Irrational Exuberance: Correcting Bias in Probability Estimates

Gareth M. James[1], Peter Radchenko[2] and Bradley Rava[1,3]

Abstract

We consider the common setting where one observes probability estimates for a large number of events, such as default risks for numerous bonds. Unfortunately, even with unbiased estimates, selecting events corresponding to the most extreme probabilities can result in systematically underestimating the true level of uncertainty. We develop an empirical Bayes approach "Excess Certainty Adjusted Probabilities" (ECAP), using a variant of Tweedie's formula, which updates probability estimates to correct for selection bias. ECAP is a flexible non-parametric method, which directly estimates the score function associated with the probability estimates, so it does not need to make any restrictive assumptions about the prior on the true probabilities. ECAP also works well in settings where the probability estimates are biased. We demonstrate through theoretical results, simulations, and an analysis of two real world data sets, that ECAP can provide significant improvements over the original probability estimates.

Keywords: Empirical Bayes; selection bias; excess certainty; Tweedie's formula.

1 Introduction

We are increasingly facing a world where automated algorithms are used to generate probabilities, often in real time, for thousands of different events. Just a small handful of examples include finance where rating agencies provide default probabilities on thousands of different risky assets (Kealhofer, 2003; Hull et al., 2005); sporting events where each season ESPN and other sites estimate win probabilities for all the games occurring in a given sport (Leung and Joseph, 2014); politics where pundits estimate the probabilities of candidates winning in congressional and state races during a given election season (Silver, 2018; Soumbatiants et al., 2006); or medicine where researchers estimate the survival probabilities of patients undergoing a given medical procedure (Poses et al., 1997; Smeenk et al., 2007). Moreover, with the increasing availability of enormous quantities of data, there are more and more automated probability estimates being generated and consumed by the general public.

Many of these probabilities have significant real world implications. For example, the rating given to a company's bonds will impact their cost of borrowing, or the estimated risk of a medical procedure will affect the patient's likelihood of undertaking the operation. This leads us to question the accuracy of these probability estimates. Let p_i and $\tilde{p}_i$ respectively represent the true and estimated probability of A_i occurring for a series of events $A_1, \ldots, A_n$. Then, we often seek an unbiased estimator such that $E(\tilde{p}_i|p_i) = p_i$, so $\tilde{p}_i$ is neither systematically too high nor too low. Of course, there are many recent examples where this unbiasedness assumption has not held. For

[1] Department of Data Sciences and Operations, University of Southern California.

[2] University of Sydney.

[3] Research is generously supported by the NSF GRFP in Mathematical Statistics.

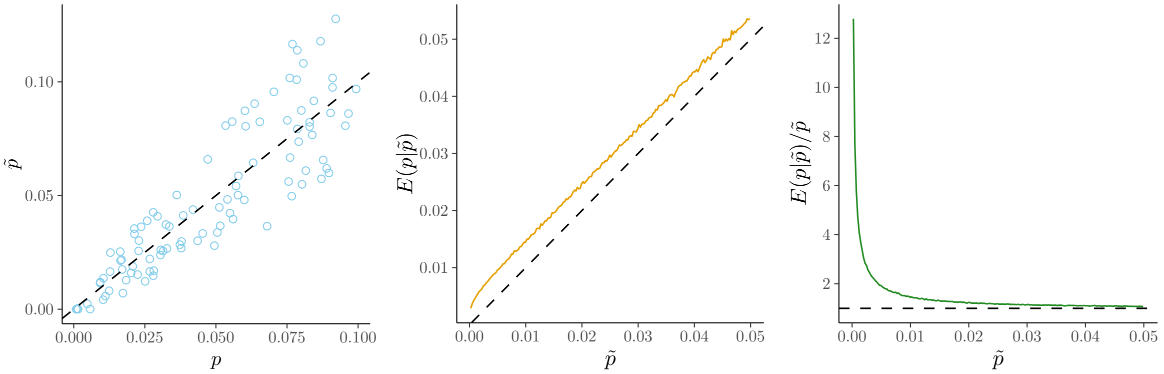


Figure 1: Left: Simulated p_i and associated $\tilde{p}_i$. The probability estimates are unbiased. Center: The average value of p_i, as a function of $\tilde{p}_i$ i.e. $E(p_i|\tilde{p}_i)$ (orange line) is systematically higher than $\tilde{p}_i$ (dashed line). Right: The ratio of $E(p_i|\tilde{p}_i)$ relative to $\tilde{p}_i$, as a function of $\tilde{p}_i$. An ideal ratio would be one (dashed line).

example, prior to the financial crisis of 2008 rating agencies systematically under estimated the risk of default for mortgage backed securities so $E(\tilde{p}_i|p_i) < p_i$. Similarly, in the lead up to the 2016 US presidential election political pundits significantly underestimated the uncertainty in which candidate would win.

However, even when unbiasedness does hold, using $\tilde{p}_i$ as an estimate for p_i can cause significant problems. Consider, for example, a conservative investor who only purchases bonds with extremely low default risk. When presented with n estimated bond default probabilities $\tilde{p}_1, \ldots, \tilde{p}_n$ from a rating agency, she only invests when $\tilde{p}_i = 0.001$. Let us suppose that the rating agency has done a careful risk assessment, so their probability estimates are unbiased for all n bonds. What then is the fraction of the investor's bonds which will actually default? Given that the estimates are unbiased, one might imagine (and the investor is certainly hoping) that the rate would be close to 0.001. Unfortunately, the true default rate may be much higher.

Figure 1 provides an illustration. We first generated a large number of probabilities p_i from a uniform distribution and then produced corresponding $\tilde{p}_i$ in such a way that $E(\tilde{p}_i|p_i) = p_i$ for $i = 1, \ldots, n$. In the left panel of Figure 1 we plotted a random sample of 100 of these probabilities, concentrating on values less than 10%. While there is some variability in the estimates, there is no evidence of bias in $\tilde{p}_i$. In the middle panel we used the simulated data to compute the average value of p_i for any given value of $\tilde{p}_i$ i.e. $E(p_i|\tilde{p}_i)$. A curious effect is observed. At every point the average value of p_i (orange line) is systematically higher than $\tilde{p}_i$ (dashed line) i.e. $E(p_i|\tilde{p}_i) > \tilde{p}_i$. Finally, in the right panel we have plotted the ratio of $E(p_i|\tilde{p}_i)$ to $\tilde{p}_i$. Ideally this ratio should be approximately one, which would, for example, correspond to the true risk of a set of bonds equalling the estimated risk. However, for small values of $\tilde{p}_i$ we observe ratios far higher than one. So, for example, our investor who only purchases bonds with an estimated default risk of $\tilde{p}_i = 0.001$ will in fact find that 0.004 of her bonds end up defaulting, a 400% higher risk level than she intended to take!

These somewhat surprising results are not a consequence of this particular simulation setting. It is in fact an instance of selection bias, a well known issue which occurs when the selection of observations is made in such a way, e.g. selecting the most extreme observations, that they can no longer be considered random samples from the underlying population. If this bias is not taken into account then any future analyses will provide a distorted estimate of the population. Consider the setting where we observe $X_1, \ldots, X_n$ with $E(X_i) = \mu_i$ and wish to estimate μ_i based on an observed X_i. Then it is well known that the conditional expectation $E(\mu_i|X_i)$ corrects for any selection bias associated with choosing X_i in a non-random fashion (Efron, 2011). Numerous approaches have been suggested to address selection bias, with most methods imposing some form of shrinkage to either explicitly, or implicitly, estimate $E(\mu_i|X_i)$. Among linear shrinkage methods, the James-Stein estimator (James and Stein, 1961) is the most well known, although many others exist (Efron and Morris, 1975; Ikeda et al., 2016). There are also other popular classes of methods, including: non-linear approaches utilizing sparse priors (Donoho and Johnstone, 1994; Abramovich et al., 2006; Bickel and Levina, 2008; Ledoit and Wolf, 2012), Bayesian estimators (Gelman and Shalizi, 2012) and empirical Bayes methods (Jiang and Zhang, 2009; Brown and Greenshtein, 2009; Petrone et al., 2014).

For Gaussian data, Tweedie's formula (Robbins, 1956) provides an elegant empirical Bayes estimate for $E(\mu_i|X_i)$, using only the marginal distribution of X_i. While less well known than the James-Stein estimator, it has been shown to be an effective non-parametric approach for addressing selection bias (Efron, 2011). The approach can be automatically adjusted to lean more heavily on parametric assumptions when little data is available, but in settings such as ours, where large quantities of data have been observed, it provides a highly flexible non-parametric shrinkage method (Benjamini and Yekutieli, 2005; Henderson and Newton, 2015).

However, the standard implementation of Tweedie's formula assumes that, conditional on μ_i, the observed data follow a Gaussian distribution. Most shrinkage methods make similar distributional assumptions or else model the data as unbounded, which makes little sense for probabilities. What then would be a better estimator for low probability events? In this paper we propose an empirical Bayes approach, called "Excess Certainty Adjusted Probability" (ECAP), specifically designed for probability estimation in settings with a large number of observations. ECAP uses a variant of Tweedie's formula which models $\tilde{p}_i$ as coming from a beta distribution, automatically ensuring the estimate is bounded between 0 and 1. We provide theoretical and empirical evidence demonstrating that the ECAP estimate is generally significantly more accurate than $\tilde{p}_i$.

This paper makes three key contributions. First, we convincingly demonstrate that even an unbiased estimator $\tilde{p}_i$ can provide a systematically sub-optimal estimate for p_i, especially in situations where large numbers of probability estimates have been generated. This leads us to develop the oracle estimator for p_i, which results in a substantial improvement in expected loss. Second, we introduce the ECAP method which estimates the oracle. ECAP does not need to make any assumptions about the distribution of p_i. Instead, it relies on estimating the marginal distribution, and conditional accuracy, of $\tilde{p}_i$, a relatively easy problem in the increasingly common situation where we observe a large number of probability estimates. Finally, we extend ECAP to the biased data setting where $\tilde{p}_i$ represents a biased observation of p_i and show that even in this setting we are able to recover systematically superior estimates of p_i.

