SUPPLEMENT TO "GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION"

By Andrew McDavid[*] Raphael Gottardo[†,‡] Noah Simon[§] and Mathias Drton[‡,¶]

Department of Biostatistics and Computational Biology[*], University of Rochester Medical Center; Rochester, New York. Vaccine and Infectuous Disease Division[†], Fred Hutchinson Cancer Research Center, Department of Statistics[‡] and Department of Biostatistics[§], University of Washington; Seattle, Washington. Department of Mathematical Sciences[¶], University of Copenhagen; Denmark.

1. Data processing. The method, and code to reproduce results in this paper is available as an R package at `https://github.com/amcdavid/HurdleNormal`.

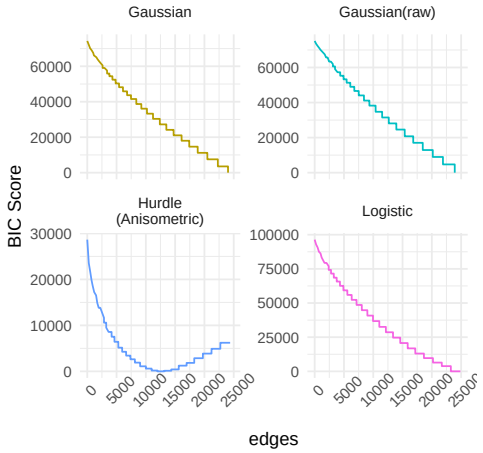
In all models and data sets, the cellular detection rate $\sum_j I_{y_{ij}>0}$ [Finak et al., 2015] was used as an unpenalized adjustment covariate in $\mathbf{W}$ as described in Algorithm 1. In the Tfh data, a separate, unpenalized intercept was fit for each donor, as well. For the Gaussian and Hurdle models, positive values were conditionally centered

$$\tilde{y}_{ij} = \begin{cases} 0 & \text{if } v_{ij} = 0, \\ y_{ij} - \bar{y}_j^+ & \text{else,} \end{cases}$$

where $\bar{y}_j^+$ is the average in a gene over positive values. This made V_j and Y_j marginally orthogonal, speeding up the convergence of the optimization algorithm and reducing the leverage of zeros in the Gaussian model. The "Gaussian(raw)" model was also fit to the untransformed data, but not always discussed as it gave similar results as the Logistic model.

The graph stability (via repeated 50% sample splitting) was used to estimate the network size. At 60% stability, the number of selected edges ranged from 11 (Hurdle) to 32 (Gaussian).

Background noise in the mouse dendritic cells (mDC) data set was thresholded as described previously [Finak et al., 2015], and filtered for low-expression and cluster-disrupted cells. Supplemental Figure 1 shows the Bayesian information criterion for the fitted path. An interior minimum fails to occur in the solution path for three of the methods.



Supplemental Figure 1: Bayesian information criterion on mDC data set

2. Singular normal distributions and Multivariate Hurdle Distributions. A random vector $\mathbf{Y}$ has singular Normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ [Rao, 1973] with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ with rank $r < m$ if the following holds for a matrix $\mathbf{U}$ with $\mathbf{U}^T\boldsymbol{\Sigma} = 0$: a) $\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\boldsymbol{\mu}$ almost surely, and b) $\mathbf{Y}$ has a density

$$f(\mathbf{y}) = \frac{(2\pi)^{-r/2}}{(\det{}^+ \boldsymbol{\Sigma})^{1/2}} \exp\{-(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^-(\mathbf{y}-\boldsymbol{\mu})/2\}, \tag{1}$$

with respect to Lebesgue measure restricted to the hyperplane $\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\boldsymbol{\mu}$. Here $\det^+$ is the pseudo-determinant (product of non-zero eigenvalues) and $\boldsymbol{\Sigma}^-$ is a pseudo-inverse, such as the Moore-Penrose inverse. In the case that $\boldsymbol{\Sigma}$ is zero outside a positive-definite submatrix of size $r \times r$, $\mathbf{U}$ can be chosen to be a diagonal selection matrix consisting of zeros and ones, and $\mathbf{Y}$ has a density with respect to the measure $\lambda^r \otimes \delta_0^{m-r}$, which is the case treated here.

