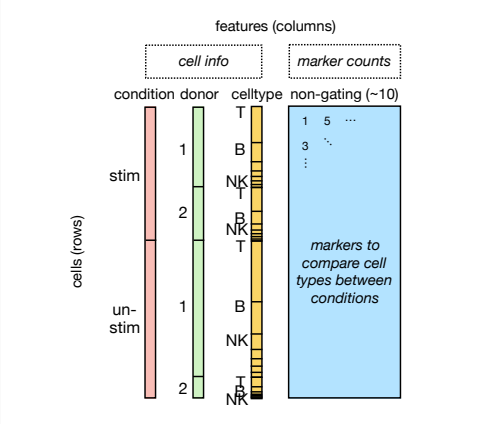
Uncertainty Quantification in Multivariate Mixed Models for Mass Cytometry Data

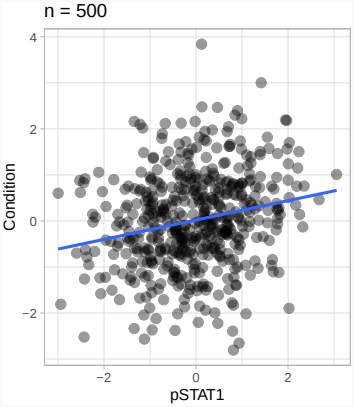
Christof Seiler
Assistant Professor of Statistics
Department of Data Science
and Knowledge Engineering
Maastricht University, The Netherlands
http://christofseiler.github.io

Second Dutch Stan Meetup 2019, Utrecht

Mass Cytometry Data

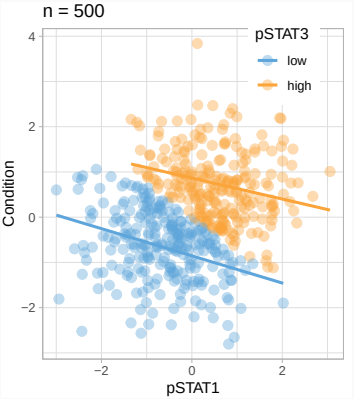


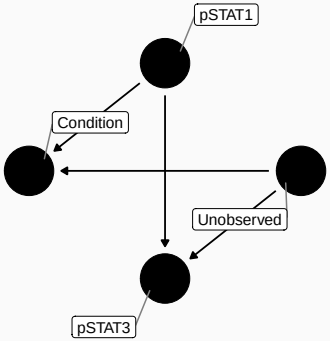
- experimental condition → non-gating markers or
- non-gating markers → experimental condition?




```
lm(formula = Condition ~ pSTAT1, data = cytof_data) %>% tidy
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)   0.0208    0.0453     0.460 0.646
## 2 pSTAT1        0.211     0.0429     4.91  0.00000122
```

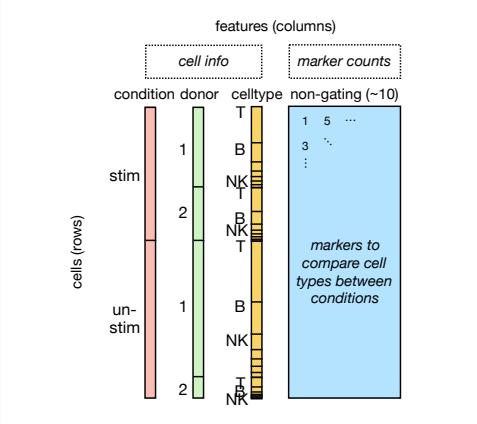





```
lm(formula = Condition ~ pSTAT1 + pSTAT3, data = cytof_data) %>% tidy
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -0.834    0.0521    -16.0  8.78e-47
## 2 pSTAT1        -0.269    0.0385     -6.99 8.90e-12
## 3 pSTAT3high     1.72     0.0814     21.1  3.01e-71
```

Mass Cytometry Data



- Rows are **clustered** and columns are **correlated**

Clustered Data

⋮

Robust methods (Huber 1964, 1973)

Varying coefficients (Hastie and Tibshirani 1993)

Mixed effects (Pinheiro and Bates 2000)

Maximin effects (Meinshausen and Bühlmann 2015)