The paper is structured as follows. In Section 2 we first formulate a model for $\tilde{p}_i$ and a loss function for estimating p_i. We then provide a closed form expression for the corresponding oracle estimator and its associated reduction in expected loss. We conclude Section 2 by proposing the ECAP estimator for the oracle and deriving its theoretical properties. Section 3 provides two extensions. First, we propose a bias corrected version of ECAP, which can detect situations where

$\tilde{p}_i$ is a biased estimator for p_i and automatically adjust for the bias. Second, we generalize the ECAP model from Section 2. Next, Section 4 contains results from an extensive simulation study that examines how well ECAP works to estimate p_i, in both the unbiased and biased settings. Section 5 illustrates ECAP on two interesting real world data sets. The first is a unique set of probabilities from ESPN predicting, in real time, the winner of various NCAA football games, and the second contains the win probabilities of all candidates in the 2018 US midterm elections. We conclude with a discussion and possible future extensions in Section 6. Proofs of all theorems are provided in the appendix.

2 Methodology

Let $\tilde{p}_1, \ldots, \tilde{p}_n$ represent initial estimates of events $A_1, \ldots, A_n$ occurring. In practice, we assume that $\tilde{p}_1, \ldots, \tilde{p}_n$ have already been generated, by previous analysis or externally, say, by an outside rating agency in the case of the investment example. Our goal is to construct estimators $\hat{p}_1(\tilde{p}_1), \ldots, \hat{p}_n(\tilde{p}_n)$ which provide more accurate estimates for $p_1, \ldots, p_n$. In order to derive the estimator we first choose a model for $\tilde{p}_i$ and select a loss function for $\hat{p}_i$, which allows us to compute the corresponding oracle estimator p_{i0}. Finally, we provide an approach for generating an estimator for the oracle $\hat{p}_i$. In this section we only consider the setting where $\tilde{p}_i$ is assumed to be an unbiased estimator for p_i. We extend our approach to the more general setting where $\tilde{p}_i$ may be a biased estimator in Section 3.1.

2.1 Modeling $\tilde{p}_i$ and Selecting a Loss Function

Given that $\tilde{p}_i$ is a probability, we model its conditional distribution using the beta distribution[1]. In particular, we model

$$\tilde{p}_i | p_i \sim Beta(\alpha_i, \beta_i), \quad \text{where} \quad \alpha_i = \frac{p_i}{\gamma^*}, \quad \beta_i = \frac{1 - p_i}{\gamma^*}, \tag{1}$$

and γ^* is a fixed parameter which influences the variance of $\tilde{p}_i$. Under (1),

$$E(\tilde{p}_i | p_i) = p_i \quad \text{and} \quad Var(\tilde{p}_i | p_i) = \frac{\gamma^*}{1 + \gamma^*} p_i (1 - p_i), \tag{2}$$

so $\tilde{p}_i$ is an unbiased estimate for p_i, which becomes more accurate as γ^* tends to zero. Figure 2 provides an illustration of the density function of $\tilde{p}_i$ for three different values of p_i. In principle, this model could be extended to incorporate observation specific variance terms γ_i^*. Unfortunately, in practice γ^* needs to be estimated, which would be challenging if we assumed a separate term for each observation. However, in some settings it may be reasonable to model $\gamma_i^* = w_i \gamma^*$, where w_i is a known weighting term, in which case only one parameter needs to be estimated.

Next, we select a loss function for our estimator to minimize. One potential option would be to use a standard squared error loss, $L(\hat{p}_i) = E\left(p_i - \hat{p}_i\right)^2$. However, this loss function is not the most reasonable approach in this setting. Consider for example the event corresponding to a bond defaulting, or a patient dying during surgery. If the bond has junk status, or the surgery is highly risky, the true probability of default or death might be $p_i = 0.26$, in which case an estimate of

[1] We consider a more general class of distributions for $\tilde{p}_i$ in Section 3.2

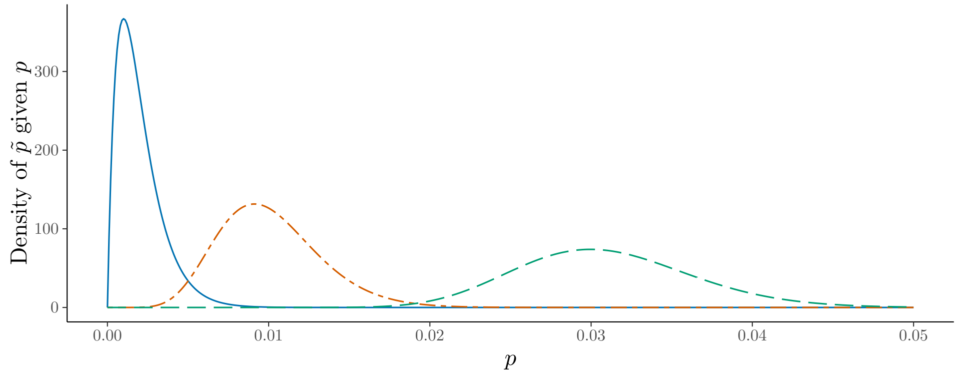


Figure 2: Density functions for $\tilde{p}_i$ given $p_i = 0.002$ (blue / solid), $p_i = 0.01$ (orange / dot-dashed), and $p_i = 0.03$ (green / dashed). In all three cases $\gamma^* = 0.001$.

$\hat{p}_i = 0.25$ would be considered very accurate. It is unlikely that an investor or patient would have made a different decision if they had instead been provided with the true probability of 0.26.

However, if the bond, or surgery, are considered very safe we might provide an estimated probability of $\hat{p}_i = 0.0001$, when the true probability is somewhat higher at $p_i = 0.01$. The absolute error in the estimate is actually slightly lower in this case, but the patient or investor might well make a very different decision when given a 1% probability of a negative outcome vs a one in ten thousand chance.

In this sense, the error between p_i and $\hat{p}_i$ *as a percentage of* $\hat{p}_i$ is a far more meaningful measure of precision. In the first example we have a percentage error of only 4%, while in the second instance the percentage error is almost 10,000%, indicating a far more risky proposition. To capture this concept of relative error we introduce as our measure of accuracy a quantity we call the "Excess Certainty", which is defined as

$$\text{EC}(\hat{p}_i) = \frac{p_i - \hat{p}_i}{\min(\hat{p}_i, 1 - \hat{p}_i)}. \tag{3}$$

In the first example EC = 0.04, while in the second example EC = 99. Note, we include $\hat{p}_i$ in the denominator rather than p_i because we wish to more heavily penalize settings where the estimated risk is far lower than the true risk (irrational exuberance) compared to the alternative where true risk is much lower.

Ideally, the excess certainty of any probability estimate should be very close to zero. Thus, we adopt the following expected loss function,

$$L(\hat{p}_i, \tilde{p}_i) = E_{p_i}\left(\text{EC}(\hat{p}_i)^2 | \tilde{p}_i\right), \tag{4}$$

where the expectation is taken over p_i, conditional on $\tilde{p}_i$. Our aim is to produce an estimator $\hat{p}_i$ that minimizes (4) conditional on the observed value of $\tilde{p}_i$. It is worth noting that if our goal was solely to remove selection bias then we could simply compute $E(p_i|\tilde{p}_i)$, which would be equivalent to minimizing $E\left[(p_i - \hat{p}_i)^2 | \tilde{p}_i\right]$. Minimizing (4) generates a similar shrinkage estimator, which also removes the selection bias, but, as we discuss in the next section, it actually provides additional shrinkage to account for the fact that we wish to minimize the relative, or percentage, error.

2.2 The Oracle Estimator

We now derive the oracle estimator, p_{i0}, which minimizes the loss function given by (4),

$$p_{i0} = \arg\min_{a} E_{p_i}\left[\mathrm{EC}(a)^2 | \tilde{p}_i\right]. \tag{5}$$

Our ECAP estimate aims to approximate the oracle. Theorem 1 below provides a relatively simple closed form expression for p_{i0} and a bound on the minimum reduction in loss from using p_{i0} relative to any other estimator.

Theorem 1 *For any distribution of $\tilde{p}_i$,*

$$p_{i0} = \begin{cases} \min\left(E(p_i|\tilde{p}_i) + \frac{Var(p_i|\tilde{p}_i)}{E(p_i|\tilde{p}_i)}\,,\,0.5\right), & E(p_i|\tilde{p}_i) \leq 0.5 \\ \max\left(0.5\,,\,E(p_i|\tilde{p}_i) - \frac{Var(p_i|\tilde{p}_i)}{1-E(p_i|\tilde{p}_i)}\right), & E(p_i|\tilde{p}_i) > 0.5. \end{cases} \tag{6}$$

Furthermore, for any $p_i' \neq p_{i0}$,

$$L(p_i', \tilde{p}_i) - L(p_{i0}, \tilde{p}_i) \geq \begin{cases} E\left(p_i^2|\tilde{p}_i\right)\left[\frac{1}{p_i'} - \frac{1}{p_{i0}}\right]^2, & p_{i0} \leq 0.5 \\ E\left([1-p_i]^2|\tilde{p}_i\right)\left[\frac{1}{1-p_i'} - \frac{1}{1-p_{i0}}\right]^2, & p_{i0} \geq 0.5. \end{cases} \tag{7}$$

Remark 1 *Note that both bounds in (7) are valid when $p_{i0} = 0.5$.*

We observe from this result that the oracle estimator starts with the conditional expectation $E(p_i|\tilde{p}_i)$ and then shifts the estimate towards 0.5 by an amount $\frac{Var(p_i|\tilde{p}_i)}{\min(E(p_i|\tilde{p}_i), 1-E(p_i|\tilde{p}_i))}$. However, if this would move the estimate past 0.5 then the estimator simply becomes 0.5.

Figure 3 plots the average excess certainty (3) from using $\tilde{p}_i$ to estimate p_i (orange lines) and from using p_{i0} to estimate p_i (green lines), for three different values of γ^*. Recall that an ideal EC should be zero, but the observed values for $\tilde{p}_i$ are far larger, especially for higher values of γ^* and lower values of $\tilde{p}_i$. Note that, as a consequence of the minimization of the expected squared loss function (4), the oracle is slightly conservative with a negative EC, which is due to the variance term in (6).

It is worth noting that Theorem 1 applies for any distribution of $\tilde{p}_i|p_i$ and does not rely on our model (1). If we further assume that (1) holds, then Theorem 2 provides explicit forms for $E(p_i|\tilde{p}_i)$ and $Var(p_i|\tilde{p}_i)$.

