Bayesian Analysis (2020) **15**, Number 4, pp. 1169–1197

Theory of Optimal Bayesian Feature Filtering

Ali Foroughi pour* and Lori A. Dalton†

Abstract. Optimal Bayesian feature filtering (OBF) is a supervised screening method designed for biomarker discovery. In this article, we prove two major theoretical properties of OBF. First, optimal Bayesian feature selection under a general family of Bayesian models reduces to filtering *if and only if* the underlying Bayesian model assumes all features are mutually independent. Therefore, OBF is optimal if and only if one assumes all features are mutually independent, and OBF is the only filter method that is optimal under at least one model in the general Bayesian framework. Second, OBF under independent Gaussian models is consistent under very mild conditions, including cases where the data is non-Gaussian with correlated features. This result provides conditions where OBF is guaranteed to identify the correct feature set given enough data, and it justifies the use of OBF in non-design settings where its assumptions are invalid.

Keywords: Bayesian decision theory, variable selection, biomarker discovery.

MSC2020 subject classifications: 62F15, 62C10, 62F07, 92C37.

1 Introduction

Biomarker discovery entails mining a small-sample high-dimensional dataset for a list of features that represent potentially interesting molecular biomarkers. The hope is that the reported features might direct future studies (Feng et al., 2004) that ultimately lead to new diagnostic or prognostic tests, better treatment recommendations, or a better understanding of the regulatory mechanisms underlying the biological phenomena or disease under study (Ilyin et al., 2004; Rifai et al., 2006; Ramachandran et al., 2008).

Unfortunately, discovering reliable and reproducible biomarkers has proven to be difficult (Diamandis, 2010). One reason is that the algorithms employed (see Ilyin et al. (2004), Saeys et al. (2007), Diamandis (2010) and Ang et al. (2016) for reviews on common methods) are typically not well suited for the biomarker discovery problem. Univariate filter methods often exhibit quirks depending on the scoring function employed (Lazar et al., 2012). For example, the popular t-test cannot detect markers based on large differences between variance alone (Foroughi pour and Dalton, 2018b), even though such markers may have an important role to play in the disease under study or help uncover previously unknown subclasses of the disease. Multivariate methods may seem to have an advantage over filters because they can account for correlations; however, rather than use this correlation information to identify *all* markers that may be of interest, they tend to avoid selecting redundant markers or reward selecting smaller fea-

*Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210, foroughipour.1@osu.edu

†Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210, dalton@ece.osu.edu

 https://doi.org/10.1214/19-BA1182

ture sets to simplify model construction or avoid overfitting (Sima and Dougherty, 2008; Awada et al., 2012; Ang et al., 2016; Li et al., 2017). This effect is so catastrophic for biomarker discovery that univariate methods often far outperform multivariate methods (Sima and Dougherty, 2006, 2008; Foroughi pour and Dalton, 2018b).

Here we examine *optimal Bayesian feature filtering* (OBF), a supervised univariate filter method designed from the ground up for exploratory biomarker discovery (Foroughi pour and Dalton, 2015). OBF assumes a finite number of classes (e.g., patients given drug A versus drug B). Under its assumed model, OBF optimally detects and ranks the set of *all* features with distributional differences between the classes. It has been shown through simulations that OBF has competitive and robust performance across Bayesian models with block-diagonal covariances, and that it enjoys particularly excellent performance when markers are individually strong with low correlations (Foroughi pour and Dalton, 2017d, 2018a). Foroughi pour and Dalton (2018b) also examined the performance of OBF when its modeling assumptions (e.g., independence, priors, and Gaussianity) are violated, provided guidance on choosing inputs and objective criteria for robust performance, and demonstrated that OBF enjoys low computation cost.

Under Gaussian models with certain non-informative priors, OBF reduces to testing each feature separately using the test statistic studied by Pearson and Neyman (1930) and Zhang et al. (2012) for the equality of two Gaussian populations. OBF does not use classification or regression in any part of its framework. While variable selection methods based on classification or regression (for instance LASSO) are useful for designing predictive models (O'Hara and Sillanpää, 2009), like most multivariate methods they are typically not suitable for biomarker discovery because their objective is model construction. Small sample sizes worsen the overfitting problem, often resulting in small feature sets. If classification is involved, error estimation bias and variance result in poor selection performance (Sima and Dougherty, 2006).

Bayesian variable selection methods like Bayesian LASSO (Park and Casella, 2008), the Bayesian extension to group LASSO by Xu and Ghosh (2015), and works by Lee et al. (2003) and Baragatti (2011) based on generalized linear models (GLMs), suffer from similar problems. Whereas OBF places priors directly on the underlying data generation model, most priors for Bayesian variable selection, for example spike and slab priors (Mitchell and Beauchamp, 1988; Madigan and Raftery, 1994; George and McCulloch, 1997; Ishwaran and Rao, 2005), place uncertainty on the classification or regression parameters, which are difficult to justify, interpret and validate in practice. Multicollinearity can be assuaged by grouping genes, but methods by Rockova and Lesaffre (2014) and Xu and Ghosh (2015) assume grouping information is known *a priori*, which is infeasible in exploratory analysis. Also, in contrast with OBF, Bayesian methods often rely on computationally intensive methods like Markov-Chain Monte-Carlo (MCMC) sampling or variational inference (Carbonetto and Stephens, 2012).

Shared kernel Bayesian screening (SKBS) by Lock and Dunson (2015) is an interesting approach that assumes all feature distributions belong to a family of mixture models with K components, and the objective is to test whether the classes have different weights in the mixture distribution. Whereas OBF treats each individual feature separately, SKBS uses the same K dictionary mixture components for all features and allows only the kernel weights to vary. When sample size is small we observed better

performance using small K, but in this case the data may not be properly modeled for all features and the detected mixture components lose interpretability. SKBS also uses MCMC, making it more computationally expensive than OBF. Bayesian non-parametric methods, for example those based on Dirichlet or Pitman-Yor processes, have also gained popularity in bioinformatics for classification, inferring gene networks, clustering, and detecting chromosomal aberrations (Shahbaba and Neal, 2009; Libbrecht and Noble, 2015; Mitra and Müller, 2015; Ni et al., 2017). Spike-and-slab Dirichlet processes avoid the need to specify the number of mixtures; however, it is still difficult to specify and justify the base distribution and priors in practice. While our focus here is on the supervised case, many works like that of Cui and Cui (2012) focus on the unsupervised case. Computation is also a key concern; Cui and Cui use Bayesian expectation-maximization, which is more demanding than OBF. Holmes et al. (2015) present a supervised method based on Pólya trees, however, the model may require larger samples than available in a typical exploratory analysis and may be sensitive to imbalanced samples.

Our main contributions are two-fold: (1) we prove optimal Bayesian feature selection under a general family of Bayesian models reduces to filtering (e.g., OBF) *if and only if* the underlying Bayesian model assumes all features are independent, and (2) we prove OBF under independent Gaussian models is a consistent estimator of the feature set we wish to select under mild conditions, including cases where the data is non-Gaussian with correlated features. Contribution (1) has two practical implications: OBF is the only filter method for which there exists a model in the general Bayesian framework where it is optimal, and OBF is optimal if and only if one assumes all features are independent. Contribution (2) is of enormous importance, since it provides conditions where OBF is guaranteed to identify the correct feature set given enough data, and it justifies the use of OBF in non-design settings where its assumptions are invalid.

We review the Bayesian model in Section 2 and optimal set selection in Section 3. In Section 4 we discuss OBF and present contribution (1) in Theorem 1, and in Section 5 we examine consistency and present contribution (2) in Theorems 2 and 3. We provide a demonstration on synthetic microarray data in Section 6, and conclude in Section 7. We provide a demonstration on real colon cancer microarray data in Sections S2 and S3 of Supplementary Material A (Foroughi pour and Dalton, 2019).

2 Bayesian Model

In Section 2.1, we describe the general three-level Bayesian model originally proposed in Dalton (2013). In Sections 2.2 and 2.3 we cover the independent case and independent Gaussian case, respectively, which are originally presented in Foroughi pour and Dalton (2015). Although not covered here, an independent categorical model and several models with correlations in the general Bayesian framework have been proposed (Dalton, 2013; Foroughi pour and Dalton, 2016a, 2017c,d).

2.1 General Bayesian Model

Consider a feature selection problem in which we are to identify all features that have distinct distributions between two classes, $y = 0$ and $y = 1$. Although we consider binary

labels here, the multiclass case is similar and has been characterized in Foroughi pour and Dalton (2017b). Let F be a set of feature indices, let each feature $f \in F$ be associated with a space, $\mathcal{X}_f$, and let $\mathcal{X} = \prod_{f \in F} \mathcal{X}_f$ be the feature space. Typically, $\mathcal{X}_f = \mathbb{R}$ for all f. We call features that we wish to select, e.g. those with distributional differences between classes, "good features." When viewed as a random quantity, we denote this set by $\bar{G}$, and we denote a realization of this random set by G. Likewise, we call features that we wish not to select "bad features," and denote them by $\bar{B} = F \backslash \bar{G}$ when random and $B = F \backslash G$ when fixed, where "$\backslash$" is the set difference. Conditioning on events like $\{\bar{G} = G\}$ or $\{f \in \bar{G}\}$ does not mean the set of good features is deterministic. Rather, this should be interpreted as merely a hypothesis that these events hold for the current $G \subseteq F$ or $f \in F$ under consideration. Furthermore, since $\bar{G} = G$ if and only if $\bar{B} = B$, conditioning on the event $\{\bar{G} = G\}$ is equivalent to conditioning on the event $\{\bar{B} = B\}$. We denote conditioning on these events by "$|G$" or "$|B$", and use these notations interchangeably throughout.

We denote a prior on $\bar{G}$ across the power set of F by $p(G) = P(\bar{G} = G)$. Given $\bar{G} = G$, let θ_y^G denote data generation parameters of class $y \in \{0, 1\}$ features in G, let θ^B denote data generation parameters of features in B, and let $\theta = \{\theta_0^G, \theta_1^G, \theta^B\}$ be the set of all data generation parameters. Define corresponding parameter spaces: Θ_y^G, Θ^B and $\Theta = \Theta_0^G \times \Theta_1^G \times \Theta^B$. We denote a prior on θ by $p(\theta|G)$, and assume θ_0^G, θ_1^G and θ^B are conditionally mutually independent, i.e.,

$$p(\theta|G) = p(\theta_0^G|G)p(\theta_1^G|G)p(\theta^B|B). \tag{2.1}$$

We assume feature selection is aided by the observation of feature-label pairs, and we denote the complete dataset, including features and labels, by S. Though we assume the data is complete here, the missing data problem has been studied for special cases of this model in Foroughi pour and Dalton (2016b). Let $x \in \mathcal{X}$ be a feature vector, and let x^G and x^B denote elements of x that correspond to features in G and B respectively. Given $\bar{G} = G$, parameter θ and class y, we also assume x^G and x^B are independent:

$$p(x|y, \theta, G) = p(x^G|\theta_y^G)p(x^B|\theta^B). \tag{2.2}$$

Assume the data is comprised of n points with n_y points in class y, that the label of each point is determined by a process independent of θ and G, and that, conditioned on the labels, sample points are independent with points belonging to the same class identically distributed. These assumptions are true in many sampling strategies, for instance random and separate sampling. Let S_y^G and S^B be the part of the data corresponding to features in G from class y and features in B from both classes, respectively. Due to independence between x^G and x^B and independence between sample points,

$$p(S|\theta, G) \propto p(S_0^G|\theta_0^G)p(S_1^G|\theta_1^G)p(S^B|\theta^B), \tag{2.3}$$

where the proportionality constant depends on the distribution of n_y for the given sampling strategy, $p(S_y^G|\theta_y^G) = \prod_{x^G \in S_y^G} p(x^G|\theta_y^G)$, and $p(S^B|\theta^B) = \prod_{x^B \in S^B} p(x^B|\theta^B)$. Thus, S_0^G, S_1^G and S^B are mutually independent given θ and G. Further, from (2.1) and (2.3), they are also independent given only G, that is,

$$p(S|G) = \int_\Theta p(\theta|G)p(S|\theta, G)d\theta \propto p(S_0^G|G)p(S_1^G|G)p(S^B|B), \tag{2.4}$$

where for $y \in \{0, 1\}$,

$$p(S_y^G|G) = \int_{\Theta_y^G} p(\theta_y^G|G)p(S_y^G|\theta_y^G)d\theta_y^G, \quad p(S^B|B) = \int_{\Theta^B} p(\theta^B|B)p(S^B|\theta^B)d\theta^B. \quad (2.5)$$

Let $p(G|S) = P(\bar{G} = G|S)$ be the posterior probability that the set G is precisely the set of good features, given our observation of the data. Applying Bayes' rule and (2.4),

$$p(G|S) \propto p(G)p(S|G) \propto p(G)p(S_0^G|G)p(S_1^G|G)p(S^B|B). \quad (2.6)$$

The marginal prior and posterior probabilities that an individual feature, $f \in F$, is in $\bar{G}$ are denoted by $\pi(f) = P(f \in \bar{G}) = \sum_{G:f\in G} p(G)$ and

$$\pi^*(f) = P(f \in \bar{G}|S) = \sum_{G:f\in G} p(G|S), \quad (2.7)$$

respectively. Note that $P(f \in \bar{B}) = 1 - \pi(f)$ and $P(f \in \bar{B}|S) = 1 - \pi^*(f)$. Also,

$$E\left(|\bar{G}|\right) = E\left(\sum_{f\in F} I(f \in \bar{G})\right) = \sum_{f\in F} P(f \in \bar{G}) = \sum_{f\in F} \pi(f), \quad (2.8)$$

where $|\cdot|$ denotes cardinality for sets, and $I(q)$ is the indicator function, equal to 1 if q holds and 0 otherwise. Similarly, $E\left(|\bar{G}||S\right) = \sum_{f\in F} \pi^*(f)$. The expected number of good features, before and after observing data, may be found from π and π^*, respectively.

