Accounting for Nuisance Covariates when Using RNA-Seq Data to Identify Differentially Expressed Genes

Yet Nguyen   Dan Nettleton

Iowa State University

March 18, 2014

Introduction

- Suppose that we have RNA sequencing (RNA-seq) read count data on J genes from an experiment with two treatment groups.
- We want to find genes that are differentially expressed (DE) between two treatment groups.
- Analyses can often be complicated by the presence of nuisance factors that arise due to experimental design limitations and heterogeneity of experimental units that can be seen in continuous covariates measured for each experimental unit and/or RNA sample.

Prototypical Dataset

	Treatment 1				Treatment 2			
	u_{11}	u_{12}	$\cdots$	u_{1n}	u_{21}	u_{22}	$\cdots$	u_{2n}
x	0.5	0.95	$\ldots$	-1.42	$-.45$	.89	$\ldots$	1.2
gene 1	56	2014	$\cdots$	28	31	975	$\cdots$	3289
gene 2	0	2	$\cdots$	1	0	0	$\cdots$	1
gene 3	1	3	$\cdots$	0	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
gene J	1701	264	$\cdots$	345	14	234	$\cdots$	34

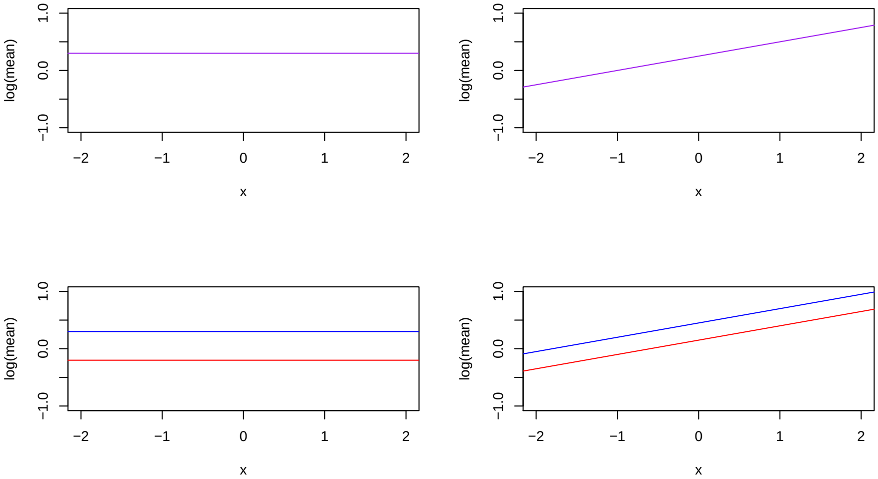
Analysis Strategies for a Single Covariate

- Ignore the covariate, i.e., assume the covariate is not associated with the transcript abundance of any gene.
- Include the covariate in a model for each gene, i.e., assume the covariate may be associated with the transcript abundance of each gene.
- Assume the covariate is associated with transcript abundance of some subset of genes and consider model selection criteria or model averaging strategies when performing inference for treatment effects.

Analysis Strategies for a Single Covariate

- Ignore the covariate, i.e., assume the covariate is not associated with the transcript abundance of any gene.
- **Include the covariate in a model for each gene**, i.e., assume the covariate may be associated with the transcript abundance of each gene.
- Assume the covariate is associated with transcript abundance of some subset of genes and consider model selection criteria or model averaging strategies when performing inference for treatment effects.

Scenarios Considered



Model for Simulated Data

$i = 1, 2;\ j = 1, \cdots, J;\ k = 1, \cdots, n$

$$y_{ijk} \sim \mathrm{NegBin}(\mu_{ijk}, \omega_j), \quad \log(\mu_{ijk}) = \tau_{ij} + \beta_j x_k,$$

where

- $n = 5$ or $n = 20$ experimential units per treatment
- $x_k \sim N(0, 1)$.

Parameters Used for Simulation

- $\tau_{1j} - \tau_{2j}$ represents the log fold change for gene j.
- ω_j is the negative binomial dispersion parameter.
- Values for $\tau_{1j} - \tau_{2j}$ and ω_j were simulated to match values estimated from real data.
- In all simulations, 80% of the $\tau_{1j} - \tau_{2j}$ values were set to zero among $J = 1000$ genes.

