Bayesian Screening for Group Differences in High-Throughput Data

Eric F. Lock

University of Minnesota Division of Biostatistics

Joint work with DB Dunson, Duke University

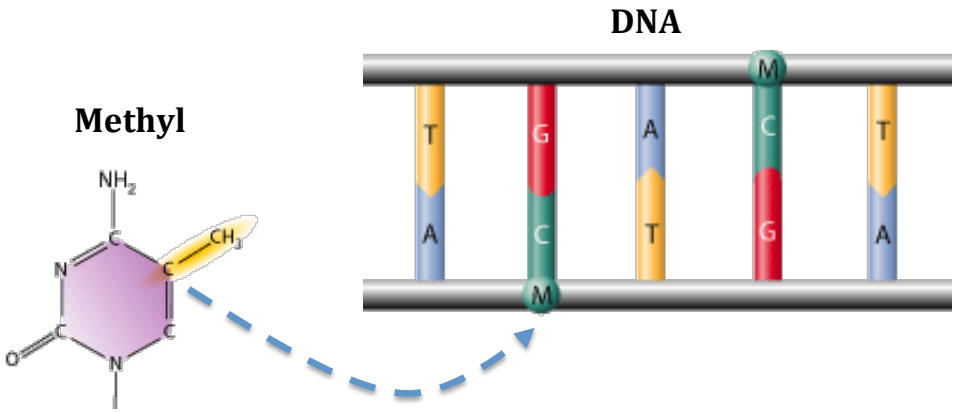
Iowa State Dept of Statistics, 11/16/2015

Organization

- Motivating example: TCGA methylation
  - Methylation array data
  - Distributional model
  - Two-group screening
- Comparison with other methods
- General framework & theory
  - General testing framework
  - Asymptotic forms
  - Consistency

DNA Methylation

- Methyl binds to CpG (cytosine-phosphate-guanine) sites



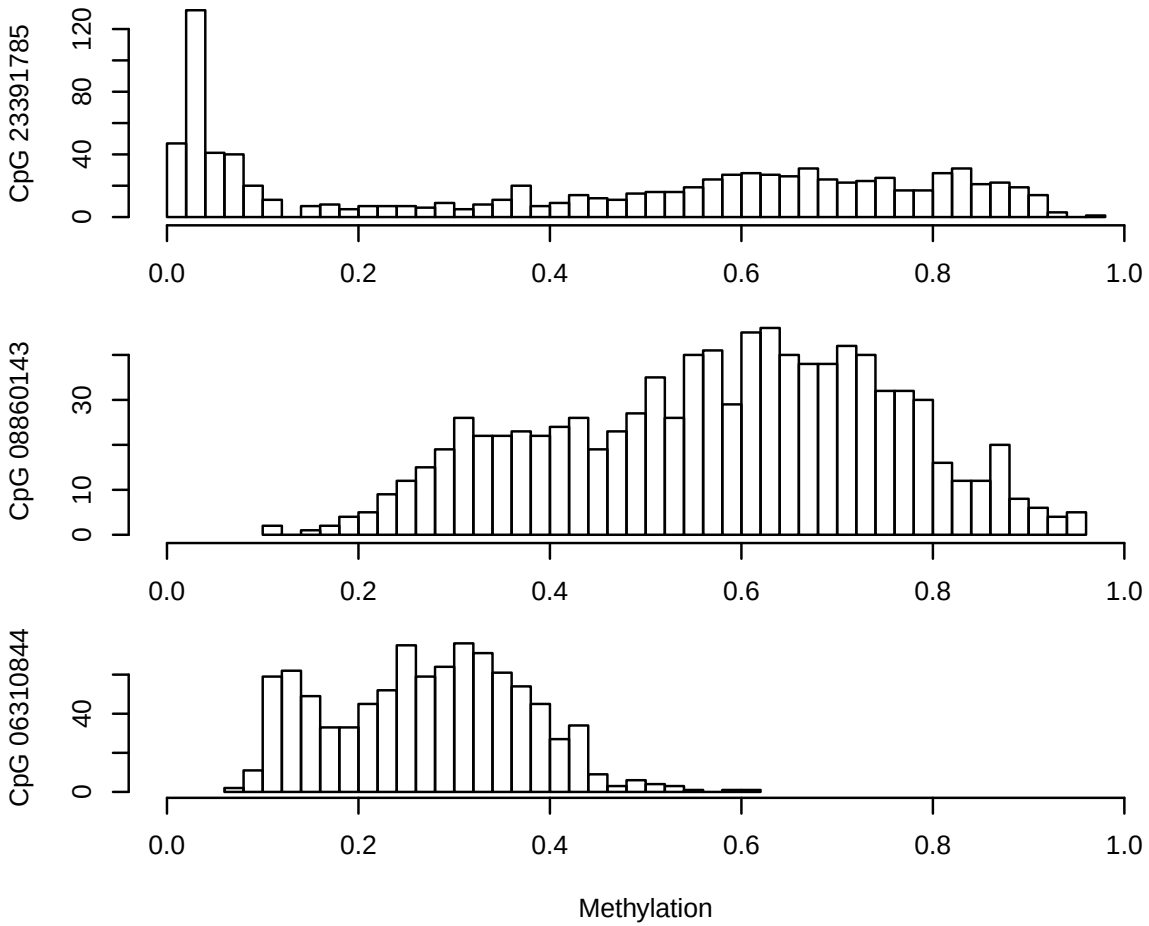
- Over 25 million CpG sites in human genome
- Methylation varies over sites / individuals / cell types
- Can affect gene transcription

TCGA array data

- $N = 597$ breast cancer tumor samples
  - From The Cancer Genome Atlas project
- Methylation measured for $M = 21,986$ CpG sites
  - Illumina 27 Beadchip array
  - Measurements from 0 (no methylation) to 1 (fully methylated)
- Goal: study role of methylation in clinical heterogeneity
  - Basal ($N_0 = 112$) vs. non-Basal ($N_1 = 485$) tumor subtypes

Example distributions

- Distribution of methylation values for select CpG sites



Kernel mixtures

- Model distribution of CpG m ($m = 1, \ldots, M$) as a mixture:

$$x_{mn} \sim \sum_{k=1}^{K} \pi_{mk} F_k$$

  - $\{F_k\}_{k=1}^{K}$ are shared kernels
  - $\Pi_m = \{\pi_{mk}\}_{k=1}^{K}$ are CpG-specific weights

