Bayesian Shrinkage Estimation of the Relative Abundance of mRNA Transcripts using SAGE

Jeffrey S. Morris, Keith Baggerly, and Kevin Coombes

University of Texas, MD Anderson Cancer Center

Outline

- Introduction
- Statistical Model for SAGE counts
- MLEs and Symmetric Dirichlet Prior
- Our method: Mixture Dirichlet Prior
- Simulation Study
- Conclusion

Goals of SAGE Experiments

1. Detect differentially expressed genes/perform clustering
2. Characterize transcriptome of tissue
   - Total number of expressed mRNA species
   - Relative abundance (proportions) of expressed genes

- Note: Question 2 underlies question 1

Statistical Model for SAGE Data

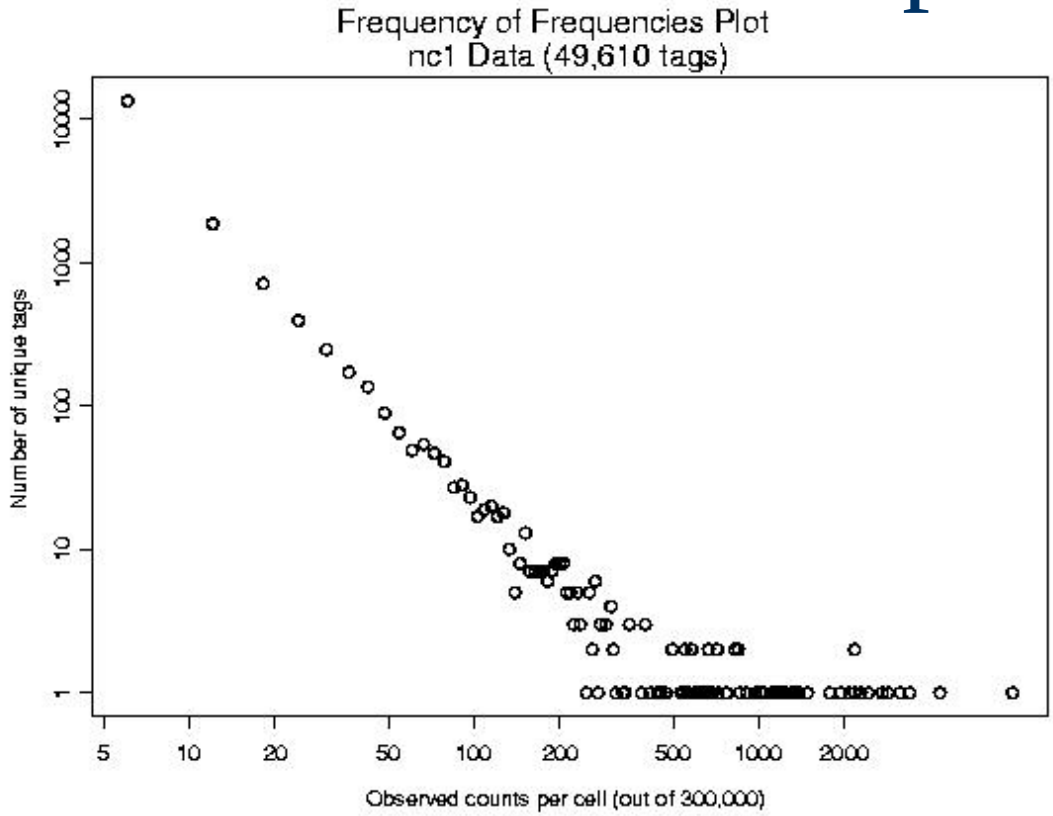
- X_i = # counts for gene i
- $\underline{\mathbf{X}} \sim$ Multinomial($n, \underline{\pi}$)
  - Dimension k= # of expressed genes
  - π_i = relative expression of gene i
- Particular characteristics:
  - $n \sim k$
  - "Distribution" of π_i heavily right skewed
- Result: Many 'missing' genes (X_i=0)

From Colon Cancer SAGE Libraries
(Velculescu, et al. 1999)

Copies/cell	% of genes	% of mass
≤5	89.9%	23%
5-50	9.2%	30%
50-500	0.8%	27%
500-5000	0.1%	20%

- Most abundant 1% of genes account for nearly 50% of total mass of mRNA

Skewed Distribution of proportions



- This linear shape in the log-log plot reflects extreme right skewness, and is characteristic of SAGE data.

Maximum Likelihood Estimation

- Maximum likelihood estimators of π_i:

$$\hat{\pi}_{i,\mathrm{MLE}} = X_i / n$$

- Problem with MLE for scarce genes
  - Underestimates π_i for 'missing genes':

$$\boxed{\hat{\pi}_{i,\mathrm{MLE}} = 0 < \pi_i}$$

  - Thus, on average, overestimates π_i for other genes.