In biomarker discovery, previously known biomarkers can be integrated into the prior to aid the discovery of new biomarkers (Foroughi pour and Dalton, 2017a). When prior knowledge is not available, improper priors for $p(\theta|G)$ may be needed and the above derivations become invalid. To circumvent this problem we: (1) require $p(\theta|G)$ to be such that the integrals in (2.5) are positive and finite, (2) require $\pi(G)$ to be proper, and (3) take (2.5), (2.6) and (2.7) as definitions with the proportionality constant in (2.6) defined such that $\sum_{G:G\subseteq F} p(G|S) = 1$. Although improper priors are controversial, see for example marginalization paradoxes described by Dawid et al. (1973), counterexamples discussed by Jaynes (2003), and discussions on the Jeffreys-Lindley paradox by Robert (1993, 2014), this guarantees the posterior $p(G|S)$ and marginal posterior $\pi^*(f)$ under improper priors are normalizable to valid densities and have definitions consistent with proper priors. See Sections S5 and S6 of Supplementary Material A for further discussions on improper priors.

2.2 Independent Bayesian Model

Assume a prior $p(G)$ on $\bar{G}$ where the events $\{f \in \bar{G}\}$ are mutually independent. Then,

$$\begin{aligned} p(G) &= P\left((\cap_{g\in G}\{g \in \bar{G}\}) \cap (\cap_{b\in B}\{b \in \bar{B}\})\right) \\ &= \prod_{g\in G} \pi(g) \prod_{b\in B} (1 - \pi(b)). \end{aligned} \quad (2.9)$$

To completely characterize this prior, note that one need only specify $\pi(f)$ for all $f \in F$. Further, if $\pi(f) = p$ is constant for all $f \in F$, then $|\bar{G}|$ is binomial$(|F|, p)$.

For every $f \in F$ we assign three parameters, θ_0^f, θ_1^f and θ^f, with parameter spaces Θ_0^f, Θ_1^f and Θ^f and densities $p(\theta_0^f)$, $p(\theta_1^f)$ and $p(\theta^f)$, respectively. Let $\theta_y^G = \{\theta_y^f : f \in G\}$ and $\theta^B = \{\theta^f : f \in B\}$ and assume parameters of individual features are mutually independent given $\bar{G} = G$. Then (2.1) becomes $p(\theta|G) = \prod_{g\in G} p(\theta_0^g)p(\theta_1^g) \prod_{b\in B} p(\theta^b)$. Finally, we assume features are mutually independent given $\bar{G} = G$, θ and y, thus the joint density in (2.2) is now of the form $p(x|y, \theta, G) = \prod_{g\in G} p(x^g|\theta_y^g) \prod_{b\in B} p(x^b|\theta^b)$, where $p(x^g|\theta_y^g)$ and $p(x^b|\theta^b)$ are the marginals of good and bad features, respectively.

As in (2.6), one can show

$$p(G|S) \propto p(G) \prod_{g\in G} p(S_0^g|g \in \bar{G})p(S_1^g|g \in \bar{G}) \prod_{b\in B} p(S^b|b \in \bar{B}), \tag{2.10}$$

where, as in (2.5),

$$p(S_y^f|f \in \bar{G}) = \textstyle\int_{\Theta_y^f} p(\theta_y^f)p(S_y^f|\theta_y^f)d\theta_y^f, \quad p(S^f|f \in \bar{B}) = \int_{\Theta^f} p(\theta^f)p(S^f|\theta^f)d\theta^f. \tag{2.11}$$

Dividing the right-hand side of (2.10) by the constant $\prod_{f\in F}(1 - \pi(f))p(S^f|f \in \bar{B})$, we have

$$p(G|S) \propto \prod_{g\in G} h(g), \tag{2.12}$$

where for all $f \in F$, we define

$$h(f) = \frac{\pi(f)}{1 - \pi(f)} \times \frac{p(S_0^f|f \in \bar{G})p(S_1^f|f \in \bar{G})}{p(S^f|f \in \bar{B})}. \tag{2.13}$$

Furthermore, from (2.7),

$$\pi^*(f) = \frac{\sum_{G:f\in G} \prod_{g\in G} h(g)}{\sum_G \prod_{g\in G} h(g)} = \frac{h(f) \sum_{G:f\notin G} \prod_{g\in G} h(g)}{(1 + h(f)) \sum_{G:f\notin G} \prod_{g\in G} h(g)} = \frac{h(f)}{1 + h(f)}. \tag{2.14}$$

Once $h(f)$ is found, $\pi^*(f)$ is obtained from (2.14). Note that $h(f) = \pi^*(f)/(1 - \pi^*(f))$. Plugging this in (2.12) and normalizing by the constant $\prod_{f\in F}(1 - \pi^*(f))$, we have

$$p(G|S) \propto \prod_{g\in G} \pi^*(g) \prod_{b\in B} (1 - \pi^*(b)). \tag{2.15}$$

In fact, (2.15) holds with equality, thus the events $\{f \in \bar{G}\}$ are mutually independent conditioned on S. Just as $\pi(f)$ characterizes $p(G)$, $\pi^*(f)$ characterizes $p(G|S)$.

When $p(\theta_y^f)$ or $p(\theta^f)$ are improper, we require $\pi(f)$ to be proper, we require the integrals in (2.11) to be positive and finite and take these equations as definitions, and we define $\pi^*(f) = h(f)/(1 + h(f))$ as in (2.14), where $h(f)$ is defined in (2.13).

2.3 Independent Gaussian Model

Now suppose all features are Gaussian with conjugate priors. If $f \in \bar{G}$ then $\theta_y^f = [\mu_y^f, \sigma_y^f]$, where μ_y^f and σ_y^f are the mean and variance of x^f in class y, respectively. Similarly, if $f \in \bar{B}$, then $\theta^f = [\mu^f, \sigma^f]$, where μ^f and σ^f are the mean and variance of x^f. To simplify notation, we drop the conventional square in variances, σ_y^f and σ^f.

Assume $p(\theta_y^f) = p(\sigma_y^f)p(\mu_y^f|\sigma_y^f)$, where $p(\sigma_y^f) = A_y^f(\sigma_y^f)^{-0.5(\kappa_y^f+2)}\exp(-0.5s_y^f/\sigma_y^f)$, $p(\mu_y^f|\sigma_y^f) = B_y^f(\sigma_y^f)^{-0.5}\exp(-0.5\nu_y^f(\mu_y^f - m_y^f)^2/\sigma_y^f)$, and s_y^f, κ_y^f, m_y^f and ν_y^f are real-valued hyper-parameters. For a proper prior we require $s_y^f, \kappa_y^f, \nu_y^f > 0$, in which case $p(\sigma_y^f)$ is an inverse-Wishart distribution with mean $s_y^f/(\kappa_y^f - 2)$ if $\kappa_y^f > 2$, and $p(\mu_y^f|\sigma_y^f)$ is Gaussian with mean m_y^f and variance σ_y^f/ν_y^f. A_y^f and B_y^f scale the two distributions, where under a proper prior $A_y^f = (0.5s_y^f)^{0.5\kappa_y^f}/\Gamma(0.5\kappa_y^f)$ and $B_y^f = (2\pi/\nu_y^f)^{-0.5}$.

The posterior, $p(\theta_y^f|S_y^f)$, is of the same form as the prior, $p(\theta_y^f)$, with updated hyper-parameters $\kappa_y^{f*} = \kappa_y^f + n_y$, $\nu_y^{f*} = \nu_y^f + n_y$, $m_y^{f*} = (\nu_y^f m_y^f + n_y\hat{\mu}_y^f)/(\nu_y^f + n_y)$, and $s_y^{f*} = s_y^f + (n_y - 1)\hat{\sigma}_y^f + \frac{\nu_y^f n_y}{\nu_y^f + n_y}(\hat{\mu}_y^f - m_y^f)^2$, where $\hat{\mu}_y^f$ and $\hat{\sigma}_y^f = \sum_{x \in S_y^f}(x - \hat{\mu}_y^f)^2/(n_y - 1)$ are the sample mean and unbiased sample variance, respectively, of feature f points in class y (Murphy, 2007). Note that $p(S_y^f|f \in \bar{G})$ is the normalization constant in finding the posterior, $p(\theta_y^f|S_y^f)$, from the prior times likelihood, $p(\theta_y^f)p(S_y^f|\theta_y^f)$:

$$p(S_y^f|f \in \bar{G}) = \frac{p(\theta_y^f)p(S_y^f|\theta_y^f)}{p(\theta_y^f|S_y^f)} = \frac{A_y^f B_y^f \Gamma(0.5\kappa_y^{f*})}{(2\pi)^{0.5(n_y-1)}(\nu_y^{f*})^{0.5}(0.5s_y^{f*})^{0.5\kappa_y^{f*}}}. \tag{2.16}$$

Moving on to bad features, we assume, $p(\theta^f) = p(\sigma^f)p(\mu^f|\sigma^f)$, where given real-valued hyper-parameters s^f, κ^f, m^f, and ν^f, $p(\sigma^f) = A^f(\sigma^f)^{-0.5(\kappa^f+2)}\exp(-0.5s^f/\sigma^f)$ and $p(\mu^f|\sigma^f) = B^f(\sigma^f)^{-0.5}\exp(-0.5\nu^f(\mu^f - m^f)^2/\sigma^f)$. For a proper prior, $s^f, \kappa^f, \nu^f > 0$, $A^f = (0.5s^f)^{0.5\kappa^f}/\Gamma(0.5\kappa^f)$ and $B^f = (2\pi/\nu^f)^{-0.5}$. The posterior has updated hyper-parameters, $\kappa^{f*} = \kappa^f + n$, $\nu^{f*} = \nu^f + n$, $m^{f*} = (\nu^f m^f + n\hat{\mu}^f)/(\nu^f + n)$, and $s^{f*} = s^f + (n-1)\hat{\sigma}^f + \frac{\nu^f n}{\nu^f + n}(\hat{\mu}^f - m^f)^2$, where $\hat{\mu}^f$ and $\hat{\sigma}^f$ are the sample mean and variance, respectively, of feature f points in both classes (Murphy, 2007). As in (2.16),

$$p(S^f|f \in \bar{B}) = \frac{A^f B^f \Gamma(0.5\kappa^{f*})}{(2\pi)^{0.5(n-1)}(\nu^{f*})^{0.5}(0.5s^{f*})^{0.5\kappa^{f*}}}. \tag{2.17}$$

Plugging (2.16) and (2.17) in (2.13),

$$h(f) = \frac{\pi(f)}{1 - \pi(f)} L^f \left(\frac{2\pi\nu^{f*}}{\nu_0^{f*}\nu_1^{f*}}\right)^{0.5} \frac{\Gamma(0.5\kappa_0^{f*})\Gamma(0.5\kappa_1^{f*})(0.5s^{f*})^{0.5\kappa^{f*}}}{\Gamma(0.5\kappa^{f*})(0.5s_0^{f*})^{0.5\kappa_0^{f*}}(0.5s_1^{f*})^{0.5\kappa_1^{f*}}}, \tag{2.18}$$

where $L^f = A_0^f B_0^f A_1^f B_1^f/(A^f B^f)$. If $\pi(f)$, L^f, ν_y^f, ν^f, κ_y^f and κ^f do not depend on f,

$$h(f) \propto \frac{(s^{f*})^{0.5\kappa^{f*}}}{(s_0^{f*})^{0.5\kappa_0^{f*}}(s_1^{f*})^{0.5\kappa_1^{f*}}}. \tag{2.19}$$

Under improper priors we require $\pi(f)$ to be proper, and to ensure (2.16) and (2.17) are positive and finite we require $s_0^{f*}, \kappa_0^{f*}, \nu_0^{f*}, s_1^{f*}, \kappa_1^{f*}, \nu_1^{f*}, s^{f*}, \kappa^{f*}, \nu^{f*} > 0$ for all $f \in F$. In addition, $L^f > 0$ becomes a separate parameter specified by the user. All theorems in this work hold under these improper priors, and set selection under proper and improper priors for the independent Gaussian case have been studied extensively in Foroughi pour and Dalton (2018b). Following Berger (1985), DeGroot (1970) and Akaike (1980), in Section S5 of Supplementary Material A we also show that $\pi^*(f)$ from these improper priors is equivalent to a limit of $\pi^*(f)$ from a sequence of proper priors.