Theorem 2 *Under (1),*

$$\begin{aligned} E(p_i|\tilde{p}_i) &= \mu_i \equiv \tilde{p}_i + \gamma^*\left[g^*(\tilde{p}_i) + 1 - 2\tilde{p}_i\right] && (8)\\ Var(p_i|\tilde{p}_i) &= \sigma_i^2 \equiv \gamma^*\tilde{p}_i(1-\tilde{p}_i) + {\gamma^*}^2\tilde{p}_i(1-\tilde{p}_i)\left[g^{*\prime}(\tilde{p}_i) - 2\right], && (9) \end{aligned}$$

where $g^*(\tilde{p}_i) = \tilde{p}_i(1-\tilde{p}_i)v^*(\tilde{p}_i)$, $v^*(\tilde{p}_i) = \frac{\partial}{\partial \tilde{p}_i}\log f^*(\tilde{p}_i)$ is the score function of $\tilde{p}_i$ and $f^*(\tilde{p}_i)$ is the marginal density of $\tilde{p}_i$.

If we also assume that the distribution of p_i is symmetric then further simplifications are possible.

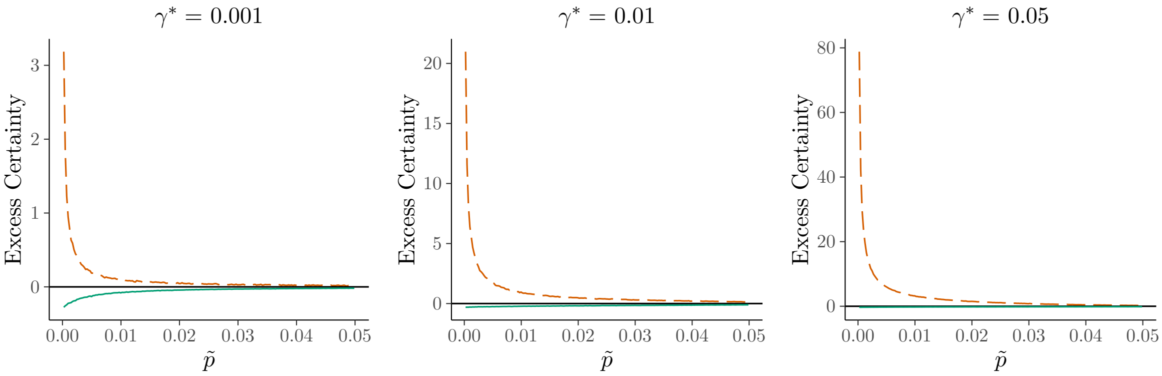


Figure 3: Average excess certainty as a function of $\tilde{p}_i$ for three different values of γ^* (orange / dashed line). All plots exhibit excess certainty far above zero but the issue grows worse as γ^* gets larger, corresponding to more variance in $\tilde{p}_i$. The green (solid) line in each plot corresponds to the average excess certainty for the oracle estimator p_{i0}.

Corollary 1 *If the prior distribution of p_i is symmetric about 0.5, then*

$$p_{i0} = \begin{cases} \min\left(E(p_i|\tilde{p}_i) + \frac{Var(p_i|\tilde{p}_i)}{E(p_i|\tilde{p}_i)}\,,\, 0.5\right), & \tilde{p}_i \le 0.5 \\ \max\left(0.5\,,\, E(p_i|\tilde{p}_i) - \frac{Var(p_i|\tilde{p}_i)}{1-E(p_i|\tilde{p}_i)}\right), & \tilde{p}_i > 0.5, \end{cases} \tag{10}$$

$$g^*(0.5) = 0, \quad \textit{and} \quad g^*(\tilde{p}_i) = -g^*(1-\tilde{p}_i). \tag{11}$$

A particularly appealing aspect of Theorem 2 and its corollary is that $g^*(\tilde{p}_i)$ is only a function of the marginal distribution of $\tilde{p}_i$, so that it can be estimated directly using the observed probabilities $\tilde{p}_i$. In particular, we do not need to make any assumptions about the distribution of p_i in order to compute $g^*(\tilde{p}_i)$.

2.3 Estimation

In order to estimate p_{i0} we must form estimates for $g^*(\tilde{p}_i)$, its derivative $g^{*\prime}(t)$, and γ^*.

2.3.1 Estimation of g

Let $\hat{g}(\tilde{p})$ represent our estimator of $g^*(\tilde{p})$. Given that $g^*(\tilde{p})$ is a function of the marginal distribution of $\tilde{p}_i$, i.e. $f^*(\tilde{p}_i)$, then one could estimate $g^*(\tilde{p}_i)$ by $\tilde{p}_i(1-\tilde{p}_i)\hat{f}'(\tilde{p}_i)/\hat{f}(\tilde{p}_i)$, where $\hat{f}(\tilde{p}_i)$ and $\hat{f}'(\tilde{p}_i)$ are respectively estimates for the marginal distribution of $\tilde{p}_i$ and its derivative. However, this approach requires dividing by the estimated density function, which can produce a highly unstable estimate in the boundary points, precisely the region we are most interested in.

Instead we directly estimate $g^*(\tilde{p})$ by choosing $\hat{g}(\tilde{p})$ so as to minimize the risk function, which is defined as $R(g) = E[g(\tilde{p}) - g^*(\tilde{p})]^2$ for every candidate function g. The following result provides an explicit form for the risk.

Theorem 3 *Suppose that model* (1) *holds, and the prior for p has a bounded density. Then,*

$$R(g) = Eg(\tilde{p})^2 + 2E\left[g(\tilde{p})(1-2\tilde{p}) + \tilde{p}(1-\tilde{p})g'(\tilde{p})\right] + C \tag{12}$$

for all bounded and differentiable functions g, where C is a constant that does not depend on g.

Remark 2 *We show in the proof of Theorem 3 that g^* is bounded and differentiable so* (12) *holds for $g = g^*$.*

Theorem 3 suggests that we can approximate the risk, up to an irrelevant constant, by

$$\hat{R}(g) = \frac{1}{n}\sum_{i=1}^{n} g(\tilde{p}_i)^2 + 2\frac{1}{n}\sum_{i=1}^{n}\left[g(\tilde{p}_i)(1-2\tilde{p}_i) + \tilde{p}_i(1-\tilde{p}_i)g'(\tilde{p}_i)\right]. \tag{13}$$

However, simply minimizing (13) would provide a poor estimate for $g^*(\tilde{p})$ because, without any smoothness constraints, $\hat{R}(g)$ can be trivially minimized. Hence, we place a smoothness penalty on our criterion by minimizing

$$Q(g) = \hat{R}(g) + \lambda \int g''(\tilde{p})^2 d\tilde{p}, \tag{14}$$

where $\lambda > 0$ is a tuning parameter which adjusts the level of smoothness in $g(\tilde{p})$. We show in our theoretical analysis in Section 2.4 (see the proof of Theorem 4) that, much as with the more standard curve fitting setting, the solution to criteria of the form in (14) can be well approximated using a natural cubic spline, which provides a computationally efficient approach to compute $g(\tilde{p})$.

Let $\mathbf{b}(x)$ represent the vector of basis functions for a natural cubic spline, with knots at $\tilde{p}_1, \ldots, \tilde{p}_n$, restricted to satisfy $\mathbf{b}(0.5) = \mathbf{0}$. Then, in minimizing $Q(g)$ we only need to consider functions of the form $g(\tilde{p}) = \mathbf{b}(\tilde{p})^T\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is the basis coefficients. Thus, (14) can be re-expressed as

$$Q_n(\boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\eta}^T\mathbf{b}(\tilde{p}_i)\mathbf{b}(\tilde{p}_i)^T\boldsymbol{\eta} + 2\frac{1}{n}\sum_{i=1}^{n}\left[(1-2\tilde{p}_i)\mathbf{b}(\tilde{p}_i)^T + \tilde{p}_i(1-\tilde{p}_i)\mathbf{b}'(\tilde{p}_i)^T\right]\boldsymbol{\eta} + \lambda\boldsymbol{\eta}^T\Omega\boldsymbol{\eta} \tag{15}$$

where $\Omega = \int \mathbf{b}''(\tilde{p})\mathbf{b}''(\tilde{p})^T d\tilde{p}$. Standard calculations show that (15) is minimized by setting

$$\hat{\boldsymbol{\eta}} = -\left(\sum_{i=1}^{n}\mathbf{b}(\tilde{p}_i)\mathbf{b}(\tilde{p}_i)^T + n\lambda\Omega\right)^{-1}\sum_{i=1}^{n}\left[(1-2\tilde{p}_i)\mathbf{b}(\tilde{p}_i) + \tilde{p}_i(1-\tilde{p}_i)\mathbf{b}'(\tilde{p}_i)\right]. \tag{16}$$

If the prior distribution of p_i is not assumed to be symmetric, then $g^*(\tilde{p}_i)$ should be directly estimated for $0 \leq \tilde{p}_i \leq 1$. However, if the prior is believed to be symmetric this approach is inefficient, because it does not incorporate the identity $g^*(\tilde{p}_i) = -g^*(1-\tilde{p}_i)$. Hence, a superior approach involves flipping all of the $\tilde{p}_i > 0.5$ across 0.5, thus converting them into $1-\tilde{p}_i$, and then using both the flipped and the unflipped $\tilde{p}_i$ to estimate $g(\tilde{p}_i)$ between 0 and 0.5. Finally, the identity $\hat{g}(\tilde{p}_i) = -\hat{g}(1-\tilde{p}_i)$ can be used to define $\hat{g}$ on $(0.5, 1]$. This is the approach we use for the remainder of the paper.

Equation (16) allows us to compute estimates for $E(p_i|\tilde{p}_i)$ and $\mathrm{Var}(p_i|\tilde{p}_i)$:

$$\begin{aligned} \hat{\mu}_i &= \tilde{p}_i + \hat{\gamma}(\mathbf{b}(\tilde{p}_i)^T\hat{\boldsymbol{\eta}} + 1 - 2\tilde{p}_i) & (17)\\ \hat{\sigma}_i^2 &= \hat{\gamma}\tilde{p}_i(1-\tilde{p}_i) + \hat{\gamma}^2\tilde{p}_i(1-\tilde{p}_i)[\mathbf{b}'(\tilde{p}_i)^T\hat{\boldsymbol{\eta}} - 2]. & (18)\end{aligned}$$

Equations (17) and (18) can then be substituted into (10) to produce the ECAP estimator $\hat{p}_i$.