Simple Example

- Population:
  - 1 abundant gene: $\pi_0 = 0.50$
  - 50 scarce genes: $\pi_i = 0.01$
- Sample n=20 mRNA transcripts
- For scarce genes:
  - If $X_i = 0$ $\Rightarrow$ $\hat{\pi}_{i,\mathrm{MLE}} = 0$
  - If $X_i = 1$ $\Rightarrow$ $\hat{\pi}_{i,\mathrm{MLE}} = 0.05$

Shrinkage Estimation

- In various multivariate contexts, it has been shown that *shrinkage estimation* can yield estimates with efficiency advantages over the MLE in estimating the joint set of parameters.
  - Shrink MLEs towards specific prior mean
- James and Stein(1961), George (1986), Gruber (1998)

Simple Dirichlet Prior

- Dirichlet prior conjugate for Multinomial

$$\underline{\pi} \sim \text{Dirichlet}(\theta, \ldots, \theta)$$

$$\underline{\pi} | \underline{\mathbf{X}} \sim \text{Dirichlet}(\theta + X_1, \ldots, \theta + X_k)$$

- Regresses MLE towards prior mean $1/k$
- In SAGE:
  - Improves estimation for scarce genes
  - Induces severe bias for abundant genes
- Prior assumption inaccurate for SAGE

Example Revisited

- Scarce Gene *i*:

	DIR(1)	MLE
$X_i = 0$:	$1/71 = \mathbf{0.014}$	$0/20 = \mathbf{0.000}$
$X_i = 1$:	$2/71 = \mathbf{0.028}$	$1/20 = \mathbf{0.050}$

- Abundant Gene:

If $X=10$:	$11/71 = \mathbf{0.15}$	$10/20 = \mathbf{0.50}$

Prior Information

- We have the following prior information on relative expression levels:
  - ***Skewness:*** Many scarce species, few abundant ones
  - ***Exchangability:*** No *a priori* knowledge assumed on which species will be seen, or are scarce or abundant.
- Our prior should satisfy relative frequency constraint $\Sigma \pi_i = 1$

Mixture Dirichlet Prior

- **Idea**: Partition genes into 2 classes:
  - ***Scarce*** and ***Abundant***, w/ different priors
- Must introduce new parameters:
  - λ_i = indicator gene i in abundant class ($i \in A$)
  - π^* = total abundant mass $= \Sigma\ \lambda_i\ \pi_i$
  - q_i = proportion of abundant/scarce mass for gene i
- NOTE: $\pi_i = \{\pi^* q_i\}^{\lambda i} \{(1-\pi^*) q_i\}^{(1-\lambda i)}$

Mixture Dirichlet Prior

- $\underline{q}_A$ = Dirichlet ($\theta_{A, \ldots,} \theta_A$), $A=\{i: \lambda_i=1\}$
  $\underline{q}_S$ = Dirichlet ($\theta_{S, \ldots,} \theta_S$) , $S=\{i: \lambda_i=0\}$
- $\lambda_i \sim$ Bernoulli (P)
- $\pi^* \sim$ Beta (a_{π^*}, b_{π^*})
- Recall $\pi_i = \{\pi^* q_i\}^{\lambda i} \{(1-\pi^*) q_i\}^{(1-\lambda i)}$

Fitting the Model

MCMC used to get posterior of $\underline{\pi}$:

1. Sample $\underline{\lambda}$ from f($\underline{\lambda}$|$\underline{X}$,π^*,P) (Bernoulli).
   Redefine A,S.
2. Sample $\underline{q}_A$ from Dirichlet(X_i+θ_A,$i \in A$)
3. Sample $\underline{q}_S$ from Dirichlet(X_i+θ_S,$i \in S$)
4. Sample π^* from Beta(a_{π^*}+n_A,b_{π^*}+n_S)

