Variational inference - theory and in Bayesian Multiple Task Multiple Kernel Learning (BMTMKL)

Henry Webel

Msc Statistics of joint program of HU Berlin, FU Berlin and TU Berlin

E-Mail: webel@posteo.eu

Outline

Focus of today

1. Mechanic understanding of *Variational Inference*
   - ▷ Directed Acyclic Graphs - Model Representation
   - ▷ Idea of Variational Inference
   - ▷ Difficulties
2. How to get to a model? (it is not LDA)
3. What does BMTMKL do?

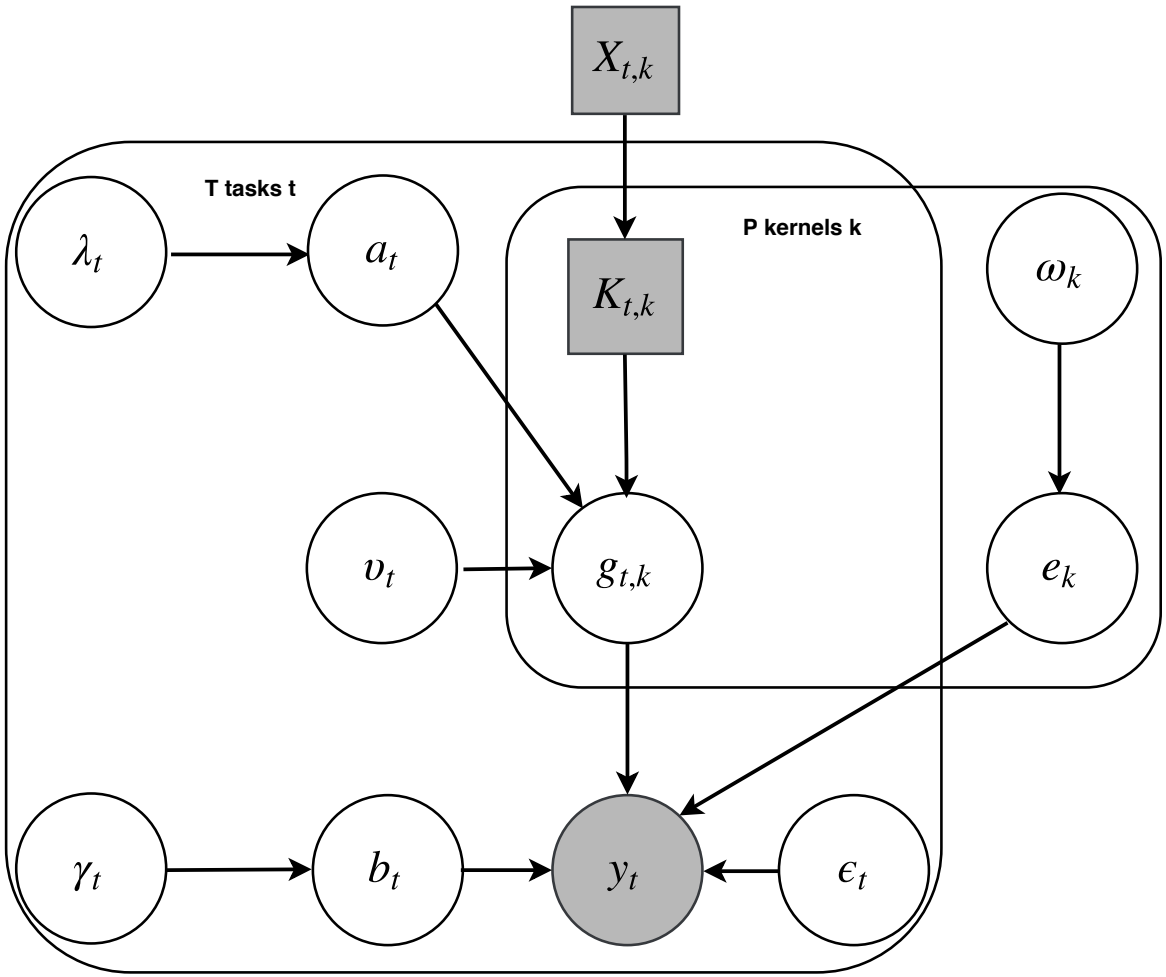
What is Variational Inference?

- A procedure to find parameters in difficult joint distributions of random variables $p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})$
- Instead of finding the posterior $p(\mathcal{Z}|\mathcal{Y}, \mathcal{X})$, one is using an easier approximation of the posterior $q(\mathcal{Z})$

Notation:

- Latent *random* variables: $\mathcal{Z}$
- measured/ observed, *non-random* explanatory data: $\mathcal{X}$
- measured/ observed, *random* outcome (one dimensional): $\mathcal{Y}$

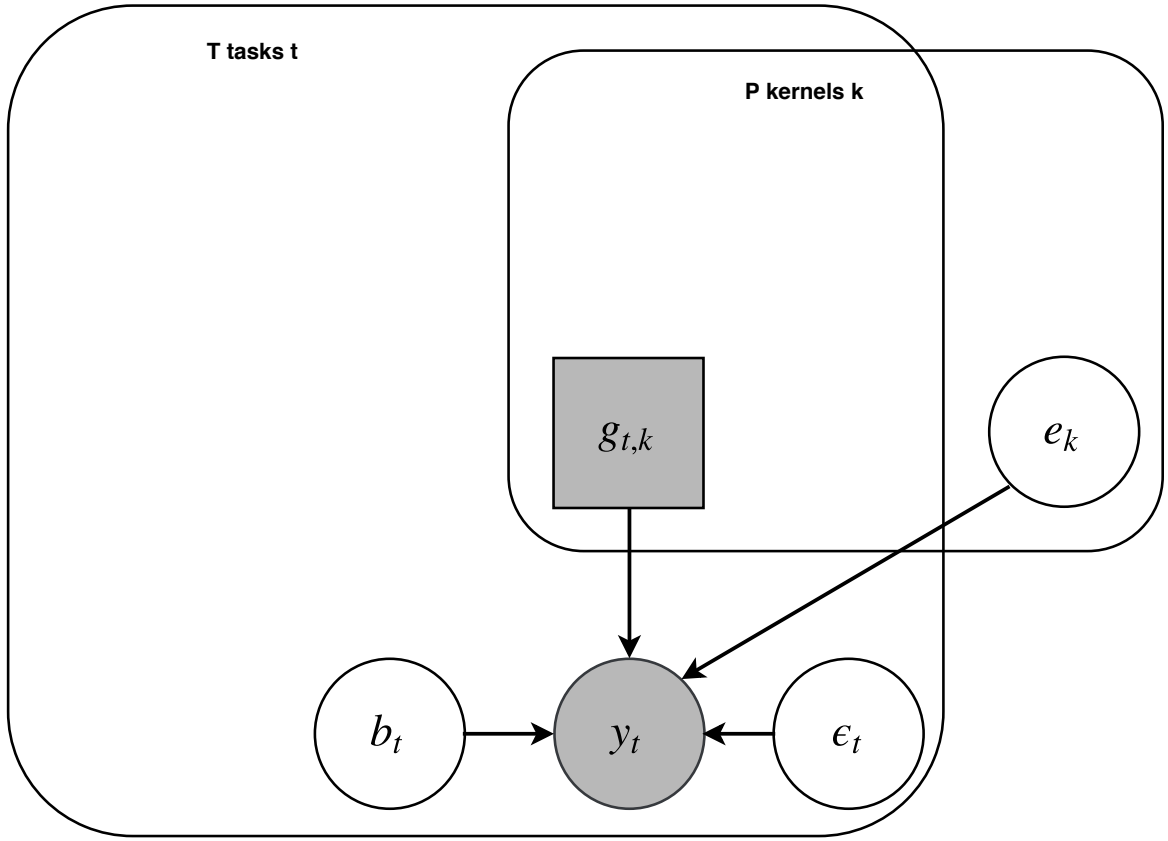
Directed Acyclic Graph (DAG) - BMTMKL



References: [Costello et. al. (2014), Gönen, 2012a]

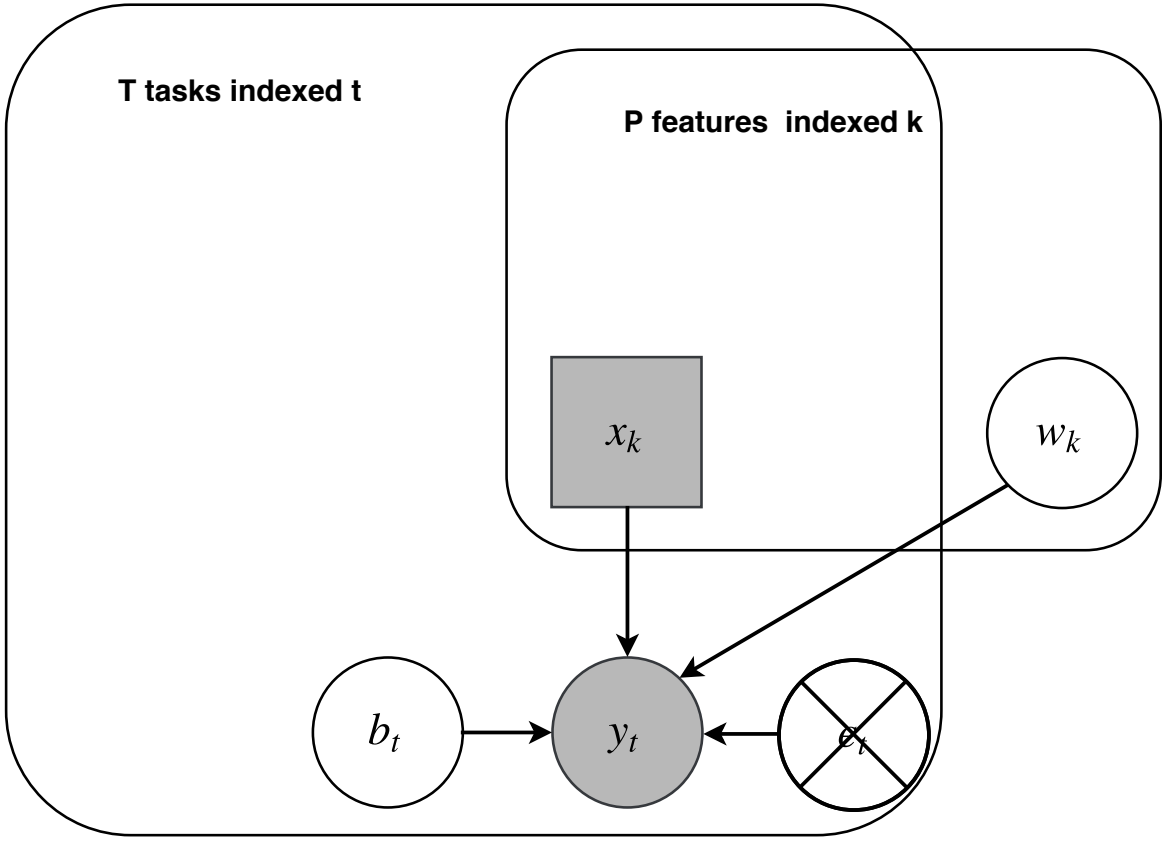
DAG - Bayesian Reg.

