Spring 2018 Syllabus for Stat 315: Modern High-Dimensional Inference

Lucas Janson

1 Syllabus

Goal of the class: preparation for research in high-dimensional inference by giving a broad overview of the field, with a particular focus on recent advances.

We have 12 class meetings including this one, and the plan is to cover a few papers on an important subject in high-dimensional inference during each weekly 90 minute class period. The presentations should aim to be 60-70 minutes so that there is plenty of time afterwards for discussion and questions. Each class will have a group assigned to present the papers that day, and that group is expected to work together to make a single coherent presentation (as opposed to each group member taking a paper and a block of presentation time and just going one after the other).

These are hard, advanced papers, partially because this is a technical field, but also because they are modern, so the field hasn't had as much time to flesh out a deeper and simpler understanding of these results—which poses both challenges and opportunities! But in order to make sure that the class can get as much out of each presentation as possible, each presenting group will be required to give a practice presentation to just me a few days before their actual presentation. This also benefits the presenters, as I will give detailed and personal feedback so that they can improve their presenting skills—an extremely valuable skill for careers in both academia and industry! Related to this, the papers we are covering can get quite technical—today's JRSSB paper is probably on the slightly less-technical side of average for the class—so it is important that you have a sufficiently strong technical background to understand *and explain* advanced statistical concepts/proofs, or else the presentations will suffer which isn't fair to the rest of the class. Since we are covering a lot of material, everyone will be expected to read all the papers each week so that the presentations can be given at a fairly high level and the discussions are especially lively and productive.

What is high-dimensional inference? It is rarely explicitly defined and definitions almost certainly vary between people, but the best characterization I can come up with is inference (i.e., making precise probabilistic statements about an unobserved quantity) regarding high-dimensional data (i.e., observations on a large set of variables) in which dependence among the variables is relevant to the inferential question. In simpler terms, if your data is arranged with observations as rows and variables as columns, and it has a lot of columns, then your data is high-dimensional (you can also have some associated low-dimensional data, like a response variable). If, further, your goal is not exclusively prediction or point estimation, and you aren't treating the columns separately from one another (e.g., just doing inference on the means of the variables), then you are probably doing high-dimensional inference. Note the exclusion of pure prediction/classification—if your goal is just to predict a response variable, this isn't really inference since no probabilistic statement is required. But if you want guarantees or uncertainty bands around your predictions or want a confidence interval for the average risk of your prediction algorithm, then it does fall under the umbrella of inference. The most common problems deal with assessing or quantifying variable relationships, such as model selection in regression.

1.1 Presentation Guidelines

Please consider all the following important questions when composing your presentation. These are generally key things to think about when reading any paper.

1. Chalk talks or slides both OK, probably ideal to have some of both (slides especially for showing graphs or pictures, chalk especially for going through proofs) but this isn't a hard rule.

2. What is main take-away (∼1-2 sentences with no formulae)?

3. What other work had been done on this before? What background do we need to fully appreciate the results?

4. What are practical ramifications? What are the intended applications?

5. What are the theoretical results? What are the main technical tools/ideas?

6. What are the limitations of the results?

7. What are interesting areas of future research (identified either by authors or you)?

8. Encouraged to look for follow-up or more recent work I haven't identified that you find interesting! Either talk about it, or at least list it as suggested reading. Google Scholar's list of citing papers is a useful way to find more recent related papers.

9. Think the paper I suggested isn't the best one for the topic? No problem, just email me to discuss it!

1.2 Schedule

For all Harvard affiliates, if you can't figure out access to a paper, try searching the article title in the Journal section of Hollis.

1/25/2018 Selective inference: selective error rates, procedures/proofs to control them, genetics application. Note Yoav Benjamini is giving the April 9 department colloquium!

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

- Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9), 5116–5121.

2/1/2018 High-dimensional regression I: learn the players—lasso, group lasso, elastic net, adaptive lasso, Dantzig; graphical models; sparse principle components and canonical correlation analysis

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.

- Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.

- Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515–534.

2/8/2018 High-dimensional regression II: model selection and prediction consistency results

- Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov), 2541–2563.
- Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The annals of statistics, 1436–1462.
- Bickel, P. J., Ritov, Y. A., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705–1732.

2/15/2018 Bayesian: computational and statistical properties of Bayesian procedures in high-dimensional regression

- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889.
- Ročková, V., & George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506), 828–846.
- Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), 1986–2018.
- Song, Q., & Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.

2/22/2018 Bootstrap: can the bootstrap be used?

- Karoui, N. E., & Purdom, E. (2016). Can we trust the bootstrap in high-dimension?. *arXiv preprint arXiv:1608.00696*.
- Dezeure, R., Bühlmann, P., & Zhang, C. H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4), 685–719.

3/1/2018 High-dimensional regression p-values I: asymptotic p-values in high-dimensional GLMs

- Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1), 2869–2909.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.

- Dezeure, R., Bhlmann, P., Meier, L., & Meinshausen, N. (2015). High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical science*, 30(4), 533–558.

3/8/2018 High-dimensional regression inference II: limitations on high-dimensional p-values and how they might be overcome

- Cai, T. T., & Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2), 615–646.
- Zhu, Y., & Bradic, J. (2017). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, (in press).

3/22/2018 Post-selection inference: valid submodel p-values after using the same data to select a submodel.

- Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802–837.
- Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597.*

3/29/2018 Knockoffs: different approach to controlled variable selection that doesn't involve p-values

- Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), 2055–2085.
- Candès, E., Fan, Y., Janson, L., & Lv, J. (2017). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B (Methodological)*, (in press).

4/5/2018 Nonparametric: inference for high-dimensional nonparametric regression (conditional mean) functions

- Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (in press).
- Lu, J., Kolar, M., & Liu, H. (2015). Post-regularization confidence bands for high dimensional nonparametric models with local sparsity. *arXiv preprint arXiv:1503.02978.*

4/12/2018 Prediction: confidence intervals for predictive risk, and prediction intervals

- Kumar, R., Lokshtanov, D., Vassilvitskii, S., & Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (pp. 27–35).

- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2017). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, (in press).

4/19/2018 Synthesis: what did we learn, what did we miss, and where to go from here [Lucas presenting again]