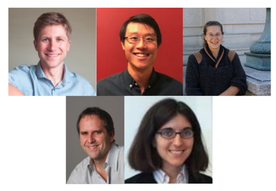
Measuring Cluster Stability for Bayesian Nonparametrics Using the Linear Bootstrap



Ryan Giordano[1*] Runjing Liu[1*] Nelle Varoquaux[1*]
Michael I. Jordan[1] Tamara Broderick[2]

[*] These authors contributed equally

[1] Department of Statistics, UC Berkeley [2] Department of EECS, MIT

Overview

- We employ a **Bayesian nonparametric** model to cluster time-course gene expression data, and do inference using **mean field variational Bayes**.
- To assess the **clustering stability** of our results, one approach is to do **bootstrap sampling**. However, this is computationally expensive and requires fitting new VB parameters to each simulated data-set.
- Therefore, we propose a *fast, automatic approximation* to a full bootstrap analysis based on the **infinitesimal jackknife** [1]. We call this alternative bootstrap analysis the **linear bootstrap**.

Data

We study data from [4] wherein mice were infected with influenza virus, and gene expressions were measured at 14 time points after infection.

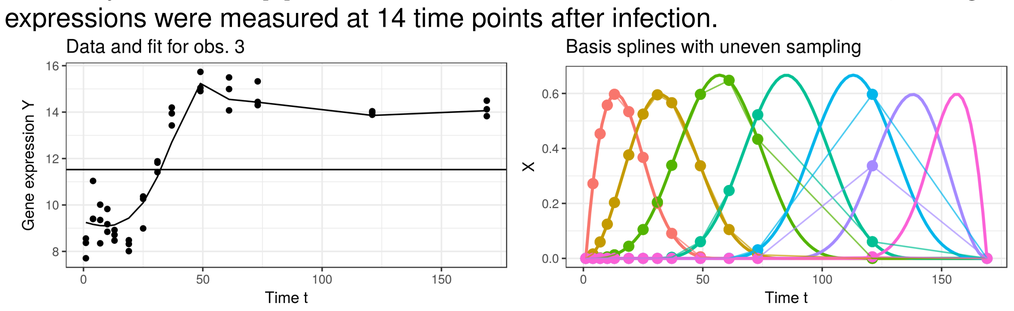


Figure 1: *The time-course gene expression data (left) and the B-splines basis (right).*

We cluster genes based on their time-course gene expression:

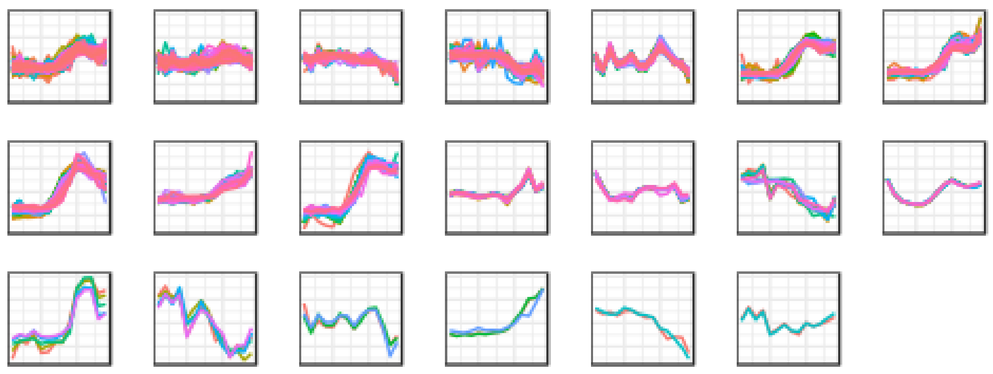


Figure 2: *Clusters recovered by our BNP model. Each curve is the measured time-course expression pattern of one gene.*

Model

We show a graphical representation of our model below:

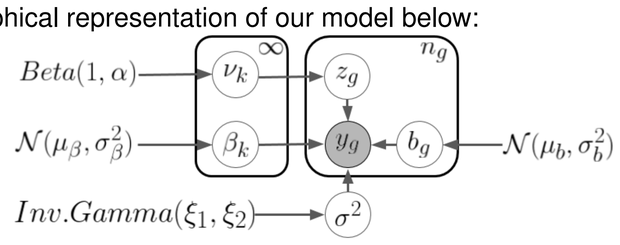


Figure 3: *y_g is the expression level for gene g, and each gene has a unique shift b_g. σ^2 is the common variance for all genes about its B-spline fit. Each cluster is defined by its B-spline coefficients β_k, which are drawn from a Dirichlet process mixture model.*

Results

A cold start refers to doing 10 random restarts for each bootstrap sample. A warm start refers to starting from a (high quality) optimum found for the full data set.