3 Optimal Bayesian Feature Selection

We define five criteria for *optimal Bayesian feature selection* under the general Bayesian model: (1) the *maximum a posteriori* (MAP) criterion selects the feature set having the highest posterior probability of being the good feature set, (2) *constrained MAP* (CMAP) uses the MAP objective but considers only feature sets of a given size, (3) the *minimal risk* (MR) criterion minimizes a notion of risk, with the *maximum number correct* (MNC) rule being a special case that minimizes the number of mislabeled features, (4) *constrained MNC* (CMNC) uses the MNC objective but considers only feature sets of a given size, and (5) the *Neyman-Pearson* (NP) criterion maximizes the expected number of good features selected given a limited expected number of bad features selected. MAP was originally presented in Dalton (2013), while MNC and an early form of CMNC constrained to selecting two features (2MNC) were originally presented in Foroughi pour and Dalton (2014); all of the other criteria are new.

3.1 Maximum *a Posteriori*

The MAP feature set is the set having maximum posterior probability:

$$G^{MAP} = \arg\max_{G \subseteq F} p(G|S). \tag{3.1}$$

We also define the CMAP feature set to be the MAP feature set under the constraint of selecting exactly D features for some user-specified constant D:

$$G^{CMAP} = \arg\max_{G \subseteq F : |G| = D} p(G|S). \tag{3.2}$$

Let $\ell(G, \bar{G})$ be a *loss* function in selecting G when $\bar{G}$ is the true set of good features, and let $E(\ell(G, \bar{G})|S)$ be the *risk* in selecting G. It can be shown that the MAP feature set minimizes risk under a zero-one loss function that assigns $\ell(G, \bar{G}) = 0$ when $\bar{G} = G$ and $\ell(G, \bar{G}) = 1$ when $\bar{G} \neq G$. Therefore, one drawback of the MAP objective is that it assigns the same loss to all incorrect feature sets, regardless of how many features are labeled incorrectly. This is remedied by the MR objective, described in the next section.

3.2 Minimal Risk

Consider the family of objective criteria with $\ell(G, \bar{G})$ of the form:

$$\ell(G, \bar{G}) = \lambda_{GG}|G \cap \bar{G}| + \lambda_{GB}|G \cap \bar{B}| + \lambda_{BG}|B \cap \bar{G}| + \lambda_{BB}|B \cap \bar{B}|, \tag{3.3}$$

where λ_{GG}, λ_{GB}, λ_{BG}, and λ_{BB} are constants such that $\lambda_{GB} \geq \lambda_{BB}$ and $\lambda_{BG} \geq \lambda_{GG}$. The MR feature set is defined as:

$$G^{MR} = \arg\min_{G \subseteq F} E(\ell(G, \bar{G})|S). \tag{3.4}$$

Observe that,

$$E\left(|G \cap \bar{G}| \,\middle|\, S\right) = E\left(\textstyle\sum_{g \in G} I(g \in \bar{G})|S\right) = \sum_{g \in G} P\left(g \in \bar{G}|S\right) = \sum_{g \in G} \pi^*(g), \tag{3.5}$$

$$E\left(|G \cap \bar{B}| \,\middle|\, S\right) = \textstyle\sum_{g \in G}(1 - \pi^*(g)). \tag{3.6}$$

Similarly, $E\left(|B \cap \bar{G}| \,\middle|\, S\right) = \sum_{b \in B} \pi^*(b)$ and $E\left(|B \cap \bar{B}| \,\middle|\, S\right) = \sum_{b \in B}(1 - \pi^*(b))$. Thus,

$$\begin{aligned} E(\ell(G, \bar{G})|S) = {} & \lambda_{GG} \textstyle\sum_{g \in G} \pi^*(g) + \lambda_{GB} \sum_{g \in G}(1 - \pi^*(g)) \\ & + \lambda_{BG} \textstyle\sum_{b \in B} \pi^*(b) + \lambda_{BB} \sum_{b \in B}(1 - \pi^*(b)). \end{aligned} \tag{3.7}$$

$E(\ell(G, \bar{G})|S)$ is minimized by considering each feature, $f \in F$, individually. In particular, f is in G^{MR} if the risk incurred by including this feature, $\lambda_{GG}\pi^*(f) + \lambda_{GB}(1 - \pi^*(f))$, is less than the risk incurred by not including it, $\lambda_{BG}\pi^*(f) + \lambda_{BB}(1 - \pi^*(f))$, or equivalently, if $(\lambda_{GB} + \lambda_{BG} - \lambda_{GG} - \lambda_{BB})\pi^*(f) > \lambda_{GB} - \lambda_{BB}$. Thus,

$$G^{MR} = \left\{f \in F : \pi^*(f) > T\right\}, \tag{3.8}$$

where $T = (\lambda_{GB} - \lambda_{BB})/(\lambda_{GB} + \lambda_{BG} - \lambda_{GG} - \lambda_{BB})$. In other words, the MR objective ranks features by their marginal posterior probability of being in $\bar{G}$, and selects those with probabilities exceeding a given threshold.

When $\lambda_{GG} = \lambda_{BB} = 0$ and $\lambda_{GB} = \lambda_{BG} = 1$, the MR cost function minimizes the expectation of the number of mislabeled features, $|G \cap \bar{B}| + |B \cap \bar{G}|$, or equivalently, maximizes the expectation of the number of correctly labeled features, $c(G, \bar{G}) = |G \cap \bar{G}| + |B \cap \bar{B}|$. This results in the MNC objective:

$$G^{MNC} = \arg\max_{G \subseteq F} E(c(G, \bar{G})|S) = \left\{f \in F : \pi^*(f) > 0.5\right\}. \tag{3.9}$$

MNC thus selects features with a posterior probability of being in $\bar{G}$ greater than 0.5.

Constrained MR (CMR) minimizes risk under the constraint of selecting exactly D features:

$$G^{CMR} = \arg\min_{G \subseteq F : |G| = D} E(\ell(G, \bar{G})|S). \tag{3.10}$$

Following a similar procedure used to derive (3.8), observe:

$$G^{CMR} = \arg\max_{G \subseteq F : |G| = D} \sum_{g \in G} \pi^*(g). \tag{3.11}$$

Thus, G^{CMR} ranks $\pi^*(f)$ and selects the D features with highest rank. Since the λ's need not be specified, we also call this criterion CMNC.

3.3 Neyman-Pearson

Viewing the number of correctly identified good features, $|G \cap \bar{G}|$, as the number of *true positives*, and the number of incorrectly identified bad features, $|G \cap \bar{B}|$, as the number of *false positives*, the NP objective maximizes the expected number of true positives while bounding the expected number of false positives by $0 \leq \alpha \leq E(|\bar{B}||S)$:

$$\begin{aligned} G^{NP} = \underset{G \subseteq F}{\arg\max} \quad & E\left(|G \cap \bar{G}||S\right) \\ \text{subject to} \quad & E\left(|G \cap \bar{B}||S\right) \leq \alpha. \end{aligned} \tag{3.12}$$

From (3.5) and (3.6), we have that

$$\begin{aligned} G^{NP} = \underset{G \subseteq F}{\arg\max} \quad & \sum_{g \in G} \pi^*(g) \\ \text{subject to} \quad & \sum_{g \in G} (1 - \pi^*(g)) \leq \alpha. \end{aligned} \tag{3.13}$$

This is solved by ranking $\pi^*(f)$ and iteratively adding features with highest rank to G^{NP} until adding a new feature results in violating the constraint. NP is closely related to MR and CMNC in that all of these methods rank features using the same scoring function, $\pi^*(f)$. However, they use different score cutoffs; in MR the cutoff is a constant threshold, in CMNC the cutoff forces a certain set size, and in NP the cutoff depends on the values of the $\pi^*(f)$. For selection rule G^k with free parameter k, plotting the pair $(E(|G^k \cap \bar{B}||S), E(|G^k \cap \bar{G}||S))$ in the $[0, E(|\bar{B}||S)] \times [0, E(|\bar{G}||S)]$ space under various k results in a curve analogous to a *receiver operating characteristic* (ROC) curve. The ROC curve for MR (varying T), CMNC (varying D) and NP (varying α) are all

$$(k - \textstyle\sum_{f=1}^{k} \pi^*_{(f)}, \sum_{f=1}^{k} \pi^*_{(f)}) \tag{3.14}$$

for $k = 0, 1, \ldots, |F|$, where the $\pi^*_{(f)}$ are the $\pi^*(f)$ ordered from largest to smallest.

4 Optimal Bayesian Feature Filtering

In the general Bayesian model, MAP and CMAP require finding $p(G|S)$ for all $G \subseteq F$, which is computationally prohibitive when $|F|$ is large. Although MR (and thus MNC), CMNC and NP always reduce to ranking features by $\pi^*(f)$ with various methods of thresholding, finding $\pi^*(f)$ also requires evaluating $p(G|S)$ for all $G \subseteq F$. In this section, we discuss how this problem is circumvented under independent Bayesian models.

Under independent Bayesian models, any method that ranks features by $\pi^*(f)$ (or equivalently $h(f)$) and selects top ranking features is considered an OBF rule. While MAP and CMAP generally do not reduce to ranking $\pi^*(f)$, in independent Bayesian models MAP reduces to MNC and CMAP reduces to CMNC by (2.15) and (3.1), thus all selection criteria covered in Section 3 reduce to OBF rules. Furthermore, since $\pi^*(f)$ can be found separately for each feature under independent Bayesian models via (2.14) (for instance using (2.18) or (2.19) in the Gaussian case), all OBF rules reduce to filtering. The fact that optimal Bayesian feature selection reduces to filtering under

independent models is not surprising, in light of similar results for Bayesian multiple comparison rules (Müller et al., 2006). By assuming independence we lose the ability to take advantage of correlations, but we greatly simplify optimal selection.

Define a *univariate filter on* F to be a feature selection rule that ranks features by a scoring function $h(f, S^f)$, which is a function of only the feature index f and the portion of the labeled data corresponding to f, and selects top ranking features using some score thresholding rule, which is based on only the set of feature scores. t-tests with Benjamini and Hochberg (1995) multiple testing correction are univariate filters. Define a *simple univariate filter on* F to be a univariate filter that uses a constant threshold, i.e., a feature selection rule that reduces to the form:

$$G = \{f \in F : h(f, S^f) > T\}, \tag{4.1}$$

where T is a constant. t-tests without multiple testing correction are simple univariate filters. By the following theorem, not only does optimal selection reduce to OBF under independent models, but optimal selection reduces to simple univariate filtering *only* under independent models, and the resulting filter must be equivalent to an OBF rule.

