

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

Magnus M. Münch^{1,2,3}  | Mark A. van de Wiel^{1,3} | Sylvia Richardson³ | Gwenaël G. R. Leday³

¹ Department of Epidemiology & Biostatistics, Amsterdam UMC, VU University, Amsterdam, The Netherlands

² Mathematical Institute, Leiden University, Leiden, The Netherlands

³ MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, United Kingdom

Correspondence

Magnus M. Münch, Department of Epidemiology & Biostatistics, Amsterdam UMC, VU University, PO Box 7057, 1007 MB Amsterdam, The Netherlands.
Email: m.munch@amsterdamumc.nl

Funding information

Amsterdam Public Health Institute's Methodology Program travel grant



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

In precision medicine, a common problem is drug sensitivity prediction from cancer tissue cell lines. These types of problems entail modelling multivariate drug responses on high-dimensional molecular feature sets in typically > 1000 cell lines. The dimensions of the problem require specialised models and estimation methods. In addition, external information on both the drugs and the features is often available. We propose to model the drug responses through a linear regression with shrinkage enforced through a normal inverse Gaussian prior. We let the prior depend on the external information, and estimate the model and external information dependence in an empirical-variational Bayes framework. We demonstrate the usefulness of this model in both a simulated setting and in the publicly available Genomics of Drug Sensitivity in Cancer data.

KEYWORDS

drug sensitivity, empirical Bayes, Genomics of Drug Sensitivity in Cancer (GDSC), variational Bayes

1 | INTRODUCTION

Recently, promising results in precision medicine have sparked an interest in cancer drug sensitivity prediction models (Iorio et al., 2016). Typically, these models predict the drug sensitivity for new patients from a set of molecular features. Development of such models is often done in well-characterised human cancer tissue cell lines. The current paper presents a novel drug sensitivity prediction model and an application to a real drug sensitivity data set.

Development of such models from cell lines has proven to be difficult (see, e.g. the DREAM 7 challenge in Costello et al. (2014)). Difficulties arise, among others, from the dimensions of the problem. Typically, the data contain hundreds

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Biometrical Journal* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

of drugs, thousands of cell lines and thousands of molecular features. An example of a large database of drug responses and molecular features is the Genomics of Drug Sensitivity in Cancer (GDSC) data (Yang et al., 2013), which we will further investigate in Section 5. Other examples of such databases include the Cancer Cell Line Encyclopedia (CCLE) (Li et al., 2019) and the US National Cancer Institute 60 human tumour cell line anticancer drug screen (NCI60) (Shoemaker, 2006). The dimensions of these data prohibit the estimation of standard regression models and typically require some form of regularisation.

The GDSC database contains additional information on both the drugs and molecular features, such as the target pathways and developmental stages of the drugs. Additional online repositories may provide extra information such as the molecular weight of the compounds or the publication signatures of the molecular features. In some cases, prior knowledge on the drug efficacies may be available, from previous experiments. We propose to include these possibly beneficial information sources in the estimation of the sensitivity prediction models in a data-driven manner. More specifically, we estimate a normal inverse Gaussian (NIG) model, where the extent of regularisation is estimated by an adaptive empirical Bayes procedure, guided by the external information.

We are not the first to work on drug sensitivity prediction models. Reviews on the topic are Azuaje (2017) and Ali and Aittokallio (2019). Zhao and Zucknick (2019) and Mai, Rønneberg, Zhao, Zucknick, and Corander (2019) consider a structured penalized multivariate regression approach. Aben, Vis, Michaut, and Wessels (2016) introduce a two-stage penalized regression model that includes two different types of molecular features. Ammad-ud din et al. (2016) and Costello et al. (2014) tackle the problem through a multiple kernel learning approach. Our solution allows for the adaptive incorporation of the external information on drugs and features. This is done by pooling information, both across drugs and features. Estimation of the model is through computationally feasible variational Bayes approximations, while empirical Bayes estimation of tuning parameters pools information across drugs and features in a data-driven manner.

The rest of the paper is structured as follows. In Section 2, we introduce our model, the estimation of which is detailed in Section 3. Section 4 describes a simulation study that investigates the estimation of hyperparameters by the proposed method. In Section 5, we analyse the GDSC data, and we end with a discussion in Section 6 on the pros and cons of the proposed method.

2 | MODEL

2.1 | Simultaneous equations model

Let y_{id} be the continuous sensitivity measures for cell lines $i = 1, \dots, n$, and drug $d = 1, \dots, D$. We predict sensitivity from molecular features x_{ij} , $j = 1, \dots, p$, collected in $\mathbf{x}_i = [x_{i1} \dots x_{ip}]^T$. We assume that both covariates and responses have been centred per drug and regress the drug sensitivities on the molecular features:

$$y_{id} = \mathbf{x}_i^T \boldsymbol{\beta}_d + \epsilon_{id}, \text{ with } \epsilon_{id} \sim \mathcal{N}(0, \sigma_d^2), \quad (1)$$

where the p -dimensional $\boldsymbol{\beta}_d = [\beta_{1d} \dots \beta_{pd}]^T$ are the drug-specific omics feature effects. Note that (1) gives rise to a system of D linear regression equations.

The cell lines used in drug response models are often taken from different tissues. In addition, other clinical covariates might be available. To obtain unbiased feature effects, one may wish to account for these. We do so by introducing unpenalized covariates, the β_{jd} coefficients of which are endowed with a flat prior. For the sake of clarity, in the following, such unpenalized covariates are omitted. However, the available software allows for their inclusion.

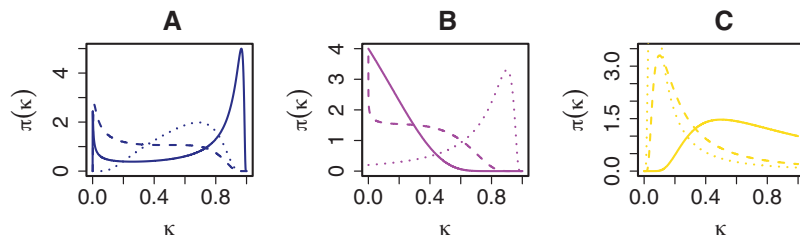
2.2 | Bayesian prior model

We carry out inference by endowing the parameters with the following priors:

$$\beta_{jd} | \gamma_{jd}^2, \tau_d^2, \sigma_d^2 \sim \mathcal{N}_p(0, \gamma_{jd}^2 \tau_d^2 \sigma_d^2), \quad (2a)$$

$$\gamma_{jd}^2 \sim \text{IG}(\phi_{jd}, \lambda_{\text{feat}}), \quad (2b)$$

FIGURE 1 Implied prior densities $\pi(\kappa_{jd})$ for the (A) NIG, (B) Student's t , and (C) lasso priors. Different line types correspond to different hyperparameter settings. The hyperparameter settings (given in Section 3 of the SM) were chosen to show some possible, distinct shapes that each of the priors can take



$$\tau_d^2 \sim IG(\chi_d, \lambda_{\text{drug}}), \quad (2c)$$

$$\sigma_d^2 \sim 1/\sigma_d^3, \quad (2d)$$

where $IG(\phi, \lambda)$ denotes an inverse Gaussian distribution with mean ϕ and shape $\lambda > 0$.

In model (2), γ_{jd}^2 in (2b) denotes a local variance component that is supposed to capture local, feature-specific variation in the model parameters β_{jd} in (2a), while the global variance components τ_d^2 in (2c) capture the drug-specific, general trend in β_d . Each drug response is endowed with a random error variance σ_d^2 , distributed according to (2d).

Prior distributions of the form (2) are often referred to as global-local shrinkage rules (Polson and Scott, 2011), due to the multiplicative separation of the prior variance into a local component γ_{jd}^2 and a global component τ_d^2 . For appropriate local shrinkage in global–local shrinkage models it is important to account for different noise levels σ_d^2 by scaling the β_{jd} variances accordingly.

