Drug sensitivity prediction with normal inverse Gaussian shrinkage informed by external data

Magnus M. Münch\textsuperscript{1,2*}, Mark A. van de Wiel\textsuperscript{1,3}, Sylvia Richardson\textsuperscript{3}, and Gwenaël G. R. Leday\textsuperscript{3}

April 30, 2020

1. Department of Epidemiology & Biostatistics, Amsterdam UMC, VU University, PO Box 7057, 1007 MB Amsterdam, The Netherlands
2. Mathematical Institute, Leiden University, Leiden, The Netherlands
3. MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom

1 Content overview

This document contains the Supplementary Material (SM) to the document ‘Drug sensitivity prediction with normal inverse Gaussian shrinkage informed by external data’. In the following, this document is referred to as Main Document (MD).

Section 2 describes the moments of the normal inverse Gaussian (NIG) prior. Section 3 describes the hyperparameter settings to generate Figure 1 in the MD and Section 4 outlines the derivation to arrive at the Jeffreys prior for the error variance in equation (2d) of the MD. Section 5 gives the variational Bayes (BV) derivations that lead to the estimating equations in the MD, Section 6 gives the evidence lower bound (ELBO) that is maximised during the VB iterations. In Section 7 efficient VB calculations are given, while Section 8 outlines a computationally stable method to calculate ratios of modified Bessel functions. Section 9 shows that the empirical Bayes (EB) scale estimates $\lambda$ are non-negative. Section 10 introduces a Gibbs sampling scheme to sample from the posterior corresponding to the full Bayesian version of the NIG model. Finally, Sections 11 and 12 contain additional results to the simulations and GDSC data analyses in Sections 4 and 5 in the MD, respectively.

2 Normal inverse Gaussian prior moments

Let $\beta \sim \mathcal{N}(0, \sigma^2 \tau^2 \gamma^2)$, where $\sigma^2$ and $\tau^2$ are fixed and $\gamma^2 \sim \mathcal{IG}(\phi, \lambda_{\text{feat}})$, then

\begin{align*}
E(\beta | \sigma^2, \tau^2) &= 0, \\
V(\beta | \sigma^2, \tau^2) &= \sigma^2 \tau^2 \phi, \\
K(\beta | \sigma^2, \tau^2) &= \frac{3\phi}{\lambda_{\text{feat}}} + 3,
\end{align*}

*Correspondence to: m.munch@amsterdamumc.nl
where $\mathcal{K}(\cdot)$ denotes the kurtosis. Now letting $\tau^2 \sim IG(\chi, \lambda_{\text{drug}})$ and integrating it out we have:

$$
E(\beta|\sigma^2) = 0,
$$

$$
\Psi(\beta|\sigma^2) = \sigma^2 \chi \phi,
$$

$$
\mathcal{K}(\beta|\sigma^2) = 3 \left( \frac{\phi}{\lambda_{\text{feat}}} - \frac{\chi}{\lambda_{\text{drug}}} + \frac{\phi}{\lambda_{\text{drug}}} \right) + 3.
$$

### 3 Hyperparameter settings

For the NIG shrinkage weights prior in Figure 1a the following hyperparameter settings (in terms of the inverse Gaussian prior) were used: $\lambda = 1$, $\phi = 10$ (solid), $\lambda = 0.1$, $\phi = 10$ (dashed), and $\lambda = 1$, $\phi = 1$ (dotted). The Student’s $t$ prior on the regression parameters $\beta_{jd}$ arises if $\gamma_{jd}^2 \sim \Gamma^{-1}(a,b)$, with $\Gamma^{-1}(a,b)$ an inverse Gamma distribution with shape $a$ and scale $b$. Figure 1b is then obtained by setting: $a = 0.9$, $b = 1.3$ (solid), $a = 1$, $b = 4$ (dashed), and $a = 1$, $b = 0.2$ (dotted). The lasso prior results from an exponential prior on $\gamma_{jd}^2$. To create Figure 1c the following settings were used for exponential rate $\lambda$: $\lambda = 1$ (solid), $\lambda = 0.2$ (dashed), and $\lambda = 0.1$ (dotted).

### 4 Jeffrey’s prior

We parametrise the data with a Gaussian distribution with unknown mean and variance, i.e., $y_i \sim N(\mu, \sigma^2)$. This results in the following Hessian (matrix of second derivatives) of the log density:

$$
H(\log f(y_i|\mu, \sigma^2)) = \begin{bmatrix}
-\frac{1}{\sigma^2} y_i - \mu & 1 \\
\frac{1}{\sigma^4} - \frac{(y_i - \mu)^2}{\sigma^6} & 0
\end{bmatrix}.
$$

Jeffreys prior (Jeffreys, 1946) is now found as:

$$
\pi(\mu, \sigma^2) \propto |I(\mu, \sigma^2)|^{1/2} = |E(H)|^{1/2} = \begin{bmatrix}
\frac{1}{\sigma^2} & 0 \\
0 & \frac{1}{2\sigma^4}
\end{bmatrix}^{1/2} \propto 1/\sigma^3,
$$

where $I(\theta)$ denotes the Fisher information of $\theta$.