Speed comparisons:

	Total time (sec) (200 bootstrap samples)	Time per fit (sec)
Initial fit (200 random restarts)	–	16100
Full bootstrap (cold start)	184000	931
Full bootstrap (warm start)	10800	53.4
Hessian inverse (for linear bootstrap)	–	12.7
Linear bootstrap (given Hessian inverse)	0.0284	0.000145

- **The linear bootstrap is orders of magnitudes faster.**
- The full bootstrap requires re-optimizing, while the linear approximation requires a one time computation and factorization of the KL Hessian [2].
- The KL Hessian can be easily computed with modern auto-differentiation tools [3].

Accuracy comparisions:

We compare distributions of cluster similarity under the full and the linear bootstrap.

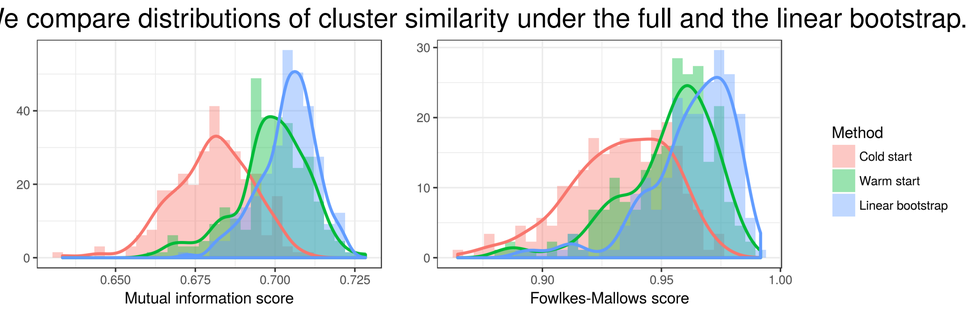


Figure 4: *We examine the Fowlkes-Mallows index and the normalized mutual information score as a measure of clustering similarity.*

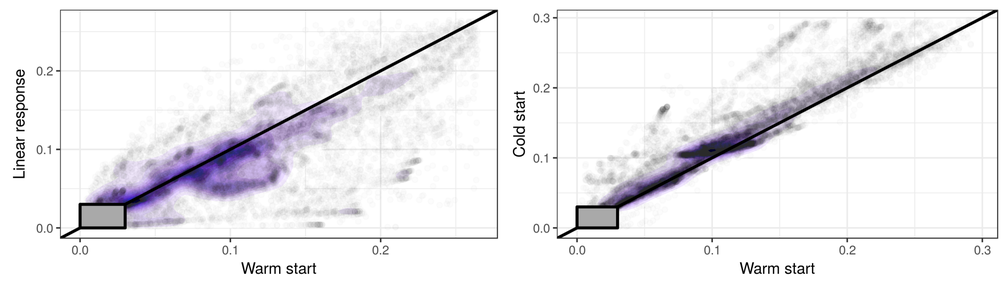


Figure 5: *Standard deviations of elements of the co-clustering matrix for a randomly selected subset of genes. Pairs with both standard deviations < 0.03 on both axes are not shown.*