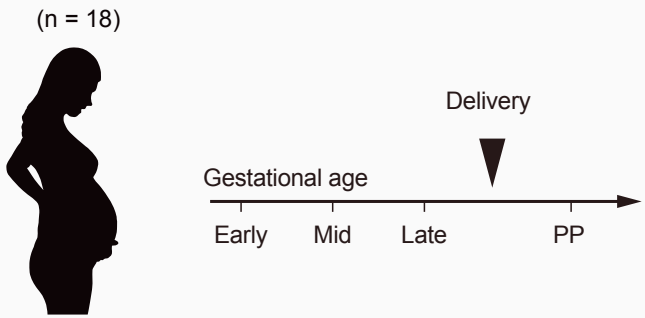
Mixtures (Tutz and Oelker 2017)

Tree-structured (Berger and Tutz 2018)

Anchor regression (Rothenhäusler et al. 2018)

⋮

Pregnancy Study by Aghaeepour et al. (2017)



- Blood samples collected during and after **pregnancy**
- **Protein expression** measured using mass cytometry (INF$-\alpha$ stimulated)
- **Differential analysis** between first and third trimester

Multivariate Poisson Log-Normal Model with Zero Inflation

Zero inflation:

- $y_{i,j}$ are counts in cell i of protein j
- For zero counts, flip a biased coin which lands Heads with probability θ
- If it comes up Heads, then set $y_{i,j} = 0$, otherwise

$$y_{i,j} \sim \text{Poisson}(\lambda_{i,j})$$

Multivariate Poisson (Chib and Winkelmann 2001):

$$\log(\lambda_{i,j}) = \beta_{\text{cond}[i],j} + b_{i,j} + u_{\text{donor}[i],j}$$

- $\boldsymbol{b}$ cell mixed effect and $\boldsymbol{u}$ donor mixed effect
- $\text{cond}[i] = 1$: first trimester and $\text{cond}[i] = 2$: third trimester

Model for Correlations

- **Cell-to-cell** variability (J total number of protein markers):

$$\begin{pmatrix} b_{i,1} \\ \vdots \\ b_{i,J} \end{pmatrix} \sim \text{Multivariate Normal}\left(\mathbf{0}, \text{diag}(\boldsymbol{\sigma})\, \boldsymbol{\Omega}\, \text{diag}(\boldsymbol{\sigma})\right)$$

- **Donor-to-donor** variability ($\boldsymbol{V}$ is a orthogonal matrix):

$$\begin{pmatrix} u_{k,1} \\ \vdots \\ u_{k,J} \end{pmatrix} \sim \text{Multivariate Normal}\left(\mathbf{0}, \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T\right)$$

Priors

- **Condition** regression coefficients:
  Weakly informative prior (typical counts range from 10^{-4} to 1096)

$$\boldsymbol{\beta} \sim \text{Normal}(0, 7^2)$$

- **Cell-to-cell** variability:
  Full rank covariance matrix distribution (Lewandowski, Kurowicka, and Joe 2009)

$$\boldsymbol{\sigma} \sim \text{Half-Cauchy}(0, 2.5)$$

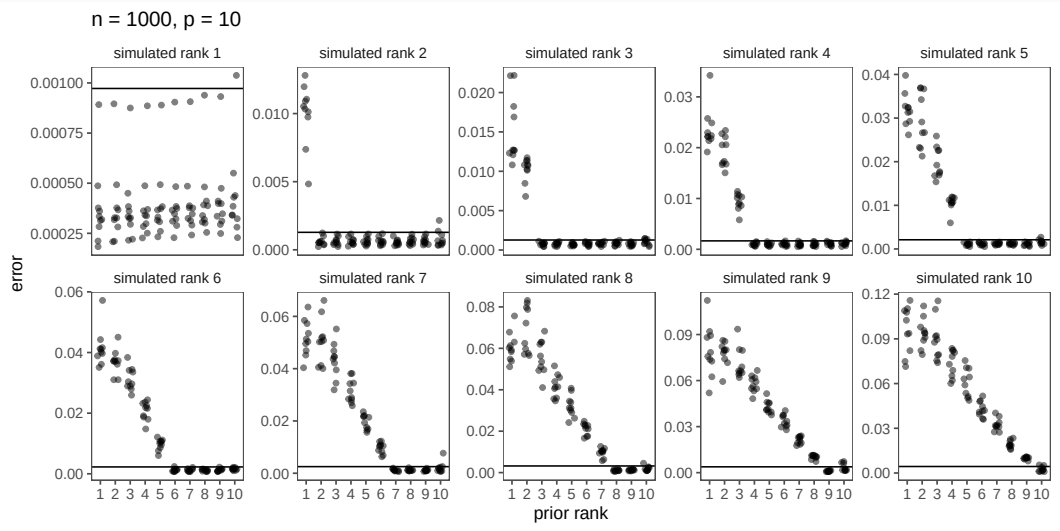
$$\boldsymbol{\Omega} \sim \text{Uniform correlation matrix}$$