### 5 Variational Bayes derivations

For clarity we have indexed the variational density functions with their respective parameters, which we omitted in the MD. In the following all expectations are with respect to the variational posterior $Q_d$. The variational posterior for $\beta_d$ is found as follows:

$$
\log q_{\beta_d}(\beta_d) \propto E[\log L(y_d|\beta_d, \sigma_d^2)] + E[\log \pi(\beta_d|\gamma_d^2, \sigma_d, \tau_d^2)]
$$

$$
\propto -\frac{1}{2} \sum_{i=1}^{n} E \left[ \left( \frac{y_{id} - \mu_{id}}{\sigma_d} \right)^2 \right] - \frac{1}{2} \sum_{j=1}^{p} E \left( \frac{\beta_{jd}^2}{\gamma_{jd}^2 \tau_d} \right),
$$

$$
q_{\beta_d}(\beta_d) \overset{D}{=} N_p(\mu_d, \Sigma_d),
$$

with $\Sigma_d = E(\sigma_d^2)^{-1} \{X^TX + E(\tau_d^{-2})\text{diag}[E(\gamma_{jd}^2)]\}^{-1}$, and $\mu_d = \{X^TX + E(\tau_d^{-2})\text{diag}[E(\gamma_{jd}^2)]\}^{-1}X^Ty_d$. 

2
The variational posterior for $\gamma_d^2$ is given by:

$$
\log q_{\gamma_d^2}(\gamma_d^2) \propto \mathbb{E}[\log \pi(\beta_d|\gamma_d^2, \sigma_d^2, \tau_d^2)] + \log \pi(\gamma_d^2)
$$

$$
\propto -\frac{1}{2} \log \gamma_d^2 - \frac{1}{2} \mathbb{E} \left( \frac{\beta_d^2}{\gamma_d^2} \right) \gamma_d^{-2} - \frac{3}{2} \log \gamma_d^2 - \frac{\lambda_{\text{feat}} (\gamma_d^2 - \phi_d)^2}{2\phi_d^2} \gamma_d^{-2}
$$

$$
\propto (-1 - 1) \log \gamma_d^2 - \frac{\lambda_{\text{feat}}}{2\phi_d^2} \gamma_d^{-2} - \frac{1}{2} \left[ \lambda_{\text{feat}} + \mathbb{E}(\tau_d^{-2}) \mathbb{E}(\sigma_d^{-2}) \mathbb{E}(\beta_d^2) \right] \gamma_d^{-2},
$$

$$
q_{\gamma_d^2}(\gamma_d^2) \overset{D}{=} GIG \left(-1, \lambda_{\text{feat}}/\phi_d^2, \delta_d \right),
$$

with $\delta_d = \mathbb{E}(\tau_d^{-2}) \mathbb{E}(\sigma_d^{-2}) \mathbb{E}(\beta_d^2) + \mathbb{V}(\beta_d) + \lambda_{\text{feat}}$.

Similarly, we derive the variational posterior for $\tau_d^2$ as:

$$
\log q_{\tau_d^2}(\tau_d^2) \propto \mathbb{E}[\log \pi(\beta_d|\gamma_d^2, \sigma_d^2, \tau_d^2)] + \log \pi(\tau_d^2)
$$

$$
\propto \frac{p}{2} \log \tau_d^2 - \frac{1}{2} \sum_{j=1}^{p} \mathbb{E} \left( \frac{\beta_d^2}{\tau_d^2} \right) \tau_d^{-2} - \frac{3}{2} \log \tau_d^2 - \frac{\lambda_{\text{drug}} (\tau_d^2 - \lambda_d)^2}{2\lambda_d^2} \tau_d^{-2}
$$

$$
\propto \left( -\frac{p + 1}{2} - 1 \right) \log \tau_d^2 - \frac{\lambda_{\text{drug}}}{2\lambda_d^2} \tau_d^{-2} - \frac{1}{2} \left[ \lambda_{\text{drug}} + \mathbb{E}(\sigma_d^{-2}) \sum_{j=1}^{p} \mathbb{E}(\gamma_j^{-2}) \mathbb{E}(\beta_j^2) \right] \tau_d^{-2},
$$

$$
q_{\tau_d^2}(\tau_d^2) \overset{D}{=} GIG \left(-\frac{p + 1}{2}, \lambda_{\text{drug}}/\lambda_d^2, \eta_d \right),
$$

with $\eta_d = \mathbb{E}(\sigma_d^{-2}) \sum_{j=1}^{p} \mathbb{E}(\gamma_j^{-2}) \mathbb{E}(\beta_j^2) + \mathbb{V}(\beta_d) + \lambda_{\text{drug}}$.