The NIG prior model was introduced in Barndorff-Nielsen (1978) and since Barndorff-Nielsen (1997), it is routinely applied in mathematical finance (see, e.g. Kalemnova, Schmid, and Werner (2007)). Here, we extend it with an additional global variance component τ_d^2 . Supplementary material (SM) Section S2 contains more details on the NIG prior. To illustrate the effect of the NIG prior on the posterior mean, we consider the prior reparametrised as in Carvalho, Polson, and Scott (2009), that is, in terms of shrinkage weights $\kappa_{jd} = 1/(1 + \gamma_{jd}^2) \in (0, 1)$. Under the (simplified) normal means model, that is, $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]^T = \mathbf{I}_p$, with fixed $\tau_d^2 = \sigma_d^2 = 1$, the resulting conditional posterior mean for the β_{jd} is $\mathbb{E}(\beta_{jd}|y_{jd}, \kappa_{jd}) = (1 - \kappa_{jd})y_{jd}$. Thus, $\kappa_{jd} = 0$ implies no shrinkage of β_{jd} and $\kappa_{jd} = 1$ implies full shrinkage towards zero. Figure 1 depicts the prior on κ_{jd} implied by several choices of β_{jd} prior.

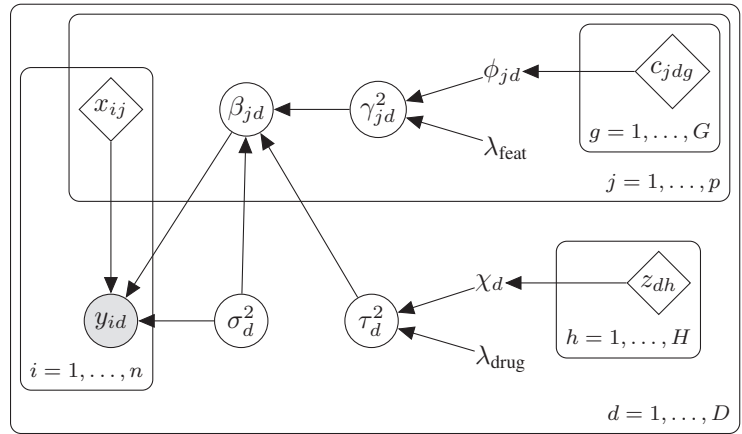
Figure 1 shows that, depending on the choice of hyperparameters, the NIG prior can behave similarly to the Student's t prior (decreasing form zero, with substantial mass close to zero and little mass close to one, like the solid lines in Figs. 1A,B), but also rather differently (dashed and dotted lines in Figs. 1A,B). Our argumentation to model the γ_{jd}^2 by an inverse Gaussian distribution, as has been suggested in Fabrizi, Greco, and Trivisano (2016) and Caron and Doucet (2008), is three-fold: (i) the NIG model is more flexible than the lasso prior (as seen from Fig. 1), (ii) the NIG prior allows to model the means of the $\gamma_{jd}^2(\phi_{jd})$ and $\tau_d^2(\chi_d)$ as a function of external data more conveniently than the Student's t prior, as explained in Section 3.2 and (iii) like the horseshoe (Carvalho et al., 2009), the NIG shrinkage weights prior can put mass both near zero and one, a desirable property of shrinkage priors (Polson and Scott, 2011).

A few remarks on the choice of error variance prior are justified here: many authors endow error variance components with vague gamma priors. Gelman (2006), among others, advises against this practice. The degree of ‘vagueness’ has a large influence on the posterior, while degree of ‘vagueness’ is a difficult parameter to set. This influence is especially pronounced if the likelihood is relatively flat, as may be reasonably expected in the large p , small n setting. We therefore model the error variance with Jeffreys objective prior (Jeffreys, 1946) that does not depend on any subjective specification of hyperparameters. In the derivation of our Jeffreys prior for the error variance, we jointly consider an unknown data mean and variance (Kass and Wasserman, 1996). This joint consideration results in the somewhat unorthodox $1/\sigma^3$ Jeffreys prior.

2.3 | External information

In drug sensitivity prediction models, external information on both the drugs and features is often available. Here, we assume this information to be available as external feature ‘covariates’ \mathbf{c}_{jdg} , for $g = 1, \dots, G$, and drug ‘covariates’ \mathbf{z}_{dh} , for $h = 1, \dots, H$. An example of a (binary) feature covariate is target pathway presence, with $c_{jdg} = 0$ if gene j is present in the target pathway of drug d and $c_{jdg} = 1$ if it is not. An example of a (ternary) drug covariate is developmental phase, with levels experimental phase, clinical development and approved by a governing agency.

FIGURE 2 Hierarchical representation of the drug sensitivity prediction model. Grey circles represent observed variables, white circles represent unobserved variables, tilted squares represent fixed data, and unenclosed letters are parameters to be estimated. Cell lines are indexed by i , features by j , drugs by d , drug covariates by h , and feature covariates by g . The y_{id} are the drug sensitivities, x_{ij} the molecular features, c_{jdg} the external feature covariates, z_{dh} the external drug covariates, β_{jd} the regression coefficients, σ_d^2 the error variances, τ_d^2 and γ_{jd}^2 the drug and feature specific variance components, respectively, and ϕ_{jd} , λ_{feat} , χ_d , and λ_{drug} the hyperparameters



The external covariates come in through our mean models for the γ_{jd}^2 and τ_d^2 hyperpriors: $\phi_{jd} = (\mathbf{c}_{jd}^T \boldsymbol{\alpha}_{\text{feat}})^{-1}$ and $\chi_d = (\mathbf{z}_d^T \boldsymbol{\alpha}_{\text{drug}})^{-1}$, with $\mathbf{c}_{jd} = [c_{jd1} \cdots c_{jdG}]$ and $\mathbf{z}_d = [z_{d1} \cdots z_{dH}]$, where categorical external covariates are dummy coded. The model now requires hyperparameters $\boldsymbol{\alpha}_{\text{feat}}$, λ_{feat} , $\boldsymbol{\alpha}_{\text{drug}}$ and λ_{drug} , which we estimate in a data-driven manner (see Section 3.2).

A representation of our model as a Bayesian DAG is given in Figure 2. We note that in many settings, the set of features might be different for different drugs. In that case the covariates are indexed by the drug d : \mathbf{X}^d , a trivial extension of model (1) and (2). This extension is included in the available software, but for clarity it is omitted in the following.

3 | ESTIMATION

3.1 | Variational Bayes

The posterior corresponding to the model described in (1) and (2) is not available in closed form. To avoid computationally intensive Markov chain Monte Carlo (MCMC) algorithms, we approximate the joint posterior by variational Bayes (see Blei, Kucukelbir, & McAuliffe, 2017, for a review), where the approximate posterior density factorises as: $p(\boldsymbol{\beta}_d, \boldsymbol{\gamma}_d^2, \tau_d^2, \sigma_d^2 | \mathbf{y}_d) \approx Q_d(\cdot) = q(\boldsymbol{\beta}_d) \cdot q(\boldsymbol{\gamma}_d^2) \cdot q(\tau_d^2) \cdot q(\sigma_d^2)$, where $\boldsymbol{\gamma}_d^2 = [\gamma_{1d}^2 \cdots \gamma_{pd}^2]^T$. For notational convenience, we slightly abuse notation and let $q(\cdot)$ denote different densities for different inputs. Under such a factorisation, the marginal variational posteriors that minimise the Kullback–Leibler divergence of the true posterior to the variational Bayes approximation (Neal and Hinton, 1998) are given by:

$$\begin{aligned}
 q(\boldsymbol{\beta}_d) &\stackrel{D}{=} \mathcal{N}_p(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d), \\
 q(\boldsymbol{\gamma}_d^2) &\stackrel{D}{=} \prod_{j=1}^p \mathcal{GIG}(-1, \lambda_{\text{feat}} / \phi_{jd}^2, \delta_{jd}), \\
 q(\tau_d^2) &\stackrel{D}{=} \mathcal{GIG}\left(-\frac{p+1}{2}, \lambda_{\text{drug}} / \chi_d^2, \eta_d\right), \\
 q(\sigma_d^2) &\stackrel{D}{=} \Gamma^{-1}\left(\frac{n+p+1}{2}, \zeta_d\right),
 \end{aligned}$$

where $\mathcal{GIG}(p, \nu, \eta)$ denotes the generalized inverse Gaussian distribution with index $p \in \mathbb{R}$, and scales $\nu > 0$ and $\eta > 0$ (Jørgensen, 1982). See SM Section S5 for the derivations. The variational parameters $\boldsymbol{\mu}_d$, $\boldsymbol{\Sigma}_d$, δ_{jd} , η_d and ζ_d contain cyclic dependencies and are iteratively updated by:

$$\boldsymbol{\Sigma}_d^{(h+1)} = \left(\mathbf{a}_d^{(h)}\right)^{-1} \left[\mathbf{X}^T \mathbf{X} + \mathbf{g}_d^{(h)} \text{diag}\left(\mathbf{b}_{jd}^{(h)}\right)\right]^{-1}, \quad (3a)$$