(In Notation of BMTMKL, see also sl. 49 in appendix)

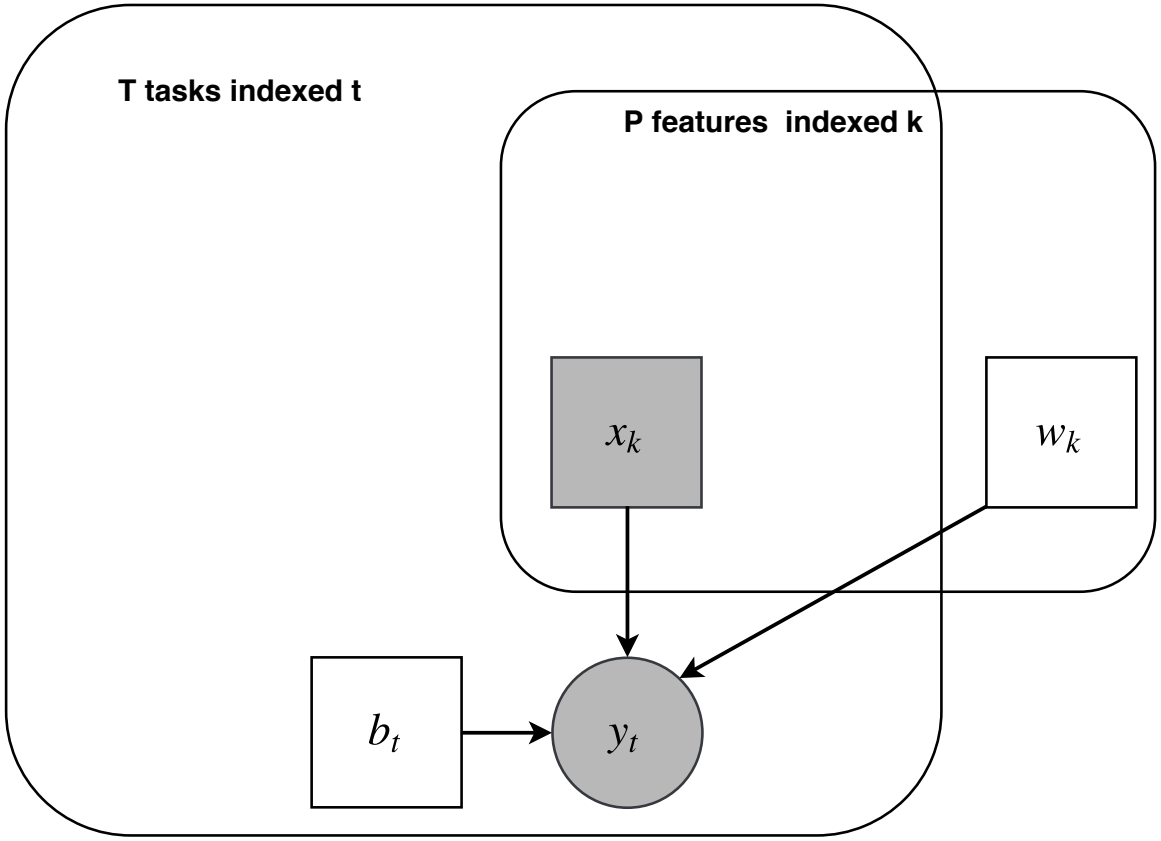


DAG - Bayesian Reg.

In notation of edwardlib.org/tutorials/supervised-regression



DAG - Maximum likelihood Reg.



Distributional assumptions: Bayesian Regression for y_t

$$w_k \sim \mathcal{N}(0, \sigma_w^2) \qquad b_t \sim \mathcal{N}(0, \sigma_b^2)$$

$$\cancel{\epsilon_t \sim \mathcal{G}am(\alpha_\epsilon, \beta_\epsilon)}$$

$$y_t | w, \mathcal{X}, b_t, \cancel{\epsilon_t} \sim \mathcal{N}\left(\underbrace{\sum_{k=1}^{P} x_k w_k + b_t \cdot \mathbb{1}_{N_t}}_{\text{mean}}, \underbrace{\cancel{\epsilon_t^{-1}} I_{N_t} \sigma_y^2}_{\text{Variance}} \right)$$

joint: $$p(\mathcal{Z}, \mathcal{Y} | \mathcal{X}) = p(w) \cdot p(b_t) \cdot \cancel{p(\epsilon_t)} \cdot p(y_t | w, \mathcal{X}, b_t, \cancel{\epsilon_t})$$

(DAGs: Conditional independence assumption to form joint.)

Variational Inference: joint, posterior and approx. posterior density in general notation

$$\text{joint} \quad p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) = p(b, e, \epsilon, \mathcal{Y}|\mathcal{X}) \quad \text{(with precision } \epsilon \text{ on } y_t\text{)}$$

$$\text{posterior} \quad p(\mathcal{Z}|\mathcal{Y}, \mathcal{X}) = \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{p(\mathcal{Y}|\mathcal{X})}$$

$$\text{approx. post.} \quad q(\mathcal{Z}) = q(b, w)q(\epsilon) \quad \text{(diff. to [Tutorial] for } b, w\text{)}$$

- Set of latent random variables $\mathcal{Z} = \{b, w, \epsilon\}$
- observed explantory data for task t: $\mathcal{X} = \{x_1, \ldots, x_P\}$
- observed outcome vector $\mathcal{Y} = y_t$

Observed data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

Objective: Kullback Leibler (KL) Divergence

Minimization of the difference between true posterior $p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and approximated posterior $q(\mathcal{Z})$ is the objective. Non-symmetric KL-divergence used

$$\mathrm{KL}\left[q(\mathcal{Z})||p(\mathcal{Z}|\mathcal{X},\mathcal{Y})\right] = \mathrm{E}_{q(\mathcal{Z})}\left[\log\frac{q(\mathcal{Z})}{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\right] \geq 0$$

- Instead Min. of KL divergence, the so called evidence lower bound (**ELBO**, $\mathcal{L}$) is **maximized**, see sl. 43
- Maximization of ELBO: (Stochastic) gradient descent [Hoffman et. al., 2013] or alternative proof [Blei et. al., 2017] (see sl. 45) can be found in more detail in my *master thesis* (Sec. 2.3, 2.5). Iterative procedure.
  $\Rightarrow$ **Full conditionals** needed

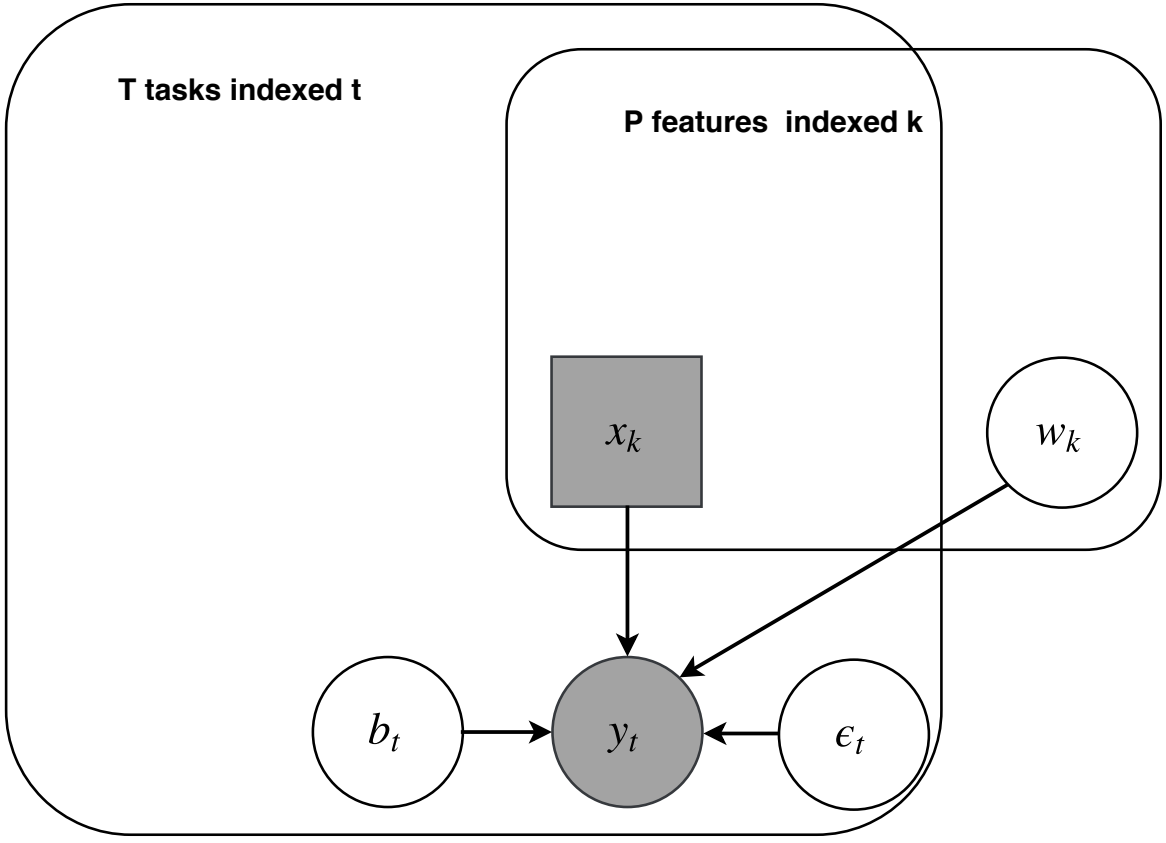
Full conditionals: $p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j})$