2.3.2 Estimation of λ and γ^*

In computing (17) and (18) we need to provide estimates for γ^* and λ. We choose λ so as to minimize a cross-validated version of the estimated risk (13). In particular, we randomly partition the probabilities into K roughly even groups: $G_1, \ldots, G_K$. Then, for given values of λ and k, $\hat{\boldsymbol{\eta}}_{k\lambda}$ is computed via (16), with the probabilities in G_k excluded from the calculation. We then compute the corresponding estimated risk on the probabilities in G_k:

$$R_{k\lambda} = \sum_{i \in G_k} \hat{h}_{ik}^2 + 2 \sum_{i \in G_k} \left[(1 - 2\tilde{p}_i)\hat{h}_{ik} + \tilde{p}_i(1 - \tilde{p}_i)\hat{h}'_{ik} \right],$$

where $\hat{h}_{ik} = \mathbf{b}(\tilde{p}_i)^T \hat{\boldsymbol{\eta}}_{k\lambda}$ and $\hat{h}'_{ik} = \mathbf{b}'(\tilde{p}_i)^T \hat{\boldsymbol{\eta}}_{k\lambda}$. This process is repeated K times for $k = 1, \ldots, K$, and

$$R_\lambda = \frac{1}{n} \sum_{k=1}^{K} R_{k\lambda}$$

is computed as our cross-validated risk estimate. Finally, we choose $\hat{\lambda} = \arg\min_\lambda R_\lambda$.

To estimate γ^* we need a measure of the accuracy of $\tilde{p}_i$ as an estimate of p_i. In some cases that information may be available from previous analyses. For example, if the estimates $\tilde{p}_i$ were obtained by fitting a logistic regression model, we could compute the standard errors on the estimated coefficients and hence form a variance estimate for each $\tilde{p}_i$. We would estimate γ^* by matching the computed variance estimates to the expression (2) for the conditional variance under the ECAP model.

Alternatively, we can use previously observed outcomes of A_i to estimate γ^*. Suppose that we observe

$$Z_i = \begin{cases} 1 & A_i \text{ occured}, \\ 0 & A_i \text{ did not occur}, \end{cases} \tag{19}$$

for $i = 1, \ldots, n$. Then a natural approach is to compute the conditional log-likelihood function for Z_i given $\tilde{p}_i$. Namely,

$$l_\gamma = \sum_{i: Z_i = 1} \log(\hat{\mu}_i^\gamma) + \sum_{i: Z_i = 0} \log(1 - \hat{\mu}_i^\gamma), \tag{20}$$

where $\hat{\mu}_i^\gamma$ is the ECAP estimate of $E(p_i|\tilde{p}_i)$ generated by substituting in a particular value of γ into (17). We then choose the value of γ that maximizes (20).

As an example of this approach, consider the ESPN data recording probabilities of victory for various NCAA football teams throughout each season. To form an estimate for γ^* we can take the observed outcomes of the games from last season (or the first couple of weeks of this season if there are no previous games available), use these results to generate a set of Z_i, and then choose the γ that maximizes (20). One could then form ECAP estimates for future games during the season, possibly updating the γ estimate as new games are played.

2.4 Large sample results

In this section we investigate the large sample behavior of the ECAP estimator. More specifically, we show that, under smoothness assumptions on the function g^*, the ECAP adjusted probabilities are consistent estimators of the corresponding oracle probabilities, defined in (5). We establish an analogous result for the corresponding values of the loss function, defined in (4). In addition to demonstrating consistency we also derive the rates of convergence. Our method of proof takes advantage of the theory of empirical processes, however, the corresponding arguments go well beyond a simple application of the existing results.

We let f^* denote the marginal density of the observed $\tilde{p}_i$ and define the $L_2(\tilde{P})$ norm of a given function $u(\tilde{p})$ as $\|u\| = [\int_0^1 u^2(\tilde{p}) f^*(\tilde{p}) d\tilde{p}]^{1/2}$. We denote the corresponding empirical norm, $[(1/n)\sum_{i=1}^n u^2(\tilde{p}_i)]^{1/2}$, by $\|u\|_n$. To simplify the presentation of the results, we define

$$r_n = n^{-4/7}\lambda_n^{-1} + n^{-2/7} + \lambda_n \qquad \text{and} \qquad s_n = 1 + n^{-4/7}\lambda_n^{-2}.$$

We write $\hat{g}$ for the minimizer of criterion (14) over all natural cubic spline functions g that correspond to the sequence of n knots located at the observed $\tilde{p}_i$. For concreteness, we focus on the case where criterion (14) is computed over the entire interval $[0, 1]$. However, all of the results in this section continue to hold if $\hat{g}$ is determined by only computing the criterion over $[0, 0.5]$, according to the estimation approach described in Section 2.3.1. The following result establishes consistency and rates of convergence for $\hat{g}$ and $\hat{g}'$.

Theorem 4 *If g^* is twice continuously differentiable on $[0, 1]$, f^* is bounded away from zero and $n^{-8/21} \ll \lambda_n \ll 1$, then*

$$\|\hat{g} - g^*\|_n = O_p\big(r_n\big), \qquad \|\hat{g}' - g^{*\prime}\|_n = O_p\big(\sqrt{r_n s_n}\big).$$

The above bounds also hold for the $\|\cdot\|$ norm.

Remark 3 *The assumption $n^{-8/21} \ll \lambda_n \ll 1$ implies that the error bounds for $\hat{g}$ and $\hat{g}'$ are of order $o_p(1)$.*

When $\lambda_n \asymp n^{-2/7}$, Theorem 4 yields an $n^{-2/7}$ rate of convergence for $\hat{g}$. This rate matches the optimal rate of convergence for estimating the derivative of a density under the corresponding smoothness conditions (Stone, 1980).

Given a value $\tilde{p}$ in the interval $(0, 1)$, we define the ECAP estimator, $\hat{p} = \hat{p}(\tilde{p})$, by replacing $\tilde{p}_i$, γ^*, and g with $\tilde{p}$,$\hat{\gamma}$ and $\hat{g}$, respectively, in the expression for the oracle estimator provided by formulas (8), (9) and (10). Thus, we treat $\hat{p}$ as a random function of $\tilde{p}$, where the randomness comes from the fact that $\hat{p}$ depends on the training sample of the observed probabilities $\tilde{p}_i$. By analogy, we define p_0 via (10), with $\tilde{p}_i$ replaced by $\tilde{p}$, and view p_0 as a (deterministic) function of $\tilde{p}$.

We define the function $W_0(\tilde{p})$ as the expected loss for the oracle estimator:

$$W_0(\tilde{p}) = E_p\big[EC\left(p_0(\tilde{p})\right)^2 | \tilde{p}\big],$$

where the expected value is taken over the true p given the corresponding observed probability $\tilde{p}$. Similarly, we define the random function $\widehat{W}(\tilde{p})$ as the expected loss for the ECAP estimator,

$$\widehat{W}(\tilde{p}) = E_p\big[EC\left(\hat{p}(\tilde{p})\right)^2 | \tilde{p}\big],$$

where the expected value is computed given the training sample $\tilde{p}_1, ..., \tilde{p}_n$ and is again taken over the true p conditional on the corresponding $\tilde{p}$. The randomness in the function $\widehat{W}(\tilde{p})$ is due to the dependence of $\hat{p}$ on the training sample.

To state the asymptotic results for $\hat{p}$ and $\hat{W}$, we implement a minor technical modification in the estimation of the conditional variance via formula (9). After computing the value of $\hat{\sigma}^2$, we set it equal to $\max\{\hat{\sigma}^2, c\sqrt{r_n s_n}\}$, where c is allowed to be any fixed positive constant. This ensures that, as the sample size grows, $\hat{\sigma}^2$ does not approach zero too fast. We note that this technical modification is only used to establish consistency of $\widehat{W}(\tilde{p})$ in the next theorem; all the other results in this section hold both with and without this modification.

Theorem 5 *If g^* is twice continuously differentiable on $[0,1]$, f^* is bounded away from zero, $n^{-8/21} \ll \lambda_n \ll 1$ and $|\hat{\gamma} - \gamma^*| = o_p(1)$, then*

$$\|\hat{p} - p_0\| = o_p(1) \qquad \text{and} \qquad \|\hat{p} - p_0\|_n = o_p(1).$$

If, in addition, $|\hat{\gamma} - \gamma^*| = O_p(\sqrt{r_n s_n})$, then

$$\int_0^1 \big|\widehat{W}(\tilde{p}) - W_0(\tilde{p})\big| f^*(\tilde{p}) d\tilde{p} = o_p(1) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} \big|\widehat{W}(\tilde{p}_i) - W_0(\tilde{p}_i)\big| = o_p(1).$$

The next result provides the rates of convergence for $\hat{p}$ and $\hat{W}$.

Theorem 6 *If g^* is twice continuously differentiable on $[0,1]$, f^* is bounded away from zero, $n^{-8/21} \ll \lambda_n \ll 1$ and $|\hat{\gamma} - \gamma^*| = O_p\big(\sqrt{r_n s_n}\big)$, then*

$$\int_\epsilon^{1-\epsilon} \big|\hat{p}(\tilde{p}) - p_0(\tilde{p})\big|^2 f^*(\tilde{p}) d\tilde{p} = O_p\big(r_n s_n\big), \qquad \int_\epsilon^{1-\epsilon} \big|\widehat{W}(\tilde{p}) - W_0(\tilde{p})\big| f^*(\tilde{p}) d\tilde{p} = O_p\big(r_n s_n\big),$$

$$\sum_{i:\, \epsilon \le \tilde{p}_i \le 1-\epsilon} \frac{1}{n}\big|\hat{p}(\tilde{p}_i) - p_0(\tilde{p}_i)\big|^2 = O_p\big(r_n s_n\big) \qquad \text{and} \qquad \sum_{i:\, \epsilon \le \tilde{p}_i \le 1-\epsilon} \frac{1}{n}\big|\widehat{W}(\tilde{p}_i) - W_0(\tilde{p}_i)\big| = O_p\big(r_n s_n\big),$$

for each fixed positive ϵ.

