Sparse Bayesian inference on gamma-distributed observations

羽村靖之[1]，鬼塚貴広[2]，橋本真太郎[2]，菅澤翔之助[3]
[1] 京都大学大学院経済学研究科
[2] 広島大学大学院先進理工系科学研究科
[3] 東京大学空間情報科学研究センター

1　はじめに

機械が故障するまでの時間やある病気の生存時間などの正値データは応用上重要である．特に，高次元の正値データに対しては平均の多くがある値 (grand mean) の周辺に集中しており，少数の重要なシグナルについては平均が大きな値をとるような (スパースな) 構造をもつことが多い．この場合に，データから適切なシグナルを検出する問題を考える．このようなスパース信号解析は，実数値や計数値のデータに対しては多くの研究があるが，正値連続データに対してはほとんど研究が行われていない．スパースな正値連続データに対する例として，Donoho and Jin (2006) ではスパースな指数分布の列モデルに対して，FDR に基づく閾値に基づく推定法が提案され，理論的な性質が研究されているが，実データの分析例などは示されていない．

一方で，ベイズ統計学の文脈ではデータに合わせた柔軟な縮小とともに不確実性の定量化を容易に行うことができる global-local shrinkage と呼ばれるテクニックがここ 10 年の間に発展してきた．実際，実数値データに対応するガウス列モデル (Carvalho et al., 2010) やカウントデータに対応するポアソン列モデル (Datta and Dunson, 2016) のスパース推定は理論・方法論ともに研究が進められている．

本研究では，正値連続データをモデル化するための代表的なモデルであるガンマ分布を扱い，新しいクラスの global-local shrinkage prior を逆ガンマ分布の形状尺度混合によって構成し，その事前分布の性質を導出する．また，それらが大きいシグナルをできるだけ縮小しない性質 (tail-robustness) とスパース性に対するカルバック・ライブラー超有効性をもつことも示す．これらの理論的性質は，ガウス列モデルやポアソン列モデルにおいても導出されているが，これらは尺度混合型の事前分布を用いており，本研究における形状尺度混合型の事前分布の場合の証明はそれよりも複雑になる．提案する事前分布に基づく事後分布の計算に関して，効率的なマルコフ連鎖モンテカルロ (MCMC) アルゴリズムを構成し，数値実験により提案手法のパフォーマンスを確認する．さらに，応用例として韓国の COVID-19 の平均入院期間と遺伝子発現データに対する分散の縮小推定についてのデータ分析を行う．

2　ガンマ列モデルに対する新たなクラスの縮小事前分布

2.1　定式化

次のような独立なガンマ列モデルを考える．

$$y_i \mid \lambda_i \sim \mathrm{Ga}\left(\delta_i, \frac{\delta_i}{\lambda_i}\right), \quad i = 1, \ldots, n.$$

ただし，$\mathrm{Ga}(\alpha, \beta)$ は，形状母数 α，尺度母数 $1/\beta$ のガンマ分布を表し，δ_i は定数であり問題により適切に設定する (5 節を参照)．また，$\mathbb{E}[y_i \mid \lambda_i] = \lambda_i$ は興味のある平均パラメータである．$\lambda = (\lambda_1, \ldots, \lambda_n)$ のベイズ推定を考えたいので，λ_i に対する事前分布が必要である．上記のガンマ分布の尺度母数に対する共役事前分布は逆ガンマ分布であるが，ここでは逆ガンマ分布に対して次のような定式化を行う．

$$\lambda_i \mid u_i \sim \mathrm{IG}(1 + \tau u_i, \beta\tau u_i), \quad i = 1, \ldots, n. \tag{1}$$

ここで，τ と u_i は以下で見るように λ_i の事後平均を縮小する役割を果たすパラメータであり，τ は大域的な縮小を行うためのパラメータ (global shrinkage parameter)，u_i は局所的な縮小を行うためのパラメータ (local shrinkage parameter) である．λ_i の事前期待値は，$\mathbb{E}[\lambda_i] = \beta$ であり，β は縮小のターゲットである grand mean を表す．このとき，λ_i の事後期待値は (1) のようなやや変わった定式化をおこなったおかげで

$$\mathbb{E}[\lambda_i \mid y] = \mathbb{E}\left[\frac{\delta_i y_i + \beta\tau u_i}{\delta_i + \tau u_i} \;\middle|\; y_i\right] = y_i - \mathbb{E}[\kappa_i \mid y_i](y_i - \beta)$$

