Variable selection in regression is a prominent problem in Statistics, and is especially relevant to inference and prediction in modern applications. In high dimensional situations, using the Lasso program for the task is common, the ordinary and most natural procedure relying on sparsity of the Lasso estimate itself. More generally, we consider variable selection procedures given by

$$\{1 \le j \le p : |\widehat{\beta}_j(\lambda)| > t\},$$

where $\widehat{\beta}(\lambda)$ is the Lasso estimate of the regression coefficients, and where λ and t may be data dependent. Thus, fixing $t = 0$ and allowing to control only λ captures ordinary Lasso selection, whereas allowing to control both λ and t captures selection by thresholding of the Lasso estimates. Figuratively, thresholded-Lasso opens up the possibility to look further down the Lasso path, which typically leads to dramatic improvement in power. This phenomenon has been quantified recently leveraging advances in approximate message-passing (AMP) theory, but the implications are actionable only if one assumes substantial knowledge of the underlying signal that is needed for control of the Type I error. In this work we study theoretically the power of a knockoffs-calibrated counterpart of thresholded-Lasso that enables us to control FDR in the realistic situation where no prior information about the signal is available. Although the basic AMP framework remains the same, the analysis requires a significant technical extension of existing theory in order to handle the pairing between original variables and their knockoffs. Relying on this extension we obtain exact asymptotic predictions for the true positive proportion achievable at a prescribed type I error level. We show that the knockoffs version of thresholded-Lasso can (still) perform much better than ordinary Lasso selection if λ is chosen by cross-validation on the augmented matrix.

Keywords: Variable selection, Lasso, False Discovery Rate, Approximate Messgae-Passing, Knockoffs