,\beta_\epsilon+\frac{c_{\epsilon,t}^2}{2}\right\}\quad ,c_{\epsilon,t}=\|y_t-G_te-b_t\cdot\mathbb{1}_{N_t}\|$

2.3 Log evidence as lower bound

Starting from $\mathrm{KL}\left[q(\mathcal{Z})||p(\mathcal{Z}|\mathcal{X},\mathcal{Y})\right] = \mathrm{E}_{q(\mathcal{Z})}\left[\log \frac{q(\mathcal{Z})}{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\right]$, which is the Kullback-Leibler (KL) divergence (c.p. appendix B) to be minimized between the true posterior of our hidden variables given the data $p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and the variational approximation of the posterior $q(\mathcal{Z})$, it can be shown that the log evidence, $\log p(\mathcal{Y}|\mathcal{X})$, is bigger than or equal to a quantity $\mathcal{L}(q(\mathcal{Z}))$, thus being a lower bound of the evidence. Often the *evidence lower bound* is abbreviated as ELBO.

The lower bound can be derived from the evidence written as a marginalization over the full joint $p(\mathcal{Z},\mathcal{Y}|\mathcal{X})$ over all hidden variables $\mathcal{Z}$: [6]

$$\begin{aligned}
\log(p(\mathcal{Y}|\mathcal{X})) &= \log\left(\int p(\mathcal{Z},\mathcal{Y}|\mathcal{X})d\mathcal{Z}\right) \\
&= \log\left(\int \frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\cdot q(\mathcal{Z})d\mathcal{Z}\right) \\
&= \log\left(\mathrm{E}_{q(\mathcal{Z})}\left[\frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right]\right) \\
\text{(Jensen)} \quad &\geq \mathrm{E}_{q(\mathcal{Z})}\left[\log\left(\frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right)\right] =: \mathcal{L}(q(\mathcal{Z})) = -\,\mathrm{KL}[q(\mathcal{Z})||p(\mathcal{Z},\mathcal{Y}|\mathcal{X})] \qquad (35)
\end{aligned}$$

After the initial reformulation of the evidence by the joint, the variational distribution is inserted in the nominator and denominator in order to reformulate the equation as an expectation w.r.t to the variational distribution. Then the Jensen inequality is applied in order to get a lower bound for the log evidence $\log p(\mathcal{Y}|\mathcal{X})$ in terms of a KL-divergence.

Since the hidden variables are inserted in order to be able to describe the data generating process, the goal is to maximize the lower bound as close as possible to the evidence.

Starting now again with the KL-divergence between posterior $p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and its desired approximation $q(\mathcal{Z})$, which one wants to minimize, the quantity $\mathcal{L}$ also appears and can be interpreted once more as a lower bound since the KL-divergence is always non-negative:

$$\begin{aligned}
\mathrm{KL}[q(\mathcal{Z})||p(\mathcal{Z}|\mathcal{X},\mathcal{Y})] &= \mathrm{E}_q\left[\log\left(\frac{q(\mathcal{Z})}{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\right)\right] \qquad \text{, note: } \mathrm{E}_{q(\mathcal{Z})}[\cdot] = \mathrm{E}_q[\cdot] \\
&= \mathrm{E}_q[\log q(\mathcal{Z})] - \mathrm{E}_q[\log p(\mathcal{Z},\mathcal{Y}|\mathcal{X})] + \log p(\mathcal{Y}|\mathcal{X}) \\
&= \log p(\mathcal{Y}|\mathcal{X}) - E_q\left[\log\frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right] \\
&= \log p(\mathcal{Y}|\mathcal{X}) - \mathcal{L}(q(\mathcal{Z})) \qquad (36)
\end{aligned}$$

The posterior is reformulated as the fraction of the joint and the evidence $p(\mathcal{Z}|\mathcal{Y},\mathcal{X}) = \frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{p(\mathcal{Y}|\mathcal{X})}$ in order to write the divergence in terms of the ELBO. As just shown, the evidence

[6] in US American context log equals the natural logarithm: log = ln

is always greater than the lower bound except for the case that the KL divergence equals zero, which would signify that the approximated posterior distribution and the posterior are identical.

One can then state the minimization of the divergence between these two distribution as a maximization of the lower bound w.r.t the variational density:

$$\begin{aligned} & \arg\min_{q(\mathcal{Z})} \quad \mathrm{KL}[q(\mathcal{Z})\|p(\mathcal{Z},\mathcal{Y}|\mathcal{X})] \\ \Rightarrow \quad & \arg\min_{q(\mathcal{Z})} \; -\mathcal{L}\left(q(\mathcal{Z})\right) \\ \Rightarrow \quad & \arg\max_{q(\mathcal{Z})} \quad \mathcal{L}\left(q(\mathcal{Z})\right) \end{aligned}$$

Since the KL-Divergence is never negative, the same inequality holds: One maximizes the lower bound in order to get as close as possible to the evidence (likelihood of our data). One gets the minimal divergence, if the lower bound $\mathcal{L}$ equals the log evidence $\log p(\mathcal{Y}|\mathcal{X})$, since the later is the maximum for the lower bound $\mathcal{L}$ as shown in equation 35. Then the variational density $q(Z)$ would equal the true posterior $p(Z|X)$ and thus their KL-divergence would be 0.

The ELBO has to be computed explicitly as a criterion for convergence during the optimization. Therefore it is written down explicitly here. Note that the variational distribution factorizes into the before described (multivariate) random variables. The expectation of a factor (or set) of random variables w.r.t the whole approximate posterior $q(\mathcal{Z})$ (eq. 3) is equivalent to taking the expectation w.r.t the factor in $q(\mathcal{Z})$ which specifies the approximated density of the hidden random variable, e.g. in the case of the factor λ this is $\mathrm{E}_{q(\mathcal{Z})}[p(\lambda)] = \mathrm{E}_{q(\lambda)}[p(\lambda)]$. As abbreviation $\mathrm{E}_{q_\lambda}[p(\lambda)] = \mathrm{E}_{q(\lambda)}[p(\lambda)]$ and $E_{q(\mathcal{Z})}[\cdot] = E_q[\cdot]$ are used:

$$\begin{aligned} \mathcal{L}(q) &= \mathrm{E}_q[\log p(\mathcal{Z},\mathcal{Y}|\mathcal{X})] - \mathrm{E}_q[\log q(\mathcal{Z})] \\ &= \mathrm{E}_{q_\lambda}[\log p(\lambda)] + \mathrm{E}_{q_\lambda,q_a}[\log p(a|\lambda)] + \mathrm{E}_{q_G q_a q_\upsilon}[\log p(G|a,\upsilon,\mathcal{X})] + \mathrm{E}_{q_\upsilon}[\log p(\upsilon)] \\ &+ \mathrm{E}_{q_\omega}[\log p(\omega)] + \mathrm{E}_{q_{b,e}q_\omega}[\log p(e|\omega)] + \mathrm{E}_{q_{b,e}q_\gamma}[\log p(b|\gamma)] + \mathrm{E}_{q_\gamma}[\log p(\gamma)] + \mathrm{E}_{q_\epsilon}[\log p(\epsilon)] \\ &+ \mathrm{E}_{q_{b,e}q_G q_\epsilon}[\log p(\mathcal{Y}|b,e,G,\epsilon)] - \mathrm{E}_{q_\lambda}[\log q(\lambda)] - \mathrm{E}_{q_a}[\log q(a)] - \mathrm{E}_{q_G}[\log q(G)] \\ &- \mathrm{E}_{q_\upsilon}[\log q(\upsilon)] - \mathrm{E}_{q_\gamma}[\log q(\gamma)] - \mathrm{E}_{q_\omega}[\log q(\omega)] - \mathrm{E}_{q_{b,e}}[\log q(b,e)] - \mathrm{E}_{q_\epsilon}[\log q(\epsilon)] \end{aligned} \tag{37}$$

The lower bound contains both expected log distributions of the priors and the expected log variational distributions. The negative expected log of a distribution, where the expectation is taken w.r.t. the distribution of the random variable itself, is in machine learning referred

Table 4: Distributions in joint

$$
\begin{aligned}
p(\lambda) &\overset{(4)}{=} \left(\frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)}\right)^{NT} \left(\prod_{t=1}^{T}\prod_{i=1}^{N_t}\lambda_{t,i}^{\alpha_\lambda-1}\right)\exp\left(-\beta_\lambda\sum_{t=1}^{T}\sum_{i=1}^{N_t}\lambda_{t,i}\right)\\
p(a|\lambda) &\overset{(10)}{=} \prod_{t=1}^{T}\left[\prod_{i=1}^{N_t}\left(\frac{\lambda_{t,i}}{2\pi}\right)^{\frac{1}{2}}\right]\cdot\exp\left\{-\tfrac{1}{2}a_t'\Lambda_t a_t\right\}\text{,where } \Lambda_t=\operatorname{diag}(\lambda_{t,i},\ldots,\lambda_{t,N_t})\\
p(G|a,\upsilon,\mathcal{X}) &\overset{(15)}{=} \prod_{t=1}^{T}\prod_{i=1}^{N_t}\left(\frac{\upsilon_t}{2\pi}\right)^{P/2}\cdot\exp\left\{-\frac{\upsilon_t}{2}(G_{t,i}'-K_{t,i}\cdot a_t)'(G_{t,i}'-K_{t,i}\cdot a_t)\right\},\\
&\quad K_{t,i}=(K_{t,1,i},\ldots,K_{t,P,i})'\in\mathbb{R}^{P\times N_t},\quad G_{t,i}'=(g_{t,1,i,\ldots,g_{t,P,i}})'\in\mathbb{R}^{P}\\
p(\upsilon) &\overset{(18)}{=} \Gamma(\alpha_\upsilon)^{-T}\beta_\upsilon^{T\alpha_\upsilon}\left(\prod_{t=1}^{T}\upsilon_t^{\alpha_\upsilon-1}\right)\cdot\exp\left(-\beta_\upsilon\sum_{t=1}^{T}\upsilon_t\right)\\
p(\omega) &\overset{(26)}{=} \left(\frac{\beta_\omega^{\alpha_\omega}}{\Gamma(\alpha_\omega)}\right)^{P}\left(\prod_{k=1}^{P}\omega_k\right)^{\alpha_\omega-1}\cdot\exp\left(-\beta_\omega\sum_{k=1}^{P}\omega_k\right)\\
p(b,e|\gamma,\omega) &\overset{(30)}{=} (2\pi)^{-(T+P)/2}|\operatorname{diag}(\gamma',\omega')|\cdot\exp\left\{-\tfrac{1}{2}(b',e')\operatorname{diag}(\gamma',\omega')(b',e')'\right\}\\
p(\gamma) &\overset{(21)}{=} \left(\frac{\beta_\gamma^{\alpha_\gamma}}{\Gamma(\alpha_\gamma)}\right)^{T}\left(\prod_{t=1}^{T}\gamma_t^{\alpha_\gamma-1}\right)\exp\left(-\beta_\gamma\sum_{t=1}^{T}\gamma_t\right)\\
p(\epsilon) &\overset{(33)}{=} \Gamma(\alpha_\epsilon)^{-T}\beta_\epsilon^{T\alpha_\epsilon}\left(\prod_{t=1}^{T}\epsilon_t^{\alpha_\epsilon-1}\right)\cdot\exp\left(-\beta_\epsilon\sum_{t=1}^{T}\epsilon_t\right)\\
p(\mathcal{Y}|b,e,G,\epsilon) &\overset{(31)}{=} \frac{(\prod_{t=1}^{T}\epsilon_t N_t)^{1/2}}{(2\pi)^{N/2}}\cdot\exp\left\{-\tfrac{1}{2}\left([y'-(b',e')B']\cdot\operatorname{diag}(\{I_{N_t}\epsilon_t\}_{t=1}^{T})\cdot[y-B(b',e')']\right)\right\}\\
&\quad ,B=(\operatorname{diag}(\{\mathbb{1}_{N_t}\}_{t=1}^{T})\vdots G)\in\mathbb{R}^{N\times(T+P)}
\end{aligned}
$$

Note: The equation numbers are given above the equal sign.

to as entropy or average information. Since extremely unlikely events hold more information in the sense that their occurrence is more important to humans than recurrent events. The entropy of a distribution represents this information (Bishop, 2006, p. 49). In the next Section the lower bound is given explicitly.

2.4 Lower bound of the model

After deriving the full conditionals and parameter updates in Section 2.2 the ELBO in 37 can be calculated specifically as a the convergence criterion of the algorithm.

As prior and approximating densities only univariate gamma and multivariate normal distributions are needed. Since the basic steps repeat, it suffices to derive the log expectation of the prior density for λ and $G_{t,i}$ and the log expectation of the variational density (its entropy or average information) for the random variables λ and a in the approximated posterior. In order to be able to compare both parts of the ELBO, both expectations for the factor λ are given first.

In the next Section the optimal approximation of the distribution of a factor will be stated in terms of the full conditional. Updates will be obtained as an expectation w.r.t the approximated posterior $q(\mathcal{Z})$ of the parameters of the full conditionals. In order to be able to

distinguish the difference between the parameters of the full conditional and parameters of the approximated distribution, the parameters of the approximated distribution will have a star: The parameters of the variational distribution for $\lambda_{t,i}$ are then $\alpha^{\star}_{\lambda_{t,i}}$ and $\beta^{\star}_{\lambda_{t,i}}$, which will be made explicit in Section 2.5 and 2.6. For the first term in the ELBO $\mathcal{L}$, one can just take the logarithm of the prior distribution for λ which is governed by the hyperparameters α_λ and β_λ (see Table 4). Taking the natural logarithm and then expectations w.r.t. $q(\lambda) = q_\lambda$ yields:

$$\begin{aligned}
\mathrm{E}_{q_\lambda}[\log p(\lambda)] &= \mathrm{E}_{q_\lambda}\left[\sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha_\lambda) + \alpha_\lambda \cdot \log\beta_\lambda + (\alpha_\lambda - 1)\cdot\log\lambda_{t,i} - \beta_\lambda\lambda_{t,i}\right] \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha_\lambda) + \alpha_\lambda \cdot \log\beta_\lambda + (\alpha_\lambda - 1)\cdot\mathrm{E}_{q_\lambda}[\log\lambda_{t,i}] - \beta_\lambda\,\mathrm{E}_{q_\lambda}[\lambda_{t,i}]
\end{aligned}$$

Now the entropy for λ is calculated, where the variational distribution of $\lambda_{t,i}$ has shape and rate parameter $\alpha^{\star}_{\lambda_{t,i}}$ and $\beta^{\star}_{\lambda_{t,i}}$:

$$\begin{aligned}
\mathrm{E}_{q_\lambda}[\log q(\lambda)] &= \mathrm{E}_{q_\lambda}\left[\sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha^{\star}_{\lambda_{t,i}}) + \alpha^{\star}_{\lambda_{t,i}}\log\beta^{\star}_{\lambda_{t,i}} + \left(\alpha^{\star}_{\lambda_{t,i}} - 1\right)\log\lambda_{t,i} - \beta^{\star}_{\lambda_{t,i}}\lambda_{t,i}\right] \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha^{\star}_{\lambda_{t,i}}) + \alpha^{\star}_{\lambda_{t,i}}\log\beta^{\star}_{\lambda_{t,i}} + \left(\alpha^{\star}_{\lambda_{t,i}} - 1\right)\mathrm{E}[\log\lambda_{t,i}] - \beta^{\star}_{\lambda_{t,i}}\,\mathrm{E}[\lambda_{t,i}]
\end{aligned}$$

The expectation operator only applies in both log-distributions to the random variables $\log(\lambda_{t,i})$ and $\lambda_{t,i}$ which are in the set of independent random variables $\{\lambda_{t,i}\}_{t=1}^{T}$. This set is, as stated before, represented by the vector λ. The log transformed gamma distributed random variable $\log(\lambda^{\star}_{t,i})$ has as mean $\mathrm{E}\left[\log\lambda^{\star}_{t,i}\right] = \psi\left(\alpha^{\star}_{\lambda_{t,i}}\right) - \log\beta^{\star}_{\lambda_{t,i}}$ and $\lambda^{\star}_{t,i}$ has as mean $\mathrm{E}\left[\lambda^{\star}_{t,i}\right] = \frac{\alpha^{\star}_{\lambda_{t,i}}}{\beta^{\star}_{\lambda_{t,i}}}$ (c.p. appendix A.1 or Bishop (2006, p. 688)). $\psi(\tau) = \frac{\partial}{\partial\tau}\log\Gamma(\tau) = \frac{\Gamma'(\tau)}{\Gamma(\tau)}$ is the digamma function. Inserting these results give

$$\mathrm{E}_{q_\lambda}[\log p(\lambda)] = \sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha_\lambda) + \alpha_\lambda\cdot\log\beta_\lambda + (\alpha_\lambda - 1)\cdot\left(\psi(\alpha^{\star}_{\lambda_{t,i}}) - \log\beta^{\star}_{\lambda_{t,i}}\right) - \beta_\lambda\cdot\frac{\alpha^{\star}_{\lambda_{t,i}}}{\beta^{\star}_{\lambda_{t,i}}}$$

and

$$\begin{aligned}
\mathrm{E}_{q_\lambda}[\log q(\lambda)] =& \sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha^{\star}_{\lambda_{t,i}}) + \alpha^{\star}_{\lambda_{t,i}}\log\beta^{\star}_{\lambda_{t,i}} \\
&+ \left(\alpha^{\star}_{\lambda_{t,i}} - 1\right)\left(\psi(\alpha^{\star}_{\lambda_{t,i}}) - \log\beta^{\star}_{\lambda_{t,i}}\right) - \beta^{\star}_{\lambda_{t,i}}\cdot\frac{\alpha^{\star}_{\lambda_{t,i}}}{\beta^{\star}_{\lambda_{t,i}}} \\
=& \sum_{t=1}^{T}\sum_{i=1}^{N_t} -\log\Gamma(\alpha^{\star}_{\lambda_{t,i}}) + \left(\alpha^{\star}_{\lambda_{t,i}} - 1\right)\psi(\alpha^{\star}_{\lambda_{t,i}}) + \log\beta^{\star}_{\lambda_{t,i}} - \alpha^{\star}_{\lambda_{t,i}}. \qquad (38)
\end{aligned}$$

Since α_λ and β_λ are fixed hyperparameters [7] , the first two terms in the expected log of the prior distribution are constant.

[7] Updating hyperparameters using the ELBO is explained later in Section 2.6.2

Now the multivariate normal distribution for the rows of the matrix of intermediate outputs G_t is considered. The density of row $G'_{t,i}$ as a column vector can be found on page 15 or in Table 4. The expected log of the density of all intermediate outputs is:

$$\mathrm{E}_q[\log p(G|\mathcal{X}, a, \upsilon)] = \sum_{t=1}^{T} \mathrm{E}_q\left[\log p(G_t|\mathcal{X}_t, a_t, v_t)\right] = \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mathrm{E}_q[\log p(G'_{t,i}|K_{t,i}, a_t, v_t)]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{N_t} \mathrm{E}_q\left[\tfrac{P}{2}\cdot\big(\log v_t - \log(2\pi)\big) - \frac{v_t}{2}(G'_{t,i} - K_{t,i}\cdot a_t)'(G'_{t,i} - K_{t,i}\cdot a_t)\right]$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{N_t} \left(\tfrac{P}{2}\cdot\big(\mathrm{E}_q[\log v_t] - \log(2\pi)\big) - \frac{1}{2}\mathrm{E}_q\left[v_t\{G_{t,i}G'_{t,i} - 2a'_tK_{t,i}G'_{t,i} + a'_tK'_{t,i}K_{t,i}a_t\}\right]\right)$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{N_t} \left(\tfrac{P}{2}\big(\mathrm{E}_q[\log v_t] - \log(2\pi)\big) - \frac{1}{2}\mathrm{E}_q\left[v_t\{G_{t,i}G'_{t,i} - 2\,\mathrm{tr}(K_{t,i}G'_{t,i}a'_t) + \mathrm{tr}(K'_{t,i}K_{t,i}a_ta'_t)\}\right]\right)$$

The trace on the scalar values is applied in order to group the squares of the random variables $G'_{t,i}a'_t$ and $a_ta'_t$. Since the kernel matrices are assumed to be fixed (non-stochastic), the expectations only apply to the independent random variables. Independence of random variables implies that the expectation of their product is just the product of expectations of each random variable.

$$\mathrm{E}_q[\log p(G|\mathcal{X}, a, \upsilon)] = \sum_{t=1}^{T} N_t\tfrac{P}{2}\cdot(\mathrm{E}_{q_\upsilon}[\log v_t] - \log(2\pi))$$

$$-\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{1}{2}\Big[\,\mathrm{E}_{q_\upsilon}[v_t]\left\{\mathrm{E}_{q_{G_{t,i}}}[G_{t,i}G'_{t,i}] - 2\,\mathrm{tr}\left(K_{t,i}\,\mathrm{E}_{q_{G_{t,i}}}[G'_{t,i}]\,\mathrm{E}_{q_{a_t}}[a'_t]\right) + \mathrm{tr}\left(K'_{t,i}K_{t,i}\,\mathrm{E}_{a_t}[a_ta'_t]\right)\right\}\Big]$$

The expected outer product of the column a_t can be expressed with the help of the Variance-Covariance:

$$\mathrm{Var}(a_t) = \mathrm{Cov}(a_t, a'_t) \overset{8}{=} \mathrm{E}_q[a_ta'_t] - \mathrm{E}_q[a_t]\,\mathrm{E}[a'_t]$$

$$\Rightarrow \quad \mathrm{E}_q[a_ta'_t] = \mathrm{Var}(a_t) + \mathrm{E}_q[a_t]\,\mathrm{E}_q[a'_t] = \Sigma^{\star}_{a_t} + \mu^{\star}_{a_t}\mu^{\star\prime}_{a_t} \tag{39}$$

The expected inner product of the vector $G'_{t,i}$ is the sum over the squared expectation of a single random variable $g_{t,k,i}$. Therefore it also holds that $\mathrm{E}[G_{t,i}G'_{t,i}] = \sum_{k=1}^{P}\big(\mathrm{E}[g_{t,k,i}]^2 + \mathrm{Var}[g_{t,k,i}]\big) = \sum_{k=1}^{P}\left(\left(\mu^{\star}_{g_{t,k,i}}\right)^2 + \sigma^{2\star}_{g_{t,k,i}}\right) = \mathrm{tr}\left[\left(\mu^{\star}_{G'_{t,i}}\mu^{\star\prime}_{G'_{t,i}} + \Sigma^{\star}_{G'_{t,i}}\right)\right]$, which gives us all the expectations needed to calculate this term of the ELBO [9].

The prior distribution of a is a product of multivariate normal distributions of the multivariate a_t, which are the kernel weights for drug t. Recall that although all random variables

[8] $\mathrm{Var}_q(a_t) = \mathrm{Cov}_q(a_t, a'_t) = \mathrm{E}_q[(a_t - \mathrm{E}_q[a_t])(a_t - \mathrm{E}_q[a_t])'] = \mathrm{E}_q[a_ta'_t] - 2\,\mathrm{E}_q[a_t]\,\mathrm{E}_q[a'_t] + \mathrm{E}_q[a_t]\,\mathrm{E}_q[a'_t]$

[9] One can show that the sum over all cell lines i for drug t is equivalent to the trace of the expected squared G_t, i.e. $\sum_{i=1}^{N_t}\mathrm{E}_q[G_{t,i}G'_{t,i}] = \mathrm{tr}\left(\mathrm{E}_q[G'_tG_t]\right)$ (see Table 1).

$a_{t,i}, i = 1, \ldots, N_t$ contained in a_t are independent, their multivariate representation is needed for calculating intermediate outputs (c.p. Section 2.2.2). Its expected log is

$$\begin{aligned}
\mathrm{E}_q\left[\log p(a|\lambda)\right] = \mathrm{E}_{q_a q_\lambda}\left[\log p(a|\lambda)\right] &= \mathrm{E}_{q_a q_\lambda}\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t}\frac{1}{2}\left(\log\lambda_{t,i} - \log(2\pi)\right)\right) - \frac{1}{2}\operatorname{tr}\left(\Lambda_t a_t a_t'\right)\right] \\
&= \sum_{t=1}^{T}\left(\sum_{i=1}^{N_t}\frac{1}{2}\left(\mathrm{E}_{q_\lambda}\left[\log\lambda_{t,i}\right] - \log(2\pi)\right)\right) - \frac{1}{2}\operatorname{tr}\left(\mathrm{E}_{q_\lambda}\left[\Lambda_t\right]\mathrm{E}_{q_a}\left[a_t a_t'\right]\right) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{1}{2}\left(\mathrm{E}_{q_{\lambda_{t,i}}}\left[\log\lambda_{t,i}\right] - \log(2\pi) + \mathrm{E}_{q_{\lambda_{t,i}}}\left[\lambda_{t,i}\right]\cdot\mathrm{E}_{q_{a_{t,i}}}\left[a_{t,i}^2\right]\right),
\end{aligned}$$

since $a_t'\Lambda_t a_t = \sum_{i=1}^{N_t}\lambda_{t,i}a_{t,i}^2$. Again one has $\mathrm{E}_{q_{\lambda_{t,i}}}[\log\lambda_{t,i}] = \psi(\alpha^\star_{\lambda_{t,i}}) - \beta^\star_{\lambda_{t,i}}$, $\mathrm{E}_{q_{\lambda_{t,i}}}[\lambda_{t,i}] = \frac{\alpha^\star_{\lambda_{t,i}}}{\beta^\star_{\lambda_{t,i}}}$, and $\mathrm{E}_{q_{a_{t,i}}}[a_{t,i}^2] = \left(\mathrm{E}_{q_{a_{t,i}}}[a_{t,i}]\right)^2 + \mathrm{Var}_{q_{a_{t,i}}}[a_{t,i}] = \left(\mu^\star_{a_{t,i}}\right)^2 + \sigma^{\star 2}_{a_{t,i}}$. In multivariate formulation it corresponds to $\mathrm{E}_q[a_t a_t'] = \mathrm{E}_q[a_t]\,\mathrm{E}_q[a_t'] + \mathrm{Var}_q[a_t] = \mu^\star_{a_t}\mu^{\star\prime}_{a_t} + \Sigma^\star_{a_t}$, $\mathrm{E}_q[\log\lambda_{t,i}] = \psi(\alpha^\star_{\lambda_{t,i}}) - \beta^\star_{\lambda_{t,i}}$ and $\mathrm{E}_q[\Lambda] = \operatorname{diag}\left(\frac{\alpha^\star_{\lambda_{t,1}}}{\beta^\star_{\lambda_{t,1}}}, \ldots \frac{\alpha^\star_{\lambda_{t,N_t}}}{\beta^\star_{\lambda_{t,N_t}}}\right)$:

$$\mathrm{E}_{q_a q_\lambda}[\log p(a|\lambda)] = \sum_{t=1}^{T}\left(\sum_{i=1}^{N_t}\frac{1}{2}\left(\psi(\alpha^\star_{\lambda_{t,i}}) - \beta^\star_{\lambda_{t,i}} - \log(2\pi)\right) - \frac{1}{2}\frac{\alpha^\star_{\lambda_{t,i}}}{\beta^\star_{\lambda_{t,i}}}\left(\mu^{\star 2}_{a_{t,i}} + \sigma^{\star 2}_{a_{t,i}}\right)\right)$$

However, the input weights for the variational approximation of G_t have to be specified in multivariate form of a_t. Therefore the results for the log expectation of the multivariate normal distributon of the random variable of weights a_t for drug t in terms of its variational mean $\mu^\star_{a_t}$ and variance $\Sigma^\star_{a_t}$ is

$$\mathrm{E}_q\left[\log(q(a_t))\right] = \mathrm{E}_q\left[-\frac{N_t}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma^\star_{a_t}| - \frac{1}{2}(a_t - \mu^\star_a)'\Sigma^{\star -1}_{a_t}(a_t - \mu^\star_a)\right]. \tag{40}$$

The first two terms are constant w.r.t. to $q(\mathcal{Z})$. The third term can be multiplied out to three terms and the linear expectation operator applied to each of it:

$$\mathrm{E}_q[\log(q(a_t))] = -\frac{N_t}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma^\star_{a_t}| - \frac{1}{2}\left(\mathrm{E}_q\left[a_t'\Sigma^{\star -1}_{a_t}a_t\right] - 2\mu^{\star\prime}_a\Sigma^{\star -1}_{a_t}\,\mathrm{E}_q[a_t] + \mu^{\star\prime}_a\Sigma^{\star -1}_{a_t}\mu^\star_a\right)$$

The last term is constant. The term before only involves the mean of the random variable, which is $E_q[a_t] = \mu^\star_{a_t}$. Therefore only the quadratic form of the random variable has to be considered in detail. The trace operator on the scalar and the trace property of invariance to cyclic permutations is applied and yields:

$$\mathrm{E}_q[a_t'\Sigma^{\star -1}_{a_t}a_t] = \mathrm{E}_q[\operatorname{tr}(a_t'\Sigma^{\star -1}_{a_t}a_t)] = \mathrm{E}_q[\operatorname{tr}(\Sigma^{\star -1}_{a_t}a_t a_t')] = \operatorname{tr}(\Sigma^{\star -1}_{a_t}\,\mathrm{E}_q[a_t a_t'])$$

Now the result for the squared random variable a_t from before is inserted (eq. 39):

$$\begin{aligned}
\mathrm{E}_q[a_t'\Sigma^{\star -1}_{a_t}a_t] &= \operatorname{tr}\left(\Sigma^{\star -1}_{a_t}\left(\mathrm{E}_q[a_t]E_q[a_t'] + \Sigma^{\star -1}_{a_t}\right)\right) \\
&= \operatorname{tr}\left(\Sigma^{\star -1}_{a_t}\,\mathrm{E}_q[a_t]\,\mathrm{E}_q[a_t']\right) + \operatorname{tr}\left(I_{N_t}\right) \\
&= \mathrm{E}_q[a_t']\,\Sigma^{\star -1}_{a_t}\,\mathrm{E}[a_t] + N_t
\end{aligned} \tag{41}$$

Therefore the expectation of the log of the density of all input weights a is

$$\begin{aligned} \mathrm{E}_q[\log q(a)] &= \sum_{t=1}^{T} \mathrm{E}_q[\log(q(a_t))] && (42) \\ &= \sum_{t=1}^{T} \left(-\frac{N_t}{2} \log(2\pi) - \frac{1}{2} \log \left|\Sigma_{a_t}^{\star}\right| - \frac{1}{2} N_t \right) \\ &= \sum_{t=1}^{T} \left(-\frac{N_t}{2} \big(\log(2\pi) + 1 \big) - \frac{1}{2} \log \left|\Sigma_{a_t}^{\star}\right| \right) . && (43) \end{aligned}$$

Note that the entropy of the normal distributed a_t does not depended on the mean.