のようにシンプルな形で書き下すことができる．この表現は縮小の意味を解釈するのに便利で，$\kappa_i = \tau u_i/(\delta_i + \tau u_i) \in (0, 1)$ は縮小係数と呼ばれる量になっていて，

$$\mathbb{E}[\lambda_i \mid y] \to \begin{cases} y_i & \text{if } \mathbb{E}[\kappa_i \mid y_i] \to 0, \\ \beta & \text{if } \mathbb{E}[\kappa_i \mid y_i] \to 1 \end{cases} \tag{2}$$

であることから，κ_i の事後平均 $\mathbb{E}[\kappa_i \mid y_i]$ は grand mean β へ縮小するかしないかを左右する役割を果たす．なお，y_i はいま考えているモデルにおける最尤推定値に対応する．

2.2　Global-local shrinkage priors

先に述べたような，大域的な縮小に加えて局所的な縮小も考慮する縮小推定はベイズ統計において global-local shrinkage prior と呼ばれる事前分布を用いることにより実現される．近年発展してきているこの方法は，Carvalho et al. (2010) によりガウス列モデルに対して提案され，Datta and Dunson (2016) によりカウントデータを扱うポアソン列モデルへと拡

張された．なお，古典的な縮小推定方法である Stein 推定やスパース推定でよく用いられる Lasso などは共に大域的な縮小を考えていることに注意する．

さて，良い global-local shrinkage prior を構成するためには u_i の事前分布が重要である．ここでいう「良い」とは

- (2) で与えられるような縮小に関する性質をもち，
- 対応する事後分布の計算を容易に行うことができる

ことをいう．ここでは，ガンマ分布に対する新しいクラスの global-local shrinkage prior として，逆ガンマ分布の形状尺度混合 (shape-scale inverse-gamma mixtures) による新たなクラスの事前分布を提案する．考える階層モデルをまとめると次のようになる：

$$y_i \mid \lambda_i \sim \mathrm{Ga}\left(\delta_i, \frac{\delta_i}{\lambda_i \delta_i}\right), \quad \lambda_i \mid u_i \sim \mathrm{IG}(1+\tau u_i, \beta\tau u_i), \quad u_i \sim \pi(\cdot). \tag{3}$$

実用上は，τ や β にも条件付き共役なガンマ事前分布を仮定し，これらのパラメータもデータから推定する．なお，各 i に対して $u_i = 1$ とした場合は global shrinkage prior と呼び，基本的な共役事前分布に基づくベイズ推定に対応する．u_i に対する事前分布として，scaled-beta (SB) prior と inverse rescaled beta (IRB) prior

$$\pi_{\mathrm{SB}}(u_i) = \frac{1}{B(a,b)} \frac{u_i^{a-1}}{(1+u_i)^{a+b}}, \quad \pi_{\mathrm{IRB}}(u_i) = \frac{1}{B(a,b)} \frac{1}{u_i(1+u_i)} \frac{\log(1+1/u_i)^{b-1}}{\{1+\log(1+1/u_i)\}}.$$

を考える．ただし，$B(a,b)$ はベータ関数であり，ハイパーパラメータ $a, b > 0$ の選択は重要である．これらの事前分布を用いた場合の λ_i の周辺事前分布は

$$p(\lambda_i) = \int_0^\infty \mathrm{IG}(\lambda_i \mid 1+\tau u_i, \beta\tau u_i)\pi(u_i)du_i$$

となり，$p(\lambda_i)$ の裾や grand mean における性質は以下で与えられる．

命題 1. $\beta = \tau = 1$ とする．モデル (3) のもとで，$\pi(u_i)$ として $\pi_{\mathrm{SB}}(u_i)$ または $\pi_{\mathrm{IRB}}(u_i)$ を考えるとき，λ_i の周辺分布 $p(\lambda_i)$ は以下の性質を満たす．

(i) $\lambda_i \to 0$ のとき，$\pi_{\mathrm{SB}}(u_i)$ のもとで $p(\lambda_i) \approx \lambda_i^{a-1}$ となり，$\pi_{\mathrm{IRB}}(u_i)$ のもとで $p(\lambda_i) \approx \lambda_i^{-1}$ となる．

