The Generalized Higher Criticism for Testing SNP-sets in Genetic Association Studies

Ian Barnett, Rajarshi Mukherjee & Xihong Lin

Harvard University

ibarnett@hsph.harvard.edu

June 24, 2014

Background

- Genome-wide association studies (GWAS): millions of common (minor allele frequency > 0.05) SNPs genotyped.
- Gene-level/pathway-level analysis can provide power to detect these types of effects by combining information over the SNPs.
- Goal: Develop powerful, computationally efficient, statistical methodology for SNP-sets that have the power to detect joint SNP effects.

Model

- n subjects, q covariates, p genetic variants.
- Y_i is phenotype for ith individual
- $\mathbf{X}_{i\cdot}$ contains q covariates for ith individual
- $\mathbf{G}_{i\cdot}$ contains SNP information (minor allele counts) in a gene/pathway/SNP-set for ith individual
- α and β contain regression coefficients.
- $\mu_i = E(Y_i|\mathbf{G}_{i\cdot}, \mathbf{X}_{i\cdot})$

Model

$$h(\mu_i) = \mathbf{X}_{i\cdot}\alpha + \mathbf{G}_{i\cdot}\beta$$

- $h(\cdot)$ is the link function.

Marginal SNP test statistics

- The marginal score test statistic for the jth variant is:

$$Z_j = \mathbf{G}_{\cdot \mathbf{j}}^{\mathbf{T}}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)$$

  where $\hat{\boldsymbol{\mu}}_0$ is the MLE of $E(\mathbf{Y}|H_0)$. Assume Z_j is normalized.
- Letting $\mathbf{U}\mathbf{U}^{\mathbf{T}} = \widehat{Cov}(\mathbf{Z}) = \hat{\boldsymbol{\Sigma}}$, define the transformed (decorrelated) test statistics:

$$\mathbf{Z}^* = \mathbf{U}^{-1}\mathbf{Z} \xrightarrow[n\to\infty]{\mathcal{L}} MVN(\mathbf{0}, \mathbf{I}_p)$$

Current popular methods

Method	**SKAT**	**MinP**
Test statistic	$\sum_{j=1}^{p} Z_j^2$	$\max_j\{\lvert Z_j\rvert\}$
Pros	High power when signal sparsity is low. Accurate p-values can be obtained quickly.	High power when signal sparsity is high.
Cons	Can have very low power when sparsity is high.	Slightly lower power when sparsity is low. Difficult to obtain accurate analytic p-values.

