Supplementary: Bayesian negative binomial regression for differential expression with confounding factors

Siamak Zamani Dadaneh, Mingyuan Zhou and Xiaoning Qian

March 27, 2018

In this supplementary file, we provide the pseudo-code of our inference algorithm for Bayesian negative binomial regression (`BNB-R`), the definition of Chinese Restaurant Table (CRT) distribution, additional results on synthetic data, and the enlarged plots for the figures in the main text.

1 `BNB-R` Pseudo-code

We first provide the pseudo code of `BNB-R`, including Bayesian inference via Gibbs sampling and the consequent differential expression (DE) analysis based on the symmetric Kullback-Leibler (KL) divergence.

Algorithm 1 `BNB-R` differential expression analysis

Inputs: gene expression counts, design matrix, N
Outputs: KL-divergence based ranking of DE genes

Initialize model parameters
\# Do Gibbs sampling:
for $iter = 1$ to N **do**
  Sample ℓ_{kj} using Chinese Restaurant Table (CRT) distribution
  Update r_j using the gamma-Poisson conjugacy
  Sample auxiliary variables ω_{kj}, using the Polya-Gamma (PG) distribution
  Update regression coefficients
  Update α_p and h
end for
Calculate KL-divergence between posterior samples

2 Chinese restaurant table (CRT) distribution

The negative binomial distribution $m \sim \mathrm{NB}(r, p)$ with the probability mass function

$$f_M(m) = \frac{\Gamma(m+r)}{m!\Gamma(r)}(1-p)^r p^m, \quad m \in \{0, 1, \ldots\}$$

can be augmented as a gamma mixed Poisson distribution as

$$m \sim \mathrm{Pois}(\lambda), \ \lambda \sim \mathrm{Gamma}(r, p/(1-p)),$$

where the gamma distribution is parametrized by its shape r and scale $p/(1-p)$. It can be augmented under a compound Poisson representation as

$$m = \sum_{t=1}^{\ell} u_t, \ u_t \sim \mathrm{Log}(p), \ \ell \sim \mathrm{Pois}(-r\ln(1-p)),$$

where $u \sim \mathrm{Log}(p)$ is the logarithmic distribution with probability generation function $C_U(z) = \ln(1-pz)/\ln(1-p)$, $|z| < p^{-1}$. We denote the conditional posterior distribution of ℓ given m and r by $(\ell|m, r) \sim \mathrm{CRT}(m, r)$ and sample it with the summation of independent Bernoulli random variables as $\ell = \sum_{n=1}^{m} b_n$, $b_n \sim \mathrm{Bernoulli}[r/(n-1+r)]$ [7].

3 Results of `BNBP` and `GNBP` on synthetic data of Section 3.1.1

Figure 1 plots the comparison of `BNBP` and `GNBP` [1] with other differential expression analysis methods, `edgeR` [6], `DESeq2` [5], and `voom` [2], on synthetic data described in Section 3.1.1 of the main text, in terms of the areas under both the receiver operating characteristic (ROC) and precision-recall (PR) curves. With Figure 1 in the main text, we can clearly see that our `BNB-R` considering the effects of covariates has the best performance with a significant margin over all the other algorithms.