(ii) $\lambda_i \to \infty$ のとき，(π_{SB}, π_{IRB} に対し) $p(\lambda_i) \approx \lambda_i^{-2}$ となる．

(iii) $\lambda_i \to 1$ のとき，(π_{SB}, π_{IRB} に対し) $b \le 1/2$ なら $p(\lambda_i) \to \infty$ となり，$b > 1/2$ なら $p(\lambda_i) \to C_1$ となる．ただし，$0 < C_1 < \infty$ は定数とする．

周辺分布は図 1 の通りである．$b \le 1/2$ とすることで，$\lambda_i = 1$ で発散し，global shrinkage prior と比べると裾が重い事前分布であることがわかる．また，$\pi_{\mathrm{SB}}(u_i)$ では，$a > 1$ とすることで原点で 0 に収束する事前分布となり，$\pi_{\mathrm{IRB}}(u_i)$ を用いると，$\pi_{\mathrm{SB}}(u_i)$ に比べ裾が厚い

事前分布であることがわかる．ハイパーパラメータ a, b の選択としてここでは，$p(\lambda_i)$ の裾の挙動の考察の帰結として $(a, b) = (2, 1/2)$ を提案，推奨する (この選択は，grand mean でスパイクをもち，厚い右裾をもつという意味でスパース性に対する主観を反映できている)．なお，$(a, b) = (1/2, 1/2)$ は有名な horseshoe 事前分布 (Carvalho et al, 2010) に対応するがここではこの選択は (SB prior においては) 原点における不要な縮小効果を引き起こすため考えない．また，IRB prior を用いる際は原点における不要な縮小は常に生じてしまうことにも注意する．

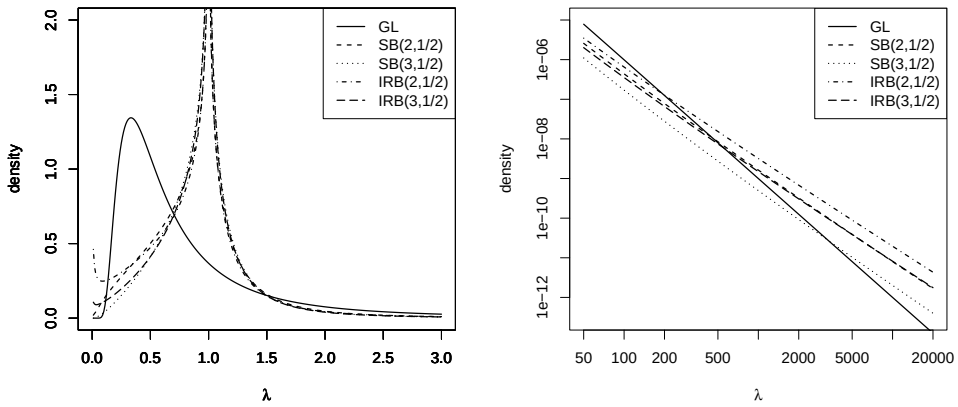

図 1: $\beta = \tau = 1$ とした場合の λ_i の周辺分布 (左) と対数スケールの裾の拡大図 (右)．

2.3 事後分布の計算アルゴリズム

提案する階層ベイズモデルに対して，事後分布の計算は以下で説明する効率的な MCMC 法により簡単に実行できる．ここでは，u_i に対する SB prior を用いた場合のみ紹介する．サンプリングを簡単にするため，$\nu_i = \tau u_i$ $(i = 1, \ldots, n)$ と変数変換し，scaled-beta 分布の積分表現を用いると y を与えたもとでの $(\lambda, \beta, \tau, \nu)$ の同時事後分布は，

$$p(\lambda, \beta, \tau, \nu \mid y) \propto \int_{(0,\infty)^n} \left[\pi(\beta)\pi(\tau) \frac{1}{\tau^{na}} \right. \\ \left. \times \prod_{i=1}^{n} \left\{ t_i^{a+b-1} e^{-t_i} \nu_i^{a-1} e^{-t_i \nu_i/\tau} \frac{\beta^{\nu_i+1} \nu_i^{\nu_i}}{\mathrm{Ga}(\nu_i)} \frac{1}{\lambda_i^{\nu_i+2}} e^{-\beta\nu_i/\lambda_i} \frac{1}{\lambda_i^{\delta_i}} e^{-(\delta_i y_i)/(\lambda_i \eta_i)} \right\} \right] dt$$