- F_k is Normal(μ_k, σ_k) truncated between 0 and 1

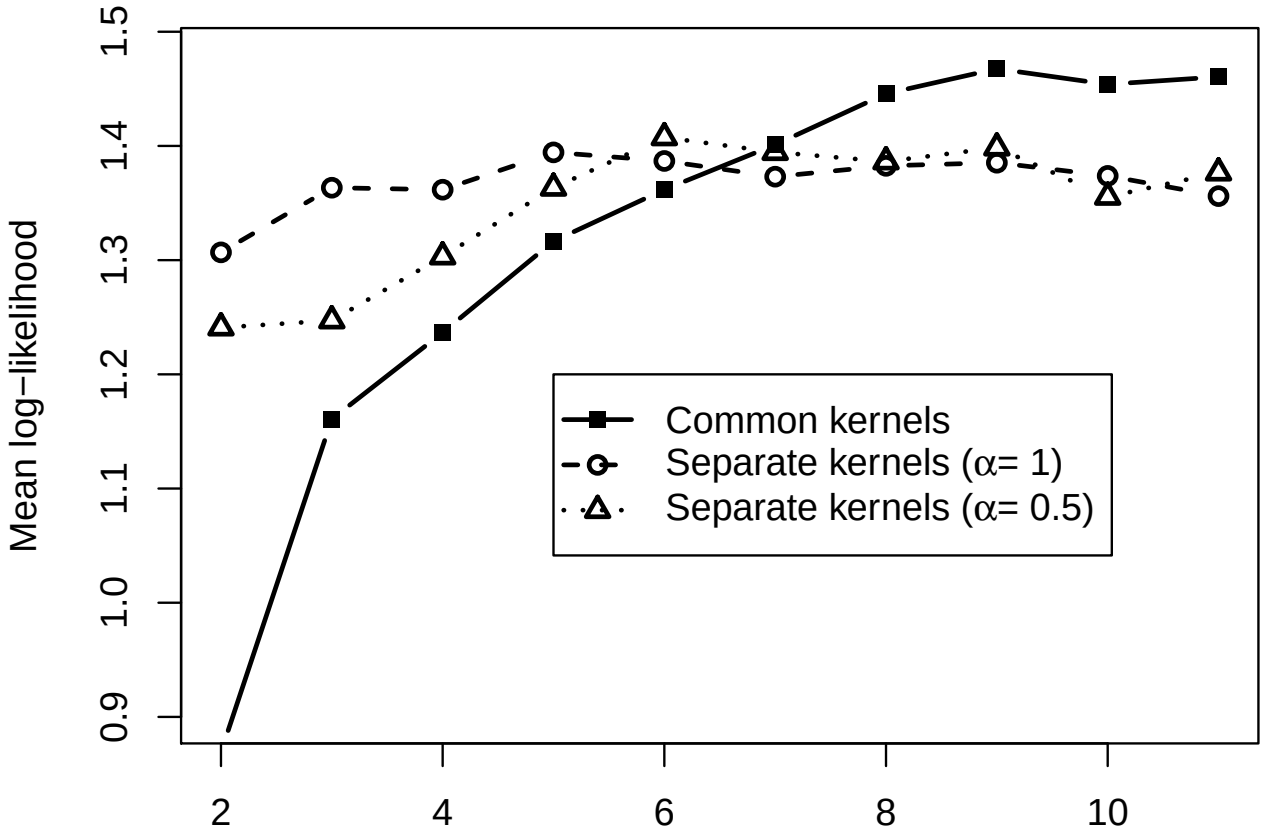
Bayesian estimation

- Use normal-inverse-gamma prior for (μ_k, σ_k)'s
- Use Dirichlet(α) prior for Π_m's
- Gibbs sample from conditional posteriors of
  - $\{(\mu_k, \sigma_k)\}_{k=1}^{K}$
  - $\{\Pi_m\}_{m=1}^{M}$
  - Kernel memberships $\{C_m\}_{m=1}^{M}$
- Estimate α via maximum likelihood during sampling

Choice of K

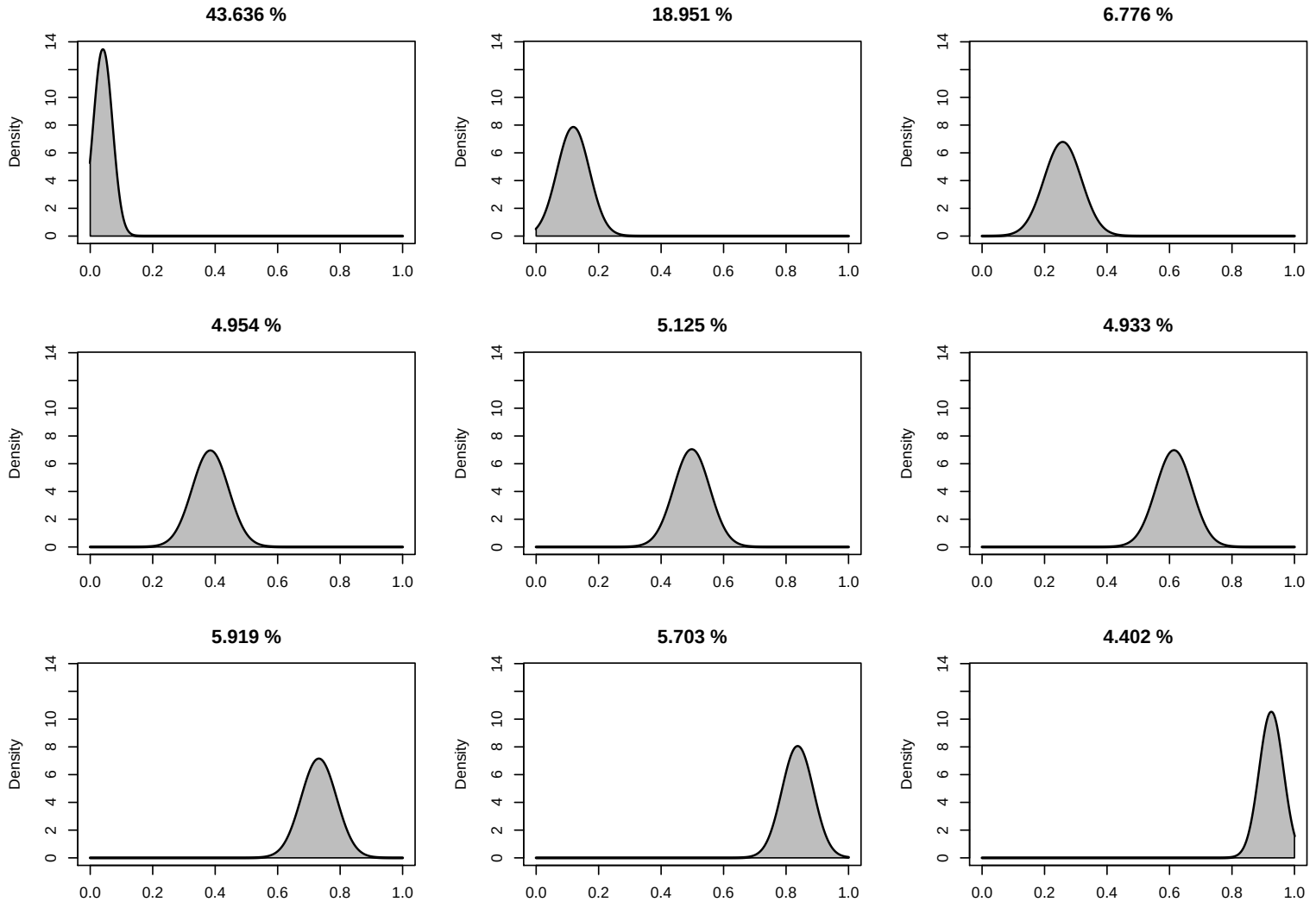
- Choose K to maximize likelihood under cross validation.
- For fixed K:
  - Estimate $F_1, \ldots, F_K$, and α from a sub-sample of CpGs
  - For each remaining CpG:
    - Hold out a random observation
    - Estimate kernel weights on $N-1$ remaining observations
    - Compute log-density for held out sample
  - Consider mean log-density for all held-out observations