Remark 4 *The assumption $n^{-8/21} \ll \lambda_n \ll 1$ ensures that all the error bounds are of order $o_p(1)$.*

In Theorem 6 we bound the integration limits away from zero and one, because the rate of convergence changes as $\tilde{p}$ approaches those values. However, we note that ϵ can be set to an arbitrarily small value. The optimal rate of convergence for $\widehat{W}$ is provided in the following result.

Corollary 2 *Suppose that λ_n decreases at the rate $n^{-2/7}$ and $|\hat{\gamma} - \gamma^*| = O_p(n^{-1/7})$. If f^* is bounded away from zero and g^* is twice continuously differentiable on $[0,1]$, then*

$$\int_\epsilon^{1-\epsilon} \big|\widehat{W}(\tilde{p}) - W_0(\tilde{p})\big| d\tilde{p} = O_p\big(n^{-2/7}\big) \qquad \text{and} \qquad \sum_{i:\, \epsilon \le \tilde{p}_i \le 1-\epsilon} \frac{1}{n}\big|\widehat{W}(\tilde{p}_i) - W_0(\tilde{p}_i)\big| = O_p\big(n^{-2/7}\big),$$

for every positive ϵ.

Corollary 2 follows directly from Theorem 6 by balancing out the components in the expression for r_n.

3 ECAP Extensions

In this section we consider two possible extensions of (1), the model for $\tilde{p}_i$. In particular, in the next subsection we discuss the setting where $\tilde{p}_i$ can no longer be considered an unbiased estimator for p_i, while in the following subsection we suggest a generalization of the beta model.

3.1 Incorporating Bias in $\tilde{p}_i$

So far, we have assumed that $\tilde{p}_i$ is an unbiased estimate for p_i. In practice probability estimates $\tilde{p}_i$ may exhibit some systematic bias. For example, in Section 5 we examine probability predictions from the FiveThirtyEight.com website on congressional house, senate, and governors races during the 2018 US midterm election. After comparing the actual election results with the predicted probability of a candidate being elected, there is clear evidence of bias in the estimates (Silver, 2018). In particular the leading candidate won many more races than would be suggested by the probability estimates. This indicates that the FiveThirtyEight.com probabilities were overly conservative, i.e., that in comparison to p_i the estimate $\tilde{p}_i$ was generally closer to 0.5; for example, $E(\tilde{p}_i|p_i) < p_i$ when $p_i > 0.5$.

In this section we generalize (1) to model situations where $E(\tilde{p}_i|p_i) \neq p_i$. To achieve this goal we replace (1) with

$$\tilde{p}_i|p_i \sim Beta(\alpha_i, \beta_i), \quad \text{where} \quad p_i = h_\theta(\alpha_i\gamma^*) = h_\theta(1 - \beta_i\gamma^*), \tag{21}$$

$h_\theta(\cdot)$ is a prespecified function, and θ is a parameter which determines the level of bias of $\tilde{p}_i$. In particular, (21) implies that for any invertible h_θ,

$$p_i = h_\theta(E(\tilde{p}_i|p_i)), \tag{22}$$

so that if $h_\theta(x) = x$, i.e., $h_\theta(\cdot)$ is the identity function, then (21) reduces to (1), and $\tilde{p}_i$ is an unbiased estimate for $\tilde{p}_i$.

To produce a valid probability model $h_\theta(\cdot)$ needs to satisfy several criteria:

1. $h_0(x) = x$, so that (21) reduces to (1) when $\theta = 0$.

2. $h_\theta(1 - x) = 1 - h_\theta(x)$, ensuring that the probabilities of events A_i and A_i^c sum to 1.

3. $h_\theta(x) = x$ for $x = 0, x = 0.5$ and $x = 1$.

4. $h_\theta(\alpha)$ is invertible for values of θ in a region around zero, so that $E(\tilde{p}_i|p_i)$ is unique.

The simplest polynomial function that satisfies all these constraints is

$$h_\theta(x) = (1 - 0.5\theta)x - \theta[x^3 - 1.5x^2],$$

which is invertible for $-4 \leq \theta \leq 2$. Note that for $\theta = 0$, we have $h_0(x) = x$, which corresponds to the unbiased model (1). However, if $\theta > 0$, then $\tilde{p}_i$ tends to overestimate small p_i and underestimate large p_i, so the probability estimates are overly conservative. Alternatively, when $\theta < 0$, then $\tilde{p}_i$ tends to underestimate small p_i and overestimate large p_i, so the probability estimates exhibit excess certainty. Figure 4 provides examples of $E(\tilde{p}_i|p_i)$ for three different values of θ, with the

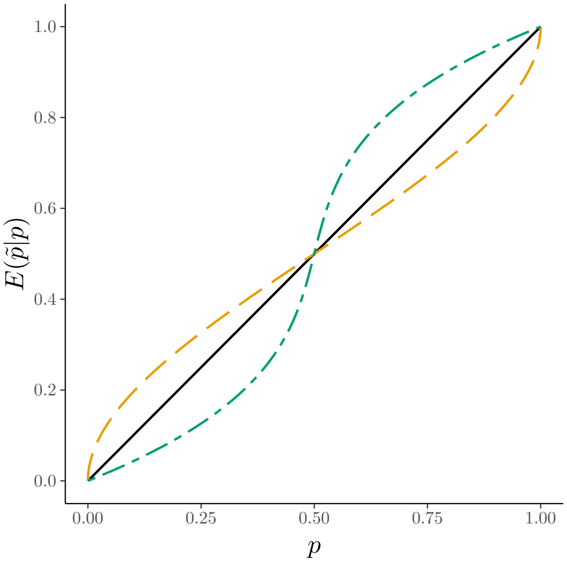


Figure 4: Plots of $E(\tilde{p}_i|p_i)$ as a function of p_i for different values of θ. When $\theta = 0$ (black / solid) the estimates are unbiased. $\theta = 2$ (orange / dashed) corresponds to a setting where $\tilde{p}_i$ systematically underestimates large values of p_i, while $\theta = -3$ (green / dot-dashed) represents a situation where $\tilde{p}_i$ is an overestimate for large values of p_i.

green line representing probabilities resulting in excess certainty, the orange line overly conservative probabilities, and the black line unbiased probabilities.

One of the appealing aspects of this model is that the ECAP oracle (10) can still be used to generate an estimator for p_i. The only change is in how $E(p_i|\tilde{p}_i)$ and $Var(p_i|\tilde{p}_i)$ are computed. The following result allows us to generalize Theorem 2 to the biased setting to compute $E(p_i|\tilde{p}_i)$ and $Var(p_i|\tilde{p}_i)$.

Theorem 7 *Suppose that model* (21) *holds, p_i has a bounded density, and μ_i and σ_i^2 are respectively defined as in* (8) *and* (9)*. Then,*

$$\begin{aligned} E(p_i|\tilde{p}_i) &= \mu_i + 0.5\theta\left[3\sigma_i^2 - 6\mu_i\sigma_i^2 + 3\mu_i^2 - \mu_i - 2\mu_i^3\right] + O\big(\theta\gamma^{*3/2}\big) && (23)\\ Var(p_i|\tilde{p}_i) &= (1-0.5\theta)^2\sigma_i^2 + \theta\sigma_i^2\big[3\mu_i(1-\mu_i)(3\theta\mu_i(1-\mu_i) - 0.5\theta + 1)\big] + O\big(\theta\gamma^{*3/2}\big). && (24)\end{aligned}$$

The remainder terms in the above approximations are of smaller order than the leading terms when γ^* is small, which is typically the case in practice. As we demonstrate in the proof of Theorem 7, explicit expressions can be provided for the remainder terms. However, the approximation error involved in estimating these expressions is likely to be much higher than any bias from excluding them. Hence, we ignore these terms when estimating $E(p_i|\tilde{p}_i)$ and $Var(p_i|\tilde{p}_i)$:

$$\begin{aligned} \widehat{E(p_i|\tilde{p}_i)} &= \hat{\mu}_i + 0.5\theta\left[3\hat{\sigma}_i^2 - 6\hat{\mu}_i\hat{\sigma}_i^2 + 3\hat{\mu}_i^2 - \hat{\mu}_i - 2\hat{\mu}_i^3\right] && (25)\\ \widehat{Var(p_i|\tilde{p}_i)} &= (1-0.5\theta)^2\hat{\sigma}_i^2 + \theta\hat{\sigma}_i^2\big[3\hat{\mu}_i(1-\hat{\mu}_i)(3\theta\hat{\mu}_i(1-\hat{\mu}_i) - 0.5\theta + 1)\big]. && (26)\end{aligned}$$

The only remaining issue in implementing this approach involves producing an estimate for θ. However, this can be achieved using exactly the same maximum likelihood approach as the one used to estimate γ^*, which is described in Section 2.3.2. Thus, we now choose both θ and γ to jointly maximize the likelihood function

$$l_{\theta,\gamma} = \sum_{i:Z_i=1} \log(\hat{\mu}_i^{\theta,\gamma}) + \sum_{i:Z_i=0} \log(1-\hat{\mu}_i^{\theta,\gamma}), \tag{27}$$

where $\hat{\mu}_i^{\theta,\gamma}$ is the bias corrected ECAP estimate of $E(p_i|\tilde{p}_i)$ from (25), generated by substituting in particular values of γ and θ. In all other respects, the bias corrected version of ECAP is implemented in an identical fashion to the unbiased version.

3.2 Mixture Distribution

We now consider another possible extension of (1), where we believe that $\tilde{p}_i$ is an unbiased estimator for p_i but find the beta model assumption to be unrealistic. In this setting one could potentially model $\tilde{p}_i$ using a variety of members of the exponential family. However, one appealing alternative is to extend (1) to a mixture of beta distributions:

$$\tilde{p}_i|p_i \sim \sum_{k=1}^{K} w_k Beta(\alpha_{ik}, \beta_{ik}), \quad \text{where} \quad \alpha_{ik} = \frac{c_k p_i}{\gamma^*}, \quad \beta_{ik} = \frac{1-c_k p_i}{\gamma^*}, \tag{28}$$

and w_k and c_k are predefined weights such that $\sum_k w_k = 1$ and $\sum_k w_k c_k = 1$. Note that (1) is a special case of (28) with $K = w_1 = c_1 = 1$.