$$\boldsymbol{\mu}_d^{(h+1)} = \left[\mathbf{X}^T \mathbf{X} + \mathbf{g}_d^{(h)} \text{diag}\left(\mathbf{b}_{jd}^{(h)}\right)\right]^{-1} \mathbf{X}^T \mathbf{y}_d, \quad (3b)$$

$$\delta_{jd}^{(h+1)} = a_d^{(h)} g_d^{(h)} \left[\left(\boldsymbol{\mu}_{jd}^{(h+1)} \right)^2 + \left(\boldsymbol{\Sigma}_d^{(h+1)} \right)_{jj} \right] + \lambda_{\text{feat}}, \quad (3c)$$

$$\eta_d^{(h+1)} = a_d^{(h)} \sum_{j=1}^p b_{jd}^{(h+1)} \left[\left(\boldsymbol{\mu}_{jd}^{(h+1)} \right)^2 + \left(\boldsymbol{\Sigma}_d^{(h+1)} \right)_{jj} \right] + \lambda_{\text{drug}}, \quad (3d)$$

$$\zeta_d^{(h+1)} = \frac{1}{2} \left[\mathbf{y}_d^T \mathbf{y}_d - 2 \mathbf{y}_d^T \mathbf{X} \boldsymbol{\mu}_d^{(h+1)} + \text{tr} \left(\mathbf{X}^T \mathbf{X} \boldsymbol{\Sigma}_d^{(h+1)} \right) + \left(\boldsymbol{\mu}_d^{(h+1)} \right)^T \mathbf{X}^T \mathbf{X} \boldsymbol{\mu}_d^{(h+1)} \right. \quad (3e)$$

$$\left. + g_d^{(h+1)} \text{tr} \left[\text{diag} \left(b_{jd}^{(h+1)} \right) \boldsymbol{\Sigma}_d^{(h+1)} \right] + g_d^{(h+1)} \left(\boldsymbol{\mu}_d^{(h+1)} \right)^T \text{diag} \left(b_{jd}^{(h+1)} \right) \boldsymbol{\mu}_d^{(h+1)} \right], \quad (3f)$$

until convergence, where $\mathbf{y}_d = [y_{1d} \cdots y_{nd}]^T$. Here, we set

$$a_d^{(h)} = \mathbb{E}_{Q^{(h)}}(\sigma_d^{-2}) = (n + p + 1) / \left(2 \zeta_d^{(h)} \right),$$

$$b_{jd}^{(h)} = \mathbb{E}_{Q^{(h)}}(\gamma_{jd}^{-2}) = \sqrt{\frac{\lambda_{\text{feat}}}{\phi_{jd}^2 \delta_{jd}^{(h)}}} \frac{K_0 \left(\sqrt{\delta_{jd}^{(h)} \lambda_{\text{feat}} / \phi_{jd}^2} \right)}{K_1 \left(\sqrt{\delta_{jd}^{(h)} \lambda_{\text{feat}} / \phi_{jd}^2} \right)} + \frac{2}{\delta_{jd}^{(h)}}, \quad (4)$$

$$g_d^{(h)} = \mathbb{E}_{Q^{(h)}}(\tau_d^{-2}) = \sqrt{\frac{\lambda_{\text{drug}}}{\chi_d^2 \eta_d^{(h)}}} \frac{K_{(p-1)/2} \left(\sqrt{\eta_d^{(h)} \lambda_{\text{drug}} / \chi_d^2} \right)}{K_{(p+1)/2} \left(\sqrt{\eta_d^{(h)} \lambda_{\text{drug}} / \chi_d^2} \right)} + \frac{p + 1}{\eta_d^{(h)}},$$

where $K_\nu(x)$ denotes the modified Bessel function of the second kind. A method for fast and numerically stable calculation of ratios of modified Bessel functions of the second kind, as in (4), is given in SM Section S8.

3.2 | Empirical Bayes

We parametrised the prior mean of the γ_{jd}^2 as $\phi_{jd} = (\mathbf{c}_{jd}^T \boldsymbol{\alpha}_{\text{feat}})^{-1}$ and the prior mean of τ_d^2 as $\chi_d = (\mathbf{z}_d^T \boldsymbol{\alpha}_{\text{drug}})^{-1}$. This parametrisation allows us to include feature and drug covariates, both continuous and discrete, into the model. Additionally, it reduces the number of hyperparameters from pD to $|\boldsymbol{\alpha}_{\text{feat}}| + |\boldsymbol{\alpha}_{\text{drug}}| + 2$. The Bayesian model then requires the specification of the hyperparameters $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_{\text{feat}}^T \boldsymbol{\alpha}_{\text{drug}}^T \right]^T$ and $\boldsymbol{\lambda} = [\lambda_{\text{feat}} \lambda_{\text{drug}}]^T$. These are abstract and hard to interpret parameters for which we generally lack expert knowledge. They do, however, have a significant influence on the shape of the posterior distribution. We therefore propose to estimate these hyperparameters by empirical Bayes. In our case, this results in an objective and data-driven inclusion of the external feature and drug covariates.

The canonical method for empirical Bayes is to maximise the marginal likelihood with respect to the hyperparameters. In Casella (2001), the marginal likelihood is maximised by an EM algorithm:

$$\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\lambda}^{(l+1)} = \underset{\boldsymbol{\alpha}, \boldsymbol{\lambda} > 0}{\text{argmax}} \mathbb{E}_{\cdot | \mathbf{Y}} [\log p(\mathbf{Y}, \mathbf{B}, \boldsymbol{\Gamma}^2, \boldsymbol{\tau}^2, \boldsymbol{\Sigma}^2) | \boldsymbol{\alpha}^{(l)}, \boldsymbol{\lambda}^{(l)}]$$

$$= \underset{\boldsymbol{\alpha}, \boldsymbol{\lambda} > 0}{\text{argmax}} \mathbb{E}_{\cdot | \mathbf{Y}} \left[\log \pi(\boldsymbol{\Gamma}^2) | \boldsymbol{\alpha}_{\text{feat}}^{(l)}, \boldsymbol{\lambda}_{\text{feat}}^{(l)} \right] + \mathbb{E}_{\cdot | \mathbf{Y}} \left[\log \pi(\boldsymbol{\tau}^2) | \boldsymbol{\alpha}_{\text{drug}}^{(l)}, \boldsymbol{\lambda}_{\text{drug}}^{(l)} \right],$$

where $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_D]$, $\mathbf{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_D]$, $\boldsymbol{\tau}^2 = [\tau_1^2 \cdots \tau_D^2]^T$, $\boldsymbol{\Sigma}^2 = [\sigma_1^2 \cdots \sigma_D^2]^T$ and $\boldsymbol{\Gamma}^2 = [\boldsymbol{\Gamma}_1^2 \cdots \boldsymbol{\Gamma}_D^2]$, and the expectation is with respect to the joint posterior. In our case, this posterior is not available in closed form, which renders the expectation difficult. While Casella (2001) suggests to approximate the expectation by a Monte Carlo sample, we propose to use the variational Bayes approximation developed in Section 3.1:

$$\boldsymbol{\alpha}^{(l+1)}, \boldsymbol{\lambda}^{(l+1)} = \underset{\boldsymbol{\alpha}, \boldsymbol{\lambda} > 0}{\text{argmax}} \mathbb{E}_{Q^{(l)}} \left[\log \pi(\boldsymbol{\Gamma}^2) | \boldsymbol{\alpha}_{\text{feat}}^{(l)}, \boldsymbol{\lambda}_{\text{feat}}^{(l)} \right] + \mathbb{E}_{Q^{(l)}} \left[\log \pi(\boldsymbol{\tau}^2) | \boldsymbol{\alpha}_{\text{drug}}^{(l)}, \boldsymbol{\lambda}_{\text{drug}}^{(l)} \right],$$

where now the expectation is with respect to the converged variational posterior $Q^{(l)} = \prod_{d=1}^D Q_d^{(l)}$. Note that the prior Γ_d^2 and τ_d^2 independence assumption results in separate optimisation problems for the feature hyperparameters (α_{feat} and λ_{feat}), and the drug hyperparameters (α_{drug} and λ_{drug}). If we stack the drug and feature covariates:

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{11}^T \\ \vdots \\ \mathbf{c}_{p1}^T \\ \vdots \\ \mathbf{c}_{1D}^T \\ \vdots \\ \mathbf{c}_{pD}^T \end{bmatrix} \text{ and } \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_D^T \end{bmatrix},$$

the empirical Bayes updates are given by:

$$\begin{aligned} \alpha_{\text{feat}}^{(l+1)} &= \left[\mathbf{C}^T \text{diag} \left(e_{jd}^{(l)} \right) \mathbf{C} \right]^{-1} \mathbf{C}^T \mathbf{1}_{pD \times 1}, \\ \lambda_{\text{feat}}^{(l+1)} &= pD \left[\sum_{d=1}^D \sum_{j=1}^p b_{jd}^{(l)} + \left(\alpha_{\text{feat}}^{(l+1)} \right)^T \mathbf{C}^T \text{diag} \left(e_{jd}^{(l)} \right) \mathbf{C} \alpha_{\text{feat}}^{(l+1)} - 2 \left(\alpha_{\text{feat}}^{(l+1)} \right)^T \mathbf{C}^T \mathbf{1}_{pD \times 1} \right]^{-1}, \\ \alpha_{\text{drug}}^{(l+1)} &= \left[\mathbf{Z}^T \text{diag} \left(f_d^{(l)} \right) \mathbf{Z} \right]^{-1} \mathbf{Z}^T \mathbf{1}_{D \times 1}, \\ \lambda_{\text{drug}}^{(l+1)} &= D \left[\sum_{d=1}^D g_d^{(l)} + \left(\alpha_{\text{drug}}^{(l+1)} \right)^T \mathbf{Z}^T \text{diag} \left(f_d^{(l)} \right) \mathbf{Z} \alpha_{\text{drug}}^{(l+1)} - 2 \left(\alpha_{\text{drug}}^{(l+1)} \right)^T \mathbf{Z}^T \mathbf{1}_{D \times 1} \right]^{-1}, \end{aligned}$$

where SM Section S9 shows that $\lambda^{(l+1)} > 0$ and

$$\begin{aligned} e_{jd}^{(l)} &= \mathbb{E}_{Q^{(l)}} \left(\gamma_{jd}^2 | \alpha_{\text{feat}}^{(l)}, \lambda_{\text{feat}}^{(l)} \right) = \left(b_{jd}^{(l)} - 2/\delta_{jd}^{(l)} \right) \cdot \delta_{jd}^{(l)} \left(\phi_{jd}^{(l)} \right)^2 / \lambda_{\text{feat}}^{(l)}, \\ f_d^{(l)} &= \mathbb{E}_{Q^{(l)}} \left(\tau_d^2 | \alpha_{\text{drug}}^{(l)}, \lambda_{\text{drug}}^{(l)} \right) = \left(g_d^{(l)} - (p+1)/\eta_d^{(l)} \right) \cdot \eta_d^{(l)} \left(\chi_d^{(l)} \right)^2 / \lambda_{\text{drug}}^{(l)}. \end{aligned}$$

To ensure proper and unbiased shrinkage, intercepts are included in α_{feat} and α_{drug} . This is achieved by appending both \mathbf{C} and \mathbf{Z} with a column of ones. These intercepts are roughly interpreted as the expected prior precisions $\mathbb{E}(\gamma_{jd}^{-2})$ and $\mathbb{E}(\tau_d^{-2})$ if the feature and drug covariates are all zero. Likewise, an α corresponding to an external covariate may be interpreted as an additive effect of the external covariate on the prior expected precision. So an $\alpha = 1$ translates to an increase in expected prior precision of 1 for every increase in the external covariate of 1, keeping all the other external covariates fixed.

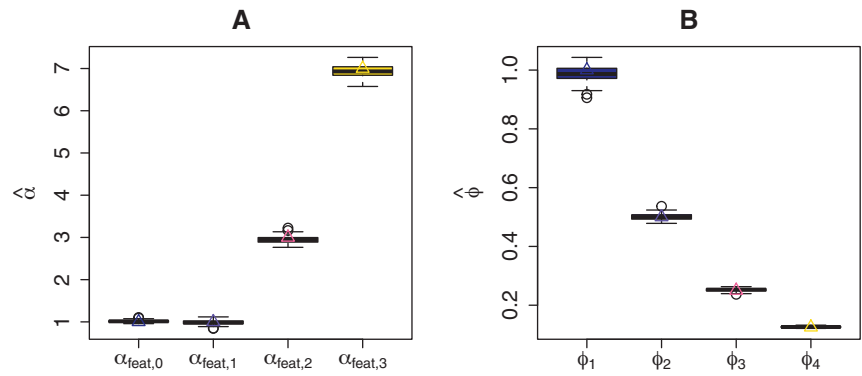
Variational Bayes approximations are known to underestimate posterior variances (Rue, Martino, and Chopin, 2009; Consonni and Marin, 2007; Bishop, 2006; Wang and Titterington, 2005). In simulation Scenario 5 in Section S11 of the SM, we compare the variational posterior to MCMC samples from the posterior with fixed hyperparameters estimates (after the procedure described in Section 3.2 has converged). In this simulation scenario and other settings (not shown), the variational approximation to the posterior is accurate. If however, the user is reluctant to trust the variational posterior variances, samples from the posterior may be generated with the Gibbs sampler in SM Section S10. Alternatively, we provide an implementation of the proposed model in stan using the R package rstan (Guo et al., 2018) at <https://github.com/magnusmunch/NIG>.

4 | SIMULATIONS

4.1 | Setup

This section investigates the empirical Bayes estimation properties of the model in a simulated setting; its main aim is to assess hyperparameter estimation. It is a data-based simulation, wherein the responses are simulated from a synthetic

FIGURE 3 Simulation results for Scenario 1 (τ_d^2 fixed): estimated and true values for (A) α_{feat} and (B) prior means ϕ_{jd}



model, but the features are taken from the real GDSC expression data introduced in Section 5. The real GDSC features contain strong collinearities. Such strong collinearities in the design matrix impede correct parameter estimation with small sample sizes. We therefore replace the ambition of correctly estimating the β_{jd} with the more modest aim of approximately correct estimation of the hyperparameters.

A pre-processing step selects 100 features with largest variance, while 251 drug sensitivities for 507 cell lines (half of the total number of cell lines) are simulated from models (1) and (2). We draw the error variances as $\forall d : \sigma_d^2 \sim \Gamma^{-1}(3, 2)$, such that the prior σ_d^2 mean and variance are both one. We consider the following four scenarios for the simulation of the drug and feature variance components:

- Scenario 1 fixes $\forall d : \tau_d^2 = 1$ and draws the γ_{jd}^2 according to model (2). We create four external dummy feature covariates that code for four approximately equally sized groups of features. We set α_{feat} such that the γ_{jd}^2 of the four groups of features have prior means $\phi_{jd} \in \{1, 1/2, 1/4, 1/8\}$ ($\alpha_{\text{feat}} = [1 \ 1 \ 3 \ 7]^T$). The prior scale parameter is set to $\lambda_{\text{feat}} = 1$.
- Scenario 2 fixes $\forall j, d : \gamma_{jd}^2 = 1$ and draws the τ_d^2 according to model (2), following a procedure similar to the procedure for the γ_{jd}^2 in Scenario 1: we create four groups of drugs with corresponding external drug dummy variables and set $\alpha_{\text{drug}} = [1 \ 1 \ 3 \ 7]^T$, such that we have $\chi_d \in \{1, 1/2, 1/4, 1/8\}$. The scale is set to $\lambda_{\text{drug}} = 1$.
- Scenario 3 combines the procedures from Scenarios 1 and 2 to draw both the γ_{jd}^2 and τ_d^2 according to (2).
- Scenario 4 is equal to Scenario 3, except that we add noise to the external covariates. Noise is supposed to mimic a low external covariate signal and is constructed by permutation of fractions $q \in \{0.1, 0.2, 0.33, 0.5, 0.67, 0.8, 1\}$ of the rows of the external covariates.

We estimate two models: (i) the NIG model that only includes an intercept in the external covariates, called NIG_f^- , NIG_d^- or NIG_{f+d}^- , depending on which variance components are estimated (feature, drug or both in Scenarios 1, 2 and 3/4, respectively), and (ii) the NIG model estimated as in Section 3 that includes all external covariates, called NIG_f , NIG_d or NIG_{f+d} , again depending on which variance components are estimated. Exclusion of the external covariates as in the NIG_f^- , NIG_d^- and NIG_{f+d}^- models amounts to direct estimation of common expected prior means ϕ and/or χ , instead of regression estimates for the ϕ_{jd} and/or χ_d as in the NIG_f , NIG_d and NIG_{f+d} models. In the language of Polson and Scott (2011) as introduced in Section 1, models NIG_f and NIG_d may be described as local and global shrinkage rules, respectively, as opposed to the global–local shrinkage models NIG_{f+d} and NIG_{f+d}^- . We repeat every simulation Scenario 100 times.