Cross-validated log-likelihood

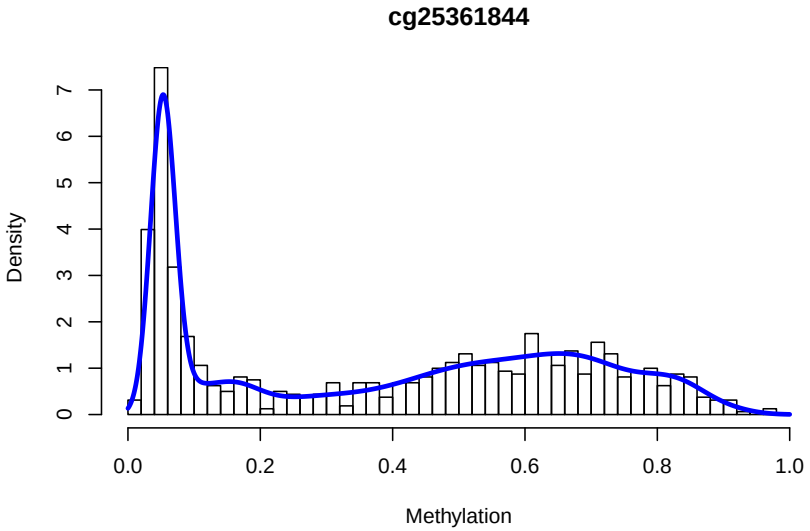


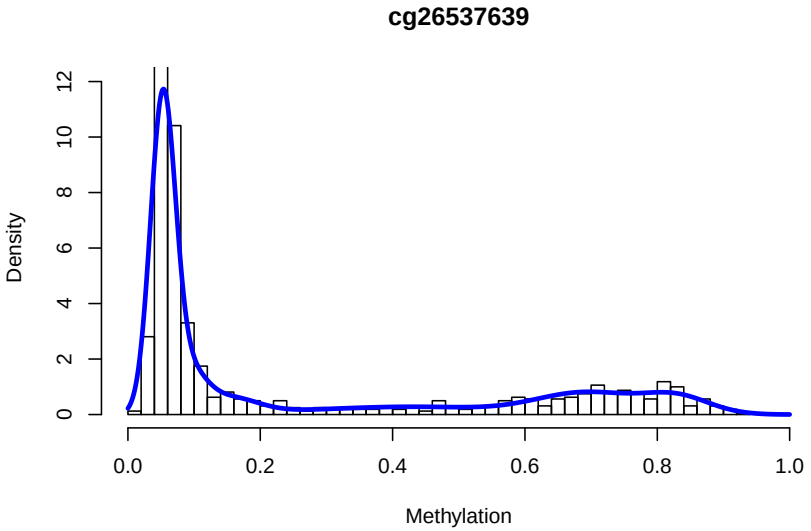
- Choose $K = 9$

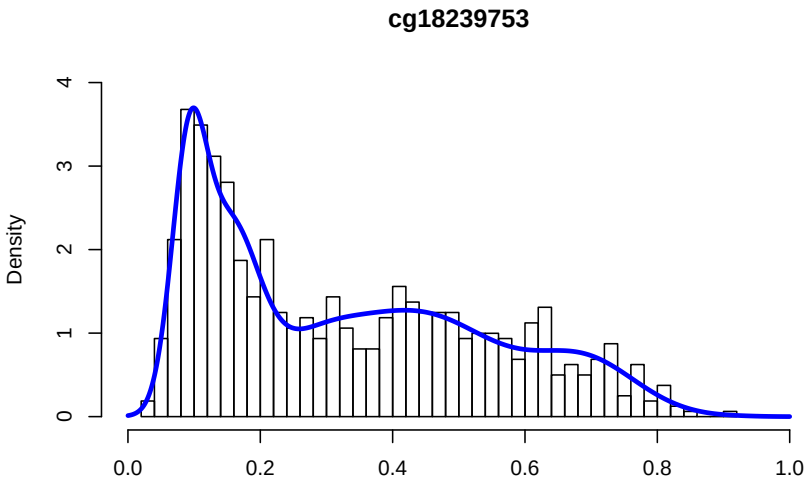
Kernel distributions

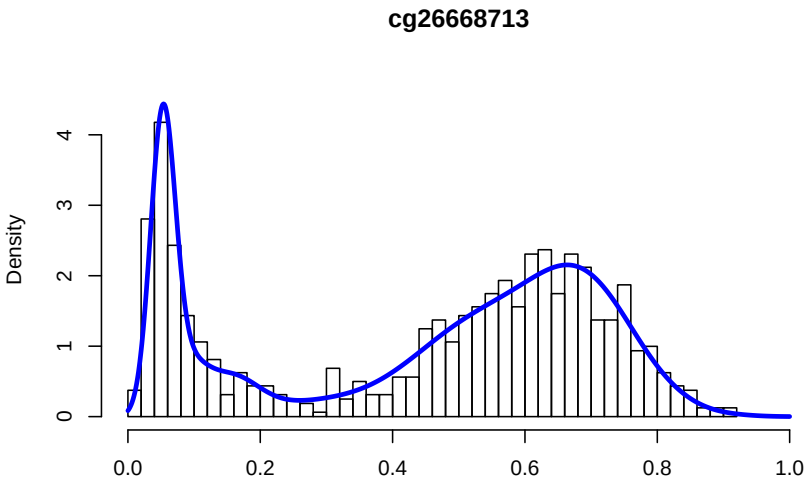


Fitted mixture examples









Test for group equality

- Compare Basal vs. non-Basal tumor subtypes at each CpG
  - Assess whether subtype distributions are different
- Subtype distributions $F_m^{(0)}, F_m^{(1)}$ are mixture of common kernels

$$F_m^{(0)} = \sum_{k=1}^{K} \pi_{mk}^{(0)} F_k \text{ and } F_m^{(1)} = \sum_{k=1}^{K} \pi_{mk}^{(1)} F_k,$$

- For each m test

$$H_{0m} : \pi_{mk}^{(0)} = \pi_{mk}^{(1)} \text{ for all } k$$
$$H_{1m} : \pi_{mk}^{(0)} \neq \pi_{mk}^{(1)} \text{ for some } k.$$

Bayesian framework

- Estimate and fix $F_1, \ldots, F_K$, and α as before.
- Under H_{0m}, $\Pi_m^{(0)} = \Pi_m^{(1)} = \Pi_m \sim \text{Dirichlet}(\alpha)$
- Under H_{1m}, $\Pi_m^{(0)}, \Pi_m^{(1)} \sim \text{Dirichlet}(\alpha)$ are independent
- P_0 is shared prior probability of equality at a given CpG
  - P_0 has Uniform$(0, 1)$ prior (see Scott & Berger 2010)

Posterior computation

- The full conditional posterior probability for H_{0m} is

$$\frac{P_0\beta(\alpha)\beta(\vec{n}_m+\alpha)}{P_0\beta(\alpha)\beta(\vec{n}_m+\alpha)+(1-P_0)\beta(\vec{n}_m^{(0)}+\alpha)\beta(\vec{n}_m^{(1)}+\alpha)}.$$

  - $\vec{n}_m^{(i)}$ gives number of realizations in group i from each kernel
  - $\vec{n}_m = \vec{n}_m^{(0)} + \vec{n}_m^{(1)}$
  - β is the multivariate beta function

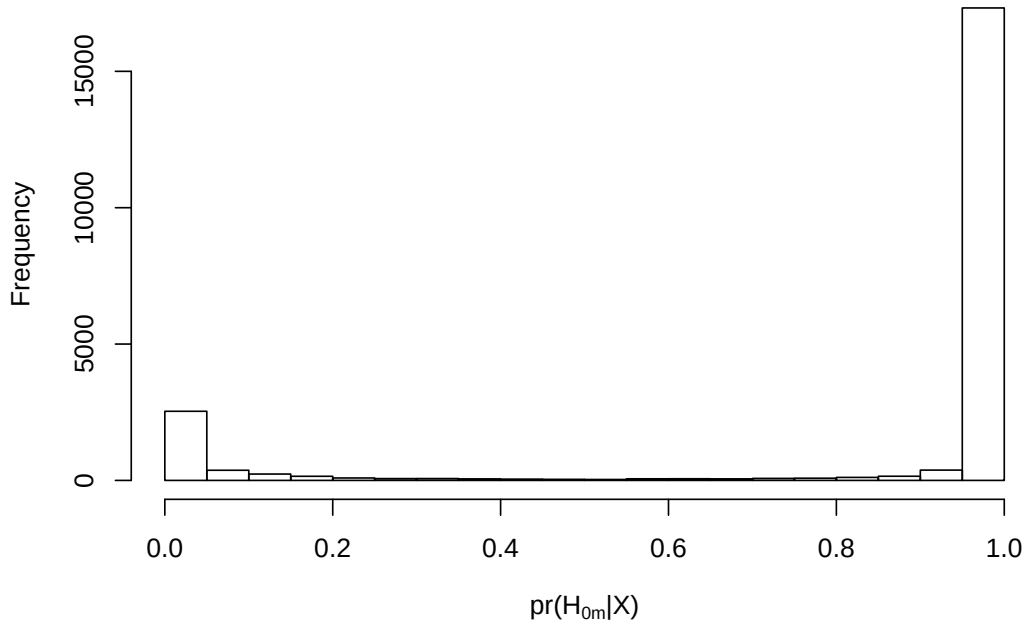
$$\beta(\alpha) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}.$$

Posterior computation

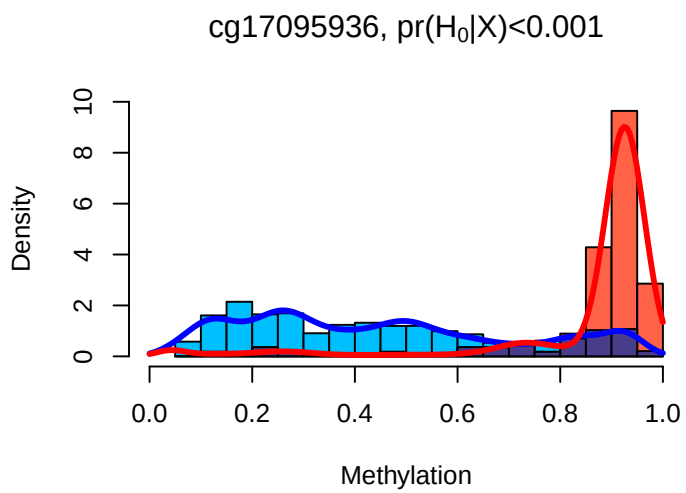
- In practice $\vec{n}_m^{(0)}$, $\vec{n}_m^{(1)}$ are unknown
- Kernel memberships are inferred probabilistically
- Gibbs sample from conditional posteriors of
  - $\{\Pi_m^{(0)}, \Pi_m^{(1)}\}_{m=1}^M$
  - $\{\vec{n}_m^{(0)}, \vec{n}_m^{(1)}\}_{m=1}^M$
  - $\{P(H_{0m} \mid \vec{n}_m^{(0)}, \vec{n}_m^{(1)})\}_{m=1}^M$
  - P_0
- Average over conditional posterior probabilities for H_{0m}