Lastly, we have the variational posterior for $\sigma_d^2$:

$$
\log q_{\sigma_d^2}(\sigma_d^2) \propto \mathbb{E}[\log \mathcal{L}(\mathbf{y}_d|\beta_d, \sigma_d^2)] + \mathbb{E}[\log \pi(\beta_d|\gamma_d^2, \sigma_d^2, \tau_d^2)] + \log \pi(\sigma_d^2)
$$

$$
\propto -\frac{n}{2} \log \sigma_d^2 - \frac{1}{2} \sum_{i=1}^{n} \mathbb{E} \left( \frac{y_i^2 - \mathbf{x}_i^T \beta_d)^2}{\sigma_d^2} \right) - \frac{p}{2} \log \sigma_d^2 - \frac{1}{2} \sum_{j=1}^{p} \mathbb{E} \left( \frac{\beta_d^2}{\tau_d^2} \right) \sigma_d^{-2} - \frac{3}{2} \log \sigma_d^2
$$

$$
= \left( -\frac{n + p + 1}{2} - 1 \right) \log \sigma_d^2
$$

$$
- \frac{1}{2} \left\{ \sum_{i=1}^{n} \mathbb{E}((y_i - \mathbf{x}_i^T \beta_d)^2) + \mathbb{E}(\tau_d^{-2}) \sum_{j=1}^{p} \mathbb{E}(\gamma_j^{-2}) \mathbb{E}(\beta_j^2) \right\} \sigma_d^{-2},
$$

$$
q_{\sigma_d^2}(\sigma_d^2) \overset{D}{=} \Gamma^{-1} \left( \frac{n + p + 1}{2}, \zeta_d \right),
$$

with $\zeta_d = \frac{1}{2} \left\{ \mathbf{y}_d^T \mathbf{y}_d - 2\mathbf{y}_d^T \mathbf{X} \mathbb{E}(\beta_d) + \text{tr}[\mathbf{X}^T \mathbf{X} \mathbb{V}(\beta_d)] + \mathbb{E}(\beta_d^T) \mathbf{X} \mathbb{E}(\beta_d)
$$

$$
+ \mathbb{E}(\tau_d^{-2}) \text{tr} \{ \text{diag}[\mathbb{E}(\gamma_j^{-2})] \mathbb{V}(\beta_d) \} + \mathbb{E}(\tau_d^{-2}) \mathbb{E}(\beta_d^T) \text{diag}[\mathbb{E}(\gamma_j^{-2})] \mathbb{E}(\beta_d) \right\}.
$$

If we fill in the expectations and variances, we arrive at the estimating equations in MD equations (3).
6 Evidence lower bound

The evidence lower bound that is maximised during the iterations is: \( \text{ELBO}^{(l)} = \sum_{d=1}^{\mathcal{D}} \text{ELBO}^{(l)}_d \), with:

\[
\text{ELBO}^{(l)}_d = - \frac{n + p + 1}{2} \log \pi + \frac{p + 1 - n}{2} \log 2 + \frac{n + 1}{2} + p + \log \Gamma \left( \frac{n + p + 1}{2} \right) \\
- \frac{n + p + 1}{2} \log \zeta^{(l)}_d + \frac{1}{2} \log \|\Sigma_d^{(l)}\| - \frac{n + p + 1}{2e_{\text{ed}}^{(l)}} \left[ y_d^T y_d - 2y_d^T X \mu_d^{(l)} + \text{tr}(X^T X \Sigma_d^{(l)}) \right] \\
+ (\mu_d^{(l)})^T X^T X \mu_d^{(l)} + g_d^{(l)} \text{tr} \left[ \text{diag}(b_{jd}^{(l)} \Sigma_d^{(l)}) \right] + g_d^{(l)} (\mu_d^{(l)})^T \text{diag}(b_{jd}^{(l)}) \mu_d^{(l)} + p \log \lambda_d^{(l)} \\
+ \frac{p + 3}{4} \log \lambda_d^{(l)} + \frac{\eta_d^{(l)} - \lambda_d^{(l)} \text{drug}}{2} g_d^{(l)} + \frac{p}{2} \sum_{j=1}^{p} \delta_{jd}^{(l)} - \frac{\lambda_d^{(l)} \text{drug}}{2} b_{jd}^{(l)} + \frac{p}{2} \sum_{j=1}^{p} \delta_{jd}^{(l)} + \sum_{j=1}^{p} \frac{\lambda_d^{(l)} \text{drug}}{\phi_{jd}^{(l)}} \\
- \frac{p + 1}{2} \log \chi_d^{(l)} - \sum_{j=1}^{p} \log \phi_{jd}^{(l)} - \frac{p + 1}{4} \log \eta_d^{(l)} - \frac{1}{2} \sum_{j=1}^{p} \log \delta_{jd}^{(l)} \\
+ \log K_{\frac{p+1}{2}} \left( \sqrt{\eta_d^{(l)}} \frac{\lambda_d^{(l)} \text{drug}}{\chi_d^{(l)}} \right) + \sum_{j=1}^{p} \log K_1 \left( \sqrt{\delta_{jd}^{(l)}} \lambda_d^{(l)} / \phi_{jd}^{(l)} \right).
\]

7 Efficient computation

The empirical Bayes updates require the quantities \( e_{jd}, b_{jd}, g_d, \) and \( f_d \), which in turn depend on the quantities \( \text{diag}(\Sigma_d), \mu_d, y_d^T X \mu_d, \text{tr}(X^T X \Sigma_d), \mu_d^2 X^T X \mu_d, \text{tr}(\text{diag}(b_{jd}) \Sigma_d), \) and \( g_d \mu_d^2 \text{diag}(b_{jd}) \mu_d \). Let \( H_d = g_d \text{diag}([b_{jd}]^{-1}) \). Then, the first two quantities are efficiently calculated as:

\[
\text{diag}(\Sigma_d^{(h+1)}) = (\eta_d^{(h)})^{-1} H_d^{(h)} - (\eta_d^{(h)})^{-1} [H_d^{(h)} X^T (I_n + X H_d^{(h)} X^T)^{-1} \circ (H_d^{(h)} X^T)] 1_{n \times 1}, \quad (1a) \\
\mu_d^{(h+1)} = H_d^{(h)} X^T y_d - H_d^{(h)} X^T (I_n + X H_d^{(h)} X^T)^{-1} X H_d^{(h)} X^T y_d. \quad (1b)
\]