2.1. *Normalizing the joint density.* The expression

$$f(\mathbf{y}) \;=\; \exp\left\{\mathbf{v}^T\mathbf{G}\mathbf{v} + \mathbf{v}^T\mathbf{H}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{K}\mathbf{y} - C(\mathbf{G},\mathbf{H},\mathbf{K})\right\}, \qquad \mathbf{y} \in \mathbb{R}^m, \tag{2}$$

that was given in (6) is a normalizable density. Let $\mathbf{K}^+ = \left(\mathcal{I}\mathbf{K}\mathcal{I}\right)^-$ and rewrite (5) as

$$\begin{aligned}\log f(\mathbf{y}|\mathbf{V}=\mathbf{v}) &= \mathbf{v}^T\mathbf{H}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{K}\mathbf{y} \\ &= \mathbf{v}^T\mathbf{H}\mathbf{y} - \mathbf{v}^T\mathbf{H}\mathbf{K}^+\mathbf{H}\mathbf{v} + \mathbf{v}^T\mathbf{H}^T\mathbf{K}^+\mathbf{K}\mathbf{K}^+\mathbf{H}\mathbf{v} - \frac{1}{2}\mathbf{y}^T\mathbf{K}\mathbf{y}\end{aligned}$$

Using the notation from (7) and applying (1), the normalizing constant of the density in (1) is found to be given by

$$C(\mathbf{G},\mathbf{H},\mathbf{K}) = \log \sum_{\mathbf{v}\in\{0,1\}^m} \exp\left[\mathbf{v}^T\mathbf{G}\mathbf{v} + \mathbf{h}^T\left(\mathcal{I}\mathbf{K}\mathcal{I}\right)^-\mathbf{h}/2\right]\left[\det{}^+\left(\frac{1}{2\pi}\mathcal{I}\mathbf{K}\mathcal{I}\right)\right]^{1/2}.$$

2.2. *The anisometric penalty is a score test of $\theta_a = 0$ for all a.*

PROPOSITION 1. *Let $\mathbf{H} = \left[\frac{\partial^2 \log f_{[b|A]}(\mathbf{y})}{\partial\theta_i\theta_j}\right]$ be the conditional information. Suppose $\mathbf{H}$ and thus also its inverse $\mathbf{H}^{-1}$ is block-diagonal. Then the anisometric group lasso penalty is equivalent to a score test of the null hypothesis that $\theta = 0$ vs. the alternative that a pre-specified subvector $\theta_a \neq 0$.*

Proof: Let $c = \mathcal{V}\setminus\{a,b\}$ and suppose that $\theta_c = 0$. From the KKT conditions, $\theta_a = 0$ is an optimum if and only if

$$\nabla_a^T H_{aa}^{-1}\nabla_a < \lambda^2,$$

where $\nabla_a = \frac{\partial \log f_{[b|A]}(\mathbf{y})}{\partial\theta_a}$ is the a-subvector of the conditional log-likelihood gradient. Taking λ^2 to be an appropriate quantile from a χ^2-distribution with $\dim(H_{aa})$ degrees of freedom results yields a score test.

3. Simulation details.

3.1. *Graphs and parametric alternatives.* In the *G-minimal* and *complete* scenarios, the underlying graph is a (perhaps incomplete) chain, with either 1.5% of nodes connected (Figure 4a) or 5% (Figures 4b and 5). In the *e. coli* scenario, the underlying graph is a 500-vertex subgraph sampled from a network described in Gama-Castro et al. [2011] and available from GeneNetWeaver [Schaffter et al., 2011]. In the G parametric alternative, given the underlying graph, the data are derived from model (6), restated in (2), with only the G interaction matrix set to non-zero. In this case, although the specified conditional independences hold exactly, an auto-logistic

(Ising) model is minimally complete, while the multivariate Hurdle model is over-parametrized. In the *complete* parametric alternative, given the underlying graph, all three interaction matrices G, H and K are non-zero simultaneously in the appropriate entries.