Full conditional of z_j: Conditional density of the set of random variables z_j given all other latent variables $\mathcal{Z}_{-z_j} = \mathcal{Z} \setminus z_j$ and data

- two constellations in presented setup of Gamma and Normal priors
  - Gamma-distributed full conditional Random Var. (RV)
    - Gamma prior and its associated normal RV
  - Normal-distributed full conditional RV
    - Normally distributed RV and its associated other RV

Bayesian Reg. incl. precison prior of y_t

In notation of edwardlib.org/tutorials/supervised-regression incl. ϵ_t



full conditional: Gamma prior ϵ_t

Example: Precision ϵ_t on outcome (drug response) y_t

$$
\begin{aligned}
p(\epsilon_t|\mathcal{D}, \mathcal{Z}_{-\epsilon_t}) &= \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{\int_{\epsilon_t} p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) d\epsilon_t} \\
&= c_1 \cdot p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) && \propto p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) \\
&= c \cdot p(\epsilon_t) p(y_t|w, \mathcal{X}, b_t, \epsilon_t) && \propto p(\epsilon_t) p(y_t|w, \mathcal{X}, b_t, \epsilon_t)
\end{aligned}
$$

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

Full conditional of ϵ_t in Bay. Regression

$$p(\epsilon_t|\mathcal{D}, \mathcal{Z}_{-\epsilon_t}) \propto \frac{1}{\Gamma(\alpha_\epsilon)} \beta_\epsilon^{\alpha_\epsilon} \epsilon_t^{\alpha_\epsilon - 1} \exp(-\beta_\epsilon \epsilon_t) \quad \text{(rate notation)}$$

$$\cdot \left(\frac{\epsilon_t}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\epsilon_t}{2} \underbrace{\left\| y_t - \sum_{k=1}^{P} w_k x_k \right\|^2}_{c_{y_t}}\right\}$$

$$\propto \epsilon_t^{\alpha_\lambda - 1} \exp(-\beta_\lambda \epsilon_t) \cdot (\epsilon_t)^{\frac{1}{2}} \exp\left\{-\frac{\epsilon_t}{2} c_{y_t}\right\}$$

$$= \epsilon_t^{\alpha_\lambda + \frac{1}{2} - 1} \cdot \exp\left\{-\left(\beta_\lambda + \frac{1}{2} c_{y_t}\right) \epsilon_t\right\}$$

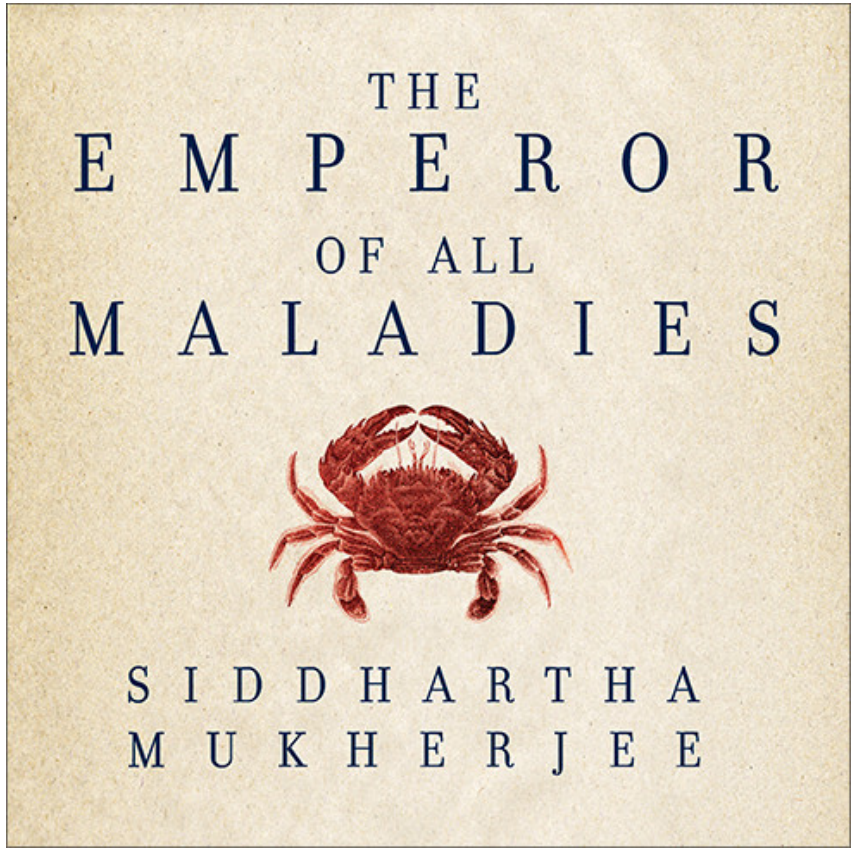
$$\propto \mathcal{G}am\left(\epsilon_t \middle| \alpha_\lambda + \frac{1}{2}, \beta_\lambda + \frac{1}{2} c_{y_t}\right)$$

Full conditionals: difficulties

- rewrite prior densities into correct (!) multidimensional form, e.g. all drug reponses in BMTMKL (c.p. with sl. 28)

$$
\begin{aligned}
p(\mathcal{Y}|e, G, b, \epsilon) &= \prod_{t=1}^{T} \mathcal{N}(y_t|G_t \cdot e + b_t \mathbb{1}_{N_t}, \epsilon_t I_{N_t}) \\
&= \ldots \quad \text{(see MA thesis, ch. 2.2.7.3)} \\
&= \frac{(\prod_{t=1}^{T} \epsilon_t N_t)^{1/2}}{(2\pi)^{N/2}} \cdot \exp\Big\{ -\tfrac{1}{2}\Big(\big[y' - (b', e')B'\big] \\
&\quad \cdot \operatorname{diag}(\{I_{N_t}\epsilon_t\}_{t=1}^{T}) \cdot \big[y - B(b', e')'\big] \Big)\Big\} \\
&\quad \text{, where: } B = (\operatorname{diag}(\{\mathbb{1}_{N_t}\}_{t=1}^{T}) \vdots G) \in \mathbb{R}^{N \times (T+P)}
\end{aligned}
$$

The Emperor of All Maladies



Topic

The research objective is

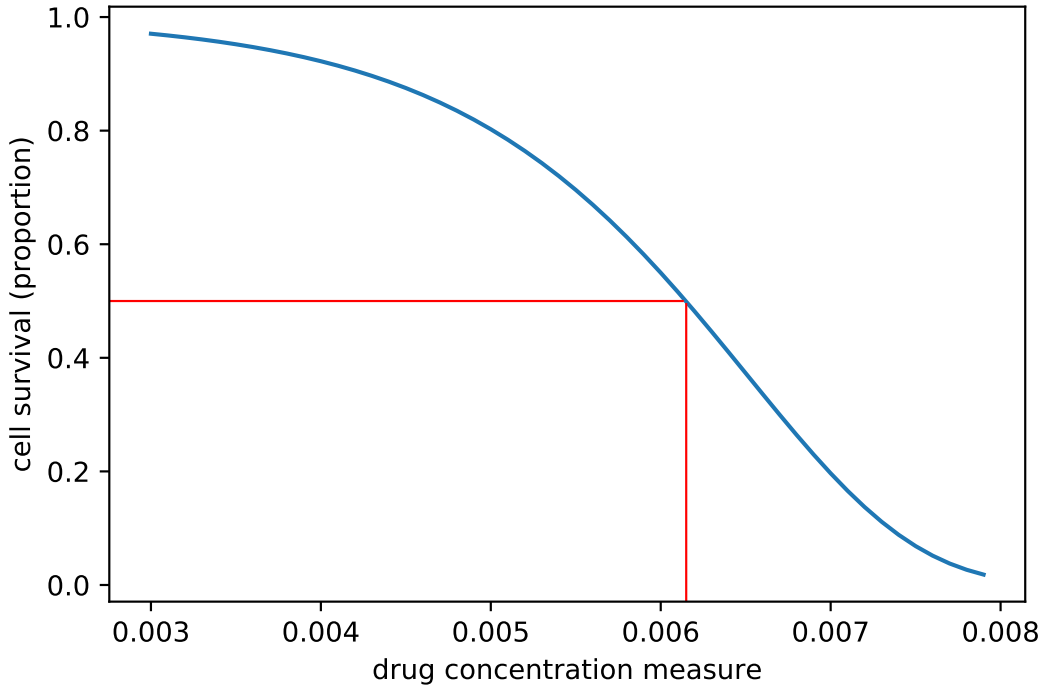
- to predict effectiveness of T different drugs on growth inhibition of cancer cells
- on the basis of P high-dimensional inputs, so called omics,
- in order to identify drugs for clinical research.

Master thesis in scope of WP 5, *Model-Based Prediction of Drug Responsiveness*, of predict-project (HU/Charité).
$\Rightarrow$ First step: Replicate and Explain in Detail winning model of *DREAM7* challenge on drug effectiveness.