SM Section S11 contains more simulation results for Scenarios 1–4 for the NIG model and the (i) frequentist lasso and (ii) ridge models. Additionally, SM Section S11 contains a comparison of MCMC and VB posteriors in simulation Scenario 3.

4.2 | Results

Figure 3 shows the estimated α_{feat} together with its true value for NIG_f in Scenario 1 of the simulation study (fixed τ_d^2). Figure 3A shows that estimation of α_{feat} is accurate. This results in accurate estimates on the ϕ_{jd} scale as well (Fig. 3B). Model NIG_f^- (that excludes the external covariates) gives a mean ϕ estimate of 0.457 (0.007) (standard deviation between parentheses), about equal to the true mean of the ϕ_{jd} , 0.469. Scale $\lambda_{\text{feat}} = 1$ is overestimated by NIG_f at 1.321 (0.08), while NIG_f^- underestimates at 0.404 (0.013).

FIGURE 4 Simulation results for Scenario 2 (γ_{jd}^2 fixed): estimated (boxplots) and true values (triangles) for (A) α_{feat} and (B) prior means ϕ_{jd}

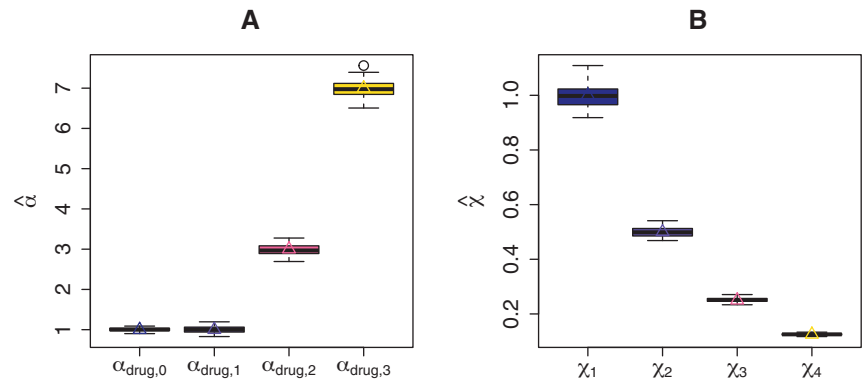


FIGURE 5 Simulation results for Scenario 3: estimated (boxplots) and true values (triangles) for (A) α_{feat} , (B) prior means ϕ_{jd} , (C) α_{drug} , and (D) prior means χ_d

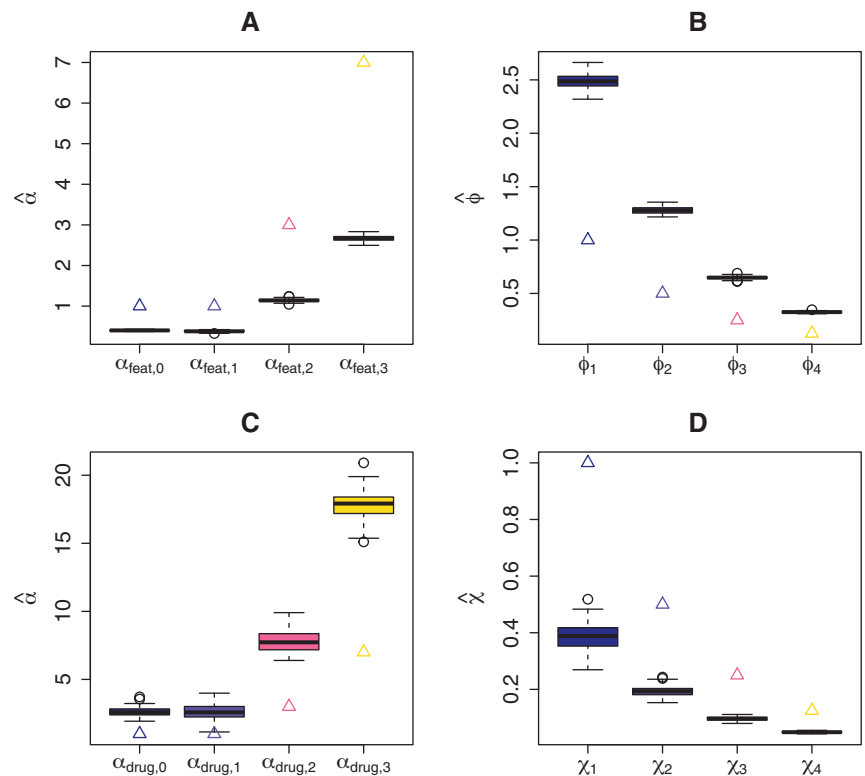


Figure 4A shows the accurately estimated α_{drug} together with the true value for NIG_d in Scenario 2 of the simulation study (fixed γ_{jd}^2). Likewise, the estimates are accurate on the χ_d scale (Fig. 4B). The mean χ estimate in the NIG_d⁻ model is 0.471 (0.012), which is about equal to the true mean 0.469. Scale $\lambda_{\text{drug}} = 1$ is overestimated by NIG_d at 12.944 (2.325) and underestimated by NIG_d⁻ at 0.592 (0.023).

In Figure 5 the α_{feat} and α_{drug} estimated by NIG_{f+d} are displayed together with their true values for simulation Scenario 3. α_{feat} are underestimated (Fig. 5A), while α_{drug} (Fig. 5C) are overestimated, resulting in overestimated ϕ_{jd} (Fig. 5B) and underestimated χ_d (Fig. 5D), respectively. The biases seem to be consistent though. The mean ratios ϕ_1 to ϕ_2 , ϕ_3 , ϕ_4 are 0.514 (0.017), 0.26 (0.009), 0.131 (0.005), while the mean ratios χ_1 to χ_2 , χ_3 , χ_4 are 0.507 (0.078), 0.253 (0.036), 0.128 (0.017). In both cases the true values are 0.5, 0.25, 0.125, so in a relative sense, the α_{feat} and α_{drug} estimates are about correct. Moreover, overestimation of the ϕ_{jd} is compensated for by the underestimation of the χ_d : the estimated mean prior variances $\mathbb{V}(\beta_{jd}) = \phi_{jd} \cdot \chi_d$ (ignoring the error variance) are unbiased (Fig. 6). The NIG_{f+d}⁻ model is also consistently over- and underestimating ϕ and χ with mean estimates 0.971 (0.018) and 0.207 (0.013), respectively (compared to the true mean 0.469). Again, on the $\mathbb{V}(\beta_{jd})$ level, this bias almost vanishes; the mean estimated $\mathbb{V}(\beta_{jd})$ (ignoring error variance) are 0.201 (0.014) while their true mean is 0.22. In Scenario 3, NIG_{f+d} overestimates $\lambda_{\text{feat}} = 1$ at 7.098 (0.621) and underestimates $\lambda_{\text{drug}} = 1$ at 0.437 (0.047). Similar results hold for NIG_{f+d}⁻ with λ_{feat} estimate 1.5 (0.066) and λ_{drug} estimate 0.195 (0.014).

