Statistical Connectomics

some, jovo,* other cep,† souls

September 11, 2014

Contents

Abstract

*yummy
†t

1 Introduction

potential neuro co-authors could include: mike milham, scott cook, mitya, bobby/jeff, clay/davi, rex jung,

we start by stating how important connectomics will be for the future of neuroscience, and how having rigorous statistical theory will enable future investigations to leverage it to substantiate their claims.

for each exploitation task, we provide:

1. rigorous definition
2. motivating application
3. R code
4. images, graphs, and graph derivatives downloads

2 Background

In classical Euclidean statistics, we we obtain n data points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. To perform *any* statistical inference, fundamental to any scientific inquiry, we (sometimes implicitly) make a number of assumptions. First, we assume each observation is a realization of a random variable that has some distribution, F which is an element of a family of distributions, $\mathcal{F}$. For example, a data point might be d-dimensional vectors, $\boldsymbol{x}_i \in \mathbb{R}^d$, F could be a multivariate Gaussian distribution, $F = \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathcal{F}$ could be the set of all possible d-dimensional Gaussian distributions, $\mathcal{F} = \{\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \Xi_d\}$, where Ξ_d denotes the set of $d \times d$ dimensional covariance matrices. Second, we assume that each data point, $\boldsymbol{x}_i$ is sampled independently and indentically (iid) from F. We denote these assumptions by writing $\boldsymbol{x}_i \overset{iid}{\sim} F \in \mathcal{F}$.

The collection of data points, $\mathcal{D}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, sometimes referred to as a point cloud, is the object upon which all statistical inference follows, including one-sample tests and unsupervised learning tasks such as mean and quantile estimation, clustering, and vector quantization. Sometimes, associated with each x_i, we know something else about instance i, which we denote y_i, yielding n tuples, $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. The y_i's could denote which condition x_i was observed under, which population instance i comes from, etc. In such settings, $y_i \in \{\widetilde{y}_1, \ldots, \widetilde{y}_C\}$ is a categorical random variable, and inference tasks of interest include 2-/multi-sample testing, and classification. If the y_i's are continuous valued, then $y_i \in \mathbb{R}$, then we have regression problems, and if y_i's are multivariate, then we have multivariate regression.

As mentioned above, many of the classical tests and estimation procedures assume that $\mathcal{D}_n$ is a collection of iid samples. In connectomics, $\mathcal{D}_n$ is *not* that, rather $\mathcal{D}_n$ is a network (or graph), G with n nodes (or vertices) and up to n^2 links (or edges, arcs). We can think of a graph as a tuple, $G = (V, E)$, where V is the set of n nodes, and E is the set of up to n^2 edges. Thus, a graph, like a point cloud, is a collection of objects. However, unlike a typical point cloud, it does not make much sense to assume objects are sampled identically and independently, except for the simplest possible random graph model. For some random graph models, we think of edges as the random variables, and assume that some are independent, but not identically sampled. Alternately, for other random graph models, we think of the vertices as random variables, and we may assume they are sampled iid, but we do not "see" the vertices, rather, we only see edges amongst the vertices. In this case, we posit the existence of latent (hidden) variables associated with each vertex. Thus, all the random graph models that we will use, our assumptions will have to relax either independence, identical, or introduce latent variables. This is similar in spirit to time-series and spatial modeling, where the independence assumptions break down.

A side-effect of relaxing the fundamental assumptions underlying our data analysis is that neither the theoretical guarantees that motivate the beloved procedures we learned from classic texts such as (1) or (2), nor the codes we have from MATLAB or R make sense out of the box. In this manuscript, we describe theoretically justified modifications/generalizations of all the above mentioned inference tasks. Moreover, we provide datasets and example code for implementing these procedures in R. We hope this helps to foster a growing community of methods and applications to this emerging field of statistical connectomics.