The higher criticism

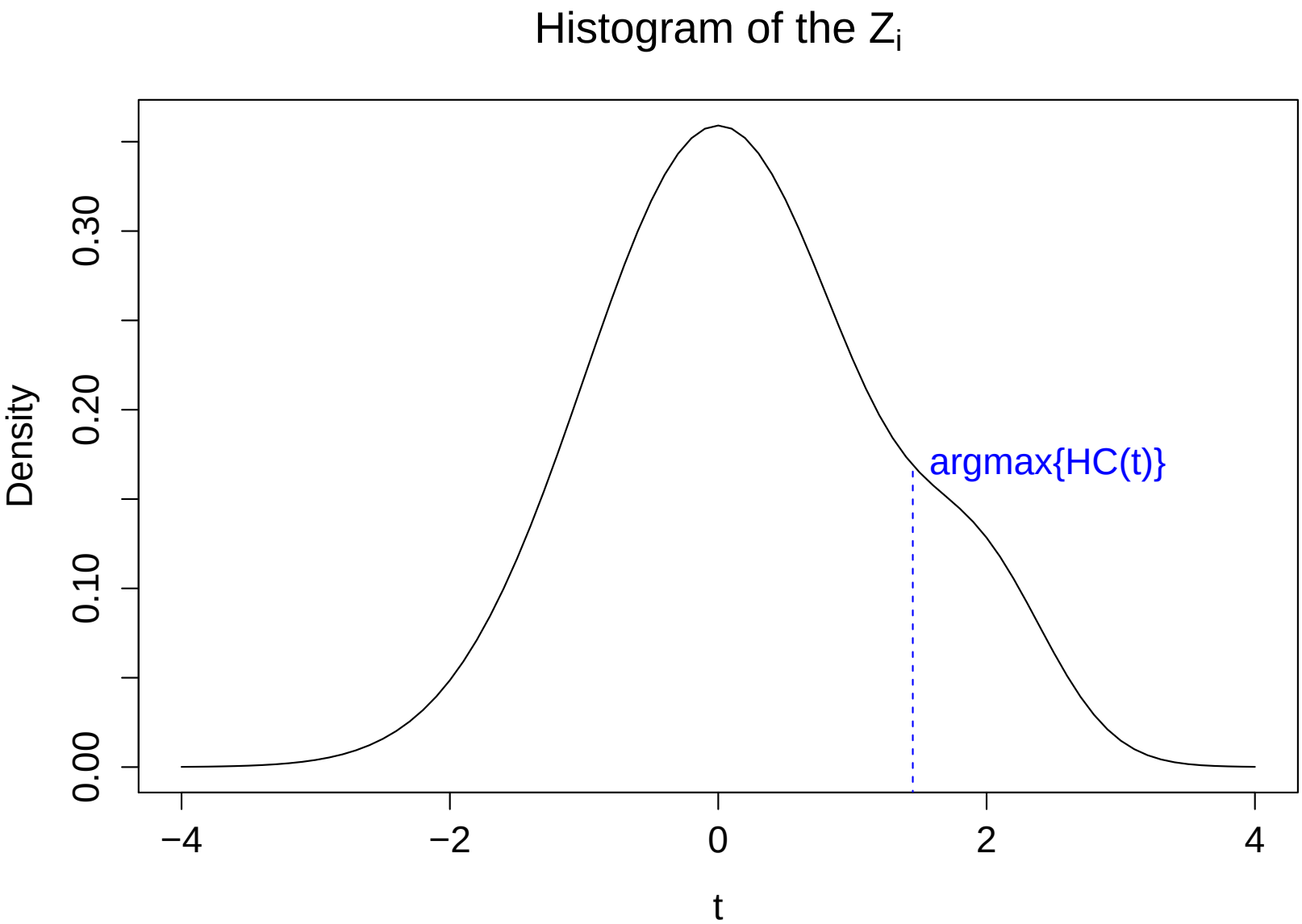
Let

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$$

- **Assumes** $\Sigma = \mathbf{I}_p$
- Under H_0, $S(t) \sim Binomial(p, 2\bar{\Phi}(t))$ where $\bar{\Phi}(t) = 1 - \Phi(t)$ is the survival function of the normal distribution.
- The Higher Criticism test statistic is:

$$HC = \sup_{t>0} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \right\}$$

The higher criticism



Adjusting for correlation

Recalling that $\mathbf{Z}^* = \mathbf{U}^{-1}\mathbf{Z}$, let

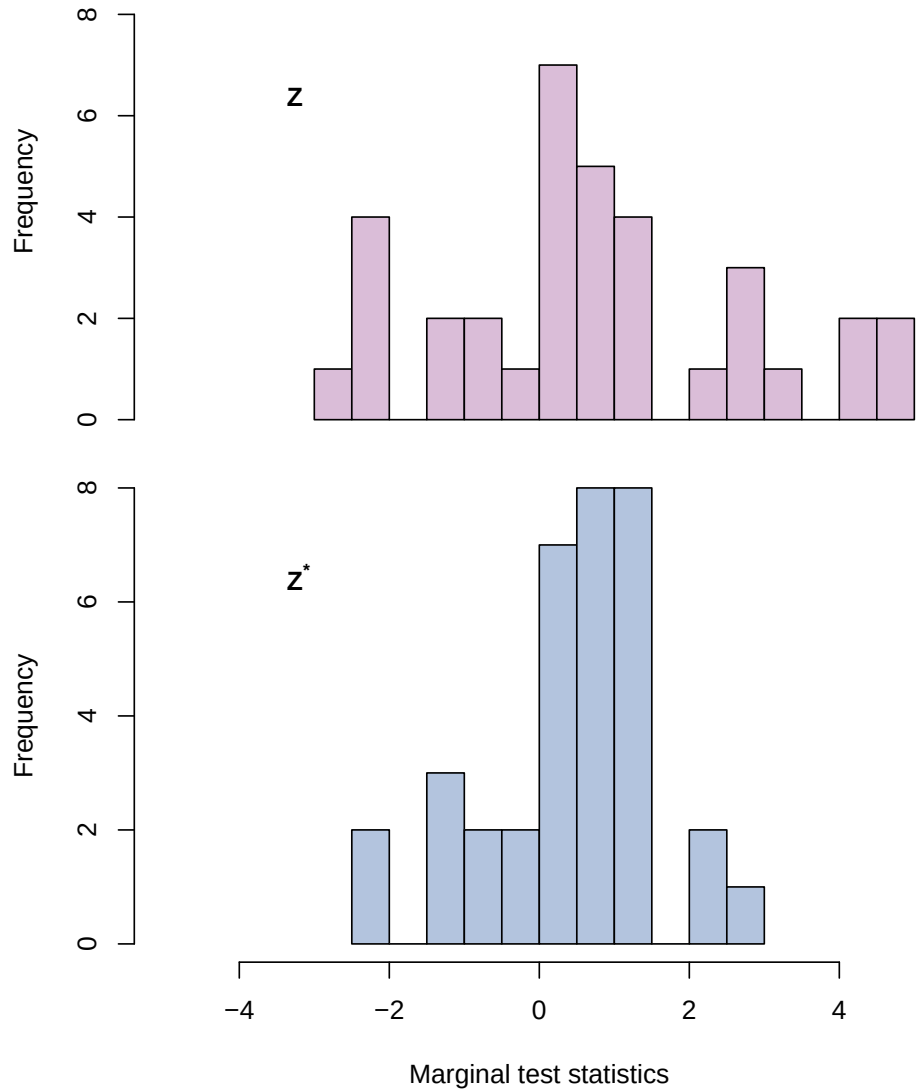
$$S^*(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j^*| \geq t\}}$$

- Note that under H_0, $S^*(t) \sim Binomial(p, 2\bar{\Phi}(t))$ **regardless for general correlated** $\mathbf{\Sigma}$.
- The innovated Higher Criticism test statistic is:

$$iHC = \sup_{t>0} \left\{ \frac{S^*(t) - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \right\}$$

Adjusting for correlation

Cancer Genetic Markers of Susceptibility (CGEM) Breast Cancer GWAS: **FGFR2 gene**



Decorrelating causes iHC to lose power.

Comparison

Method	MinP	SKAT	iHC
Robust to signal sparsity	✓		✓
Robust to correlation/LD structure	✓		✓
Computationally efficient		✓	✓
Does not require decorrelating test statistics	✓	✓	

Comparison

Method	MinP	SKAT	iHC	GHC
Robust to signal sparsity	✓		✓	✓
Robust to correlation/LD structure	✓		✓	✓
Computationally efficient		✓	✓	✓
Doesn't require decorrelating test statistics	✓	✓	✓	✓

*We will also consider the omnibus test, OMNI, in our power simulations. It is based on the minimum p-value of the SKAT, MinP, and GHC.

Our contribution: the generalized higher critcism (GHC)

Recall

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$$

- Now **we allow Σ to have arbitrary correlation structure.**
- $S(t)$ is no longer binomial. Instead we approximate with Beta-binomial, matching on first two moments.
- The Generalized Higher Criticism test statistic is:

$$GHC = \sup_{t>0} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\widehat{Var}(S(t))}} \right\}$$

The variance estimator $\widehat{Var}(S(t))$

Theorem 1

Let $\overline{r^n} = \frac{2}{p(1-p)} \sum_{1 \leq k < l \leq p} (\Sigma_{kl})^n$ *and let* $\mathcal{H}_i(t)$ *be the Hermite polynomials:* $\mathcal{H}_0(t) = 1$, $\mathcal{H}_1(t) = t$, $\mathcal{H}_2(t) = t^2 - 1$ *and so on. Then*

$$Cov\left(S(t_k), S(t_j)\right) = p[2\bar{\Phi}(max\{t_j, t_k\}) - 4\bar{\Phi}(t_j)\bar{\Phi}(t_k)]$$
$$+4p(p-1)\phi(t_j)\phi(t_k) \sum_{i=1}^{\infty} \frac{\mathcal{H}_{2i-1}(t_j)\mathcal{H}_{2i-1}(t_k)\overline{r^{2i}}}{(2i)!}$$