The remaining terms of the ELBO in equation (37) are stated in the appendix in Table 25 explicitly and in Table 26 using expectation on random variables.

2.5 Parameter updating

Up to now we have assumed that an approximation $q(\mathcal{Z})$ for the posterior $p(\mathcal{Z}|\mathcal{Y}, \mathcal{X})$ exists, without deriving its exact form. As stated before, the full conditionals have been introduced first in order to familiarize the model and as they are integral for finding a solution. Updating the parameters is simplified in the case of a fully conjugated model where the full conditionals of the hidden random variables stay in the same class of exponential family distribution as the prior. In the following it is shown that in these models the update for a parameter $\upsilon_j^{\star}$ governing the variational distribution of the random variable j reduces to $\upsilon_j^{\star} = E_{q_{-z_j}} \left[\eta_j \left(x, \mathcal{Z}_{-z_j} \right) \right]$, which is the expectation w.r.t. random variables different from z_j, i.e. $\mathcal{Z}_{-z_j} = \{ \mathcal{Z} \setminus z_{-j} \}$, of the (natural) parameters η_j of the full conditional of z_j.

In the approximated, or variational, distribution the latent variables are assumed to be independent from each other, which is why they are called factors. Therefore, the expectation w.r.t the latent variables' approximated distribution can be applied to each hidden random variable separately. By definition of a probability distribution, all distributions, except the one for the hidden random variable of interest, just integrate to one[10].

Finding the optimal $q(z_j)$ can be seen as a problem of functional analysis, where one tries to find the optimal function from an infinite space of functions in order to avoid imposing any restrictions on the form of $q(z_j)$ (Bishop, 2006; Sasane, 2017). However, choosing the exponential family for prior (or model) distributions simplifies the approximation significantly.

First, the proof outlined in Hoffman et al. (2013) is presented. Variables specific to data points, which is needed for topic models considered by Hoffman et al. (2013), are denoted by the vector of hidden local variables $z = (z_1, \ldots, z_N)$ and the vector of global hidden variables

[10] Recall the example from the previous Section for the factor λ: $\mathrm{E}_{q(\mathcal{Z})}[p(\lambda)] = \mathrm{E}_{q(\lambda)}[p(\lambda)]$

as $\beta = (\beta_1, \ldots, \beta_K)$. The local or global distinction is not important for this paper, but is kept for reference. Important is only to have two sets of hidden random variables. Note that the notation has been used as in Blei et al. (2017, p. 868 f.). Thus, the hidden variables are in the set $\mathcal{Z} = \{\beta, z\}$ and the observed data is the vector x (i.e. $\mathcal{D} = x$). The ELBO in general notation is then

$$\begin{aligned}\mathcal{L}(q(\mathcal{Z})) = \mathcal{L}\left(q(\beta, z)\right) &= \mathrm{E}_q[\log p(\beta, z, x)] - \mathrm{E}_q[\log q(\beta, z)] \qquad , \mathrm{E}_q[\cdot] = \mathrm{E}_{q(\mathcal{Z})}[\cdot] \\ &= \mathrm{E}_q[\log p(\beta, z|x)] + \mathrm{E}_q[\log p(x)] - \mathrm{E}_q[\log q(\beta, z)]\end{aligned}$$

For now it is assumed that in the fully conjugated model the variational distribution is in the same family as the full conditional, i.e. that the sufficient statistics $t(\beta)$ are the same and they are part of the exponential family. The distributional assumptions will be proven in the next step. The density of the variational distribution is set to the form $q(z, \beta) = q(\beta|\lambda) \cdot \prod_{i=1}^{N} q(z_i|\phi_i) = q(\beta) \cdot q(z)$, which is slightly simpler than the one considered in Hoffman et al. (2013) in order to facilitate the derivation.

λ are the natural parameters of the global variables β, and $\phi_{1:n}$ are the natural parameters of the local variables $z_{1:n}$ in this variational distribution. The ELBO can then be expressed in terms of the natural parameters λ of the variational distribution[11] and the derivative w.r.t them:

$$\begin{aligned}\mathcal{L}(\lambda, \phi_{1:n}) &= \mathrm{E}_q[\log p(\beta, z, x)] - \mathrm{E}_q[\log q(\beta, z;\, \lambda, \phi_{1:n})] \\ &= \mathrm{E}_q[\log p(\beta|z, x) + \log p(z, x)] - \mathrm{E}_q[\log q(\beta, z;\, \lambda, \phi_{1:n})] \\ \Rightarrow \quad \nabla_\lambda \mathcal{L} &= \nabla_\lambda \mathrm{E}_q[\log p(\beta|z, x)] + \underbrace{\nabla_\lambda E_q[\log p(z, x)]}_{=0 \text{ since no } \beta} - \underbrace{\nabla_\lambda E_q[\log q(\beta, z;\, \lambda, \phi_{1:n})]}_{\nabla_\lambda \mathrm{E}_{q_\beta}[q(\beta)]}\end{aligned}$$

The relation between the full conditional and the variational approximations is achieved by the equality between the derivative of the log normalizer and the expectation of the sufficient statistics of exponential family distribution, i.e. $\nabla_\lambda a(\lambda) = \mathrm{E}_{q_\beta}[t(\beta)]$. Given in exponential family notation (c.p. A.1.2 in the appendix), $p(\beta|z, x) = h(\beta) \exp\{\eta'(z, x) \cdot t(\beta) - a\left(\eta(z, x)\right)\}$ is the full conditional of β given the other latent variables z and the observed data x, where this conditional distribution of β is governed by the natural parameters $\eta(z, x) = \eta_\beta$. The

[11] Here the dependence on other random variables and the natural parameters is distinguished by using a semicolon for the later. Note that the natural parameters can be formed by random variables itself (see next Section 2.6).

identity is now used to infer the update:

$$
\begin{aligned}
&= \nabla_\lambda \, \mathrm{E}_q \left[\log h(\beta) + \eta(z,x)' t(\beta) - a(\eta(z,x)) \right] - \nabla_\lambda \, \mathrm{E}_q \left[\log h(\beta) + \lambda' t(\beta) - a(\lambda) \right] \\
&= \nabla_\lambda (\mathrm{E}_{q_{-\beta}}[\eta(z,x)]' \, \mathrm{E}_{q_\beta}[t(\beta)] - \nabla_\lambda (\lambda' \, \mathrm{E}_{q_\beta}[t(\beta)] - a(\lambda)) \qquad , \lambda, a(\lambda) = const. \\
&= \nabla_\lambda (\mathrm{E}_{q_{-\beta}}[\eta(z,x)]' \nabla_\lambda a(\lambda)) - \nabla_\lambda (\lambda' \nabla_\lambda a(\lambda) - a(\lambda)) \qquad \text{,where } \mathrm{E}_{q_\beta}[t(\beta)] = \nabla_\lambda a(\lambda) \\
&= \mathrm{E}_{q_{-\beta}}[\eta(z,x)]' \nabla_\lambda^2 a(\lambda) - \nabla_\lambda a(\lambda) - \lambda' \nabla_\lambda^2 a(\lambda) + \nabla_\lambda a(\lambda) \\
&= (\mathrm{E}_{q_{-\beta}}[\eta(z,x)] - \lambda])' \, \nabla_\lambda^2 a(\lambda) = (E_{q_{-\beta}}[\eta(z,x)] - \lambda])' \, \nabla_\lambda^2 a(\lambda)
\end{aligned}
$$

In the first line the base measure $\log h(\beta)$ and the log-normalizer $a(\eta(z,x))$ of the full conditional of β do not depend on the natural parameter λ of the variational distribution and can be considered constant after taking expectations w.r.t λ. In the last line the expectation on the natural parameter of the full conditional $\eta(z,x)$ do not depend on the random variable β, i.e. $\mathrm{E}_q[\eta(z,x)] = \mathrm{E}_{q_{-\beta}}[\eta(z,x)] = \mathrm{E}_{q_z}[\eta(z,x)]$. If the derivative is then set to zero, the result is

$$
\begin{aligned}
\nabla_\lambda \mathcal{L} = \nabla_\lambda^2 a(\lambda) \cdot (\mathrm{E}_q[\eta(z,x)] - \lambda) &\overset{!}{=} 0 \\
\Leftrightarrow \quad \lambda = \mathrm{E}_q[\eta(z,x)] \qquad &(= \mathrm{E}_{q_{-\beta}}[\eta(z,x)] = \mathrm{E}_{q_z}[\eta(z,x)])
\end{aligned}
$$

After deriving the optimal solution for one vector of latent variables in an abstract model using coordinate ascent, a different proof is chosen with the model at hand, where the functional relation between the full conditionals and the variational densities will be established directly. However, the first approach is important as it easily yields the so called natural gradient used for stochastic optimization, i.e. only using mini-batches of the possibly huge data sets (Hoffman et al., 2013, p. 1314- 1318). Blei et al. (2017, p. 869f.) give a brief summary on stochastic coordinate ascent in Subsection 4.3 of their review on variational inference.

The general idea of the second type of proof is presented in Blei et al. (2017, p. 10) or Bishop (2006, ch. 10). Starting again from the ELBO given in equation (35):

$$
\mathcal{L}(q(\mathcal{Z})) = \mathrm{E}_q[\log p(\mathcal{Z}, \mathcal{Y} | \mathcal{X})] - \mathrm{E}_q[\log q(\mathcal{Z})] \qquad , \mathrm{E}_q[\cdot] = \mathrm{E}_{q(\mathcal{Z})}[\cdot]
$$

Taking the lower bound, consider only the density of a set of random variables z_j w.r.t which one wants to maximize the ELBO. The solution can be identified directly:

$$
\begin{aligned}
\mathcal{L}(q) &= \mathrm{E}_q[\log p(z_j | \mathcal{D}, \mathcal{Z}_{-z_j})] + \mathrm{E}_q[\log p(\mathcal{D}, \mathcal{Z}_{-z_j})] - \mathrm{E}_q[\log q(\mathcal{Z})] \\
&= \mathrm{E}_{q_j}[\mathrm{E}_{q_{-j}}[\log p(z_j | \mathcal{D}, \mathcal{Z}_{-z_j})]] - E_{q_j}[\log(q_j(z_j))] + const. \\
&= \mathrm{E}_{q_j} \left[\log \frac{\exp \left\{ \mathrm{E}_{q_{-j}} \left[\log p(z_j | \mathcal{D}, \mathcal{Z}_{-z_j}) \right] \right\}}{q_j(z_j)} \right] + const \\
&= -\mathrm{KL}[q(z_j) || \tilde{p}(z_j | \mathcal{D}, \mathcal{Z}_{-z_j})] + const,
\end{aligned}
$$

where $\tilde{p}(z_j|\mathcal{D}, \mathcal{Z}_{-z_j}) \propto \exp\{\mathrm{E}_{q_{-j}}\left[\log p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j})\right]\}$ is the exponential of the expectation of the log full conditional of z_j w.r.t the variational distribution. The KL-divergence between the approximated distribution q_j and $\tilde{p}(z_j|\mathcal{D}, \mathcal{Z}_{-z_j})$ is minimized when both are the same:

$$q_j(z_j) = \tilde{p}(z_j|\mathcal{D}, \mathcal{Z}_{-z_j}) \propto \exp\left(E_{q_{-j}}[\log(p(z_j|\mathcal{Z}_{-j}, \mathcal{X}, \mathcal{Y}))]\right) \propto \exp\left(E_{q_{-j}}[\log p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})]\right)$$

The variational density of z_j corresponds to the exponential expectation of the log *full conditional* w.r.t. the variational distribution. The variational densities are of the same form as the corresponding full conditionals. In exponential family notation it is easily shown that q_j itself is a density integrating to one by inserting the general exponential family notation of the full conditional $p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j}) = h(z_j) \cdot \exp\{\eta'(\mathcal{D}, \mathcal{Z}_{z_j}) \cdot t(z_j) - a(\eta(\mathcal{D}, \mathcal{Z}_{z_j}))$:

$$\begin{aligned} q_j^\star(z_j) =& \exp\left\{E_{q_{-j}}\left[\log(h(z_j)) + \eta'\left(\mathcal{D}, \mathcal{Z}_{-j}\right) t(z_j) - a\big(\eta(\mathcal{D}, \mathcal{Z}_{-j})\big)\right]\right\} \\ =& h(z_j) \cdot \exp\left\{E_{q_{-j}}\left[\eta'\left(\mathcal{D}, \mathcal{Z}_{-j}\right)\right] t(z_j) - a\big(\eta(\mathcal{D}, \mathcal{Z}_{-j})\big)\right\} \end{aligned}$$

In the factorized or so called (structured) *mean-field*-set-up of VI taking expectations is greatly simplified, since the random variables are assumed independent in the approximated posterior $q(\mathcal{Z}) = \prod_j z_j$. This allows updating the parameters of the variational distribution separately for the sets of random variables z_j.

$$q_j^\star(z_j) \propto \exp\{\mathrm{E}_{-j}[\log p(z_j|z_{-j}, x)]\} = \exp\left(\int \prod_{l \neq j} q(z_l) \log p(z_j|z_{-j}, x)\, dz_l\right)$$

If the conditional probability does not depend on certain $q_l(z_l)$ with $l \neq j$, they integrate to one as a probability density function. If the full conditional $p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j})$ is stated in the exponential family notation, i.e. $p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j}) = h(z_j) \exp\{\eta'(\mathcal{D}, \mathcal{Z}_{-z_j}) \cdot t(z_j) - a\left(\eta(\mathcal{D}, \mathcal{Z}_{-z_j})\right)\}$, the update can be related to the natural parameters instead of the canonical parameters. Since the canonical parameters form the natural parameters as shown in appendix A.1, the optimal updates can be described in both forms. The canonical parameters are the mean and variance in case of a normal or the shape and rate parameters in case of a gamma distribution.

Note that the dependence of the variational densities on full conditionals introduces an overall dependence of the model on the hyperparameters.

Updating the parameters in the ELBO is done iteratively until a convergence criterion of relative change is met. The variational parameters have to be initialized randomly at the start as they are needed computing an update. The variational parameters are marked with a star, e.g. $\eta_{z_j}^\star$, which corresponds to an update at a certain iteration.

2.6 Parameter updates for the proposed model

After deriving the updates of the optimal approximated distribution for the groups of hidden random variables, their specific parameter updates are stated, before updating the hyperparameters is considered.

2.6.1 Updating the distributional parameters of the factors

The approximated density function for all gamma distributed latent variables λ is denoted as $q(\lambda) = \prod_{t=1}^{T}\prod_{i=1}^{N_t} \mathcal{G}am\Big(\lambda_{t,i}|\, \alpha^{\star}_{\lambda_{t,i}}, \beta^{\star}_{\lambda_{t,i}}\Big)$, which is defined by the set of distributional parameters of each univariate full conditional. The resulting approximations can be obtained, as it is shown in the previous Section 2.5, by taking the expectations w.r.t to the variational distribution of the parameters of the full conditionals, which are for each $\lambda_{t,i}$ shape $\alpha(\lambda_{t,i}) = \alpha_{\lambda_{t,i}}$ and rate $\beta(\lambda_{t,i}) = \beta_{\lambda_{t,i}}$:

$$\begin{aligned} q(\lambda_{t,i}) &= \exp \mathrm{E}_q\left[\log p(\lambda_{t,i}|\mathcal{D}, \mathcal{Z}_{-\lambda_{t,i}})\right] \\ &= \exp \mathrm{E}_q\left[\log \mathcal{G}am(\lambda_{t,i}|\, \alpha_{\lambda_{t,i}}, \beta_{\lambda_{t,i}})\right] = \mathcal{G}am\big(\lambda_{t,i}|\ \mathrm{E}_q[\alpha_{\lambda_{t,i}}], \mathrm{E}_q[\beta_{\lambda_{t,i}}]\big)\,. \end{aligned}$$

Made explicitly this is

$$q(\lambda) = \prod_{t=1}^{T}\prod_{i=1}^{N_t} \mathcal{G}am\left(\lambda_{t,i}\middle|\, \alpha_\lambda + \frac{1}{2}\,,\, \frac{1}{2}\cdot \mathrm{E}_{q(a_{t,i})}\left[a^2_{t,i}\right]\right)$$

where $E_{q_{-\lambda_{t,i}}}[a^2_{t,i}] = E_q[a^2_{t,i}] = E_{q_{a_{t,i}}}[a^2_{t,i}] = E[a^2_{t,i}]$ and thus after the algebraic reformulation in terms of mean and variance[12] $E_{q_{a_{t,i}}}[a^2_{t,i}] = \mathrm{E}[a_{t,i}]^2 + \mathrm{Var}[a_{t,i}] = \mu^{\star 2}_{a_{t,i}} + \sigma^{2\star}_{a_{t,i}}$. In order to update the relevant parameters for $\lambda_{t,i}$ only the mean of distribution of weight $a_{t,i}$ and not a realization is needed. Note that the updated shape parameter $\alpha^{\star}_{\lambda_{t,i}} = \alpha^{\star}_{\lambda} = \alpha_\lambda + \frac{1}{2}$ is a fixed value for all iterations and for all $\lambda_{t,i}$. If the shape parameter should be updated, it has to be done by updating the hyperparameter α_λ.

Next the parameter updates for the kernel weights are calculated in the same manner, i.e. taking expectations of the distributional parameters of the full conditional for factor a, which is:

$$p(a|\mathcal{D}, \mathcal{Z}_{-a}) = \prod_{t=1}^{T} \mathcal{N}\left(a_t\,\middle|\, \Sigma_{a_t}\left(\sum_{k=1}^{P} K_{t,k} g_{t,k}\right) v_t\,,\, \left(\Lambda_t + v_t \sum_{k=1}^{P} K_{t,k} K_{t,k}\right)^{-1}\right)$$

The update depends only on the expectations w.r.t to the relevant variational densities $q(\upsilon)$, $q(\lambda)$ and $q(G)$. Firstly one has to update the precision of a_t in order to be able to compute

[12] $\mathrm{E}[X^2] = \mathrm{E}[X]^2 + \mathrm{Var}[X]$ for a random variable X

the update for the mean. It involves $\mathrm{E}_q[\lambda_{t,i}] = \mathrm{E}_{q_{\lambda_{t,i}}}[\lambda_{t,i}] = \alpha^\star_{\lambda_{t,i}}/\beta^\star_{\lambda_{t,i}}$ as the mean of a gamma distribution, thus $\mathrm{E}[\Lambda_t] = \mathrm{diag}(\alpha^\star_{\lambda_{t,1}}/\beta^\star_{\lambda_{t,1}}, \ldots, \alpha^\star_{\lambda_{t,N_t}}/\beta^\star_{\lambda_{t,N_t}})$, and the corresponding mean of υ_t is $\mathrm{E}_{q_\upsilon}[\upsilon_t] = \alpha^\star_{\upsilon_t}/\beta^\star_{\upsilon_t}$. The expected sum $\mathrm{E}_q\left[\sum_{k=1}^P K_{t,k} g_{t,k}\right] = \sum_{k=1}^P K_{t,k}\,\mathrm{E}_q\left[g_{t,k}\right]$ involves the mean of the columns of G_t, i.e. $\mathrm{E}_q[g_{t,k}] = \mu^\star_{g_{t,k}}$. The sum could be vectorized by stacking the kernel matrices $K_t = (K_{t,1}, \ldots, K_{t,P})'$ and the vector of stacked columns of means of G_t, $\mu^\star_{g_t} = (\mu^{\star\prime}_{g_{t,1}}, \ldots, \mu^{\star\prime}_{g_{t,P}})'$. Therefore the sum is equal to $\sum_{k=1}^P K_{t,k}\mu^\star_{g_{t,k}} = K_t' \cdot \mu^\star_{g_t}$.

Finally, a more tedious example is calculated. The full conditional for ϵ is given in equation (34) where the parameter updates depend on the fixed hyperparameters α_ϵ and β_ϵ and the expectation of the norm c_{ϵ_t}.

$$p(\epsilon|\mathcal{D}, \mathcal{Z}_{-\epsilon}) \propto \prod_{t=1}^T \mathcal{G}am\left\{\epsilon_t \,\middle|\, \alpha_\epsilon + \frac{N_t}{2}\,,\, \beta_\epsilon + \frac{c^2_{\epsilon_t}}{2}\right\} \quad , c_{\epsilon_t} = \|y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\| \tag{44}$$

The expectation only applies to the random variables in $c^2_{\epsilon_t}$, therefore:

$$\begin{aligned}
\mathrm{E}_q[c^2_{\epsilon_t}] = \; & E_{q_G q_{b,e}}\|y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\|^2 = \mathrm{E}_{q_G q_{b,e}}\left[(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t})'(y_t - G_t e - b_t \cdot \mathbb{1}_{N_t})\right] \\
= & \,\mathrm{E}_{q_G q_{b,e}}\Big[y_t'y_t - y_t'G_t e - y_t' b_t \cdot \mathbb{1}_{N_t} - (G_t e)'y_t + (G_t e)'G_t e + (G_t e)' b_t \cdot \mathbb{1}_{N_t} \\
& \qquad - b_t \mathbb{1}'_{N_t} y_t + b_t \mathbb{1}'_{N_t} G_t e + b_t \mathbb{1}'_{N_t}\mathbb{1}_{N_t} b_t\Big] \\
= & \,\mathrm{E}_{q_G q_{b,e}}\left[y_t'y_t - 2e'G_t'y_t - 2y_t'\mathbb{1}_{N_t} b_t + e'G_t'G_t e + 2e'G_t'\mathbb{1}_{N_t} b_t + b_t^2 N_t\right] \\
= & \, y_t'y_t - 2\,\mathrm{E}_{q_e}[e']\,\mathrm{E}[G_t']y_t - 2y_t'\mathbb{1}_{N_t}\,\mathrm{E}_{q_{b_t}}[b_t] + \mathrm{tr}\left(\mathrm{E}_{q_{G_t}}[G_t'G_t]\,\mathrm{E}_{q_e}[ee']\right) \\
& + 2\,\mathrm{E}_{q_e}[e']\,\mathrm{E}_{q_{G_t}}[G_t']\mathbb{1}_{N_t}\,\mathrm{E}_{q_{b_t}}[b_t] + \mathrm{E}_{q_{b_t}}[b_t^2]N_t
\end{aligned}$$

The variational distribution is only given over the hidden variables which is why no expectation w.r.t to y_t is applied and one just plugs in the observed data as for the kernels $K_{t,k}$. Most expectations are similar to the one defined above for a_t or $\lambda_{t,i}$, depending only on the type of variational distribution, i.e. $E[b_t] = \mu^{\star 2}_{b_t} + \sigma^{2\star}_{b_t}$ and $\mathrm{E}[ee'] = \mu^\star_e \mu^{\star\prime}_e + \Sigma^\star_e$, which have to be extracted both from their common variational density $q(b, e)$, which is possible as it is a normal distribution (Bishop, 2006, p. 89).

The exception is the expectation of the product of the matrix of intermediate outputs $G_t'G_t$. The expectation of the inner product of an row $G_{t,i}$ of G_t is given by equation (39) on page 30. The equivalence of the matrix product $G_t'G_t$, where $G_t \in \mathbb{R}^{N_t \times P}$ to the sum over cell lines i of outer products of the row vectors $G_{t,i}$ in G_t representing all intermediate outputs for cell line i tested on drug t, i.e. $G_{t,i} = (g_{t,1,i}, \ldots, g_{t,P,i}) \in \mathbb{R}^P$, has to be shown. The column vector $g_{t,k} \in \mathbb{R}^{N_t}$ represents the column k of G_t and contains all intermediate

outputs for all cell lines for input k (c.p. Table 1 on page 4):

$$G_t' G_t = \begin{pmatrix} g_{t,1}' \\ \vdots \\ g_{t,P}' \end{pmatrix} \begin{pmatrix} g_{t,1} & \dots & g_{t,P} \end{pmatrix} = \begin{pmatrix} g_{t,1}' g_{t,1} & \dots & g_{t,1}' g_{t,P} \\ \vdots & \ddots & \vdots \\ g_{t,P}' g_{t,1} & \dots & g_{t,P}' g_{t,P} \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{i=1}^{N_t} g_{t,1,i}^2 & \dots & \sum_{i=1}^{N_t} g_{t,1,i} g_{t,P,i} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N_t} g_{t,P,i} G_{t,1,i} & \dots & \sum_{i=1}^{N_t} g_{t,P,i}^2 \end{pmatrix} = \sum_{i=1}^{N_t} \begin{pmatrix} g_{t,1,i} \\ \vdots \\ g_{t,P,i} \end{pmatrix} \begin{pmatrix} g_{t,1,i} & \dots & g_{t,P,i} \end{pmatrix} = \sum_{i=1}^{N_t} G_{t,i}' G_{t,i}$$

The expected outer product of row vectors is $\mathrm{E}\left[G_{t,i}' G_{t,i}\right] = \mathrm{E}[G_{t,i}'] \left(\mathrm{E}[G_{t,i}']\right)' + \mathrm{Var}[G_{t,i}'] = \mu_{G_{t,i}'}^{\star} \mu_{G_{t,i}'}^{\star\prime} + \Sigma_{G_{t,i}'}^{\star}$, where $\Sigma_{G_{t,i}'}^{\star} = \Sigma_{G_{t,\cdot}'}^{\star}$ is the same for all cell lines. Therefore one has the sum over the outer product of means over all cell lines, which is equivalent to the matrix product of $\mu_{G_t}^{\star\prime} \mu_{G_t}^{\star}$, and N_t times the variance of each $\Sigma_{G_{t,i}'}^{\star}$, i.e. $\mathrm{E}\left[G_t' G_t\right] = \mu_{G_t}^{\star\prime} \mu_{G_t}^{\star} + N_t \Sigma_{G_{t,\cdot}'}^{\star}$.