Theorem 1. *MR under a general Bayesian model* $\mathcal{M}$ *on feature set* F *is a simple univariate filter on* F *for all thresholds* T *if and only if there exists an independent Bayesian model* $\mathcal{M}'$ *on* F *such that* $\pi^*(f|\mathcal{M}') = \pi^*(f|\mathcal{M})$ *for all* $f \in F$ *and all labeled datasets* S.

Proof. Suppose an independent Bayesian model, $\mathcal{M}'$, exists as characterized above. Let T be an arbitrary constant. MR simplifies to $G^{MR} = \{f \in F : \pi^*(f|\mathcal{M}') > T\}$ by (3.8), where $\pi^*(f|\mathcal{M}')$, given in (2.14), depends only on f and S^f (note that S^f is comprised of S_0^f and S_1^f, along with the labels). Thus, MR reduces to a simple univariate filter on F under both $\mathcal{M}'$ and $\mathcal{M}$ for all T.

Now suppose that MR under $\mathcal{M}$ is a simple univariate filter on F for all T. Suppose there exist samples $S_\bullet \neq S_\circ$ and $f \in F$ such that $S_\bullet^f = S_\circ^f$, but $P(f \in \bar{G}|S_\bullet, \mathcal{M}) > P(f \in \bar{G}|S_\circ, \mathcal{M})$. Let T be the midpoint between $P(f \in \bar{G}|S_\bullet, \mathcal{M})$ and $P(f \in \bar{G}|S_\circ, \mathcal{M})$. MR at threshold T selects f under $S_\bullet$, but does not select f under $S_\circ$. This contradicts the premise that MR is a simple univariate filter. Thus, for all triplets $S_\bullet$, $S_\circ$ and f such that $S_\bullet \neq S_\circ$ and $S_\bullet^f = S_\circ^f$, we must have $P(f \in \bar{G}|S_\bullet, \mathcal{M}) = P(f \in \bar{G}|S_\circ, \mathcal{M})$. Fix $f_0 \in F$. Assume that $P(f_0 \in \bar{G}|S, \mathcal{M})$, which is in general a function of S, cannot be written as a function of only S^{f_0}. Then there exists a pair of samples $S_\bullet$ and $S_\circ$ such that $S_\bullet \neq S_\circ$, $S_\bullet^{f_0} = S_\circ^{f_0}$ and $P(f_0 \in \bar{G}|S_\bullet, \mathcal{M}) \neq P(f_0 \in \bar{G}|S_\circ, \mathcal{M})$. By contradiction, $P(f_0 \in \bar{G}|S, \mathcal{M})$ can be written as a function of only S^{f_0}. Since f_0 is arbitrary, we must have that the marginal posterior for each feature can be expressed as $\pi^*(f|\mathcal{M}) \equiv P(f \in \bar{G}|S, \mathcal{M}) = P(f \in \bar{G}|S^f, \mathcal{M})$ for all $f \in F$ and all S. From Bayes rule,

$$\pi^*(f|\mathcal{M}) = \frac{p_0}{p_0 + p_1}, \tag{4.2}$$

where $p_0 = P(f \in \bar{G}|\mathcal{M}) \prod_{y \in \{0,1\}} p(S_y^f | f \in \bar{G}, \mathcal{M})$, $p_1 = P(f \in \bar{B}|\mathcal{M}) p(S^f | f \in \bar{B}, \mathcal{M})$,

$$p(S_y^g | g \in \bar{G}, \mathcal{M}) = \sum_{G: g \notin G} P(\bar{G} = G \cup \{g\} | g \in \bar{G}, \mathcal{M}) p(S_y^g | G \cup \{g\}, \mathcal{M}), \tag{4.3}$$

$$p(S^b|b \in \bar{B}, \mathcal{M}) = \sum_{B: b \notin B} P(\bar{B} = B \cup \{b\}|b \in \bar{B}, \mathcal{M}) p(S^b|B \cup \{b\}, \mathcal{M}), \tag{4.4}$$

$p(S_y^g|G, \mathcal{M}) = \int_{\Theta_y^G} p(\theta_y^G|G, \mathcal{M}) p(S_y^g|\theta_y^G, \mathcal{M}) d\theta_y^G$ and $p(S^b|B, \mathcal{M}) = \int_{\Theta^B} p(\theta^B|B, \mathcal{M})$ $p(S^b|\theta^B, \mathcal{M}) d\theta^B$. We now construct an independent Bayesian model, $\mathcal{M}'$. The idea is to create auxiliary random variables for each $f \in F$ that are independent from other features and yet sufficient to describe $\pi^*(f|\mathcal{M})$. Define $P(f \in \bar{G}|\mathcal{M}') = P(f \in \bar{G}|\mathcal{M})$, define the data generation parameters $\phi_y^g = \{\bar{H}, \theta_y^{\bar{H} \cup \{g\}}\}$ for each $g \in F$, and define priors on a realization of $H \subseteq F \backslash \{g\}$ and $\theta_y^{H \cup \{g\}} \in \Theta_y^{H \cup \{g\}}$ from $\mathcal{M}$ by,

$$p(\phi_y^g|\mathcal{M}') = P(\bar{G} = H \cup \{g\}|g \in \bar{G}, \mathcal{M}) p(\theta_y^{H \cup \{g\}}|H \cup \{g\}, \mathcal{M}). \tag{4.5}$$

Similarly, for all $b \in F$, define $\phi^b = \{\bar{H}, \theta^{\bar{H} \cup \{b\}}\}$, and define priors on $H \subseteq F \backslash \{b\}$ and $\theta^{H \cup \{b\}} \in \Theta^{H \cup \{b\}}$ from $\mathcal{M}$ by,

$$p(\phi^b|\mathcal{M}') = P(\bar{B} = H \cup \{b\}|b \in \bar{B}, \mathcal{M}) p(\theta^{H \cup \{b\}}|H \cup \{b\}, \mathcal{M}). \tag{4.6}$$

In this way, for each feature $f \in F$ we merge the identity of features excluding f with the data generation parameters. Finally, we define the distributions $p(x^g|\phi_y^g, \mathcal{M}') = p(x^g|\theta_y^{H \cup \{g\}}, \mathcal{M})$ and $p(x^b|\phi^b, \mathcal{M}') = p(x^b|\theta^{H \cup \{b\}}, \mathcal{M})$ using the marginal distributions of x^f under $\mathcal{M}$. Note that $p(S_y^g|\phi_y^g, \mathcal{M}') = p(S_y^g|\theta_y^{H \cup \{g\}}, \mathcal{M})$ and $p(S^b|\phi^b, \mathcal{M}') = p(S^b|\theta^{H \cup \{b\}}, \mathcal{M})$. Applying (2.14), the definition of $h(f)$, and the definition of $P(f \in \bar{G}|\mathcal{M}')$, $\pi^*(f|\mathcal{M}')$ is of the form in (4.2) with $p_0 = P(f \in \bar{G}|\mathcal{M}) \prod_{y \in \{0,1\}} p(S_y^f|f \in \bar{G}, \mathcal{M}')$ and $p_1 = P(f \in \bar{B}|\mathcal{M}) p(S^f|f \in \bar{B}, \mathcal{M}')$, where

$$p(S_y^g|g \in \bar{G}, \mathcal{M}') = \sum_{H: g \notin H} \int_{\Theta_y^{H \cup \{g\}}} p(\{H, \theta_y^{H \cup \{g\}}\}|\mathcal{M}') p(S_y^g|\{H, \theta_y^{H \cup \{g\}}\}, \mathcal{M}') d\theta_y^{H \cup \{g\}}, \tag{4.7}$$

$$p(S^b|b \in \bar{B}, \mathcal{M}') = \sum_{H: b \notin H} \int_{\Theta^{H \cup \{b\}}} p(\{H, \theta^{H \cup \{b\}}\}|\mathcal{M}') p(S^b|\{H, \theta^{H \cup \{b\}}\}, \mathcal{M}') d\theta^{H \cup \{b\}}. \tag{4.8}$$

Plugging in $p(\phi_y^g|\mathcal{M}')$, $p(\phi^b|\mathcal{M}')$, $p(S_y^g|\phi_y^g, \mathcal{M}')$ and $p(S^b|\phi^b, \mathcal{M}')$, and comparing $p(S_y^g|g \in \bar{G}, \mathcal{M}')$ and $p(S^b|b \in \bar{B}, \mathcal{M}')$ with counterparts in $\mathcal{M}$, we have $\pi^*(f|\mathcal{M}') = \pi^*(f|\mathcal{M})$. □

5 Consistency

A key property of any estimator is consistency: as data are collected, will the estimator converge to the quantity it is to estimate? We are now interested in frequentist asymptotics, that is, the behavior of an estimator under a fixed set of good features, $\bar{G}$, a fixed set of parameters, $\bar{\theta}$, and the corresponding sampling distribution.

Let S_∞ denote a countably infinite labeled dataset, and let S_n denote the first n observations. In general, a sequence of estimators, $\hat{\theta}_n(S_n)$ for $n \geq 1$, of a parameter, $\bar{\theta}$, is said to be strongly consistent at $\bar{\theta}$ if

$$P(\hat{\theta}_n(S_n) \to \bar{\theta} \big| \bar{\theta}) = 1, \tag{5.1}$$

where convergence is understood with respect to a distance metric d, and this probability is taken with respect to the infinite sampling distribution on S_∞ under some true data generation parameter, $\bar{\theta}$. For feature selection, we will use $d(\bar{G}, G) = I(\bar{G} \neq G)$. Under this metric, $G_n \to \bar{G}$ if and only if $G_n = \bar{G}$ for all but finitely many n. The following theorem addresses the convergence of MR, CMNC and NP under any sequence of posteriors, $p(G|S_n)$. The posteriors may be based on any general Bayesian model.

Theorem 2. *Fix S_∞. If $\lim_{n\to\infty} p(\bar{G}|S_n) = 1$, then $G^{MR} \to \bar{G}$ if $T \in (0,1)$, $G^{CMNC} \to \bar{G}$ if $D = |\bar{G}|$, and $G^{NP} \to \bar{G}$ if $\alpha \in (0,1)$.*

Proof. By (2.7), $\lim_{n\to\infty} p(\bar{G}|S_n) = 1$ implies $\pi^*(g) \to 1$ and $\pi^*(b) \to 0$ for all $g \in \bar{G}$ and $b \in \bar{B}$. The consistency of MR and NP follow immediately for the range of T and α specified, and the consistency of CMNC follows if $D = |\bar{G}|$. □

By Theorem 2, if $p(G|S_n)$ converges *almost surely* (a.s.), i.e., with probability 1, to a point mass at $\bar{G}$, then MR (and thus MNC) and NP are strongly consistent and CMNC is strongly consistent when constrained to select the correct number of features. In Section 5.1 we prove that $p(G|S_n)$ converges almost surely for independent Gaussian models under very mild conditions; the data need not be independent or Gaussian.

5.1 Convergence of $p(G|S_n)$ Under Independent Gaussian Models

For fixed $\bar{G}$, let $F_\infty^{\bar{G}}$ be the infinite sampling distribution on S_∞. For fixed S_n, define $\rho = n_0/n$, $c_y^f = s_y^{f*}/(n_y - 1)$ for all $f \in F$ and $y = 0, 1$, and $c^f = s^{f*}/(n-1)$ for all $f \in F$. Throughout this section, we assume $p(G|S_n)$ is calculated under an independent Gaussian model with proper or improper priors on $p(\theta_y^f)$ and $p(\theta^f)$, and (in a slight generalization) allow $p(G)$ to be arbitrary. Allowing $p(G)$ to be arbitrary, equations analogous to (2.12) and (2.18) are straightforward to derive. We have:

$$p(G|S_n) \propto a(G, S_n) z(G, S_n), \tag{5.2}$$

where $z(G, S_n) = p(G) \prod_{f\in G} l(f, S_n)$,

$$l(f, S_n) = L^f(n_0, n_1) \frac{\Gamma(0.5\kappa_0^{f*})\Gamma(0.5\kappa_1^{f*})}{\Gamma(0.5\kappa^{f*})} \left(\frac{2\pi\nu^{f*} 0.5^{\kappa^f - \kappa_0^f - \kappa_1^f} (n-1)^{\kappa^{f*}}}{\nu_0^{f*}\nu_1^{f*}(n_0-1)^{\kappa_0^{f*}}(n_1-1)^{\kappa_1^{f*}}} \right)^{0.5} \tag{5.3}$$

and

$$a(G, S_n) = \prod_{f\in G} \left(\frac{(c^f)^{\kappa^{f*}}}{(c_0^f)^{\kappa_0^{f*}}(c_1^f)^{\kappa_1^{f*}}} \right)^{0.5}. \tag{5.4}$$

We write L^f as a function of n_0 and n_1 to emphasize that it may be allowed to change depending on the sample size. We assume all other inputs and hyper-parameters of the independent Gaussian model are constant across all samples sizes.

