Using Probabilistic Knockoffs of Binary Variables to Control the False Discovery Rate

Aaron Maurer
Advisor: Rina Foygel Barber

July 29th, 2015

Overview

1. Original Knockoffs: What They Do and Where They Fail
2. Making Knockoffs Work With GLMs
3. Random Binary Knockoffs: The Theory
4. Random Binary Knockoffs: Performance
5. Where to next?

Variable Selection in Linear Regression

Assume

$$\mathbf{y} = X\beta + \mathbf{z}$$

where $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, and $\mathbf{z}$ is Gaussian noise. Also, assume sparsity:

$$\beta_i = 0 \quad \forall i \notin S$$

How do we pick estimate $\hat{S}$?

False Discover Rate

A common goal for a method that generates $\hat{S}$ is to control the false discovery rate

$$\mathrm{FDR} = \mathrm{E}\left[\frac{|\{j : \beta_j = 0 \,\&\, j \in \hat{S}\}|}{\max\{|\hat{S}|, 1\}}\right]$$

In other words, control portion of elements in $\hat{S}$ which aren't in S.

FDR is controlled at level q if $q <$FDR irrespective of true β.

Knockoff Features

Knockoff variables can be used to control FDR in linear regression.

- The idea is to create a forgery of each variable; if the forgeries seem about as good predictors as the originals, the originals are lousy predictors.
- For each variable X_i, create a knockoff feature $\tilde{X}_i$ such that, where $X^TX = G$, $\mathrm{diag}\{X^TX\} - s$ and

$$\tilde{X}^T\tilde{X} = G \quad \& \quad X^T\tilde{X} = G - \mathrm{diag}\{\mathbf{s}\}$$

- For $\tilde{X}$ to exist, it must be the case that

$$G_L = [X\ \tilde{X}]^T[X\ \tilde{X}] = \begin{bmatrix} G & G - \mathrm{diag}\{\mathbf{s}\} \\ G - \mathrm{diag}\{\mathbf{s}\} & G \end{bmatrix} \succeq 0$$

- $\tilde{X}_i$ and X_i will have same correlation with other variables, but only low correlation with each other.
- Given $\mathbf{s}$, $\tilde{X}$ can be generated via a rotation of X.

Knockoff Filter

These knockoffs can be used in the knockoff filter method.

- Fit full path of LASSO regression on $[\, X\ \tilde{X}\,]$.

$$\beta(\lambda) = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2} \|\mathbf{y} - X_L \mathbf{b}\|_2^2 + \lambda \|b\|_1 \right\}$$

- Z_i, $\tilde{Z}_i$ largest λ such that X_i, $\tilde{X}_i$ have nonzero coefficient.
- $W_i = Z_i$ if $Z_i > \tilde{Z}_i$, otherwise $W_i = -\tilde{Z}_i$.
- Since G_L & $[\, X\ \tilde{X}\,]^T \mathbf{y}$ sufficient statistics for $\beta(\lambda)$, W_i symmetrically distributed around 0 when X_i null predictor.
- Thus, FDR controlled when $\hat{S} = \{i : W_i > T\}$ for

$$T = \min \left\{ t > 0 \ :\ \frac{|\{j : W_j \leq -t\}|}{\max\{|\{j : W_j \geq t\}|, 1\}} \leq q \right\}$$

Variable Selection in GLMs

Knockoffs work great for linear regression, but what about GLMs?

Now, assume, for some link function g and $y_i, \ldots, y_n$ from a exponential family distribution,

$$\mathrm{E}(\mathbf{y}) = g(X\beta)$$

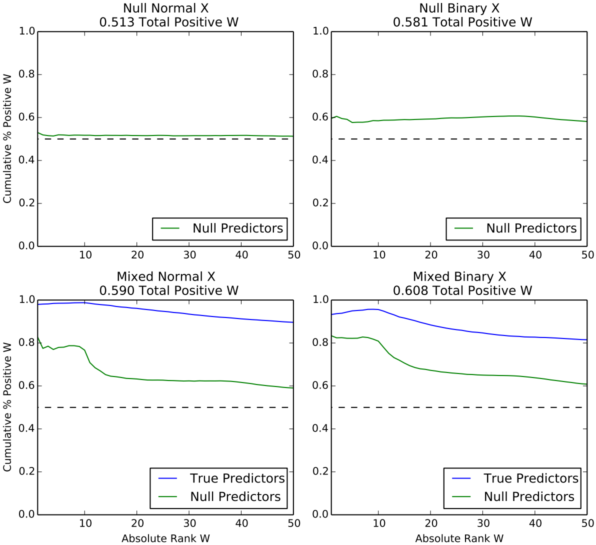
where $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$. Also, assume sparsity:

$$\beta_i = 0 \quad \forall i \notin S$$

How do we pick estimate $\hat{S}$?

Where Knockoff Filter Fails

Knockoff filter don't work for other GLMs.



Can Knockoffs Be Fixed for GLMs?