As K grows, the mixture model can provide as flexible a model as desired, but it also has a number of other appealing characteristics. In particular, under this model it is still the case that $E(\tilde{p}_i|p_i) = p_i$. In addition, Theorem 8 demonstrates that simple closed form solutions still exist for $E(p_i|\tilde{p}_i)$ and $Var(p_i|\tilde{p}_i)$, and, hence, also the oracle ECAP estimator p_{i0}.

Theorem 8 *Under* (28),

$$E(p_i|\tilde{p}_i) = \mu_i \sum_{k=1}^{K} \frac{w_k}{c_k} \tag{29}$$

$$Var(p_i|\tilde{p}_i) = (\sigma_i^2 + \mu_i^2) \sum_{k=1}^{K} \frac{w_k}{c_k^2} - \mu_i^2 \left(\sum_{k=1}^{K} \frac{w_k}{c_k} \right)^2, \tag{30}$$

where μ_i and σ_i^2 are defined in (8) *and* (9).

The generalized ECAP estimator can thus be generated by substituting $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, given by formulas (17) and (18), into (29) and (30). The only additional complication involves computing values for w_k and c_k. For settings with a large enough sample size, this could be achieved using a variant of the maximum likelihood approach discussed in Section 2.3.2. However, we do not explore that approach further in this paper.

4 Simulation Results

In Section 4.1 we compare ECAP to competing methods under the assumption of unbiasedness in $\tilde{p}_i$. We further extend this comparison to the setting where $\tilde{p}_i$ represents a potentially biased estimate in Section 4.2.

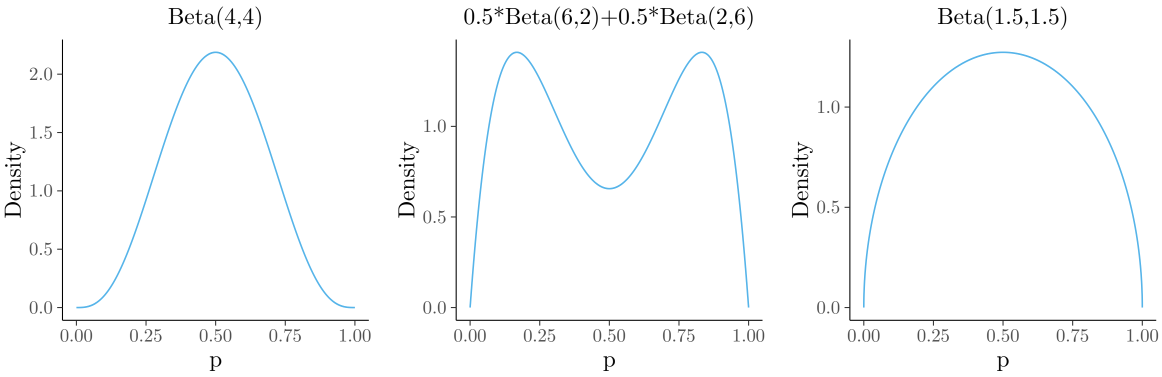


Figure 5: Distributions of p used in the simulation

4.1 Unbiased Simulation Results

In this section our data consists of $n = 1{,}000$ triplets $(p_i, \tilde{p}_i, Z_i)$ for each simulation. The p_i are generated from one of three possible prior distributions; $\text{Beta}(4,4)$, an equal mixture of $\text{Beta}(6,2)$ and $\text{Beta}(2,6)$, or $\text{Beta}(1.5,1.5)$. The corresponding density functions are displayed in Figure 5.

Recall that ECAP models $\tilde{p}_i$ as coming from a beta distribution, conditional on p_i. However, in practice there is no guarantee that the observed data will exactly follow this distribution. Hence, we generate the observed data according to:

$$\tilde{p}_i = p_i + p_i^q(\tilde{p}_i^{\text{o}} - p_i), \tag{31}$$

where $\tilde{p}_i^{\text{o}} | p_i \sim \text{Beta}(\alpha, \beta)$ and q is a tuning parameter. In particular for $q = 0$ (31) generates observations directly from the ECAP model, while larger values of q provide a greater deviation from the beta assumption. In practice we found that setting $q = 0$ can result in $\tilde{p}$'s that are so small they are effectively zero ($\tilde{p}_i = 10^{-20}$, for example). ECAP is not significantly impacted by these probabilities but, as we show, other approaches can perform extremely poorly in this scenario. Setting $q > 0$ prevents pathologic scenarios and allows us to more closely mimic what practitioners will see in real life. We found that $q = 0.05$ typically gives a reasonable amount of dispersion so we consider settings where either $q = 0$ or $q = 0.05$. We also consider different levels of the conditional variance for $\tilde{p}_i$, by taking γ^* as either 0.005 or 0.03. Finally, we generate Z_i, representing whether event A_i occurs, from a Bernoulli distribution with probability p_i.

We implement the following five approaches: the *Unadjusted* method, which simply uses the original probability estimates $\tilde{p}_i$, two implementations of the proposed *ECAP* approach (ECAP Opt and ECAP MLE), and two versions of the James Stein approach (JS Opt and JS MLE). For the proposed ECAP methods, we select λ via the cross-validation procedure in Section 2.3.2. ECAP Opt is an oracle-type implementation of the ECAP methodology, in which we select γ to minimize the average expected loss, defined in (4), over the training data. Alternatively, ECAP MLE makes use of the Z_i's and estimates γ^* using the maximum likelihood approach described in Section 2.3.2. The James-Stein method we use is similar to its traditional formulation. In particular the estimated probability is computed using

$$\hat{p}_i^{JS} = \bar{\tilde{p}} + (1 - c)\left(\tilde{p}_i - \bar{\tilde{p}}\right), \tag{32}$$

Table 1: Average expected loss for different methods over multiple unbiased simulation scenarios. Standard errors are provided in parentheses.

γ^*	q	Method Type	Beta(4, 4)	0.5*Beta(6,2) + 0.5*Beta(2,6)	Beta(1.5, 1.5)
0.005	0	Unadjusted	0.0116 (0.0001)	44.9824 (43.7241)	3.9×10^{12} (3.9×10^{12})
		ECAP Opt	0.0095 (0.0001)	0.0236 (0.0002)	0.0197 (0.0001)
		JS Opt	0.0100 (0.0001)	0.0241 (0.0002)	0.0204 (0.0002)
		ECAP MLE	0.0120 (0.0004)	0.0326 (0.0008)	0.0294 (0.0009)
		JS MLE	0.0121 (0.0003)	1.1590 (0.8569)	4.8941 (4.7526)
	0.05	Unadjusted	0.0100 (0.0001)	0.0308 (0.0006)	0.0273 (0.0006)
		ECAP Opt	0.0085 (0.0000)	0.0196 (0.0001)	0.0166 (0.0001)
		JS Opt	0.0090 (0.0000)	0.0201 (0.0001)	0.0172 (0.0001)
		ECAP MLE	0.0022 (0.0005)	0.0073 (0.0010)	0.0084 (0.0011)
		JS MLE	0.0105 (0.0002)	0.0265 (0.0006)	0.0245 (0.0007)
0.03	0	Unadjusted	2.1×10^{8} (2.1×10^{8})	2.4×10^{14} (1.6×10^{14})	1.6×10^{15} (5.5×10^{14})
		ECAP Opt	0.0391 (0.0002)	0.0854 (0.0004)	0.0740 (0.0004)
		JS Opt	0.0537 (0.0002)	0.0986 (0.0005)	0.0899 (0.0005)
		ECAP MLE	0.0435 (0.0009)	0.1202 (0.0136)	0.1203 (0.0152)
		JS MLE	0.0636 (0.0019)	1.4×10^{13} (1.4×10^{13})	1.2×10^{14} (1.1×10^{14})
	0.05	Unadjusted	0.0887 (0.0010)	0.3373 (0.0047)	0.2780 (0.0043)
		ECAP Opt	0.0364 (0.0002)	0.0765 (0.0004)	0.0665 (0.0004)
		JS Opt	0.0488 (0.0002)	0.0874 (0.0005)	0.0801 (0.0005)
		ECAP MLE	0.0022 (0.0004)	0.0075 (0.0010)	0.0078 (0.0014)
		JS MLE	0.0558 (0.0011)	0.1213 (0.0066)	0.1235 (0.0071)

where $\bar{\tilde{p}} = \frac{1}{n}\sum_{j=1}^{n}\tilde{p}_j$ and c is a tuning parameter chosen to optimize the estimates.[2] Equation (32) is a convex combination of $\tilde{p}_i$ and the average observed probability $\bar{\tilde{p}}$. The JS Opt implementation selects c to minimize the average expected loss in the same fashion as for ECAP Opt, while the JS MLE implementation selects c using the maximum likelihood approach described in Section 2.3.2. Note that ECAP Opt and JS Opt represent optimal situations that can not be implemented in practice because they require knowledge of the true distribution of p_i.

In each simulation run we generate both training and test data sets. Each method is fit on the training data. We then calculate $EC(\hat{p}_i)^2$ for each point in the test data and average over these observations. The results for the three prior distributions, two values of γ^*, and two values of q, averaged over 100 simulation runs, are reported in Table 1. Since the ECAP Opt and JS Opt approaches both represent oracle type methods, they should be compared with each other. The ECAP Opt method statistically significantly outperforms its JS counterpart in each of the twelve settings, with larger improvements in the noisy setting where $\gamma^* = 0.03$. The ECAP MLE method is statistically significantly better than the corresponding JS approach in all but five settings. However, four of these settings, correspond to $q = 0$ and actually represent situations where JS MLE has failed because it has extremely large excess certainty, which impacts both the mean and standard error. Alternatively, the performance of the ECAP approach remains stable even in the presence of extreme outliers. Similarly, the ECAP MLE approach statistically significantly

[2] To maintain consistency with ECAP we flip all $\tilde{p}_i > 0.5$ across 0.5 before forming $\hat{p}_i^{JS}$ and then flip the estimate back.

Table 2: Average expected loss for different methods over multiple biased simulation scenarios.