The remaining updates can be found in Table 23 on page 79.

2.6.2 Updating hyperparameters

The considered BMTMKL model has ten hyperparameters for the gamma distributed random variables which are used as precision for the normally distributed random variables. Hyperparameters have been introduced in the beginning of Section 2 on page 7.

Generally, hyperparameters for the gamma priors are fixed in advance and a optimal set can be determined by running a grid search. However, updates for hyperparameters could also be learned using derivatives of the ELBO w.r.t to these hyperparameters. In this subsection it will become apparent why analytically deriving hyperparameters is only possible using some sort of constraints between variables as the updates of the variational parameters are intertwined. As an example it will be shown that the hyperparamters of the precisions λ for the kernel weights a propagate through the graphical model. This means that several terms of the ELBO depend on the hyperparameters of the gamma distributed random variables which are used as precision for normally distributed variables (see Table 24 or 25 in appendix). Optimizing the hyperparameters would involve log prior and entropy of λ, a and υ as well as the log prior of G. Since the updates of the variational distributions depend on each other.

These difficulties are illustrated by calculating the gradients for the two hyperparameters of $\lambda = \lambda_{1:T,1:N_t}$. Here the shape and scale notation is used in order to be comparable to the implementation of Gönen, who used the scale notation[13]. The expected log prior stated in

[13] In Bishop (2006) β_τ, where $\tau \sim \mathcal{G}am(\alpha_\tau, \beta_\tau)$, denotes the rate parameter, whereas Gönen uses β_τ for the scale parameter, which is in Bishop (2006) described as $\theta_\lambda = \frac{1}{\beta_\lambda}$.

terms of the scale instead of shape parameter, i.e. β_λ is set to $\frac{1}{\theta_\lambda}$, is:

$$\begin{aligned}\mathrm{E}\left[\log p(\lambda_{t,i})\right] =& \sum_{t,i} \Big(-\log\Gamma(\alpha_\lambda) + \alpha_\lambda \log\beta_\lambda + (\alpha_\lambda - 1)\cdot \mathrm{E}_q\left[\log\lambda_{t,i}\right] - \beta_\lambda \cdot \mathrm{E}_q\left[\lambda_{t,i}\right] \Big) \\ =& \sum_{t,i} \left(-\log\Gamma(\alpha_\lambda) - \alpha_\lambda \log\theta_\lambda + (\alpha_\lambda - 1)\cdot \mathrm{E}_q\left[\log\lambda_{t,i}\right] - \frac{1}{\theta_\lambda}\cdot \mathrm{E}_q\left[\lambda_{t,i}\right] \right)\end{aligned}$$

The corresponding variational parameters in scale notation for $\lambda_{t,i}$ have been introduced in Section 2.2.1 and can be found in Gönen (2012a). The variational shape parameters is in both notations the same. In order to use both notations, the variational scale parameter is given along the corresponding version in rate notation. Expectations of the random variable $\lambda_{t,i}$ w.r.t. the variational distributions then:

$$\mathrm{E}_q\left[\lambda_{t,i}\right] = \frac{\alpha^\star_\lambda}{\beta^\star_{\lambda_{t,i}}} = \alpha^\star_\lambda \cdot \theta^\star_{\lambda_{t,i}} \qquad \theta^\star_{\lambda_{t,i}} = \left(\tfrac{1}{\theta_\lambda} + \tfrac{1}{2}\,\mathrm{E}_q\left[a^2_{t,i}\right]\right)^{-1} \qquad \beta^\star_{\lambda_{t,i}} = \beta_\lambda + \tfrac{1}{2}\,\mathrm{E}_q\left[a^2_{t,i}\right]$$

$$\begin{aligned}\mathrm{E}_q\left[\log\lambda_{t,i}\right] =& \psi\left(\alpha^\star_\lambda\right) - \log\beta^\star_{\lambda_{t,i}} = \psi\left(\alpha^\star_\lambda\right) - \log\left(\beta_\lambda + \tfrac{1}{2}\,\mathrm{E}_q\left[a^2_{t,i}\right]\right) \qquad , \alpha^\star_\lambda = \alpha_\lambda + \tfrac{1}{2} \\ =& \psi\left(\alpha^\star_\lambda\right) + \log\theta^\star_{\lambda_{t,i}} = \psi\left(\alpha^\star_\lambda\right) - \log\left(\tfrac{1}{\theta_\lambda} + \tfrac{1}{2}\,\mathrm{E}_q\left[a^2_{\lambda_{t,i}}\right]\right)\end{aligned}$$

Now consider for inner derivatives, i.e. the variational parameters derived w.r.t to both hyperparameters for λ:

$$\frac{\partial\alpha^\star_\lambda}{\partial\alpha_\lambda} = 1 \qquad \frac{\partial\alpha^\star_\lambda}{\partial\theta_\lambda} = 0 \qquad \frac{\partial\theta^\star_{\lambda_{t,i}}}{\partial\alpha_\lambda} = 0 \qquad \frac{\partial\theta^\star_{\lambda_{t,i}}}{\partial\theta_\lambda} = \frac{\left(\theta^\star_{\lambda_{t,i}}\right)^2}{\theta^2_\lambda}$$

Next the outer derivatives, i.e. the expectations of the (log) random variable $\lambda_{t,i}$, are:

$$\begin{aligned}&\frac{\partial}{\partial\alpha_\lambda}\mathrm{E}_q\left[\lambda_{t,i}\right] = \theta^\star_{\lambda_{t,i}} = 1/\beta^\star_{\lambda_{t,i}} && \frac{\partial}{\partial\alpha_\lambda}\mathrm{E}_q\left[\log\lambda_{t,i}\right] = \frac{\partial}{\partial\alpha_\lambda}\psi(\alpha^\star_\lambda) = \psi'(\alpha^\star_\lambda) \\ &\frac{\partial}{\partial\theta_\lambda}\mathrm{E}_q\left[\lambda_{t,i}\right] = \alpha^\star_\lambda \cdot \frac{\left(\theta^\star_{\lambda_{t,i}}\right)^2}{\theta^2_\lambda} && \frac{\partial}{\partial\theta_\lambda}\mathrm{E}_q\left[\log\lambda_{t,i}\right] = \frac{1}{\theta_{\lambda^\star_{t,i}}}\cdot\frac{\partial}{\partial\theta_\lambda}\theta^\star_{\lambda_{t,i}} = \frac{\theta^\star_{\lambda_{t,i}}}{\theta^2_\lambda} \\ &\frac{\partial}{\partial\beta_\lambda}\mathrm{E}_q\left[\lambda_{t,i}\right] = -\alpha^\star_\lambda \cdot \left(\beta^\star_{\lambda_{t,i}}\right)^{-2} && \frac{\partial}{\partial\beta_\lambda}\mathrm{E}_q\left[\log\lambda_{t,i}\right] = -\frac{1}{\beta_{\lambda^\star_{t,i}}}\cdot\frac{\partial}{\partial\beta_\lambda}\beta^\star_{\lambda_{t,i}} = -\frac{1}{\beta_{\lambda^\star_{t,i}}}\end{aligned}$$

$\psi'(\cdot)$ is the trigamma function. Then the derivatives of the expected log prior $\lambda_{t,i}$ w.r.t. to the hyperparameters can be derived with the help of just stated preliminary results:

$$\begin{aligned}\frac{\partial\,\mathrm{E}\left[\log p(\lambda_{t,i})\right]}{\partial\alpha_\lambda} =& N\left(-\psi(\alpha_\lambda) - \log\theta_\lambda\right) \\ &+ (\alpha_\lambda - 1)\cdot\sum_{t,i}\frac{\partial}{\partial\alpha_\lambda}\mathrm{E}_q\left[\log\lambda_{t,i}\right] + \sum_{t,i}\mathrm{E}_q\left[\log\lambda_{t,i}\right] - \frac{1}{\theta_\lambda}\cdot\sum_{t,i}\frac{\partial}{\partial\alpha_\lambda}\mathrm{E}_q\left[\lambda_{t,i}\right] \\ =& N\left(-\psi(\alpha_\lambda) - \log\theta_\lambda\right) \\ &+ (\alpha_\lambda - 1)\cdot\sum_{t,i}\psi'(\alpha^\star_{\alpha_\lambda}) + \sum_{t,i}\left(\psi(\alpha^\star_\lambda) + \log\theta^\star_{\lambda_{t,i}}\right) - \frac{1}{\theta_\lambda}\cdot\sum_{t,i}\theta^\star_{\lambda_{t,i}}\end{aligned}$$

$$\begin{aligned}
\frac{\partial \operatorname{E}\left[\log p(\lambda_{t,i})\right]}{\partial \theta_\lambda} =& \sum_{t,i} \left(-\frac{\alpha_\lambda}{\theta_\lambda} + (\alpha_\lambda - 1) \cdot \frac{\partial}{\partial \theta_\lambda} \operatorname{E}_q\left[\log \lambda_{t,i}\right] + \frac{1}{\theta_\lambda^2} \operatorname{E}_q\left[\lambda_{t,i}\right] - \frac{1}{\theta_\lambda} \cdot \frac{\partial}{\partial \theta_\lambda} \operatorname{E}_q\left[\lambda_{t,i}\right] \right) \\
=& \sum_{t,i} \left(-\frac{\alpha_\lambda}{\theta_\lambda} + (\alpha_\lambda - 1) \cdot \frac{\theta^\star_{\lambda_{t,i}}}{\theta_\lambda^2} + \frac{1}{\theta_\lambda^2} \left(\alpha^\star_\lambda \cdot \theta^\star_{\lambda_{t,i}} \right) - \frac{1}{\theta_\lambda} \cdot \frac{\left(\theta^\star_{\lambda_{t,i}}\right)^2 \alpha^\star_\lambda}{\theta_\lambda^2} \right) \\
=& - N \frac{\alpha_\lambda}{\theta_\lambda} + \frac{\alpha_\lambda - 1}{\theta_\lambda^2} \sum_{t,i} \theta^\star_{\lambda_{t,i}} + \frac{\alpha^\star_\lambda}{\theta_\lambda^2} \sum_{t,i} \theta^\star_{\lambda_{t,i}} - \frac{\alpha^\star_\lambda}{\theta_\lambda^3} \cdot \sum_{t,i} \left(\theta^\star_{\lambda_{t,i}}\right)^2
\end{aligned}$$

Next the entropy of λ is considered, which is easy to state in terms of the scale parameters $\theta^\star_{t,i}$ instead of the rate $\beta^\star_{\lambda_{t,i}}$:

$$\begin{aligned}
-E_q\left[\log q(\lambda)\right] =& \sum_{t,i} \left(\log \Gamma(\alpha^\star_\lambda) + (1 - \alpha^\star_\lambda) \cdot \psi(\alpha^\star_\lambda) + \alpha^\star_\lambda - \log \beta^\star_{\lambda_{t,i}} \right) \\
=& \sum_{t,i} \left(\log \Gamma(\alpha^\star_\lambda) - (\alpha^\star_\lambda - 1) \cdot \psi(\alpha^\star_\lambda) + \alpha^\star_\lambda + \log \theta^\star_{\lambda_{t,i}} \right)
\end{aligned}$$

Using the previous results, the derivatives w.r.t. to both hyperparamters are:

$$\begin{aligned}
\frac{\partial - E_q\left[\log q(\lambda)\right]}{\partial \alpha_\lambda} =& \sum_{t,i} \left(\psi(\alpha^\star_\lambda) - \psi(\alpha^\star_\lambda) - (\alpha^\star_\lambda - 1) \cdot \frac{\partial}{\partial \alpha_\lambda} \psi(\alpha^\star_\lambda) + 1 \right) \\
=& \sum_{t,i} \left((1 - \alpha^\star_\lambda) \cdot \psi'(\alpha^\star_\lambda) + 1 \right) \\
\frac{\partial - E_q\left[\log q(\lambda)\right]}{\partial \theta_\lambda} =& \frac{1}{\theta_\lambda^2} \sum_{t,i} \theta^\star_{\lambda_{t,i}}
\end{aligned}$$

The expected log prior $\operatorname{E}_q[\log p(\lambda)]$ and the entropy $\operatorname{E}_q[\log q(\lambda)]$ depend in their parameters directly on the hyperparameters. Now one has to consider where the moments of the variables in λ are used for defining parameters of other variables. The next term in the ELBO containing the random variables λ is the expected log prior of the kernel weights a (see Table 26):

$$\begin{aligned}
\operatorname{E}_q[\log p(a|\lambda)] =& \frac{1}{2} \sum_{t,i} \left(\operatorname{E}_q\left[\log \lambda_{t,i}\right] - \log(2\pi) - \operatorname{E}_q\left[\lambda_{t,i}\right] \cdot \operatorname{E}_q\left[a^2_{t,i}\right] \right) \\
=& \frac{1}{2} \sum_t \left(\mathbb{1}'_{N_t} \cdot \operatorname{E}_q\left[\log \lambda_t\right] - \frac{N_t}{2} \log(2\pi) - \operatorname{tr}\left[\operatorname{diag}\left(\underbrace{\operatorname{E}_q\left[\lambda_{t,1}\right], \ldots, \operatorname{E}_q\left[\lambda_{t,N_t}\right]}_{=\operatorname{E}_q[\lambda_t]} \right) \cdot \operatorname{E}_q\left[a_t a_t'\right] \right] \right)
\end{aligned}$$

The derivative w.r.t. α_λ and θ_λ do apply to $\operatorname{E}_q\left[\lambda_t\right]$, $\operatorname{E}_q\left[\log \lambda_t\right]$ and $\operatorname{E}_q\left[a_t a_t'\right]$, which would give the derivative using the product rule:

$$\begin{aligned}
\frac{\partial \operatorname{E}_q[\log p(a|\lambda)]}{\partial \alpha_\lambda} =& \frac{1}{2} \sum_t \Bigg(\mathbb{1}'_{N_t} \frac{\partial}{\partial \alpha_\lambda} \operatorname{E}_q\left[\log \lambda_t\right] \\
& - \operatorname{tr}\left[\operatorname{diag}\left(\frac{\partial}{\partial \alpha_\lambda} \operatorname{E}_q\left[\lambda_t\right] \right) \cdot \operatorname{E}_q\left[a_t a_t'\right] + \operatorname{diag}\left(\operatorname{E}_q\left[\lambda_t\right] \right) \cdot \frac{\partial}{\partial \alpha_\lambda} \operatorname{E}_q\left[a_t a_t'\right] \right] \Bigg)
\end{aligned}$$

where $\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[a_t a_t'\right]$ can be calculated using Table 24 in the appendix or the previous Section 2.6.1:

$$\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[a_t a_t'\right] = \frac{\partial}{\partial \alpha_\lambda} \left(\mathrm{E}_q \left[a_t\right] \mathrm{E}_q \left[a_t'\right] + \mathrm{Var}_q \left[a_t\right]\right)$$

The variational mean and variance of a_t are $\mathrm{E}_q \left[a_t\right] = \Sigma_{a_t}^\star \cdot K_t' \cdot \mathrm{E}_q \left[g_t\right] \mathrm{E}_q \left[\upsilon_t\right]$ and $\mathrm{Var}_q \left[a_t\right] = \Sigma_{a_t}^\star = \left(\mathrm{diag}\left(\mathrm{E}_q \left[\lambda_t\right]\right) + \mathrm{E}_q \left[\upsilon_t\right] \cdot \sum_{k=1}^P K_{t,k} K_{t,k}\right)^{-1}$. Here it now becomes apparent that the updates point to each other. $\mathrm{E}_q \left[\upsilon_t\right]$ does involve $\mathrm{E}_q \left[a_t\right]$ and vice versa, as can be seen in Table 24 in the appendix. Continuing to calculate $\frac{\partial \mathrm{E}_q[\log p(a|\lambda)]}{\partial \alpha_\lambda}$ for now, the variational variance $\Sigma_{a_t}^\star = \left(\Lambda_{a_t}^\star\right)^{-1}$ of a_t has to be differentiated using the rule for derivation of inverses $\frac{\partial}{\partial \alpha} \Lambda^{-1} = -\Lambda^{-2} \frac{\partial}{\partial \alpha} \Lambda$ (Petersen and Pedersen, 2012, ch. 2.2) knowing that the inverse exists from previous sections.

$$\frac{\partial}{\partial \alpha_\lambda} \Sigma_{a_t}^\star = -\left(\Sigma_{a_t}^\star\right)^2 \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[\lambda_t\right]\right)$$

The first part of the derivative of the mean of the squared random variable a_t w.r.t to the hyperparameters of λ depends only on the variational variance of a_t:

$$\begin{aligned}
\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[a_t\right] \mathrm{E}_q \left[a_t'\right] &= \frac{\partial}{\partial \alpha_\lambda} \left(\Sigma_{a_t}^\star \cdot K_t' \cdot \mathrm{E}_q \left[g_t\right] \mathrm{E}_q \left[\upsilon_t\right] \cdot \left(\Sigma_{a_t}^\star \cdot K_t' \cdot \mathrm{E}_q \left[g_t\right] \mathrm{E}_q \left[\upsilon_t\right]\right)'\right) \\
&= \left(\frac{\partial}{\partial \alpha_\lambda} \left(\Sigma_{a_t}^\star\right)^2\right) \cdot \underbrace{K_t' \mathrm{E}_q \left[g_t\right] \mathrm{E}_q \left[g_t'\right] K_t \left(\mathrm{E}_q \left[\upsilon_t\right]\right)^2 \mathrm{E}_q \left[\upsilon_t\right]}_{=C_{a_t}}
\end{aligned}$$

K_t is the stacked matrix of input kernels and g_t the stacked vector of intermediate outcomes. C_{a_t} is a $N_t \times N_t$ matrix which is constant w.r.t. the hyperparameter α_λ and θ_λ as it does not depend in any parameters of random variables λ.

$$\begin{aligned}
\frac{\partial}{\partial \alpha_\lambda} \left(\Sigma_{a_t}^\star\right)^2 &= 2 \cdot \Sigma_{a_t}^\star \cdot -\left(\Sigma_{a_t}^\star\right)^2 \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[\lambda_t\right]\right) \cdot C_{a_t} \\
&= -2 \left(\Sigma_{a_t}^\star\right)^3 \cdot C_{a_t} \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[\lambda_t\right]\right)
\end{aligned}$$

Combing the results for $\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[a_t a_t'\right]$ yields

$$\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[a_t a_t'\right] = -\left(2 \left(\Sigma_{a_t}^\star\right)^3 \cdot C_{a_t} + \left(\Sigma_{a_t}^\star\right)^2\right) \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda} \mathrm{E}_q \left[\lambda_t\right]\right)$$

The derivative w.r.t to the hyerparameter θ_λ is very similar to the one of α_λ. Using for drug t the vector of variational scale parameters $\theta_{\lambda_t}^\star = \left(\theta_{\lambda_{t,1}}^\star, \ldots, \theta_{\lambda_{t,N_t}}^\star\right)$, both derivatives of

$\mathrm{E}_q[\log p(a|\lambda)]$ w.r.t. the hyperparameters of λ are:

$$\begin{aligned}\frac{\partial \mathrm{E}_q[\log p(a|\lambda)]}{\partial \alpha_\lambda} &= \frac{1}{2}\sum_t \Bigg(N_t \cdot \psi'\left(\alpha_\lambda^\star\right) - \mathrm{tr}\left(\mathrm{diag}\left(\theta_{\lambda_t^\star}\right) \cdot \mathrm{E}_q\left[a_t a_t'\right]\right) \\ &\quad + \mathrm{tr}\left(\mathrm{diag}\left(\mathrm{E}_q\left[\lambda_t\right]\right) \cdot \left(2\left(\Sigma_{a_t}^\star\right)^3 \cdot C_{a_t} + \left(\Sigma_{a_t}^\star\right)^2\right) \cdot \mathrm{diag}\left(\theta_{\lambda_t}^\star\right)\right)\Bigg) \\ \frac{\partial \mathrm{E}_q[\log p(a|\lambda)]}{\partial \theta_\lambda} &= \frac{1}{2}\sum_t \Bigg(N_t \cdot \frac{\theta_{\lambda_{t,i}}^\star}{\theta_\lambda^2} - \mathrm{tr}\left(\mathrm{diag}\left(\alpha_\lambda^\star \frac{\left(\theta_{\lambda_{t,i}}^\star\right)^2}{\theta_\lambda^2}\right) \cdot \mathrm{E}_q\left[a_t a_t'\right]\right) \\ &\quad + \mathrm{tr}\left(\mathrm{diag}\left(\mathrm{E}_q\left[\lambda_t\right]\right) \cdot \left(2\left(\Sigma_{a_t}^\star\right)^3 \cdot C_{a_t} + \left(\Sigma_{a_t}^\star\right)^2\right) \cdot \mathrm{diag}\left(\theta_{\lambda_t}^\star\right)\right)\Bigg)\end{aligned}$$

Continuing to consider the squared kernel weights a_t, they appear further in $\log p(G|a, \upsilon, \mathcal{X})$:

$$\begin{aligned}\mathrm{E}_q[\log p(G|a,\upsilon,\mathcal{X})] =& \frac{1}{2}\sum_{t=1}^T \left(N_t \cdot P \cdot \mathrm{E}_q\left[\log \upsilon_t\right] - N_t P \log(2\pi) - \mathrm{E}_q\left[\upsilon_t\right] \cdot c_{G_t}\right) \\ , c_{G_t} =& \sum_{i=1}^{N_t}\left[\mathrm{tr}\left(\mathrm{E}_q\left[G_{t,i}' G_{t,i}\right]\right) - 2\,\mathrm{E}_q\left[G_{t,i}'\right]' K_{t,i}\,\mathrm{E}_q\left[a_t\right] + \mathrm{tr}\left(K_{t,i}' K_{t,i}\,\mathrm{E}_q\left[a_t a_t'\right]\right)\right]\end{aligned}$$

Recalling $\sum_k K_{t,i} g_{t,k} = K_t g_t$ and $\sum_{i=1}^{N_t} K_{t,i}\prime K_{t,i} = \sum_{k=1}^P K_{t,k}^2$ (see Table 2), one gets:

$$\frac{\partial}{\partial \alpha_\lambda}\mathrm{E}_q\left[\log p(G|a,\upsilon,\mathcal{X})\right] = -\sum_t \mathrm{E}_q\left[\upsilon_t\right] \cdot \mathrm{tr}\left(\sum_k K_{t,k} K_{t,k} \cdot \frac{\partial}{\partial \alpha_\lambda}\mathrm{E}_q\left[a_t a_t'\right]\right)$$

Next to consider is the entropy of $q(a_t)$:

$$-\mathrm{E}_{q_G}[\log q(G)] = \sum_{t=1}^T \frac{1}{2}\left(N_t(\log(2\pi+1)) + \log\left|\Sigma_{G_{t,\cdot}'}^\star\right|\right) \qquad , \Sigma_{G_{t,\cdot}'}^\star = \mathrm{Var}_q\left[G_{t,\cdot}'\right] = \mathrm{Var}_q\left[G_{t,i}'\right]$$

The derivative of the determinant of the variance using the rule $\frac{\partial}{\partial\alpha}|\Sigma| = |\Sigma|\,\mathrm{tr}\left(\Sigma^{-1}\frac{\partial}{\partial\alpha}\Sigma\right)$ (Petersen and Pedersen, 2012, ch. 2.1) is:

$$\begin{aligned}\frac{\partial}{\partial \alpha_\lambda}\log\left|\Sigma_{a_t}^\star\right| &= \frac{1}{\left|\Sigma_{a_t}^\star\right|} \cdot \left|\Sigma_{a_t}^\star\right| \cdot \mathrm{tr}\left(\left(\Sigma_{a_t}^\star\right)^{-1}\left(\Sigma_{a_t}^\star\right)^2 \cdot (-1) \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda}\mathrm{E}_q\left[\lambda_t\right]\right)\right) \\ &= \mathrm{tr}\left(-\Sigma_{a_t}^\star \cdot \mathrm{diag}\left(\frac{\partial}{\partial \alpha_\lambda}\mathrm{E}_q\left[\lambda_t\right]\right)\right)\end{aligned}$$

and thus equivalently $\frac{\partial}{\partial\theta_\lambda}\log\left|\Sigma_{a_t}^\star\right| = \mathrm{tr}\left(-\Sigma_{a_t}^\star \cdot \mathrm{diag}\left(\frac{\partial}{\partial\theta_\lambda}\mathrm{E}_q\left[\lambda_t\right]\right)\right)$.

The next term in the ELBO would be $\mathrm{E}_q[\log p(\upsilon)]$, where $\theta_{\upsilon_t}^\star$ does depend on a_t and thus the hyperparameters of λ. Since the propagation would continue, deriving the gradient for hyperparameters α_λ and θ_λ is stopped here. It has become apparent that one needs to restrict the search or try to determine the hyperparameters empirically using the differentiation quotient. The latter is unfortunately difficult in the current implementation of BMTMKL as the ELBO $\mathcal{L}$ depends on the variational parameters which have already been updated. It

is easier to determine, both analytically and empirically, the gradients of the expected log prior and entropy of only gamma distributed random variables which depends directly on the hyperparameters. These are $\lambda, \upsilon, \gamma, \omega$ and ϵ. These could then be optimized using gradient descent with a small step size. Using the criterion going back to Robbins and Monro (1951) and using the implementation as in Blei et al. (2017), an according update is then

$$\alpha_{upsilon}^{l+1} = \alpha_{upsilon}^{l} + l^{\kappa} \cdot \frac{\partial \mathcal{L}^{l}}{\partial \alpha_{\upsilon}}, \quad \kappa \in (0.5, 1],$$

where the superscript l represents the number of iterations already performed. Hyperparameters are not updated each iteration to guarantee a smoother optimization. Unfortunately, this approach did not lead to better results than setting all hyperparameters to one. Due to time constraints analytically optimizing hyperparameters has been discarded.

2.7 Predictions or fitted values:

Finally, obtaining predictions for the test data or fitted values for the training data is described. In terms of random variables of the BMTMKL model and having N_t cell lines tested for a drug t, the mean of the normal distribution of y_t is formed as follows: Individual observations of P kernels $K_{t,k} \in \mathbb{R}^{N_t \times N_t}$ are weighted by a_t of dimension $(N_t \times 1)$ to form intermediate outputs. Intermediate outputs are then multiplied with kernel specific coefficients e_k and a bias term b_t is added, yielding the distributional parameters of the random outcome y_t. In order to obtain a prediction for an out-of-sample cell line, the kernels of a new cell line with all N_t cell lines in the training sample have to be computed. As point predictions the mean of normally distributed random variables is used.