Definition 1. *$\bar{G}$ is an independent unambiguous set of good features if, for each $g \in \bar{G}$ μ_y^g and σ_y^g exist and are finite such that either $\mu_0^g \neq \mu_1^g$ or $\sigma_0^g \neq \sigma_0^g$, and for each $b \in \bar{B}$ μ_y^b and σ_y^b exist and are finite such that $\mu^b = \mu_0^b = \mu_1^b$ and $\sigma^b = \sigma_0^b = \sigma_0^b$.*

Definition 2. *S_∞ is called a balanced sample if the label of sample points are such that* $\liminf_{n\to\infty} \rho > 0$ *and* $\limsup_{n\to\infty} \rho < 1$*, and, conditioned on the labels, sample points are independent with points belonging to the same class identically distributed.*

Definition 3. *$p(\theta|G)$ is called semi-proper if, for all $f \in F$, there exists $c > 0$ and $p < 1$ such that*

$$L^f(n_0, n_1) \sim cn^p \tag{5.5}$$

as $n \to \infty$. $f \sim g$ as $n \to \infty$ means $\lim_{n\to\infty} f(n)/g(n) = 1$.

The following theorem proves our desired result. Three lemmas used in the proof are provided in Section S1 of Supplementary Material A. The conditions assumed by the theorem are very mild. Condition (i) is based on Definition 1 and essentially says that $\bar{G}$ is really the feature set we wish to select, i.e., good features must truly have different means or variances between the classes, and bad features must truly have the same means and the same variances between the classes. Conditions (i) and (ii) require certain moments to exist, but there is no requirement for the data to be Gaussian or for features to be independent from each other. Condition (iii) is based on Definition 2 and addresses the sampling strategy; the assumptions are similar to those made by most finite sample data generation models for classification, with an additional requirement on the infinite sample that the proportion of points observed in either class must not converge to zero. Conditions (iv) and (v) place constraints on the inputs to OBF. Condition (iv) requires that OBF assign a non-zero probability prior to the feature set we ultimately wish to select, which is easily achieved by setting $0 < \pi(f) < 1$ for all $f \in F$. Condition (v) is based on Definition 3 and addresses the possibility that one might input different L^f for an improper prior to OBF depending on sample size. Condition (v) is always satisfied with $p = 0$ under proper priors, and under improper priors with L^f set to a positive constant across all samples sizes. By Theorems 2 and 3, under these conditions and posteriors computed based on the independent Gaussian model, we have that MR (and thus MNC and MAP) is strongly consistent, and CMNC (and thus CMAP) is strongly consistent when constrained to select the correct number of features.

The proof of Theorem 3 also characterizes the rate of convergence of the posterior. Under the conditions stated in the theorem, there exist $R > 1$ and $N > 0$ such that $h(g) > R^n$ (a.s.) for all $n > N$ and all good features $g \in \bar{G}$. Equivalently, there exist $0 < r < 1$ and $N > 0$ such that $\pi^*(g) > 1 - r^n$ (a.s.) for all $n > N$ and all $g \in \bar{G}$; thus the marginal posterior of good features converges to 1 at least exponentially (a.s.). Further, there exist $c, N > 0$ such that $h(b) < n^{-c}$ (a.s.) for all $n > N$ and all bad features $b \in \bar{B}$. Equivalently, there exist $c, N > 0$ such that $\pi^*(b) < n^{-c}$ (a.s.) for all $n > N$ and all $b \in \bar{B}$; thus the marginal posterior of bad features converges to 0 at least polynomially (a.s.). Extending these facts to the full posterior on feature sets, there exist $0 < r < 1$ and $N > 0$ such that

$$\frac{p(G|S_n)}{p(\bar{G}|S_n)} < r^n \quad \text{a.s.} \tag{5.6}$$

for all $n > N$ and all G missing at least one feature in $\bar{G}$, and there exist $c, N > 0$ such that

$$\frac{p(G|S_n)}{p(\bar{G}|S_n)} < n^{-c} \quad \text{a.s.} \tag{5.7}$$

for all $n > N$ and all $G \neq \bar{G}$. More discussions on rates of convergence are provided in Section S6 of Supplementary Material A.

Theorem 3. *Suppose the following are true: (i) $\bar{G}$ is an independent unambiguous set of good features, (ii) Fourth order moments exist and are finite for all features $b \in \bar{B}$, (iii) S_∞ is a balanced sample with probability 1, (iv) $p(\bar{G}) \neq 0$, and (v) $p(\theta|G)$ is semi-proper. Then $\lim_{n\to\infty} p(\bar{G}|S_n) = 1$ for $F_\infty^{\bar{G}}$-almost all sequences.*

Proof. It suffices to show that for all $G \subseteq F$ such that $G \neq \bar{G}$,

$$\lim_{n\to\infty} \frac{p(G|S_n)}{p(\bar{G}|S_n)} = 0 \quad \text{a.s.} \tag{5.8}$$

Let $G \neq \bar{G}$. If $p(G) = 0$, then (5.8) holds trivially. Thus, assume $p(G) \neq 0$. Note that

$$\frac{p(G|S_n)}{p(\bar{G}|S_n)} = \frac{z(G,S_n)}{z(\bar{G},S_n)} \prod_{g\in B\cap\bar{G}} \left(\frac{(c_0^g)^{\kappa_0^{g*}} (c_1^g)^{\kappa_1^{g*}}}{(c^g)^{\kappa^{g*}}} \right)^{0.5} \prod_{b\in G\cap\bar{B}} \left(\frac{(c^b)^{\kappa^{b*}}}{(c_0^b)^{\kappa_0^{b*}} (c_1^b)^{\kappa_1^{b*}}} \right)^{0.5}. \tag{5.9}$$

Since $p(\theta|G)$ is semi-proper, by Lemma S1 in Supplementary Material A, there exists $L_1 > 0$ and $q > 0$ such that

$$\frac{z(G,S_n)}{z(\bar{G},S_n)} \sim L_1 n^{q(|\bar{G}|-|G|)} \tag{5.10}$$

as $n \to \infty$ (a.s.), where $\sim$ denotes asymptotic equivalence. Therefore, it suffices to show that for each $g \in B \cap \bar{G}$ and each $b \in G \cap \bar{B}$ we have

$$\lim_{n\to\infty} n^q \left(\frac{(c_0^g)^{\kappa_0^{g*}} (c_1^g)^{\kappa_1^{g*}}}{(c^g)^{\kappa^{g*}}} \right)^{0.5} = 0 \quad \text{a.s.}, \tag{5.11}$$

$$\lim_{n\to\infty} n^{-q} \left(\frac{(c^b)^{\kappa^{b*}}}{(c_0^b)^{\kappa_0^{b*}} (c_1^b)^{\kappa_1^{b*}}} \right)^{0.5} = 0 \quad \text{a.s.} \tag{5.12}$$

First, we prove (5.11). Let $g \in B \cap \bar{G}$. Consider a fixed sample in which $\hat{\mu}_y^g$ converges to μ^g and $\hat{\sigma}_y^g$ converges to σ_y^g for $y = 0, 1$. Since sample points in a class are independent and identically distributed with finite first and second order moments, this event occurs almost surely by the strong law of large numbers. By Lemma S2 in Supplementary Material A, there exists $\epsilon > 0$ and $L_2 > 0$ such that for n large enough

$$n^q \left(\frac{(c_0^g)^{\kappa_0^{g*}} (c_1^g)^{\kappa_1^{g*}}}{(c^g)^{\kappa^{g*}}} \right)^{0.5} < n^q L_2 (1-\epsilon)^{0.5n}. \tag{5.13}$$

Since the limit of the right-hand side is zero, so is that of left-hand side.

Now we prove (5.12). Let $b \in G \cap \bar{B}$. Observe that

$$\frac{c^b}{\hat{\sigma}^b} = 1 + \frac{s^b}{(n-1)\hat{\sigma}^b} + \frac{\nu^b n(\hat{\mu}^b - m^b)^2}{\hat{\sigma}^b (n-1)(\nu^b + n)}. \tag{5.14}$$

Consider a fixed sample in which $\hat{\mu}_y^b$ and $\hat{\mu}^b$ are bounded and $\hat{\sigma}_y^b$ and $\hat{\sigma}^b$ converge to σ^b, which occurs almost surely. There exists $L_4 > 0$ such that for n large enough,

$$1 < \frac{c^b}{\hat{\sigma}^b} < 1 + \frac{L_4}{n}. \tag{5.15}$$

Similarly, there exists $L_{50}, L_{51} > 0$ such that for n large enough

$$1 < \frac{c_0^b}{\hat{\sigma}_0^b} < 1 + \frac{L_{50}}{n} \quad \text{and} \quad 1 < \frac{c_1^b}{\hat{\sigma}_1^b} < 1 + \frac{L_{51}}{n}. \tag{5.16}$$

From (5.15) and (5.16) we conclude there exists $L_6 > 0$ such that for n large enough:

$$\left(\frac{c^b}{(c_0^b)^\rho (c_1^b)^{1-\rho}}\right)^{0.5n} < L_6 \left(\frac{\hat{\sigma}^b}{(\hat{\sigma}_0^b)^\rho (\hat{\sigma}_1^b)^{1-\rho}}\right)^{0.5n}. \tag{5.17}$$

Furthermore, as c^b and c_y^b converge, there exists $L_7 > 0$ such that for n large enough,

$$\left(\frac{(c^b)^{\kappa^b}}{(c_0^b)^{\kappa_0^b} (c_1^b)^{\kappa_1^b}}\right)^{0.5} < L_7. \tag{5.18}$$

Therefore, for n large enough we may write

$$n^{-q}\left(\frac{(c^b)^{\kappa^{b*}}}{(c_0^b)^{\kappa_0^{b*}} (c_1^b)^{\kappa_1^{b*}}}\right)^{0.5} < \frac{L_6 L_7}{n^q}\left(\frac{\hat{\sigma}^b}{(\hat{\sigma}_0^b)^\rho (\hat{\sigma}_1^b)^{1-\rho}}\right)^{0.5n}. \tag{5.19}$$

The following property of sample variance holds, provided that sample moments exist:

$$\begin{aligned}\hat{\sigma}^b &= \rho\hat{\sigma}_0^b + (1-\rho)\hat{\sigma}_1^b + \frac{\rho(1-\rho)n}{n-1}(\hat{\mu}_0^b - \hat{\mu}_1^b)^2 - \frac{1-\rho}{n-1}\hat{\sigma}_0^b - \frac{\rho}{n-1}\hat{\sigma}_1^b \\ &\leq \rho\hat{\sigma}_0^b + (1-\rho)\hat{\sigma}_1^b + \frac{\rho(1-\rho)n}{n-1}(\hat{\mu}_0^b - \hat{\mu}_1^b)^2. \end{aligned} \tag{5.20}$$

Let us consider the sample mean term in (5.20). Since $\hat{\sigma}_0^b, \hat{\sigma}_1^b \to \sigma^b$, for n large enough,

$$\frac{n\rho(1-\rho)(\hat{\mu}_0^b - \hat{\mu}_1^b)^2}{(n-1)(\hat{\sigma}_0^b)^\rho (\hat{\sigma}_1^b)^{1-\rho}} < \frac{2\rho(1-\rho)(\hat{\mu}_0^b - \hat{\mu}_1^b)^2}{\sigma^b}. \tag{5.21}$$