とかける．ただし，$\nu = (\nu_1, \ldots, \nu_n)$．ここで，$t = (t_1, \ldots, t_n) \in \mathbb{R}^n_{>0}$ を潜在変数と見て，共役な $\pi(\beta) = \mathrm{Ga}(\beta \mid a_\beta, b_\beta)$，$\pi(\tau) = \mathrm{Ga}(\tau \mid a_\tau, b_\tau)$ を仮定すると，$\lambda, \beta, \tau, t, \nu$ は以下でサンプリングできる．

- $i = 1, \ldots, n$ に対して，$\lambda_i \sim \mathrm{IG}(\delta_i + \nu_i + 1, \delta_i y_i/\eta_i + \beta\nu_i)$ を独立に生成する．
- $\beta \sim \mathrm{Ga}\big(\sum_{i=1}^n \nu_i + n + a_\beta, \sum_{i=1}^n \nu_i/\lambda_i + b_\beta\big)$ を生成する．
- $\tau \sim \mathrm{GIG}\,(-na + a_\tau, 2b_\tau, 2\sum_{i=1}^n t_i\nu_i)$ を独立に生成する．

- $i = 1, \ldots, n$ に対して，$t_i \sim \mathrm{Ga}(a+b, 1+\nu_i/\tau)$ を独立に生成する．

- ν_i の完全条件付き分布は，$\prod_{i=1}^{n}\{\mathrm{Ga}(\nu_i|a, t_i/\tau)\mathrm{Ga}(1/\lambda_i|\nu_i, \beta\nu_i)\}$ に比例する形となりこれは標準的な確率分布とはならないが，Miller (2019) による近似分布を独立型 MH の提案分布として用いる．

実際の利用では，$\pi(\beta) = \mathrm{Ga}(\beta \mid 0.1, 0.1)$，$\pi(\tau) = \mathrm{Ga}(\tau \mid 0.1, 0.1)$ をデフォルトな事前分布として用いる．

3　理論的性質

ここでは，$\beta = \tau = 1$ として，性質 (2) に対する理論的な結果を紹介する．まず，$\pi(u_i)$ に対して，以下の条件を仮定する：

(C1) $\sup_{u \le 1} u\pi(u) < \infty$.

(C2) $u \to 0$ のとき，ある $\alpha \ge 0$，$\gamma \ge -1$ に対して，

$$\pi(u) \sim C \frac{u^{\alpha-1}}{\{1 + \log(1+1/u)\}^{1+\gamma}}.$$

ただし，C は定数で，$f(x) \sim g(x)$ は $\lim_{x\to 0} f(x)/g(x) = 1$ を意味する．提案事前分布である $\pi_{\mathrm{SB}}(u)$ と $\pi_{\mathrm{IRB}}(u)$ が (C1), (C2) を満たすことは容易に確認できる．また，u の事前分布が improper になるケースである，$\alpha = 0$ と $\gamma \le 0$ は除外する．

定理 1. ある関数 $\kappa^* : (0,\infty) \to (0,\infty)$ が存在して，$y_i \to \infty$ のとき，

$$\mathbb{E}[\kappa_i \mid y_i] \sim \frac{1}{\delta_i}(1+\alpha)\kappa^*(\delta_i y_i) \to 0.$$

なお，global shrinkage prior ($u_i = 1$) については，$\mathbb{E}[\kappa_i \mid y_i] = 1/(\delta_i + 1) \nrightarrow 0$ ($y_i \to \infty$) であることから上の定理のような性質は成り立たないことに注意.