Gönen (2012a, p. 358) gives the prediction for the multiple regression model presented in this paper without conditioning on all relevant variables. Here I formulate the mean w.r.t to variational distribution as $E_q[\tau]$ for a random variable τ. The general result of obtaining a predictive distribution can be found in Bishop (2006, ch. 3.3.2, p.156f.). The density of the intermediate outputs is

$$p\left(G_* \middle| \{\{k_{t,k,*}, K_{t,k}\}_{k=1}^{P}\}_{t=1}^{T}, \mathcal{Y}\right) = \prod_{t=1}^{T} \prod_{k=1}^{K} \mathcal{N}\left(g_{t,k,*} \middle| k_{t,k,*} \, \mathrm{E}_q\left[a_t\right], \frac{1}{\mathrm{E}_q\left[\upsilon_t\right]} + k'_{t,k,*} \Sigma_{a_t} k_{t,k,*}\right)$$

The asterisk $*$ symbol denotes the new or testing cell line. The dimension of a vector of kernels between training data cell lines and a test data cell line $k_{t,k,*}$ is in $\mathbb{R}^{N_t}$, since it gives all the kernels between the training data and the test cell line, i.e. $k_{t,k,*} = \left(k(x_{t,k,1}, x_{t,k,*}), \ldots, k(x_{t,k,N_t}, x_{t,k,*})\right) = \left(k_{t,k,1}, \ldots, k_{t,k,N_t}\right)$, where $k(\cdot)$ is some function taking two inputs of the same dimension

and maps it to a real value $\mathbb{R}^F \times \mathbb{R}^F \mapsto \mathbb{R}$ (see Section 3.4). The random vector[14] of intermediate outputs for the test data point is $g_{t,*} = (g_{t,1,*}, \ldots, g_{t,P,*})'$ and its mean is $\mathrm{E}_q[g_{t,*}] = (\mathrm{E}_q[g_{t,1,*}], \ldots, \mathrm{E}_q[g_{t,P,*}])'$. The predictive density for a single out-of-sample outcome $y_{t,*}$ is

$$p(y_{t*} \mid G_*, \mathcal{X}, \mathcal{Y}) = \prod_{t=1}^{T} \mathcal{N}\left(y_{t*} \,\middle|\, \begin{bmatrix} 1 & \mathrm{E}_q\left[g'_{t,*}\right] \end{bmatrix} \mathrm{E}_q[b_t, e] \;,\; \frac{1}{\mathrm{E}_q[\epsilon_t]} + \begin{bmatrix} 1 & \mathrm{E}_q\left[g'_{t,*}\right] \end{bmatrix} \Sigma_{b_t,e} \begin{bmatrix} 1 \\ \mathrm{E}_q[g_{t,*}] \end{bmatrix} \right).$$

$\mathrm{E}_q[b_t, e]$ is the mean of the variational distribution of the random variables $(b_t, e')'$ which is extracted from the variational normal distribution of (b, e). The same is done for its variance $\Sigma_{b_t,e}$. In the end, the value of interest is the mean of $y_{t,*}$ as the point prediction with the specified variance:

$$\mathrm{E}_q[y_{t,*}] = \hat{y}_{t,*} = (1, \mathrm{E}_q\left[{g_{t,*}}'\right]) \cdot \mathrm{E}_q[b_t, e] \quad , \; \hat{\Sigma}_{y_{t,*}} = \frac{1}{\mathrm{E}_q[\epsilon_t]} + \begin{bmatrix} 1 & \mathrm{E}_q\left[g'_{t,*}\right] \end{bmatrix} \Sigma_{b_t,e} \begin{bmatrix} 1 \\ \mathrm{E}_q[g_{t,*}] \end{bmatrix}$$

In order to obtain fitted values for a training cell line i the kernels computed for training can be used, i.e. $K_{t,k,i}$ (c.p. Table 2). Predictions for training cell lines of drug response using the variational parameters are denoted as fitted values, whereas out of sample, or test, cell line values of drug response are denoted as predictions.

[14] again specified in terms of a column vector

3 Data

The used data is obtained from the DREAM7 challenge, starting with the initial profiling data sets. The 28 drugs modelled in the challenge form 8 groups (Costello et al., 2014, Figure 5) and the analyzed cell lines are a set of groups formed by different breast cancer types (ibid., Online Methods). These five subgroups are luminal, basal-like, cladion-low, ERBB2-ampliefied breast cancer cell lines and non-cancerous, normal-like (non-malignant). Cell line data is a subsample obtained from the data published by Daemen et al. (2013, 2015). 35 cell lines are available for training and 18 cell lines form the test dataset. In both datasets some drug sensitivity measures are missing for each cell line, which can be seen in Table 40 regarding the number of not available values (NA). In Section 3.1 the outcomes, i.e. drug responses [15] or tasks in BMTMKL denomination, are described. In Section 3.2 and 3.3 the available and added inputs are summarized, which are then transformed into kernel matrices as described in 3.4. The kernel matrices are the inputs to BMTMKL.

3.1 Drug data

The drug names have not been known to the challenge participants. 28 out of 31 drug names can been recovered from the the supplementary material of Costello et al. (2014, S2: Table 8) and are given in Table 39 in the appendix. Drugs can be grouped into 8 subsets (Costello et al., 2014, Figure 5, S2: Table 8) and their class names are indicating their target: Signaling RTK, Singaling NFkB, Signaling growth, Cell cycle, Proteasome, Metabolism, Regulation and Autophagy. For example, signaling drugs block some receptor needed for cell reproduction, autophagy drugs aim at the normal cell death mechanism or metabolism drugs inhibit enzymes necessary for chemical reactions needed in cells.

Higher numbers of $-\log_{10} GI_{50}$ drug values indicate higher effectiveness. Drug sensitivity is indicating the rate of survival of cells exposed to it. The survival rate is measured for different concentrations of a drug. The concentration for which 50 percent of the cells survive is interpolated using non-linear regression and then transformed by $-\log_{10}$ in order to get a positive number with reversed proportions: The less concentrations of a drug needed in a cell culture to achieve a growth inhibition of 50 percent, the higher the $-\log_{10} GI_{50}$. For example a $-\log_{10} GI_{50}$ value of 5 denotes the concentration of 0.00001 in the units of measuring concentration, which is unknown.

[15] Drug outcome, response, sensitivity or susceptibility are interchangeable denominations (see appendix B).

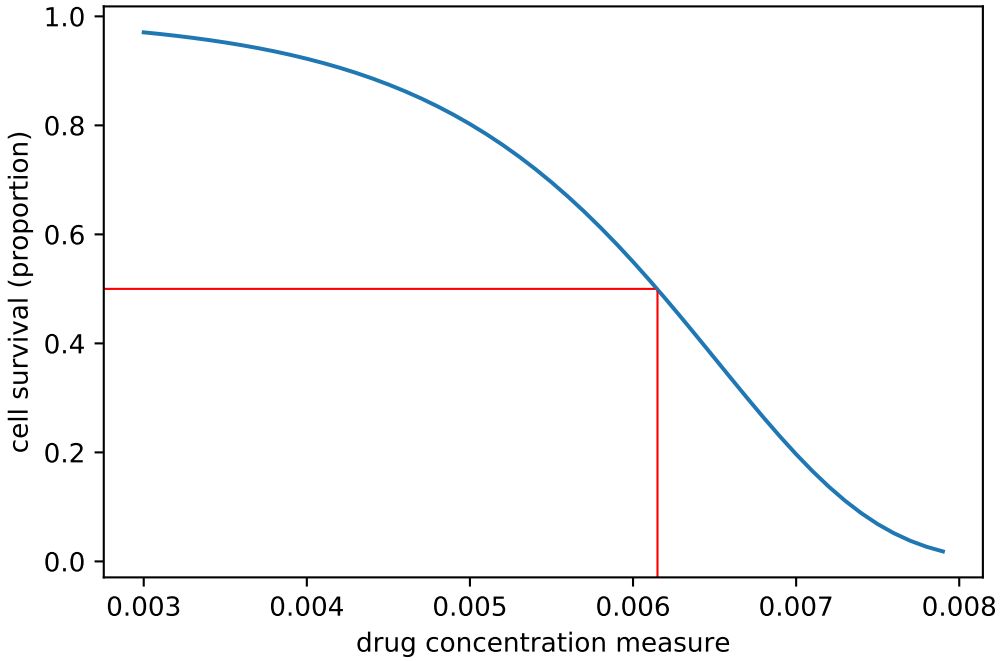
Drug effectiveness was established under laboratory conditions, i.e. withouth an immune system, by 5 doses over 72 hours with different concentrations. The curve fitting to obtain the 50 percent value is described in Monks et al. (1991) and illustrated for an artificial drug in Figure 2. For drugs where the GI_{50} could not be reached, but the data quality was sufficient, the maximal concentration has been used (Costello et al., 2014, see *Online Methods*). Looking at Table 40 in the appendix, drug 26 has the same value for all cell lines, i.e. the cut-off has not been observed at any tested level of concentration. The maximal tested concentration of $GI_{50} = 10^{-5.48} = 3.3110^{-6}$, which still do not lead to a 50 percent growth inhabitation, is thus assigned to all tested cell lines for drug 26. Costello et al. (2014) excluded drugs number 12, 26 and 27 in the challenge evaluation due to lack of predictive power: "The NCI-DREAM data as presented to participants contained 31 drugs; however, 3 of these drugs had completely flat profiles (i.e., the GI_{50} for all cell lines were the same) and were thus unscorable" (Costello et al., 2014, S1: p.76). The 28 remaining drugs are identified with the help of the given Drug ZScores[16] used in later evaluation. However, Table 40 only shows drug 26 to be completely flat. Drug 12 and 27 show little variation, but it is not obvious why drug 11, 21, 22 or 29 have not been excluded instead or additionally. Since the teams participating in the DREAM7 challenge have only known the training data, summary statistics of the training drug response data are given in Table 41 in the appendix, suggesting to exclude besides drug 26 as well drug 5 and 24 (c.p. Costello et al. 2014, Figure 1b). The measure of drug concentration is common but might pose problems if gene view specific kernel coefficients have to be learned. Since drugs can be categorized into 8 subsets using different mechanisms in the cell lines, it is hardly sensible to compare them directly: The toxicity of one drug cannot be compared to a hormone which suppresses cell growths. This thought of line suggests standardizing the drug data to mean zero and variance one. In Section 4 performance between standardized and non-standardized drug responses is compared.

3.2 Six profiling datasets

Having six profiling datasets is exceeding the scope of regularly available data in this field of research. Often only gene expression data is available. These data sets are described in Table 5 including their subject of measurement, which is either Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA) or proteins. The Table also gives the number of features (# feat.) and the assigned kernel coefficient e_k in the BMKMTL model. These datasets can for

[16] www.synapse.org/#!Synapse:syn2785845

Figure 2: Fitted growth inhibition curve for a hypothetical drug



example capture gene amplification, which is an increase in the number of copies of a gene, common to cancer cells. There might be also an increase in the RNA and proteins made in conjunction with amplified genes. Therefore an amplified gene may cause a cancer cell to grow or become resistant to a drug.

Briefly the data might be described as following: The three DNA based datasets give a hint if there are less or more than the two expected copies of information in DNA. Copy Number Variations (CNV) give the variation of proteines A, C,T or G in the DNA, whereas methylation is focusing on "CG" combinations and their changes in DNA. Both give some kind of proportion of DNA changed for a gene as amplification value. Exome Sequencing is comparing genes on the whole genome and indicates alterations.

RNA sequencing data is using the RNA genome. The genome of a cancer cell is compared to the one of a non-cancerous cell. Gene expression or microarry data is quantifying the activation of single genes in RNA directly.

Reverse Phase Protein Array (RPPA) is a measure to quantify protein occurrence using antibodies. Only full validated proteins are taken into consideration.

Two datasets have been provided additionally dichotomized, increasing the total number of provided datasets to eight.

Table 5: Six profiling data sets used in DREAM7 challenge (Costello et al., 2014)

	name	# feat.	short description as given in data files	e_k
RNA	Gene Expression[17]	18632	Transcript expression values. Affymetrix GeneChip Human Gene 1.0 ST microarrays were processed using the R package aroma.affymetrix (over 18,000 expression values)	e_1
DNA	CNV	27234	DNA Copy-Number Variation (CNV) relativ to normal cells. Affymetrix Genome-Wide Human SNP6.0 Array.	e_2
DNA	Methylation	27551	DNA methylation data. The Illumina Infinium Human Methylation27 BeadChip Kit was used for the genome-wide detection of 27,578 CpG loci, spanning 14,495 genes. GenomeStudio Methylation Module v1.0 was used to express the methylation for each CpG locus as a value between 0 (completely unmethylated) and to 1 (completely methylated) (over 27,000 CpGs)	e_3
RNA	RNA-seq	36953	RNA sequencing data (RNA-seq). RNA-seq libraries were prepared using the TruSeq RNA Sample Preparation Kit (Illumina) and Agilent Automation NGS system per manufacturers instructions. Expression analysis was performed with the ALEXA-seq software package	e_4
			discretized (provided by DREAM7 challenge)	e_7
DNA	Exome seq	10607	Whole exome sequencing (seq). Mutation status was obtained from exome-capture sequencing (Agilent Sure Select system).	e_5
			discretized (provided by DREAM7 challenge)	e_8
	RPPA	66	An antibody-based method to quantitatively measure protein abundance (131 proteins assayed, 66 fully validated)	e_6

3.3 Additional data sets: pathway information and interactions

3.3.1 Pathway information

Additional data has been matched using databases which assign a combinations of genes to some cell functionality. So called pathways of genes are a set of genes involved in a certain chain of chemical reactions within the cell. If a gene in this chain is altered by

cancer certain reactions fail or change their behavior (Vaske et al., 2010). Using a pathway, i.e. an information on a set of genes for a given dataset, some statistic has to be chosen as summary, e.g. the mean of the set of genes involved in the pathway. Therefore using additional information regarding pathways is technically equivalent to subdividing the data and adding as a new observations some statistic of theses values for the cell line.

The challenge is to identify the activation of pathways which are known from clinical research to be important for the (sub-) types of cancer. Gönen and his team used MSigDB for adding curated (C2) and canonical (CP) pathways for the three DNA data sets (gene expression, methylation and copy number variation) and the exome-seq data. Additionally two datasets were created as a combination of the original ones using the PARADIGM algorithm (Costello et al. 2014, p. 1206f. ,Vaske et al. 2010). As MSigDB and PARADIGM have not been available, different pathway information has been added in this thesis (c.p. Table 6).

Pathway information can be based on five of the six original profiling datasets as they have gene identifiers. It can be assumed that identified pathway activations on the basis of different profiling datasets are correlated. However, as they are used to provide additional information along with the original data, only the additional information is used by BMTMKL.

In R gene sets are provided amongst others by *reactom.db* (Ligtenberg, 2017) and *org.Hs.eg.db* (Carlson, 2017) databases, which are part of the bioconductor project for open source software in R for bioinformatics. Pathways are mapped using the HUGO Gene Nomenclature Committee Identifier (HGNC_ID)[18] in the data, extracting all activation values of genes in the specified pathway. The average or maximal activation value is saved as a new information for the involved cell lines. It would be possible to use other statistics, e.g. the median.

3.3.2 Interactions between profiling datasets

Information of genes is interacted by multiplying kernels for cell lines. As kernels, which are some sort of similarity defined between cell lines given some data and which are presented in the following Subsection, are additive, kernels for cell lines could be multiplied directly. It would also be possible to match only common genes using the HGNC_ID and then multiply features of DNA directly before computing kernels.

[18]HUGO is short for Human Genome Organization

Table 6: Pathways as reported in Costello et al. (2014) and in this thesis

Costello et al. (2014)	this thesis	statistic	e_k
Gene Expression C2	gene expression - Reactom	average	e_9
Expression CP	gene expression - Org.Hs.Eg		e_{10}
Methylation C2	Methylation - Reactom	maximum	e_{11}
Methylation CP	Methylation - Org.Hs.Eg		e_{12}
Copy Number V. C2	Copy Number Variation - Reactom	maximum	e_{13}
Copy Number V. CP	Copy Number Variation - Org.Hs.Eg		e_{14}
Exome-seq C2	Exome-seq - Reactom	maximum	e_{15}
Exome-seq CP	Exome-seq - Org.Hs.Eg		e_{16}
PARADIGM genes	RNASeq - Reactom	maximum	e_{17}
PARADIGM sets	RNASeq - OrgHsEg		e_{18}

Table 7: Interacted profiling datasets

Gene Expression · Methylation	e_{19}
Gene Expression · Copy Number Variation (CNV)	e_{20}
Copy Number Variation (CNV) · Methylation	e_{21}
Gene Expression · Copy Number Variation (CNV) · Methylation	e_{22}

3.4 Kernels

Two kernels have been calculated, one for real valued outcomes, one for binary data.

The Gaussian kernel calculates the scalar product between two vectors of real valued entries $x_{t,k,i}$ and $x_{t,k,j}$, normalizes it by $-2\sigma^2_{t,k}$ and then transforms all values into positive values using the exponential function. The Gaussian kernel corresponds to a infinite dimensional feature space (Bishop, 2006, p. 297, 321) and is defined as

$$k_{t,k}(x_{t,k,i}, x_{t,k,j}) = \exp\left(-\frac{1}{2\sigma_k^2}\|x_{t,k,i} - x_{t,k,j}\|^2\right) \quad \forall(t,k,i,j),$$

where σ_k^2 is set to the dimensionality, i.e. the number of features, of the corresponding genomic view, as describe in Costello et al. (2014, S1, p. 6). Choosing the right bandwidth σ_k^2 can be an optimization problem itself. One could for example perform feature selection for each drug and thus have $\sigma^2_{t,k}$ specific to each drug t. Although differently noted down (ibid.) the winning team, Mr. Gönen in collaboration with Ammad-Ud-din computed only an overall kernel between cell lines corresponding to the dimensionality of the genomic view

(c.p. Table 5) . If feature selection is performed one has to adapt the bandwidth parameter accordingly.

For binary data the Jaccard similarity is used. The scalar product between two vectors of zeros and ones, $x_{t,k,i}$ and $x_{t,k,j}$, is giving the number of identical entries with one. In this thesis, a one e.g. represents an activation in RNA data on a specific genome. The number is normalized by the sum of the number of ones in $x_{t,k,i}$, the number of ones in $x_{t,k,j}$ and the just described number of shared entries of one.

$$k_{t,k}(x_{t,k,i}, x_{t,k,j}) = \frac{x'_{t,k,i} x_{t,k,j}}{x'_{t,k,i} x_{t,k,i} + x'_{t,k,i} x_{t,k,j} + x'_{t,k,j} x_{t,k,j}}$$

The six profiling data sets and the pathway data are processed according to the following steps:

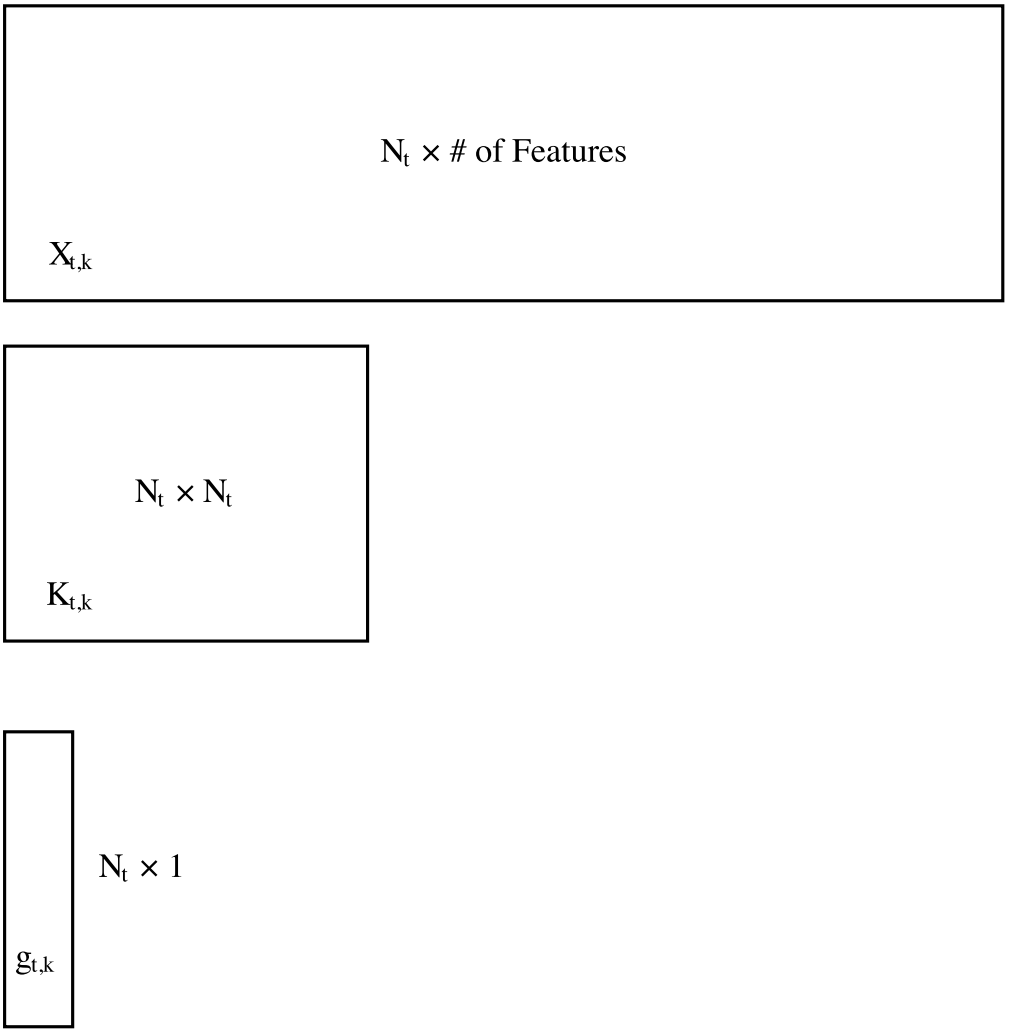
1. Augment all input data to all training cell lines by adding missing ones with NAs.
2. Real valued data: Standardize data
3. Replace NAs with column mean, i.e. zero after standardizing the data
4. Compute kernel matrices as described
5. Select for each drug the cell lines in all kernels for which valid drug information is provided, resulting in a varying number of cell line information used for each drug

In total 22 kernel matrices are computed on the basis of the six provided profiling data sets, of which two are also additionally dichotomized. These kernel matrices are the0 first dimensionality reduction performed in BMTMKL. This is illustrated in Figure 3 using the N_t cell lines of drug t and their corresponding entries of the kernel matrix k, which are extracted from the overall 35×35 dimensional kernel matrix on all training cell lines (c.p. Section 2.1).

3.5 Data Limitations

Drug data has been anonymized in the DREAM7 challenge, prohibiting to use additional prior information on dependencies or similarities between drugs. However, similarity between provided drug effectiveness could be measured empirically and incorporated in an adapted model in the future.

Figure 3: Dimensionality reduction performed in BMTMKL for all cell lines with measured GI_{50} concentration for drug t



Specific gene defects which indicate a possibility for treatment, e.g. to block a malfunction with some sort of receptor, cannot be modelled directly using kernel matrices, as they are based on overall similarity between cell lines. However, kernels could be computed on pathways only.

4 Results

After deriving the model and describing the available data, some results are highlighted. First, different hyperparameter settings are discussed and compared in Section 4.2. The evaluation in the DREAM7 challenge described in Costello et al. (2014, S1, Supplementary Note 4, p. 77ff.) is presented in Section 4.2 and the performance of different models in comparison to the obtained score by BMTMKL in the challenge is discussed.

4.1 Model evaluations - different models of grid search

A total of ten hyperparameters have to be set for the gamma distributed random variables, namely α_λ, β_λ, α_υ, β_υ, α_γ, β_γ, α_ω, β_ω, α_ϵ, β_ϵ using the shape and rate notation introduced in Section 2.1. Gönen uses the scale instead of the rate notation (c.p. appendix A.1.1) for his implementation, e.g. θ_λ instead of β_λ. As one is the inversion of the other, the model does not change fundamentally, as explained exemplary for the full conditional and thus the variational parameters of λ in Section 2.2.1.

Trying 3 values for each hyperparameter results into 10^3 possible combinations, i.e. the need to train the model 59049 times. On a computer with 12 logical processing cores, this takes around 12 hours performing 200 iterations and using 22 kernel matrices.

In Table 8 five combinations of hyperparameters are reported: The default setting for small sample sizes as recommended in the code by Mr. Gönen, all set to one, the set with maximal ELBO, with minimal ELBO and a randomly chosen set. Gönen notes in his algorithm used for Costello et al. (2014) that a good choice for hyperparameters of a gamma distributed random variable τ, which are shape α_τ and scale θ_τ, is 10^{-10}. This results in an initial precision for the normally distributed random variables $a, G, b, e, \mathcal{Y}$ of $E_q[\tau] \overset{19}{=} 10^{-20}$ and thus in a variance of 10^{20}. It means that no prior knowledge of the underlying solution for the normally distributed variables is assumed. Accordingly, setting all hyperparameters to one results in an initial variance of one. In case of standardized drug outcomes setting α_ϵ or θ_ϵ to one can be justified as a good choice. Grid selection by combining values for the shape and scale parameters is arbitrary, which is why for different combinations of hyperparameters the ELBO, the mean-squared error and rankings as performance measures are compared.

Comparing the results in Table 8 suggests that the level of the ELBO depends on the chosen hyperparameters. This claim is plausible regarding the updates in Table 24 and

[19] See appendix A.1.1 for details on gamma distribution

Table 8: Five configurations of grid search on hyperparameters shape $\alpha \in \{10^{-10}, 1, 10\}$ and scale $\theta \in \{10^{-10}, 0.01, 1\}$ and their ELBO after 200 iterations using standardized drug response data

	α_λ	θ_λ	α_υ	θ_υ	α_γ	θ_γ	α_ω	θ_ω	α_ϵ	θ_ϵ	ELBO
default	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	-31936
ones	1	1	1	1	1	1	1	1	1	1	-2001
max	10	1	1	0.01	10	1	10	1	1	1	-1508
min	10^{-10}	10^{-10}	10	1	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	-41772
random	10^{-10}	10^{-10}	1	10^{-10}	10	1	1	1	1	1	-20823

Table 9: *Rescaled* fitted values of drug ten for five *training* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses and corresponding ranking

	true	default	ones	max	min	random	ranks					
UACC812	7.491	6.243	6.829	6.243	6.243	7.125	2	1	2	1	3	2
MCF10F	6.003	6.243	5.907	6.243	6.243	5.853	4	2	5	5	4	5
M.134VI	7.265	6.243	6.410	6.243	6.243	6.694	3	3	3	4	2	3
HCC1419	8.401	6.243	7.153	6.243	6.243	8.032	1	4	1	2	5	1
HCC1143	5.833	6.243	6.022	6.243	6.243	5.936	5	5	4	3	1	4

the ELBO in Table 25, both in the appendix. As the updated shape parameters $\alpha_\tau^\star, \tau \in \{\lambda, \gamma, \omega, \upsilon, \epsilon\}$ are fixed, and they occur in the log expected prior for the gamma distributed random variables, these terms influence the level of the ELBO as constants. Only scale parameters of gamma and parameters of normally distributed random variables are updated iteratively. The BMTMKL model has many hidden random variables and is very flexible in adapting to different settings. However, the result regarding the ELBO in Section 2.3 identify the setting with maximal ELBO as the best model as it is closest to the logarithmic evidence $\log p(\mathcal{Y}|\mathcal{X})$. In terms of the ELBO one would chose the *max*imal model.

Table 9 gives the rescaled fitted $-\log_{10} GI_{50}$ values for five sample cell lines from the training data for drug ten additionally to the true standardized drug response data. In Section 2.7 the procedure to obtain these fits is described in detail.

The level of fitted values to training cell lines clearly differs. Setting all hyperparameters to one gives the closest fits to observed values of drug response $-\log_{10} GI_{50}$ for drug ten. The values of the default, minimal and maximal ELBO model is constant looking at the

Table 10: Fitted values of drug ten for five *training* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses and corresponding ranking

	true	default	ones	max	min	rand.	ranks					
UACC812	0.95	1.94E-16	0.45	-7.61E-06	7.97E-06	0.68	2	3	2	1	3	2
MCF10F	-0.18	1.99E-16	-0.26	-3.71E-05	3.67E-06	-0.30	4	1	5	5	4	5
MD.134VI	0.78	1.91E-16	0.13	-3.12E-05	9.81E-06	0.34	3	4	3	4	2	3
HCC1419	1.65	1.98E-16	0.70	-1.01E-05	1.60E-07	1.37	1	2	1	2	5	1
HCC1143	-0.31	1.88E-16	-0.17	-2.11E-05	1.71E-05	-0.23	5	5	4	3	1	4

first three decimal values. Rescaling was performed using the empirical mean and standard deviation of the training data of drug ten. In order of the models stated, for fitted values their corresponding rankings for the five sample cell lines are given. As the rescaled default fits are constant, the cell lines are ranked in order of occurrence.

In order to see why the rescaled values are nearly constant, the standardized observed values of drug response $-\log_{10} GI_{50}$ for drug ten along the model fits are reported in Table 10. The default model yields fitted values in the range of 10^{-16}, whereas the min and max model yield fitted values in the range of 10^{-7} to 10^{-5}. Since the scale and mean can be given precisely in the range of 10^{-5}, the rescaled default fitted values appear to be constant due to too imprecise storing. Therefore only the ranking using the not rescaled fitted values for the default model yields sensible results. The model setting all hyperparameters to one and the random model yield the closest fits as they are in the range of the observed standardized values.