- Other GLMs don't have the same sufficient statistics as linear regression.
- Original Knockoffs don't remotely have same distribution as X, so "look" different than real variables.
- Knockoffs will likely work better if they have the same marginal distribution as originals.
- For X_i with arbitrary distribution, unclear how this might be accomplished.

Random Binary Notation

- Binary data is common in data analysis and a much more manageable family of distributions for X.
- We can think of observations in X as observations of random binary vector $\mathbf{x} \in \{0,1\}^p$.
- The full family for $\mathbf{x}$ is multinomial on 2^p outcomes.
- Still useful to consider first two moments:

$$\mathrm{E}(\mathbf{x}) = \mathbf{m} \in [0,1]^p \quad \& \quad \mathrm{E}(\mathbf{x}\mathbf{x}^T) = M \in [0,1]^{p \times p}$$

- For arbitrary M to correspond to a random binary vector, must be case that $M - \mathbf{m}\mathbf{m}^T = \Sigma \succeq 0$

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

Random Binary Knockoffs

- Integer programing is np-hard, making finding finding $\tilde{X} \in \{0,1\}^{n \times p}$ to fit correlations exactly difficult.
- Instead, introduce a relaxed problem where $\tilde{X} \mid X$ is drawn randomly such that, where $\Sigma = \mathrm{Cov}(\mathbf{x})$

$$\mathrm{Cov}(\tilde{\mathbf{x}}, \mathbf{x}) = \Sigma - \mathrm{diag}\{\mathbf{s}\} \quad \& \quad \mathrm{Cov}(\tilde{\mathbf{x}}) = \Sigma$$

- For this to correspond to a random binary vector, must be the case

$$\Sigma_L = \mathrm{Cov}([\,\mathbf{x}\,\tilde{\mathbf{x}}\,]) = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{\mathbf{s}\} \\ \Sigma - \mathrm{diag}\{\mathbf{s}\} & \Sigma \end{bmatrix} \succeq 0$$

- Almost same correlation condition as before, just only holds in expectation.
- Switch from Gramian matrix to correlation matrix makes moment condition less likely to be violated.

Quadratic Programing

- Simplest approach to Random Binary Knockoffs is to draw the entries of $\tilde{X}$ independently based on $P \in [0,1]^{n \times p}$.
- The best possible P for the task would satisfy

$$\begin{array}{rl} \text{minimize} & \|X^T P - (M - \mathrm{diag}\{s\})\|_{fro}^2 + \sum_{i \neq j} (P_i^T P_j - M_{ij})^2 \\ \text{subject to} & \mathbf{1}^T P = \mathbf{m} \\ & 0 \leq P \leq 1 \end{array}$$

- Can be formulated as a quadratic program with slack variables

$$\begin{array}{rl} \text{minimize} & \|W\|_{fro}^2 + \|V\|_{fro}^2 \\ \text{subject to} & -W \leq X^T P - (M - \mathrm{diag}\{s\}) \leq W \\ & -V_{ij} \leq P_i^T P_j - M_{ij} \leq V_{ij} \quad \forall i \neq j \\ & \mathbf{1}^T P = \mathbf{m} \\ & 0 \leq P \leq 1 \end{array}$$

- Huge optimization problem, likely computationally difficult.

Ising Model

- Instead, what if we found a random binary vector variable $\mathbf{x}_L$ that had cross-moments M_L corresponding to Σ_L?
- The Ising model can match any M_L. If A lower triangular matrix and L logistic link

$$\mathrm{P}(\mathbf{x} = \gamma) \propto L(\gamma^T A \gamma)$$

- The Ising model binary analog of normal distribution; maximum entropy for given covariance matrix.
- Very easy to draw successive entries

$$\mathrm{P}(x_i = 1 \mid x_1, \dots, x_{i-1}) = L\left(A_{ii} + \sum_{k=1}^{i-1} A_{ik} x_i\right)$$

- Once fit, can draw $\tilde{\mathbf{x}} \mid \mathbf{x}$ easily.

Fitting Ising Model

- If we were just trying to fit $A\ X$, we could do so via successive logistic regression to fit row $\mathbf{a}_i$.
- Instead, simulate $\mathbf{m}_i = f(\mathbf{a}_i)$ and fit via Newton-Raphson. Let $\mathbf{x}_{-i}^{(k)} \sim \mathbf{x}_{-i}$

$$f\left(\mathbf{a}_i\right) \approx \frac{1}{K}\sum_{k=1}^{K} L\left(\mathbf{a}_i^T \begin{bmatrix} \mathbf{x}_{-i}^{(k)} \\ 1 \end{bmatrix}\right) \begin{bmatrix} \mathbf{x}_{-i}^{(k)} \\ 1 \end{bmatrix}$$

$$J\left(\mathbf{a}_i\right) \approx \frac{1}{K}\sum_{k=1}^{K} L'\left(\mathbf{a}_i^T \begin{bmatrix} \mathbf{x}_{-i}^{(k)} \\ 1 \end{bmatrix}\right) \begin{bmatrix} \mathbf{x}_{-i}^{(k)} \\ 1 \end{bmatrix} \begin{bmatrix} [x_{-i}^{(k)}]^T & 1 \end{bmatrix}$$