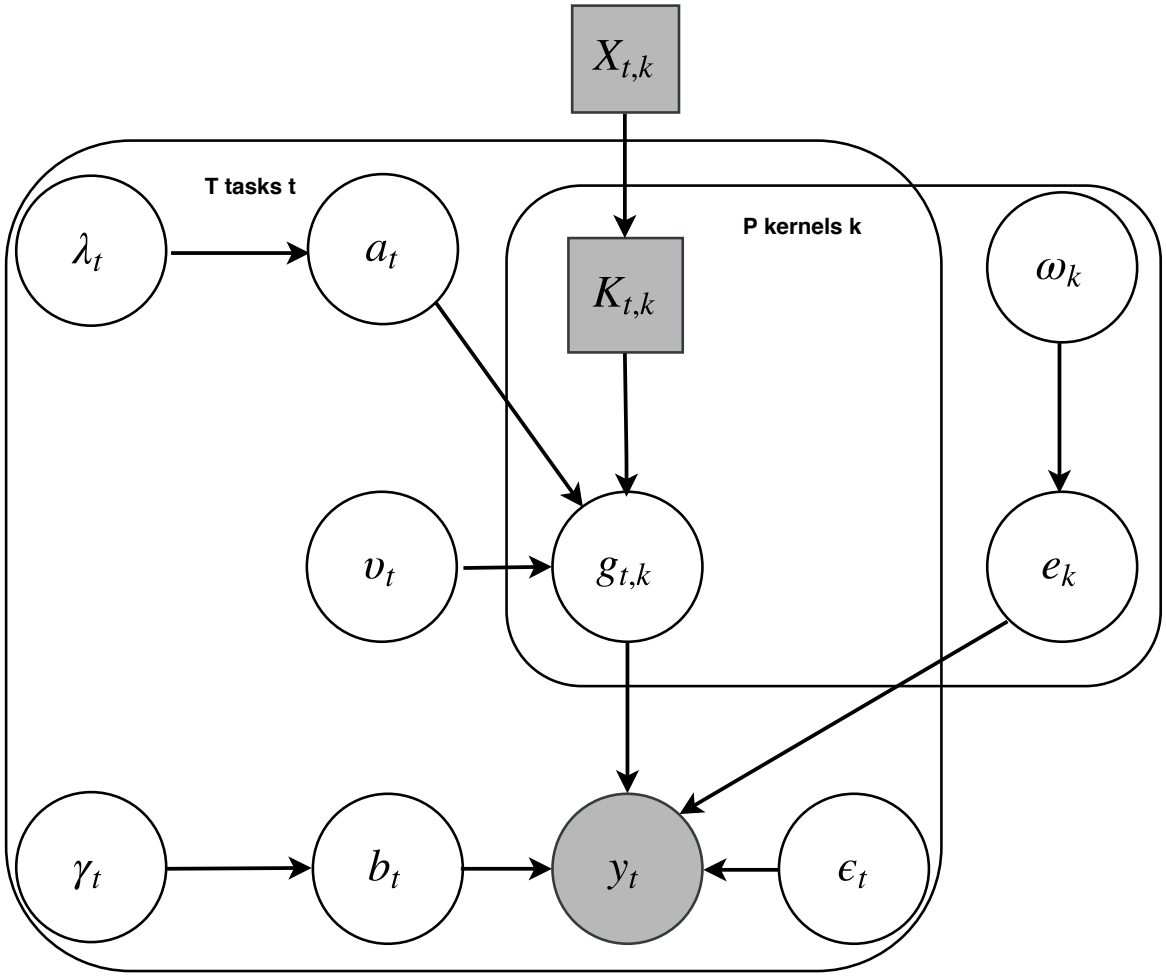
Good Reference for start: [Nic et. al. (2016)]

drug response (susceptibility or sensitivity)

y_t: Higher numbers indicate higher drug effectiveness, because it is constructed as a measure of drug concentrations transformed by $-\log_{10}$ - which corresponds to a reversion of proportion.

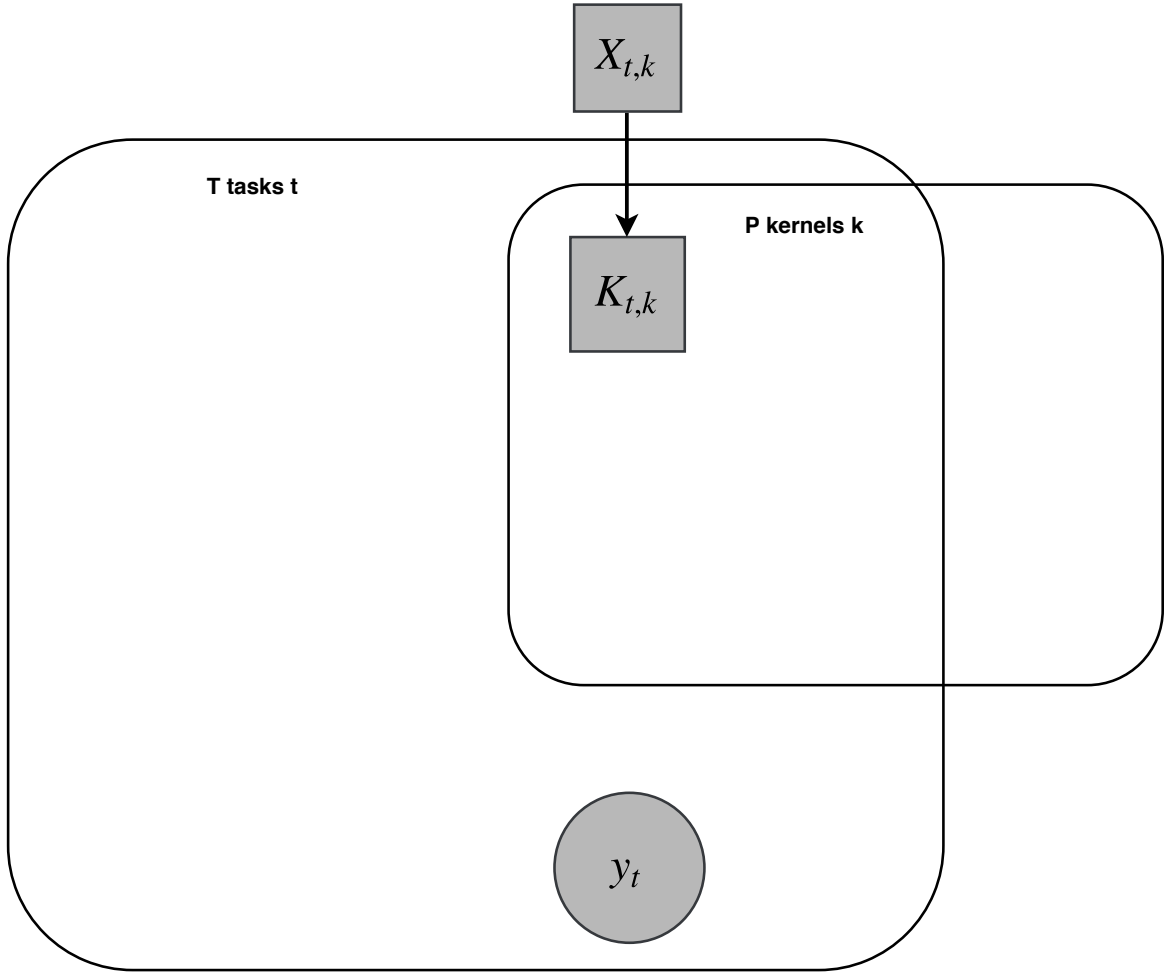


Model as Directed Acyclic Graph



References: [Costello et. al. (2014), Gönen, 2012a]

Observed deterministic and random variables



After all a linear regression

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

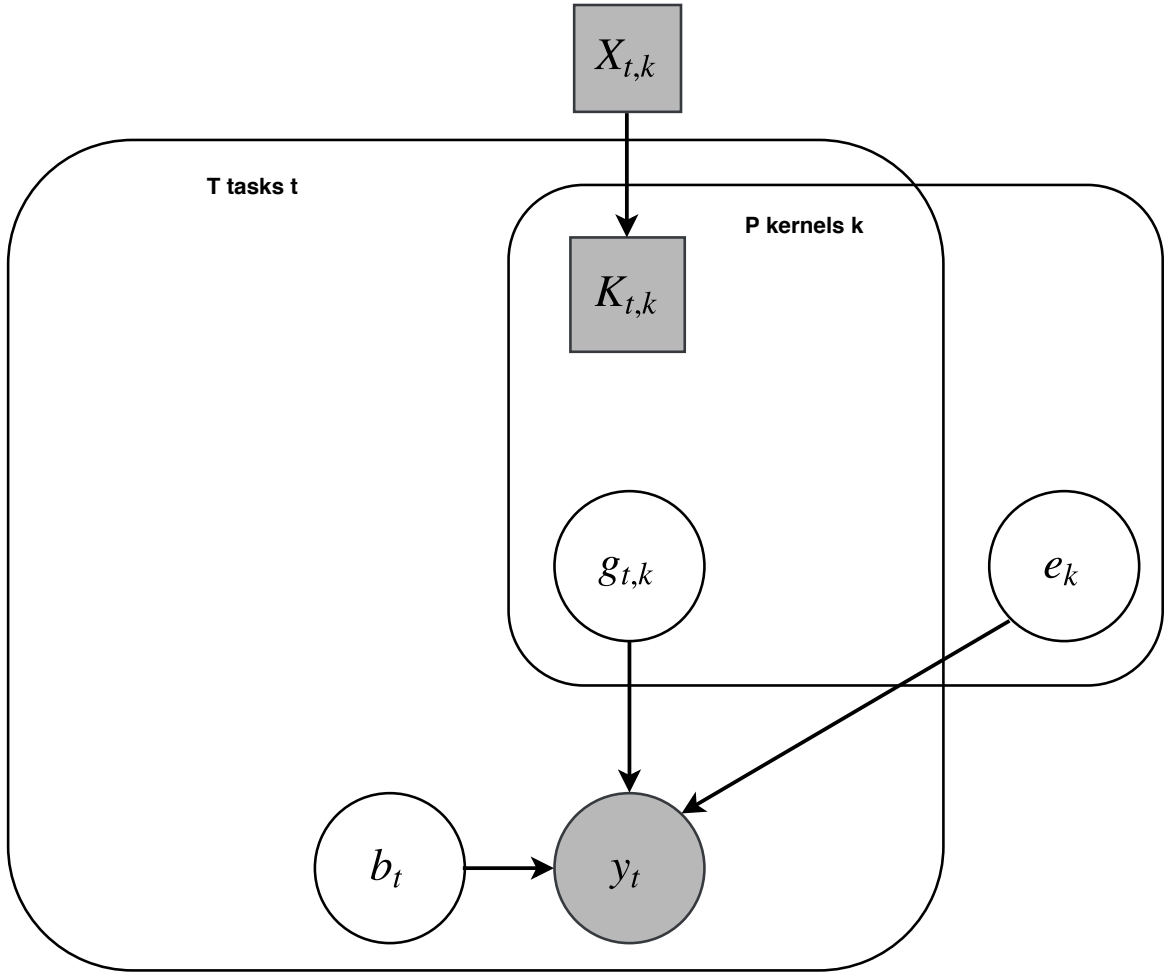
probability density function for output y_t depends on several hidden variables. In matrix notation replacing $\sum_{k=1}^{P} g_{t,k} e_k = G_t \cdot e$:

$$y_t|e, G_t, b_t, \epsilon_t \sim \mathcal{N}\left(\, G_t \cdot e + b_t \cdot \mathbb{1}_{N_t},\, \epsilon_t^{-1} I_{N_t}\right) \quad , e = (e_1, \ldots, e_P)'$$

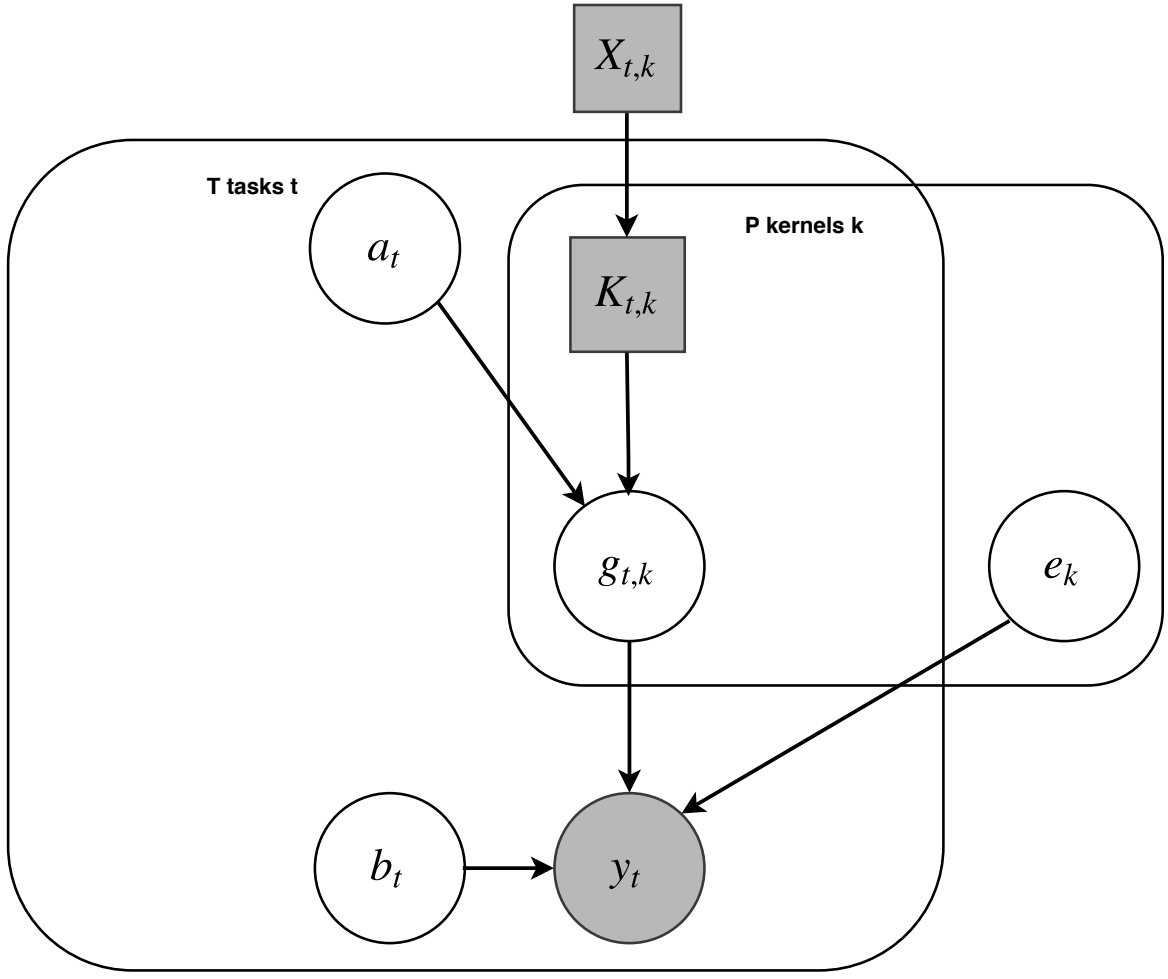
$$\sim \mathcal{N}\left(\, \begin{bmatrix}\mathbb{1}_{N_t} & G_t\end{bmatrix} \cdot \begin{pmatrix} b_t \\ e \end{pmatrix},\, \epsilon_t^{-1} I_{N_t}\right)$$

where: $G_t = (g_{t,1}, \ldots, g_{t,P})$

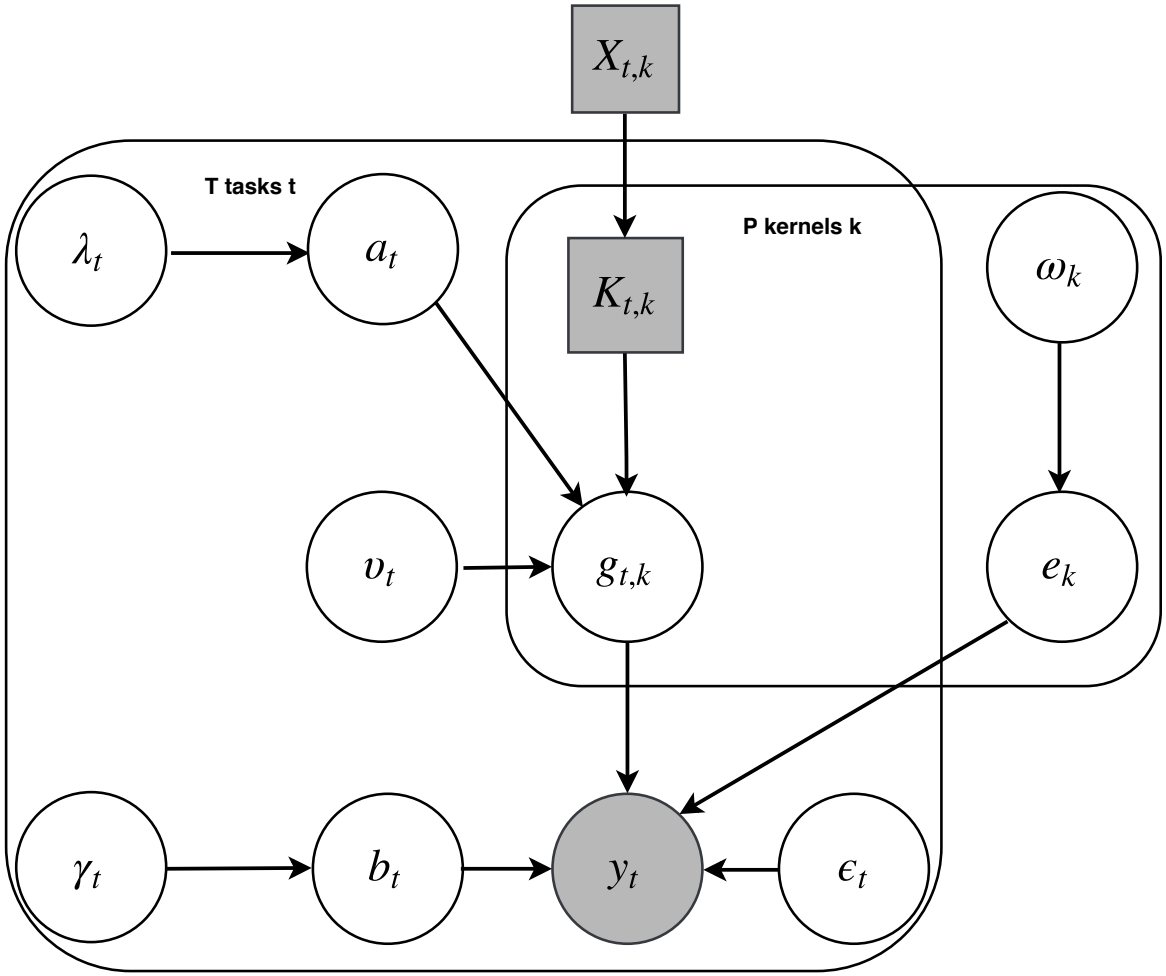
Bayesian regression



Bayesian reg. with intermediate outputs



... and adding prior on precision of normals



References: [Costello et. al. (2014), Gönen, 2012a]

$K_{t,k}$: **Input kernels matrices** $k = 1, \ldots, P$ for each drug t of dimension $N_t \times N_t$ containing some kind of similarity measure between cell lines $i = 1, \ldots, N_t$

a_t: **input kernel weights** (for all kernels $k = 1, \ldots, P$ specific for each drug t) of dimension $N_t \times 1$