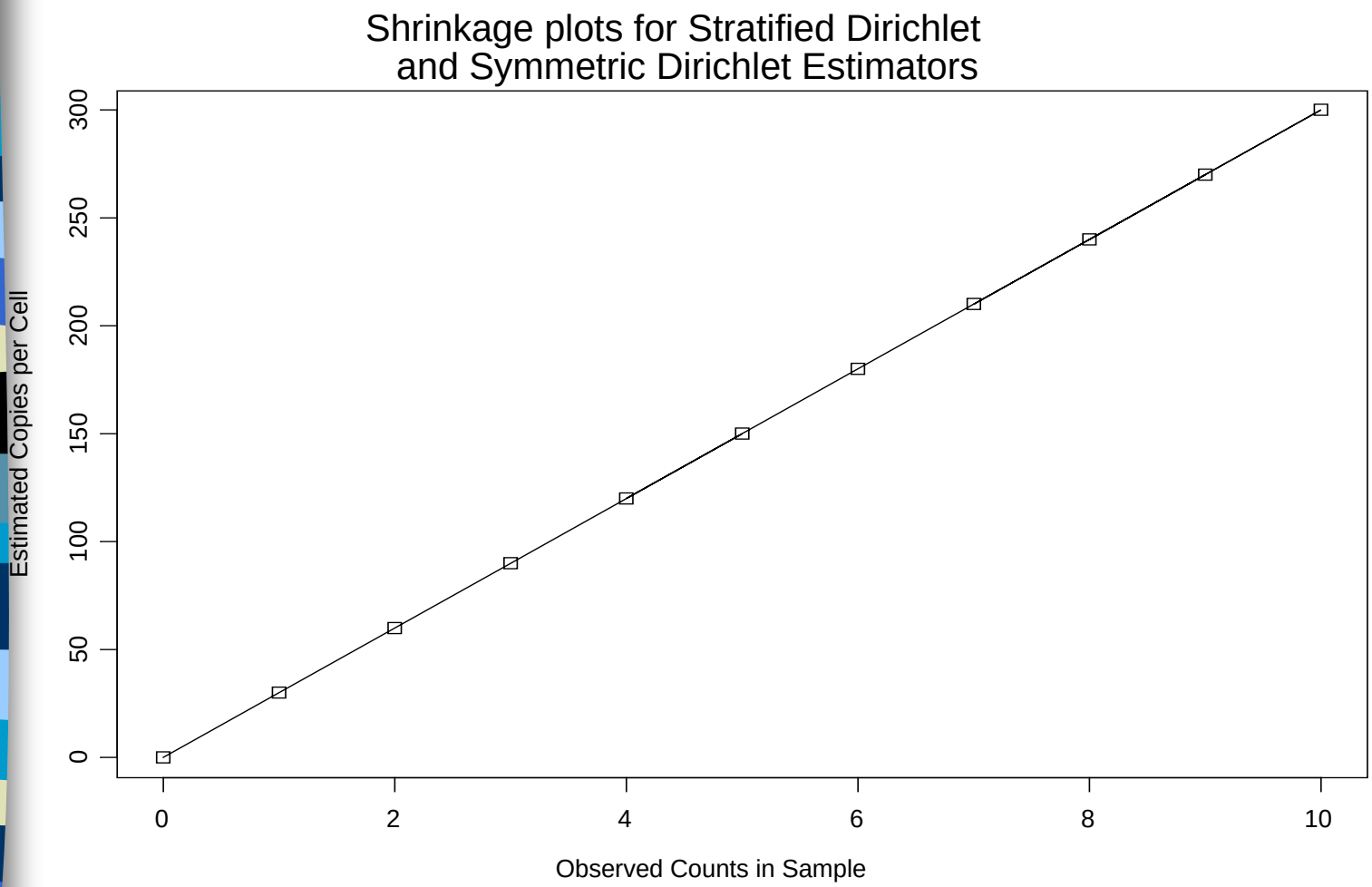
Fitting the Model

- λ_i sampled one at a time.
  - Generate $u \sim U(0,1)$
  - Set $\lambda_i=1$ if $u < \alpha_i = \Pr(\lambda_i=1 \mid \underline{\lambda}_{(-i)}, \underline{X}, P, \pi^*)$
- α_i available in closed form
  - Computationally expensive -- must calculate for all k~50,000 genes at each MCMC iteration
  - Some computation tricks necessary to speed calculations.