Simulated Slope Coefficients

- $\beta_j \sim \frac{\nu}{2} \times \text{Unif}(L, U) + \frac{\nu}{2} \times (-\text{Unif}(L, U)) + (1 - \nu) \times \delta_{\{0\}}$
  - $\nu \in \{0, 0.25, 0.50, 0.75\}$
  - $(L, U) \in \{(0.1, 0.5), (1, 1.5)\}$

Testing for Treatment and Covariate Effects

- The QuasiSeq R package (Lund et al., 2012) was used to obtain a *p*-value for each test.
- $p_{j\tau}$ is the *p*-value for the test of $H_{0j} : \tau_{1j} = \tau_{2j}$ for the no-covariate model where $\log(\mu_{ijk}) = \tau_{ij}$.
- $p_{j\tau|\beta}$ is the *p*-value for the test of $H_{0j} : \tau_{1j} = \tau_{2j}$ for the covariate model where $\log(\mu_{ijk}) = \tau_{ij} + \beta_j x_k$.
- $p_{j\beta|\tau}$ is the *p*-value for the test of $H_{0j} : \beta_j = 0$ for the covariate model where $\log(\mu_{ijk}) = \tau_{ij} + \beta_j x_k$.

Methods for Identification of DE Genes

We consider 5 methods for converting $p_{j\tau}, p_{j\tau|\beta}, p_{j\beta|\tau}$ to a decision about the DE status of gene j.

1. nocov: Convert $p_{j\tau}$ ($j = 1, \cdots, J$) to q-values.
2. cov: Convert $p_{j\tau|\beta}$ ($j = 1, \cdots, J$) to q-values.

Methods for Identification of DE Genes

3. ebp: Convert

$$p_{j\tau} I[\text{EBP}(p_{j\beta|\tau}) > 0.5] + p_{j\tau|\beta} I[\text{EBP}(p_{j\beta|\tau}) \leq 0.5]$$

$(j = 1, \ldots, J)$ to q-values.

4. aic: Convert

$$p_{j\tau} I[\text{AIC}_{j\tau} < \text{AIC}_{j\tau\beta}] + p_{j\tau|\beta} I[\text{AIC}_{j\tau} \geq \text{AIC}_{j\tau\beta}]$$

$(j = 1, \ldots, J)$ to q-values.

5. aaa: Compute

$$\text{EBP}_j = \text{EBP}(p_{j\tau})\text{EBP}(p_{j\beta|\tau}) + \text{EBP}(p_{j\tau|\beta})[1 - \text{EBP}(p_{j\beta|\tau})]$$

$(j = 1, \ldots, J)$.

Computing Empirical Bayes Probability (EBP)

- Suppose p_j is a p-value for testing a null hypothesis H_{0j} for $j = 1, \cdots, J$.

- The EBP for p_j is computed as

$$\mathrm{EBP}(p_j) := \widehat{\mathrm{P}}(H_{0j} | p_j) = \frac{\hat{\pi}_0}{\hat{f}(p_j)},$$

  where $\hat{f}$ is the Grenander estimator of the pdf of p-values, which is the nonparametric MLE of decreasing pdf of p-values, and $\hat{\pi}_0 = \hat{f}(1)$.