The quantities \( y_d^T X \mu_d, \mu_d^2 X^T X \mu_d, \text{tr}(\text{diag}(b_{jd}) \Sigma_d), \) and \( g_d \mu_d^2 \text{diag}(b_{jd}) \mu_d \) are easily calculated from the first two. For the remaining one we have:

\[
\text{tr}(X^T X \Sigma_d^{(h+1)}) = (a_d^{(h)})^{-1} 1_{1 \times p} H_d^{(h)} X^T 1_{n \times 1} \\
- (a_d^{(h)})^{-1} 1_{1 \times p} [(H_d^{(h)} X^T (I_n + X H_d^{(h)} X^T)^{-1} X H_d^{(h)} X^T) \circ X^T] 1_{n \times 1}. \quad (2)
\]

Both (1) and (2) are operations of complexity \( \mathcal{O}(pm^2) \) instead of \( \mathcal{O}(p^3) \) for naive calculation.

8 Ratios of modified Bessel functions

Ratios of modified Bessel functions of the second kind \( K_{\alpha-1}(x)/K_{\alpha}(x) \) are prone to under- and overflow for large \( \alpha \). In our case, \( \alpha \) increases linearly with \( p \). Since \( p \) may be large, this may cause numerical issues in the calculation of various quantities. We alleviate the numerical issues through the following.
We let \( n_1 = \frac{p}{2} \) and \( n_2 = \frac{(p - 1)}{2} \) and use the well-known recursive relation:

\[
K_\alpha(x) = K_{\alpha-2}(x) + \frac{2(\alpha - 1)}{2} K_{\alpha-1}(x),
\]

to rewrite the ratio:

\[
\frac{K_{p-1}(x)}{K_{p+1}(x)} = \left\{ \begin{array}{ll}
\left( \ldots \left( \left( 1 + \frac{2a-1}{x} \right)^{-1} + \frac{\alpha+1}{x} \right)^{-1} + \cdots + \frac{2n-1}{x} \right)^{-1}, & \text{for } p \text{ even}, \\
\left( \ldots \left( \frac{K_\alpha(x)}{K_{\alpha-2}(x)} + \frac{2}{x} \cdot 1 \right)^{-1} + \frac{\alpha}{x} \cdot 2 \right)^{-1} + \cdots + \frac{2}{x} \cdot n_2 \right)^{-1}, & \text{for } p \text{ odd}.
\end{array} \right.
\]

The ratio \( K_0(x)/K_1(x) \) is well-behaved, so that an arbitrary ratio \( K_{\alpha-1}(x)/K_\alpha(x) \) may be calculated as a sequence of numerically stable scalar sums, products and inverses.

### 9 Empirical Bayes

For the feature scale parameter \( \lambda_{\text{feat}} \) updates to be non-negative, we have to show that:

\[
\sum_{d=1}^{D} \sum_{j=1}^{p} b_{jd} + \alpha_{\text{feat}}^T C^T \text{diag}(e_{jd}) C \alpha_{\text{feat}} - 2\alpha_{\text{feat}}^T C^T 1_{p \times 1} = \sum_{d=1}^{D} \sum_{j=1}^{p} (b_{jd} + \phi_{jd}^{-2} e_{jd} - 2\phi_{jd}^{-1}) \geq 0, \tag{3}
\]

which holds if all summands in the left-hand side of (3) are non-negative.

We note that through Jensen’s inequality \( e_{jd}^{-1} = E(\gamma_{jd}^2)^{-1} \leq E(\gamma_{jd}^{-2}) = b_{jd} \), we may lowerbound the summands in left-hand side of (3):

\[
b_{jd} + \phi_{jd}^{-2} e_{jd} - 2\phi_{jd}^{-1} \geq e_{jd}^{-1} + \phi_{jd}^{-2} e_{jd} - 2\phi_{jd}^{-1},
\]

which we use to show that:

\[
\begin{align*}
    & e_{jd}^{-1} + \phi_{jd}^{-2} e_{jd} - 2\phi_{jd}^{-1} \geq 0 \\
    & 1 + \phi_{jd}^{-1} e_{jd}^2 - 2\phi_{jd}^{-1} e_{jd} \geq 0 \\
    & \left( 1 - \phi_{jd}^{-1} e_{jd} \right)^2 \geq 0.
\end{align*}
\]

Similar reasoning holds for the drug scale parameter \( \lambda_{\text{drug}} \).