- **Donor-to-donor** variability:
  Low rank covariance matrix distribution (Easton 1989; Jauch, Hoff, and Dunson 2019)

$$\text{diag}(\boldsymbol{D}) \sim \text{Half-Cauchy}(0, 2.5)$$

$$\boldsymbol{V} \sim \text{Uniform orthogonal matrix}$$

Priors: Choosing the Rank



Posterior Inference

- Parameters of interest:

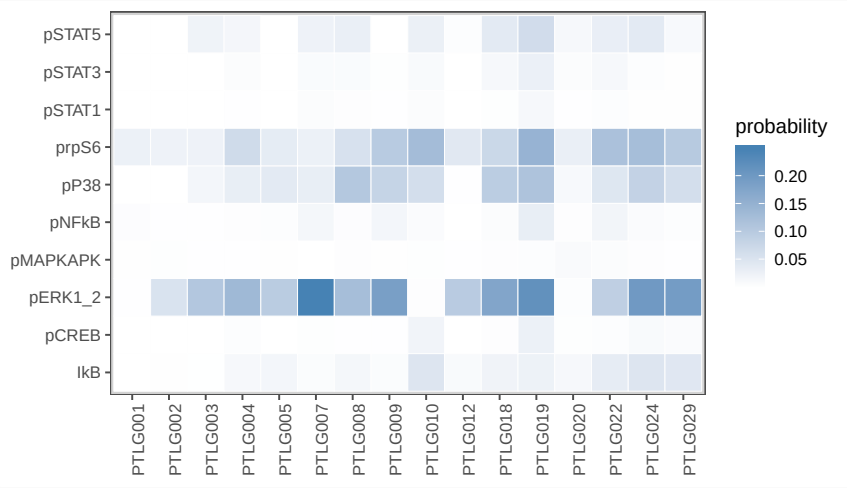
$$\textbf{pars} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\Omega}, \boldsymbol{V}, \boldsymbol{D}\}$$

- Sample from **posterior distribution** using Stan (Carpenter et al. 2017):

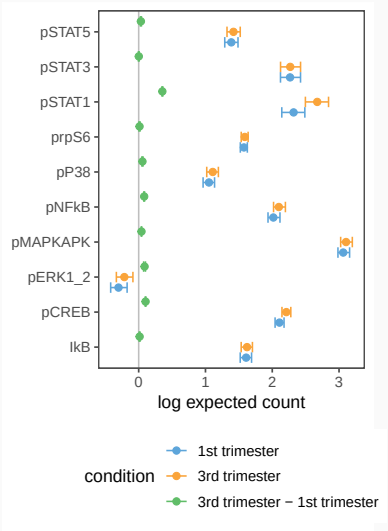
$$p(\textbf{pars}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\textbf{pars})\, p(\textbf{pars})}{p(\boldsymbol{y})}$$

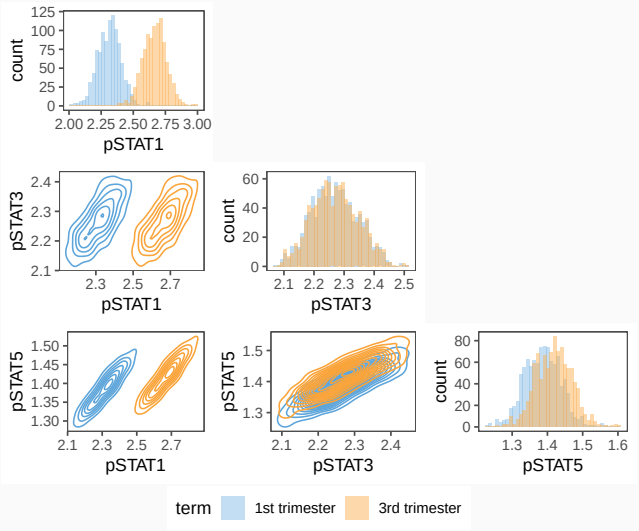
- Complete Stan model available on GitHub: poisson.stan
- **Summarize** posterior samples (e.g. median, credible intervals, MDS)