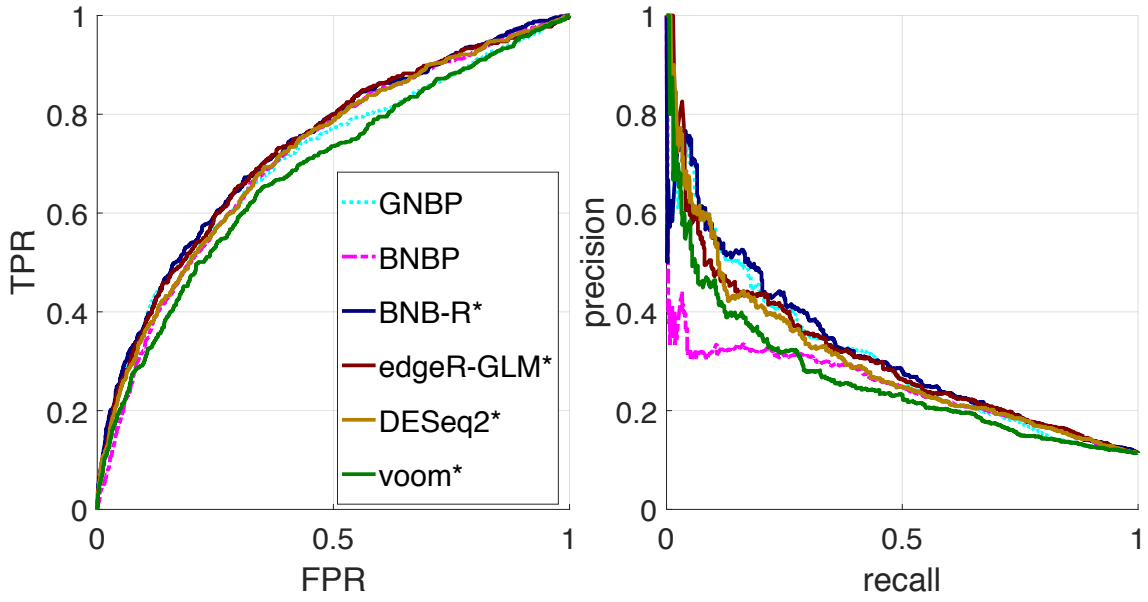

Figure 1: **Left panel**: ROC curve, **Right panel**: PR curve. Performance comparison of different methods in detecting differentially expressed genes generated under a negative binomial regression model with covariates: *condition*, *gender* and *dosage*. The curves correspond to the case that only the condition covariate in used in differential expression analysis. In particular, the results for `BNBP` and `GNBP`, which were omitted in the original paper, are included here.

4 Figures in the paper

We plot the figures in the main text with larger size to better visualize the performance difference between different DE analysis methods.

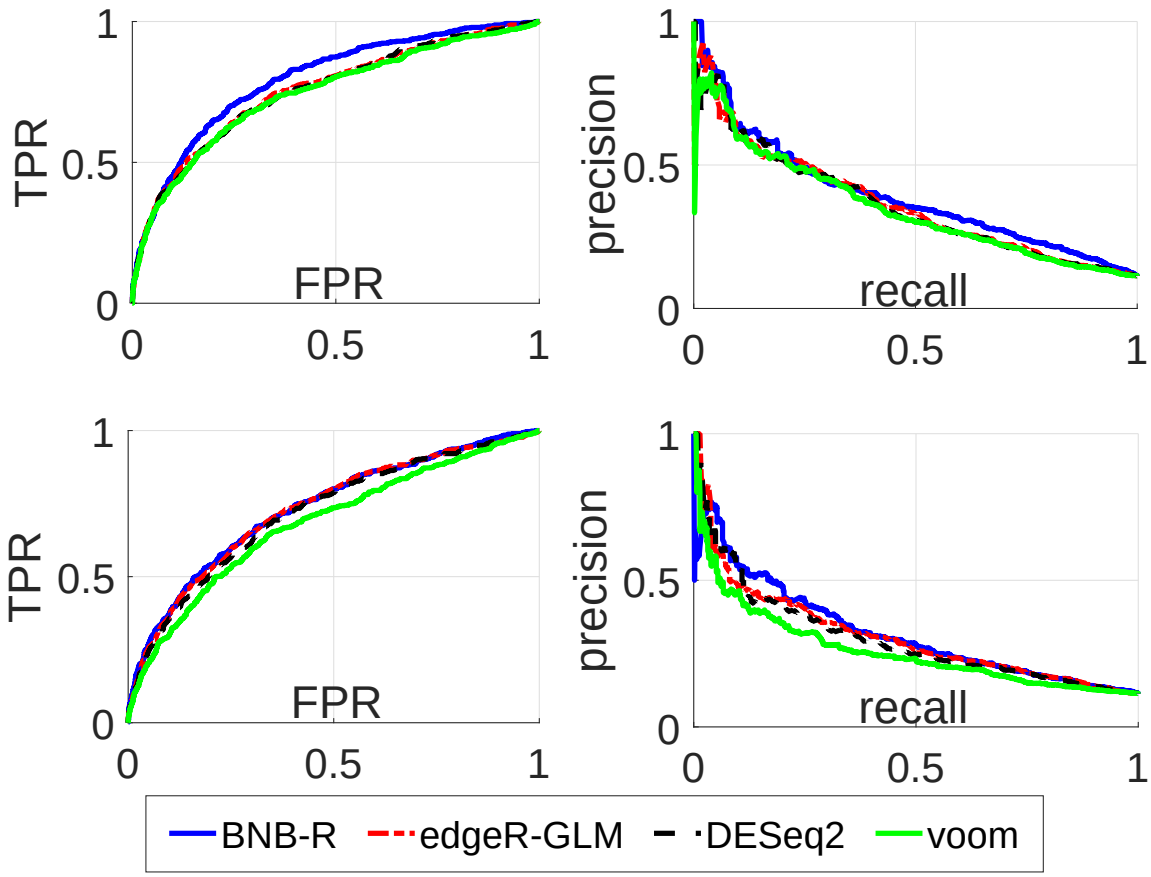

Figure 2: **Left panel**: ROC curve, **Right panel**: PR curve. Performance comparison of different methods in detecting differentially expressed genes generated under a negative binomial regression model with covariates: *condition*, *gender* and *dosage*. Panels in the top row correspond to the case that full covariate information is used in differential expression analysis. Panels in the bottom row correspond to the case that only condition covariate is used in differential expression analysis.