FIGURE 6 Simulation results for Scenario 3: mean estimated prior variances $\hat{V}(\beta_{jd})$ versus true values, with line of identity (dotted)

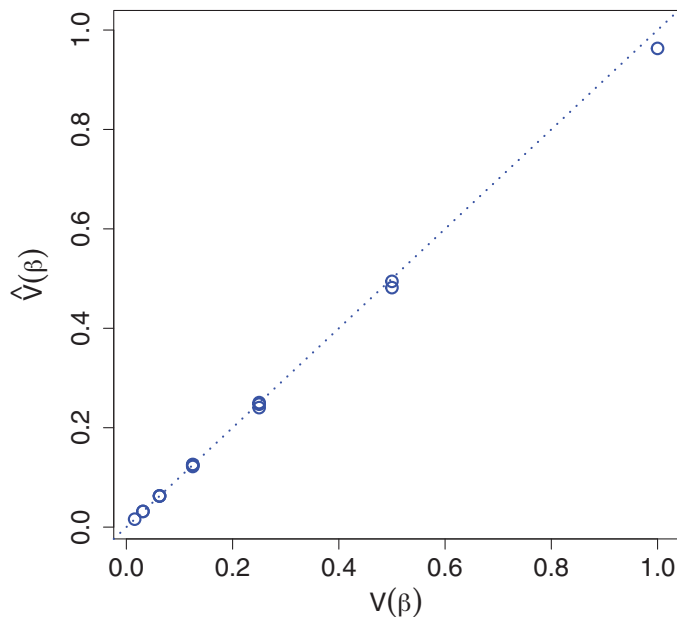


FIGURE 7 Simulation results for Scenario 3: mean estimated prior means (A) $\hat{\phi}_{jd}$, and (B) $\hat{\chi}_d$ for different levels of noise in the external covariates

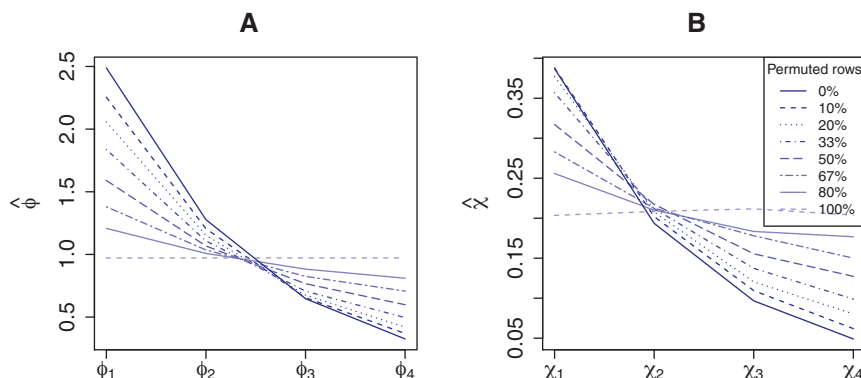


Figure 7 displays the mean $\hat{\phi}_{jd}$ and $\hat{\chi}_d$ estimates for different noise levels in simulation Scenario 4. The simulation shows that with increasing noise level, the estimated prior means $\hat{\phi}_{jd}$ and $\hat{\chi}_d$ for the four groups of external covariates become more and more alike. In other words, noise in the external covariates impedes estimation of α_{feat} and α_{drug} , as expected.

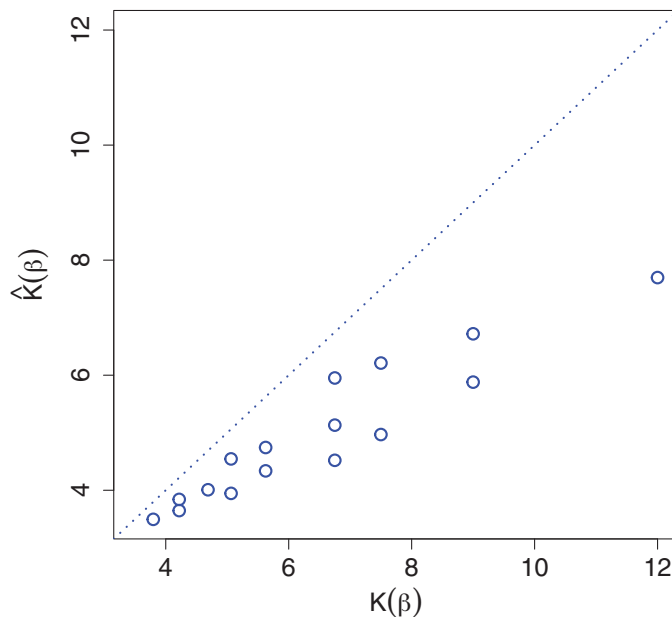
To summarise, estimation of only α_{feat} or α_{drug} (and consequently $\hat{\phi}_{jd}$ and $\hat{\chi}_d$) by NIG_f and NIG_d , respectively is relatively unbiased, as evident from simulation Scenarios 1 and 2 (Figs. 3 and 4). In contrast, simultaneous estimation in Scenario 3 results in overestimated $\hat{\phi}_{jd}$ and underestimated $\hat{\chi}_d$ by NIG_{f+d} (Fig. 5). We conjecture that this interplay of drug and feature variance components is due to near-unidentifiability. In any case, the consequences are limited, since the variances on the β_{jd} level are left unbiased (Fig. 6). If separately estimated, scale parameters λ_{feat} and λ_{drug} are overestimated by NIG_f and NIG_d , respectively. If estimated simultaneously NIG_{f+d} overestimates λ_{feat} and underestimated λ_{drug} . On the β_{jd} level, λ influences kurtoses $\mathcal{K}(\beta_{jd})$. Figure 8 shows the mean estimated kurtoses versus their true values. Kurtoses seem to be underestimated by NIG_{f+d} . In contrast, NIG_{f+d}^{-1} overestimates the true mean of the $\mathcal{K}(\beta_{jd})$, 6.4716797, at 10.234 (3.696). Lastly, the results from Scenario 3 in SM Section S11 show that the VB approximation is quite good as compared to standard MCMC.

5 | GDSC DATA

5.1 | Primary data

The GDSC project's (Yang et al., 2013) aim is 'to improve cancer treatments by discovering therapeutic biomarkers that can be used to identify patients most likely to respond to anticancer drug'. Part of the project is to screen > 1000 human cancer

FIGURE 8 Simulation results for Scenario 3: mean estimated prior Kurtoses $\hat{\mathcal{K}}(\beta_{jd})$ versus true values, with line of identity (dotted)



cell lines for drug sensitivities. The cell lines have been genetically characterised and several drug sensitivity measures are recorded. The data is freely available from Garnett et al. (2012) and consist of: (i) the sensitivity measures of the cell lines to the drugs, (ii) annotation of the screened compounds and (iii) the cell lines' genomic profile (mutations, copy numbers, methylation profiles and gene expression). We will attempt to predict drug sensitivities of the cell lines, as quantified by half maximal inhibitory concentration (IC50), using the gene expression and gene mutation data. Other choices of sensitivity measures than IC50 are possible, but a discussion on the pros and cons of different sensitivity measures is beyond the aim of this paper. We have used the version of the data that is presented in Iorio et al. (2016). We averaged repeated measures over cell line-drug combinations and model the logarithm of the IC50 values. In the following, IC50 refers to these log-transformed values. After removing all cell lines with missing values, we end up with 388 to 1043 IC50 estimates for 251 drugs. Differences in the number of cell lines between drugs occur, because not all drug and cell line combinations are available. The pre-processed expression and mutation data consist of 17,737 and 300 genes, respectively.

5.2 | External data

Two ternary drug covariates are available: the developmental stage (experimental, in clinical development or clinically approved) of the drugs and the action (unknown, cytotoxic or targeted) of the drugs. These drug covariates are taken directly from the GDSC database's annotation file and dummy coded with reference categories clinically approved drugs and cytotoxic drugs. We expect that drugs that have been clinically approved are easiest to predict and hence yield the largest prior β_{jd} variances, followed by the drugs in clinical development, and the experimental drugs. Likewise, we expect the targeted drugs to yield the largest prior β_{jd} variances, followed by the cytotoxic drugs, and the unknown target drugs. Note that large β_{jd} variances translate to large prior γ_{jd}^2 and τ_d^2 means.

Furthermore, we have a binary feature covariate available that indicates whether a gene belongs to the drug target pathway. The feature covariate was created by comparing the target pathways in the GDSC annotation to the KEGG (Kanehisa and Goto, 2000) and reactome (Fabregat et al., 2018) repositories. The reference category here is features that are not in the target pathway. For this external covariate, we expect that genes that are in the pathway of the drug are more predictive than genes that are not, that is, they have larger prior β_{jd} variances than drugs that are not in the pathway.

The type of molecular marker may be included as external covariate, that is, whether the feature is a gene expression or gene mutation. As an alternative to direct inclusion of the mutation data, we use p -values from the mutations as external covariate. These were obtained from a t -test comparing IC50 values of mutated and unmutated genes. We expect that lower mutation p -values result in a larger prior β_{jd} variances.

Lastly, p -values from an analysis of the CCLE data (Li et al., 2019), a database similar to the GDSC, are included as external covariate. These p -values are obtained from a simple correlation between the IC50 values and the gene expressions

TABLE 1 α estimates from analysis 1

	Feature intercept	p -value	Drug intercept
NIG_{f+d}^-	2.616		280.408
NIG_{f+d}	2.551	0.179	282.068

Note. Empty cells correspond to fixed zero parameters.