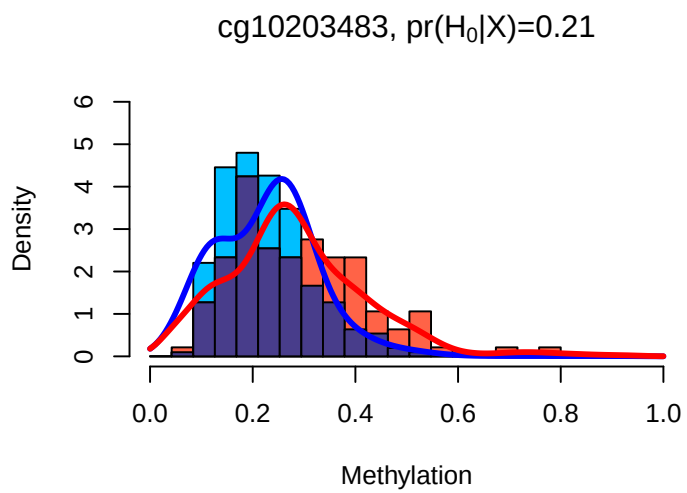
Basal vs. non-Basal groups

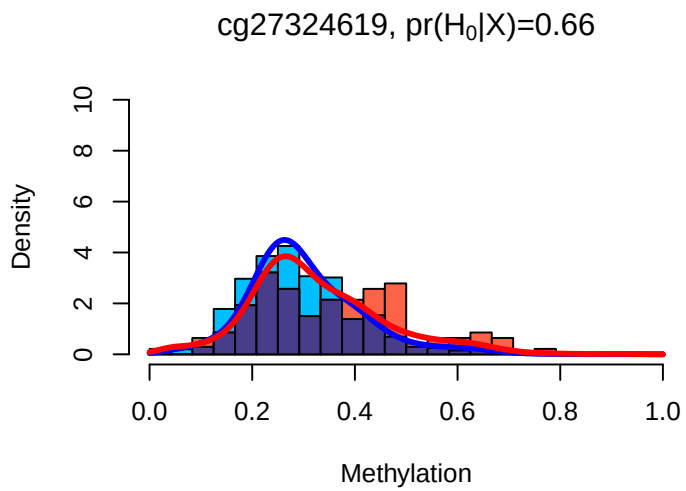
- Prior probability of equality: $\hat{P}_0 = 0.82$
- Distribution of posterior probabilities:

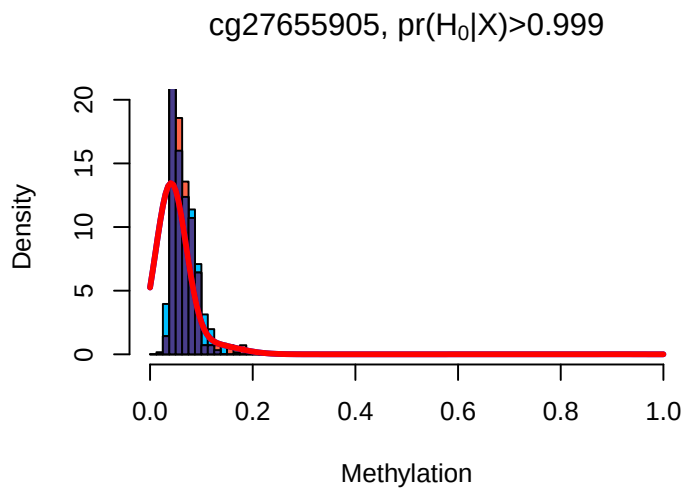


Basal vs. non-Basal groups



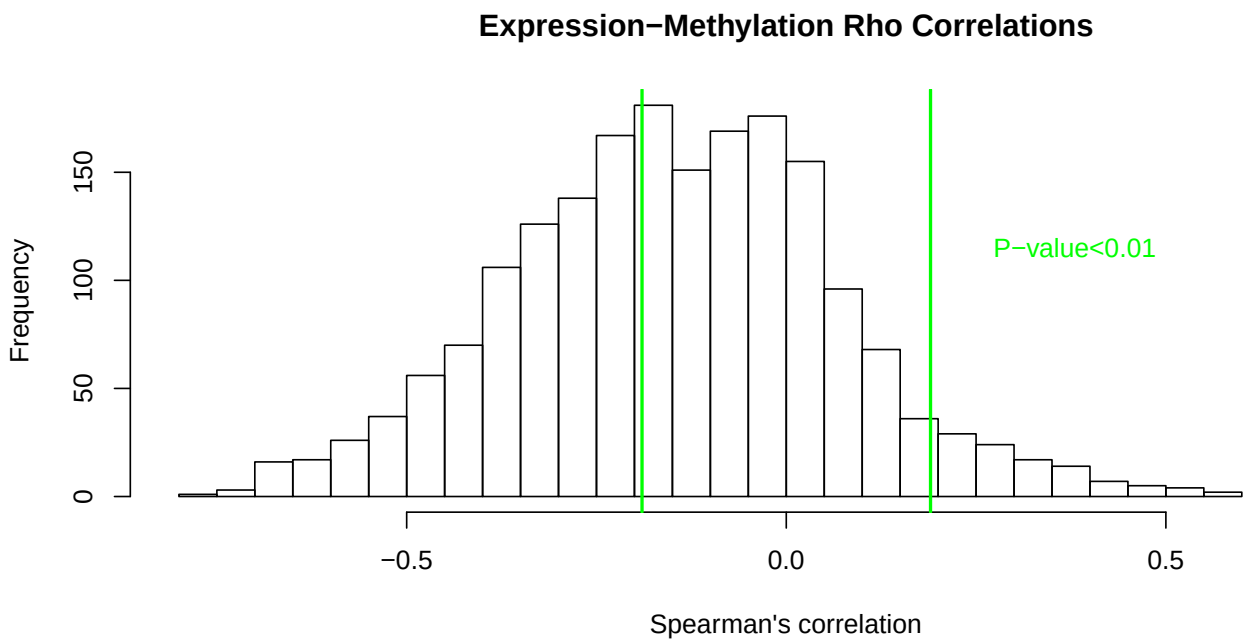






Basal vs. non-Basal groups

- 2117 CpG sites with $P(H_{0m}|X) < 0.01$
- Consider association with expression at their gene:



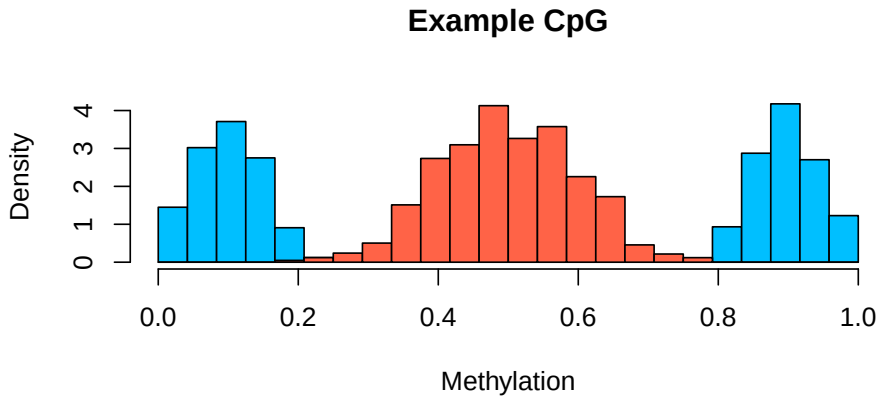
- Negative association & in PAM50 signature (Parker, 2009):
  - *MYBL2*, *EGFR*, *MIA*, *SFRP1* and *MLPH*

Related work: Methylation

- Multi-modality of methylation widely noted
  - Qiu & Zhang 2012, Izirray et al. 2008, Gervin et al 2011.
- Arbitrary thresholds define "methylated" vs "unmethylated"
  - Qiu & Zhang 2012 use 0.2, Chen et al. 2011 use 0.33
- Mixture models have been used for clustering
  - Kormaksson et al. 2012, Zhang et al 2012
- For group comparisons, t- and Wilcoxon tests most common
  - Bock 2012, Laird 2013

Related work: Methylation

- General tests for distributional equality are rarely used
- But they are well motivated...
  - Cancer & normal cells show different variability (Hansen 2011)
  - Groups may have differential "stability" across cells:



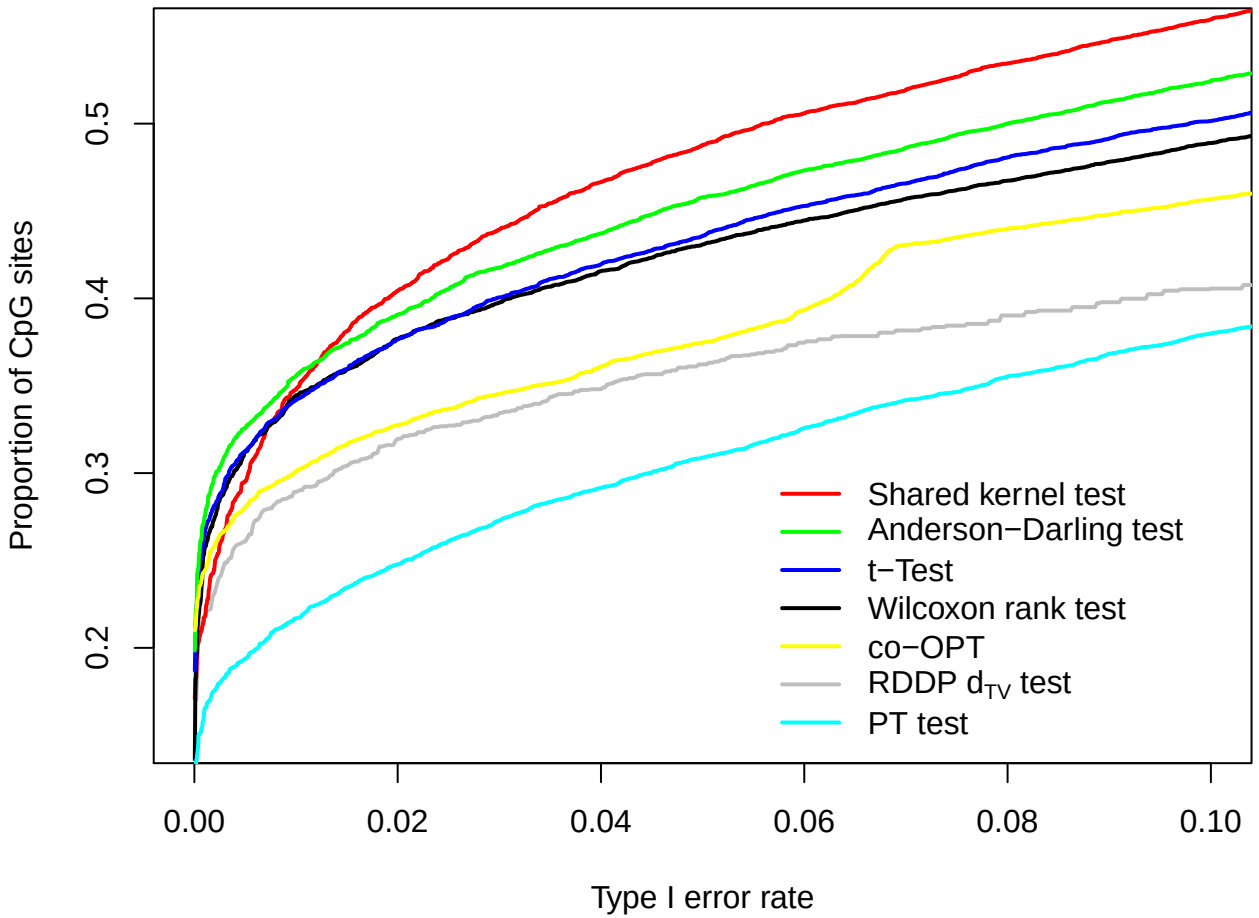
Related work

- Frequentist tests for distributional equality
  - Anderson-Darling, Shapiro-Wilk
- Bayesian nonparametric tests using Dirichlet processes
  - Dunson & Peddada 2008, Pennell & Dunson 2008
- Bayesian nonparametric tests using Polya trees
  - Ma & Wang 2011, Holmes et al 2014

Methods comparison for TCGA data

- Apply several methods to TCGA data
  - t-test, Wilcoxon test, Anderson-Darling test, Dunson & Peddada (RDDP), Ma & Wang (co-OPT), Holmes et al. (PT), and shared kernel test with fixed $P_0 = 0.5$.
- Permute class labels for each CpG and apply again.
- Permutation creates a null model to assess type I error
- Compare distribution of results (p-values or Bayes factors) for true and permuted data.

Methods comparison for TCGA data



Abstract testing framework

- Two distributions $F^{(0)}, F^{(1)}$ are mixtures

$$F^{(0)} = \sum_{k=1}^{K} \pi_k^{(0)} F_k \text{ and } F^{(1)} = \sum_{k=1}^{K} \pi_k^{(1)} F_k,$$

- Test whether $\pi_k^{(0)} = \pi_k^{(1)} \ \forall$ k.
- $F^{(0)}, F^{(1)}$ describe two populations with same strata
  - Test whether strata have different proportions

Abstract testing framework

- If strata/kernel memberships are known:
  - Test for association in $2 \times K$ table
  - Frequentist approaches: Chi-Square, Fisher's exact test
  - Bayesian Approaches: Good & Crook 1987, Albert 1997
- If memberships (and perhaps the F_k's) are unknown:
  - Little statistical literature
  - Addressed partly in Xu et al 2010

Asymptotic forms

- Consider behavior of the full conditional for H_0:

$$\frac{P_0\beta(\alpha)\beta(\vec{n}+\alpha)}{P_0\beta(\alpha)\beta(\vec{n}_m+\alpha)+(1-P_0)\beta(\vec{n}^{(0)}+\alpha)\beta(\vec{n}^{(1)}+\alpha)}$$

  as $N \to \infty$.

- For the following assume:
  - $\lambda_0 = \frac{N_0}{N_0+N_1}$ is fixed
  - $\vec{n}^{(0)}, \vec{n}^{(1)}$ are known

Asymptotic forms

- THEOREM: Can derive a closed asymptotic form for the full conditional
- CORROLARY: Can fully characterize asymptotic distribution under H_0 and H_1
- Under $H_0 : \Pi^{(0)} = \Pi^{(1)} = \Pi$, the log Bayes factor has order

$$\frac{K-1}{2}\log(N) + O_p(1)$$

- Under $H_1 : \Pi^{(0)} \neq \Pi^{(1)}$, let $\Pi^* = \lambda_0 \Pi^{(0)} + (1-\lambda_0)\Pi^{(1)}$. The log of the Bayes factor has order

$$-N \sum \left\{ \lambda_0 \pi_k^{(0)} \log\left(\frac{\pi_k^{(0)}}{\pi_k^*}\right) + (1-\lambda_0)\pi_k^{(1)} \log\left(\frac{\pi_k^{(1)}}{\pi_k^*}\right) \right\} + O_p\left(N^{1/2}\right),$$

Asymptotic forms

- Posterior probability of H_0 converges
  - Sublinearly to 1 under H_0
  - Exponentially to 0 under H_1
- Such rates have been observed for several Bayesian tests
  - Kass & Raftery 1995; Walker 2004; Johnson & Rossell 2010.
- Often such models are "local prior densities"
  - The parameter space under H_0 has positive density under H_1

Asymptotic behavior simulation

- Simulate hundreds of two-group univariate Gaussian mixture datasets
- Vary N for each simulated dataset
- Each simulation dataset generated under either H_0 or H_1
- Gibbs sample to estimate kernels, weights, and $\text{pr}(H_0)$

Asymptotic behavior simulation (details)

1. Draw N uniformly on a log-scale from 10 to $1,000,000$.
2. Draw K uniformly from $\{2, \ldots, 9\}$.
3. Draw $\mu_1, \ldots, \mu_K$ independently from $\text{Un}(0, 1)$.
4. Draw $\sigma_1, \ldots, \sigma_K$ independently from $\text{Un}(0, \frac{1}{K})$
5. Draw H_0 from Bernoulli(0.5)
6. If $H_0 = 1$
   - Draw Π from a uniform, K-dimensional Dirichlet distribution
   - For $n = 1, \ldots, N$ assign x_n to class 0 or 1 with equal probability
   - Draw $x_1, \ldots, x_N \in \mathbb{X}$ from $\sum_{k=1}^{K} \pi_k \text{Tnorm}(\mu_k, \sigma_k, [0, 1])$,
7. If $H_0 = 0$
   - Draw $\Pi^{(0)}$ and $\Pi^{(1)}$ independently from a uniform, K-dimensional Dirichlet distribution
   - For $n = 1, \ldots, N$ assign x_n to class 0 or 1 with equal probability
   - Draw $x_1, \ldots, x_{N_0} \in \mathbb{X}^{(0)}$ from $\sum_{k=1}^{K} \pi_k^{(0)} \text{Tnorm}(\mu_k, \sigma_k, [0, 1])$
   - Draw $x_1, \ldots, x_{N_1} \in \mathbb{X}^{(1)}$ from $\sum_{k=1}^{K} \pi_k^{(1)} \text{Tnorm}(\mu_k, \sigma_k, [0, 1])$.

Asymptotic behavior simulation

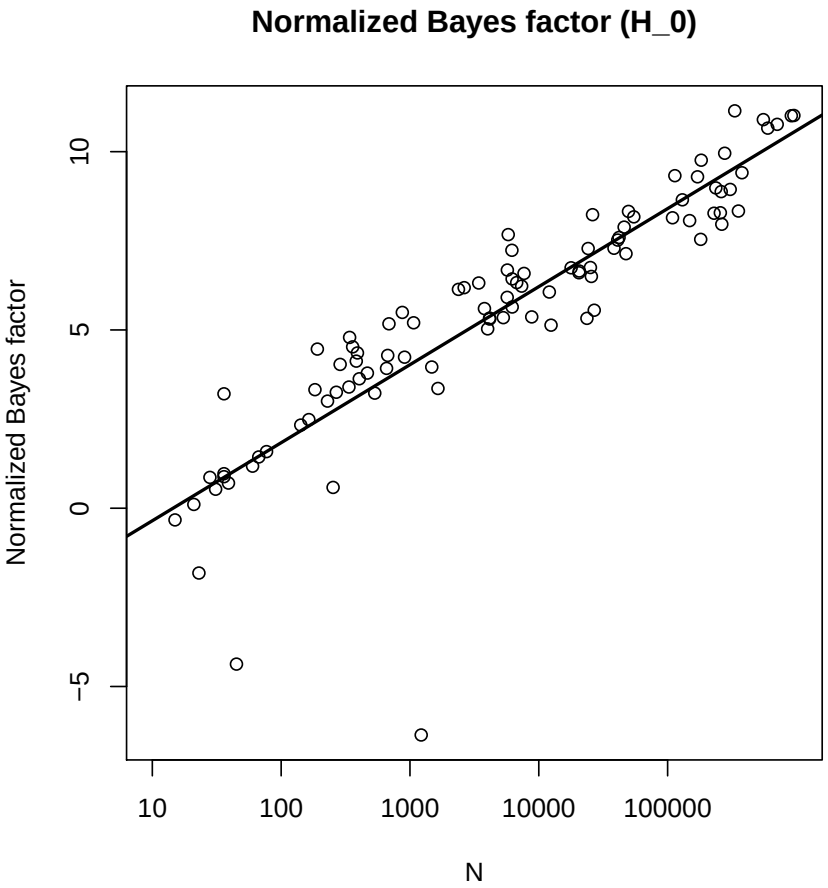
- Normalize log Bayes factor by dominant asymptotic term
- For H_0 simulations:

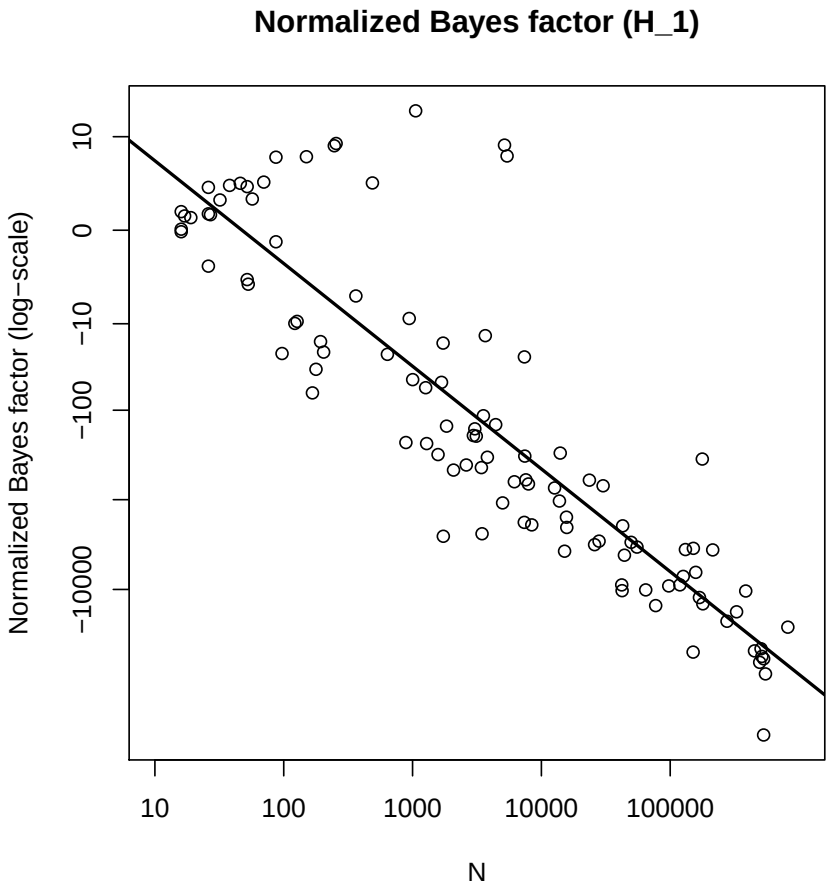
$$\frac{2}{K-1}\log\left\{\frac{\text{pr}(H_0|X)}{\text{pr}(H_1|X)}\right\}$$

- For H_1 simulations:

$$\frac{\log\left\{\frac{\text{pr}(H_0|X)}{\text{pr}(H_1|X)}\right\}}{\sum\left\{\lambda_0\pi_k^{(0)}\log\left(\frac{\pi_k^{(0)}}{\pi_k^*}\right)+(1-\lambda_0)\pi_k^{(1)}\log\left(\frac{\pi_k^{(1)}}{\pi_k^*}\right)\right\}}.$$

Simulation results





Consistency under misspecification

- Bayesian context:
  - True distribution is not within support of prior
- E.g: data may not result from a finite Gaussian mixture
- Misspecified models not "fully" consistent
- May still be consistent as a test for distributional equality

Consistency under misspecification

- Use work of Kleijn & Van der Vaaart (2006)
- General behavior under Bayesian misspecification:
  - Let $\mathbb{F}$ be space of all distributions admitted by prior
  - Let F_0 be data generating distribution
  - Let F^* be distribution in $\mathbb{F}$ minimizing KL-divergence to F_0
  - Posterior concentrates on F^* as $N \to \infty$
- Little work on misspecification asymptotics for Bayesian tests

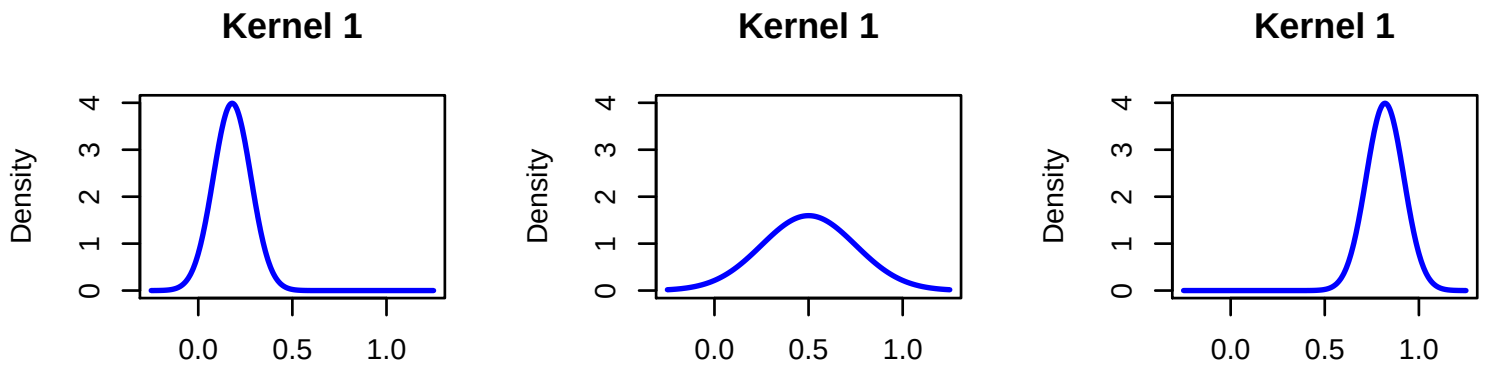
Misspecification for finite mixtures

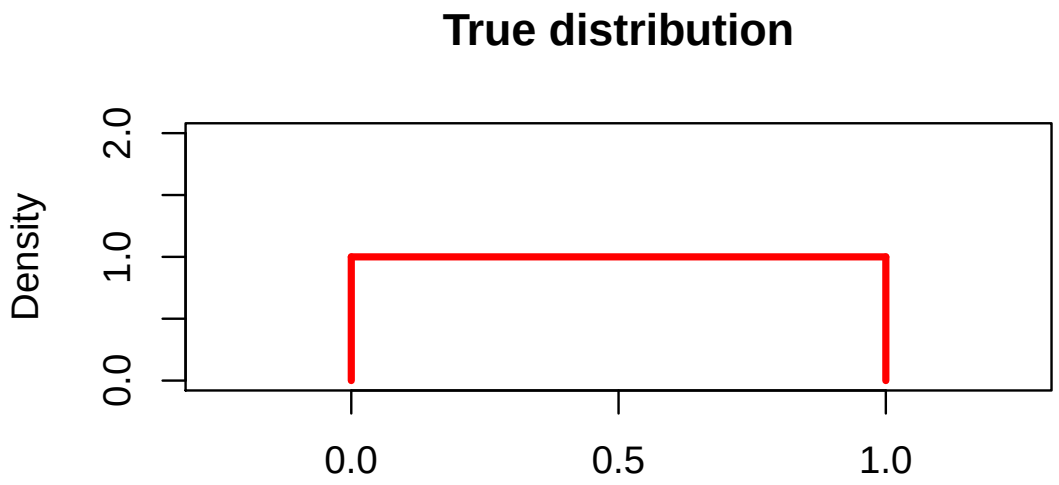
- Let $x_1, \ldots, x_N$ be independent with density f_0.
- Let $\mathbb{F}$ be define all convex combinations of densities $\{f_k\}_{k=1}^K$
- Let P define a prior with positive support over $\mathbb{F}$.
- Let $f^* = \underset{f \in \mathbb{F}}{\text{argmin}}\ \text{KL}(f_0 || f^*)$
- THEOREM: let $\Pi^* = (\pi_1^*, \ldots, \pi_K^*)$ be the component weights corresponding to f^*. Assume Π^* is unique in that $\sum \pi_k f_k = \sum \pi_k^* f_k = f^*$ only if $\Pi = \Pi^*$. Then, for any fixed $\epsilon > 0$,

$$\text{pr}(\Pi \in \mathbb{S}^{K-1} : ||\Pi - \Pi^*|| \geq \epsilon \mid x_1, \ldots, x_N) \to 0.$$