Comparing the rankings, the default model gets the maximal value of the five samples wrong, but ranks the rest in order of the true results. The model of ones as the random model get the first three sample values right, and only swaps the order of the last two. Although the level of fitted values for the default is very small, the ranking is not totally wrong.

Table 11 and Table 12 give five examples for predicted values for test or out of sample cell lines which have not been used for training. Two cell lines of the example have no reported drug sensitivity value, which is designated by NA. Again, for the rescaled version the default, max and min model have constant values regarding the first three decimal values. The ranking for the default model is not sensible because of the unprecise rescaling described before.

The ranking can only be judged for the cell lines for which the values are observed, i.e. the

Table 11: *Rescaled* predictions of drug ten for five *test* cell lines given by models trained on standardized $-\log_{10} GI_{50}$ drug responses

cell line	true	default	ones	max	min	random	ranks					
ZR75B		6.243	6.391	6.243	6.243	6.476	4	1	1	2	1	1
21NT		6.243	6.076	6.243	6.243	6.021	5	2	2	4	4	2
HCC3153	4.848	6.243	5.940	6.243	6.243	5.742	3	3	3	1	2	4
184B5	5.550	6.243	5.928	6.243	6.243	5.801	2	4	4	5	3	3
MCF10A	6.082	6.243	5.691	6.243	6.243	5.452	1	5	5	3	5	5

Table 12: Predictions of drug ten for five *test* cell lines given by models trained on standardized drug responses

cell line	true	default	ones	max	min	rand.	ranks					
ZR75B	NA	1.96E-16	0.11	-2.84E-05	1.54E-06	0.18	4	1	1	2	1	1
21NT	NA	1.93E-16	-0.13	-4.52E-05	-3.06E-06	-0.17	5	4	2	4	4	2
HCC3153	-1.067	1.89E-16	-0.23	-2.75E-05	6.93E-08	-0.38	3	5	3	1	2	4
184B5	-0.530	1.94E-16	-0.24	-5.05E-05	-3.00E-06	-0.34	2	2	4	5	3	3
MCF10A	-0.123	1.94E-16	-0.42	-3.96E-05	-6.66E-06	-0.60	1	3	5	3	5	5

values of cell line MCF10A to 184B5 and HCC3153 as well as the one of 184B5 to HCC3153 are to be compared. Looking only at the ranking of the default and random model compared to the true drug sensitivity values, one notices that both models did rank correctly cell line 184B5 before cell line HCC3153 and both failed to rank MCF10A before 184B5. However, the default model correctly ranked MCF10A before HCC3153, whereas the random model failed to do so, indicating that the default model is again still suitable for ranking although its drug sensitivity predictions are in the level worse than the random model ones. The ones model gives a ranking in reversed order to the true ranking.

As the rankings and the level of fitted values are two totally different measures, the squared error will be used to access the fit. The mean squared error (MSE) over T drugs is defined as

$$MSE_j = \sum_{t=1}^{T} \frac{(y_t^j - \hat{y}_t^j)^2}{N_t^j} \quad , j = 1, 2 \quad ,$$

where y_t^j is either the non-missing drug data from the test set ($j = 2$) or training set ($j = 1$) (c.p. Corts-Ciriano et al. 2016, p. 88f.).

Table 13: Maximal and minimal ELBO configuration after 200 iterations of grid search of hyperparameters for shape $\alpha \in \{10^{-10}, 1, 10\}$ and scale $\theta \in \{10^{-10}, 0.01, 1\}$ using non-standardized drug response data

	α_λ	θ_λ	α_υ	θ_υ	α_γ	θ_γ	α_ω	θ_ω	α_ϵ	θ_ϵ	ELBO
max*	10	1	1	0.01	10	0.01	10	1	10	1	-843
min	10^{-10}	10^{-10}	10	1	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	-41772

As a reference the MSE using the mean of the drug sensitivity measure for each drug has been used as fit or prediction. The MSE using mean values is 9.12 on the training set and 29.16 on the test set. The MSE on the test set is mainly influenced by one outlier on drug 24, which drives 60 percent of the out-of-sample MSE. See the highlighted maximum value for drug 24 in Table 40 in the appendix.

Standardized and non-standardized drug sensitivity data is compared in order to check the reasoning presented on drug sensitivity data in Section 3.1. In Table 14 and 15 the mean squared error (MSE) over all drugs for the five models listed in Table 8 are reported, both for training models on standardized and non-standardized drug sensitivity data, along the MSE of the reference mean model. There is one difference between both tables. The maximal setting for non-standardized drug responses is different from the maximal setting for standardized data as given in Table 8 and denoted as max*. It is given in Table 13. The minimal setting on the defined grid is the same both for models trained on standardized as non-standardized drug responses.

The MSE of the rescaled fitted values in Table 14 can be compared directly with the MSE on the train set in Table 15. Using standardized differences, drug responses are weighted in the sense that the overall variance in the drug sensitivity measure $-\log_{10} GI_{50}$ of drug t is accounted for.

The default and minimal ELBO model have in both cases the same MSE on the training set and the same ELBO, which indicates that the level of the ELBO is mainly driven by hyperparameters. The MSE on the train set using non-standardized drug response data is as good as or worse than using the mean of the training data. The MSE with rescaled values is smaller for the ones and random model than using training means. The out-of-sample error using standardized drug data is for both models smaller than the one trained on non-standardized drug sensitivity data, but a lot less pronounced. Even the ones model as best performing model only decreases the MSE on the training data by less than three percent.

Table 14: Configurations of Table 8 with different Mean Squared Error criteria adding a model using training means as a reference

	mean	default	ones	max	min	random
MSE train set	28.87	28.87	8.65	28.87	28.87	1.61
MSE train set (rescaled)	9.12	9.12	3.33	9.12	9.12	0.71
MSE test set (rescaled predictions)	29.16	29.16	28.34	29.16	29.16	28.09
ELBO after 200 iterations	NA	-31936	-2001	-1509	-41772	-20823

Table 15: Configurations of Table 8 replacing the maximal model as given in Table 13 trained with *non-standardized* drug responses

	mean	default	ones	max*	min	random
MSE on train set	9.12	9.12	9.12	9.12	9.12	12.20
MSE on test set	29.16	29.22	29.17	29.16	29.22	40.17
ELBO after 200 iterations		-31936	-1432	-843	-41772	-21046

Excluding the outlier which drives 60 percent of this MSE, yields an increase still under ten percent.

Regarding the fitted values and predictions in Table 16 and 17, the models trained on non-standardized outcome data provide a less precise fit for the drug response on drug ten for the sample cell lines than the ones obtained from models trained on standardized drug data reported in Table 9 and Table 11. Using standardized drug data yields better fits as drug responses are ensured to be comparable.

Left out up to now has been the variance of the fitted values or predictions. Table 18 gives for the model of ones the fitted values in standardized and rescaled version with variance and standard deviation for the sample cell lines. The variance for the standardized version is obtained as described in Section 2.7 and then transformed for the rescaled version by the squared standard deviation of drug ten.

Having introduced three criteria of model selection, the question arises which one to use for selecting a final model, which the team around Mr. Gönen submitted in the DREAM 7 challenge.

Selecting a model could be done by three fold cross validation using the mean-squared error for out-of-sample predictions and using the complete data. As this thesis aims to be

Table 16: Fitted values for five *training* cell lines given by models trained on *non-standardized* $-\log_{10} GI_{50}$ drug responses and corresponding ranking

cell line	true	default	ones	max	min	random	ranks					
UACC812	7.49	6.23	6.22	5.55	6.23	6.35	2	1	2	4	2	3
MCF10F	6.00	6.23	6.22	5.57	6.23	5.65	4	2	4	3	4	5
MDAMB134VI	7.26	6.23	6.22	6.23	6.23	6.36	3	3	1	1	3	2
HCC1419	8.40	6.23	6.22	5.98	6.23	6.98	1	4	3	2	5	1
HCC1143	5.83	6.23	6.22	5.44	6.23	5.66	5	5	5	5	1	4

Table 17: Predictions of drug ten for five *test* cell lines given by model trained on *non-standardized* drug responses

cell line	true	default	ones	max	min	random	ranks					
ZR75B		6.234	6.215	6.283	6.234	6.466	4	1	1	1	1	1
21NT		6.234	6.215	6.171	6.234	6.283	5	2	5	3	3	2
HCC3153	4.848	6.234	6.215	5.539	6.234	5.645	3	3	3	5	2	4
184B5	5.550	6.234	6.215	6.196	6.234	6.160	2	4	4	2	4	3
MCF10A	6.082	6.234	6.215	5.901	6.234	5.595	1	5	2	4	5	5

comparable to Costello et al. (2014) only the training split and test split of the data as provided in the challenge is used. Furthermore the comparison with the out-of-sample mean squared error has not been possible to challenge participants as they lacked drug responses for test cell lines.

The second criterion, the selection could be based on, is the in sample MSE which indicates how well the model fits the observed drug sensitivity values on the training set. In Table 30 in the appendix three settings of hyperparameters both for the minimal MSE on the training data (in sample) and test data (out-of-sample) are reported considering the 59049 possible combination of hyperparameters of the grid constructed of $\alpha \in \{10^{-10}, 1, 10\}$ and $\theta \in \frac{1}{\beta} = \{10^{-10}, 0.01, 1\}$. Comparing the sets of ten hyperparameters, one notices that only hyperparameters for the shape parameter for one gamma distributed random variable differs, which is highlighted in bold fond. This suggests that the shape parameter does not change the fit, although it clearly has an influence on the ELBO. This comparison shows that at least more than one setting of hyperparameters can lead to equal performance in predicting the real valued outcomes.

Table 18: Fitted values of drug ten for five *training* cell lines, their variance and standard deviation in ones model trained on standardized $-\log_{10} GI_{50}$ drug responses

	fit	VAR	SD	fit res.	VAR res.	SD res.
UACC812	0.448	0.310	0.557	6.829	0.531	0.728
MCF10F	-0.257	0.310	0.557	5.907	0.530	0.728
MDAMB134VI	0.128	0.310	0.557	6.410	0.530	0.728
HCC1419	0.697	0.311	0.557	7.153	0.531	0.729
HCC1143	-0.169	0.310	0.557	6.022	0.530	0.728

Table 19: Relative changes between ten iterations in the ones model of Figure 5

Iteration	5	15	25	35	45	55	65	75	85	95
ELBO	-2882	-2487	-2385	-2322	-2276	-2238	-2206	-2178	-2153	-2131
Δ ELBO		-0.137	-0.041	-0.026	-0.020	-0.017	-0.014	-0.013	-0.011	-0.010
Δ MSE		-0.828	-0.261	-0.039	0.056	0.101	0.122	0.131	0.133	0.130
Δ MSE res.		-0.813	-0.231	-0.020	0.071	0.114	0.134	0.141	0.141	0.137
Δ MSE ofs.		-0.020	-0.003	-0.001	-0.000	0.000	0.000	0.000	0.001	0.001

The third criterion is the relative ranking for which a measure is introduced in the next Section. The previous comparison of rankings between the default, random or ones model on the five sample cell lines from the training or test data have indicated that the ranking is not necessarily related to the goodness of fits or predictions. Fortunately, in Section 4.2.2 it will be shown that better predictions give better overall rankings.

The MSE and the ELBO have been compared for training the model with a fixed number of 200 iterations. In Figure 4, the dependence of the ELBO and of the three MSE estimates on the number of performed iterations is presented for the random model trained on standardized drug data. The Figure's first values are at iteration five as the random initialization leads to noisy initial values.

First, one notices that the model takes some initial iterations where in terms of the MSE hardly any improvement is observed. In the first iterations, however, the ELBO increases drastically, before it increases more incrementally. Within roughly ten iterations the different MSE estimates drop significantly, as the ELBO makes a brief pronounced increasement, before the ELBO does continue to increase smoothly and the MSE estimates stay nearly constant. Note that in terms of MSE around iteration 100, the random model is overfitting.

Figure 4: MSEs and ELBO for random model using standardized drug responses

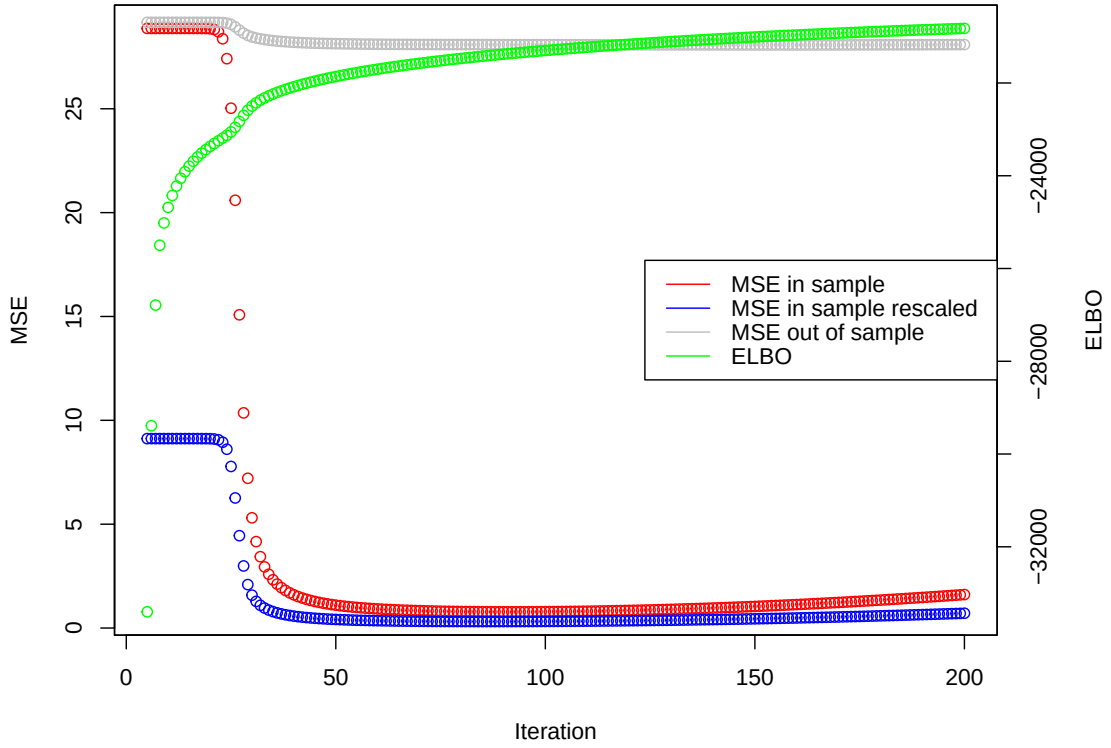


Figure 5: MSEs and ELBO for ones model using standardized drug responses

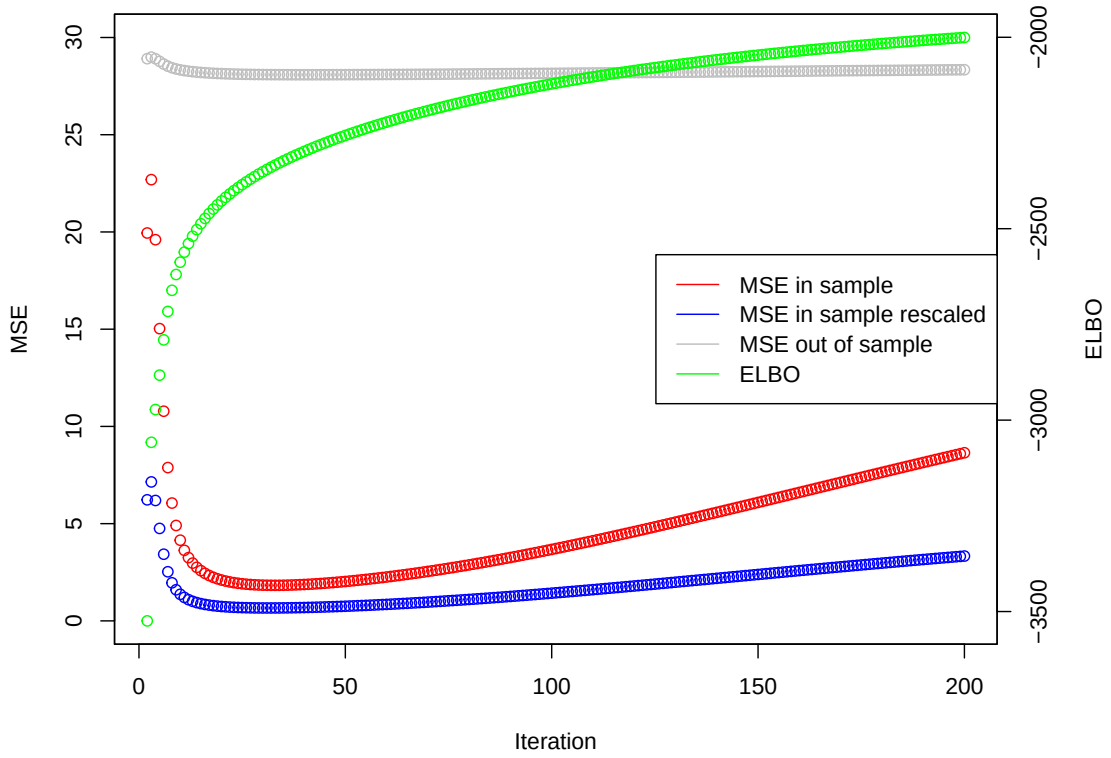


Figure 6: MSEs and ELBO for default model using standardized drug responses

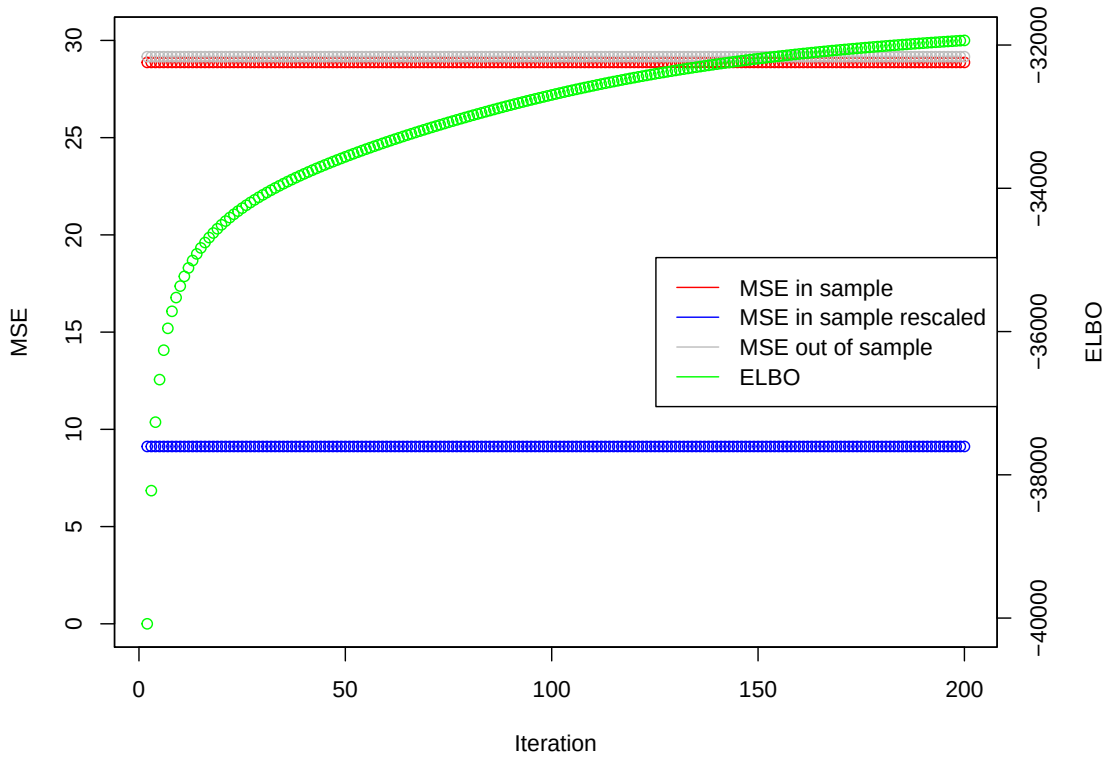


Table 20: Number of iterations for minimal MSEs of default, ones and random model

	Default	ones	random
MSE train set	120	33	91
MSE train set (rescaled)	145	31	87
MSE test set (rescaled predictions)	106	41	158

Overfitting is even more pronounced for the ones model, which is shown in Figure 5. Overfitting starts around iteration 30 whereas the ELBO still increases significantly afterwards. In Table 19 relative changes for the ones model are given by a step size of 10 iterations starting at 5 and ending at 95 iterations. The left vertical axis gives MSE values, whereas the right vertical axis gives ELBO values.

In the default model, see Figure 6 the MSE estimators are flat for mainly all iterations showing that setting the prior precisions too small yields no good performance in terms of the squared errors.

In Table 20 the number of iterations needed to obtain the minimum of the three different MSEs for the three previously discussed models are listed. The minimal MSE on the train-

ing set is not obtained on the same number of iterations for the standardized and rescaled version, however for the ones and random model both are close. As one can see in Figures 4 and 5, the change around the minimal MSE is not large.

Above only the drug sensitivity measure is considered. Now, the change in parameters of the hidden random variables is discussed for models with different settings of hyperparameters, which are trained on standardized drug sensitivity data 200 times. Although this leads to overfitting in terms of the MSE as shown in the previous Section, the analysis does not change in its general results. As estimate of a random variable for a cell line, its expectation is taken.

In Table 31 in the appendix the expectations of the precision and the kernel weights are reported. The precision is the inverse variance and thus has to be inverted in order to use it as posterior value in the model equations (c.p. Table 4). These values are used to calculate fits or predictions as described in Section 2.7.

The random model yields the most extreme results, having a very low precision, i.e. high variance, and strongly varying kernel weights. The default model yields very small values for both, whereas the maximal and minimal ELBO either have small means of the precisions $\lambda_{t,i}$ or the kernel weights $a_{t,i}$. The ones model is more moderate in the sense that variance is quite precise on the kernel weights as well as obtaining good variation. The hyperparameters therefore determine in which range these two random variables have the mean of their distribution.

The same picture can be drawn looking at the expected intermediate outputs for the sampled training cell lines and their expected precisions of the default, ones and random model in Table 32, 33 and 34 in the appendix. The default model has the smallest values, whereas the ones model give intermediate outputs around zero and the random model yields extremer and larger values. This is shown in the summary statistics for intermediate outputs of these models for drug 10 in Tables 35, 36 and 37 in the appendix, reporting additionally the expectations of the kernel coefficient $\mathrm{E}_q[e_k]$ and its precision $\mathrm{E}_q[\omega_k]$ for view k. The standard deviation of the intermediate outputs increases from default to ones to random model.

The corresponding views to the first six intermediate outputs are the six initial profiling sets given in Table 5, which are in order gene expression, CNV, methylation, RNA sequence, Exome sequence and RPPA datasets. The extremer and larger values of intermediate outputs for the random model are compensated with relatively small kernel coefficients.

If these kernel coefficients are interpreted as weights as proposed in Gönen (2012a), the

gene expression data has the most impact in the ones and random model, which corresponds to the results found by Costello et al. (2014). However, although all high dimensional inputs and response variables are standardized, the influence is hard to read directly form the magnitude of coefficients. Several transformations are performed, leading to the mentioned variation in intermediate outputs. Furthermore, regarding the ones model in Table 36 the precision of around 1.5 gives a standard deviation of around 0.82. Having a normally distributed variable e_k with mean of around 0.04 and a standard deviation of 0.81 could be statistically interpreted as having zero mean. Significance of views has been established by random simulations of different test data using the ranking performance measure in Costello et al. (2014, Figure 4).

Since possibly new data obtained within the predict project, in which scope this thesis is written, will be only gene expression data, further analysis on relevance of inputs is left out.

Finally Table 38 gives biases and their corresponding precision for the first ten drugs for different models. The level of biases corresponds to the level of fitted values (see Table 10). Although models have been trained on standardized drug data, there is variation in the expected values of biases.

4.2 Model evaluation in DREAM7

In Costello et al. (2014) a weighted probabilistic concordence index (wpc-index) has been calculated to compare the results between different models. It is based on the ranking of drug effectiveness of cell lines for each drug. All concordence measures for each drug are than weighted according to their share in variation. The scoring method is presented in detail in subsection 4.2.1.

BMTMKL has scored first with an wpc-index of 0.583, which is the upper benchmark. The difficulty is however to choose the best model for evaluation. This could be done solely on the basis of the obtained ELBO as it presents the optimum regarding the proof based on the evidence $p(\mathcal{Y}|\mathcal{X})$ in Section 2.3. However, shown in the comparison with evaluating model accuracy in terms of the mean squared error, the maximal ELBO does not present the best fit to the data. The ELBO and the MSE are both used for selecting the best model for ranking.

4.2.1 Challenge scoring method - wpc-index

The used evaluation method in Costello et al. (2014) is now presented. As a reference of drug response a Goldstandard for each drug is calculated, yielding a "gold standard list of

dose response values" (Costello et al., 2014, S1, Supplementary Note 4, p. 77ff.). Replicate measurements of $-\log_{10}(GI_{50})$ give mean values of drug response for each drug t. The gold standard for the training data is the drug response data provided and used as outcomes in the challenge DREAM7.

The ranked list of 18 test lines is denoted $R_t = \{r_{t,1}, \ldots, r_{t,18}\}$ for drug t, the first being the best. The Gold standard is a list of mean values for each cell line of dose response values in the test set, or out-of-sample set, obtained from replicate measurements of $-\log_{10}(GI_{50})$ values of a drug is $G_t = \{g_{t,1}, \ldots, g_{t,18}\}$, where the largest value is the best. The concordence between the predicted rank and the mean $-\log_{10}(GI_{50})$ for these 18 cell lines can be calculated as c-index $= c(G_t, R_t) = \frac{2}{n(n-1)} \cdot \sum_{i<j} h\big(g_{t,i}, g_{t,j}, r_{t,i}, r_{t,j}\big)$, where

$$h\big(g_{t,i}, g_{t,j}, r_{t,i}, r_{t,j}\big) = \begin{cases} 1 & , \text{if } \big(g_{t,i} > g_{t,j} \cap r_{t,i} < r_{t,j}\big) \cup \big(g_{t,i} < g_{t,j} \cap r_{t,i} > r_{t,j}\big) \\ 0.5 & , \text{if } \big(g_{t,i} = g_{t,j}\big) \\ 0 & , \text{if } \big(g_{t,i} > g_{t,j} \cap r_{t,i} > r_{t,j}\big) \cup \big(g_{t,i} < g_{t,j} \cap r_{t,i} < r_{t,j}\big) \end{cases}$$

The concordence index thus works as follows: If the ranking is false, nothing is added to the sum. If the cell lines have the same $-\log_{10} GI_{50}$ mean, 0.5 is added to the sum and if it is true a 1 is added. The total sum is normalized by the inverse of total comparisons made between both lists.

In order to be able to account for variation $\mathrm{Var}(t) = s_t^2$ in the drug t, a probabilistic concorcance index (pc-index) is used by Costello et al. (2014): $pc\big(G_t, R_t, s_t^2\big)\frac{2}{n(n-1)} \cdot \sum_{i<j} h_p\big(g_{t,i}, g_{t,j}, r_{t,i}, r_{t,j}, s_t\big)$, where

$$h_p\big(g_{t,i}, g_{t,j}, r_{t,i}, r_{t,j}, s_t\big) = \begin{cases} \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{g_{t,i} - g_{t,j}}{2s_t}\right)\right) & , \text{if } \big(r_{t,i} < r_{t,j}\big) \\ 0.5 & , \text{if } \big(r_{t,i} = r_{t,j}\big) \\ \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{g_{t,j} - g_{t,i}}{2s_t}\right)\right) & , \text{if } \big(r_{t,i} > r_{t,j}\big) \end{cases}$$

and using the Gauss error function $\mathrm{erf}(b) = \frac{2}{\sqrt{\pi}} \int_0^b e^{-t^2}\, dt$, $b \in \mathbb{R}$, which yields the probability of a normally distributed variable to be in the interval of $[-a, a]$ for $a = |b|$. In the presented formulation, it yields a negative value for negative b and a positive value for positive b in the range of $[-1, 1]$. Thus correctly ordered rankings now get assigned a value in the range of $(0.5, 1]$ and wrongly ranked values between $[0, 0.5)$.