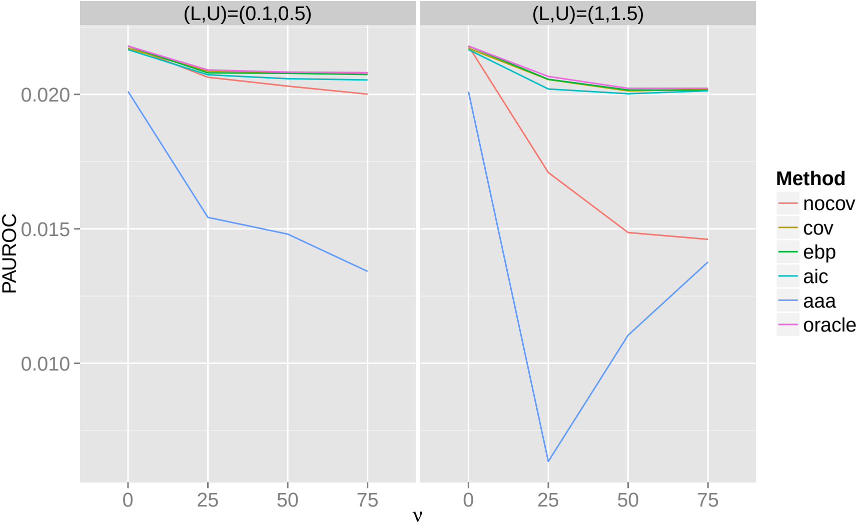
Citations for Methods

- Storey (2002)
- Efron et al. (2001)
- Strimmer (2008)
- Grenander (1956)
- Pounds and Rai (2009)
- Pounds et al. (2012)

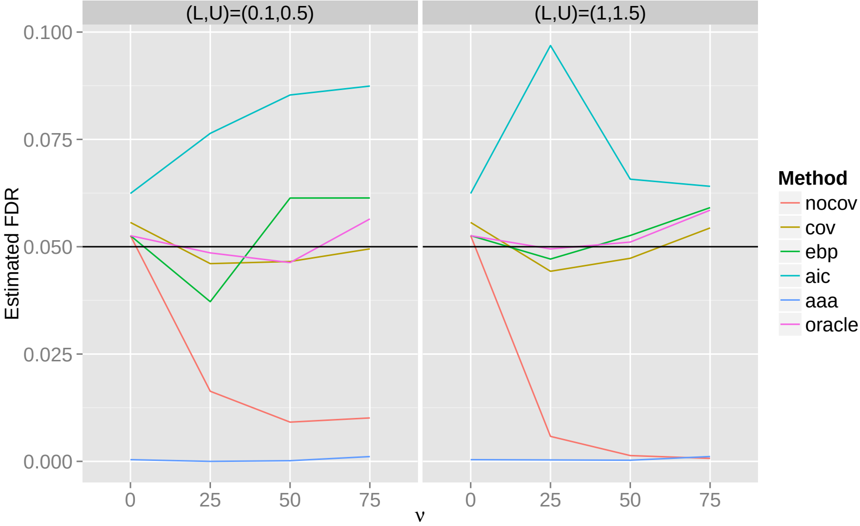
Evaluation of Methods

- For 100 replications of each simulation setting, we compute
  - Average Area Under the Receiver Operating Characteristic Curve
  - Estimated FDR when FDR is nominally controlled at 0.05
- We use Storey's FDR control procedure based on q-values for Method 1, 2, 3, and 4, and average accummulative EBP for Method 5.

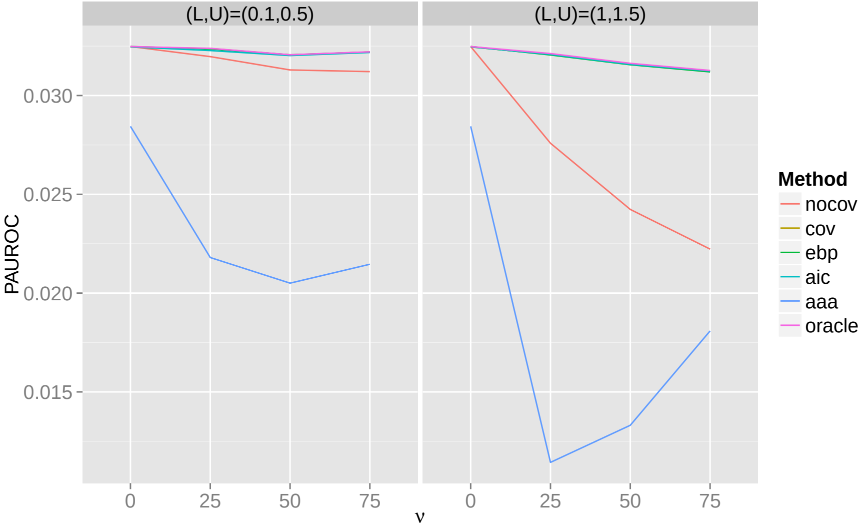
Partial Area under ROC (FPR ≤ 0.05), $n = 5$