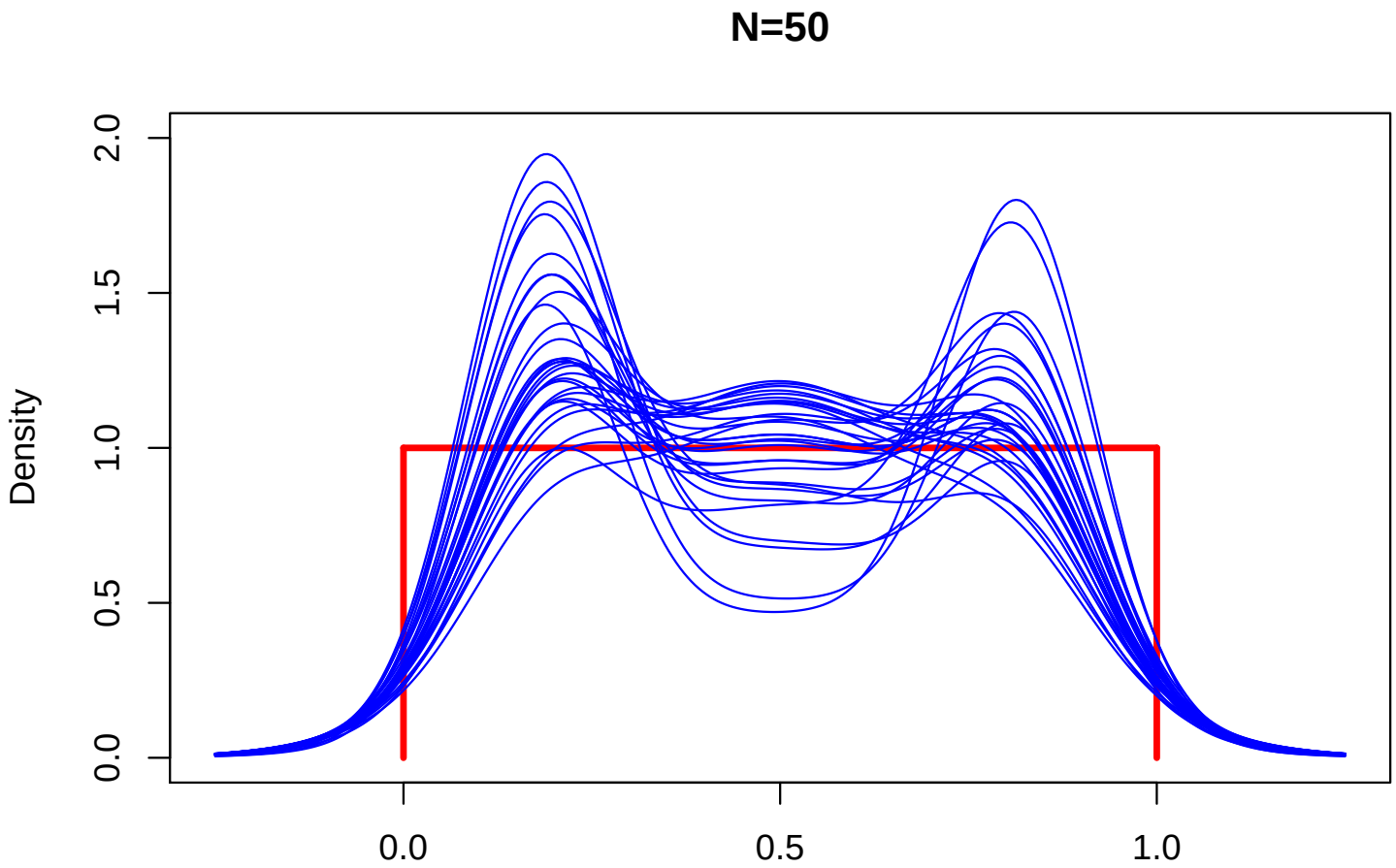
- Π^* is generally unique for normal $f_k's$ (Yakowitz 1968)