Rankings are considered to equal, but this makes in my opinion little sense. If one would allow for rankings to equal, i.e. $r_{t,i} = r_{t,j}$, some values might have a greater value by definition instead of a value smaller than 0.5 in case of a wrong ranking. If measured $-\log_{10} GI_{50}$ drug

values are equal $g_{t,i} = g_{t,j}$, 0.5 is added for the comparison to the index, as in the previous c-index. As $r_{t,i}$ is a ranking, equal values are just ranked in their order of appearance in this thesis.

Finally an overall *weighted probabilistic concordence* (wpc-) index is calculated for all drugs using weights w_t, i.e.

$$\text{wpc-index} = \frac{\sum_{t=1}^{T} w_t \cdot pc_t}{\sum_{t=1}^{T} w_t} \quad .$$

The weights are obtained using the mean μ_t^r and standard deviation σ_t^r of 10000 random rankings and their corresponding pc-indices, compared with the pc-index using the ranked gold standard R_t^*, i.e. $w_t = \frac{pc_t^* - \mu_t^r}{\sigma_t^r}$ where $pc_t^* = pc\big(G_t, R_t^*, s_t^2\big)$. In a statistical sense these weights should be rather called importance, as the sum of w_t is not guaranteed to sum to one, which is why they are normalized. A random prediction for each drug resulting in $pc_t = 0.5$ results in a wpc index of 0.5.

The described evaluation method has been developed by Costello (2018) for the DREAM7 challenge and therefore no further reference is available.

4.2.2 Applying the wpc- index to models trained in Section 4.2.1

A perl script [20] for the scoring indices is used for evaluating the competition. These Perl scripts have a Python wrapper which is available on github. For comparison the models trained on standardized drug responses as given in Table 8 are trained once for 200 and 40 iterations. The later number of iterations has been choosen as a second heuristic in addition to the default 200 iterations suggested in the BMTMKL implementation of Mr. Gönen. The wpc-indices for the models in Table 8 trained on standardized drug responses are given in Table 21. Using the goldstandard, i.e. empirically obtained true drug responses, would yield a wpc-index of 0.823.

The model of hyperparameters set all to one nearly obtains the benchmark set in Costello et al. (2014) by the team around Gönen with a wpc-index of 0.582. As 40 iterations are in the range of the number of iterations obtaining the maximum for the MSEs of the model of ones this is plausible. Since setting hyperparameters to one can be justified at least of ϵ_λ and ϵ_θ, choosing the ones model as a prototyp for trails and new data is suggested.

The default model yields in terms of squared error as well as in terms of rankings no good results, indicating that smaller MSE means better rankings. No models trained on

[20] See file "calculate_zscore_weights.pl" on www.synapse.org/#!Synapse:syn2785785

Table 21: wpc- index for different models trained on standardized drug responses

	iterations		
model	40	200	truth
default	0.5066	0.461093	
ones	**0.5823**	0.570913	
maxELBO	0.5741	0.556977	0.8229
minELBO	0.5048	0.556977	
random	0.5806	0.580159	
minMSE ins	0.5737		
minMSE ofs	0.5750		
Gönen	0.583		

Table 22: Number of iterations and wpc-index for minimal MSEs of default, ones and random model

	Default	ones	random
MSE train set	120	33	91
wpc-index	0.4930	0.5837	0.5776
MSE test set (rescaled predictions)	29.163	28.088	28.087
MSE train set (rescaled)	145	31	87
wpc-index	0.4666	0.5836	0.5767
MSE test set (rescaled predictions)	29.163	28.091	28.088
MSE test set (rescaled predictions)	106	41	158
wpc-index	0.5091	0.5821	0.5710
MSE test set (rescaled predictions)	29.163	28.083	28.074

standardized drug responses performed well in terms of the wpc-index or MSE, and have thus been discarded from further analysis.

If Gönen used standardized or non-standardized drug responses, which hyperparameters and how many iterations he used, has not been published and an attempt to acquire these information was unsuccessful.

Considering to train the optimal number of iterations for the MSEs reported in Table 20 of the default, ones and random model yields the results reported in Table 22.

The ones and default model perform better for less iterations using the optimal number of iterations calculated on the training data. The default model is for all three configurations either worse than random or nearly random. Using the optimal solutions in terms of the training MSEs for the ones model trained on standardized drug data, outperforms the result obtained by Gönen. However, the wpc-index has been developed solely for the challenge and is not applied on different applications. Its theoretical properties thus are open to be studied. Notice that the ones and random model obtain better wpc-indices not for the optimal number of iterations in terms of the MSE on the test cell lines. Although, all MSEs of the test data are close for both, it shows that the two measures on the training cell lines do not match perfectly. Therefore, it depends on the choice of the performance measure which model setting would be termed best.

5 Conclusions

BMTMKL provides highly flexible models for fitting drug responses to a nearly perfect match on the training data, although these are then not best generalizing in terms of out-of-sample predictions. Using the BMTMKL model with fixed hyperparameters set to one and using the best fit to the training data, the ranking reported in Costello et al. (2014) by the team of Gönen could be reproduced. The ranking measure has been developed for the DREAM7 challenge and BMTMKL performs better than random prediction. As it is a probabilistic ranking, the scoring cannot be interpreted directly in terms of correct ranked combinations of cell lines. Therefore the MSE has been considered as a second criterion to assess the performance of predicting drug response for test cell lines. Predictions can be improved compared to using only training means of drug responses, but the gain is less than ten percent even when discarding the one outlier on drug 24 which is driving 60 percent of the MSE of test cell lines.

The detailed description of the theoretical derivation of the DAG called BMTMKL optimized with VI is provided in order to check the accuracy of implementations. It makes the dimensionality reduction performed on the highdimensional genetic inputs accessible and stresses the dependency of fits based on the training cell lines. Having a balanced set of cell lines for the subtypes of cancer of interest is essential for the BMTMKL approach as it uses similarities in kernels to predict drug responses. These drug responses should be standardized. Drug concentrations are not directly comparable as drugs target different mechanisms in cells. The scale of concentration is not provided in the challenge data and thus drug concentrations measured under laboratory conditions might be rejected by doctors as non reproducible in the human body.

Despite the caveat that a cell line in a laboratory cannot be set equal to a human body, ranking cell lines and predicting the level of concentration needed to inhibit 50 percent cell growth can help doctors to select patients into clinical trails. Statistical information is then one additional guidance for finding new combinations of drugs. In future, genetic dispositions of patients might indicate which combinations of drugs should be used for treatment.

Besides being a very flexible modeling approach, BMTMKL has nonetheless limitations. First, it cannot model gene specific relations directly as the method is based on the similarities between cell lines measured in the feature space defined by kernels. A kernel on pathways including the specific gene has to be computed to model influences of single genes. For example, the original datasets could be subdivided to cover only pathways known to be

important from medical research, i.e. to be involved in the abnormal growth of a cancer cell of the specific subtypes of cancer. A subset of the data only containing genes related to the pathway of importance would then yield a kernel coefficient e_k related to the specific gene. Second, BMTMKL is unable to calculate interactions of drugs. Two drugs combined could have a longer lasting effect, whereas a single drug is often not useful regarding long term success in therapy. Or, chemical reactions along pathways might be blocked only by using two or more drugs simultaneously. For a comprehensive introduction to the search of drug combinations, see Mukherjee (2011).

Understanding the data seems as important as using an appropriate algorithm. After having described a very adaptive algorithm with many parameters, it becomes apparent that the underlying data has to be understood more thoroughly. How comparable are the subsets of drugs? Are there pathways describing the subset of cell lines better as they are connected to the classification of cancer subtypes? For example, Iorio et al. (2016) map pathway information on human cell lines and then identify drug response for the found human cell line. Are there better alternatives to the $-\log_{10} GI_{50}$ drug response measure which are more comparable between types of drugs? At least drugs which show effectiveness in a range of concentration which doctors deem unpractical considering the human body should be removed from the sample.

Further statistical research could also evaluate the differences between models found by VI in comparison to Monte Carlo Markov Chain inference approaches.

This master thesis covered one promising model for predicting drug response on the basis of gene information. It will be interesting to see if this approach or others used in the field will be revealed as robust enough for clinical applications (Costello et al., 2014; De Niz et al., 2016).

References

Beal, M. J. (2003): "Variational algorithms for approximate bayesian inference," Phd. Thesis.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*, vol. 4, New York: Springer.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017): "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877.

Carlson, M. (2017): *org.Hs.eg.db: Genome wide annotation for Human*, r package version 3.5.0.

Corts-Ciriano, I., G. J. P. van Westen, G. Bouvier, M. Nilges, J. P. Overington, A. Bender, and T. E. Malliavin (2016): "Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel," *Bioinformatics*, 32, 85–95.

Costello, J. (2018): personal communication.

Costello, J. C., L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S. a. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky (2014): "A community effort to assess and improve drug sensitivity prediction algorithms." *Nature biotechnology*, 32, 20–23.

Daemen, A., O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, F. Pepin, S. Durinck, J. E. Korkola, M. Griffith, J. S. Hur, N. Huh, J. Chung, L. Cope, M. J. Fackler, C. Umbricht, S. Sukumar, P. Seth, V. P. Sukhatme, L. R. Jakkula, Y. Lu, G. B. Mills, R. J. Cho, E. A. Collisson, L. J. van't Veer, P. T. Spellman, and J. W. Gray (2013): "Modeling precision treatment of breast cancer," *Genome Biology*, 14, R110.

——— (2015): "Modeling precision treatment of breast cancer," *Genome Biology*, 14, 95.

De Niz, C., R. Rahman, X. Zhao, and R. Pal (2016): "Algorithms for Drug Sensitivity Prediction," *Algorithms*, 9.

Gönen, M. (2012a): "A Bayesian Multiple Kernel Learning Framework for Single and Multiple Output Regression," *20Th European Conference on Artificial Intelligence (Ecai 2012)*, 242, 354–359.

Gönen, M. (2012b): "Bayesian Efficient Multiple Kernel Learning," *Proceedings of the 29th International Conference on Machine Learning*.

Gönen, M. (2017): personal communication.

Hoffman, M. D., D. M. Blei, C. Wang, J. Paisley, and J. Edu (2013): "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347.

Iorio, F., T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. Wessels, J. Saez-Rodriguez, U. McDermott, and M. J. Garnett (2016): "A Landscape of Pharmacogenomic Interactions in Cancer," *Cell*, 166, 740 – 754.

Ligtenberg, W. (2017): *reactome.db: A set of annotation maps for reactome*, r package version 1.62.0.

Monks, A., D. Scudiero, P. Skehan, R. Shoemaker, K. Paull, D. Vistica, C. Hose, J. Langley, P. Cronise, and A. Vaigro-Wolff (1991): "Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines." *Journal of the National Cancer Institute*, 83, 757–66.

Mukherjee, S. (2011): *The Emperor of all maladies. A biography of cancer*, London: Forth Estate.

Murphy, P. K. (2012): *Machine Learning: A Probabilistic Perspective*, Cambridge, MA and London, England: MIT Press.

Petersen, K. B. and M. S. Pedersen (2012): "The Matrix Cookbook," Version of 15th Nov. 2012.

Robbins, H. and S. Monro (1951): "A Stochastic Approximation Method," *Ann. Math. Statist.*, 22, 400–407.

Sasane, A. (2017): "Functional analysis and its applications," .

Vaske, C., S. C Benz, J. Zachary Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. Stuart (2010): "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," 26, 237–45.

Wainwright, M. J. and M. I. Jordan (2008): "Graphical Models, Exponential Families, and Variational Inference," *Found. Trends Mach. Learn.*, 1, 1–305.

A Background Information

A.1 Distributions

A.1.1 Gamma distribution

A gamma distribution of $\lambda \in (0, \infty)$ with shape parameter α and rate parameter β, which is the inverse scale ($\beta = 1/\theta$), has the following density function:

$$p(\lambda|\, \alpha_\lambda, \beta_\lambda) = \frac{1}{\Gamma(\alpha_\lambda)} \beta_\lambda^{\alpha_\lambda} \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda) \tag{45}$$

The gamma function is defined as $\Gamma(\alpha_\lambda) = \int_0^\infty x^{\alpha_\lambda - 1} \exp(-x)\, dx$, where α_λ is a positive real integer in the model trained here. For a positive integer n the gamma function is the factorial function shifted by one $\Gamma(n) = (n-1)!$.

A gamma distributed random variable is used as the precision for a second normally distributed random variable. In the model used this is for example the case for the normally distributed kernel weight $a_{t,i}$ of cell line i and drug sample t, which has as precision the gamma distribute random variable $\lambda_{t,i}$. Note that the variance or precision is always non-negative, which corresponds with the support of a gamma distributed random variable .
Using not $\lambda_{t,i}$ as an example from the model, the density can be expressed in exponential notation,i.e. the form $p(\lambda|\eta) = h(\lambda) \exp\{\eta' t(\lambda) - a(\eta)\}$:

$$\begin{aligned}
p(\lambda_{t,i}|\alpha_\lambda, \beta_\lambda) &= \mathcal{G}am(\lambda_{t,i}\,|\alpha_\lambda, \beta_\lambda) \\
&= \frac{1}{\Gamma(\alpha_\lambda)} \beta_\lambda^{\alpha_\lambda} \lambda_{t,i}^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda_{t,i}) \\
&= \exp\{-\log\Gamma(\alpha_\lambda) + \alpha_\lambda \log(\beta_\lambda) + (\alpha_\lambda - 1)\log(\lambda_{t,i}) - \beta_\lambda \lambda_{t,i}\} \\
&= \underbrace{\lambda_{t,i}^{-1}}_{h(\lambda_{t,i})} \exp\{\underbrace{(-\beta_\lambda, \alpha_\lambda)}_{\eta'(\lambda_{t,i})} \underbrace{(\lambda_{t,i}, \log(\lambda_{t,i}))'}_{t(\lambda_{t,i})} \underbrace{- \log\Gamma(\alpha_\lambda) + \alpha_\lambda \log(\beta_\lambda)}_{-a(\eta(\lambda_{t,i}))}\} \\
&= h(\lambda_{t,i}) \exp\{\eta'(\lambda_{t,i}) \cdot t(\lambda_{t,i}) - a(\eta(\lambda_{t,i}))\}\ , \eta'(\lambda_{t,i}) = \eta'_{\lambda_{t,i}} = (-\beta_\lambda, \alpha_\lambda)
\end{aligned}$$

Sometimes the gamma distribution is parameterized with the so called scale parameter $\theta_\lambda = 1/\beta_\lambda$, which is the inverse of the rate parameter β_λ used in the exponential family defintion. Then we have

$$\begin{aligned}
\lambda_{t,i} \sim \mathcal{G}am(\alpha_\lambda, \beta_\lambda) &= \mathcal{G}am\left(\alpha_\lambda, \tfrac{1}{\theta_\lambda}\right) \\
\Rightarrow \quad p\left(\lambda_{t,i}|\alpha_\lambda, \tfrac{1}{\theta_\lambda}\right) &= \mathcal{G}am\left(\lambda_{t,i}\,|\alpha_\lambda, \tfrac{1}{\theta_\lambda}\right) = \frac{1}{\Gamma(\alpha_\lambda)} \frac{1}{\theta_\lambda^{\alpha_\lambda}} \lambda_{t,i}^{\alpha_\lambda - 1} \exp\left(-\frac{\lambda_{t,i}}{\theta_\lambda}\right)
\end{aligned}$$

Both can be rewritten into a standard gamma distribution with scale or rate parameter fixed to one:

$$\mathcal{G}am(x|\alpha,\theta) = \frac{1}{\theta}\cdot\mathcal{G}am\left(\frac{x}{\theta}\middle|\alpha,1\right) = \beta\cdot\mathcal{G}am(\beta x|\alpha,1) = \mathcal{G}am(x|\alpha,\beta)$$

The expectation of (the log of) a gamma distributed random variable $\lambda_{t,i}$ governed by its shape parameter α_λ and rate parameter $\beta_{\lambda_{t,i}}$ are $\mathrm{E}[\lambda_{t,i}] = \alpha_{\lambda_{t,i}}/\beta_{\lambda_{t,i}}$ and $\mathrm{E}[\log\lambda_{t,i}] = \psi(\alpha_{\lambda_{t,i}}) - \log\beta_{\lambda_{t,i}}$, where $\psi(\alpha_{\lambda_{t,i}}) = \frac{\partial}{\partial\lambda_{t,i}}\log\Gamma(\lambda_{t,i})$ is the digamma function. Using the scale notation instead with $\theta_{\lambda_{t,i}} = 1/\beta_{\lambda_{t,i}}$ the expectations are $\mathrm{E}[\lambda_{t,i}] = \alpha_{\lambda_{t,i}}\cdot\theta_{\lambda_{t,i}}$ and $\mathrm{E}[\log\lambda_{t,i}] = \psi(\alpha_{\lambda_{t,i}}) + \log\theta_{\lambda_{t,i}}$, A brief discussion of the gamma distribution can be found in Bishop (2006, p.688) where the rate notation was used.

In Costello et al. (2014) and other papers of M. Gönen the gamma distribution is parametrized with the scale parameter instead of the rate parameter, however, the scale is denoted as β, not θ. To avoid confusion thus both relevant parameterizations have been defined.

A.1.2 Normal distribution

A univariate normally distributed random variable x with support $x \in \mathbb{R}$, mean μ and variance σ^2 or precision $\lambda = 1/\sigma^2$ has as density function:

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) = p(x|\mu,\lambda^{-1}) \quad (46)$$

A mulitvariate normally distributed random variable x with support $x \in \mathbb{R}^d$, mean $\mu \in \mathbb{R}^d$ and variance $\Sigma \in \mathbb{R}^{d\times d}$ or precision $\Lambda = \Sigma^{-1}$ has as density function:

$$\begin{aligned}
p\left(x|\mu,\Sigma\right) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\
&= \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(x-\mu)'\Lambda(x-\mu)\right) = p(x|\mu,\Lambda) && (47)\\
&= (2\pi)^{-d/2}|\Lambda|^{1/2}\exp\left\{-\tfrac{1}{2}\left(x'\Lambda x - 2\mu'\Lambda x + \mu'\Lambda\mu\right)\right\} && (48)\\
&\propto \exp\left\{\mu'\Lambda x - \frac{1}{2}x'\Lambda x\right\} && (49)
\end{aligned}$$

The density of a multivariate normal can be expressed in exponential family notation as:

$$\begin{aligned}
p\left(x|\mu,\Sigma\right) &= (2\pi)^{-d/2}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left(\Lambda xx'\right) - 2\mu'\Lambda x - \log|\Lambda| - \mu'\Lambda\mu\right)\right\} && (50)\\
&\propto \exp\left\{\mu'\Lambda x - \frac{1}{2}\mathrm{tr}(\Lambda xx')\right\} && (51)
\end{aligned}$$

In the last line the property of traces on scalars (often referred to as the *trace trick*) $x'\Lambda x = \mathrm{tr}(x'\Lambda x) = \mathrm{tr}(xx'\Lambda) = \mathrm{tr}(\Lambda xx')$ of invariance under cyclic permutations is used. The terms

of the exponential family formulation $p(x|\eta) = h(x)\exp\{\eta' t(x) - a(\eta)\}$ are

$$\eta = (\eta_1, \eta_2)' = (\mu'\Lambda, -\tfrac{1}{2}\Lambda)' \qquad (\Rightarrow\ \mu = \Sigma\eta_1 \ \text{ and } \Sigma = \Lambda^{-1}) \tag{52}$$

$$t(x) = (x, xx')' \tag{53}$$

$$h(x) = (2\pi)^{-d/2} \tag{54}$$

$$a(\eta) = \frac{1}{2}\log|\Lambda| + \frac{1}{2}\mu'\Lambda\mu \tag{55}$$

The identity $\nabla_\eta a(\eta) = \mathrm{E}_q[t(x)]) = (\mu, \mu\mu' + \Sigma)'$ used in section 2.5 can be proven with the help of the matrix cookbook (Petersen and Pedersen, 2012).

Both the canonical and the exponential family notation can be used to identify the mean and covariance of the full conditionals in section 2.2.

A.1.3 Proportionality

Proportionality of a distribution is defined in terms of its density as

$$p(z|x) \propto \tilde{p}(z|x)$$

if there exists a constant $c \in \mathbb{R}$, sometimes referred to as normalizing constant, such that

$$p(z|x) = c \cdot \tilde{p}(z|x).$$

Since a distribution in the exponential family is characterized completely by its parameters, it suffices to identify the relevant (natural) parameters.

A.2 Basic (Bayesian) Regression

Let's assume to have a real valued random variable y_{ti} which is normally distributed. It's realization is observed and the outcome one wants to predict. Furthermore we have a bias (intercept) term b_t, coefficients $e = (e_1, \ldots, e_P)$ and inputs $x_{t,i} = (x_{t,i,1}, \ldots, x_{t,i,P})$, where b_t, e_k, $x_{t,i,k}$ all real-valued. We want to estimate the mean function $\mu(\cdot)$

$$y_{t,i} = \mu\left(x_{t,i}, e, b_t\right) + u_{t,i} = x_{t,i} \cdot e + b_t + u_{t,i},$$

where $u_{t,i}$ is a the true error. The distributional assumption for all $y_{t,i}$ for task t lead to the density of $y_t = (y_{t,1}, \ldots, y_{t,N_t})$, called likelihood, as

$$p(y_t|X_t, e, b_t, \epsilon_t) = \prod_{i=1}^{N_t} \mathcal{N}\big(y_{t,i}|x_i \cdot e + b_t\,,\ \epsilon_t\big),$$

where $X_t = (x'_{t,1}, \ldots, x'_{t,N_t})'$ is the matrix of inputs where each row represents one row vector $x_{t,i}$ and each $y_{t,i}$ is real valued. The precision ϵ_t of $y_{t,i}$ is shared between all outcomes $i = 1, \ldots, N_t$ for drug t. The log-likelihood of the data would then be

$$\begin{aligned}\log p(y_t|X_t, e, b_t, \epsilon_t) =& \sum_{i=1}^{N_t} \frac{1}{2} \left[\log \epsilon_t - \log(2\pi) - \epsilon_t (y_{t,i} - x_{t,i} \cdot e - b_t)^2\right] \\ =& \frac{1}{2} N_t \left[\log \epsilon_t - \log(2\pi)\right] - \frac{1}{2} \epsilon_t \sum_{i=1}^{N_t} (y_{t,i} - x_{t,i} \cdot e - b_t)^2\end{aligned}$$

One would then compute the maximum log likelihood by setting the derivatives of $\log(p|\cdot)$ w.r.t. (b_t), e and ϵ_t to zero in order to get the optimal solutions for b_t, e and ϵ_t.

A Bayesian approach is then to introduce (conjugate) priors (or model distributions) for the coefficients $(b_t, e')' = w$ or ϵ_t. For simplicity only an isotrope prior of the precision α of coefficient vector is introduced

$$p(b, e) = p(w) = \mathcal{N}(w|\mu_w, \Sigma_w) = \mathcal{N}\big(w \mid 0\,,\ \alpha^{-1} I_{(1+P)}\big)\ ,$$

where we set the prior mean to zero of the coefficients to zero and assign each coefficient a precision (c.p. Bishop, 2006, ch. 3.3). The normal distribution is chosen as it is the conjugate prior for the mean of normal distribution. Their joint density is now the likelihood of the outcomes as before times the prior for the coefficients:

$$p(y_t, w|X_t, \epsilon_t) = p(y_t|X_t, w, \epsilon_t) \cdot p(w|\alpha)$$

In Bayesian inference the posterior of the coefficients given the observed outcomes is one integral part of most learning algorithms, as for example expectation maximization (EM):

$$p(w|y_t, X_t, \epsilon_t) = \frac{p(y_t, w|X_t, \epsilon_t)}{p(y_t|X_t, \epsilon_t)} = c_1 \cdot p(y_t, w|X_t, \epsilon_t)$$

The evidence $p(y_t|X_t, \epsilon_t) = \int p(y_t, w|X_t, \epsilon_t) dw$ is constant w.r.t. the parameter vector w. Again we take the natural logarithm of the posterior in order to find optimal solution for the parameters:

$$\begin{aligned}\log p(w|y_t, X_t, \epsilon_t) =& \log p(y_t|X_t, \epsilon_t) + \log p(w|\alpha) + c_1 \\ =& c_2 - \frac{1}{2} \epsilon_t \sum_{i=1}^{N_t} \big(y_{t,i} - (1, x_{t,i}) w\big) + c_3 - \frac{1}{2} \alpha\, w' w + c_1\end{aligned}$$

The log of the posterior is up to a constant equivalent with the least-squares objective with a quadratic regularization term, sometimes called Ridge Regression. Here, again derivatives are

taken, this time w.r.t. w, ϵ_t and α. If one would add a prior on the precision ϵ_t the posterior normalizing evidence would be needed to compute the posterior explicitly, which is already intracable. In that case a type 2 maximum likelihood or so called evidence approximation has to be performed (Bishop, 2006, ch. 3.5).

A.3 Expectation operator:

The approximated variational distribution $q_{\lambda_{t,i}}$ has as parameters

$$
\begin{aligned}
E_{q_{-\lambda_{t,i}}}\left[\alpha_\lambda + \frac{1}{2}\right] &= \int \left(\prod_{j \in \{Z \setminus \lambda_{t,i}\}} q(z_j) \cdot \left(\alpha_\lambda + \frac{1}{2}\right) \, dz_{-\lambda_{t,i}} \right) \\
&= \left(\alpha_\lambda + \frac{1}{2}\right) \int \left(\prod_{j \in \{Z \setminus \lambda_{t,i}\}} q(z_j) \, dz_{-\lambda_{t,i}} \right) = \alpha_\lambda + \frac{1}{2} \\
E_{q_{-\lambda_{t,i}}}\left[\beta_\lambda + \frac{1}{2} a_{t,i}^2\right] &= \int \prod_{j \in \{Z \setminus \lambda_{t,i}\}} q(z_j) \cdot \left(\beta_\lambda + \frac{1}{2} a_{t,i}^2\right) dz_{-\lambda_{t,i}} \\
&= \beta_\lambda + \frac{1}{2} \int \prod_{j \in \{Z \setminus \lambda_{t,i}\}} q(z_j) \cdot a_{t,i}^2 \, dz_{-\lambda_{t,i}} \\
&= \beta_\lambda + \frac{1}{2} \int_{-\infty}^{\infty} q(a_{t,i}) \cdot a_{t,i}^2 \, da_{t,i} = \beta_\lambda + \frac{1}{2} E_{q_{a_{t,i}}}\left[a_{t,i}^2\right] \\
&= \beta_\lambda + \frac{1}{2}\left(\mu_{a_{t,i}}^2 + \sigma_{a_{t,i}}^2\right)
\end{aligned}
$$

The integral over the product of variational approximations of each factor $q_j = q(z_j)$, $z_j \in \{Z \setminus \lambda_{t,i}\}$, is an product of integrals over each approximated distributions support.