3 One Sample Tests

A classical one sample test is defined as followed: Given $\mathcal{D}_n$, test whether $x_i \overset{iid}{\sim} F$, for some given F. For example, imagine that we want to test whether $\mathcal{D}_n$ is a sample from a mixture of two multivariate Gaussians. Let $\mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denote a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$, and let ω_j be the weight of Gaussian j, where each $\omega_j \geq 0$ and $\sum_j \omega_j = 1$. Let $\mathcal{H}_0$ denote the set of mixtures of two d-dimensional Gaussians, and let $\mathcal{H}_A$ denote all other distributions. We can therefore define the one-sample test:

$$H_0 \colon \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{iid}{\sim} F \in \mathcal{H}_0 \tag{1}$$

$$H_A \colon \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{iid}{\sim} F' \in \mathcal{H}_A, \tag{2}$$

To conduct this test, we must first choose a test-statistic, $T_n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, to measure the goodness-of-fit of the data to the model, and a way of estimating the null distribution. We choose likelihood, and adopt a parametric bootstrap to estimate the null distribution. The likelihood of a data point, given a distribution F, is denoted, $\ell(x_1; F)$. Let the likelihood of x_i being sampled from the j^{th} Gaussian distribution be $\phi_j(x_i) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}\big(x_i-\mu)^\intercal\boldsymbol{\Sigma}^{-1}(x_i-\mu)\big)$. For n samples, each sampled independently and identically from the same distribution, the likelihood of the dataset is given simply by $\ell(\mathcal{D}_n; F) = \prod_i^n \ell(x_i; F)$. The steps of a parametric boostrap under a Gaussian mixture model are described in Pseudocode 1.

Pseudocode 1 Parametric bootstrap test for a mixture of Gaussians

Input: $\{X_1, \ldots, X_n\}$
Output: p-value
1: Obtain an estimate $\hat{F}$ of the Gaussian mixture model (say, using `mclust`)
2: Compute the likelihood of the observed data, given the estimated model, $\ell(\mathcal{D}_n; \hat{F})$.
3: **for** $i \in \{1, 2, \ldots, 1000\}$ **do**
4: Sample $\mathcal{D}_n^{(i)}$, that is, n samples from $\hat{F}$
5: Compute $\ell^{(i)}(\mathcal{D}_n^{(i)}; \hat{F})$
6: **end for**
7: Compute, $\hat{F}_\ell$, empirical cumulative distribution of $\ell^{(i)}$
8: Let p-value equal $F_\ell(\ell(\mathcal{D}_n; \hat{F}))$

The homologue of this test for connectomics is the following. Given a connectome, represented purely by its vertex set and connectivity structure, $G = (V, E)$, where G is a graph comprised of an vertex set V and a set of edges amongst them E, we desire to determine

$$H_0 \colon G \sim F \in \mathcal{H}_0 \tag{3}$$

$$H_A \colon G \sim F' \in \mathcal{H}_A, \tag{4}$$

3.1 tests for model fit

hsbm on fly optic lobe data (3) or c elegans.

3.2 tests for independence between connectivity and vertex attributes (such as direction preference, excitatory vs. inhibitory, etc.)

bock11 (4) dataset, testing independence of tuning direction vs connectivity, using residual error of regression o ase as test statistic, permutation test to obtain null

3.3 tests for independence between space and connectivity

kasthuri11 dataset (no cite yet, coming soon), touches vs. synapses, using whatever we do (probably importance sampling to obtain null distribution)

4 2-sample tests for comparative connectomics

4.1 comparing 2 different connectomes

elegans electrical vs. chemical & elegans vs. pacificus & elegans male vs. herm. See (5) for the most clear description of these graphs.

4.2 2 populations of connectomes

(6, 7) describes two different populations of subjects collected for two different studies, both of which are useful.

5 Connectome Unsupervised Learning

5.1 mean estimation

(8, 9) are two papers proving that Stein's paradox does not occur in finite spaces, in other words, $\bar{A}$ is admissible under squared error loss. nonetheless, it seems likely that some smoothing/regularizing of $\bar{A}$ would be advantageous for finite sample sizes. in particular, spectral and constrained estimates of latent vectors. we can use any number of MR datasets, such as those MRN-111 in (10).

5.2 robust mean (eg, median, or Lq) estimation

we can again use the MRN-111 dataset, the theory is motivated by (11, 12).

5.3 Clustering

using tensor factorizations (13–15), or DELTACON (16, 17), which is just hclust with a different dissimilarity function.

5.4 errorbars around mean estimation, eg, estimation variance

bayesian nonparametric model (18)

5.5 Canonical Cortical Circuits

tensor factorization of NKI-Enhanced (13–15).

6 connectome coding

6.1 classifying connectomes

signal subgraphs paper (19), or using ASE or tensor factorization, followed by classical classification.