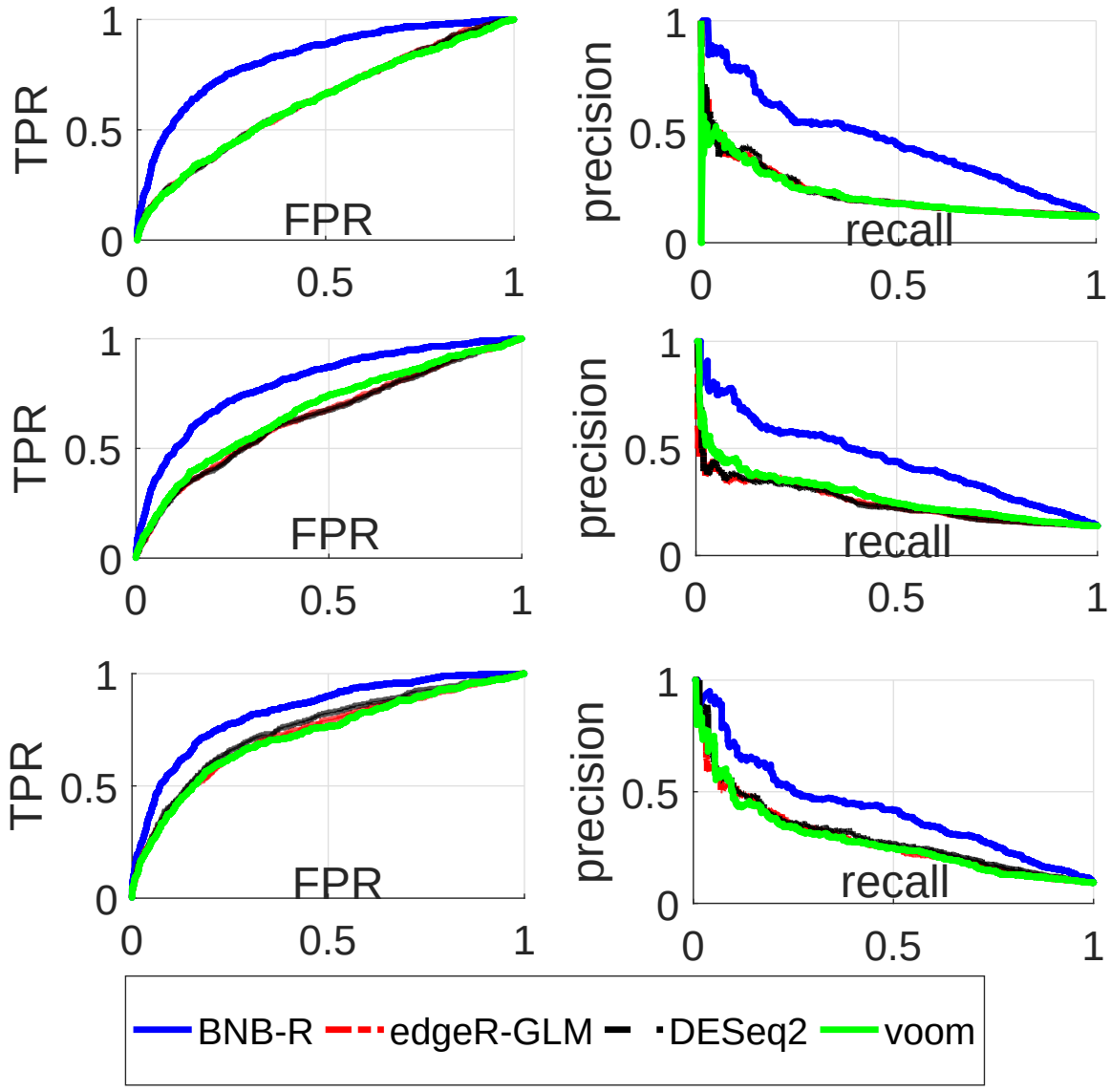

Figure 3: **Left panels**: ROC curve, **Right panels**: PR curve. Performance comparison of different methods in detecting differentially expressed genes generated under the negative binomial regression model with covariates: *condition*, *gender*, *dosage*, and interaction of *condition* and *gender*. The panels in the top and middle rows correspond to differentially expressed genes across conditions for males and females, respectively. The panels in the bottom row correspond to differentially expressed genes for the case that full covariate information is not employed, with the interaction term excluded from differential expression analyses by all the methods.