B Definitions

- **drug response**, susceptibility, sensitivity: ability of a microorganism to be inhibited or killed by the drug. Concentration of a drug in a cell line achieving 50 percent growth inhibition is used as reference. For the challenge the concentration is transformed by $-\log_{10}$
- **prior:** Density of a random variable in the initial model (data generating process)
- **posterior:** Posterior is short for posterior probability density of all latent variables given the evidence
- **evidence:** Observed variables, drug sensitivity and the different gene information data, summarized in set $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

- **Set minus abbreviation:** $\mathcal{Z}_{-\lambda_{t,i}} = \{\mathcal{Z} \setminus \lambda_{t,i}\}$
- **Expectation w.r.t. variational distribution of $\mathcal{Z}$:** $\mathrm{E}_q[p(\tau)] = \mathrm{E}_{q_\tau}[p(\tau)]$ for the set of random variables or factor τ, since density $q(\mathcal{Z})$ is a product of densities of independent random variables.
- **L2 norm** or euclidean norm (distance) $\|x\|$ is defined as $\sqrt{x'x}$ for random variable x.
- **Fully conjugated model** All random variables are in the exponential family and the conditional density of a latent variable given all other latent variables and the data, i.e. the full or complete conditional, is defined by the same exponential family distribution as the prior (Beal, 2003, ch. 2.4, p. 64ff.).
- **Kullback-Leibler-Divergence**: $\mathrm{KL}(q(\tau)|p(\tau)) = \mathrm{E}_{q(\tau)}\left[\log \frac{q(\tau)}{p(\tau)}\right] = -\mathrm{E}_q\left[\log \frac{p(\tau)}{q(\tau)}\right]$ and has the property of being always positive, $\mathrm{KL}(q(\tau)|p(\tau)) \geq 0$ and only equal to 0 almost everywhere (w.r.t to q) iff q=p. Proof:

$$\begin{aligned} \mathrm{E}_{q(\tau)}\left[\log \frac{q(\tau)}{p(\tau)}\right] &= \mathrm{E}_q\left[-\log \frac{p(\tau)}{q(\tau)}\right] \\ \text{(Jensen inequality)} \qquad &\geq \log \mathrm{E}_q\left[\frac{p(\tau)}{q(\tau)}\right] = \log \mathrm{E}[p(\tau)] = \log 1 = 0 \end{aligned}$$

  which holds since $p(\tau)$ is a probability density function which integrates to one over its support and both $p(\tau)$ and $q(\tau)$ are as probability density functions always non-negative. Note that Jensen is defined for convex function, which is why the negative logarithm has to be used.

C Tables

Table 23: Variational densities - full conditionals with updated parameters

r.v.	variational densities ($\widetilde{k} = \mathrm{E}_{q_k}[k]$)
λ	$q(\lambda) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} \mathcal{G}am\left(\lambda_{t,i} \middle\vert\, \alpha_\lambda + \frac{1}{2},\ \beta_\lambda + \frac{1}{2}\widetilde{a_{t,i}^2}\right)$
a	$q(a) = \prod_{t=1}^{T} \mathcal{N}\left(a_t \,\middle\vert\, \Sigma_{a_t}\left(\sum_{k=1}^{P} K_{t,k}\widetilde{g_{t,k}}\right)\widetilde{v}_t\,,\ \left(\widetilde{\Lambda}_t + \widetilde{v}_t \sum_{k=1}^{P} K_{t,k}K_{t,k}\right)^{-1}\right)$
G	$q(G) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} \mathcal{N}(G'_{t,i} \,\vert\, \Sigma_{G'_{t,i}} \cdot \left[\widetilde{v}_t K_{t,i}\widetilde{a}_t + \widetilde{\epsilon}_t(y_{t,i}\widetilde{e} - \widetilde{b_t e})\right]\,,\ (\widetilde{v}_t I_P + \widetilde{\epsilon}_t\widetilde{ee'})^{-1})$
υ	$q(\upsilon) = \prod_{t=1}^{T} \mathcal{G}am\left(\upsilon_t \middle\vert\, \alpha_\upsilon + \frac{P \cdot N_t}{2}\,,\ \beta_\upsilon + \frac{\tilde{c}_{\upsilon_t}}{2}\right), c_{\upsilon_t} = \sum_{k=1}^{P} \lVert g_{t,k} - K_{t,k}a_t\rVert^2$
γ	$q(\gamma) = \prod_{t=1}^{T} \mathcal{G}am\left(\gamma_t \middle\vert\, \alpha_\gamma + \frac{1}{2}\,,\ \beta_\gamma + \frac{1}{2}\tilde{b}_t^2\right)$
ω	$q(\omega) = \prod_{k=1}^{P} \mathcal{G}am\left(\omega_k \middle\vert\, \alpha_\omega + \frac{1}{2}\,,\ \beta_\omega + \frac{1}{2}\tilde{e}_k^2\right)$
b, e	$q(b,e) = \mathcal{N}\left(\begin{matrix} b \\ e \end{matrix} \middle\vert\, \Sigma_{(b',e')'} \cdot \begin{pmatrix} y_1'\tilde{\epsilon}_1\mathbb{1}_{N_1} \\ \vdots \\ y_T'\tilde{\epsilon}_T\mathbb{1}_{N_T} \\ \sum_t \tilde{\epsilon}_t\tilde{G}_t'y_t \end{pmatrix},\ \begin{pmatrix} \tilde{\gamma}_1 + N_1\tilde{\epsilon}_1 & & & \tilde{\epsilon}_1\mathbb{1}_{N_1}'\tilde{G}_1 \\ & \ddots & & \vdots \\ & & \tilde{\gamma}_T + N_T\tilde{\epsilon}_T & \tilde{\epsilon}_T\mathbb{1}_{N_T}'\tilde{G}_T \\ \tilde{\epsilon}_1\tilde{G}_1'\mathbb{1}_{N_1} & \dots & \tilde{\epsilon}_T\tilde{G}_T'\mathbb{1}_{N_T} & \mathrm{diag}(\tilde{\omega}) + \sum_t \tilde{G}_t'\tilde{\epsilon}_t\tilde{G}_t \end{pmatrix}^{-1}\right)$
ϵ	$q(\epsilon) = \prod_{t=1}^{T} \mathcal{G}am\left\{\epsilon_t \,\middle\vert\, \alpha_\epsilon + \frac{N_t}{2}\,,\ \beta_\epsilon + \frac{\tilde{c}_{\epsilon_t}^2}{2}\right\}\quad, c_{\epsilon_t} = \lVert y_t - G_t e - b_t \cdot \mathbb{1}_{N_t}\rVert$

Table 24: Parameter updates of variational distributions

$$
\begin{aligned}
\alpha^{\star}_{\lambda_{t,i}} = \alpha^{\star}_{\lambda} = \alpha_{\lambda} + \tfrac{1}{2} \qquad & \beta^{\star}_{\lambda_{t,i}} = \beta_{\lambda} + \tfrac{1}{2}(\mu^{\star 2}_{a_{t,i}} + \sigma^{2\star}_{a_{t,i}}) \\
\alpha^{\star}_{\upsilon_t} = \alpha^{\star}_{\upsilon} = \alpha_{\upsilon} + \tfrac{P \cdot N}{2} \qquad & \beta^{\star}_{\upsilon_t} = \beta_{\upsilon} + \tfrac{1}{2}\Bigg(\operatorname{tr}\left[\mu^{\star\prime}_{G_t}\mu^{\star}_{G_t} + N_t \cdot \Sigma^{\star}_{G'_{t,\cdot}}\right] - 2\mu^{\star\prime}_{g_t} K_t \mu^{\star}_{a_t} \\
& \qquad + \operatorname{tr}\left[\left(\sum_{k=1}^{P} K_{t,k}K_{t,k}\right)\left(\mu^{\star}_{a_t}\mu^{\star\prime}_{a_t} + \Sigma^{\star}_{a_t}\right)\right]\Bigg) \\
\alpha^{\star}_{\gamma_t} = \alpha^{\star}_{\gamma} = \alpha_{\gamma} + \tfrac{1}{2} \qquad & \beta^{\star}_{\gamma_t} = \beta_{\gamma} + \tfrac{1}{2}(\mu^{\star 2}_{b_t} + \sigma^{2\star}_{b_t}) \\
\alpha^{\star}_{\omega_k} = \alpha^{\star}_{\omega} = \alpha_{\omega} + \tfrac{1}{2} \qquad & \beta^{\star}_{\omega_k} = \beta_{\omega} + \tfrac{1}{2}(\mu^{\star 2}_{e_k} + \sigma^{2\star}_{e_k}) \\
\alpha^{\star}_{\epsilon_t} = \alpha_{\epsilon} + \tfrac{N_t}{2} \qquad & \beta^{\star}_{\epsilon_t} = \beta_{\epsilon} + \tfrac{1}{2}\Bigg(y_t'y_t - 2y_t'\mu^{\star}_{G_t}\mu^{\star}_{e} - 2y_t'\mathbb{1}_{N_t}\mu^{\star}_{b_t} + 2\mu^{\star}_{b_t}\mathbb{1}'_{N_t}\mu^{\star}_{G_t}\mu^{\star}_{e} \\
& \qquad - \operatorname{tr}\left[\left(\mu^{\star\prime}_{G_t}\mu^{\star}_{G_t} + N_t \cdot \Sigma^{\star}_{G'_{t,\cdot}}\right)\left(\mu^{\star}_{e}\mu^{\star\prime}_{e} + \Sigma^{\star}_{e}\right)\right]\Bigg)
\end{aligned}
$$

the updates of the normal do not invole expecatations of squared random variables:

$$
\begin{aligned}
\mu^{\star}_{a_t} &= \Sigma^{\star}_{a_t} \cdot K_t' \cdot \mu^{\star}_{g_t} \frac{\alpha^{\star}_{\upsilon_t}}{\beta^{\star}_{\upsilon_t}} \quad , \quad \Sigma^{\star}_{a_t} = \left(\operatorname{diag}\left(\frac{\alpha^{\star}_{\lambda_{t,1}}}{\beta^{\star}_{\lambda_{t,1}}}, \ldots, \frac{\alpha^{\star}_{\lambda_{t,N_t}}}{\beta^{\star}_{\lambda_{t,N_t}}}\right) + \frac{\alpha^{\star}_{\upsilon_t}}{\beta^{\star}_{\upsilon_t}} \sum_{k=1}^{P} K_{t,k}K_{t,k}\right)^{-1} \\
\mu^{\star}_{G_t} &= \frac{\alpha^{\star}_{\upsilon_t}}{\beta^{\star}_{\upsilon_t}} \left(K_{t,1}\mu^{\star}_{a_t}, \ldots, K_{t,K}\mu^{\star}_{a_t}\right) \cdot \Sigma^{\star}_{G'_{t,\cdot}} + \frac{\alpha^{\star}_{\epsilon_t}}{\beta^{\star}_{\epsilon_t}} \left(y_t - \mathbb{1}_{N_t}\frac{\alpha^{\star}_{b_t}}{\beta^{\star}_{b_t}}\right)\mu^{\star\prime}_{e} \quad , \\
\Sigma^{\star}_{G'_{t,\cdot}} &= \Sigma^{\star}_{G'_{t,i}} = \left(\frac{\alpha^{\star}_{\upsilon_t}}{\beta^{\star}_{\upsilon_t}} I_P + \frac{\alpha^{\star}_{\epsilon_t}}{\beta^{\star}_{\epsilon_t}}\left(\mu^{\star}_{e}\mu^{\star\prime}_{e} + \Sigma^{\star}_{e}\right)\right)^{-1} \\
\mu^{\star}_{b,e} &= \Sigma^{\star}_{b,e} \cdot \left(\frac{\alpha^{\star}_{\epsilon_1}}{\beta^{\star}_{\epsilon_1}} y_1'\mathbb{1}_{N_1}, \ldots, \frac{\alpha^{\star}_{\epsilon_T}}{\beta^{\star}_{\epsilon_T}} y_T'\mathbb{1}_{N_T}, \sum_{t=1}^{T} \frac{\alpha^{\star}_{\epsilon_t}}{\beta^{\star}_{\epsilon_t}} y_t'\mu^{\star}_{G_t}\right)' \quad , \\
\Sigma^{\star}_{b,e} &= \begin{pmatrix}
\frac{\alpha^{\star}_{\gamma_1}}{\beta^{\star}_{\gamma_1}} + N_1\frac{\alpha^{\star}_{\epsilon_1}}{\beta^{\star}_{\epsilon_1}} & & & \frac{\alpha^{\star}_{\epsilon_1}}{\beta^{\star}_{\epsilon_1}}\mathbb{1}'_{N_1}\mu^{\star}_{G_1} \\
& \ddots & & \vdots \\
& & \frac{\alpha^{\star}_{\gamma_T}}{\beta^{\star}_{\gamma_T}} + N_T\frac{\alpha^{\star}_{\epsilon_T}}{\beta^{\star}_{\epsilon_T}} & \frac{\alpha^{\star}_{\epsilon_T}}{\beta^{\star}_{\epsilon_T}}\mathbb{1}'_{N_T}\mu^{\star}_{G_T} \\
\frac{\alpha^{\star}_{\epsilon_1}}{\beta^{\star}_{\epsilon_1}}\mu^{\star\prime}_{G_1}\mathbb{1}_{N_1} & \cdots & \frac{\alpha^{\star}_{\epsilon_T}}{\beta^{\star}_{\epsilon_T}}\mu^{\star\prime}_{G_T}\mathbb{1}_{N_T} & \operatorname{diag}(\frac{\alpha^{\star}_{\omega}}{\beta^{\star}_{\omega}}) + \sum_t \mu^{\star\prime}_{G_t}\frac{\alpha^{\star}_{\epsilon_t}}{\beta^{\star}_{\epsilon_t}}\mu^{\star}_{G_t}
\end{pmatrix}^{-1}
\end{aligned}
$$

Table 25: Terms of ELBO expressed as variational parameters of random variables

$$
\begin{aligned}
\mathrm{E}_{q_\lambda}[\log p(\lambda)] =& \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Bigg(-\log\Gamma(\alpha_\lambda) + \alpha_\lambda\cdot\log\beta_\lambda \\
&+(\alpha_\lambda-1)\cdot\left(\psi(\alpha^\star_{t,i}) - \log\beta^\star_{\lambda_{t,i}}\right) - \beta_\lambda\left(\alpha^\star_{\lambda_{t,i}}/\beta^\star_{\lambda_{t,i}}\right)\Bigg) \\
\mathrm{E}_{q_a,q_\lambda}[\log p(a|\lambda)] =& \tfrac{1}{2}\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t}\left[\psi(\alpha^\star_{\lambda_{t,i}}) - \log\beta^\star_{\lambda_{t,i}} - \log(2\pi) - \frac{\alpha^\star_{\lambda_{t,i}}}{\beta^\star_{\lambda_{t,i}}}\left(\mu^{\star 2}_{a_{t,i}} + \sigma^{\star 2}_{a_{t,i}}\right)\right]\right) \\
=& \tfrac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\psi(\alpha^\star_{\lambda_{t,i}}) - \log\beta^\star_{\lambda_{t,i}}\right) - \tfrac{N}{2}\log(2\pi) \\
&-\sum_{t=1}^{T}\mathrm{tr}\left[\mathrm{diag}\left(\frac{\alpha^\star_{\lambda_{t,1}}}{\beta^\star_{\lambda_{t,1}}},\ldots,\frac{\alpha^\star_{\lambda_{t,N_t}}}{\beta^\star_{\lambda_{t,N_t}}}\right)\left(\mu^\star_{a_t}\mu^{\star\prime}_{a_t} + \Sigma^\star_{a_t}\right)\right] \\
\mathrm{E}_{q_G q_a q_\upsilon}[\log p(G|a,\upsilon,\mathcal{X})] =& \sum_{t=1}^{T}\tfrac{N_t\cdot P}{2}\cdot\left(\psi(\alpha^\star_{\upsilon_t}) - \log\beta^\star_{\upsilon_t}\right) - NP\log(2\pi) \\
&-\tfrac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{\alpha^\star_{\upsilon_t}}{\beta^\star_{\upsilon_t}}\Bigg[\mathrm{tr}\left[\left(\mu^\star_{G'_{t,i}}\mu^{\star\prime}_{G'_{t,i}} + \Sigma^\star_{G'_{t,i}}\right)\right] - 2\mu'_{G'_{t,i}}K_{t,i}\,\mu_{a_t} \\
&+\mathrm{tr}\left[K'_{t,i}K_{t,i}\left(\mu^\star_{a_t}\mu^{\star\prime}_{a_t} + \Sigma^\star_{a_t}\right)\right]\Bigg] \\
=& \sum_{t=1}^{T}\tfrac{1}{2}\Bigg(N_t\cdot P\cdot\left(\psi(\alpha^\star_{\upsilon_t}) - \log\beta^\star_{\upsilon_t}\right) - \sum_{t=1}^{T}N_t\log(2\pi) \\
&-\frac{\alpha^\star_{\upsilon_t}}{\beta^\star_{\upsilon_t}}\Bigg[\mathrm{tr}\left[\mu^{\star\prime}_{G_t}\mu^\star_{G_t} + N_t\cdot\Sigma^\star_{G'_{t,\cdot}}\right] - 2\mu'_{g_t}K_t\mu_{a_t} \\
&+\mathrm{tr}\left[\sum_{k=1}^{K}K_{t,k}K_{t,k}\left(\mu^\star_{a_t}\mu^{\star\prime}_{a_t} + \Sigma^\star_{a_t}\right)\right]\Bigg]\Bigg) \\
\mathrm{E}_{q_\upsilon}[\log p(\upsilon)] =& T\left[\alpha_\upsilon\log\beta_\upsilon - \log\Gamma(\alpha_\upsilon)\right] + (\alpha_\upsilon-1)\sum_{t=1}^{T}\left(\psi(\alpha^\star_{\upsilon_t}) - \log\beta^\star_{\upsilon_t}\right) \\
&-\beta_\upsilon\cdot\sum_{t=1}^{T}\frac{\alpha^\star_{\upsilon_t}}{\beta^\star_{\upsilon_t}} \\
\mathrm{E}_{q_\omega}[\log p(\omega)] =& P\left[\alpha_\omega\log\beta_\omega - \log\Gamma(\alpha_\omega)\right] + (\alpha_\omega-1)\sum_{k=1}^{P}\left(\psi(\alpha^\star_{\omega_k}) - \log\beta^\star_{\omega_k}\right) \\
&-\beta_\omega\cdot\sum_{k=1}^{P}\frac{\alpha^\star_{\omega_k}}{\beta^\star_{\omega_k}} \\
\mathrm{E}_{q_{b,e}q_\gamma q_\omega}[\log p(b,e|\gamma,\omega)] =& -\tfrac{1}{2}(T+P)\log(2\pi) + \tfrac{1}{2}\sum_{t=1}^{T}\left(\psi(\alpha^\star_{\gamma_t}) - \log\beta^\star_{\gamma_t}\right) \\
&+\tfrac{1}{2}\sum_{k=1}^{P}\left(\psi(\alpha^\star_{\omega_k}) - \log\beta^\star_{\omega_k}\right) \\
&-\tfrac{1}{2}\mathrm{tr}\left[\mathrm{diag}\left(\frac{\alpha^\star_{\gamma_1}}{\beta^\star_{\gamma_1}},\ldots,\frac{\alpha^\star_{\gamma_T}}{\beta^\star_{\gamma_T}},\frac{\alpha^\star_{\omega_1}}{\beta^\star_{\omega_1}},\ldots,\frac{\alpha^\star_{\omega_P}}{\beta^\star_{\omega_P}}\right)\left(\mu^\star_{b,e}\mu^{\star\prime}_{b,e} + \Sigma^\star_{b,e}\right)\right] \\
\mathrm{E}_{q_\gamma}[\log p(\gamma)] =& T\left[\alpha_\gamma\log\beta_\gamma - \log\Gamma(\alpha_\gamma)\right] + (\alpha_\gamma-1)\sum_{t=1}^{T}\left(\psi(\alpha^\star_{\gamma_t}) - \log\beta^\star_{\gamma_t}\right) \\
&-\beta_\gamma\cdot\sum_{t=1}^{T}\frac{\alpha^\star_{\gamma_t}}{\beta^\star_{\gamma_t}} \\
\mathrm{E}_{q_\epsilon}[\log p(\epsilon)] =& T\left[\alpha_\epsilon\log\beta_\epsilon - \log\Gamma(\alpha_\epsilon)\right] + (\alpha_\epsilon-1)\sum_{t=1}^{T}\left(\psi(\alpha^\star_{\epsilon_t}) - \log\beta^\star_{\epsilon_t}\right) \\
&-\beta_\epsilon\cdot\sum_{t=1}^{T}\frac{\alpha^\star_{\epsilon_t}}{\beta^\star_{\epsilon_t}} \\
\mathrm{E}_{q_{b,e}q_G q_\epsilon}[\log p(\mathcal{Y}|b,e,G,\epsilon)] =& \sum_{t=1}^{T}\tfrac{1}{2}\Bigg(N_t\left(\left(\psi(\alpha^\star_{\epsilon_t}) - \log\beta^\star_{\epsilon_t}\right) - \log(2\pi)\right) \\
&-\frac{\alpha^\star_{\epsilon_t}}{\beta^\star_{\epsilon_t}}\Bigg(y'_t y_t - 2y'_t\mu_{G_t}\mu_e - 2y'_t\mathbb{1}_{N_t}\mu_{b_t} + 2\mu_{b_t}\mathbb{1}'_{N_t}\mu_{G_t}\mu_e \\
&-\mathrm{tr}\left[\left(\mu^{\star\prime}_{G_t}\mu^\star_{G_t} + N_t\cdot\Sigma^\star_{G'_{t,\cdot}}\right)\left(\mu^\star_e\mu^{\star\prime}_e + \Sigma^\star_e\right)\right]\Bigg)\Bigg) \\
-\mathrm{E}_{q_\lambda}[\log q(\lambda)] =& \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(\log\Gamma(\alpha^\star_\lambda) + (1-\alpha^\star_\lambda)\psi(\alpha^\star_\lambda) + \alpha^\star_\lambda - \log\beta^\star_{\lambda_{t,i}}\right) \\
-\mathrm{E}_{q_a}[\log q(a)] =& \sum_{t=1}^{T}\tfrac{1}{2}\left(N_t(\log(2\pi+1)) + \log|\Sigma^\star_{a_t}|\right) \\
-\mathrm{E}_{q_G}[\log q(G)] =& \sum_{t=1}^{T}\tfrac{1}{2}\left(N_t(\log(2\pi+1)) + \log\left|\Sigma^\star_{G'_{t,\cdot}}\right|\right) \\
-\mathrm{E}_{q_\upsilon}[\log q(\upsilon)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha^\star_\upsilon) + (1-\alpha^\star_\upsilon)\psi(\alpha^\star_\upsilon) + \alpha^\star_\upsilon - \log\beta^\star_{\upsilon_t}\right) \\
-\mathrm{E}_{q_\gamma}[\log q(\gamma)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha^\star_\gamma) + (1-\alpha^\star_\gamma)\psi(\alpha^\star_\gamma) + \alpha^\star_\gamma - \log\beta^\star_{\gamma_t}\right) \\
-\mathrm{E}_{q_\omega}[\log q(\omega)] =& \sum_{k=1}^{P}\left(\log\Gamma(\alpha^\star_\omega) + (1-\alpha^\star_\omega)\psi(\alpha^\star_\omega) + \alpha^\star_\omega - \log\beta^\star_{\omega_k}\right) \\
-\mathrm{E}_{q_{b,e}}[\log q(b,e)] =& \tfrac{1}{2}\left((T+P)(\log(2\pi+1)) + \log\left|\Sigma^\star_{b,e}\right|\right) \\
-\mathrm{E}_{q_\epsilon}[\log q(\epsilon)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha^\star_{\epsilon_t}) + (1-\alpha^\star_{\epsilon_t})\psi(\alpha^\star_{\epsilon_t}) + \alpha^\star_{\epsilon_t} - \log\beta^\star_{\epsilon_t}\right)
\end{aligned}
$$

Table 26: Terms of ELBO expressed in expectations of random variables

$$
\begin{aligned}
\mathrm{E}_{q_\lambda}[\log p(\lambda)] =& \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Big(-\log\Gamma(\alpha_\lambda)+\alpha_\lambda\cdot\log\beta_\lambda+(\alpha_\lambda-1)\cdot\mathrm{E}_q[\log\lambda_{t,i}]-\beta_\lambda\,\mathrm{E}_q[\lambda_{t,i}]\Big)\\
\mathrm{E}_{q_a,q_\lambda}[\log p(a|\lambda)] =& \tfrac{1}{2}\sum_{t=1}^{T}\Big(\sum_{i=1}^{N_t}\left[\mathrm{E}_q[\log\lambda_{t,i}]-\log(2\pi)-\mathrm{E}_q[\lambda_{t,i}]\cdot\mathrm{E}_q\left[a_{t,i}^2\right]\right]\Big)\\
\mathrm{E}_{q_G q_a q_\upsilon}[\log p(G|a,\upsilon,\mathcal{X})] =& \tfrac{1}{2}\sum_{t=1}^{T}\Big(N_t\cdot P\cdot\mathrm{E}_q[\log\upsilon_t]-N_tP\log(2\pi)-\mathrm{E}_q[\upsilon_t]\cdot c_{G_t}\Big)\\
&, c_{G_t}=\sum_{i=1}^{N_t}\Big[\operatorname{tr}\left(\mathrm{E}_q\left[G_{t,i}'G_{t,i}\right]\right)-2\,\mathrm{E}_q\left[G_{t,i}'\right]'K_{t,i}\,\mathrm{E}_q[a_t]+\operatorname{tr}\left(K_{t,i}'K_{t,i}\,\mathrm{E}_q[a_t a_t']\right)\Big]\\
\mathrm{E}_{q_\upsilon}[\log p(\upsilon)] =& T\left[\alpha_\upsilon\log\beta_\upsilon-\log\Gamma(\alpha_\upsilon)\right]+(\alpha_\upsilon-1)\sum_{t=1}^{T}\mathrm{E}_q[\log\upsilon_t]-\beta_\upsilon\cdot\sum_{t=1}^{T}\mathrm{E}_q[\upsilon_t]\\
\mathrm{E}_{q_\omega}[\log p(\omega)] =& P\left[\alpha_\omega\log\beta_\omega-\log\Gamma(\alpha_\omega)\right]+(\alpha_\omega-1)\sum_{k=1}^{P}\mathrm{E}_q[\log\omega_k]-\beta_\omega\cdot\sum_{k=1}^{P}\mathrm{E}_q[\omega_k]\\
\mathrm{E}_{q_{b,e}q_\gamma q_\omega}[\log p(b,e|\gamma,\omega)] =& -\tfrac{1}{2}(T+P)\log(2\pi)+\tfrac{1}{2}\sum_{t=1}^{T}\mathrm{E}_q[\log\gamma_t]+\tfrac{1}{2}\sum_{k=1}^{P}\mathrm{E}_q[\log\omega_k]\\
&-\tfrac{1}{2}\operatorname{tr}\left[\operatorname{diag}\left(\mathrm{E}_q[\gamma_1],\ldots,\mathrm{E}_q[\gamma_T],\mathrm{E}_q[\omega_1],\ldots,\mathrm{E}_q[\omega_P]\right)\cdot\mathrm{E}_q\left[(b',e')'(b',e')\right]\right]\\
\mathrm{E}_{q_\gamma}[\log p(\gamma)] =& T\left[\alpha_\gamma\log\beta_\gamma-\log\Gamma(\alpha_\gamma)\right]+(\alpha_\gamma-1)\sum_{t=1}^{T}\mathrm{E}_q[\log\gamma_t]-\beta_\gamma\cdot\sum_{t=1}^{T}\mathrm{E}_q[\gamma_t]\\
\mathrm{E}_{q_\epsilon}[\log p(\epsilon)] =& T\left[\alpha_\epsilon\log\beta_\epsilon-\log\Gamma(\alpha_\epsilon)\right]+(\alpha_\epsilon-1)\sum_{t=1}^{T}\mathrm{E}_q[\log\epsilon_t]-\beta_\epsilon\cdot\sum_{t=1}^{T}\mathrm{E}_q[\epsilon_t]\\
\mathrm{E}_{q_{b,e}q_G q_\epsilon}[\log p(\mathcal{Y}|b,e,G,\epsilon)] =& \sum_{t=1}^{T}\tfrac{1}{2}\Big(N_t\left(\mathrm{E}_q[\log\epsilon_t]-\log(2\pi)\right)-\mathrm{E}_q[\epsilon_t]\cdot c_{y_t}\Big)\\
&, c_{y_t}=y_t'y_t-2y_t'\,\mathrm{E}_q[G_t]\,\mathrm{E}_q[e]-2y_t'\mathbb{1}_{N_t}\,\mathrm{E}_q[b_t]+2\,\mathrm{E}_q[b_t]\,\mathbb{1}_{N_t}'\,\mathrm{E}_q[G_t]\,\mathrm{E}_q[e]\\
&-\operatorname{tr}\left(\mathrm{E}_q[G_t'G_t]\,\mathrm{E}_q[ee']\right)\\
-\mathrm{E}_{q_\lambda}[\log q(\lambda)] =& \sum_{t=1}^{T}\sum_{i=1}^{N_t}\Big(\log\Gamma(\alpha_\lambda^\star)+(1-\alpha_\lambda^\star)\psi(\alpha_\lambda^\star)+\alpha_\lambda^\star-\log\beta_{\lambda_t,i}^\star\Big)\\
-\mathrm{E}_{q_a}[\log q(a)] =& \sum_{t=1}^{T}\tfrac{1}{2}\left(N_t(\log(2\pi+1))+\log|\Sigma_{a_t}^\star|\right)\\
-\mathrm{E}_{q_G}[\log q(G)] =& \sum_{t=1}^{T}\tfrac{1}{2}\left(N_t(\log(2\pi+1))+\log\left|\Sigma_{G_{t,\cdot}'}^\star\right|\right)\\
-\mathrm{E}_{q_\upsilon}[\log q(\upsilon)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha_\upsilon^\star)+(1-\alpha_\upsilon^\star)\psi(\alpha_\upsilon^\star)+\alpha_\upsilon^\star-\log\beta_{\upsilon t}^\star\right)\\
-\mathrm{E}_{q_\gamma}[\log q(\gamma)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha_\gamma^\star)+(1-\alpha_\gamma^\star)\psi(\alpha_\gamma^\star)+\alpha_\gamma^\star-\log\beta_{\gamma t}^\star\right)\\
-\mathrm{E}_{q_\omega}[\log q(\omega)] =& \sum_{k=1}^{P}\left(\log\Gamma(\alpha_\omega^\star)+(1-\alpha_\omega^\star)\psi(\alpha_\omega^\star)+\alpha_\omega^\star-\log\beta_{\omega_k}^\star\right)\\
-\mathrm{E}_{q_{b,e}}[\log q(b,e)] =& \tfrac{1}{2}\left((T+P)(\log(2\pi+1))+\log\left|\Sigma_{b,e}^\star\right|\right)\\
-\mathrm{E}_{q_\epsilon}[\log q(\epsilon)] =& \sum_{t=1}^{T}\left(\log\Gamma(\alpha_{\epsilon_t}^\star)+(1-\alpha_{\epsilon_t}^\star)\psi(\alpha_{\epsilon_t}^\star)+\alpha_{\epsilon_t}^\star-\log\beta_{\epsilon_t}^\star\right)
\end{aligned}
$$