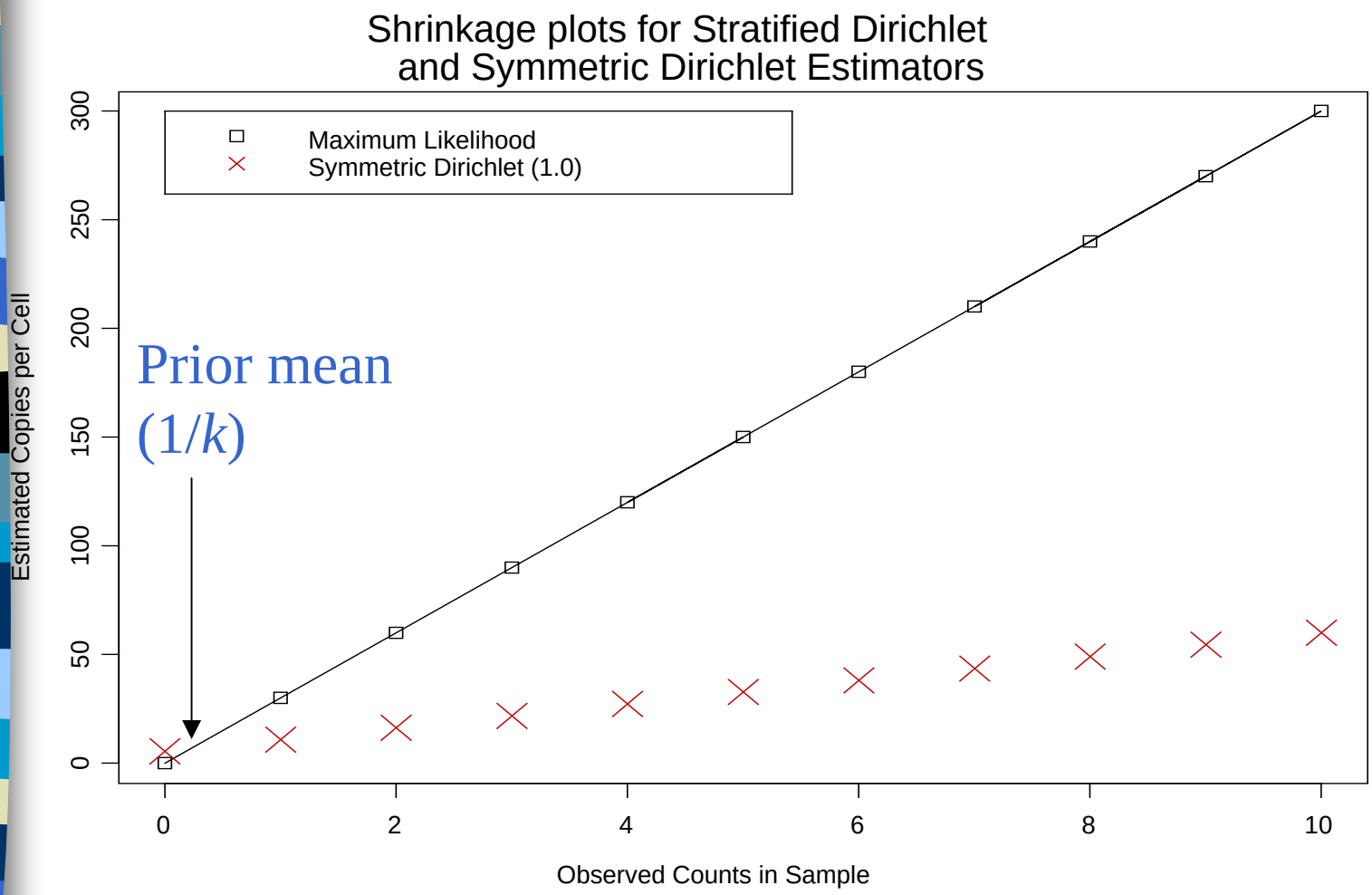
Fitting the Model

- Quicker option:
  - $k_A \sim \text{Binomial}(k,P)$
  - Assumes monotonicity in tags' abundance
- Draw samples from $f(k_A|\underline{X},P)$ using M-H
  - Abundant set = k_A tags with largest counts
- MCMC time = 1 minute vs. 1 hour.