6.2 regressing connectomes

MRN114 via NTF followed by regression onto CCI

6.3 multivariate regression for connectomes

Adelstein (20) using JoFC on 5-factor personality test.

7 Discussion

general issues:

7.1 bias variance trade-off: num params $>$ num subjects

7.2 nuisance signals: age, sex, batch

7.3 Graph Matching

oh, which papers to list, how about (21–24)

7.4 Future Work

Acknowledgments

The author thanks the anonymous authors whose work largely constitutes this sample file. He also thanks the INFO-TeX mailing list for the valuable indirect assistance he received.

8 Bibliography

[1] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *BeiJing: Publishing House of Electronics Industry*, 2004. 2

[2] J. A. Rice, *Mathematical statistics and data analysis*. Duxbury Press, 1995. [Online]. Available: http://www.citeulike.org/user/tarjeiha/article/1691927 2

[3] S.-Y. Takemura, A. Bharioke, Z. Lu, A. Nern, S. Vitaladevuni, P. K. Rivlin, W. T. Katz, D. J. Olbris, S. M. Plaza, P. Winston, T. Zhao, J. A. Horne, R. D. Fetter, S. Takemura, K. Blazek, L.-A. Chang, O. Ogundeyi, M. A. Saunders, V. Shapiro, C. Sigmund, G. M. Rubin, L. K. Scheffer, I. A. Meinertzhagen, and D. B. Chklovskii, "A visual motion detection circuit suggested by Drosophila connectomics," *Nature*, vol. 500, no. 7461, pp. 175–181, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1038/nature12450 3

[4] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. Andermann, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, G. Hood, and R. C. Reid, "Network anatomy and in vivo physiology of visual cortical neurons," *Nature*, vol. 471, no. 7337, pp. 177–182, Mar. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3095821&tool=pmcentrez&rendertype=abstract 3

[5] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, D. B. Chklovskii, C. Spring, and J. Farm, "Structural Properties of the Caenorhabditis elegans Neuronal Network," *PLoS Computational Biology*, vol. 7, no. 2, pp. 1–41, Feb. 2011. 3

[6] K. B. Nooner, S. J. Colcombe, R. H. Tobe, M. Mennes, M. M. Benedict, A. L. Moreno, L. J. Panek, S. Brown, S. T. Zavitz, Q. Li, S. Sikka, D. Gutman, S. Bangaru, R. T. Schlachter, S. M. Kamiel, A. R. Anwar, C. M. Hinz, M. S. Kaplan, A. B. Rachlin, S. Adelsberg, B. Cheung, R. Khanuja, C. Yan, R. C. Craddock, V. D. Calhoun, W. Courtney, M. King, D. Wood, C. L. Cox, A. M. C. Kelly, A. Di Martino, E. Petkova, P. T. Reiss, N. Duan, D. Thomsen, B. Biswal, B. Coffey, M. J. Hoptman, D. C. Javitt, N. Pomara, J. J. Sidtis, H. S. Koplewicz, X. F. Castellanos, B. L. Leventhal, and M. P. Milham, "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry." *Frontiers in neuroscience*, vol. 6, p. 152, Jan. 2012. [Online]. Available: http://www.frontiersin.org/Brain_Imaging_Methods/10.3389/fnins.2012.00152/full 4

[7] B. a. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. a. D. Farrell, J. a. Bogovic, J. Hua, M. Chen, S. Jarso, S. a. Smith, S. Joel, S. Mori, J. J. Pekar, P. B. Barker, J. L. Prince, and P. C. M. van Zijl, "Multi-Parametric Neuroimaging Reproducibility: A 3T Resource Study." *NeuroImage*, Nov. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21094686 4

[8] B. M. Johnson, "On the Admissible Estimators for Certain Fixed Sample Binomial Problems," *The Annals of Mathematical Statistics*, vol. 42, no. 5, pp. 1579–1587, Oct. 1971. [Online]. Available: http://projecteuclid.org/euclid.aoms/1177693156 4

[9] S. Gutmann, "Stein's Paradox is Impossible in Problems with Finite Sample Space," *The Annals of Statistics*, vol. 10, no. 3, pp. 1017–1020, Sep. 1982. [Online]. Available: http://projecteuclid.org/euclid.aos/1176345893 4