- Make successive updates

$$\mathbf{a}_i^{(k+1)} = \mathbf{a}_i^{(k)} - \left[J\left(\mathbf{a}_i^{(k)}\right)\right]^{-1}\left[f\left(\mathbf{a}_i^{(k)}\right) - \mathbf{m}_i\right]$$

Computational Issues

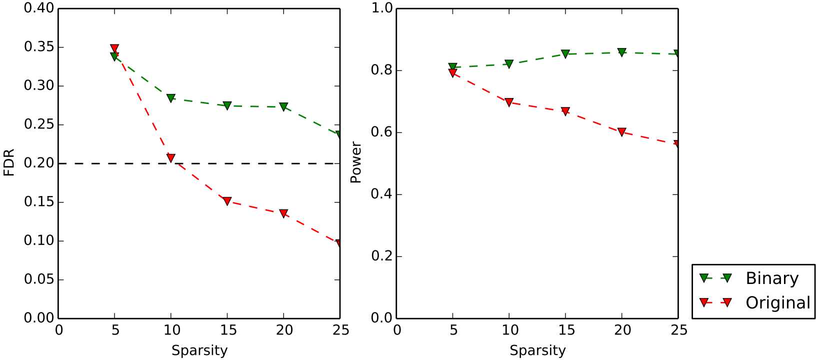
- K must be very large for big p and high correlation.
- This makes $J(\mathbf{a}_i)$ and even $f(\mathbf{a}_i)$ very expensive to calculate.
- This makes quasi-Newtonian methods, where $J(\mathbf{a}_i)$ is approximated appealing.
- In particular, Anderson Mixing, where f approximated with secant hyperplane through $\mathbf{a}_i^k, \ldots, \mathbf{a}_i^{(k-h+1)}$ works well
- When K too small, can instead solve relaxed problem

$$\mathbf{m}_i^*(\tau) = (1-\tau)\mathbf{m}_i + \tau \begin{bmatrix} 0 & \ldots & 0 & M_{ii} \end{bmatrix}^T$$

- n doesn't matter, but this method is also fairly impractical for large p.

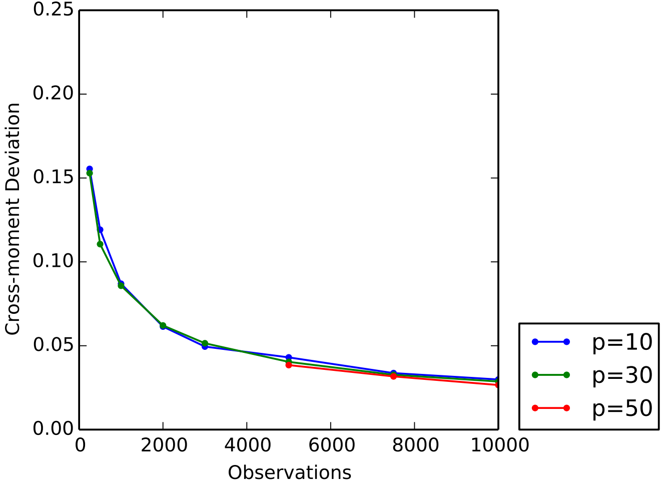
Random Binary Knockoffs in Linear Regression

Unfortunately, random binary Knockoffs only provide approximate FDR control for linear regression.



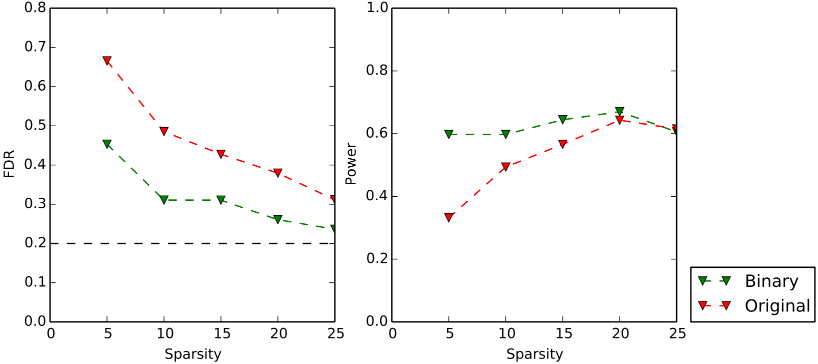
Distortion from M_L

Since $\tilde{X}$ is randomly generated $\frac{1}{n}X_L^TX_L$ deviates from M_L.



Random Binary Knockoffs in Logistic Regression

Still, Random Binary Knockoffs seem to do way better than original knockoffs in logistic regression.



Discussion

- Random Binary Knockoffs seem to offer promise as a useful technique, but have outstanding issues.
- Seem to offer a method to extend Knockoffs for one type of variable to GLMs.
- Computational complexity prohibitive; simpler method, perhaps by good approximation of P, would be helpful.
- Random distortion from desired cross-moments might be compensated for by ensemble method based on multiple $\tilde{X}$.
- Might build higher order interactions into Ising model to allow for nasty data.