$g_{t,k}$: **intermediate output** $g_{t,k} = K_{t,k} \cdot a_t$ of dimension $N_t \times 1$ for drug t and kernel k

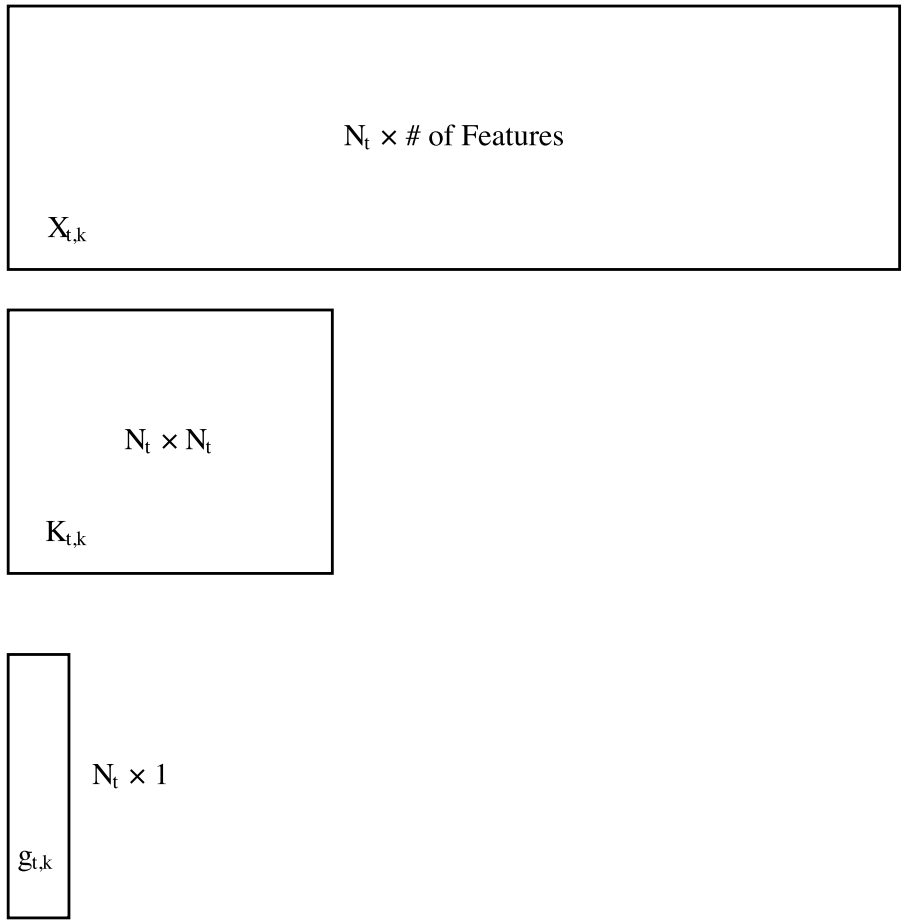
b_t: bias or **intercept** for drug t capturing the drug specific level of sensitivity

e_k: kernel **coefficients** for each input kernel $k = 1, \ldots, P$

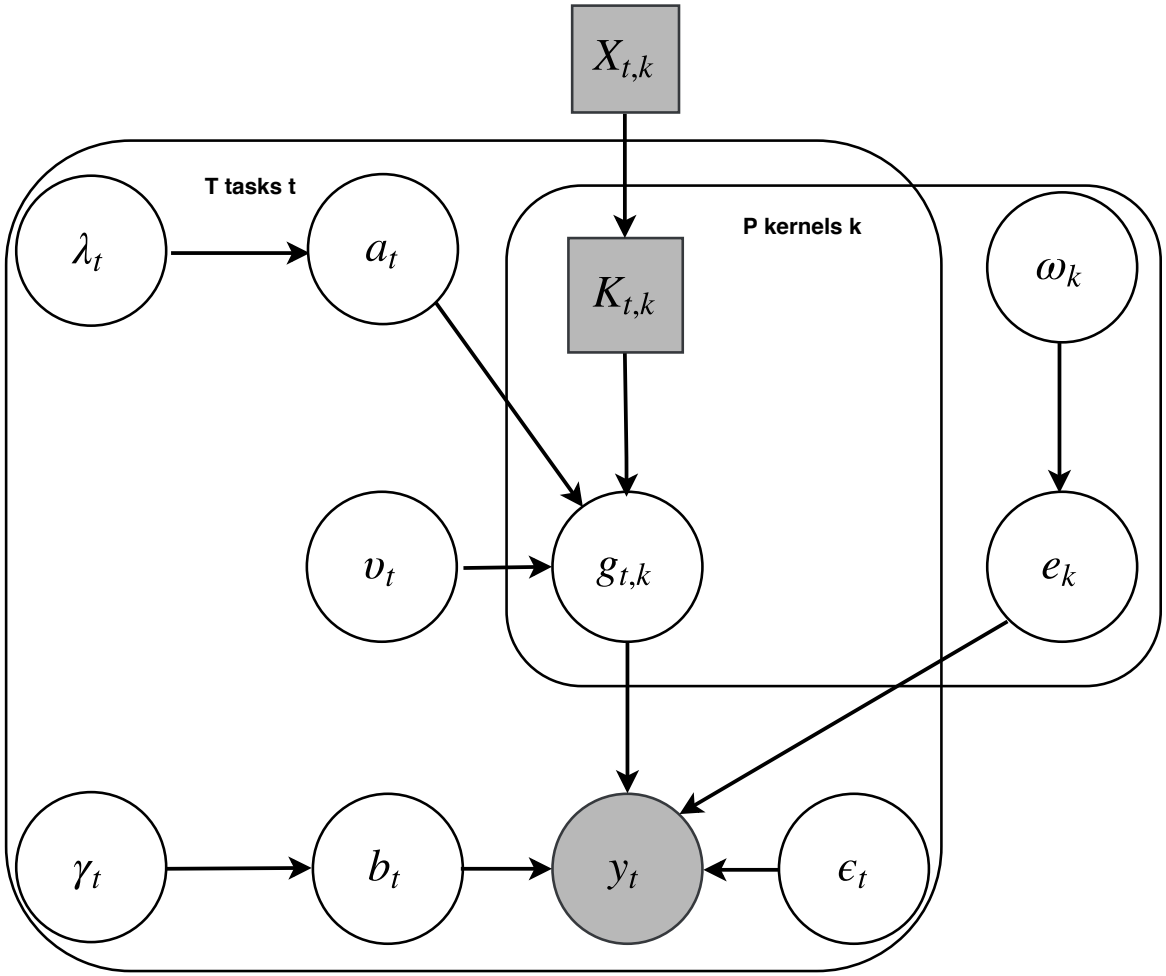
$$y_t | e, G_t, b_t, \epsilon_t \sim \mathcal{N}\left(\sum_{k=1}^{P} g_{t,k} e_k + b_t \cdot \mathbb{1}_{N_t} \ , \ \epsilon_t^{-1} I_{N_t}\right)$$

$$\sim \mathcal{N}\left(\sum_{k=1}^{P} (K_{t,k} \cdot a_t) e_k + b_t \cdot \mathbb{1}_{N_t} \ , \ \epsilon_t^{-1} I_{N_t}\right)$$

Dimensionality reduction from input to intermediate output



Model - Directed Acyclic Graph



References: [Costello et. al. (2014), Gönen, 2012a]

Distributional assumptions

$$\upsilon_t \sim \mathcal{G}am(\alpha_\upsilon, \beta_\upsilon) \qquad a_{t,i}|\lambda_{t,i} \sim \mathcal{N}(0, \lambda_{t,i}^{-1})$$
$$g_{t,k}|K_{t,k}, a_t, \upsilon_t \sim \mathcal{N}(K_{t,k}a_t, \upsilon_t^{-1} I_{N_t}) \qquad \lambda_{t,i} \sim \mathcal{G}am(\alpha_\lambda, \beta_\lambda)$$
$$\omega_k \sim \mathcal{G}am(\alpha_\omega, \beta_\omega) \qquad \gamma_t \sim \mathcal{G}am(\alpha_\gamma, \beta_\gamma)$$
$$e_k|\omega_k \sim \mathcal{N}(0, \omega_k^{-1}) \qquad b_t|\gamma_t \sim \mathcal{N}(0, \gamma_t^{-1})$$
$$\epsilon_t \sim \mathcal{G}am(\alpha_\epsilon, \beta_\epsilon)$$

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

$$\text{joint: } p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) = p(\lambda)p(a|\lambda)p(G|a, \upsilon, \mathcal{X})p(\upsilon)p(\omega)p(e|\omega) \cdot p(b|\gamma)p(\gamma)p(\epsilon)p(\mathcal{Y}|b, e, G, \epsilon)$$

Data Specifications for drugs

Data specification for drug response:

- drug information is standardized:
  - Standardization allows comparability of drugs despite of different level and variation of effectiveness (pooling).
- drug data is not imputed: not all drugs have been - sucessfully - tested on all cell-lines

Standardization of Input data used for constructing kernels:

- see sl. 55ff. in appendix

Implementation

- algorithm, data loading, imputation and kernel computation implemented in R
- optimization w.r.t hyperparameters of model and hyperparameters of kernel in progress
  - hyper parameters for Gamma priors
  - number of Features used of one data set
  - hyperparameters of kernels

Models by hyperparameters

Table 1: Four configurations of a grid search of hyperparameters with final ELBO value after 200 iterations using scaled drug data for shape $\alpha \in \{10^{-10}, 1, 10\}$ and scale $\theta \in \{10^{-10}, 0.01, 1\}$

	α_λ	β_λ	α_υ	β_υ	α_γ	β_γ	α_ω	β_ω	α_ϵ	β_ϵ
default	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$
ones	1	1	1	1	1	1	1	1	1	1
max	10	1	1	0.01	10	1	10	1	1	1
min	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	10	1	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$
random	$\frac{1}{10^{10}}$	$\frac{1}{10^{10}}$	1	$\frac{1}{10^{10}}$	10	1	1	1	1	1