Recall that (5.15) through (5.19) and (5.21) hold when the sample means are bounded and the sample variances converge to σ^b. We now consider the rate of convergence of the means and variances. Suppose x_i, $i = 1, \ldots, n_0$, are the values of feature b for points in class 0. Observe that $(x_i - \mu^b)/\sqrt{\sigma^b}$ are independent random variables with zero mean and unit variance. By the law of the iterated logarithm (Kolmogorov, 1929),

$$\limsup_{n_0 \to \infty} \left| (n_0 \log\log n_0)^{-0.5} \sum_{i=1}^{n_0} \frac{x_i - \mu^b}{\sqrt{\sigma^b}} \right| = \sqrt{2} \quad \text{a.s.} \tag{5.22}$$

Further,

$$|\hat{\mu}_0^b - \mu^b| = \frac{\sqrt{\sigma^b}}{n_0}\left|\sum_{i=1}^{n_0}\frac{x_i - \mu^b}{\sqrt{\sigma^b}}\right|. \tag{5.23}$$

Hence for n large enough,

$$|\hat{\mu}_0^b - \mu^b| < 2\sqrt{\frac{\sigma^b \log\log \rho n}{\rho n}} < 2\sqrt{\frac{\sigma^b \log\log n}{\rho n}} \quad \text{a.s.} \tag{5.24}$$

Similarly, for n large enough,

$$|\hat{\mu}_1^b - \mu^b| < 2\sqrt{\frac{\sigma^b \log\log (1-\rho) n}{(1-\rho) n}} < 2\sqrt{\frac{\sigma^b \log\log n}{(1-\rho) n}} \quad \text{a.s.} \tag{5.25}$$

By the triangle inequality, for n large enough,

$$|\hat{\mu}_0^b - \hat{\mu}_1^b| < 2\sqrt{\sigma^b}\left(\rho^{-0.5} + (1-\rho)^{-0.5}\right)\sqrt{\frac{\log\log n}{n}} \quad \text{a.s.} \tag{5.26}$$

Note that for all $0 < \rho < 1$,

$$\rho(1-\rho)\left(\rho^{-0.5} + (1-\rho)^{-0.5}\right)^2 \leq 2. \tag{5.27}$$

Combining (5.21), (5.26), and (5.27), we see that for n large enough,

$$\frac{n\rho(1-\rho)(\hat{\mu}_0^b - \hat{\mu}_1^b)^2}{(n-1)(\hat{\sigma}_0^b)^{\rho}(\hat{\sigma}_1^b)^{1-\rho}} < 16\frac{\log\log n}{n} \quad \text{a.s.} \tag{5.28}$$

Now, consider variance terms in (5.20). We have another property of sample variance:

$$\begin{aligned}
|\hat{\sigma}_0^b - \sigma^b| &= \left|\frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(x_i - \hat{\mu}_0^b)^2 - \sigma^b\right| \\
&= \left|\frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(x_i - \mu^b + \mu^b - \hat{\mu}_0^b)^2 - \sigma^b\right| \\
&= \left|\frac{1}{n_0 - 1}\sum_{i=1}^{n_0}\left((x_i - \mu^b)^2 - \sigma^b\right) - \frac{n_0}{n_0 - 1}(\mu^b - \hat{\mu}_0^b)^2 + \frac{1}{n_0 - 1}\sigma^b\right| \\
&\leq \left|\frac{1}{n_0 - 1}\sum_{i=1}^{n_0}\left((x_i - \mu^b)^2 - \sigma^b\right)\right| + \frac{n_0}{n_0 - 1}(\mu^b - \hat{\mu}_0^b)^2 + \frac{1}{n_0 - 1}\sigma^b.
\end{aligned} \tag{5.29}$$

Under balanced sampling, ρn increases with n. Note that $\lim_{n\to\infty} \rho n/(\rho n - 1) = 1$, thus $\rho n/(\rho n - 1) < 2$ for n large enough (a.s.). Also, $1/(\rho n - 1) < (\log\log n)/(\rho n)$ for n large enough (a.s.). In addition, we can use (5.24) to bound $(\mu^b - \hat{\mu}_0^b)^2$. Hence, for n large enough,

$$\frac{\rho n}{\rho n - 1}(\mu^b - \hat{\mu}_0^b)^2 + \frac{1}{\rho n - 1}\sigma^b < 9\sigma^b\frac{\log\log n}{\rho n} \quad \text{a.s.} \tag{5.30}$$

Since we assume fourth order (and thus lower order) moments of features in $\bar{B}$ are finite, the variance of $(x_i - \mu^b)^2/\sigma^b$ is finite, and we call this variance K_0. Again applying the law of the iterated logarithm,

$$\frac{1}{\rho n - 1}\sum_{i=1}^{\rho n}\left(\left(\frac{x_i - \mu^b}{\sqrt{\sigma^b}}\right)^2 - 1\right) < 2\sqrt{K_0\frac{\log\log \rho n}{\rho n}} \quad \text{a.s.} \tag{5.31}$$

Combining (5.29), (5.30), and (5.31) we conclude that for n large enough,

$$|\hat{\sigma}_0^b - \sigma^b| < 2\sigma^b\sqrt{K_0\frac{\log\log \rho n}{\rho n}} + 9\sigma^b\frac{\log\log n}{\rho n} \leq 4\sigma^b\sqrt{K_0\frac{\log\log n}{\rho n}} \quad \text{a.s.} \tag{5.32}$$

Similarly, we can show there exists $K_1 > 0$ such that for n large enough,

$$|\hat{\sigma}_1^b - \sigma^b| < 4\sigma^b\sqrt{K_1\frac{\log\log n}{(1-\rho)n}} \quad \text{a.s.} \tag{5.33}$$

Now, observe that

$$\frac{\rho\hat{\sigma}_0^b + (1-\rho)\hat{\sigma}_1^b}{(\hat{\sigma}_0^b)^\rho(\hat{\sigma}_1^b)^{1-\rho}} = \rho\left(\frac{\hat{\sigma}_0^b}{\hat{\sigma}_1^b}\right)^{1-\rho} + (1-\rho)\left(\frac{\hat{\sigma}_0^b}{\hat{\sigma}_1^b}\right)^{-\rho}. \tag{5.34}$$

Using (5.32) and (5.33), we can show that for n large enough,

$$\begin{aligned}\left|\frac{\hat{\sigma}_0^b}{\hat{\sigma}_1^b} - 1\right| &= \left|\frac{\hat{\sigma}_0^b - \hat{\sigma}_1^b}{\hat{\sigma}_1^b}\right| \quad \text{a.s.} \\ &\leq \frac{2}{\sigma^b}\left|\hat{\sigma}_0^b - \hat{\sigma}_1^b\right| \quad \text{a.s.} \\ &\leq \frac{2}{\sigma^b}\left(\left|\hat{\sigma}_0^b - \sigma^b\right| + \left|\hat{\sigma}_1^b - \sigma^b\right|\right) \quad \text{a.s.} \\ &\leq K\left(\frac{1}{\sqrt{\rho}} + \frac{1}{\sqrt{1-\rho}}\right)\sqrt{\frac{\log\log n}{n}} \quad \text{a.s.},\end{aligned} \tag{5.35}$$

where $K = 8\max\{\sqrt{K_0}, \sqrt{K_1}\}$. By Lemma S3 in Supplementary Material A, there exists $r > 0$ such that for all $t \in (0,1)$ and $x \in (1-r, 1+r)$,

$$tx^{1-t} + (1-t)x^{-t} \leq 1 + t(1-t)(x-1)^2. \tag{5.36}$$

Using (5.34), (5.35), and (5.36), we see that for n large enough,

$$\begin{aligned}\frac{\rho\hat{\sigma}_0^b + (1-\rho)\hat{\sigma}_1^b}{(\hat{\sigma}_0^b)^\rho(\hat{\sigma}_1^b)^{1-\rho}} &\leq 1 + K^2\rho(1-\rho)\left(\frac{1}{\sqrt{\rho}} + \frac{1}{\sqrt{1-\rho}}\right)^2\frac{\log\log n}{n} \quad \text{a.s.} \\ &\leq 1 + 2K^2\frac{\log\log n}{n} \quad \text{a.s.},\end{aligned} \tag{5.37}$$

where in the last inequality we have used (5.27). Combining (5.20), (5.28), and (5.37) we see that for n large enough,

$$n^{-q}\left(\frac{\hat{\sigma}^b}{(\hat{\sigma}_0^b)^{\rho}(\hat{\sigma}_1^b)^{1-\rho}}\right)^{0.5n} < n^{-q}\left(1+\left(16+2K^2\right)\frac{\log\log n}{n}\right)^{0.5n} < n^{-q}(\log n)^{K^2+8} \quad \text{a.s.}, \tag{5.38}$$

where in the last inequality we have used the fact that for all $x, t > 0$, $(1 + t/x)^x < e^t$. Since the limit of the right-hand side is 0 whenever $q > 0$, so is that of the left-hand side. Combining (5.19) and (5.38) we see that (5.12) holds almost surely. □

6 Performance and Consistency on Synthetic Data

Here we implement OBF and several other feature selection methods on synthetically generated microarray data. An application on real colon cancer microarray data is provided in Sections S2 and S3 of Supplementary Material A. Although OBF assumes all features are independent with Gaussian class-conditional distributions, the data generation model employed violates these assumptions by generating correlated and non-Gaussian features. Remarkably, OBF is still theoretically consistent by Theorems 2 and 3. Since the main contributions of this paper are theoretical, and numerous extensive simulation studies have already shown that OBF has competitive and robust performance (Foroughi pour and Dalton, 2017d, 2018a,b), our primary objective in this section is to simply observe whether OBF is indeed consistent, i.e. whether it eventually selects the correct feature set as sample size grows. Our secondary objective is to provide new examples showing that OBF enjoys competitive performance, running time, and memory consumption compared with popular Bayesian and non-Bayesian feature selection algorithms, including several methods that OBF has not been compared with before.

The data is generated using a variant of the "synergetic" model originally proposed in Hua et al. (2009). For a fixed sample size, n, in each iteration we assign an equal number of points to class 0 and 1 (n is always even). We generate $|F| = 20{,}000$ features, including a random assignment of 20 *global markers*, 80 *heterogeneous markers*, 11,900 *low-variance non-markers* and 8,000 *high-variance non-markers*. Markers have distinct class conditional distributions, non-markers have identical distributions in both classes, and heterogeneous markers and high-variance non-markers account for unknown subclasses in the data. Global markers, heterogeneous markers, and low-variance non-markers are randomly partitioned into blocks of size $k = 5$. All features within a block are correlated, while all blocks of markers, all blocks of low-variance non-markers, and all high-variance non-markers are independent from each other. All features are also randomly assigned to one of four groups, $i = 0, 1, 2, 3$, such that each group contains one block of global markers, four blocks of heterogeneous markers, 595 blocks of low-variance non-markers and 2,000 high-variance non-markers.