Proof follows from Schwartzman and Lin (2009) where they showed:

$$P(Z_k > t_i, Z_l > t_j) = \bar{\Phi}(t_i)\bar{\Phi}(t_j) + \phi(t_i)\phi(t_j) \sum_{n=1}^{\infty} \frac{\Sigma_{kl}^n}{n!} \mathcal{H}_{n-1}(t_i)\mathcal{H}_{n-1}(t_j)$$

Analytic p-values for the GHC

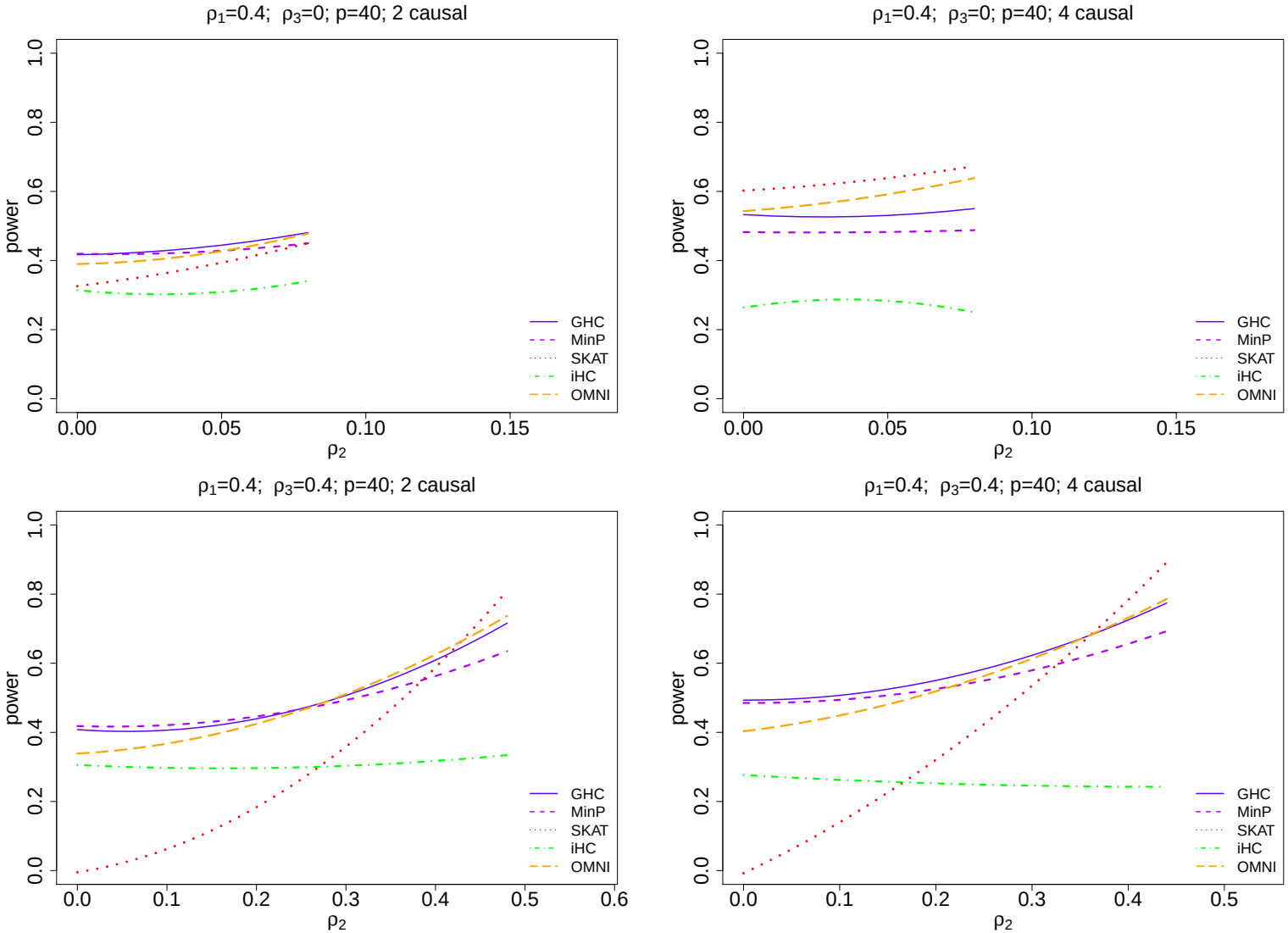
- Letting h be the observed GHC statistic:

$$\text{p-value} = pr\left(\sup_{t>0}\left\{\frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\widehat{Var}(S(t))}}\right\} \geq h\right)$$