### 10 Gibbs sampler

MCMC samples from the posterior corresponding to model (2) may be generated for each equation independently. MCMC samples may be generated by iteratively sampling the following conditional
distributions:

$$\beta_d | \gamma_d^2, \sigma_d^2, \tau_d^2, y_d \sim N(\mu_d, \Sigma_d),$$  

(4a)

with $$\Sigma_d = \sigma_d^2 [X^T X + \tau_d^{-2} \text{diag}(\gamma_d^{-2})]^{-1},$$  

(4b)

and $$\mu_d = [X^T X + \tau_d^{-2} \text{diag}(\gamma_d^{-2})]^{-1} X^T y_d,$$  

(4c)

$$\gamma_{jd}^2 | \beta_{jd}, \sigma_{jd}^2, \tau_{jd}^2, y \sim GIG(-1, \lambda_{\text{feat}}/\phi_{jd}^2, \delta_{jd}),$$  

(4d)

with $$\delta_{jd} = \sigma_{jd}^{-2} \tau_{jd}^{-2} \beta_{jd}^2 + \lambda_{\text{feat}},$$  

(4e)

$$\tau_{jd}^2 | \beta_{jd}, \sigma_{jd}^2, \gamma_{jd}^2, y \sim \Gamma\left(-\frac{n + p + 1}{2}, \frac{1}{2} \left(y_d^T y_d - 2 y_d^T X \beta_d + \beta_d X^T X \beta_d + \tau_d^{-2} \beta_d^T \text{diag}(\gamma_d^{-2}) \beta_d \right)\right),$$  

(4f)

with $$\zeta_d = \frac{1}{2} \left[ y_d^T y_d - 2 y_d^T X \beta_d + \beta_d X^T X \beta_d + \tau_d^{-2} \beta_d^T \text{diag}(\gamma_d^{-2}) \beta_d \right].$$  

(4g)

In high dimensional space, the $$p \times p$$ matrix inversions in (4b) and (4c) are significant computational bottlenecks. Bhattacharya et al. (2016) describe a method to sample from (4a) without explicit calculation of this inverse, thereby offering a significant speed-up compared to naive sampling.

Sampling from (4) either requires estimating (as in MD Section 3.2) or specifying hyperparameters $$\alpha_{\text{feat}}, \alpha_{\text{drug}}, \lambda_{\text{feat}},$$ and $$\lambda_{\text{drug}},$$ or to endow them with and extra layer of hyperpriors. In general, specifying them requires rarely available, detailed subject knowledge, so we may resort to endowing them with hyperpriors (as in Upadhyay and Agrawal (1996)):

$$\alpha_{\text{feat}, g} | \lambda_{\text{feat}} \sim N(0, \nu_{\text{feat}}^2 / \lambda_{\text{feat}}),$$

$$\alpha_{\text{drug}, h} | \lambda_{\text{drug}} \sim N(0, \nu_{\text{drug}}^2 / \lambda_{\text{drug}}),$$

$$\lambda_{\text{feat}} \sim \Gamma(\kappa_{\text{feat}}, \xi_{\text{feat}}),$$

$$\lambda_{\text{drug}} \sim \Gamma(\kappa_{\text{drug}}, \xi_{\text{drug}}),$$

with hyperparameters $$\nu_{\text{feat}}^2, \nu_{\text{drug}}^2, \kappa_{\text{feat}}, \xi_{\text{feat}}, \kappa_{\text{drug}},$$ and $$\xi_{\text{drug}}.$$ Improper, flat priors occur if we let
\( \nu_{\text{feat}}, \nu_{\text{drug}} \to \infty \) and \( \kappa_{\text{feat}}, \kappa_{\text{drug}}, \xi_{\text{feat}}, \xi_{\text{drug}} \to 0 \). The corresponding full conditionals are given by:

\[
\alpha_{\text{feat}} | \lambda_{\text{feat}}, \Gamma \sim \mathcal{N}(m_{\text{feat}}, S_{\text{feat}})
\]

with \( S_{\text{feat}} = \lambda_{\text{feat}}^{-1}[C^T \text{diag}(\gamma_d^2) C + \nu_{\text{feat}}^{-1} \mathbf{I}_G]^{-1} \), and \( m_{\text{feat}} = [C^T \text{diag}(\gamma_d^2) C + \nu_{\text{feat}}^{-1} \mathbf{I}_G]^{-1} C^T 1_{pD \times 1} \),

\[
\alpha_{\text{drug}} | \lambda_{\text{drug}}, \tau \sim \mathcal{N}(m_{\text{drug}}, S_{\text{drug}}),
\]

with \( S_{\text{drug}} = \lambda_{\text{drug}}^{-1}[Z^T \text{diag}(\tau_d^2) Z + \nu_{\text{drug}}^{-2} \mathbf{I}_H]^{-1} \), and \( m_{\text{drug}} = [Z^T \text{diag}(\tau_d^2) Z + \nu_{\text{drug}}^{-2} \mathbf{I}_H]^{-1} Z^T 1_{D \times 1} \),

\[
\lambda_{\text{feat}} | \alpha_{\text{feat}}, \Gamma \sim \Gamma \left( \frac{pD}{2} + \kappa_{\text{feat}}, \nu_{\text{feat}} \right),
\]

with \( \nu_{\text{feat}} = \frac{1}{2} [C \alpha_{\text{feat}} - \text{diag}(\gamma_d^2) \mathbf{1}_{pD \times 1}]^T \text{diag}(\gamma_d^2) [C \alpha_{\text{feat}} - \text{diag}(\gamma_d^2) \mathbf{1}_{pD \times 1}] + \xi_{\text{feat}} \),

\[
\lambda_{\text{drug}} | \alpha_{\text{drug}}, \tau \sim \Gamma \left( \frac{D}{2} + \kappa_{\text{drug}}, \nu_{\text{drug}} \right),
\]

with \( \nu_{\text{drug}} = \frac{1}{2} [Z \alpha_{\text{drug}} - \text{diag}(\tau_d^2) \mathbf{1}_{D \times 1}]^T \text{diag}(\tau_d^2) [Z \alpha_{\text{drug}} - \text{diag}(\tau_d^2) \mathbf{1}_{D \times 1}] + \xi_{\text{drug}} \),

which may be sampled together with (4).