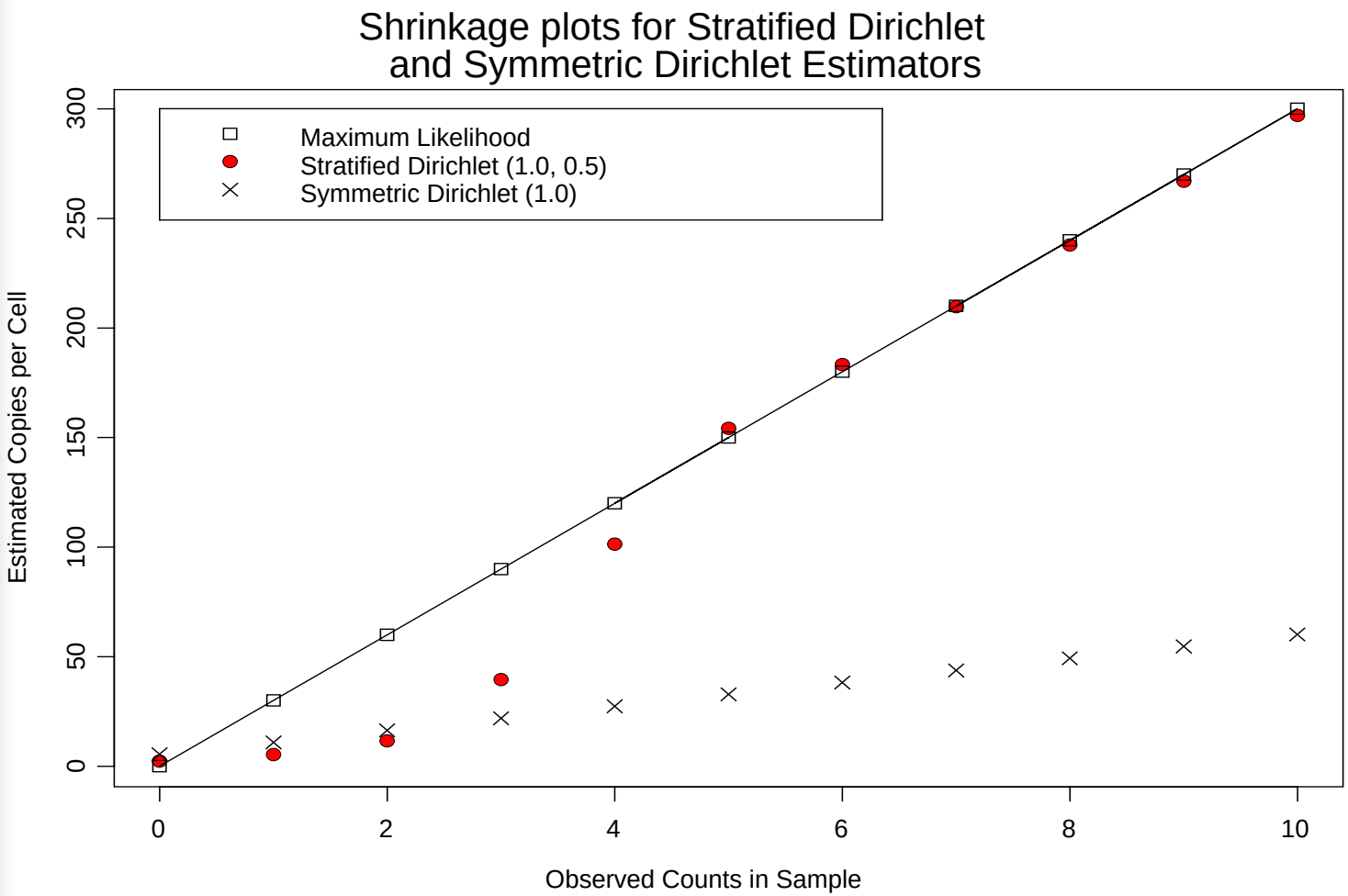
Shrinkage Curves



Shrinkage Curves



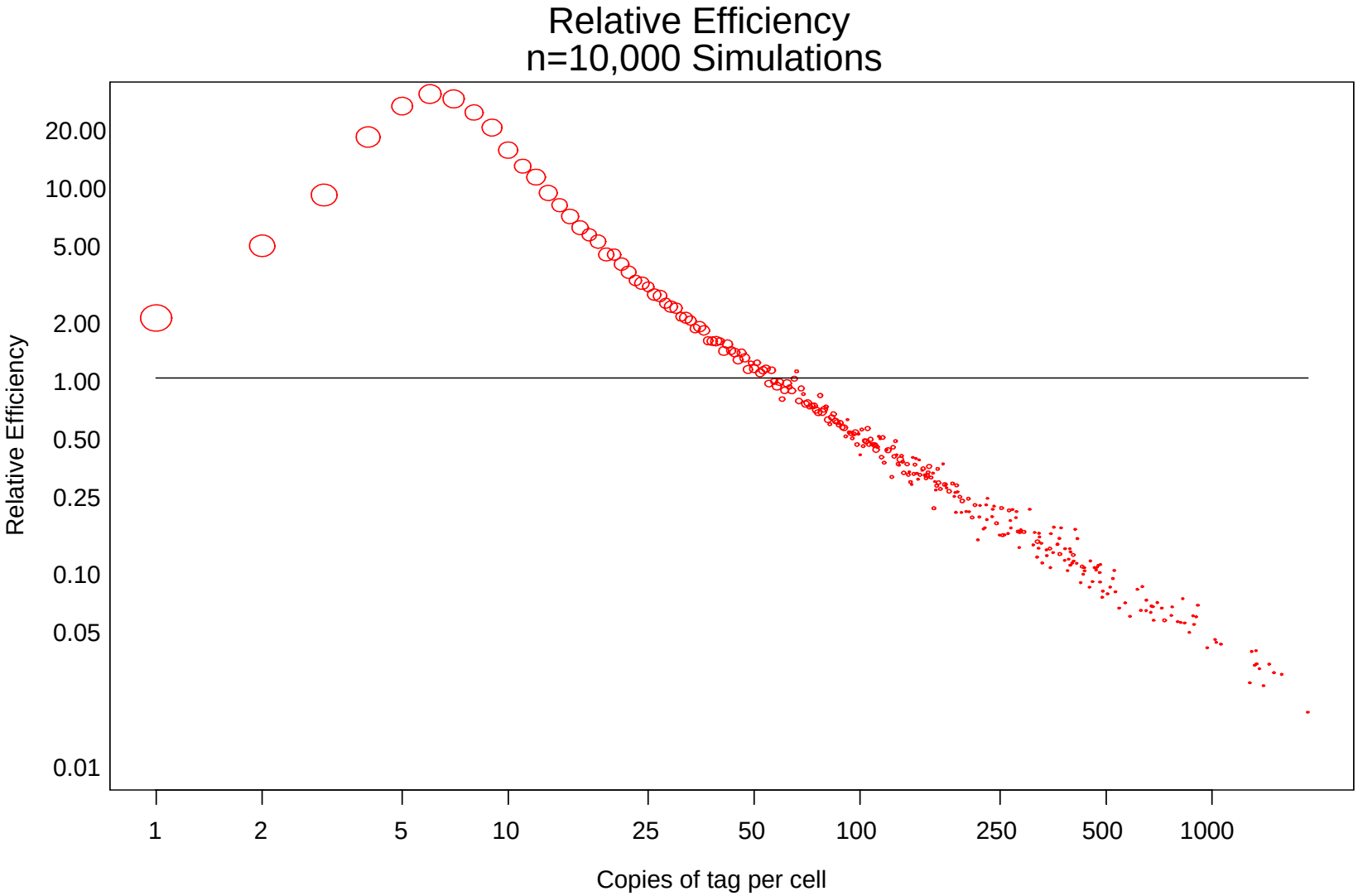
Shrinkage Curves



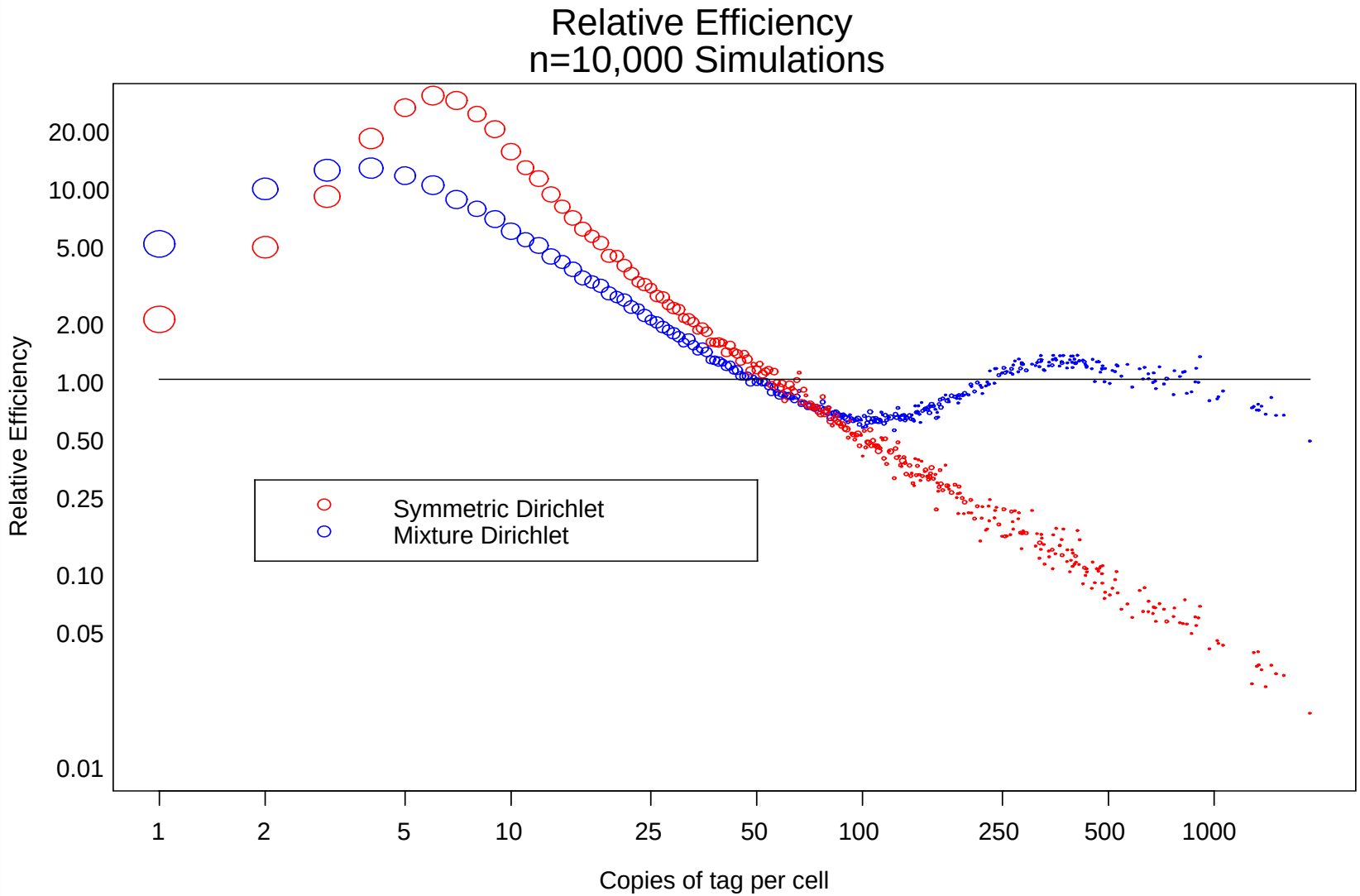
Simulation Study: SAGE data

- SAGE samples generated from “true population” of mRNA transcripts (k=45,984)
  - 100 datasets with n = 10,000
  - 100 datasets with n = 50,000
- 3 Estimators evaluated by MSE, IMSE