Illustrative example

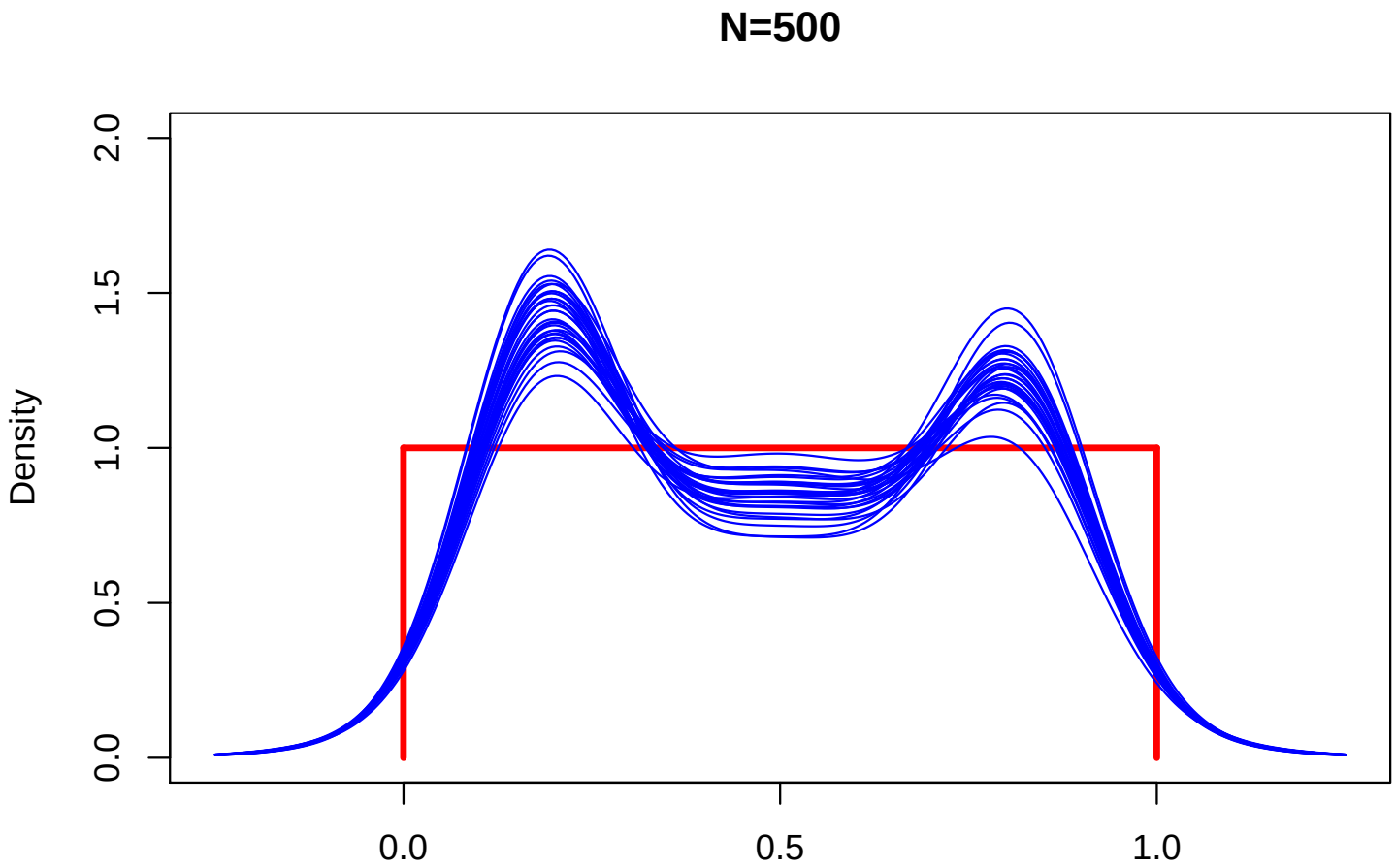




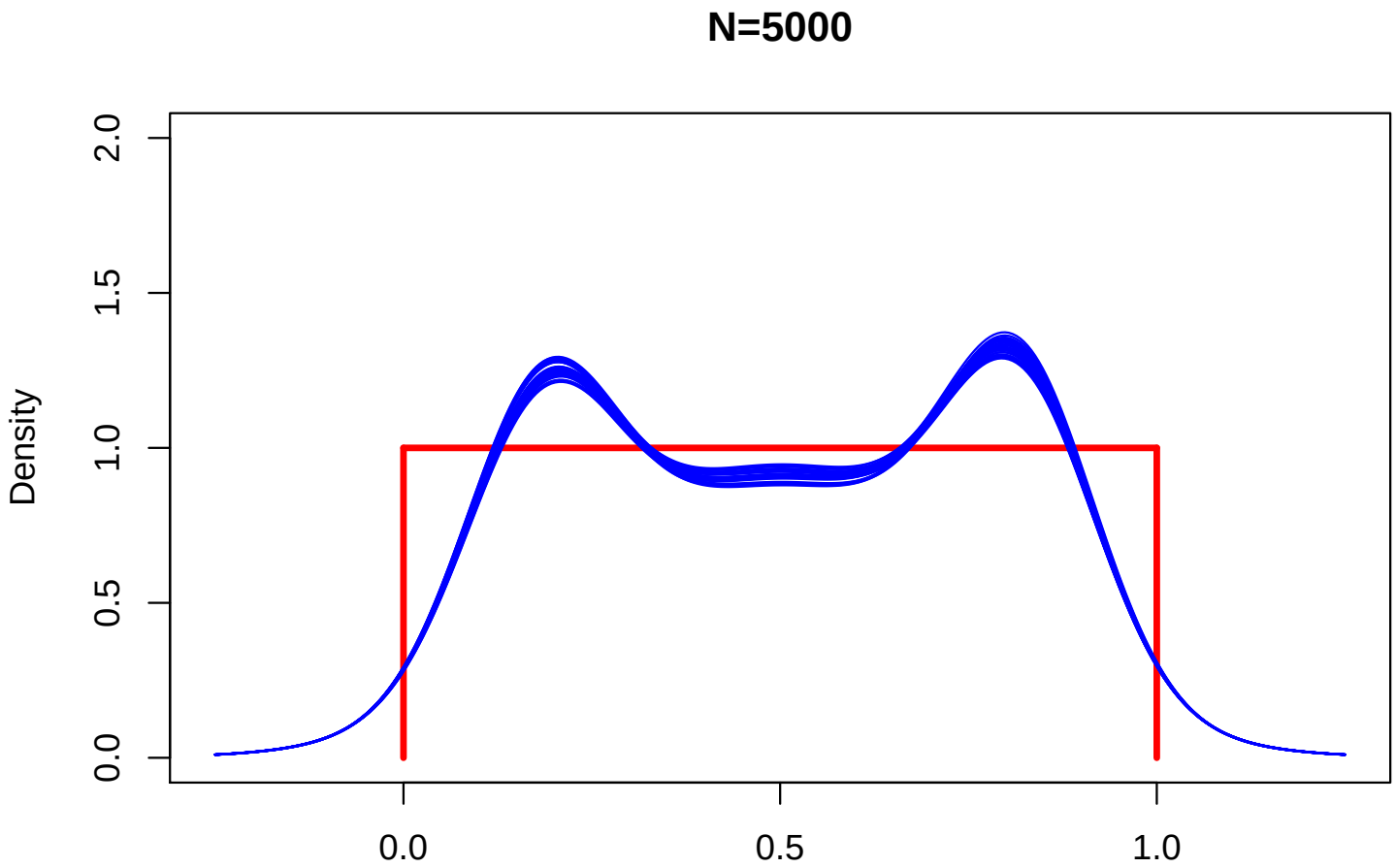
Illustrative example



Illustrative example



Illustrative example



Misspecification for finite mixtures

- REMARK: Assume $\pi_k^* > 0$ for $k = 1, \ldots, K$ and $\sum \pi_k^* = 1$. Then, $f^* = \sum \pi_k^* f_k$ achieves the minimum KL-divergence in $\mathbb{F}$ with respect to f_0 if and only if

$$\int \frac{f_1}{f^*} f_0 = \ldots = \int \frac{f_K}{f^*} f_0.$$

  If some $\pi_k^* = 0$, the minimum KL-divergence is achieved where $\int \frac{f_k}{f^*} f_0$ are equivalent for all $\pi_k^* > 0$.

Consistency under misspecification

- THEOREM: Assume $x_1^{(0)}, \ldots, x_{N_0}^{(0)}$ are independent with density $f^{(0)}$, $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ are independent with density $f^{(1)}$, and let

$$f^{*(0)} = \underset{f \in \mathbb{F}}{\text{argmin}}\ \text{KL}(f^{(0)} || f)\ ,\ \ f^{*(1)} = \underset{f \in \mathbb{F}}{\text{argmin}}\ \text{KL}(f^{(1)} || f).$$

  Under uniqueness assumptions for $f^{*(0)}$ and $f^{*(1)}$,

  - if $f^{(0)} = f^{(1)}$, $\text{pr}(H_0 \mid X) \to 1$ as $N \to \infty$ and
  - if $f^{*(0)} \neq f^{*(1)}$, $\text{pr}(H_0 \mid X) \to 0$ as $N \to \infty$.

Future directions

- Consider shared kernel model for other contexts
  - Negative binomial kernels for RNA-Seq data
- Extend to multi-group testing problems
- More sophisticated dependence models
  - Hierarchical model with gene-level $P_0's$
  - Spatial dependence
- Data-driven alternative hypotheses

Thank you!

- Reference:
  - EF Lock and DB Dunson. Shared kernel Bayesian screening. doi: 10.1093/biomet/asv032, 2015
- R code to reproduce application to TCGA data:
  - http://www.tc.umn.edu/~elock/MethTestingTCGA.zip
- Email: elock@umn.edu

Simulation study

- M variables and N observations
- Simulate data from a Gaussian mixture
- Mixture components shared across variables
- Two groups, with equal weights on $M \times P$ variables
- Five repeated simulations for each combination of
  - $M = \{10, 60, 360\}$
  - $N = \{30, 120, 480\}$
  - $P = 0.1, 0.2, \ldots, 0.9$

Data generating details

- Draw $\mu_1, \ldots, \mu_5$ independently from Ga$(1, 1)$.
- Draw $\sigma_1, \ldots, \sigma_5$ independently from Un$(0, 1/2)$.
- For variables $m = 1$ through $m = PM$, draw data under H_0
  - Draw Π from a uniform Dirichlet distribution
  - Draw $x_{m1}, \ldots, x_{mN}$ from $\sum_{k=1}^{K} \pi_k \mathrm{N}(\mu_k, \sigma_k)$.
- For variables $m = PM + 1$ through $m = M$, draw data for two groups of size $N/2$
  - Draw $\Pi^{(0)}$ and $\Pi^{(1)}$ independently from a uniform Dirichlet distribution
  - Draw $x_{m1}, \ldots, x_{m(N/2)}$ from $\sum_{k=1}^{K} \pi_k \mathrm{N}(\mu_k, \sigma_k)$.
  - Draw $x_{m(N/2+1)}, \ldots, x_{mN}$ from $\sum_{k=1}^{K} \pi_k \mathrm{N}(\mu_k, \sigma_k)$.

Simulation study

- For each simulated dataset perform
  - Shared kernels and shared estimate for P_0 among variables
  - Shared kernels among variables and fixed $P_0 = 0.5$
  - Independently estimated kernels and fixed $P_0 = 0.5$
  - The co-OPT method (Ma & Wang 2011)
- Compute Bayes error for each method:

$$\sum_{m=1}^{M} [\{1 - \mathbb{1}(H_{0m})\}\text{pr}(H_{0m} \mid X) + \mathbb{1}(H_{0m})\{1 - \text{pr}(H_{0m} \mid X)\}]/M.$$

Results

		M = 10	M = 60	M = 360
N = 30	Shared kernels and estimated P_0	**0.40** $\pm$ 0.03	**0.32** $\pm$ 0.02	**0.31** $\pm$ 0.02
	Shared kernels and $P_0 = 0.5$	**0.41** $\pm$ 0.02	**0.36** $\pm$ 0.02	**0.36** $\pm$ 0.01
	Separate kernels and $P_0 = 0.5$	**0.47** $\pm$ 0.02	**0.47** $\pm$ 0.01	**0.47** $\pm$ 0.01
	co-OPT test	**0.46** $\pm$ 0.02	**0.49** $\pm$ 0.01	**0.49** $\pm$ 0.02
N = 120	Shared kernels and estimated P_0	**0.20** $\pm$ 0.04	**0.19** $\pm$ 0.03	**0.16** $\pm$ 0.01
	Shared kernels and $P_0 = 0.5$	**0.20** $\pm$ 0.03	**0.20** $\pm$ 0.02	**0.18** $\pm$ 0.01
	Separate kernels and $P_0 = 0.5$	**0.32** $\pm$ 0.02	**0.30** $\pm$ 0.04	**0.30** $\pm$ 0.01
	co-OPT test	**0.40** $\pm$ 0.02	**0.40** $\pm$ 0.02	**0.43** $\pm$ 0.03
N = 480	Shared kernels and estimated P_0	**0.07** $\pm$ 0.02	**0.09** $\pm$ 0.02	**0.08** $\pm$ 0.01
	Shared kernels and $P_0 = 0.5$	**0.08** $\pm$ 0.02	**0.09** $\pm$ 0.02	**0.09** $\pm$ 0.01
	Separate kernels and $P_0 = 0.5$	**0.12** $\pm$ 0.05	**0.14** $\pm$ 0.02	**0.13** $\pm$ 0.01
	co-OPT test	**0.29** $\pm$ 0.07	**0.28** $\pm$ 0.03	**0.29** $\pm$ 0.04

Results: Estimated P_0's