Contents

Chapter 1

Synthetic Dataset Generation

Synthetic datasets have been generated for use in the hyperparameter tuning process for the various regression methods. These synthetic datasets have been designed to be very similar to real epigenetic datasets. The assumption is that the various regression methods would continue performing well in the context of real epigenetic data after having been tuned on similar generated datasets.

The benefit of using synthetic data for parameter tuning is the existence of ground truth about the relationship between predictors and the target variable. This ground truth enables comparing the various regression methods not only in terms of prediction error, but also with regard to the sensitivity, specificity and precision of their variable selection.

The sections of this chapter describe in detail the synthetic dataset generation process. The gene network, simulated expression levels of all genes and the primary simulation setups of the response variable are implemented as suggested by Li and Li [8]. Four secondary simulation setups are derived from each of the four primary setups, resulting in a total of 20 independent simulation setups.

1.1 Predictor network generation

All simulation setups share the following common predictor network. Consider a setup, for which 50 transcription factors regulate 10 independent genes each. In the gene network corresponding to this scenario, there would be edges between all transcription factors (TFs) and their 10 regulated genes.

The resulting network contains 550 nodes and 500 edges, all edge weights set to 1. This graph consists of 50 star-shaped connected components of 11 nodes, the central node of each representing the corresponding TF.

1.2 Generation of predictor observations

As a consequence of using the shared predictor network described in the previous section, all synthetic datasets contain 550 predictors. The expression levels for each of the 50 transcription factors follow a standard normal distribution $X_{TF_j} \sim N(\mu = 0, \sigma = 1)$.

The expression level of the regulated genes (RG) is dependent on the expression level of their corresponding TF_j and follows the normal distribution $X_{RG} \sim N(\mu = 0.7 * X_{TF_j}, \sigma = 0.71)$. This means that the expression levels of a TF and each of its RG are jointly distributed as a bivariate normal with a correlation of 0.7.

1.3 Response variable generation

Values of the response variable y are generated according to a linear model $y = X\beta + \epsilon$, where ϵ is added noise and the coefficient vector β is specified by the current simulation setup.

The added noise follows a normal distribution $\epsilon \sim N(\mu = 0, \sigma = F(\beta)))$, whose shape is calculated from the coefficient vector β for the current setup according to equation 1.1.

$$\sigma_\epsilon = F(\beta) = \sqrt{\frac{\sum_{j=1}^{p} \beta_j^2}{4}} \tag{1.1}$$

1.3.1 Primary simulation setups

The primary simulation setups assume that four transcription factors and their regulated genes are related to the response variable y.

Setup 1

$$\begin{aligned}\beta = (5,\ & \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}},\\ -5,\ & \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}},\\ 3,\ & \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}},\\ -3,\ & \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}},\\ 0,\ & \ldots, 0)\end{aligned} \tag{1.2}$$

Setup 2

$$\begin{aligned}\beta = (5,\ & \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}},\\ -5,\ & \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}}, \tfrac{-5}{\sqrt{10}},\\ 3,\ & \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}},\\ -3,\ & \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}}, \tfrac{-3}{\sqrt{10}},\\ 0,\ & \ldots, 0)\end{aligned} \tag{1.3}$$

Setup 3

$$\begin{aligned}\beta = (5,\ & \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10},\\ -5,\ & \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10},\\ 3,\ & \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10},\\ -3,\ & \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10},\\ 0,\ & \ldots, 0)\end{aligned} \tag{1.4}$$

Setup 4

$$\begin{aligned}\beta = (5,\ & \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10},\\ -5,\ & \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10}, \tfrac{-5}{10},\\ 3,\ & \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10},\\ -3,\ & \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10}, \tfrac{-3}{10},\\ 0,\ & \ldots, \quad 0)\end{aligned} \tag{1.5}$$

1.3.2 Secondary simulation setups

Bibliography

[1] Robin Holliday. Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80, 2006.

[2] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254, 2003.

[3] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, 2004.

[4] Manel Esteller. Epigenetics in cancer. *New England Journal of Medicine*, 358(11):1148–1159, 2008.

[5] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[7] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[8] Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[9] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498, 2010.

[10] Wei Pan, Benhuai Xie, and Xiaotong Shen. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484, 2010.

[11] Chong Luo, Wei Pan, and Xiaotong Shen. A two-step penalized regression method with networked predictors. *Statistics in biosciences*, 4(1):27–46, 2012.

[12] Sunkyung Kim, Wei Pan, and Xiaotong Shen. Network-based penalized regression with application to genomic data. *Biometrics*, 69(3):582–593, 2013.

[13] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[14] Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

[15] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014.

[16] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.