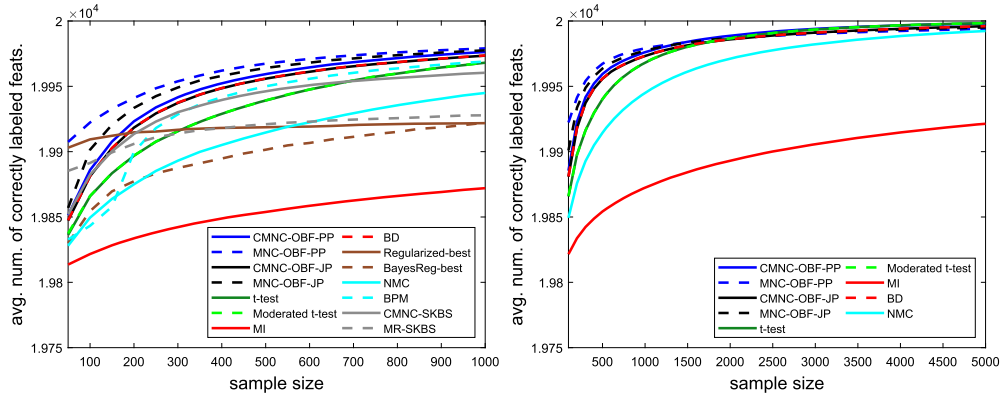
We now focus on how data is generated in group i. The single block of global markers is jointly Gaussian in class $y = 0, 1$ with mean μ_y and covariance matrix $\Sigma_{y,i} = \sigma_{y,i}\Sigma$,

where $\mu_0 = [0, \ldots, 0]$, $\mu_1 = [1, 1/2, \ldots, 1/k]$, diagonal elements of Σ are 1, and off-diagonal elements are $\rho = 0.8$. To generate heterogeneous markers, points in class 1 are further partitioned into $c = 2$ roughly equal size subclasses (when $n_1 = n/2$ is odd, subclass 0 is assigned one more point than subclass 1). For two blocks of heterogeneous markers, points in class 0 or subclass 0 of class 1 are drawn from $\mathcal{N}(\mu_0, \Sigma_{0,i})$ and points in subclass 1 of class 1 are drawn from $\mathcal{N}(\mu_1, \Sigma_{1,i})$. For the remaining two blocks, points in class 0 or subclass 1 of class 1 are drawn from $\mathcal{N}(\mu_0, \Sigma_{0,i})$ and points in subclass 0 of class 1 are drawn from $\mathcal{N}(\mu_1, \Sigma_{1,i})$. Each block of low-variance non-markers is jointly Gaussian with mean μ_0 and covariance matrix $\Sigma_{0,i}$ in both classes. High-variance non-markers are independent and drawn from the mixture of Gaussians $p\mathcal{N}(0, \sigma_{0,i}) + (1 - p)\mathcal{N}(1, \sigma_{1,i})$, where p is independently drawn from a uniform distribution over $(0, 1)$ for each feature. We set $\sigma_{0,0} = \sigma_{1,0} = 0.16$, $\sigma_{0,1} = \sigma_{1,1} = 0.49$, $\sigma_{0,2} = 0.09$, $\sigma_{1,2} = 0.25$, $\sigma_{0,3} = 0.49$ and $\sigma_{1,3} = 0.64$. These values were originally suggested in Hua et al. (2009). Also note that in Hua et al. (2009), there is only one group, and low-variance non-markers are all independent rather than being assigned to blocks.

We implement four variants of Gaussian OBF: MNC-OBF-PP, CMNC-OBF-PP, MNC-OBF-JP and CMNC-OBF-JP. PP refers to a proper prior with $s_0^f = s_1^f = s^f = 0.5$, $\kappa_0^f = \kappa_1^f = \kappa^f = 3$, $m_0^f = m^f = 0$, $m_1^f = 0.2$ and $\nu_0^f = \nu_1^f = \nu^f = 0.1$ for all $f \in F$. These κ's are the smallest integer values where $E(\sigma_0^f)$, $E(\sigma_1^f)$ and $E(\sigma^f)$ exist. JP is based on Jeffreys non-informative prior, and sets $L^f = 0.1$, $s_0^f = s_1^f = s^f = 0$, $\kappa_0^f = \kappa_1^f = \kappa^f = 0$ and $\nu_0^f = \nu_1^f = \nu^f = 0$ for all f. When ν's are 0, m's need not be specified. We set $\pi(f) = 0.005$ for all f under PP and JP. Under MNC, we select all features f such that $\pi^*(f) = h(f)/(1 + h(f)) > 0.5$, where $h(f)$ is given in (2.18). For MNC, the choice of $\pi(f)$ (and L^f under improper priors) affects the average number of features selected; larger $\pi(f)$ and L^f produce larger feature sets. Under CMNC we select the $D = 100$ features maximizing the right-hand side of (2.19). CMNC-OBF-JP reduces to minimizing $(\hat{\sigma}_0^f)^{0.5n_0}(\hat{\sigma}_1^f)^{0.5n_1}/(\hat{\sigma}^f)^{0.5n}$, which is essentially the Pearson and Neyman (1930) statistic. For CMNC, as long as $\pi(f)$ and L^f are constant for all f, their values do not affect the rank of features and thus need not be specified.

In addition to OBF, we implement: Welch's t-test (t-test), a moderated t-test from the `limma` package in R (Smyth, 2004) (Moderated t-test), the Bhattacharyya distance between Gaussian distributions with sample means and variances computed from each class (BD), the mutual information between features and class labels computed from a non-parametric entropy estimator based on sample spacings of order $m = 1$ (Beirlant et al., 1997) (MI), and a bolstered error estimate (Braga-Neto and Dougherty, 2004) under nearest mean classification (NMC). In each case, we output the $D = 100$ top ranked features. Note that these methods are all univariate filters.

We also implement 84 regularized regression methods, using three link functions (linear regression, a GLM with logit link, and a GLM with probit link), two penalty families (LASSO and elastic net), and 14 regularization parameters (using MATLAB's `lassoglm` function we set $\lambda = 0.1, 0.2, 0.5$, $\lambda = \gamma/\sqrt{n}$ for $\gamma = 0.1, 0.2, 0.5, 1, 2, 5, 10$, $\lambda = (\log n)^\gamma/n$ for $\gamma = 0.5, 1, 1.5, 2$, and $\alpha = 0.5$). See Zou (2006) for properties of LASSO under these families of regularization parameters. For each regression method, we output the set of features used in the regression model.



(a) (b)

Figure 1: Average number of correctly labeled features versus sample size for a synthetic microarray model. (a) All feature selection algorithms for n up to 1,000; (b) Univariate filters for n up to 5,000.

Finally, we implement three types of Bayesian variable selection methods: the univariate filter method SKBS (Lock and Dunson, 2015), a regression method using a slab-and-spike prior and probit link (Lee et al., 2003), hereafter called Bayesian probit model (BPM), and a regression method by Makalic and Schmidt (2016) (BayesReg). Due to the high computation cost of these methods, we run each on the top 300 features as ranked by BD, rather than on the full set of 20,000 features. We implement SKBS with $K = 2$ to $K = 7$ Gaussian mixture kernels. We observed best performance with $K = 2$ and report on only this case. As in Lock and Dunson (2015), we find the marginal posterior probability of each feature having distributional differences using a Gibbs sampler with a burn-in period of 1,000 steps and a sampling period of 5,000 steps. We report the $D = 100$ features having largest marginal posteriors with ties broken by BD (CMNC-SKBS), and the set of all features with marginal posteriors greater than $T = 0.9$ (MR-SKBS). We also implemented $T = 0.5$ (the threshold of MNC) and $T = 0.75$, but observed best performance with $T = 0.9$. We implement BPM using default settings in the published code, except we initialize the MCMC chain with the top $D = 100$ features ranked by BD, forgo the burn-in period, and directly generate 5,000 samples. Similar to CMNC-SKBS, we then report the $D = 100$ features having largest marginal posteriors with ties broken by BD. We implement four variations of BayesReg using default settings in the published MATLAB code. Each variant corresponds to one combination of prior (L_1 or horseshoe) and link function (linear or logit). BayesReg outputs a t-statistic, and for each variant of BayesReg we report the $D = 100$ features with largest absolute t-statistic.

This procedure is iterated 600 times for each n, where n increases from 50 to 1,000 in steps of 50. For each algorithm, reported features are labeled markers and unreported features are labeled non-markers. Figure 1(a) shows the average number of correctly labeled features over iterations with respect to n. For each n, Regularized-best

presents the best performance observed among all 84 regularized regression methods, and BayesReg-best presents the best performance among all four BayesReg methods.

In general, the best performing algorithm is MNC-OBF-PP, which is followed by MNC-OBF-JP, CMNC-OBF-PP, then CMNC-OBF-JP and BD. OBF and BD perform well because they can detect differences between both means and variances (Foroughi pour and Dalton, 2018b). Observe that PP outperforms JP. In general, an informed prior like PP can have better performance than a non-informative prior like JP when assumptions are accurate, but may be less robust when assumptions are inaccurate. Also observe that MNC outperforms CMNC. In general, MNC outperforms CMNC when the sample size is small, and CMNC slightly outperforms MNC when the sample size is large. It may seem counterintuitive for MNC to outperform CMNC, since CMNC is directly informed with the true number of markers to select (via D) and MNC is not. However, MNC is given some information about the number of markers through $\pi(f)$—recall that the expected number of good features given $\pi(f)$ can be found in (2.8). In addition, MNC outputs a variable number of features, and under small samples it can be beneficial to output a smaller feature set to avoid selecting features that one is uncertain about. Also note that CMNC-OBF-JP and BD make similar assumptions, and typically have very similar performance, as seen here.

Regularized-best and MR-SKBS appear to perform very well under small samples; however, these methods are the only methods besides MNC-OBF-PP and MNC-OBF-JP that output a variable number of features, and they perform very close to the trivial algorithm that outputs no features (which always labels 19,900 features correctly). CMNC-SKBS also performs fairly well under small samples, but drops below BPM at around $n = 300$ and below Moderated t-test at around $n = 650$. This may be due to insufficient sampling iterations of the Gibbs sampler, or an issue with selecting the number of kernels. Since SKBS models mixtures of Gaussians, it can detect differences between means and variances like OBF and BD, and potentially differences between higher order moments, but performance may be sensitive to the number of kernels used.

Under large samples, BD is followed by BPM, t-test, CMNC-SKBS, then NMC. BPM appears to perform close to BD under large samples because its MCMC chain is initialized with BD. Unlike OBF and BD, t-test and NMC struggle to detect features with similar means but different variances, which usually results in some loss in performance relative to BD, with NMC performing worse than t-test (Foroughi pour and Dalton, 2018b). BayesReg-best has comparable performance to Regularized-best and MR-SKBS, while MI has the poorest performance across all sample sizes. Although MI does not perform well here, as a non-parametric method it can detect any distributional differences, and it has been observed that MI can shine under large differences in skewness (Foroughi pour and Dalton, 2018b).

All univariate filters (OBF, t-test, Moderated t-test, BD, MI and NMC) do not account for correlations between features, while all regression based methods we implemented (Regularized-best, BPM and BayesReg) do account for correlations. Regression-based methods do not perform particularly well, except Regularized-best under small samples (where it tends to output very few features) and BPM under large samples (where performance tracks BD because the MCMC chain is initialized with BD). As discussed in Section 1, classification and regression based methods tend to miss weak

Method	OBF	t-test	NMC	SKBS(300, $K = 2$)	SKBS(300, $K = 7$)	BPM(300)	BayesReg-best(300)
Running time	1	1	7	400	800	2000	300
Memory	< 5MB	< 5MB	< 7MB	20MB	33MB	30MB	25MB
Method	BD	MI	Regularized-best	SKBS(5000, $K = 2$)	SKBS(5000, $K = 7$)	BPM(5000)	BayesReg-best(5000)
Running time	0.9	5	30	> 10000	> 10000	> 20000	> 10000
Memory	< 5MB	< 50MB	< 5MB	300MB	> 500MB	350MB	75MB

*OBF running time is taken as the unit of time.

Table 1: Computation Cost of Feature Selection Algorithms.

features in the presence of strong features, and miss strong features that are correlated to other stronger features, because these features are not very useful in improving the predictive capacity of the model. See Section S4 of Supplementary Material A for more discussion on this.

Table 1 lists the average running time and maximum memory requirement of several methods for $n = 200$ over 10 iterations. MNC-OBF-PP, CMNC-OBF-PP, MNC-OBF-JP and CMNC-ONF-JP have similar computation cost and are reported in the table as "OBF." t-test and Moderated t-test have comparable computation cost and are averaged together in the table and reported as "t-test." "Regularized-best" reports the average computation cost for all 84 regularized regression models. We implement SKBS with $K = 2$ kernels, SKBS with $K = 7$ kernels, BPM, and the four earlier variants of BayesReg after filtering out all but the top 300 features with BD, and again after filtering out all but the top 5,000 features. "BayesReg-best" reports the average computation cost of all four variants of BayesReg. OBF is not only the best performing, but also stands among the fastest methods with low memory requirements. SKBS, BPM and BayesReg all have running times that are orders of magnitude higher than that of OBF and require several times the amount of memory, particularly when run on a larger number of features. Our code is vectorized, which tends to reduce running time at the cost of higher memory consumption.

We conclude this section with a simulation similar to that of Figure 1(a), except we do not implement computationally intensive methods and we let sample size increase from 100 to 5,000 in steps of 100. Figure 1(b) plots the average number of correctly labeled features with respect to sample size. The curves for BD, t-test, Moderated t-test, and all methods based on OBF appear to converge to 20,000, which suggests that these methods are consistent under the current data model. It is also interesting that t-test becomes more competitive for very large sample sizes.