TABLE 2 α estimates from analysis 2

	Feature intercept	Mutation	Drug intercept
NIG_{f+d}^-	3.025		171.623
NIG_{f+d}	3.884	-1.781	153.217

Note. Empty cells correspond to fixed zero parameters.

from the CCLE data. The harmonic mean per gene is then used as external covariate for the GDSC data analysis. Again, we expect a positive relation between these external p -values and the larger prior β_{jd} variances.

5.3 | Analyses

Four analyses were conducted:

- Analysis 1 includes gene expressions as predictors and p -values from the gene mutation data as external covariate ($G = 1$, $H = 0$). A pre-processing step selects between 221 and 280 genes per drug for which both the expression as well as a mutation p -value is available.
- Analysis 2 includes both gene expressions and mutations as predictors with the feature type as external covariate, that is, whether the feature is an expression or mutation ($G = 1$, $H = 0$). For this analysis, we pre-select 300 gene expressions with maximum variance and 295 gene mutations for which there are both mutated and wild-type cell lines available.
- Analysis 3 uses 500 gene expressions as features, selected based on maximum variance. The CCLE p -values are included as external covariates ($G = 1$, $H = 0$).
- Analysis 4 includes the gene expressions as features and both the annotated drug variables and pathway status of the genes as external covariates ($G = 1$, $H = 2$, before dummy coding). We pre-select 500 gene expressions based on maximum variance.

In all analyses we estimated the same models as in the simulations (Section 4 and SM Section S11): (i) NIG_{f+d}^- , (ii) NIG_{f+d} and (iii) frequentist lasso and (iv) ridge models.

In all analyses we use all cell lines to estimate the hyperparameters presented in Section 5.4. Mean prediction mean squared errors (PMSE) and its standard error are estimated by 10-fold cross validation, where $\text{PMSE} = D^{-1}n^{-1} \sum_{d=1}^D \sum_{i=1}^n (\mathbf{y}_d - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_d)^2$, with $\hat{\boldsymbol{\beta}}_d$ the estimator for $\boldsymbol{\beta}_d$. In the NIG model, that provides full posteriors, the posterior mean $\mathbb{E}(\boldsymbol{\beta}_d | \mathbf{y}_d)$ is used as point estimate.

5.4 | Results

The non-zero NIG_{f+d} α estimates in Tables 1–5 show that there is an effect of the external covariates.

- Analysis 1 results in a positive additive effect of the mutation p -values on the prior $\boldsymbol{\beta}_d$ precisions (Table 1). This translates to more $\boldsymbol{\beta}_d$ shrinkage towards zero with increasing p -value, as expected.
- Analysis 2 shows that gene mutations are more predictive than gene expressions, as observed from the negative effect of mutation dummy on prior precisions (Table 2): mutations are shrunken less than expressions.
- Analysis 3 indicates that CCLE p -values are positively related to prior precision (Table 3), that is, higher CCLE p -values results in more shrinkage of $\boldsymbol{\beta}_d$, as expected.

TABLE 3 α estimates from analysis 3

	Feature intercept	p-value	Drug intercept
NIG_{f+d}^-	2.84		330.614
NIG_{f+d}	2.55	0.658	337.086

Note. Empty cells correspond to fixed zero parameters.

TABLE 4 α estimates from analysis 4

	Feature intercept	Pathway
NIG_{f+d}^-	2.841	
NIG_{f+d}	2.845	-0.261

Note. Empty cells correspond to fixed zero parameters.

- Analysis 4 gives a negative effect for the pathway dummy (Table 4), indicating less shrinkage for genes that are in the drugs' target pathway, as expected. According to expectation, experimental and developmental drugs prior precisions are shrunken more than the reference category, approved drugs (Table 5). Somewhat surprisingly, targeted and unknown drugs are shrunken more than the reference, cytotoxic drugs (Table 5), although the differences on the β_d variance scale are small: $\hat{E}(\tau_d^2) = \{0.0035, 0.0034, 0.0031\}$ for cytotoxic, targeted, and unknown drugs, respectively (ignoring other variance components).

The mean PMSE, calculated on the test data, are displayed in Table 6. Note that due to standardisation, an empty reference model has a PMSE of one. In terms of PMSE, ridge outperforms the other models in all Analyses 1, 3 and 4, while NIG_{f+d} α performs best in Analysis 2. In general however, all models perform very similarly. The ranges of mean PMSE over all methods are 0.0146, 0.0179, 0.0163 and 0.0163, for Analyses 1–4, respectively, indicating that the differences in predictive performance are very small. Furthermore, the difference between NIG_{f+d} , that includes external covariates, and NIG_{f+d}^- , that excludes the external covariates is small; an indication that the external covariates are not very informative here.

Part of the differences in performance between NIG and ridge may be explained with the different levels of sparsity in the solution. Although NIG does not automatically select features, as opposed to the lasso, its β_d prior has larger kurtosis than the ridge prior (see SM Section S2). The resulting heavy-tailedness as compared to the ridge prior facilitates features selection. Selection of features from a sparse prior may be achieved through the decoupling shrinkage and selection approach (DSS) introduced in Hahn and Carvalho (2015). We have applied DSS in our analyses, where we select either (approximately) the same number of features as lasso (about 50 in all analyses), 25 or 100 features. Table 7 compares the resulting mean PMSE values. The table shows that NIG_{f+d} plus DSS outperforms lasso, for all three numbers of selected features, in all analyses.

To assess model fit, Section S12 in the SM displays the conditional predictive ordinates (CPO) for the NIG_{f+d} model for the four analyses. A visual inspection of the CPOs learns that no extreme outliers occur.

TABLE 5 α estimates from analysis 4

	Drug intercept	Experimental	Development	Targeted	Unknown
NIG_{f+d}^-	331.403				
NIG_{f+d}	289.290	43.985	53.864	2.845	34.964

Note. Empty cells correspond to fixed zero parameters (continued).

TABLE 6 Mean (standard deviation) of cross-validated PMSE for GDSC data

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
NIG_{f+d}^-	0.796 (0.006)	0.805 (0.007)	0.783 (0.005)	0.785 (0.006)
NIG_{f+d}	0.796 (0.006)	0.803 (0.007)	0.783 (0.005)	0.785 (0.006)
ridge	0.795 (0.005)	0.807 (0.005)	0.782 (0.005)	0.783 (0.005)
lasso	0.809 (0.005)	0.821 (0.005)	0.798 (0.005)	0.799 (0.006)

Note. Best performing model (per analysis) in bold.

TABLE 7 Mean (standard deviation) of cross-validated PMSE for selection methods (number of selected features between parentheses) on GDSC data

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
lasso (25)	0.828 (0.004)	0.823 (0.005)	0.812 (0.005)	0.812 (0.005)
lasso _{λ} ^{cv} (~50) ^a	0.809 (0.005)	0.821 (0.005)	0.798 (0.005)	0.799 (0.006)
lasso (100)	0.826 (0.008)	0.841 (0.007)	0.809 (0.006)	0.81 (0.007)
NIG _{f+d} +DSS (25)	0.826 (0.004)	0.814 (0.005)	0.809 (0.005)	0.811 (0.005)
NIG _{f+d} +DSS (~50) ^a	0.806 (0.004)	0.808 (0.006)	0.792 (0.005)	0.794 (0.006)
NIG _{f+d} +DSS (100)	0.798 (0.006)	0.799 (0.006)	0.786 (0.005)	0.787 (0.006)

Note. Best performing model (per analysis) in bold.

^aLasso with cross validated λ selects, on average, 42–56 features in all analyses (indicated with ~50).

6 | DISCUSSION

The preceding presents a novel model for drug sensitivity prediction from a set of high-dimensional molecular features. The model allows for the inclusion of discrete and continuous external covariates on both the drugs and features. Inclusion of the external information is through data-driven and adaptive empirical Bayes estimation of the hyperparameters in the NIG prior model (2). Variational Bayes estimation is efficient and scales well with the number of features and samples. Estimation of NIG_{f+d} in the GDSC data analyses in Section 5 took 20, 24, 17 and 125 min on a 2016 MacBook Pro with 2 GHz Dual-Core Intel Core i5 processor and 8 GB of memory, running macOS 10.15.1.

Simulation Scenarios 1 and 2 in Section 4 show that estimation of drug- and feature-specific hyperparameters is, in principle, fairly accurate. However, when estimated jointly, biases may occur due to the interplay between the two sources of information. Fortunately, these biases cancel out on the β_{jd} level, such that the prior variance estimates $\mathbb{V}(\beta_{jd})$ are accurate. Ultimately, predictive performance benefits from the inclusion of external covariates, according to the results in Section 11 of the SM.