	Method Type	Beta(4,4)	0.5*Beta(6,2) + 0.5*Beta(2,6)	Beta(1.5, 1.5)
$\theta = -3$	Unadjusted	0.1749 (0.0005)	0.7837 (0.0025)	0.6052 (0.0030)
	ECAP Opt	0.0019 (0.0000)	0.0109 (0.0000)	0.0086 (0.0000)
	JS Opt	0.0609 (0.0002)	0.2431 (0.0005)	0.1526 (0.0003)
	ECAP MLE	0.0036 (0.0002)	0.0135 (0.0003)	0.0111 (0.0003)
	JS MLE	0.0633 (0.0003)	0.2712 (0.0014)	0.1707 (0.0011)
$\theta = -1$	Unadjusted	0.0319 (0.0001)	0.1389 (0.0007)	0.1130 (0.0008)
	ECAP Opt	0.0051 (0.0000)	0.0150 (0.0000)	0.0124 (0.0001)
	JS Opt	0.0142 (0.0000)	0.0477 (0.0001)	0.0361 (0.0001)
	ECAP MLE	0.0065 (0.0002)	0.0176 (0.0002)	0.0158 (0.0004)
	JS MLE	0.0155 (0.0002)	0.0541 (0.0008)	0.0413 (0.0010)
$\theta = 0$	Unadjusted	0.0099 (0.0000)	0.0305 (0.0002)	0.0275 (0.0003)
	ECAP Opt	0.0084 (0.0000)	0.0195 (0.0001)	0.0164 (0.0001)
	JS Opt	0.0088 (0.0000)	0.0199 (0.0001)	0.0171 (0.0001)
	ECAP MLE	0.0102 (0.0003)	0.0229 (0.0003)	0.0206 (0.0005)
	JS MLE	0.0094 (0.0001)	0.0233 (0.0005)	0.0219 (0.0005)
$\theta = 2$	Unadjusted	0.0652 (0.0001)	0.2419 (0.0003)	0.1776 (0.0003)
	ECAP Opt	0.0240 (0.0001)	0.0614 (0.0002)	0.0502 (0.0001)
	JS Opt	0.0652 (0.0001)	0.2419 (0.0003)	0.1776 (0.0003)
	ECAP MLE	0.0256 (0.0002)	0.0739 (0.0013)	0.0591 (0.0009)
	JS MLE	0.0652 (0.0001)	0.2419 (0.0003)	0.1776 (0.0003)

outperforms the Unadjusted approach, often by large amounts, except for the five settings with large outliers, which result in extremely bad average performance for the latter method.

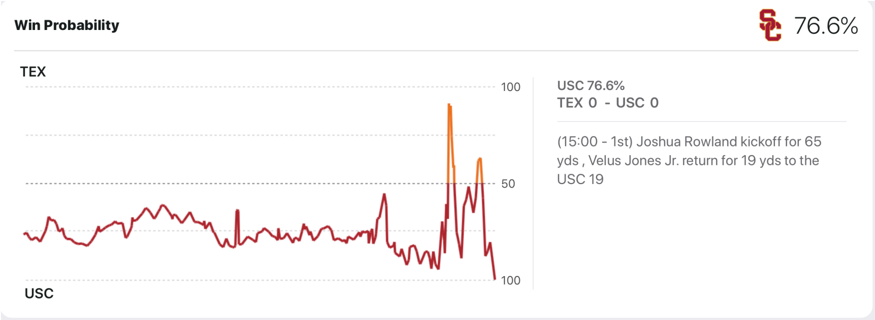
4.2 Biased Simulation

In this section we extend the results to the setting where the observed probabilities may be biased, i.e., $E(\tilde{p}_i|p_i) \neq p_i$. To do this we generate $\tilde{p}_i$ according to (21) using four different values for θ, $\{-3, -1, 0, 2\}$. Recall that $\theta < 0$ corresponds to anti-conservative data, where $\tilde{p}_i$ tends to be too close to 0 or 1, $\theta = 0$ represents unbiased observations, and $\theta > 0$ corresponds to conservative data, where $\tilde{p}_i$ tends to be too far from 0 or 1. In all other respects our data is generated in an identical fashion to that of the unbiased setting.[3]

To illustrate the biased setting we opted to focus on the $q = 0.05$ with $\gamma^* = 0.005$ setting. We also increased the sample size to $n = 5{,}000$ because of the increased difficulty of the problem. The two ECAP implementations now require us to estimate three parameters: λ, γ and θ. We estimate λ in the same fashion as previously discussed, while γ and θ are now chosen over a two-dimensional grid of values, with θ restricted to lie between -4 and 2. The two JS methods remain unchanged.

The results, again averaged over 100 simulation runs, are presented in Table 2. In the two settings where $\theta < 0$ we note that the unadjusted and JS methods all exhibit significant deterioration in their performance relative to the unbiased $\theta = 0$ scenario. By comparison, the two ECAP methods significantly outperform the JS and unadjusted approaches. A similar pattern is observed for $\theta > 0$. In this setting all five methods deteriorate, but ECAP is far more robust to the biased setting than unadjusted and JS.

[3] Because the observed probabilities are now biased, we replace p_i in (31) with $E(\tilde{p}_i|p_i)$.



USC 76.6%
TEX 0 - USC 0

(15:00 - 1st) Joshua Rowland kickoff for 65 yds , Velus Jones Jr. return for 19 yds to the USC 19

Figure 6: A screenshot of the NCAA football win probabilities publicly available on ESPN's website. USC vs. Texas (2017)

It is perhaps not surprising that the bias corrected version of ECAP outperforms the other methods when the data is indeed biased. However, just as interestingly, even in the unbiased setting ($\theta = 0$) we still observe that ECAP matches or slightly outperforms its JS counterpart, despite the fact that ECAP must estimate θ. This is likely a result of the fact that ECAP is able to accurately estimate θ. Over all simulation runs and settings, ECAP Opt and ECAP MLE respectively averaged absolute errors of only 0.0582 and 0.2016 in estimating θ.

5 Empirical Results

In this section we illustrate ECAP on two real world data sets. Section 5.1 contains our results analyzing ESPN's probability estimates from NCAA football games, while Section 5.2 examines probability estimates from the 2018 US midterm elections. Given that for real data p_i is never observed, we need to compute an estimate of $EC(\hat{p}_i)$. Hence, we choose a small window δ, for example $\delta = [0, 0.02]$, and consider all observations for which $\tilde{p}_i$ falls within δ.[4] We then estimate p_i via $\bar{p}_\delta = \frac{1}{n_\delta}\sum_{i=1}^n Z_i\delta_i$, where $\delta_i = I(\tilde{p}_i \in \delta)$, $n_\delta = \sum_{i=1}^n \delta_i$ and Z_i is defined as in (19). Hence we can estimate EC using

$$\widehat{EC}_\delta(\bar{\hat{p}}_\delta) = \frac{\bar{p}_\delta - \bar{\hat{p}}_\delta}{\bar{\hat{p}}_\delta}, \tag{33}$$

where $\bar{\hat{p}}_\delta = \frac{1}{n_\delta}\sum_{i=1}^n \hat{p}_i\delta_i$.

5.1 ESPN NCAA Football Data

Each year there are approximately 1,200 Division 1 NCAA football games played within the US. For the last several seasons ESPN has been producing automatic win probability estimates for every game. These probabilities update in real time after every play. Figure 6 provides an example of a fully realized game between the University of Southern California (USC) and the University of Texas at Austin (TEX) during the 2017 season. For most of the game the probability of a USC win hovers around 75% but towards the end of the game the probability starts to oscillate wildly, with both teams having high win probabilities, before USC ultimately wins.[5] These gyrations are

[4] In this section, for simplicity of notation, we have flipped all probabilities greater than 0.5, and the associated Z_i around 0.5 so $\delta = [0, 0.02]$ also includes probabilities between 0.98 and 1.

[5] The game was not chosen at random.

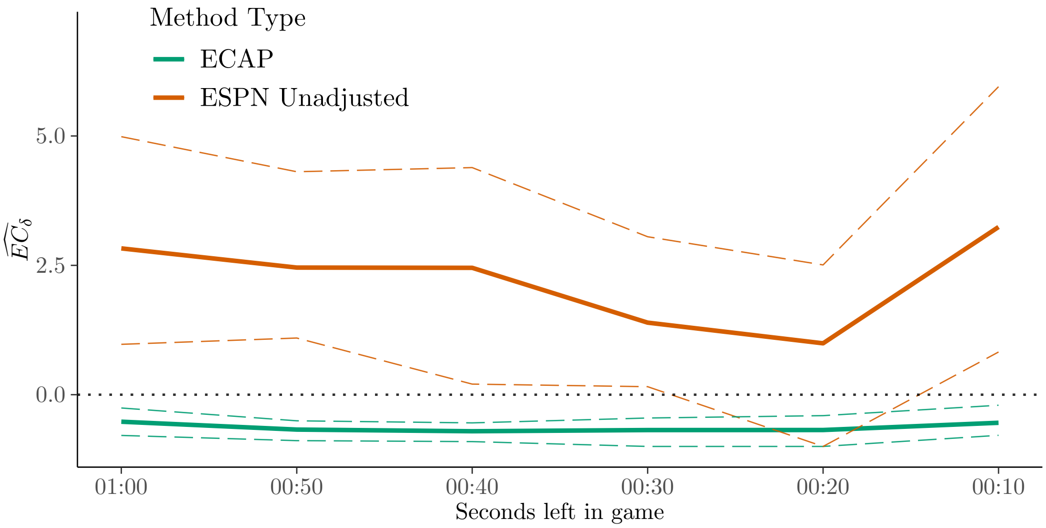


Figure 7: Empirical EC in both the unadjusted and ECAP setting with $\delta = [0, 0.02]$.

quite common and occasionally result in a team with a high win probability ultimately losing. Of course even a team with a 99% win probability will end up losing 1% of the time so these unusual outcomes do not necessarily indicate an error, or selection bias issue, with the probability estimates.