次に，予測問題におけるカルバック・ライブラー (KL) リスクを考察する．次のモデルを考える：

$$y \sim \mathrm{Ga}\left(\delta, \frac{\delta}{\lambda}\right), \quad \lambda \sim \mathrm{IG}(1+u, u), \quad u \sim \pi(u).$$

このとき，$f(y \mid \lambda) = \mathrm{Ga}(y \mid \delta, \delta/\lambda)$ とし，λ_0 を真値とすると，$f(y \mid \lambda)$ と $f(y \mid \lambda_0)$ の KL ダイバージェンスは，

$$D^{\mathrm{KL}}(\lambda_0, \lambda) = \delta\left(\frac{1/\lambda}{1/\lambda_0} - 1 - \log\frac{1/\lambda}{1/\lambda_0}\right) = \delta\left(\frac{\lambda_0}{\lambda} - 1 - \log\frac{\lambda_0}{\lambda}\right).$$

である．Barron (1987) により，Cesáro-mean risk R_n について，

$$R_n = n^{-1}\sum_{k=1}^{n} D^{\mathrm{KL}}(f(y \mid \lambda_0) \mid \hat{f}_k(\lambda)) \leq \varepsilon - \frac{1}{n}\log \Pi(\lambda \in A_\varepsilon(\lambda_0))$$

が成り立つ．ただし，$A_\varepsilon(\lambda_0) = \left\{\lambda \in (0,\infty) \mid D^{\mathrm{KL}}(\lambda_0, \lambda) < \varepsilon\right\}$ は λ_0 の KL 近傍であり，$\hat{f}_k(\lambda)$ は観測値 $y_1, \ldots, y_k$ $(k \leq n)$ に基づくベイズ予測密度である．

定理 2. 真のモデル $\mathrm{Ga}(\delta, \delta/\lambda_0)$ に対して，$\lambda_0 \neq 1$ のとき，

$$R_n = O\left(n^{-1}\log n\right).$$

また，$\lambda_0 = 1$ で $u \to \infty$ のとき $0 < b \leq 1/2$ について $\pi(u) \propto u^{-1-b}$ であれば，

$$R_n = O\left\{n^{-1}\left(\log n - \log\log n\right)\right\}.$$

この現象は真の grand mean において高次のオーダーでリスクを改善することから KL 超有効性と呼ばれ，Carvalho et al. (2010) や Datta and Dunson (2016) でも同様の結果が示されている．

4　数値実験

ここでは，ガンマ分布 $y_i \sim \mathrm{Ga}(\delta_i, \delta_i/\lambda_i)$，$\delta_i = 5$，$(i = 1, \ldots, n(= 200))$ に従うデータを以下の 6 つのシナリオに関して生成し，提案手法と既存手法の比較を行う：

(Scenario 1) $\lambda_i \sim 0.95\delta_\mu + 0.05\mathrm{Ga}(20\mu, 2)$,　(Scenario 2) $\lambda_i \sim 0.9\delta_\mu + 0.1\mathrm{Ga}(20\mu, 2)$,
(Scenario 3) $\lambda_i \sim 0.95\delta_\mu + 0.05\mu|t_3|$,　(Scenario 4) $\lambda_i \sim 0.9\mathrm{Ga}(5\mu, 5) + 0.1\mu|t_1|$,
(Scenario 5) $\lambda_i \sim 0.9\delta_\mu + 0.1\mathrm{Ga}(10\mu, 2)$,　(Scenario 6) $\lambda_i \sim 0.85\delta_\mu + 0.15\mathrm{Ga}(10\mu, 2)$.

ただし，$\mu = 5$ で，δ_a と t_c はそれぞれ a における一点分布と自由度 c の t-分布を表す．比較手法は，提案手法 (SB, IRB)，global shrinkage estimate (GL)，DasGupta (1986) による古典的な縮小推定 (DG)，Lu and Stephens (2016) による adaptive variance shrinkage 法，最尤推定値 y_i (ML) の 6 つの手法である．1000 回の繰り返しによる mean absolute percentage error (MAPE) $n^{-1}\sum_{i=1}^{n}\lambda_i^{-1}|\lambda_i - \hat{\lambda}_i|$ を図 2 に示す．数値実験の詳細は当日報告する．