The model is put into practice on the GDSC data in Section 5. The comparison of NIG_{f+d} to NIG_{f+d}⁻ (that excludes the external covariates) shows that although the inclusion of external covariates substantially modifies the hyperparameters, predictive performance as measured by PMSE is only slightly better in one of four analyses. The NIG model is competitive with conventional, penalized methods like lasso and ridge, but all three methods achieve PMSE of only 0.80 in all analyses, a 20% reduction compared to the empty model. In three of the four analyses, ridge slightly outperforms NIG, which in turn outperforms lasso. In Analysis 2, NIG slightly outperforms ridge and lasso. The indications of the above are two-fold: (i) the GDSC data do not contain a lot of signal overall and (ii) the external covariates are not very informative for the GDSC data. We note however, that NIG seems to have a small advantage in terms of feature selection. If we follow the DSS approach in Hahn and Carvalho (2015) for feature selection, PMSE after selection is slightly better than lasso in all four analyses.

The penalized regression methods estimate penalty parameters by cross-validation. Cross-validation directly minimises the (approximate) PMSE, as opposed to empirical Bayes in the NIG that maximises the (approximate) marginal likelihood, a measure of model fit. To achieve maximal predictive accuracy, direct prediction error optimisation by cross-validation is preferred. However, direct prediction error optimisation by cross-validation is not feasible in the external covariates setting, due to the large number of hyperparameters. A caveat with penalized regression methods is that they do not give measures of parameter uncertainty. NIG, on the other hand, gives the full posterior of the parameters, either through a variational Bayes approximation or with the Gibbs sampler from SM Section S10. The full posterior gives direct access to the parameter uncertainties for a better interpretable model. In addition, given that the linear predictor is a linear combination of β_{jd} , and we have access to the approximate multivariate posterior of β_d , NIG also allows to assess uncertainty of the predictions.

An alternative strategy to include external covariates is the varying coefficient (VC) model (Hastie and Tibshirani, 1993). The VC model treats the mean of the regression coefficients as a deterministic function of external covariates, as opposed to our probabilistic model for the variance of the regression coefficients. Ni, Stingo, Ha, Akbani, and Baladandayuthapani (2019) introduces a Bayesian VC model where the relation between the regression coefficients and external covariates is no longer deterministic, but still based on the mean of the coefficients. Besides our computationally more feasible VB-EM estimation, we advocate for a more indirect model for the relation between external covariates and regression coefficients,

that is, through their random variance components. This indirect model assumes less structure about the relation between external covariates and regression coefficients than the direct VC mean model approach. In particular, a VC mean model describes both magnitude and direction of the external covariates effects, while our variance model only describes the magnitude and is invariant to the direction of the effects. Nonetheless, a combination of the two approaches (both mean and variance modelled as functions of the external covariates) may be a fruitful future research direction. Other methods that consider an external covariate model for the variance components of the regression coefficients are bSEM (Leday et al., 2017; Kpogbezan, Vaart, Wieringen, Leday, and van de Wiel, 2017; and xtune (Zeng, Thomas, and Lewinger, 2020). Here, bSEM is designed to include only dichotomous external covariates, so is not applicable in most of the applications and simulations that we have considered here. Inclusion of multiple external features is allowed by xtune, but not on the drug level, so xtune has limited applicability in our simulations and applications.

A possible criticism of the NIG model is the treatment of α as fixed hyperparameters instead of random. A Bayesian could argue that endowment of α with a hyperprior results in propagation of uncertainty about α and as a result improved regression parameter uncertainty quantification. Van de Wiel, te Beest, and Münch (2019) show in a similar setting that EB estimation of hyperparameters does not necessarily lead to worse uncertainty quantification as measured by frequentist coverage of Bayesian credible intervals, as compared to a full Bayes treatment of the hyperparameters.

Possible directions of future research are applications of the NIG model to different data types. Suitable applications are eQTL studies, in which gene expressions are regressed on SNPs. Several interesting external covariates are available, both on the genes as well as the SNPs. An example of such an external covariate for the genes is gene length, where we suspect that longer genes are harder to predict. The distance of the SNP to the gene is an example of an external SNP covariate, where the expectation is that SNPs further from the gene are less predictive of that gene's expression.


ACKNOWLEDGEMENTS

We thank the referees for their helpful comments and suggestions on an earlier version of this manuscript. MM visited SR and GL on a travel grant funded by the Amsterdam Public Health institute's Methodology program.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Magnus M. Münch  <https://orcid.org/0000-0001-9659-2044>

REFERENCES

- Aben, N., Vis, D. J., Michaut, M., & Wessels, L. F. (2016). TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17), i413–i420.
- Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews*, 11(1), 31–39.
- Ammad-ud din, M., Khan, S. A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T., & Kaski, S. (2016). Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, 32(17), i455–i463.
- Azuaje, F. (2017). Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics*, 18(5), 820–829.
- Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics*, 5(3), 151–157.
- Barndorff-Nielsen, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24(1), 1–13.
- Bishop, C. M., (2006). *Pattern recognition and machine learning*. New York: Springer.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

- Caron, F., & Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 88–95), ICML '08. New York, NY, USA: ACM.
- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *AISTATS* (Vol. 5, pp. 73–80).
- Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4), 485–500.
- Consonni, G., & Marin, J.-M. (2007). Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis*, 52(2), 790–798.
- Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., ... Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12), 1202–1212.
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., ... D'Eustachio, P. (2018). The reactome pathway knowledge base. *Nucleic Acids Research*, 46(D1), D649–D655.
- Fabrizi, E., Greco, F., & Trivisano, C. (2016). On the specification of prior distributions for variance components in disease mapping models. *Statistica*, 76(1), 93–111.
- Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., ... Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391), 570–575.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), 515–534.
- Guo, J., Gabry, J., Goodrich, B., Lee, D., Sakrejda, K., Martin, M., ... Oehlschlaegel-Akiyoshi (R/pairs.R), J. (2018). rstan: R Interface to Stan.
- Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509), 435–448.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4), 757–796.
- Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., ... Garnett, M. J. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3), 740–754.
- 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(1007), 453–461.
- Jørgensen, B., (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Lecture notes in statistics ; 9. 830800018, New York, Heidelberg, Berlin: Springer.
- Kalemanova, A., Schmid, B., & Werner, R. (2007). The normal inverse Gaussian distribution for synthetic CDO pricing. *The Journal of Derivatives*, 14(3), 80–94.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435), 1343–1370.
- Kpogbezan, G. B., van der Vaart, A. W., van Wieringen, W. N., Leday, G. G. R., & van de Wiel, M. A. (2017). An empirical Bayes approach to network recovery using external knowledge. *Biometrical Journal*, 59(5), 932–947.
- Leday, G. G. R., Gunst, M. C. M. d., Kpogbezan, G. B., van der Vaart, A. W., van Wieringen, W. N., & van de Wiel, M. A. (2017). Gene network reconstruction using global-local shrinkage priors. *The Annals of Applied Statistics*, 11(1), 41–68.
- Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., ... Sellers, W. R. (2019). The landscape of cancer cell line metabolism. *Nature Medicine*, 25(5), 850–860.
- Mai, T. T., Rønneberg, L., Zhao, Z., Zucknick, M., & Corander, J. (2019). Composite local low-rank structure in learning drug sensitivity. *arXiv:1905.00095 [stat]*. arXiv: 1905.00095.
- Neal, R. M., & Hinton, G. E. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 355–368), number 89 in NATO ASI Series. Springer Netherlands.
- Ni, Y., Stingo, F. C., Ha, M. J., Akbani, R., & Baladandayuthapani, V. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, 114(525), 48–60.
- Polson, N. G., & Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 9* (pp. 501–538). Oxford University Press.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392.
- Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10), 813–823.
- van de Wiel, M. A., Te Beest, D. E., & Münch, M. M. (2019). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 46(1), 2–25.
- Wang, B., & Titterton, D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS* (pp. 373–380).
- Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., ... Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1), D955–D961.
- Zeng, C., Thomas, D. C., & Lewinger, J. P. (2020). Incorporating prior knowledge into regularized regression. preprint, Bioinformatics.
- Zhao, Z., & Zucknick, M. (2019). Structured penalized regression for drug sensitivity prediction. *arXiv:1902.04996 [stat]*. arXiv: 1902.04996.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Münch MM, van de Wiel MA, Richardson S, Leday GGR. Drug sensitivity prediction with normal inverse Gaussian shrinkage informed by external data. *Biometrical Journal*. 2021;63:289–304. <https://doi.org/10.1002/bimj.201900371>