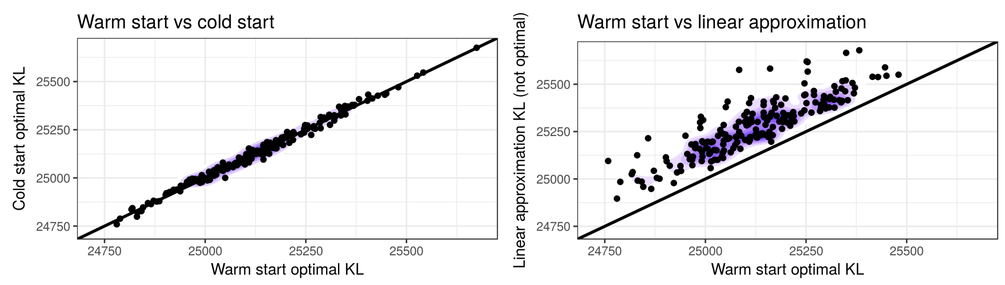


Figure 6: *Distribution of KL divergence relative to the warm start*

The linear bootstrap

- Augment the model by adding per-gene weights $W = (w_1, ..., w_{n_g}), w_i \geq 0$.
- The likelihood becomes $\sum_g w_g \log p\left(y_g|\theta\right)$.
- The optimal variational parameter $\eta^*(W)$ is now a function of W.
- Note letting $W = W_1 := (1, ..., 1)$ we recover the original fit on the whole data set.
- A bootstrap sample draws $W_b \sim Multinomial(n_g, n_g^{-1})$, and evaluates $\eta^*(W_b)$.
- The **infinitesimal jackknife** approximates this with a first order Taylor expansion:

$$\eta^*(W_b) \approx \eta^*_{lin}(W_b) := \eta^*(W_1) + \frac{d\eta^*}{dW}|_{W_1}(W_b - W_1)$$

- Auto-differentiation and numerical linear algebra software can evaluate $\frac{d\eta^*}{dW}|_{W_1}$.
- For a stability measure ϕ, we approximate the full bootstrap distribution $\phi(\eta^*(W_b))$ using $\phi(\eta^*_{lin}(W_b))$.
- We call $\phi(\eta^*_{lin}(W_b))$ **the linear bootstrap**. Note that, unlike the infinitesimal jackknife, we do not assume linearity of ϕ.

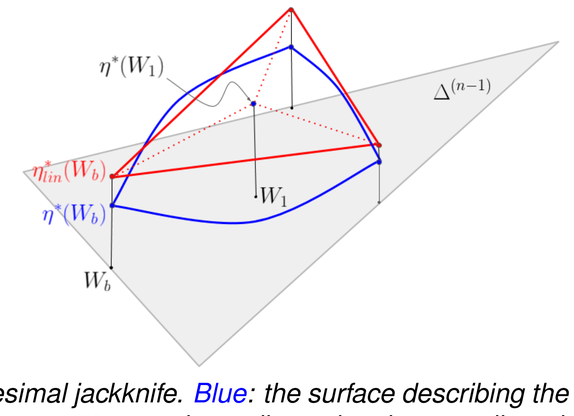


Figure 7: *The infinitesimal jackknife. Blue: the surface describing the parameter η as a function of the weights vector W over the n-dimensional resampling simplex Δ^{n-1}. Red: the linear approximation, i.e., the tangent plane at the original weight vector $W_1 = (1, ..., 1)$.*

Conclusion

- The linear bootstrap is a fast and reasonable alternative to the full bootstrap.
- However, the linear bootstrap underestimates the variance of the full bootstrap.
- A main reason for the underestimation is the presence of local optima.

References

[1] B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[2] R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness, and variational bayes. *arXiv preprint arXiv:1709.02536*, 2017.

[3] D. Maclaurin, D. Duvenaud, and R. P. Adams. Autograd: Effortless gradients in numpy. In *International Conference on Machine Learning 2015 AutoML Workshop*, 2015.

[4] J. E. Shoemaker, S. Fukuyama, A. J. Eisfeld, D. Zhao, E. Kawakami, S. Sakabe, T. Maemura, T. Gorai, H. Katsura, Y. Muramoto, S. Watanabe, T. Watanabe, K. Fuji, Y. Matsuoka, H. Kitano, and Y. Kawaoka. An Ultrasensitive Mechanism Regulates Influenza Virus-Induced Inflammation. *PLoS Pathogens*, 11(6):1–25, 2015.

Acknowledgments: Ryan Giordano and Nelle Varoquaux's research was funded in full by the Gordon and Betty Moore Foundation through Grant GBMF3834 and by the Alfred P. Sloan Foundation through Grant 2013-10-27 to the University of California, Berkeley. Runjing Liu's research was funded by the NSF graduate research fellowship. Tamara Broderick's research was supported in part by a Google Faculty Research Award and the Office of Naval Research under contract/grant number N00014-17-1-2072.

Contact: rgiordano@berkeley.edu, runjing_liu@berkeley.edu, nelle@berkeley.edu