7 Conclusion

OBF should not be used in applications where the objective is dimensionality reduction to design a simpler model or avoid overfitting. Rather, it is designed for applications where *all* features that exhibit distributional differences between the classes should be ranked and reported. That being said, as a filter method, OBF cannot identify a feature that is itself indistinguishable between the classes, while being highly correlated with other features that do have distributional differences. Such features are of interest in biomarker discovery because: (1) they might be paired with other biomarkers to develop better tests for the biological condition of interest, and (2) strong correlations between

genes or gene products suggest possible links in the underlying biological mechanisms, and understanding these links is an important part of the discovery process. Therefore, a major thrust of our future work is in developing models and methods that can take advantage of correlations. A few suboptimal methods have been proposed in prior works (Foroughi pour and Dalton, 2014, 2016a, 2017d, 2018a), however, more work is needed in identifying conditions under which these algorithms are consistent, and in understanding performance and robustness properties of these algorithms.

Finally, note that the OBF framework makes it possible to conduct a Bayesian error analysis for feature selection, much like Bayesian error estimation in classification (Dalton and Dougherty, 2011a,b). For instance, one may find the probability $p(G|S) = P(\bar{G} = G|S)$ in (2.15) or the expectation $E(\ell(G, \bar{G})|S)$ in (3.7) for an arbitrary feature set G, or find the ROC curve defined in (3.14) for an arbitrary feature selection rule. We plan to study Bayesian error analysis under the OBF framework in future work, and to develop and study methods of error analysis that also take into account correlations.

Supplementary Material

Theory of Optimal Bayesian Feature Filtering: Supplementary Material A (DOI: 10.1214/19-BA1182SUPP; .pdf). Here we present three lemmas used in Theorem 3 along with their proofs (Section S1), an example using colon cancer data (Sections S2 and S3), a discussion on regression and classification based feature selection (Section S4), a proof that Gaussian OBF with improper priors is equivalent to a limit based on proper priors (Section S5), and a discussion on the Jeffreys-Lindley paradox (Section S6).

References

Akaike, H. (1980). "The interpretation of improper prior distributions as limits of data dependent proper prior distributions." *Journal of the Royal Statistical Society, Series B (Methodological)*, 42(1): 46–52. MR0567200. 1176

Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5): 971–989. 1169, 1170

Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., and Napolitano, A. (2012). "A review of the stability of feature selection techniques for bioinformatics data." In *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI)*, 356–363. 1170

Baragatti, M. (2011). "Bayesian variable selection for probit mixed models applied to gene selection." *Bayesian Analysis*, 6(2): 209–229. MR2806242. doi: https://doi.org/10.1214/11-BA607. 1170

Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). "Nonpara-

metric entropy estimation: An overview." *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39. MR1471870. 1188

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B (Methodological)*, 289–300. MR1325392. 1179

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer Science & Business Media, second edition. MR0804611. doi: https://doi.org/10.1007/978-1-4757-4286-2. 1176

Braga-Neto, U. and Dougherty, E. R. (2004). "Bolstered error estimation." *Pattern Recognition*, 37(6): 1267–1281. 1188

Carbonetto, P. and Stephens, M. (2012). "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies." *Bayesian Analysis*, 7(1): 73–108. MR2896713. doi: https://doi.org/10.1214/12-BA703. 1170

Cui, K. and Cui, W. (2012). "Spike-and-slab Dirichlet process mixture models." *Open Journal of Statistics*, 2(5): 512–518. 1171

Dalton, L. A. (2013). "Optimal Bayesian feature selection." In *Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 65–68. 1171, 1176

Dalton, L. A. and Dougherty, E. R. (2011a). "Bayesian minimum mean-square error estimation for classification error–Part I: Definition and the Bayesian MMSE error estimator for discrete classification." *IEEE Transactions on Signal Processing*, 59(1): 115–129. MR2789269. doi: https://doi.org/10.1109/TSP.2010.2084572. 1192

Dalton, L. A. and Dougherty, E. R. (2011b). "Bayesian minimum mean-square error estimation for classification error–Part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions." *IEEE Transactions on Signal Processing*, 59(1): 130–144. 1192

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). "Marginalization paradoxes in Bayesian and structural inference." *Journal of the Royal Statistical Society, Series B (Methodological)*, 189–233. MR0365805. 1173

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw Hill. MR0356303. 1176

Diamandis, E. P. (2010). "Cancer biomarkers: Can we turn recent failures into success?" *Journal of the National Cancer Institute*, 102(19): 1462–1467. 1169

Feng, Z., Prentice, R., and Srivastava, S. (2004). "Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective." *Pharmacogenomics*, 5(6): 709–719. 1169

Foroughi pour, A. and Dalton, L. A. (2014). "Optimal Bayesian feature selection on high dimensional gene expression data." In *Proceedings of the 2014 IEEE Global*

Conference on Signal and Information Processing (GlobalSIP), 1402–1405. 1176, 1192

Foroughi pour, A. and Dalton, L. A. (2015). "Optimal Bayesian feature filtering." In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 651–652. 1170, 1171

Foroughi pour, A. and Dalton, L. A. (2016a). "Multiple sclerosis biomarker discovery via Bayesian feature selection." In *Proceedings of the 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 540–541. 1171, 1192

Foroughi pour, A. and Dalton, L. A. (2016b). "Optimal Bayesian feature selection with missing data." In *Proceedings of the 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 35–39. 1172

Foroughi pour, A. and Dalton, L. A. (2017a). "Integrating prior information with Bayesian feature selection." In *Proceedings of the 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB)*, 610–610. 1173

Foroughi pour, A. and Dalton, L. A. (2017b). "Multiclass Bayesian feature selection." In *Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 725–729. 1172

Foroughi pour, A. and Dalton, L. A. (2017c). "Optimal Bayesian feature filtering for single-nucleotide polymorphism data." In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2290–2292. 1171

Foroughi pour, A. and Dalton, L. A. (2017d). "Robust feature selection for block covariance Bayesian models." In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2696–2700. 1170, 1171, 1187, 1192

Foroughi pour, A. and Dalton, L. A. (2018a). "Heuristic algorithms for feature selection under Bayesian models with block-diagonal covariance structure." *BMC Bioinformatics*, 19(3): 70. 1170, 1187, 1192

Foroughi pour, A. and Dalton, L. A. (2018b). "Optimal Bayesian filtering for biomarker discovery: Performance and robustness." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 1169, 1170, 1176, 1187, 1190

Foroughi pour, A. and Dalton, L. A. (2019). "Theory of Optimal Bayesian Feature Filtering: Supplementary Material A." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1182SUPP. 1171

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–373. 1170

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). "Two-sample Bayesian nonparametric hypothesis testing." *Bayesian Analysis*, 10(2): 297–320. MR3420884. doi: https://doi.org/10.1214/14-BA914. 1171

Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). "Performance of feature-selection methods in the classification of high-dimension data." *Pattern Recognition*, 42(3): 409–424. 1187, 1188

Ilyin, S. E., Belkowski, S. M., and Plata-Salamán, C. R. (2004). "Biomarker discovery and validation: Technologies and integrative approaches." *Trends in Biotechnology*, 22(8): 411–416. 1169

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: Frequentist and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 1170

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, U.K: Cambridge University Press. MR1992316. doi: https://doi.org/10.1017/CBO9780511790423. 1173

Kolmogorov, A. (1929). "Über das Gesetz des iterierten Logarithmus." *Mathematische Annalen*, 101(1): 126–135. MR1512520. doi: https://doi.org/10.1007/BF01454828. 1184

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., and Nowe, A. (2012). "A survey on filter techniques for feature selection in gene expression microarray analysis." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4): 1106–1119. 1169

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). "Gene selection: A Bayesian variable selection approach." *Bioinformatics*, 19(1): 90–97. 1170, 1189

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). "Feature selection: A data perspective." *ACM Computing Surveys*, 50(6): 94. 1170

Libbrecht, M. W. and Noble, W. S. (2015). "Machine learning applications in genetics and genomics." *Nature Reviews Genetics*, 16(6): 321–332. 1171

Lock, E. F. and Dunson, D. B. (2015). "Shared kernel Bayesian screening." *Biometrika*, 102(4): 829–842. MR3431556. doi: https://doi.org/10.1093/biomet/asv032. 1170, 1189

Madigan, D. and Raftery, A. E. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's window." *Journal of the American Statistical Association*, 89(428): 1535–1546. 1170

Makalic, E. and Schmidt, D. F. (2016). "High-dimensional Bayesian regularised regression with the BayesReg package." *ArXiv e-prints*. 1189

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 1170

Mitra, R. and Müller, P. (eds.) (2015). *Nonparametric Bayesian inference in biostatistics*. Switzerland: Springer. MR3382176. doi: https://doi.org/10.1007/978-3-319-18968-0. 1171

Müller, P., Parmigiani, G., and Rice, K. (2006). "FDR and Bayesian multiple comparisons rules." In *Proceedings of the Valencia/ISBA 8th World Meeting on Bayesian Statistics*. MR2433200. 1179

Murphy, K. P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution." Technical report. 1175

Ni, Y., Müller, P., Zhu, Y., and Ji, Y. (2017). "Heterogeneous reciprocal graphical models." *Biometrics*, 74(2). MR3825347. doi: https://doi.org/10.1111/biom.12791. 1171

O'Hara, R. B. and Sillanpää, M. J. (2009). "A review of Bayesian variable selection methods: What, how and which." *Bayesian Analysis*, 4(1): 85–117. MR2486240. doi: https://doi.org/10.1214/09-BA403. 1170

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 1170

Pearson, E. S. and Neyman, J. (1930). "On the problem of two samples." In Neyman, J. and Pearson, E. S. (eds.), *Joint Statistical Papers (1967)*, 99–115. MR0208706. 1170, 1188

Ramachandran, N., Srivastava, S., and LaBaer, J. (2008). "Applications of protein microarrays for biomarker discovery." *Proteomics – Clinical Applications*, 2(10–11): 1444–1459. 1169

Rifai, N., Gillette, M. A., and Carr, S. A. (2006). "Protein biomarker discovery and validation: The long and uncertain path to clinical utility." *Nature Biotechnology*, 24(8): 971–983. 1169

Robert, C. P. (1993). "A note on Jeffreys-Lindley paradox." *Statistica Sinica*, 601–608. MR1243404. 1173

Robert, C. P. (2014). "On the Jeffreys-Lindley paradox." *Philosophy of Science*, 81(2): 216–232. MR3235417. doi: https://doi.org/10.1086/675729. 1173

Rockova, V. and Lesaffre, E. (2014). "Incorporating grouping information in Bayesian variable selection with applications in genomics." *Bayesian Analysis*, 9(1): 221–258. MR3188306. doi: https://doi.org/10.1214/13-BA846. 1170

Saeys, Y., Inza, I., and Larrañaga, P. (2007). "A review of feature selection techniques in bioinformatics." *Bioinformatics*, 23(19): 2507–2517. 1169

Shahbaba, B. and Neal, R. (2009). "Nonlinear models using Dirichlet process mixtures." *Journal of Machine Learning Research*, 10: 1829–1850. MR2540778. 1171

Sima, C. and Dougherty, E. R. (2006). "What should be expected from feature selection in small-sample settings." *Bioinformatics*, 22(19): 2430–2436. 1170

Sima, C. and Dougherty, E. R. (2008). "The peaking phenomenon in the presence of feature-selection." *Pattern Recognition Letters*, 29(11): 1667–1674. 1170

Smyth, G. K. (2004). "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology*, 3(1): 1–25. MR2101454. doi: https://doi.org/10.2202/1544-6115.1027. 1188

Xu, X. and Ghosh, M. (2015). "Bayesian variable selection and estimation for group lasso." *Bayesian Analysis*, 10(4): 909–936. MR3432244. doi: https://doi.org/10.1214/14-BA929. 1170

Zhang, L., Xu, X., and Chen, G. (2012). "The exact likelihood ratio test for equality of two normal populations." *The American Statistician*, 66(3): 180–184. MR2993221. doi: https://doi.org/10.1080/00031305.2012.707083. 1170

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. 1188

Acknowledgments

This work is supported by the National Science Foundation (CCF-1422631 and CCF-1453563).