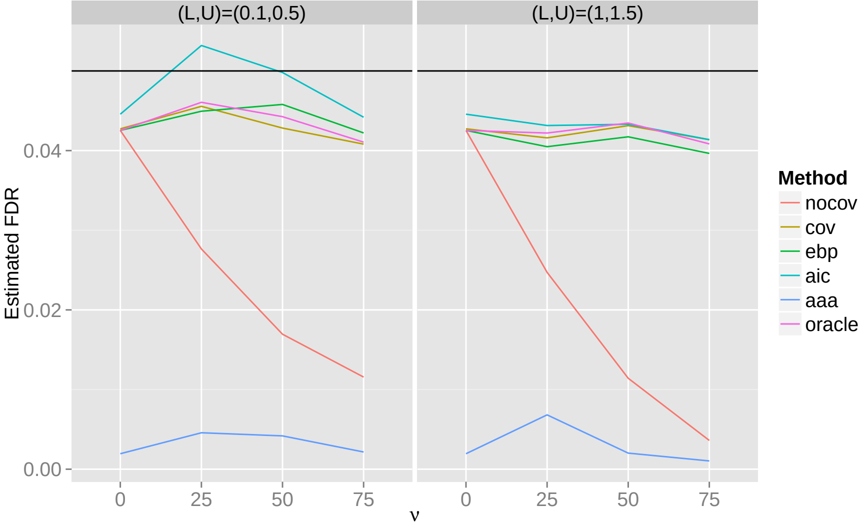
Control FDR at 5%, $n = 5$



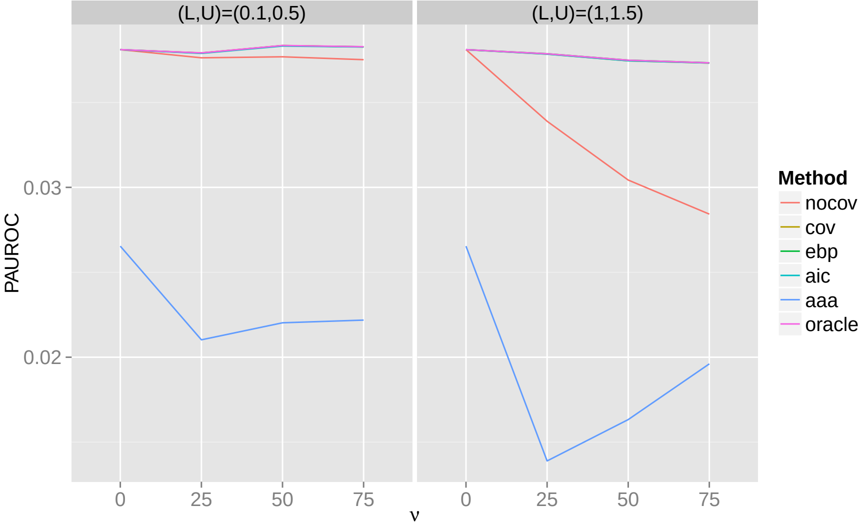
Partial Area under ROC (FPR ≤ 0.05), $n = 20$