3.2. *Generative models.* The Hurdle generative model, and deviations from it are considered. In the *exact* case, observations are generated through Gibbs sampling from model (6) using the full conditional distributions available in (8). Samples from conditional distributions are generated simply as Bernoulli and Normal random variates. A 2000 iteration burn-in phase, and sample thinning was employed. Thinned samples exhibited only mild auto-correlation. In the *contaminated* case, a matrix of exact variates $\mathbf{Y}$ are sampled, and onto them (given $Y_{ij} \neq 0$) is added t_8-distributed noise. So the final variates remain zero-inflated, but are heavier-tailed than a Normal distribution. In the *selection* case, a matrix $\tilde{\mathbf{Y}}$ of latent, non-zero-inflated Gaussian variates are sampled that follow the graphical model implied by the K-interaction matrix. These are zero-inflated through a selection model

$$P\left(\tilde{V}_j|\mathbf{Y}=\mathbf{y}\right) = \text{logit}(a_j + b_j y_j),$$
$$\mathbf{Y} = \tilde{\mathbf{Y}}\tilde{V}.$$

The parameters a_j and b_j are chosen to keep $P(\tilde{V}_j)$ away from the boundary values 0 and 1.

Lastly, in some cases, we consider *in-silico 10-cell* replicates. Given a desired sample size n, draw $10n$ observations $\mathbf{Y}$ from model (6), and let the observed data $\mathbf{Y^{(10)}}$ follow

$$\mathbf{Y^{(10)}} = \log_2 \sum_{i=1}^{10} 2^{\mathbf{Y}_i}/10.$$

References.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, dec 2015.

Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muniz-Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S. García-Sotelo, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Shirley Alquicira-Hernández, Alejandra Medina-Rivera, Irma Martínez-Flores, Kevin Alquicira-Hernández, Ruth Martínez-Adame, César Bonavides-Martínez, Juan Miranda-Ríos,

Graph topologies and parametric models

1. G-minimal chain graphs, with tri-diagonal G-interaction matrix with off-diagonal entries set to 1, and diagonal H, K. In this case, an Ising/logistic model is minimally complete.
2. G-H-K-complete chain graphs, with off diagonal $G = .2, H = -.75, K = -.4$. The proposed model is thus minimally complete.
3. *e. coli*-networks: 500 edges from a semi-empirical *e. coli* network and pairwise hurdle likelihood. 50% of edge weights are G-minimal, 25% K-minimal and 25% complete.
4. 10-cell versions of 1-3. The 10-cell observation $\mathbf{Y^{(10)}}$ is generated as $\mathbf{Y^{(10)}} = \log_2 \sum_{i=1}^{10} 2^{\mathbf{Y}_i}/10$ and $\mathbf{Y}$ is generated as under model 1-3.
5. 1-3 with non-zero observations contaminated with t_8 noise.
6. 1-3 with the following latent Gaussian/logistic selection model:

$$\begin{aligned} \tilde{\mathbf{Y}} &\sim \mathcal{N}(\mu, \mathbf{K}), \\ P\left(\tilde{V}_j | \tilde{\mathbf{Y}} = \tilde{\mathbf{y}}\right) &= \text{logit}(a + b\tilde{y}_j), \\ \mathbf{Y} &= \tilde{\mathbf{Y}}\tilde{V}. \end{aligned} \tag{3}$$

Methods

1. Aracne [Margolin et al., 2006]: connects genes with significant pairwise mutual information and applies pruning rules to suppress indirect effects.
2. Gaussian: neighborhood selection with ℓ_1-penalized linear regression [Meinshausen and Bühlmann, 2006].
3. Logistic: neighborhood selection with ℓ_1-penalized logistic regression[Ravikumar et al., 2010].
4. NPN: neighborhood selection with ℓ_1-penalized linear regression on Gaussian-quantile transformed responses [Liu et al., 2009].
5. Hurdle (isometric): neighborhood selection with model (6) and isometric group-lasso penalty.
6. Hurdle (anisometric): neighborhood selection with model (6) and anisometric group-lasso penalty.

Supplemental Table 1

Overview of simulation scenarios and methods compared.

Araceli M. Huerta, Alfredo Mendoza-Vargas, Leonardo Collado-Torres, Blanca Taboada, Leticia Vega-Alvarado, Maricela Olvera, Leticia Olvera, Ricardo Grande, Enrique Morett, and Julio Collado-Vides. RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research*, 39(SUPPL. 1), 2011.

Han Liu, John Lafferty, and Larry Wasserman. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10(10):2295–2328, 2009.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, mar 2006.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

C. Radhakirshna Rao. *Linear Statistical Inference and its Applications*. John Wiley & Sons, Ltd., 2nd edition, 1973.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, jun 2010.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

E-MAIL: andrew_mcdavid@urmc.rochester.edu

E-MAIL: rgottard@fredhutch.org

E-MAIL: nrsimon@uw.edu

E-MAIL: md5@uw.edu