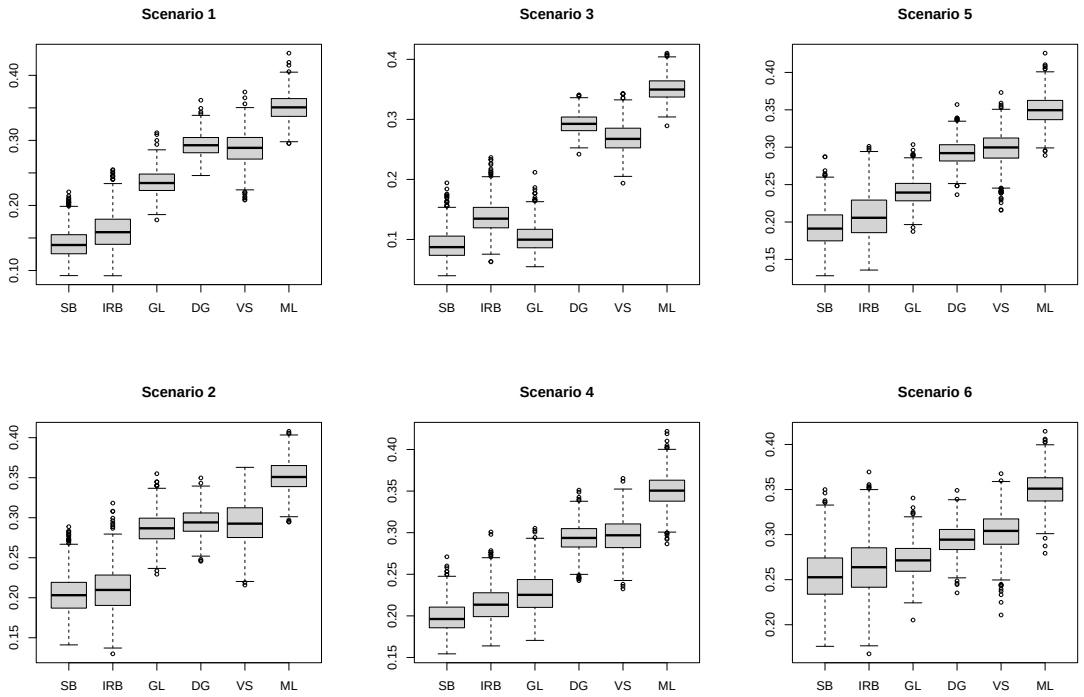

図 2: MAPE のボックスプロット

5　実データへの応用例

5.1　韓国の COVID-19 の平均入院期間

韓国の 2020 年初旬の COVID-19 患者の平均入院期間に関するデータを扱う[1]．1587 人の患者データ (死亡や打ち切りが生じたデータは除外) について，98 の町と 3 つの年齢階級でグループ化した $n = 185$ のデータについて，各患者の平均入院期間がグループごとに平均 λ_i の指数分布に従うとすると，グループごとの標本平均 y_i は，

$$y_i \sim \mathrm{Ga}(n_i, n_i/\lambda_i) \quad (i = 1, \ldots, n)$$

となる．ただし，n_i は i 番目のグループのサンプルサイズである．したがって，(3) で $\delta = n_i$ とおくけばよい．このデータに提案手法 (SB) と GL, DG を用いて，結果を比較する．

5.2　遺伝子発現データの分散の縮小推定

Singh et al. (2002) による前立腺癌 (prostate cancer) の遺伝子発現データの分散推定に適用する．50 のコントロール群に対する $n = 6033$ の遺伝子について，n 個の遺伝子の発現

[1] `https://www.kaggle.com/kimjihoo/coronavirusdataset`

量が正規分布に従うと仮定したとき，その標本分散を y_i とおくと

$$y_i \sim \mathrm{Ga}(n_i/2, n_i/(2\lambda_i)) \quad (i = 1, \ldots, n)$$

となる．ただし，λ_i は i 番目の遺伝子の発現量の真の分散である．したがって，(3) で $\delta_i = n/2$ とおけばよい．このデータに提案手法 (SB, IRB) と GL, VS を用いて，結果を比較する．

参考文献

[1] Barron, A. R. (1987). Are Bayes rules consistent in information?. In Open problems in communication and computation (pp. 85-91). Springer, New York, NY.

[2] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.

[3] Datta, J. and Dunson, D. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4), 971–983.

[4] DasGupta, A. (1986). Simultaneous estimation in the multiparameter gamma distribution under weighted quadratic losses. *The Annals of Statistics*, 14(1), 206–219.

[5] Donoho, D. and Jin, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics*, 34, 2980–3018.

[6] Lu, M. and Stephens, M. (2016). Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances. *Bioinformatics*, 32, 3428–3434.

[7] Miller, J.W. (2019). Fast and accurate approximation of the full conditional for gamma shape parameters. *Journal of Computational and Graphical Statistics*, 28, 476–480.

[8] Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* 9, 501–538.

[9] Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D ' Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2), 203–209.