Posterior Summaries: θ

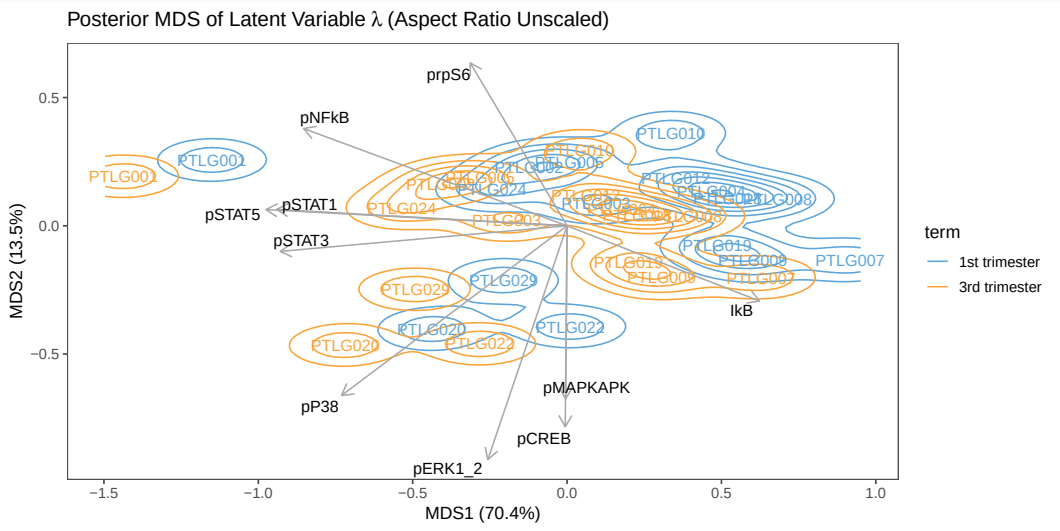


Posterior Summaries: β

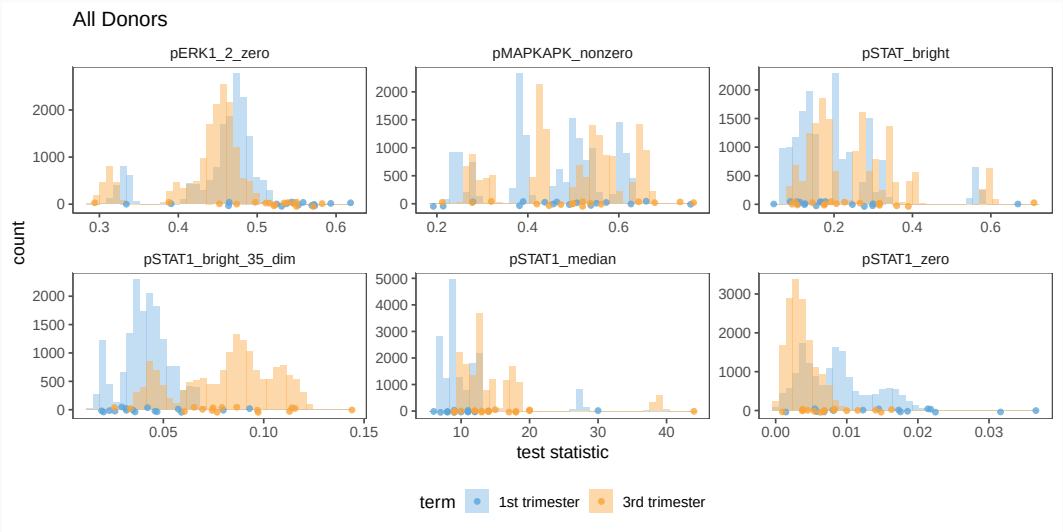




Posterior Summaries: λ



Goodness of Fit: Posterior Predictive Checks



Related Workflows/Packages

1. **Response variable**: Mean counts and cell type abundance
   - R package **diffcyt** (Weber et al. 2018):
     - High-resolution clustering and empirical Bayes moderated tests adapted from transcriptomics
   - F1000 CyTOF Workflow (Nowicka et al. 2017):
     - Manual gating and univariate analyses
2. **Response variable**: experimental condition
   - R package **Citrus** (Bruggner et al. 2014):
     - Hierarchical clustering and regularized regression to select predictive features
   - Python package **CellCnn** (Arvaniti and Claassen 2017):
     - Convolutional neural networks to detect rare cell populations