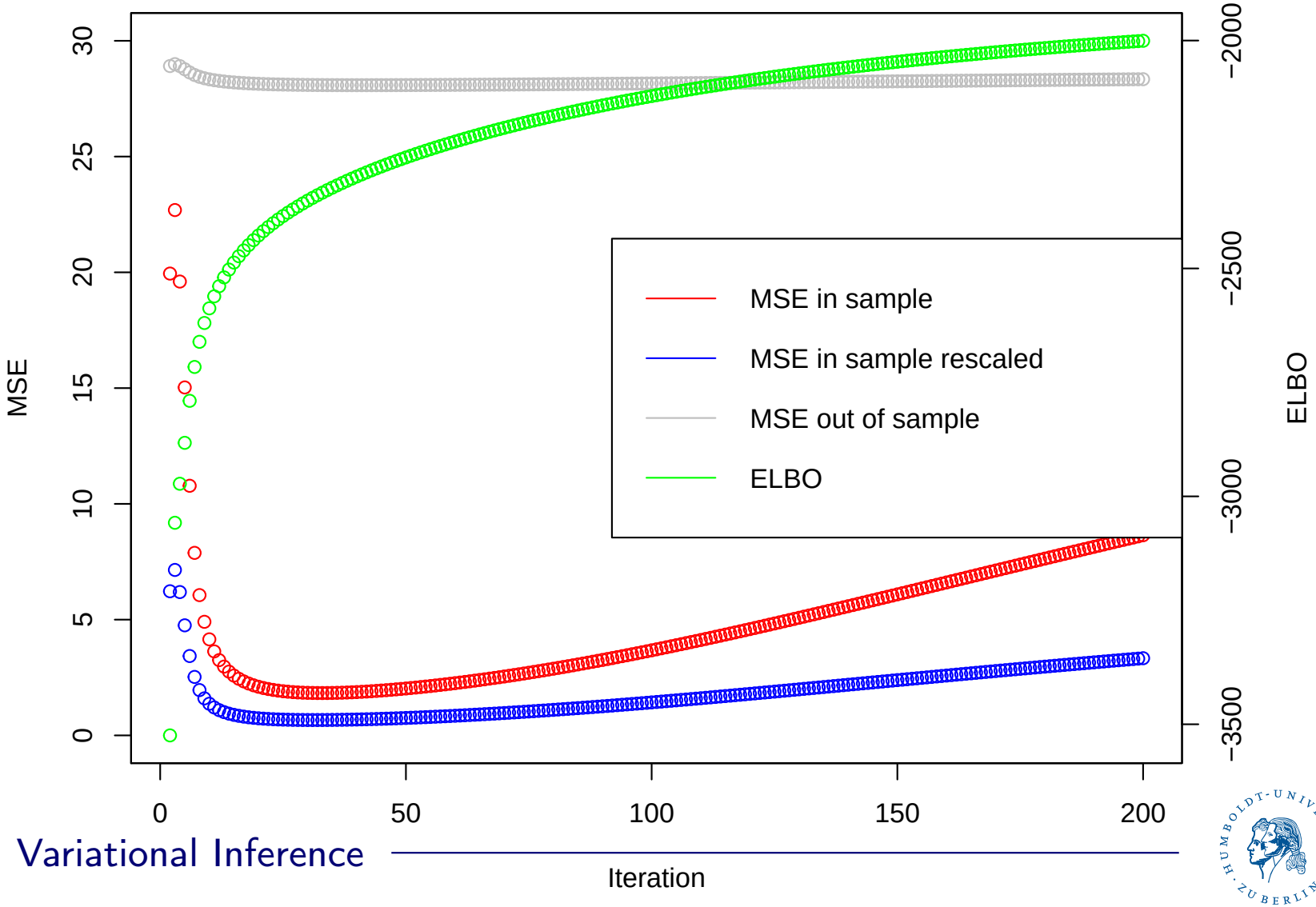
Descriptives on outcome for 5 drugs in training sample

	Drug1	Drug8	Drug16	Drug23	Drug25
N_t	30.00	30.00	34.00	32.00	17.00
No. NA	5.00	5.00	1.00	3.00	18.00
min	3.78	6.26	3.82	4.18	4.48
max	6.43	9.83	5.81	6.25	8.23
range	2.66	3.57	1.99	2.07	3.75
median	4.90	6.93	4.73	4.48	4.93
mean	4.89	7.02	4.73	4.49	5.87
var	0.23	0.45	0.11	0.14	2.13
std.dev	0.48	0.67	0.34	0.38	1.46

Results: out-of-sample predictions

	D1t	D1p	D8t	D8p	D23t	D23p
21NT	5.20	4.82	6.74	6.86	-	4.64
HCC3153	4.75	4.91	6.95	6.95	4.48	4.51
SUM225CWN	-	4.80	-	7.11	4.48	4.56
SUM149PT	5.73	4.90	6.72	6.99	4.91	4.61
ZR75B	4.74	4.94	7.40	6.93	4.48	4.48
SUM1315MO2	5.33	4.97	7.56	7.00	4.60	4.49
184B5	-	4.80	-	6.87	5.52	4.68
184A1	-	4.79	-	6.90	5.12	4.69
SUM159PT	-	4.90	-	7.00	6.33	4.57
MCF10A	5.02	4.86	7.01	6.96	5.87	4.62
LY2	4.34	4.88	6.46	6.73	4.48	4.46

Trade-off ELBO vs. MSE



Conclusion: DAGs, VI, BMTMKL

We have

- seen how to construct a Bayesian Regression Model incorporating dimensionality reduction, called BMTMKL
- a broad idea of Variational Inference (VI)
- seen an application modeling cancer drug responses given high dimensional inputs

Difficulties, Challenges and ToDos

- derivation needs multivariate reformulation of model distributions (priors)
- efficent learning of hyperparameters (Gamma prior parameters)
- Mean-Squared Error vs ELBO trade-off
- Implementation using Edward

Edward - Linear Bayesian Regression

Edward Tutorial
Supervised Learning (Regression)
edwardlib.org/tutorials/supervised-regression

For Further Reading - BMKMTL

Costello et. al. (2014)
A community effort to assess and improve drug sensitivity prediction algorithms
Nature Biotechnology, Vol. 32, no. 12,

Mehmet Gönen (2012)
A Bayesian Multiple Kernel Learning Framework for Single and Multiple Output Regression
ECAI 2012

Mehmet Gönen (2012)
Bayesian Efficient Multiple Kernel Learning
Proceedings of the 29th International Conference on Machine Learning

For Further Reading - Variational Inference


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


Hoffman, M. D., D. M. Blei, C. Wang, J. Paisley, and J. Edu (2012)
Stochastic Variational Inference
Journal of Machine Learning Research, 14, p. 1303-1347.

For Further Reading - Introduction


De Niz, Carlos and Rahman, Raziur and Zhao, Xiangyuan and Pal, Ranadip (2016)
Algorithms for Drug Sensitivity Prediction
Algorithms, Vol. 9 , no. 4

For Further Reading - Master thesis

Webel, Henry (2018)
Drug response in cancer treatments in a Bayesian Multiple Task Multiple Kernel Learning framework using Variational Inference for updating
Master thesis in statistics, visit
github.com/enryH/bmtmkl_dream7_thesis/

Variational Inference: joint, posterior and approx. posterior density in general notation

$$\text{joint} \quad p(\mathcal{Z}, \mathcal{Y}|\mathcal{X}) = p(\lambda, a, G, \upsilon, \gamma, \omega, b, e, \epsilon, \mathcal{Y}|\mathcal{X})$$

$$\text{posterior} \quad p(\mathcal{Z}|\mathcal{Y}, \mathcal{X}) = \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{p(\mathcal{Y}|\mathcal{X})} = \frac{p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})}{\int p(\mathcal{Z}, \mathcal{Y}|\mathcal{X})\, d\mathcal{Z}}$$

$$\text{approx. post.} \quad q(\mathcal{Z}) = q(\lambda)q(a)q(G)q(\upsilon)q(\gamma)q(\omega)q(b, e)q(\epsilon)$$

Set of latent random variables $\mathcal{Z} = \{\lambda, a, G, \upsilon, \gamma, \omega, b, e, \epsilon\}$, observed data for one drug t is $\mathcal{X}_t = \{K_{t,1}, \ldots, K_{t,P}\} = \{K_{t,k}\}_{k=1}^{P}$ with corresponding outcome vector y_t, for all T drugs at once $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_T\}$ and $\mathcal{Y} = \{y_1, \ldots, y_T\} = \{y_t\}_{t=1}^{T}$. Observed data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

Kullback Leibler (KL) Divergence

Minimization of the difference between true posterior $p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and approximated posterior $q(\mathcal{Z})$ is the objective. Non-symmetric KL-divergence used

$$\mathrm{KL}\left[q(\mathcal{Z})||p(\mathcal{Z}|\mathcal{X},\mathcal{Y})\right] = \mathrm{E}_{q(\mathcal{Z})}\left[\log\frac{q(\mathcal{Z})}{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\right] \geq 0$$

$p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and $q(\mathcal{Z})$ are non-negative as density functions. Thus it follows by Jensens inequality for convex functions:

$$\begin{aligned}\mathrm{E}_{q(\mathcal{Z})}\left[-\log\frac{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}{q(\mathcal{Z})}\right] &\geq -\log\mathrm{E}_{q(\mathcal{Z})}\left[\frac{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}{q(\mathcal{Z})}\right] \\ &= -\log\mathrm{E}_{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\left[1\right] = 0\end{aligned}$$