### 11 Simulations

The lasso and ridge models in this section were fit separately per drug, using the R package \texttt{glmnet} (Friedman et al., 2010) with cross validated penalty parameters.

Tables 1-3 display averaged estimation mean squared error (EMSE), and prediction mean squared error (PMSE), calculated on independent test data, for simulation Scenarios 1-3. PMSE is calculated as in Section 5 of the MD, while EMSE is calculated as: \( \text{EMSE} = D^{-1} p^{-1} \sum_{d=1}^{D} \sum_{j=1}^{p} (\beta_{jd} - \hat{\beta}_{jd})^2 \), with \( \hat{\beta}_{jd} \) the estimator for \( \beta_{jd} \). In the NIG models, that provide the full posteriors, the posterior mean \( \mathbb{E}(\beta_{jd} | y_d) \) is used as point estimate. EMSE is further split into the contribution of the bottom 10% of the \( \beta_{jd} \) in terms of size (in absolute value) and the top 10% of the \( \beta_{jd} \) in size.

Focussing on estimation, we see that the NIG, NIGd, and NIGf+d models (depending on the Scenario), that include the external covariates, outperform the other methods in terms of EMSE and PMSE. The lower EMSEs confirm that the NIG models that include the external data (NIGf, NIGd, and NIGf+d) learn the underlying structure in the data better than the models that do not include the external data (NIGf−, NIGd−, and NIGf−d). Furthermore, as expected, the sparser models lasso and NIG are better able to learn the \( \beta_{jd} \) with larger magnitude, as seen from the lower EMSEtop. On the other hand, ridge is better able to capture the smaller \( \beta_{jd} \), as evident from the lower EMSEbottom. Figure 1 shows that, as expected, the performance of the NIGf+d model deteriorates with increasing noise level in the external covariates and converges to the NIGf−d model performance.
Table 1: Mean EMSE (separated for bottom and top 10% of $|\hat{\beta}_{jd}|$) and PMSE for simulation Scenario 1, estimated on test data (lowest in bold).

|       | EMSE$_{\text{total}}$ | EMSE$_{\text{top}}$ | EMSE$_{\text{bottom}}$ | PMSE   |
|-------|------------------------|----------------------|------------------------|--------|
| NIG$^{-}_t$ | 0.349 (0.024)          | 2.135 (0.163)        | 0.001 (0)              | 0.033 (0.001) |
| NIG$^t$    | **0.348 (0.024)**      | **2.133 (0.163)**    | **0.001 (0)**          | **0.033 (0.001)** |
| ridge      | 0.371 (0.025)          | 2.284 (0.168)        | 0.002 (0)              | 0.045 (0.001)  |
| lasso      | 0.349 (0.024)          | 2.134 (0.163)        | 0.001 (0)              | 0.033 (0.001)  |

Table 2: Mean EMSE (separated for bottom and top 10% of $|\hat{\beta}_{jd}|$) and PMSE for simulation Scenario 2, estimated on test data (lowest in bold).

|       | EMSE$_{\text{total}}$ | EMSE$_{\text{top}}$ | EMSE$_{\text{bottom}}$ | PMSE   |
|-------|------------------------|----------------------|------------------------|--------|
| NIG$^{-}_d$ | 0.359 (0.037)          | 1.552 (0.161)        | **0.003 (0)**          | 0.053 (0.001) |
| NIG$^d$    | **0.359 (0.037)**      | 1.552 (0.161)        | **0.003 (0)**          | **0.053 (0.001)** |
| ridge      | 0.376 (0.038)          | 1.626 (0.164)        | 0.003 (0)              | 0.062 (0.002)  |
| lasso      | 0.359 (0.037)          | **1.548 (0.161)**    | 0.003 (0)              | 0.053 (0.001)  |

Table 3: Mean EMSE (separated for bottom and top 10% of $|\hat{\beta}_{jd}|$) and PMSE for simulation Scenario 3, estimated on test data (lowest in bold).

|       | EMSE$_{\text{total}}$ | EMSE$_{\text{top}}$ | EMSE$_{\text{bottom}}$ | PMSE   |
|-------|------------------------|----------------------|------------------------|--------|
| NIG$_{f+d}^{-}$ | 0.157 (0.025)          | 0.955 (0.159)        | 0.001 (0)              | 0.127 (0.004) |
| NIG$_{f+d}$    | **0.157 (0.025)**      | 0.953 (0.159)        | **0.001 (0)**          | **0.126 (0.004)** |
| ridge      | 0.167 (0.026)          | 1.025 (0.163)        | 0.001 (0)              | 0.134 (0.004)  |
| lasso      | 0.157 (0.025)          | **0.951 (0.159)**    | **0.001 (0)**          | **0.128 (0.004)** |

