Binary Knockoffs Notes

Aaron Maurer

June 8, 2015

1 Preliminaries

Some early investigation into how the deterministic knockoffs as described in Rina's paper work with regularized logistic regression revealed that the answer is "not very well". Even when X is a null predictor of y, the X still tend to enter the model prior to $\tilde{X}$. The issue is that even when $X_i \sim N_p(\mathbf{0}, \Sigma)$ for some $\Sigma \succeq 0$, $\tilde{X}$ is not normally distributed. This can be seen from producing qq plots of X_i vs $\tilde{X}_j$Of course, when X_i is a binary vector, $\tilde{X}_i$ completely doesn't match its distribution, causing the original X to beat the knockoffs into the model. This indicates that a new method of generating $\tilde{X}$ must be created to use with FDR via knockoffs for regularized logistic regression.

2 Probabilistic Random Bernoulli Knockoffs

My idea is to generate $\tilde{X}$ randomly such that, approximately, $\tilde{X}_i \sim X_i$. In particular, they should have similar marginal densities, expectations, and first moments. However, $\tilde{X} \mid X$ should also have desired knockoff property that $\mathrm{E}(\tilde{X}'X \mid X) = X'X - s$, where $\mathrm{diag}(X'X) - s$ is small. In the general case, this is likely infeasible; however, if X is a binary vector, as is often the case, we know we are dealing with a much more limited class of random variables, and it should be possible to randomly generate $\tilde{X} \mid X$ so as to have the desired properties. At worst, this method will provide a suitable replacement for deterministic $\tilde{X}$ as described in Rina's paper for LASSO, and if we are lucky, it will work reasonably for other regularized GLMs.

3 Random Bernoulli Generation

Thankfully, there has been a reasonable amount of work on how one can generate random Bernoulli vectors with some kind of correlation among among the values. A random Bernoulli vector X can be easiest represented with a mean vector $\mathrm{E}(X) = m \in (0,1)^p$ and $\mathrm{E}(XX') = M \in (0,1)^{p \times p}$, called the cross moment matrix. Obviously, $m_i = \mathrm{P}(X_i = 1)$, $M_{ij} = \mathrm{P}(X_i = X_j = 1)$, and $m = \mathrm{diag}(M)$. For an arbitrary symmetric M to be valid cross-moment matrix, $M - \mathrm{diag}(M)\mathrm{diag}(M)'$ must be PSD, and

$$\max\{0, m_i + m_j - 1\} \leq M_{ij} \leq \min\{m_i, m_j\}$$

for all $i \neq j$[1]. Given a qualifying M, or observed X, there are a few ways of generating more random X.

3.1 Gaussian Copula Family

Since multivariate normal distributions are easy to randomly draw, the idea is to find some random normal variable $Z\ N_p(\mathbf{0}, \Sigma)$ such that,for $X_j = I(Z_j < 0)$, X has the desired properties. There are a number of ways to do this[2][3], but it turns out that there is only certain to exist a working Σ in the bivariate case.

3.2 μ-Conditionals family

It turns out that there exists a more flexible family which will always work for arbitrary M called μ-conditionals. The basic idea is that the X is generate sequentially as

$$X_j \mid X_j, ..., X_{j-1} \sim \mathrm{B}\left(1, \mu\left(a_{jj} + \sum_{k=1}^{j-1} a_{kj} X_j\right)\right)$$

for some monotone function $\mu : \mathbb{R} \to [0, 1]$. This is essentially a binomial family GLM for a link function μ. If one takes all of the a_{kj}, they can form a lower triangular matrix A, and then the joint density can be expressed as

$$\mathrm{P}(X_j = \gamma) \propto \mu(\gamma' A \gamma)$$

If μ is chose such that it is a bijection and differentiable, there is a unique M such that $\mathrm{E}(X_i X_i') = M$[4]. As one might guess, the natural μ is the logistic link function, which is the "binary analogue of the multivariate normal distribution which is the maximum entropy distribution on $\mathbb{R}^p$ having a given covariance matrix." Additionally, it has the usual benefit that the coefficients can be viewed as a log odds ratio:

$$A_{ij} = \log\left(\frac{\mathrm{P}(X_j = X_k = 1)\mathrm{P}(X_j = X_k = 0)}{\mathrm{P}(X_j = 0, X_k = 1)\mathrm{P}(X_j = 1, X_k = 0)}\right)$$

when $i \neq j$. I think this dictates that if $A_{jk} = 0$, then X_k and X_j are independent.

4 Generating Knockoffs

[1] "On parametric families for sampling binary data with specified mean and correlation" - http://arxiv.org/abs/1111.0576

[2] "On the Generation of Correlated Artificial Binary Data" - http://epub.wu.ac.at/286/1/document.pdf

[3] "On parametric families for sampling binary data with specified mean and correlation"

[4] "On parametric families for sampling binary data with specified mean and correlation"