- There exists $0 < t_1 < \cdots < t_p$, such that

$$\text{p-value} = 1 - pr\left(\bigcap_{k=1}^{p}\{S(t_k) \leq p - k\}\right)$$

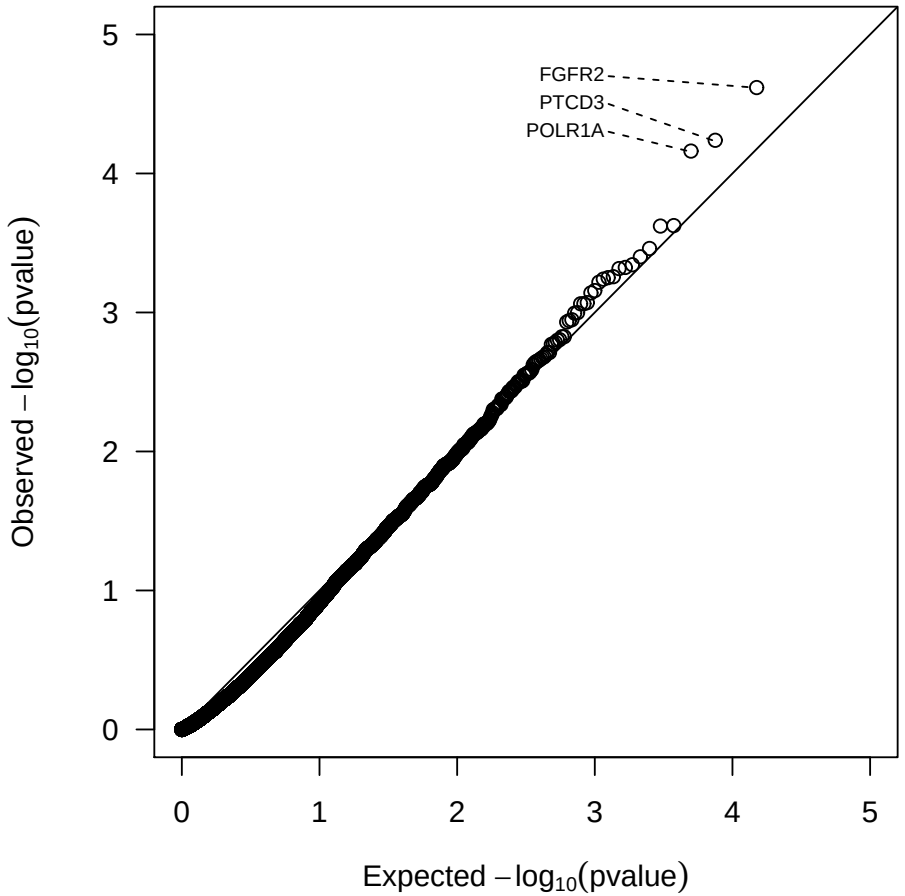
Power simulations



- ρ_1: correlation within causal variants.
- ρ_2: correlation between causal and noncausal variants.
- ρ_3: correlation within non-causal variants.

Data analysis

The National Cancer Institute's Cancer Genetic Markers of Susceptibility (CGEM) breast cancer GWAS. Sample has 1145 cases, 1142 controls with european ancestry.



Next (and final) step

- Thresholding tests (GHC and MinP) and summing tests (SKAT) are good complements
- Combining these classes of tests in a more principled way (than OMNI) is to use the following test statistic:

$$\sup_{\gamma,t} \left\{ \sum_{j=1}^{p} |Z_j|^{\gamma} I_{\{|Z_j|>t\}} \right\}$$

- $\gamma = 2,\ t = 0 \rightarrow$ SKAT
- $\gamma = 0 \rightarrow$ GHC
- We label this test as OPT.

Simulations $p = 20$, exchangeable correlation ρ

- For OPT, the supremum is selected from $t \in (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4)$ and $\gamma \in (0, 0.5, 1, 1.5, 2)$.
- Non-zero β decrease with ρ.