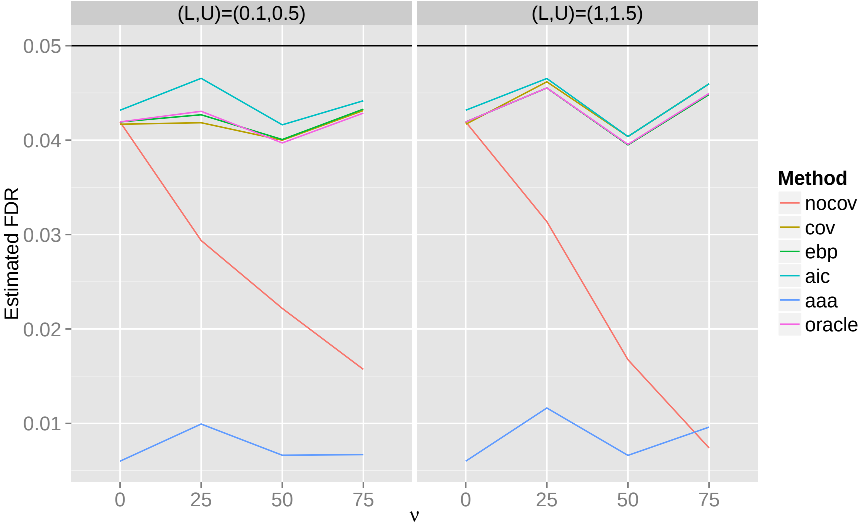
Control FDR at 5%, $n = 20$



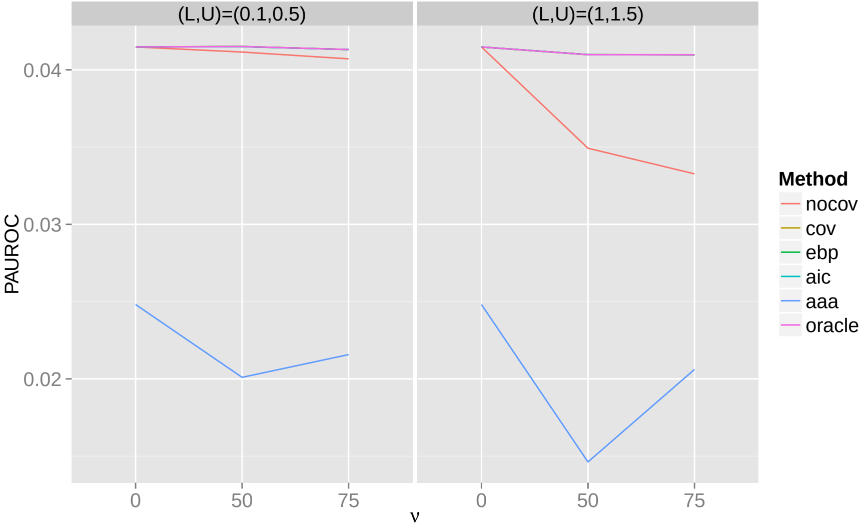
Partial Area under ROC (FPR ≤ 0.05), $n = 50$