To assess the accuracy of ESPN's estimation procedure we collected data from the 2016 and 2017 NCAA football seasons. We obtained this unique data set by scrapping the win probabilities, and ultimate winning team, for a total of 1,722 games (about 860 per season), involving an average of approximately 180 probabilities per game. Each game runs for 60 minutes, although the clock is often stopped. For any particular time point t during these 60 minutes, we took the probability estimate closest to t in each of the individual games. We used the entire data set, 2016 and 2017, to compute $\bar{p}_\delta$, which represents the ideal gold standard. However, this estimator is impractical in practice because we would need to collect data over two full years to implement it. By comparison, we used only the 2016 season to fit ECAP and ultimately to compute $\bar{\hat{p}}_\delta$. We then calculated $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ for both the raw ESPN probabilities and the adjusted ECAP estimates. The intuition here is that $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ provides a comparison of these estimates to the ideal, but unrealistic, $\bar{p}_\delta$.

In general we found that $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ computed on the ESPN probabilities was not systematically different from zero, suggesting ESPN's probabilities were reasonably accurate. However, we observed that, for extreme values of δ, $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ was well above zero towards the end of the games. Consider, for example, the solid orange line in Figure 7, which plots $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ using $\delta = [0, 0.02]$ at six different time points during the final minute of these games. We observe that excess certainty is consistently well above zero. The 90% bootstrap confidence intervals (dashed lines), generated by sampling with replacement from the probabilities that landed inside δ_i, demonstrate that the difference from zero is statistically significant for most time points. This suggests that towards the end of the game ESPN's probabilities are too extreme i.e. there are more upsets then would be predicted by their estimates.

Next we applied the unbiased implementation of ECAP, i.e. with $\theta = 0$, separately to each of these six time points and computed $\widehat{EC}_\delta(t)$ for the associated ECAP probability estimates. To estimate the out of sample performance of our method, we randomly picked half of the 2016 games

Table 3: Bias corrected ECAP adjustment of FiveThirtyEight's 2018 election probabilities. Reported average $\widehat{EC}_\delta$.

Method	Adjustment	δ_1	δ_2
Classic	Unadjusted	-0.6910	-0.8361
	ECAP	-0.2881	-0.0758
Deluxe	Unadjusted	-0.4276	-0.8137
	ECAP	-0.0371	0.1814
Lite	Unadjusted	-0.8037	-0.8302
	ECAP	-0.3876	-0.1118

to estimate γ^*, and then used ECAP to produce probability estimates on the other half. We repeated this process 100 times and averaged the resulting $\widehat{EC}_\delta(\bar{\hat{p}}_\delta, t)$ independently for each time point. The solid green line in Figure 7 provides the estimated excess certainty. ECAP appears to work well on this data, with excess certainty estimates close to zero. Notice also that ECAP is consistently producing a slightly negative excess certainty, which is actually necessary to minimize the expected loss function (4), as demonstrated in Figure 3. Interestingly this excess certainty pattern in the ESPN probabilities is no longer apparent in data for the 2018 season, suggesting that ESPN also identified this as an issue and applied a correction to their estimation procedure.

5.2 Election Data

Probabilities have increasingly been used to predict election results. For example, news organizations, political campaigns, and others, often attempt to predict the probability of a given candidate winning a governors race, or a seat in the house, or senate. Among other uses, political parties can use these estimates to optimize their funding allocations across hundreds of different races. In this section we illustrate ECAP using probability estimates produced by the FiveThirtyEight.com website during the 2018 US midterm election cycle. FiveThrityEight used three different methods, *Classic, Deluxe*, and *Lite*, to generate probability estimates for every governor, house, and senate seat up for election, resulting in 506 probability estimates for each of the three methods.

Interestingly a previous analysis of this data (Silver, 2018) showed that the FiveThirtyEight probability estimates appeared to be overly conservative i.e. the leading candidate won more often than would have been predicted by their probabilities. Hence, we should be able to improve the probability estimates using the bias corrected version of ECAP from Section 3.1. We first computed $\widehat{EC}_\delta(\bar{\hat{p}}_\delta)$ on the unadjusted FiveThirtyEight probability estimates using two different values for δ i.e. $\delta_1 = [0, 0.1]$ and $\delta_2 = [0.1, 0.2]$. We used wider windows for δ in comparison to the ESPN data because we only had one third as many observations. The results for the three methods used by FiveThirtyEight are shown in Table 3. Notice that for all three methods and both values of δ the unadjusted estimates are far below zero and several are close to -1, the minimum possible value. These results validate the previous analysis suggesting the FiveThirtyEight estimates are systematically conservatively biased.

Next we applied ECAP separately to each of the three sets of probability estimates, with the value of θ chosen using the MLE approach previously described. Again the results are provided in Table 3. ECAP appears to have significantly reduced the level of bias, with most values of $\widehat{EC}_\delta(\bar{\hat{p}}_\delta)$ close to zero, and in one case actually slightly above zero. For the Deluxe method with δ_1, ECAP

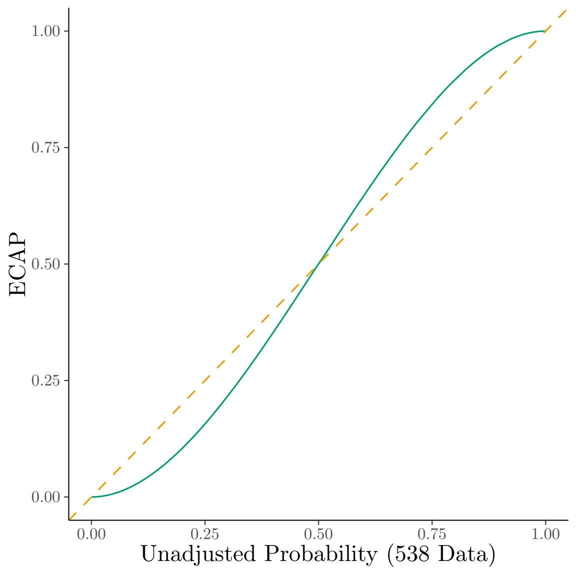


Figure 8: ECAP bias corrected probabilities vs original FiveThirtyEight probability from classic method.

has an almost perfect level of excess certainty. For the Classic and Lite methods, $\theta = 2$ was chosen by ECAP for both values of δ, representing the largest possible level of bias correction. For the Deluxe method, ECAP selected $\theta = 1.9$. Figure 8 demonstrates the significant level of correction that ECAP applies to the classic method FiveThirtyEight estimates. For example, ECAP adjusts probability estimates of 0.8 to 0.89 and estimates of 0.9 to 0.97.

6 Discussion

In this article, we have convincingly demonstrated both theoretically and empirically that probability estimates are subject to selection bias, even when the individual estimates are unbiased. Our proposed ECAP method applies a novel non-parametric empirical Bayes approach to adjust both biased and unbiased probabilities, and hence produce more accurate estimates. The results in both the simulation study and on real data sets demonstrate that ECAP can successfully correct for selection bias, allowing us to use the probabilities with a higher level of confidence when selecting extreme values.

There are a number of possible areas for future work. For example, the ESPN data contains an interesting time series structure to the probabilities, with each game consisting of a probability function measured over 60 minutes. Our current method treats each time point independently and adjusts the probabilities accordingly. However, one may be able to leverage more power by incorporating all time points simultaneously using some form of functional data analysis. Another potential area of exploration involves the type of data on which ECAP is implemented. For example, consider a setting involving a large number of hypothesis tests and associated p-values, $\tilde{p}_1, \ldots, \tilde{p}_n$. There has been much discussion recently of the limitations around using p-values. A superior approach would involve thresholding based on the posterior probability of the null hypothesis being true i.e. $p_i = P(H_{0i}|X_i)$. Of course, in general, p_i is difficult to compute which is why we use the p-value $\tilde{p}_i$. However, if we were to treat $\tilde{p}_i$ as a, possibly biased, estimate of p_i, then it may be possible to use a modified version of ECAP to estimate p_i. If such an approach could be implemented it would likely have a significant impact in the area of multiple hypothesis testing.

Acknowledgements

We would like to thank the Editor, the Associate Editor, and two referees for their useful comments and suggestions.

References

Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist. 34*(2), 584–653.

Benjamini, Y. and D. Yekutieli (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association 100*(469), 71–81.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Ann. Statist. 36*(6), 2577–2604.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

Gelman, A. and C. R. Shalizi (2012). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Henderson, N. C. and M. A. Newton (2015). Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(4), 781–804.

Hull, J., M. Predescu, and A. White (2005). Bond prices, default probabilities and risk premiums. *Journal of Credit Risk 1*, 53–60.

Ikeda, Y., T. Kubokawa, and M. S. Srivastava (2016). Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions. *Computational Statistics and Data Analysis 95*, 95 – 108.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., pp. 361–379. University of California Press.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist. 37*(4), 1647–1684.

Kealhofer, S. (2003). Quantifying credit risk i: Default prediction. *Financial Analysts Journal 59*(1), 30–44.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist. 40*(2), 1024–1060.

Leung, C. K. and K. W. Joseph (2014). Sports data mining: Predicting results for the college football games. *Procedia Computer Science 35*, 710 – 719. Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.

Petrone, S., S. Rizzelli, J. Rousseau, and C. Scricciolo (2014). Empirical bayes methods in classical and bayesian inference. *METRON 72*(2), 201–215.

Poses, R. M., W. R. Smith, D. K. McClish, E. C. Huber, F. L. W. Clemo, B. P. Schmitt, D. Alexander-Forti, E. M. Racht, I. Colenda, Christopher C., and R. M. Centor (1997, 05). Physicians' Survival Predictions for Patients With Acute Congestive Heart Failure. *JAMA Internal Medicine 157*(9), 1001–1007.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Math. Statistic. and Prob. 1*, 157–163.

Silver, N. (2018, Dec). How fivethirtyeight's 2018 midterm forecasts did. https://fivethirtyeight.com/features/how-fivethirtyeights-2018-midterm-forecasts-did/. Online; accessed 04 September 2019.

Smeenk, R. M., V. J. Verwaal, N. Antonini, and F. A. N. Zoetmulder (2007). Survival analysis of pseudomyxoma peritonei patients treated by cytoreductive surgery and hyperthermic intraperitoneal chemotherapy. *Annals of Surgery 245(1)*, 104–109.

Soumbatiants, S., H. W. Chappell, and E. Johnson (2006, Apr). Using state polls to forecast u.s. presidential election outcomes. *Public Choice 127*(1), 207–223.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 1348–1360.