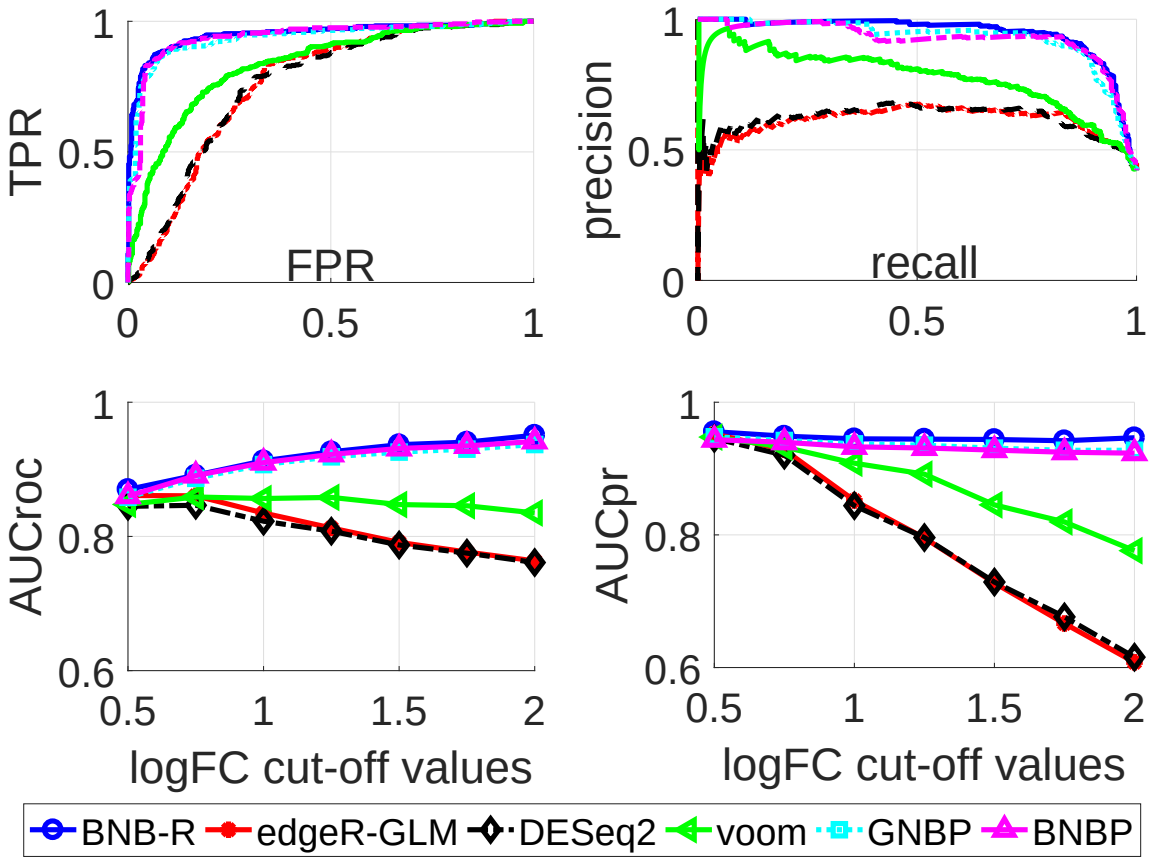

Figure 4: **Top row**: ROC and PR curves for a fixed cut-off, **Bottom row**: AUC of ROC and PR curves for different cut-off values. Performance comparison of different methods in detecting differentially expressed genes on real-world benchmark RNA-seq data from the SEQC project. `edgeR`, `DESeq2`, and `voom` are applied in conjunction with SVA with two surrogate variables.

5 Surrogate Variable Analysis

Figure 5 illustrates the performance of `edgeR` [6], `DESeq2` [5], and `voom` [2] on the SEQC benchmark data, with and without surrogate variable analysis [4, 3].