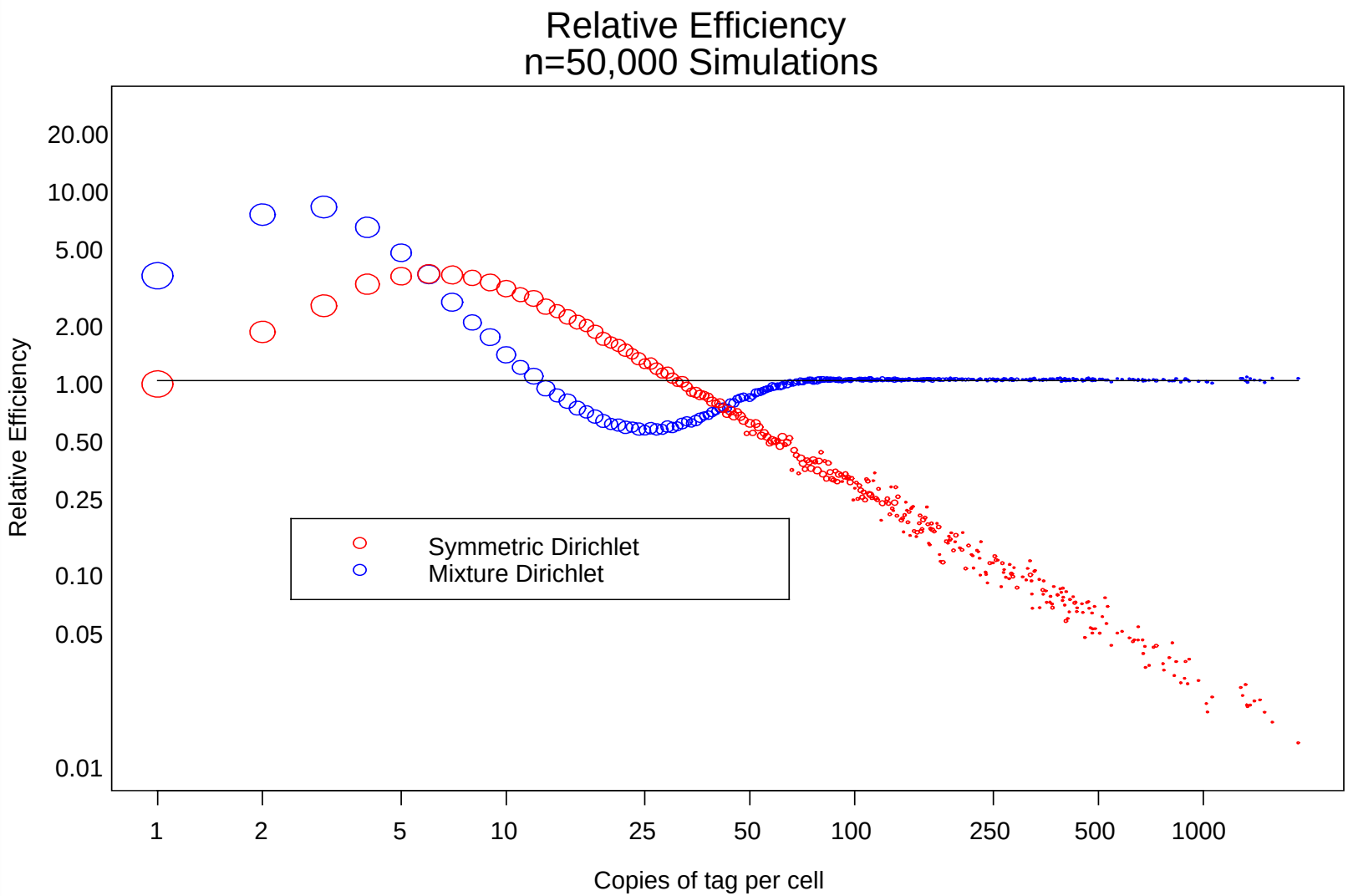
Simulation Results (*n=10,000*)



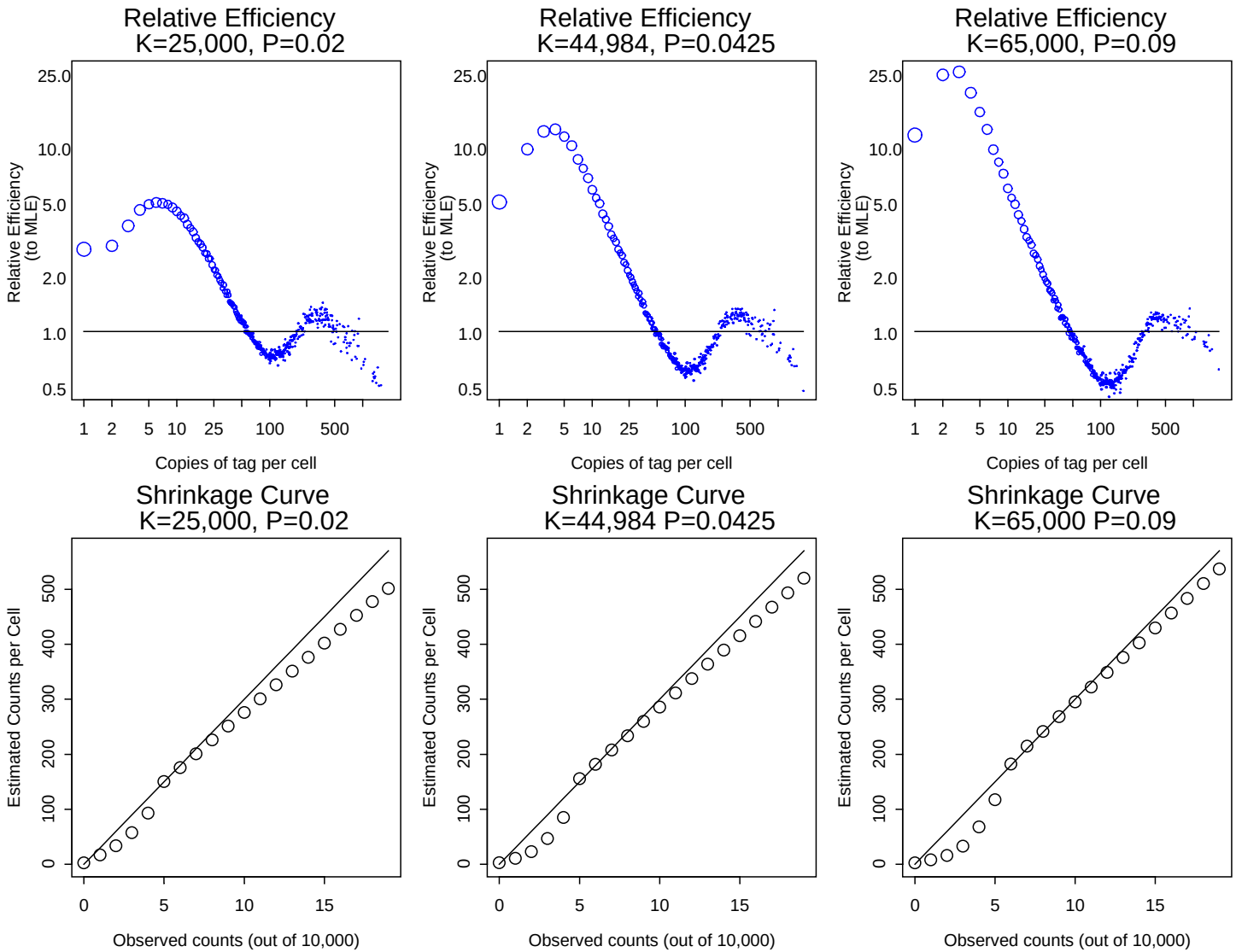
Simulation Results (*n=10,000*)