Figure 2 shows instances of both MCMC and VB $\beta_{jd}$ posteriors from simulation Scenario 3 for the NIG$_{f+d}$ model. The chosen $\beta_{jd}$ represent every combination of four drugs and four feature groups (for a total of 16 $\beta_{jd}$). The Figure shows that the MCMC and variational posteriors are in agreement. This leads us to conclude that the variational approximation to the posterior is quite accurate in this Scenario. The simulation Scenario was setup to mimic the real GDSC data, so that we expect an accurate variational posterior in the real GDSC data as well.

For completeness, Figure 3 displays samples from the full Bayesian NIG posterior, as introduced in Section 10. Unfortunately, we were not able to generate enough samples for a reliable posterior using the Gibbs procedure as described in the Section. We generated the samples from the full Bayesian posterior using stan through the R package rstan (Guo et al., 2018) (see https://github.com/magnusmunch/NIG for the implemented stan code). The Figure shows that the posteriors for $\phi$ and $\chi$ have unexpected shapes and seem to have explored only a small part of the posterior space. This indication that posterior inference is not reliable is confirmed by the average and minimum effective sample sizes of 66 and 2over all parameters, respectively. For reliable inference, Kruschke (2015) recommends an effective sample size of at least 10 000. Inspection of the computation times and a quick calculation learns that we have to sample from the posterior for at least 37 hours to obtain an average effective sample size of 10 000. Moreover, to achieve a
Figure 1: (a) Mean EMSE, separated for (b) bottom and (c) top 10% of $|\beta_{jd}|$, and (d) PMSE for simulation Scenario 4 versus the proportion of permuted rows in the external data, estimated on test data.

minimum effective sample size of 10 000 for all parameters, the sampler has to be run for about 7 weeks.
12 GDSC data

the CPO for cell line $i$ and drug $d$ is

$$CPO_{id} = p(y_{id} | y_{-id}) = \left[ \int_{\beta_d} \int_{\sigma_d^2} \mathcal{L}(y_{id} | \beta_d, \sigma_d^2)^{-1} q(\beta_d | y_d) q(\sigma_d^2 | y_d) d\beta_d d\sigma_d^2 \right]^{-1}$$  (5)
Figure 3: Simulation results for Scenario 3: Samples from full Bayesian NIG posteriors (see Section 10) with true value superimposed (red line) for (top row) group 1, (second row) group 2, (third row) group 3, and (bottom row) group 4 for both features and drugs, and (left column) $\phi$ and (second column) $\chi$. Note that the red line is not present in most of the Figures, indicating that the posterior is far of from the true value.

and is the Bayesian version of the leave-one-out cross-validated likelihood. Congdon (2005) suggests as outlier cut-off 0.01. Figure 4 displays these CPOs (calculated by numerical integration of (5)) for the four Analyses. For $y_{id}|y_{-id}$ following a Gaussian distribution with standard deviation 0.83 (the average posterior mean of $\sigma^2$ from our analyses) this implies an expected proportion of outliers of about 0.003. In our analyses we find 0.006, 0.006, 0.006, and 0.006, for the four analyses respectively. So about double the amount of expected outliers under a Gaussian posterior

11
predictive distribution. However, in reality, due to posterior uncertainty of $\beta_d$ and $\sigma_d^2$, the predictive distribution is more heavy-tailed than the Gaussian and the corresponding expected proportion of outliers higher than 0.003. Unfortunately, calculation of this true expected proportion is infeasible due to intractibility of the posterior predictive distribution.

Figure 4: distribution of the CPO values for Analysis (a) 1, (b) 2, (c) 3, and (d) 4 in Section 5 of the MD, with outlier cutoff value 0.01 (red line).

References

Bhattacharya, A. et al. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103, 985–991.
Congdon, P.D. (2005). *Bayesian models for categorical data*. Wiley series in probability and mathematical statistics 810651580. John Wiley, Chichester [etc.].

Friedman, J. et al. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Guo, J. et al. (2018). rstan: R Interface to Stan. URL https://CRAN.R-project.org/package=rstan

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186, 453–461.

Kruschke, J.K. (2015). *Doing Bayesian data analysis*. Elsevier, second edition edition.

Upadhyay, S.K. and Agrawal, R. (1996). Bayesian Analysis of Inverse Gaussian Non-Linear Regression by Simulation. *Sankhyā: The Indian Journal of Statistics, Series B*, 58, 363–378.