Minimizing KL betw. approx. and posterior

Evidence and evidence lower bound form KL-divergence between posterior $p(\mathcal{Z}|\mathcal{Y},\mathcal{X})$ and approximated posterior $q(\mathcal{Z})$:

$$\begin{aligned}
\mathrm{KL}[\underset{(approx.)}{q(\mathcal{Z})}\,||\underset{(posterior)}{p(\mathcal{Z}|\mathcal{X},\mathcal{Y})}] &= \mathrm{E}_q\left[\log\left(\frac{q(\mathcal{Z})}{p(\mathcal{Z}|\mathcal{Y},\mathcal{X})}\right)\right] \quad \text{, note: } \mathrm{E}_{q(\mathcal{Z})}[\cdot] = \mathrm{E}_q[\cdot] \\
&= \mathrm{E}_q[\log q(\mathcal{Z})] - \mathrm{E}_q[\log p(\mathcal{Z},\mathcal{Y}|\mathcal{X})] + \log p(\mathcal{Y}|\mathcal{X}) \\
&= \log p(\mathcal{Y}|\mathcal{X}) - E_q\left[\log \frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right] \\
&= \log p(\mathcal{Y}|\mathcal{X}) - \mathcal{L}(q(\mathcal{Z})) \quad \text{, } \mathcal{L}(q(\mathcal{Z})) \text{ is lower bound}
\end{aligned}$$

$$\min_{q(\mathcal{Z})} \quad \mathrm{KL}[q(\mathcal{Z})\|p(\mathcal{Z},\mathcal{Y}|\mathcal{X})]$$

$$\Rightarrow \quad \arg\min_{q(\mathcal{Z})} \; -\mathcal{L}\left(q(\mathcal{Z})\right) \quad \Rightarrow \max_{q(\mathcal{Z})} \mathcal{L}\left(q(\mathcal{Z})\right)$$

ELBO - $\mathcal{L}(q(\mathcal{Z}))$ - evidence lower bound

The log evidence, $\log p(\mathcal{Y}|\mathcal{X})$, is bigger than or equal to $\mathcal{L}(q(\mathcal{Z}))$, which is thus called the lower bound of the evidence (or ELBO):

$$
\begin{aligned}
\log(p(\mathcal{Y}|\mathcal{X})) &= \log\left(\int p(\mathcal{Z},\mathcal{Y}|\mathcal{X})dz\right) \\
&= \log\left(\int \frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\cdot q(\mathcal{Z})dz\right) \\
&= \log\left(\mathsf{E}_{q(\mathcal{Z})}\left[\frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right]\right) \\
\text{(Jensen)} \quad &\geq \mathsf{E}_{q(\mathcal{Z})}\left[\log\left(\frac{p(\mathcal{Z},\mathcal{Y}|\mathcal{X})}{q(\mathcal{Z})}\right)\right] =: \underset{(ELBO)}{\mathcal{L}(q(\mathcal{Z}))} \qquad (1) \\
&= -\,\mathsf{KL}[q(\mathcal{Z})||p(\mathcal{Z},\mathcal{Y}|\mathcal{X})]
\end{aligned}
$$

Updating - inspect ELBO [Blei et. al., 2017]

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

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:

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

$p(z_j|\mathcal{D}, \mathcal{Z}_{-z_j})$ is called full conditional. Hidden random var. minus z_j is $\mathcal{Z}_{-z_j} = \{\mathcal{Z} \setminus z_j\}$ and data is $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

Could we compute the likelihood instead ?

Specify density for random variable y_t (density $\mathcal{N}(y_t|\mu_{y_t}, \Sigma_{y_t})$):

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

Input weights a_t, coeff. e_k, intercept b_t and precision ϵ_t latent:

- N_t values have to be inferred for each a_t (which is proportional to the number of cell lines) $\Rightarrow$ **No**
- P values would have to be inferred for $e = (e_1, \ldots, e_P)'$
- T values for both $b = (b_1, \ldots, b_T)'$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$

joint den (stating dep. on random variables)

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

joint density (stating parameters)

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

(Bayesian) Linear Regression

Lets consider a row vector of inputs $x_{t,i} = (x_{t,i,1}, \ldots, x_{t,i,P}) \in \mathbb{R}^P$, a vector of coefficients $e = (e_1, \ldots, e_P)'$, a bias (intercept) b_t and a precision (inverse variance) term ϵ_t. The outcome vector of observations we want to predict is $y_t = (y_{t,1}, \ldots, y_{t,N_t})' \in \mathbb{R}^{N_t}$. Its conditional distribution is a normal with mean $\mu_{y_t}(x, e, b_t)$ and variance $\Sigma_{y_t}(\epsilon_t)$:

$$y_{t,i}|x_{t,i}, e, b_t, \epsilon_t \sim \mathcal{N}\left(y_t \,\middle|\, \sum_{k=1}^{P} x_{t,i,k} \cdot e_k + b_t\,,\ \epsilon_t^{-1}\right) \tag{2}$$

One classic approach: Compute log-likelihood and take its derivatives OR take sum of squared residuals and take these derivatives.

- Bayesian Linear Regression: Introduce a normal prior on coefficient vector $w = (b_t, e')'$ and eventually also a gamma prior on precision ϵ_t.

$$p(w) = \mathcal{N}(w|\mu_w, \Sigma_w) = \mathcal{N}(w \,|\, 0\,,\, \epsilon_t^{-1} I_{P+1}) \tag{3}$$

$$p(\epsilon_t) = \mathcal{G}am(\epsilon_t|\alpha_\epsilon, \beta_\epsilon) \tag{4}$$

where α_ϵ and β_ϵ are the so called shape and rate parameters.

Kernels in BMTMKL (Non-linear similarity measure)

Gaussian kernel (real-valued inputs)

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

where $\sigma_{t,k}^2$ is set to the dimensionality (i.e. the number of features or variables) of the corresponding genomic view.

Jaccard similarity kernel (binary inputs)

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

DREAM7 challenge

The NCI-DREAM Drug Sensitivity Prediction Challenge

- 35 training cell lines, 18 test cell lines
- 6 profiling datasets (omics): RPPA, Expression, DNA Methylation, RNA-seq, Copy Number Variation (CNV), Exome-seq
- 28 unknown drugs
- several missing information for inputs (omics) and drug responses y_t

abriviation	short description as given by [Costello et. al. (2014)] in data files
RPPA	An antibody-based method to quantitatively measure protein abundance. RPPA data were generated and pre-processed as previously described (131 proteins assayed)
Expression [1]	Transcript expression values. Affymetrix GeneChip Human Gene 1.0 ST microarrays were processed using R package aroma.affymetrix (over 18,000 expression values)
Methylation	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)

abriviation	short description as given by [Costello et. al. (2014)] in data files
RNA-seq	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 (just under 37,000 RNAs)
Copy number	DNA copy-number variation (CNV). Affymetrix Genome-Wide Human SNP6.0 Array.
Exome-seq	Whole exome sequencing (exome seq). Mutation status was obtained from exome-capture sequencing (Agilent Sure Select system).

Data Specifications for omics (gene views)

Overall, non-task-specific input matrices $X_{\cdot,k}$:

- missing data of gene information has to be imputed
  - different cell-lines have varying set of available information: all cell-lines need to be represented in all inputs by adding a place holder for missing values (NA)
  - data is standardized w.r.t. columns containing the features, i.e. imputing the mean for NAs: Set missing values to zero
- further data can be added, also by adding interactions and pathway information (c.p. sl. 58)

$\Rightarrow$ kernel coefficients $e_{1:P}$ (the coefficients for the linear regression on intermediate outputs) for each input data set (aka kernels/ views) can be learned jointly over all drugs.

Data Specifications for omics (gene views)

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

type	name	# feat.	e_k
RNA	Gene Expression	18632	e_1
DNA	Copy Number Variation	27234	e_2
	Methylation	27551	e_3
RNA	RNA-seq	36953	e_4
	discretized RNA-seq		e_7
DNA	Exome seq	10607	e_5
	discretized Exome seq		e_8
Proteins	RPPA	66	e_6
pathways	Reactom.db und Hs.Org.db		e_9- e_{18}
Interactions	Multiplication of kernels		e_{19} -e_{22}

add so called pathway information: subsetting the data

Genes form biological pathways for certain chemical reactions: Activation on the basis of the values of certain genes can be computed for these pathways.

Pathway: Subset of data (here: genes 1, 23 and 2340)

	gene 1	gene 23	gene 2340
cell 1	value 1,1	value 1, 23	value 1, 2340
$\vdots$	$\vdots$	$\vdots$	$\vdots$
cell N_t	gene$N_t, 1$	value$N_t, 23$	value $N_t, 2340$

Table 2: Pathways

pathway	statistic	e_k
gene expression - Reactom	average	e_9
gene expression - Org.Hs.Eg		e_{10}
Methylation - Reactom	maximum	e_{11}
Methylation - Org.Hs.Eg		e_{12}
Copy Number Variation - Reactom	maximum	e_{13}
Copy Number Variation - Org.Hs.Eg		e_{14}
Exome-seq - Reactom	maximum	e_{15}
Exome-seq - Org.Hs.Eg		e_{16}
RNASeq - Reactom	maximum	e_{17}
RNASeq - OrgHsEg		e_{18}

Table 3: Interacted kernels of original datasets

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

BMTMKL - out-of-sample predictions - intermediate outputs

The density of the intermediate outputs is

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

The asterisk $*$ symbol denotes the new or testing data. The dimension of a vector of kernels between train and a test data point $k_{t,k,*}$ is $\mathbb{R}^{N_t}$. Further definition see [Webel, Henry (2018)].

BMTMKL - out-of-sample predictions - predictive density

The density for a out-of-sample outcome $y_{t,*}$ is

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

In the end, the value of interest is the mean of $y_{t,*}$ as the point prediction with the specified variance:

$$\mathrm{E}_q\left[y_{t,*}\right] = \hat{y}_{t,*} = \left(1, \mathrm{E}_q\left[{g_{t,*}}'\right]\right) \cdot \mathrm{E}_q\left[b_t, e\right] \quad ,$$

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