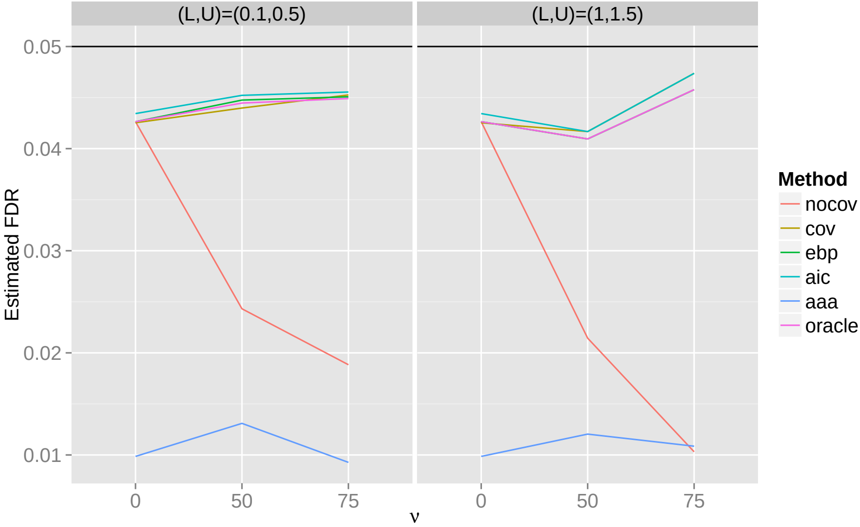
Control FDR at 5%, $n = 50$



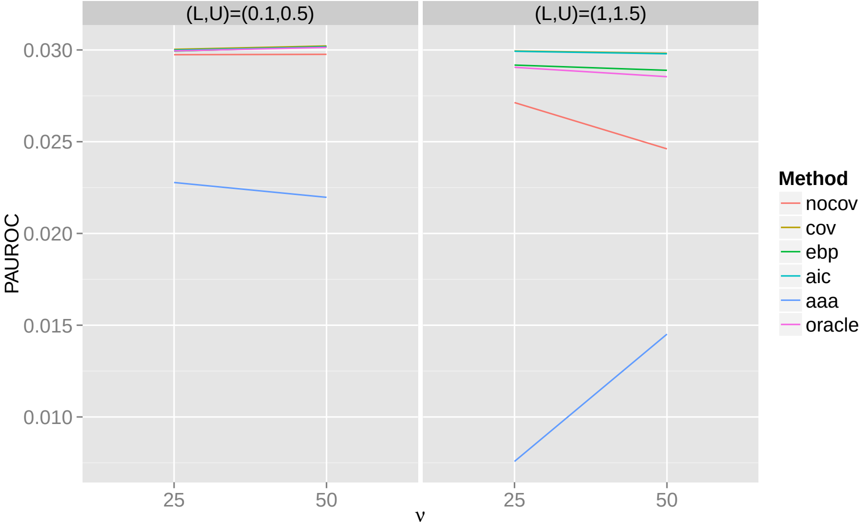
Partial Area under ROC (FPR ≤ 0.05), $n = 100$



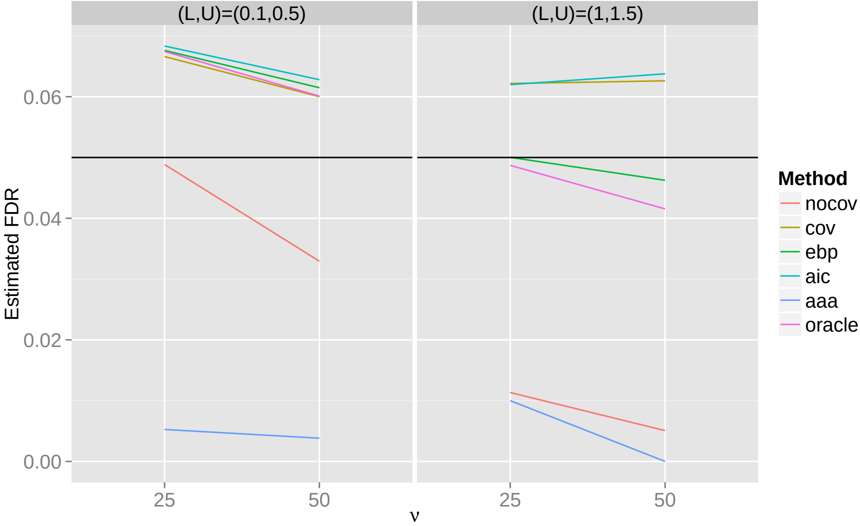
Control FDR at 5%, $n = 100$



EdgeR, Partial Area under ROC (FPR ≤ 0.05), $n = 20$



Edger, Control FDR at 5%, $n = 20$



Comments on the Results

- For the simulation settings we considered, the cov method based on $p_{j\tau|\beta}$ $(j = 1, \ldots, J)$ performed best overall.

- The price paid for including an irrelevant covariate in the model was far less than the cost of excluding an important covariate.

- All other methods relied on $p_{j\tau}$ $(j = 1, \ldots, J)$ to some extent and performed poorly whenever a substantial portion of genes were associated with the covariate.

Problems with $p_{j\tau}$

- When $\beta_j \neq 0$, $p_{j\tau}$ is likely to be inflated because variation in the response unexplained by the model pushes the test of $H_{0j} : \tau_{1j} = \tau_{2j}$ toward nonsignificance.
- Even when $\beta_j = 0$, $p_{j\tau}$ is likely to be inflated if $\beta_{j^*} \neq 0$ for a substantial portion of genes $j^* \neq j$.
- Borrowing information across genes to estimate gene-specific dispersions can lead to overestimation of dispersions when $\beta_{j^*} \neq 0$ for many genes and the covariate is excluded from the model for each gene.

Conclusions

- When using existing software for RNA-seq analysis that requires the model for each gene to have the same design matrix, it may be best to favor flexibility over simplicity.
- More sophisticated empirical Bayes of fully Bayesian strategies for combining model selection and inference are needed.

Acknowledgements

- This material is based upon work supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30336 from the USDA National Institute of Food and Agriculture, and the National Science Foundation (NSF) under Grant No. 0922746.

- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the USDA or NSF.