**Session info**

```r
devtools::session_info()
```

### - Session info

| setting  | value                                      |
|----------|--------------------------------------------|
| version  | R version 3.6.3 (2020-02-29)              |
| os       | macOS Catalina 10.15.3                    |
| system   | x86_64, darwin15.6.0                      |
| ui       | X11                                        |
| language | (EN)                                       |
| collate  | en_US.UTF-8                               |
| ctype    | en_US.UTF-8                               |
| tz       | Europe/Brussels                           |
| date     | 2020-04-30                                |

### - Packages

| package     | version | date      | lib     | source |
|-------------|---------|-----------|---------|--------|
| assertthat  | 0.2.1   | 2019-03-21| [1] CRAN| (R 3.6.0) |
| backports   | 1.1.6   | 2020-04-05| [1] CRAN| (R 3.6.2) |
| callr       | 3.4.3   | 2020-03-28| [1] CRAN| (R 3.6.2) |
| cli         | 2.0.2   | 2020-02-28| [1] CRAN| (R 3.6.0) |
| colorspace  | 1.4-1   | 2019-03-18| [1] CRAN| (R 3.6.0) |
| crayon      | 1.3.4   | 2017-09-16| [1] CRAN| (R 3.6.0) |
| desc        | 1.2.0   | 2018-05-01| [1] CRAN| (R 3.6.0) |
| devtools    | 2.3.0   | 2020-04-10| [1] CRAN| (R 3.6.3) |
| digest      | 0.6.25  | 2020-02-23| [1] CRAN| (R 3.6.0) |
| ellipsis    | 0.3.0   | 2019-09-20| [1] CRAN| (R 3.6.0) |
| evaluate    | 0.14    | 2019-05-28| [1] CRAN| (R 3.6.0) |
| fansi       | 0.4.1   | 2020-01-08| [1] CRAN| (R 3.6.0) |
| formatR     | 1.7     | 2019-06-11| [1] CRAN| (R 3.6.0) |
| fs          | 1.4.1   | 2020-04-04| [1] CRAN| (R 3.6.2) |
| glue        | 1.4.0   | 2020-04-03| [1] CRAN| (R 3.6.2) |
```
## hms 0.5.3 2020-01-08 [1] CRAN (R 3.6.0)
## htmltools 0.4.0 2019-10-04 [1] CRAN (R 3.6.0)
## httr 1.4.1 2019-08-05 [1] CRAN (R 3.6.0)
## kableExtra 1.1.0 2019-03-16 [1] CRAN (R 3.6.0)
## knitr * 1.28 2020-02-06 [1] CRAN (R 3.6.0)
## lattice 0.20-38 2018-11-04 [1] CRAN (R 3.6.3)
## lifecycle 0.2.0 2020-03-06 [1] CRAN (R 3.6.0)
## magrittr 1.5 2014-11-22 [1] CRAN (R 3.6.0)
## memoise 1.1.0 2017-04-21 [1] CRAN (R 3.6.0)
## munsell 0.5.0 2018-06-12 [1] CRAN (R 3.6.0)
## pillar 1.4.3 2019-12-20 [1] CRAN (R 3.6.0)
##pkgbuild 1.0.6 2019-10-09 [1] CRAN (R 3.6.0)
## pkgconfig 2.0.3 2019-09-22 [1] CRAN (R 3.6.0)
## pkgload 1.0.2 2018-10-29 [1] CRAN (R 3.6.0)
## prettyunits 1.1.1 2020-01-24 [1] CRAN (R 3.6.0)
## processx 3.4.2 2020-02-09 [1] CRAN (R 3.6.0)
## ps 1.3.2 2020-02-13 [1] CRAN (R 3.6.0)
## R6 2.4.1 2019-11-12 [1] CRAN (R 3.6.0)
## Rcpp 1.0.4.6 2020-04-09 [1] CRAN (R 3.6.3)
## readr 1.3.1 2018-12-21 [1] CRAN (R 3.6.0)
## remotes 2.1.1 2020-02-15 [1] CRAN (R 3.6.0)
## rlang 0.4.5 2019-12-20 [1] CRAN (R 3.6.0)
## markdown 2.1 2020-01-20 [1] CRAN (R 3.6.0)
## rmarkdown 1.3-2 2018-01-03 [1] CRAN (R 3.6.0)
## rstudioapi 0.11 2020-02-07 [1] CRAN (R 3.6.0)
## rvest 0.3.5 2019-11-08 [1] CRAN (R 3.6.0)
## scales 1.1.0 2019-11-18 [1] CRAN (R 3.6.0)
## sessioninfo 1.1.1 2018-11-05 [1] CRAN (R 3.6.0)
## sp * 1.4-1 2020-02-28 [1] CRAN (R 3.6.0)
## stringi 1.4.6 2020-02-17 [1] CRAN (R 3.6.0)
## stringr 1.4.0 2019-02-10 [1] CRAN (R 3.6.0)
## testthat 2.3.2 2020-03-02 [1] CRAN (R 3.6.0)
## tibble 3.0.0 2020-03-30 [1] CRAN (R 3.6.0)
## usethis 1.6.0 2020-04-09 [1] CRAN (R 3.6.0)
## vctrs 0.2.4 2020-03-10 [1] CRAN (R 3.6.0)
## viridisLite 0.3.0 2018-02-01 [1] CRAN (R 3.6.0)
## webshot 0.5.2 2019-11-22 [1] CRAN (R 3.6.0)
## withr 2.1.2 2018-03-15 [1] CRAN (R 3.6.0)
## xfun 0.13 2020-04-13 [1] CRAN (R 3.6.0)
## xml2 1.3.1 2020-04-09 [1] CRAN (R 3.6.0)

## [1] /Library/Frameworks/R.framework/Versions/3.6/Resources/library