Take Home

- **Multivariate models**
  - describe marker correlations
  - avoid biases
- **Mixed models**
  - describe individual donor effects
  - avoid reporting overconfident results
- R package **cytoeffect** with vignettes:
  https://christofseiler.github.io/cytoeffect/
- **Preprint**: Seiler et al. (2019)

Acknowledgements

- **Holmes Lab (Stanford):** Susan Holmes, Simon Rubinstein-Salzado, Lan Huong Nguyen, Kris Sankaran, Julia Fukuyama, Pratheepa Jeganathan, Claire Donnat, Joey McMurdie, and Benjamin Callahan
- **Blish Lab (Stanford):** Catherine Blish, Lisa Kronstad, Laura Simpson, Mathieu Le Gars, and Elena Vendrame
- **Funding:** Swiss NSF, NIH, and CZI

Thanks for your attention!

References i

Aghaeepour, Nima, Edward A. Ganio, David Mcilwain, Amy S. Tsai, Martha Tingle, Van GassenSofie, Dyani K. Gaudilliere, et al. 2017. "An Immune Clock of Human Pregnancy." *Science Immunology* 2 (15): eaan2946. https://doi.org/10.1126/sciimmunol.aan2946.

Arvaniti, Eirini, and Manfred Claassen. 2017. "Sensitive Detection of Rare Disease-Associated Cell Subsets via Representation Learning." *Nature Communications* 8: 14825.

Berger, Moritz, and Gerhard Tutz. 2018. "Tree-Structured Clustering in Fixed Effects Models." *Journal of Computational and Graphical Statistics*, 1–13.

Bruggner, Robert V, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. 2014. "Automated Identification of Stratifying Signatures in Cellular Subpopulations." *Proceedings of the National Academy of Sciences* 111 (26): E2770–E2777.

References ii

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Chib, Siddhartha, and Rainer Winkelmann. 2001. "Markov Chain Monte Carlo Analysis of Correlated Count Data." *Journal of Business & Economic Statistics* 19 (4): 428–35.

Easton, Morris L. 1989. "Chapter 7: Random Orthogonal Matrices." In *Group Invariance in Applications in Statistics*, Volume 1:100–107. Regional Conference Series in Probability and Statistics. Haywood CA; Alexandria VA: Institute of Mathematical Statistics; American Statistical Association. https://projecteuclid.org/euclid.cbms/1462061037.

Hastie, Trevor, and Robert Tibshirani. 1993. "Varying-Coefficient Models." *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–96.

References iii

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35 (1): 73–101.

———. 1973. "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *The Annals of Statistics* 1 (5): 799–821.

Jauch, Michael, Peter D Hoff, and David B Dunson. 2019. "Monte Carlo Simulation on the Stiefel Manifold via Polar Expansion." *arXiv Preprint arXiv:1906.07684*.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100 (9): 1989–2001.

Meinshausen, Nicolai, and Peter Bühlmann. 2015. "Maximin Effects in Inhomogeneous Large-Scale Data." *The Annals of Statistics* 43 (4): 1801–30.

References iv

Nowicka, M, C Krieg, LM Weber, FJ Hartmann, S Guglietta, B Becher, MP Levesque, and MD Robinson. 2017. "CyTOF Workflow: Differential Discovery in High-Throughput High-Dimensional Cytometry Datasets [Version 2; Referees: 2 Approved]." *F1000Research* 6 (748). https://doi.org/10.12688/f1000research.11622.2.

Pinheiro, José C, and Douglas M Bates. 2000. "Linear Mixed-Effects Models: Basic Concepts and Examples." *Mixed-Effects Models in S and S-Plus*, 3–56.

Rothenhäusler, Dominik, Peter Bühlmann, Nicolai Meinshausen, and Jonas Peters. 2018. "Anchor Regression: Heterogeneous Data Meets Causality." *arXiv Preprint arXiv:1801.06229*.

Seiler, Christof, Lisa M Kronstad, Laura J Simpson, Mathieu Le Gars, Elena Vendrame, Catherine A Blish, and Susan Holmes. 2019. "Uncertainty Quantification in Multivariate Mixed Models for Mass Cytometry Data." *arXiv Preprint arXiv:1903.07976*.

Tutz, Gerhard, and Margret-Ruth Oelker. 2017. "Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures." *International Statistical Review* 85 (2): 204–27.

Weber, Lukas M, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson. 2018. "Diffcyt: Differential Discovery in High-Dimensional Cytometry via High-Resolution Clustering." *bioRxiv*. https://doi.org/10.1101/349738.