Table 27: Expectation of gamma distributed random variable τ in terms of the variational parameters $\alpha^\star_{\tau_t}$ and $\beta^\star_{\tau_t} = \frac{1}{\theta^\star_{\tau_t}}$

$$\mathrm{E}_q\left[\log \tau_t\right] = \psi\left(\alpha^\star_{\tau_t}\right) + \log \theta^\star_{\tau_t} \qquad = \psi\left(\alpha^\star_{\tau_t}\right) - \log \beta^\star_{\tau_t}$$
$$\mathrm{E}_q\left[\tau_t\right] = \alpha^\star_{\tau_t} \cdot \theta^\star_{\tau_t} \qquad = \frac{\alpha^\star_{\tau_t}}{\beta^\star_{\tau_t}}$$

Table 28: Derivatives of variational parameters w.r.t hyperparameters α_τ and β_τ of the gamma distributed random variable τ_t in rate notation

$$\frac{\partial \alpha^\star_\tau}{\partial \beta_\tau} = \frac{\partial \beta^\star_{\tau_t}}{\partial \alpha_\tau} = 0 \qquad \frac{\partial \beta^\star_{\tau_t}}{\partial \beta_\tau} = \frac{\partial \alpha^\star_\tau}{\partial \alpha_\tau} = 1$$
$$\frac{\partial}{\partial \alpha_\tau}\mathrm{E}_q\left[\tau_t\right] = \frac{1}{\beta^\star_{\tau_t}} \qquad \frac{\partial}{\partial \alpha_\tau}\mathrm{E}_q\left[\log \tau_t\right] = \frac{\partial}{\partial \alpha_\tau}\psi(\alpha^\star_\tau) = \psi'(\alpha^\star_\tau)$$
$$\frac{\partial}{\partial \beta_\tau}\mathrm{E}_q\left[\tau_t\right] = -\alpha^\star_\tau \cdot \left(\beta^\star_{\tau_t}\right)^{-2} \qquad \frac{\partial}{\partial \beta_\tau}\mathrm{E}_q\left[\log \tau_t\right] = -\frac{1}{\beta_{\tau^\star_t}} \cdot \frac{\partial}{\partial \beta_\tau}\beta^\star_{\tau_t} = -\frac{1}{\beta_{\tau^\star_t}}$$

Table 29: Derivatives of variational parameters w.r.t hyperparameters α_τ and θ_τ of the gamma distributed random variable τ_t in scale notation

$$\frac{\partial \alpha^\star_\tau}{\partial \theta_\tau} = 0 \qquad \frac{\partial \theta^\star_{\tau_t}}{\partial \alpha_\tau} = 0 \qquad\qquad \frac{\partial \theta^\star_{\tau_t}}{\partial \theta_\tau} = \frac{\left(\theta^\star_{\tau_t}\right)^2}{\theta^2_\tau} \qquad \frac{\partial \alpha^\star_\tau}{\partial \alpha_\tau} = 1$$

$$\frac{\partial}{\partial \theta_\tau \partial \theta_\tau} \theta^\star_{\tau_t} = \frac{\partial}{\partial \theta_\tau} \frac{\left(\theta^\star_{\tau_t}\right)^2}{\theta^2_\tau} = \frac{\left(\theta^\star_{\tau_t}\right)^3}{\theta^4_\tau} - \frac{\left(\theta^\star_{\tau_t}\right)^2}{\theta^3_\tau}$$

$$\frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\tau_t\right] = \theta^\star_{\tau_t} \qquad\qquad \frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\log \tau_t\right] = \frac{\partial}{\partial \alpha_\tau} \psi(\alpha^\star_\tau) = \psi'(\alpha^\star_\tau)$$

$$\frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\tau_t\right] = \alpha^\star_\tau \cdot \frac{\left(\theta^\star_{\tau_t}\right)^2}{\theta^2_\tau} \qquad\qquad \frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\log \tau_t\right] = \frac{\partial}{\partial \theta_\tau} \log \theta^\star_{\tau_t} = \frac{1}{\theta_{\tau^\star_t}} \cdot \frac{\partial}{\partial \theta_\tau} \theta^\star_{\tau_t} = \frac{\theta^\star_{\tau_t}}{\theta^2_\tau}$$

$$\frac{\partial}{\partial \theta_\tau \partial \theta_\tau} \mathrm{E}_q\left[\log \tau_t\right] = \frac{\partial}{\partial \theta_\tau} \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} = \frac{\left(\theta^\star_{\tau_t}\right)^2}{\theta^4_\tau} - \frac{\theta^\star_{\tau_t}}{\theta^3_\tau}$$

$$\mathrm{E}_q\left[\log p(\tau_t)\right] = \sum_t -\log \Gamma(\alpha_\tau) - \alpha_\tau \log \theta_\tau + (\alpha_\tau - 1)\, \mathrm{E}_q\left[\log \tau_t\right] - \frac{1}{\theta_\tau} \mathrm{E}_q\left[\tau\right]$$

$$\frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\log p(\tau_t)\right] = \sum_t -\log \theta_\tau - \psi(\alpha_\tau) + \mathrm{E}_q\left[\log \tau_t\right] + (\alpha_\tau - 1) \frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\log \tau_t\right] - \frac{1}{\theta_\tau} \frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\tau\right]$$

$$\frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\log p(\tau_t)\right] = \sum_t -\frac{\alpha_\tau}{\theta_\tau} + (\alpha_\tau - 1) \frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\log \tau_t\right] + \frac{1}{\theta^2_\tau} \mathrm{E}_q\left[\tau\right] - \frac{1}{\theta_\tau} \frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\tau\right]$$

$$-\mathrm{E}_q\left[\log q(\tau_t)\right] = \sum_t \log \Gamma(\alpha^\star_{\tau_t}) + \alpha^\star_{\tau_t} \log \theta^\star_{\tau_t} - (\alpha^\star_{\tau_t} - 1)\, \mathrm{E}_q\left[\log \tau_t\right] + \frac{1}{\theta^\star_{\tau_t}} \mathrm{E}_q\left[\tau\right]$$

$$\begin{aligned}\frac{\partial}{\partial \alpha_\tau} -\mathrm{E}_q\left[\log q(\tau_t)\right] &= \sum_t \psi\left(\alpha^\star_{\tau_t}\right) + \log \theta^\star_{\tau_t} - \mathrm{E}_q\left[\log \tau_t\right] - (\alpha^\star_{\tau_t} - 1) \frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\log \tau_t\right] + \frac{1}{\theta^\star_{\tau_t}} \frac{\partial}{\partial \alpha_\tau} \mathrm{E}_q\left[\tau\right] \\ &= \sum_t -(\alpha^\star_{\tau_t} - 1)\psi'\left(\alpha^\star_{\tau_t}\right) + 1\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \theta_\tau} -\mathrm{E}_q\left[\log q(\tau_t)\right] &= \sum_t \alpha^\star_{\tau_t} \cdot \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} - (\alpha^\star_{\tau_t} - 1) \cdot \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} - \frac{1}{\theta^2_\tau} \mathrm{E}_q\left[\tau\right] + \frac{1}{\theta^\star_{\tau_t}} \frac{\partial}{\partial \theta_\tau} \mathrm{E}_q\left[\tau\right] \\ &= \sum_t \alpha^\star_{\tau_t} \cdot \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} - (\alpha^\star_{\tau_t} - 1) \cdot \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} - \alpha^\star_{\tau_t} \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} + \alpha^\star_{\tau_t} \cdot \frac{\theta^\star_{\tau_t}}{\theta^2_\tau} \qquad = \frac{\theta^\star_{\tau_t}}{\theta^2_\tau}\end{aligned}$$

Table 30: Three combinations of hyperparameters obtaining minimal MSE on training and test data

	min MSE in sample			min MSE out of sample		
shape lambda α_λ	$\mathbf{10^{-10}}$	**1**	**10**	1	1	1
scale lambda θ_λ	10^{-10}	10^{-10}	10^{-10}	1	1	1
shape upsilon α_υ	10	10	10	10	10	10
scale upsilon θ_υ	1	1	1	1	1	1
shape gamma α_γ	10	10	10	$\mathbf{10^{-10}}$	**1**	**10**
scale gamma θ_γ	1	1	1	10^{-10}	10^{-10}	10^{-10}
shape omega α_ω	10	10	10	10	10	10
scale omega θ_ω	1	1	1	1	1	1
shape epsilon α_ϵ	10	10	10	10^{-10}	10^{-10}	10^{-10}
scale epsilon θ_ϵ	1	1	1	1	1	1
iteration	200	200	200	200	200	200
MSE train set	0.08035	0.08035	0.08035	0.8420	0.8420	0.8420
MSE test set	28.645	28.645	28.645	27.98	27.98	27.98
MSE train set rescaled	0.03	0.03	0.03	0.31	0.31	0.31
final ELBO	-30536	-11508	-10435	-4376	-3706	-3668

Table 31: Expected kernel weights and their precisions for models of Table 8

	default		ones		maxELBO		minELBO		random	
cell line i	$E_q[\lambda_{10,i}]$	$E_q[a_{10,i}]$	$E_q[\lambda_{10,i}]$	$E_q[a_{10,i}]$	$E_q[\lambda_{10,i}]$	$E_q[a_{10,i}]$	$E_q[\lambda_{10,i}]$	$E_q[a_{10,i}]$	$E_q[\lambda_{10,i}]$	$E_q[a_{10,i}]$
UACC812	2.2E-11	9.1E-04	1.15	0.450	10.026	5.1E-04	5.0E-11	-0.170	2.5E-11	117135.6
MCF10F	2.5E-11	8.2E-04	1.29	-0.147	10.034	-1.2E-03	5.0E-11	-0.171	4.1E-11	-5237.4
M.134VI	2.3E-11	1.0E-03	1.21	0.290	10.038	-1.1E-04	5.0E-11	-0.175	2.5E-11	114722.3
HCC1419	2.3E-11	9.3E-04	0.87	1.002	10.033	9.6E-04	5.0E-11	0.285	8.0E-12	314795.6
HCC1143	2.3E-11	8.5E-04	1.25	-0.180	10.025	-1.0E-03	5.0E-11	-0.397	3.8E-11	-30703.5

Table 32: Expected intermediate outputs of drug ten for five *test* cell lines - default model

cell line i	$\mathrm{E}_q\left[g_{10,1,i}\right]$	$\mathrm{E}_q\left[g_{10,2,i}\right]$	$\mathrm{E}_q\left[g_{10,3,i}\right]$	$\mathrm{E}_q\left[g_{10,4,i}\right]$	$\mathrm{E}_q\left[g_{10,5,i}\right]$	$\mathrm{E}_q\left[g_{10,6,i}\right]$
UACC812	0.012	0.011	0.015	0.013	0.016	0.012
MCF10F	0.011	0.014	0.019	0.014	0.018	0.012
MDAMB134VI	0.011	0.017	0.013	0.019	0.017	0.018
HCC1419	0.010	0.012	0.019	0.013	0.020	0.011
HCC1143	0.010	0.010	0.011	0.013	0.018	0.011
$\mathrm{E}_q\left[\upsilon_{10}\right]$	1.3885E-12					

Table 33: Expected intermediate outputs of drug ten for five *test* cell lines - ones model

	$\mathrm{E}_q\left[g_{10,1}\right]$	$\mathrm{E}_q\left[g_{10,2}\right]$	$\mathrm{E}_q\left[g_{10,3}\right]$	$\mathrm{E}_q\left[g_{10,4}\right]$	$\mathrm{E}_q\left[g_{10,5}\right]$	$\mathrm{E}_q\left[g_{10,6}\right]$
UACC812	0.113	0.461	0.010	0.552	0.337	0.689
MCF10F	-0.920	-0.398	-0.550	-0.301	-0.043	-0.775
MDAMB134VI	0.216	-0.068	0.179	0.203	0.273	-0.091
HCC1419	1.165	0.921	-0.140	1.096	0.267	1.587
HCC1143	-1.049	-0.543	-0.496	-0.369	-0.147	-0.956
$\mathrm{E}_q\left[\upsilon_{10}\right]$	0.1131					

Table 34: Expected intermediate outputs of drug ten for five *test* cell lines of drug ten - random model

	$\mathrm{E}_q\left[g_{10,1}\right]$	$\mathrm{E}_q\left[g_{10,2}\right]$	$\mathrm{E}_q\left[g_{10,3}\right]$	$\mathrm{E}_q\left[g_{10,4}\right]$	$\mathrm{E}_q\left[g_{10,5}\right]$	$\mathrm{E}_q\left[g_{10,6}\right]$
UACC812	4351.1	150101.3	32207.3	202331.2	193799.8	167337.6
MCF10F	-149456.0	50801.7	-21708.0	67697.2	160204.3	-84209.3
MDAMB134VI	48541.9	75570.8	81768.4	175899.1	195934.6	40766.7
HCC1419	246109.6	263778.1	-12388.4	325972.8	178186.9	374869.1
HCC1143	-173831.6	-19885.5	-36486.4	49418.0	136131.8	-127793.7
$\mathrm{E}_q\left[\upsilon_{10}\right]$	1.2548E-11					

Table 35: Summary statistics expected intermediate outputs of drug ten - default model

stat	$E_q[g_{10,1}]$	$E_q[g_{10,2}]$	$E_q[g_{10,3}]$	$E_q[g_{10,4}]$	$E_q[g_{10,5}]$	$E_q[g_{10,6}]$
min	0.0058	0.0078	0.0070	0.0063	0.0114	0.0087
max	0.0169	0.0170	0.0187	0.0188	0.0205	0.0177
range	0.0111	0.0093	0.0116	0.0125	0.0090	0.0090
median	0.0112	0.0111	0.0143	0.0137	0.0166	0.0119
mean	0.0115	0.0116	0.0137	0.0141	0.0164	0.0125
std.dev	0.0023	0.0024	0.0037	0.0025	0.0022	0.0028
$E_q[e_k]$	8.79E-17	2.49E-16	2.01E-16	-7.10E-18	1.47E-16	1.20E-16
$E_q[\omega_k]$	5.00E-11	5.00E-11	5.00E-11	5.00E-11	5.00E-11	5.00E-11

Table 36: Summary statistics expected intermediate outputs of drug ten - ones model

stat	$E_q[g_{10,1}]$	$E_q[g_{10,2}]$	$E_q[g_{10,3}]$	$E_q[g_{10,4}]$	$E_q[g_{10,5}]$	$E_q[g_{10,6}]$
min	-2.050	-1.300	-1.393	-1.071	-0.853	-1.731
max	1.165	0.972	0.635	1.096	0.859	1.587
range	3.215	2.272	2.028	2.168	1.711	3.318
median	-0.740	-0.331	-0.443	-0.182	-0.073	-0.315
mean	-0.578	-0.251	-0.389	-0.072	-0.094	-0.197
std.dev	0.907	0.679	0.539	0.593	0.476	0.977
$E_q[e_k]$	0.0431	0.0399	0.0346	0.0331	0.0272	0.0348
$E_q[\omega_k]$	1.4986	1.4988	1.4991	1.4991	1.4994	1.4991

Table 37: Summary statistics expected intermediate outputs of drug ten - random model

stat	$E_q[g_{10,1}]$	$E_q[g_{10,2}]$	$E_q[g_{10,3}]$	$E_q[g_{10,4}]$	$E_q[g_{10,5}]$	$E_q[g_{10,6}]$
min	-348980.4	-152453.7	-197341.4	-76542.3	-33072.3	-273436.3
max	246109.6	263778.1	144942.5	325972.8	293777.2	374869.1
range	595090.1	416231.8	342283.9	402515.1	326849.5	648305.4
median	-119546.6	30570.7	-28911.5	106772.8	145856.2	11074.9
mean	-88018.2	38507.2	-9243.6	110914.2	125176.5	32374.5
std.dev	155701.4	116118.5	83821.8	110296.8	90271.4	181941.3
$E_q[e_k]$	2.94E-07	2.82E-07	2.42E-07	2.39E-07	1.99E-07	2.25E-07
$E_q[\omega_k]$	1.50	1.50	1.50	1.50	1.50	1.50

Table 38: Expected biases and their precision for fitted values or predicitions for drug one to ten for five models of grid search trained on standardized drug responses

	default	ones	minELBO	maxELBO	random
$\mathrm{E}_q\left[b_1\right]$	-5.449E-16	0.066	3.755E-06	5.063E-05	0.011
$\mathrm{E}_q\left[b_2\right]$	-5.137E-16	0.096	-1.919E-06	5.909E-05	0.011
$\mathrm{E}_q\left[b_3\right]$	2.133E-16	0.217	2.967E-06	4.432E-05	0.105
$\mathrm{E}_q\left[b_4\right]$	-1.012E-15	-0.068	-6.246E-06	-7.320E-06	-0.022
$\mathrm{E}_q\left[b_5\right]$	-3.140E-14	-0.117	-5.679E-06	1.039E-05	-0.017
$\mathrm{E}_q\left[b_6\right]$	-1.447E-15	0.202	8.433E-06	3.027E-05	0.024
$\mathrm{E}_q\left[b_7\right]$	1.540E-16	-0.071	1.624E-06	2.955E-05	-0.057
$\mathrm{E}_q\left[b_8\right]$	-5.812E-16	-0.030	1.392E-06	7.388E-05	0.008
$\mathrm{E}_q\left[b_9\right]$	9.876E-16	0.090	-2.462E-06	-1.567E-05	0.083
$\mathrm{E}_q\left[b_{10}\right]$	1.599E-16	0.302	-2.286E-06	6.166E-05	0.082
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathrm{E}_q\left[\gamma_1\right]$	4.994E-11	1.490	4.994E-11	1.037E+01	10.472
$\mathrm{E}_q\left[\gamma_2\right]$	4.994E-11	1.486	4.994E-11	1.037E+01	10.472
$\mathrm{E}_q\left[\gamma_3\right]$	4.995E-11	1.460	4.995E-11	1.038E+01	10.420
$\mathrm{E}_q\left[\gamma_4\right]$	4.987E-11	1.479	4.987E-11	1.033E+01	10.445
$\mathrm{E}_q\left[\gamma_5\right]$	4.995E-11	1.487	4.995E-11	1.038E+01	10.479
$\mathrm{E}_q\left[\gamma_6\right]$	4.979E-11	1.454	4.979E-11	1.030E+01	10.431
$\mathrm{E}_q\left[\gamma_7\right]$	4.995E-11	1.488	4.995E-11	1.038E+01	10.451
$\mathrm{E}_q\left[\gamma_8\right]$	4.994E-11	1.494	4.994E-11	1.037E+01	10.473
$\mathrm{E}_q\left[\gamma_9\right]$	4.994E-11	1.485	4.995E-11	1.037E+01	10.433
$\mathrm{E}_q\left[\gamma_{10}\right]$	4.993E-11	1.428	4.993E-11	1.036E+01	10.434
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 39: Drug number in data and its name. Drugs number 12, 26 and 27 have been excluded in Costello et al. (2014)

No.		No.	
1	4-HC (DNA alkylator)	17	B581 (FTPase)
2	4-HC+Dox (Combination)	18	Methylglyoxol (Pyruvate)
3	Baicalein (CYP2C9)	19	MG-132 (Proteasome)
4	Bromopyruvate (Glycolysis)	20	Nelfinavir (Protease)
5	Cetuximab (EGFR)	21	Nilontinib (BCR-ABL)
6	Chloroquine (Autophagy)	22	Olomoucine II (CDK1)
7	Disulfiram (ALDH2)	23	PD184352 (MEK)
8	Doxorubicin (TOP2A)	24	PS-1145 (IKK)
9	FR180304 (ERK)	25	QNZ (NFkB)
10	Everolimus (mTOR)	26	NA
11	Mebendazole (Tubulin)	27	NA
12	NA	28	TCS PIM-11 (PIM1)
13	GW5074 (RAF1)	29	Valproate (HDAC)
14	Trastuzumab (ERBB2)	30	MG-132b (Proteasome)
15	IKK 16 (IKK2)	31	MG-115 (Proteasome)
16	Imatinib (BCR-ABL)		

Table 40: Summary statistics of $-\log_{10} GI_{50}$ drug responses for each drug for all observations (training and testing data). Costello et al. (2014) excluded drug 12, 26 and 27.

drug no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
no. obs	42	42	49	25	47	20	46	42	46	43	47	26	24	30	43	50
no. NA	11	11	4	28	6	33	7	11	7	10	6	27	29	23	10	3
min	3.78	3.78	3.78	6.42	6.16	4.86	4.18	6.26	4.00	4.27	4.18	4.48	4.48	5.14	5.09	3.82
max	6.43	6.18	4.94	7.88	7.61	7.92	7.30	9.83	5.23	8.40	5.00	5.29	6.14	8.87	6.05	5.81
range	2.66	2.40	1.16	1.46	1.45	3.05	3.13	3.57	1.23	4.13	0.83	0.81	1.67	3.73	0.96	1.99
median	4.91	4.96	4.30	7.11	6.16	5.06	4.99	6.88	4.48	6.08	4.48	4.48	4.48	5.14	5.48	4.70
mean	4.90	4.94	4.28	7.12	6.26	5.25	5.54	6.98	4.43	6.23	4.40	4.51	4.67	5.45	5.51	4.71
var	0.20	0.18	0.10	0.17	0.09	0.47	1.37	0.35	0.10	1.49	0.04	0.03	0.21	0.59	0.07	0.08
std.dev	0.45	0.43	0.32	0.41	0.30	0.68	1.17	0.59	0.32	1.22	0.20	0.16	0.46	0.77	0.26	0.29

drug no.	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
no. obs	49	39	48	50	48	48	47	50	26	49	25	49	49	41	42
no. NA	4	14	5	3	5	5	6	3	27	4	28	4	4	12	11
min	4.18	2.52	6.26	4.48	3.70	4.95	4.18	2.72	4.48	5.48	4.48	3.88	2.16	5.27	5.52
max	6.96	5.37	7.79	5.74	4.22	5.98	6.33	19.50	8.87	5.48	5.71	5.61	3.45	7.32	7.43
range	2.78	2.86	1.53	1.26	0.52	1.03	2.16	16.78	4.40	0.00	1.24	1.73	1.29	2.05	1.90
median	6.30	3.22	6.78	5.01	3.70	5.23	4.48	2.72	4.94	5.48	4.48	3.88	2.78	6.60	6.28
mean	5.98	3.32	6.82	5.00	3.74	5.30	4.61	3.07	5.85	5.48	4.56	4.17	2.79	6.62	6.29
var	0.61	0.40	0.11	0.08	0.02	0.04	0.24	5.63	2.20	0.00	0.08	0.16	0.06	0.19	0.15
std.dev	0.78	0.64	0.33	0.29	0.12	0.20	0.49	2.37	1.48	0.00	0.29	0.40	0.25	0.43	0.39

Table 41: Summary statistics of $-\log_{10} GI_{50}$ drug responses for each drug for training data. Drugs 5, 24 and 26 have nearly no variation

drug no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
no. obs.	30	30	34	20	33	16	33	30	31	28	33	17	15	22	30	34
no. NA	5	5	1	15	2	19	2	5	4	7	2	18	20	13	5	1
min	3.78	3.78	3.78	6.42	6.16	4.86	4.18	6.26	4.00	4.27	4.18	4.48	4.48	5.14	5.09	3.82
max	6.43	6.18	4.94	7.88	6.22	5.78	7.30	9.83	5.23	8.40	4.48	5.29	6.14	8.87	6.00	5.81
range	2.66	2.40	1.16	1.46	0.06	0.91	3.12	3.57	1.23	4.13	0.30	0.81	1.67	3.73	0.91	1.99
median	4.90	4.96	4.31	7.08	6.16	5.01	5.85	6.93	4.48	6.05	4.48	4.48	4.48	5.14	5.43	4.73
mean	4.89	4.94	4.29	7.05	6.17	5.12	5.67	7.02	4.40	6.24	4.35	4.53	4.59	5.45	5.47	4.73
var	0.23	0.20	0.10	0.15	0.00	0.09	1.46	0.45	0.12	1.71	0.02	0.04	0.19	0.75	0.05	0.11
std.dev	0.48	0.45	0.32	0.39	0.01	0.30	1.21	0.67	0.34	1.31	0.15	0.20	0.43	0.86	0.23	0.34

drug no.	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
no. obs	34	27	32	35	33	32	32	34	17	34	17	33	33	28	29
no. NA	1	8	3	0	2	3	3	1	18	1	18	2	2	7	6
min	4.18	2.52	6.26	4.48	3.70	4.95	4.18	2.72	4.48	5.48	4.48	3.88	2.16	5.78	5.62
max	6.96	5.37	7.32	5.74	4.21	5.72	6.25	2.76	8.23	5.48	5.71	4.57	3.15	7.32	7.43
range	2.78	2.86	1.06	1.26	0.51	0.77	2.07	0.04	3.75	0.00	1.24	0.70	0.99	1.54	1.81
median	6.29	3.00	6.73	4.99	3.70	5.23	4.48	2.72	4.93	5.48	4.48	3.88	2.76	6.58	6.26
mean	6.01	3.25	6.76	4.98	3.74	5.27	4.49	2.73	5.87	5.48	4.60	4.03	2.74	6.64	6.30
var	0.51	0.39	0.07	0.09	0.01	0.03	0.14	0.00	2.13	0.00	0.12	0.05	0.06	0.11	0.13
std.dev	0.72	0.62	0.26	0.31	0.12	0.17	0.38	0.01	1.46	0.00	0.34	0.23	0.25	0.33	0.37

D Declaration of academic honesty

I, Henry Webel, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables and images), have been acknowledged by me as such. I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

Berlin, 30th January 2018

H. Webel