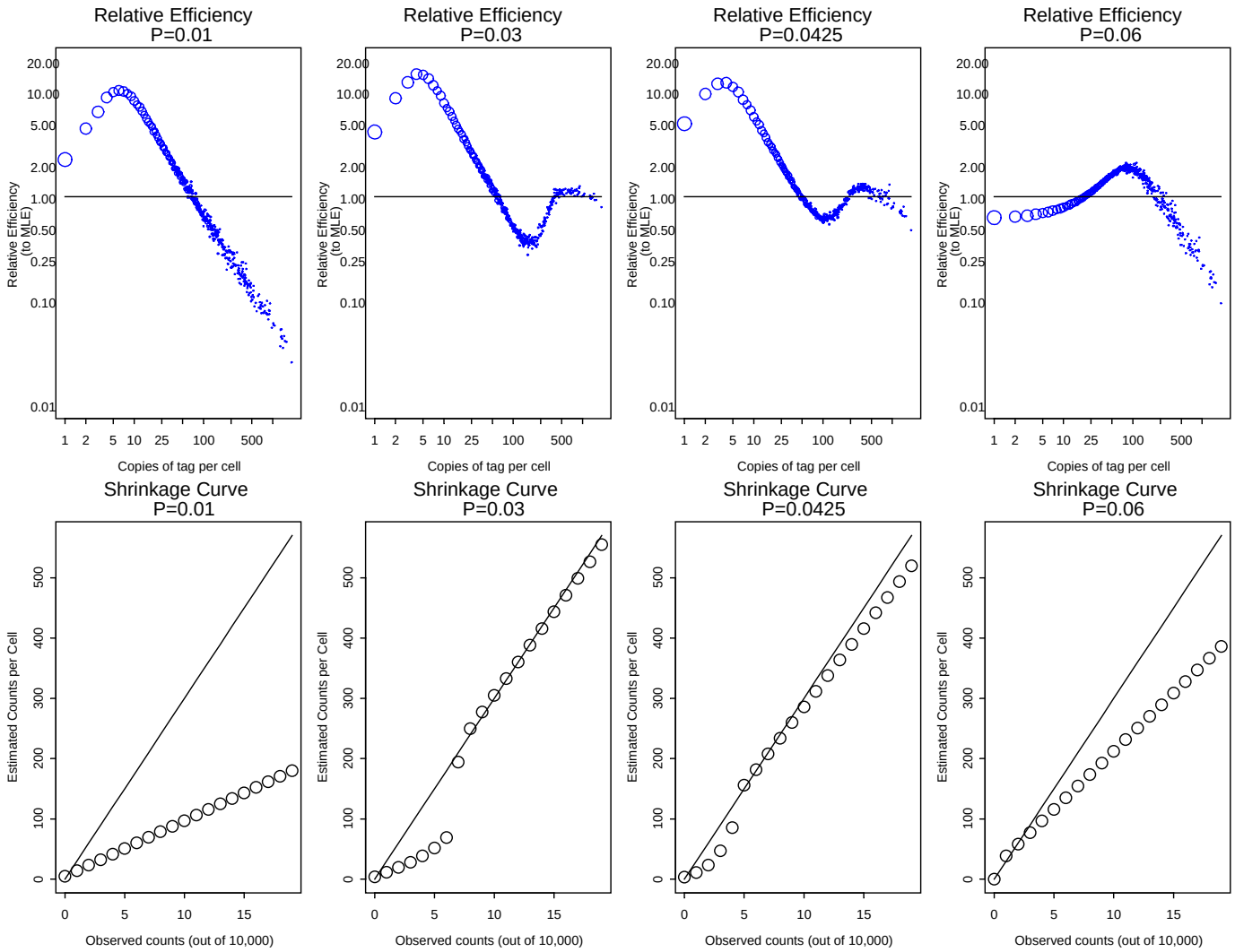
Simulation Results (*n=50,000*)



Sensitivity to k



Sensitivity to P



Conclusions

- SAGE: Method for quantifying gene expression
  - Yields multinomial data
    - Many scarce, few abundant genes
    - Sample size ~ number of genes
  - MLE can be improved upon by taking advantage of known prior information (Shrinkage estimation)
  - Improved estimators for scarce genes – some trade off in intermediate region.
- Benefit of joint probability model: Inference
  - Posterior probabilities of fold-differences
  - Model can be extended with other features
- Sensitivity to (P,θ_A,θ_S) a practical problem