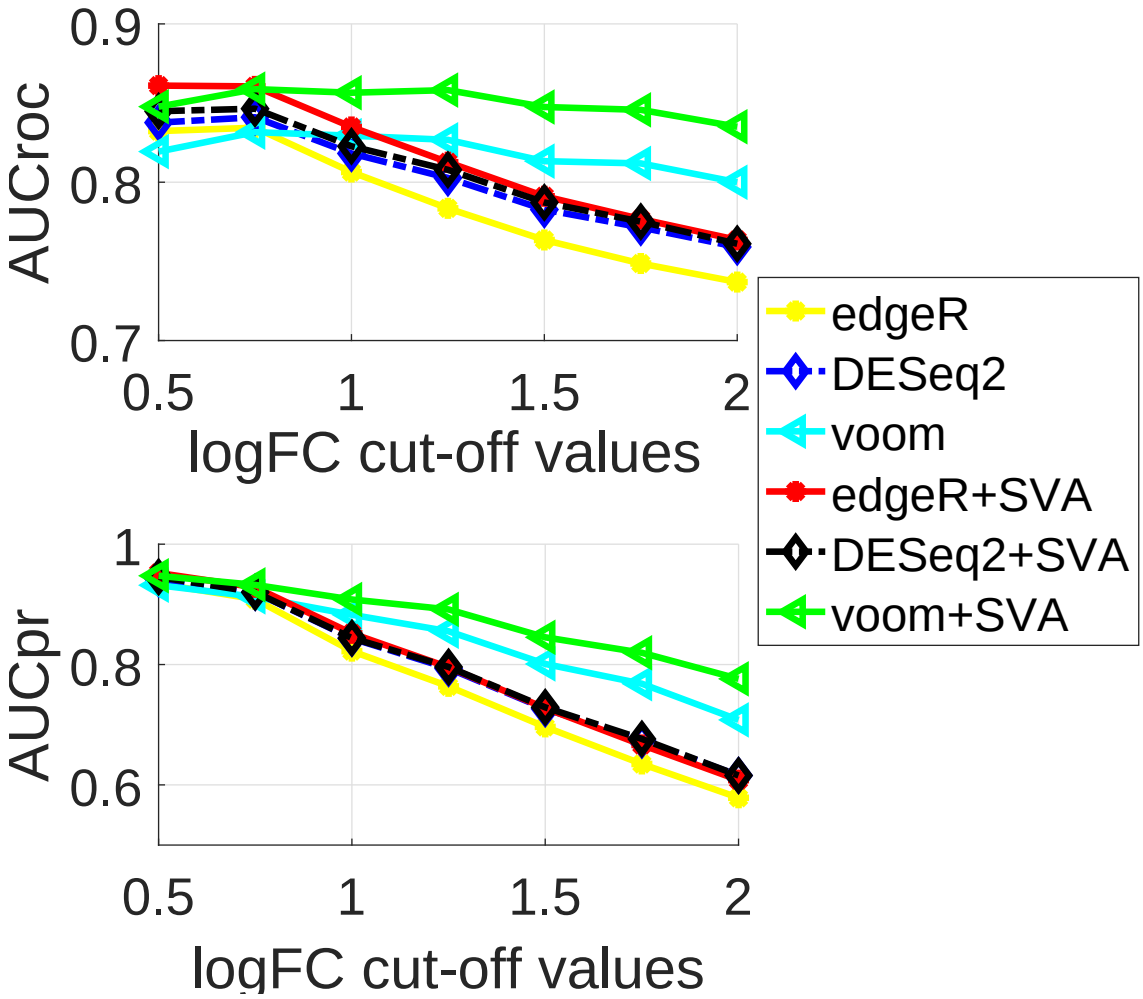

Figure 5: AUC of ROC and PR curves for edgeR, DESeq2 and voom methods applied to SEQC benchmark data, with and without surrogate variable analysis.

As shown in Figure 5, `edgeR`, `DESeq2`, and `voom` together with `sva` indeed achieve better performance on differential expression analysis compared to the corresponding original methods. However, as shown in Figure 4 here or Figure 3 of the main text, their performances are still not as good as `BNB-R`, which explicitly model the influence from the covariates of interest.

References

[1] Siamak Zamani Dadaneh, Xiaoning Qian, and Mingyuan Zhou. BNP-Seq: Bayesian nonparametric differential expression analysis of sequencing count data. *Journal of the American Statistical Association*, (in-press, doi:10.1080/01621459.2017.1328358), 2017.

[2] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014.

[3] Jeffrey T Leek. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research*, 42(21):e161–e161, 2014.

[4] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[5] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

[6] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[7] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.