[10] W. Gray Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, A. R. Bowles, D. K. Donavos, S. Ryman, R. E. Jung, L. Wu, V. D. Calhoun, and R. J. Vogelstein, "MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics," *Global Conference on Signal and Information Processing*, 2013. [Online]. Available: http://arxiv.org/abs/1312.4875 4

[11] D. Ferrari and Y. Yang, "Maximum L q -likelihood estimation," *The Annals of Statistics*, vol. 38, no. 2, pp. 753–783, Apr. 2010. [Online]. Available: http://projecteuclid.org/euclid.aos/1266586613 4

[12] Y. Qin and C. E. Priebe, "Maximum L q -Likelihood Estimation via the Expectation-Maximization Algorithm: A Robust Estimation of Mixture Models," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 914–928, Sep. 2013. [Online]. Available: http://www.tandfonline.com.proxy3.library.jhu.edu/doi/abs/10.1080/01621459.2013.787933#.UuJtQmQo7og 4

[13] N. H. Lee, C. E. Priebe, R. Tang, and M. Rosen, "Using non-negative factorization of time series of graphs for learning from an event-actor network," Dec. 2013. [Online]. Available: http://arxiv.org/abs/1312.7559 4

[14] N. H. Lee, I.-J. Wang, Y. Park, C. E. Priebe, and M. Rosen, "Automatic Dimension Selection for a Non-negative Factorization Approach to Clustering Multiple Random Graphs," Jun. 2014. [Online]. Available: http://arxiv.org/abs/1406.6315

[15] N. H. Lee, I.-J. Wang, R. Tang, M. Rosen, and C. E. Priebe, "A rank estimation criterion using an NMF algorithm under an inner dimension condition," Jun. 2014. [Online]. Available: http://arxiv.org/abs/1406.6319 4

[16] D. Koutra, J. T. Vogelstein, and C. Faloutsos, "DeltaCon: Measuring Connectivity Differences in Large Networks," in *SIAM International Conference on Data Mining*, 2013. [Online]. Available: http://knowledgecenter.siam.org/105SDM/1 4

[17] D. Koutra, N. Shah, J. T. Vogelstein, B. J. Gallagher, and C. Faloutsos, "DELTACON: A Principled Massive-Graph Similarity Function and Applications," *in preparation*. 4

[18] D. Durante, D. B. Dunson, and J. T. Vogelstein, "Nonparametric Bayes Modeling of Populations of Networks," Jun. 2014. [Online]. Available: http://arxiv.org/abs/1406.7851 4

[19] J. T. Vogelstein, W. Gray Roncal, R. J. Vogelstein, and C. E. Priebe, "Graph Classification using Signal Subgraphs: Applications in Statistical Connectomics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1539 – 1551, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6341752 4

[20] J. S. Adelstein, Z. Shehzad, M. Mennes, C. G. DeYoung, X.-N. Zuo, C. Kelly, D. S. Margulies, A. Bloomfield, J. R. Gray, F. X. Castellanos, and M. P. Milham, "Personality is reflected in the brain's intrinsic functional architecture," *PLoS ONE*, vol. 6, no. 11, 2011. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-82455217247&partnerID=40&md5=15a8e206a9ebdcbe0e137dd8a34cc332 4

[21] J. T. Vogelstein, J. C. M. Conroy, L. J. Podrazik, S. G. Kratzer, D. E. Fishkind, R. J. Vogelstein, and C. E. Priebe, "(Brain) Graph Matching via Fast Approximate Quadratic Programming," *Submitted to Computational Statistics and Data Analysis*, 2013. [Online]. Available: http://arxiv.org/abs/1112.5507 4

[22] D. E. Fishkind, S. Adali, and C. E. Priebe, "Seeded Graph Matching," *arXiv preprint*, p. 1209.0367v1, 2012. [Online]. Available: http://arxiv.org/abs/1209.0367

[23] V. Lyzinski, S. Adali, J. T. Vogelstein, Y. Park, and C. E. Priebe, "Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability," *Submitted to Journal of Classification*.

[24] V. Lyzinski, D. L. Sussman, D. E. Fishkind, H. Pao, and C. E. Priebe, "Seeded graph matching for large stochastic block model graphs," *arXiv preprint*, p. 1310.1297, Oct. 2013. [Online]. Available: http://arxiv.org/abs/1310.1297 4