Statistics and Probability Letters 118 (2016) 8–15

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Demystifying the bias from selective inference: A revisit to Dawid's treatment selection problem

Jiannan Lu *, Alex Deng

Analysis and Experimentation, Microsoft Corporation, United States

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form 13 April 2016
Accepted 9 June 2016
Available online 16 June 2016

Keywords:
Bayesian inference
Posterior mean
Selection paradox
Multivariate truncated normal

ABSTRACT

We extend the heuristic discussion in Senn (2008) on the bias from selective inference for the treatment selection problem (Dawid, 1994), by deriving the closed-form expression for the selection bias. We illustrate the advantages of our theoretical results through numerical and simulated examples.

1. Introduction

Selective inference gained popularity in recent years (e.g., Lockhart et al., 2014; G'Sell et al., 2016; Reid and Tibshirani, 2016). To quote (Dawid, 1994), "...a great deal of statistical practice involves, explicitly or implicitly, a two stage analysis of the data. At the first stage, the data are used to identify a particular parameter on which attention is to focus; the second stage then attempts to make inferences about the selected parameter". Consequently, the results (e.g., point estimates, *p*-values) produced by selective inference are generally "cherry-picked" (Taylor and Tibshirani, 2015), and therefore it is of great importance for practitioners to conduct "exact post-selection inference" (e.g., Tibshirani et al., 2014; Lee et al., 2016).

To demonstrate the importance of "exact post-selection inference", in this paper we focus on the "bias" of the posterior mean associated with the most extreme observation (formally defined later, and henceforth referred to as "selection bias") in the treatment selection problem (Dawid, 1994), which is not only fundamental in theory, but also of great practical importance in, e.g., agricultural studies, clinical trials, and large-scale online experiments (Kohavi et al., 2013). In an illuminating paper, Senn (2008) provided a heuristic explanation that the existence of selection bias depended on the prior distribution, and upheld Dawid's claim that the fact that selection bias did not exist in some standard cases was a consequence of using certain conjugate priors. In this paper, we relax the modeling assumptions in Senn (2008) and derive the closed-form expression for the selection bias. Consequently, our work can serve as a complement of the heuristic explanation provided by Senn (2008), and is useful from both theoretical and practical perspectives.

The paper proceeds as follows. Section 2 reviews the treatment selection problem, defines the selection bias, and describes the Bayesian inference framework which the remaining parts of the paper are based on. Section 3 derives the closed-form expression for the selection bias. Section 4 highlights numerical and simulated examples that illustrates the advantages of our theoretical results. Section 5 concludes and discusses future directions.

* Correspondence to: Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA.
E-mail address: jiannl@microsoft.com (J. Lu).

http://dx.doi.org/10.1016/j.spl.2016.06.007

2. Bayesian inference for treatment selection problem

2.1. Treatment selection problem and selection bias

Consider an experiment with $p \geq 2$ treatment arms. For $i = 1, \ldots, p$, let μ_i denote the mean yield of the ith treatment arm. After running the experiment, we observe the sample mean yield of the ith treatment arm, denoted as X_i. Let

$$i^* = \underset{1 \leq i \leq p}{\operatorname{argmax}} X_i$$

denote the index of the largest observation. The focus of selective inference is on μ_{i^*}, which relies on $X_1, \ldots, X_p$. We let $\mathrm{E}(\mu_{i^*} \mid X_{i^*})$ be the posterior mean of μ_{i^*} as if it were selected before the experiment, and

$$\mathrm{E}(\mu_{i^*} \mid X_{i^*}, X_{i^*} = \max X_i)$$

be the "exact post-selection" posterior mean of μ_{i^*}, which takes the selection into account. Following Senn (2008), we define the selection bias as

$$\Delta = \mathrm{E}(\mu_{i^*} \mid X_{i^*}) - \mathrm{E}(\mu_{i^*} \mid X_{i^*}, X_{i^*} = \max X_i). \tag{1}$$

Having defined the selection bias, we briefly discuss the "selection paradox" in Dawid (1994), i.e., "since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data". In other words, if we define the selection bias as

$$\tilde{\Delta} = \mathrm{E}(\mu_{i^*} \mid X_1, \ldots, X_p) - \mathrm{E}(\mu_{i^*} \mid X_1, \ldots, X_p, X_{i^*} = \max X_i),$$

then indeed $\tilde{\Delta} = 0$.

2.2. The normal–normal model

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\boldsymbol{X} = (X_1, \ldots, X_p)'$. Following Dawid (1994), we treat them as random vectors. We generalize Senn (2008) and assume that

$$\boldsymbol{\mu} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{X} \mid \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

where

$$\boldsymbol{\Sigma}_0 = \gamma^2 \boldsymbol{I}_p + (1-\gamma^2)\mathbf{1}_p\mathbf{1}_p', \qquad \boldsymbol{\Sigma} = \sigma^2\{\eta^2\boldsymbol{I}_p + (1-\eta^2)\mathbf{1}_p\mathbf{1}_p'\}, \quad 0 \leq \gamma, \eta \leq 1. \tag{3}$$

To interpret (3) we let $X_i = \mu_i + \epsilon_i$, where μ_i is generated by

$$\phi \sim N\left(0, 1-\gamma^2\right), \qquad \mu_i \mid \phi \sim N\left(\phi, \gamma^2\right),$$

and ϵ_i is generated by

$$\xi \sim N\{0, (1-\eta^2)\sigma^2\}, \qquad \epsilon_i \mid \xi \sim N(\xi, \eta^2\sigma^2).$$

Note that $\eta = 1$ in Senn (2008), and we relax this assumption by allowing correlated errors.

2.3. Posterior mean

To derive the posterior mean of μ_p given $X_1, \ldots, X_p$, we rely on the following classic result.

Lemma 1 (*Normal Shrinkage*)**.** *Let*

$$\mu \sim N(\mu_0, \nu^2), \qquad Z_i \mid \mu \overset{i.i.d.}{\sim} N(\mu, \tau^2) \quad (i = 1, \ldots, n).$$

Then the posterior mean of μ *is*

$$\mathrm{E}(\mu \mid Z_1, \ldots, Z_n) = \frac{\tau^2\mu_0 + \nu^2 \sum_{i=1}^{n} Z_i}{\tau^2 + n\nu^2},$$

Proposition 1. *The posterior mean of μ_p given X_p is*

$$\mathrm{E}(\mu_p \mid X_p) = \frac{1}{1+\sigma^2} X_p. \tag{4}$$

Furthermore, let

$$a = \gamma^2 + \sigma^2\eta^2, \qquad b = 1 - \gamma^2 + \sigma^2(1-\eta^2)$$

and

$$r_1, \dots, r_{p-1} = \frac{\sigma^2(\eta^2-\gamma^2)}{a(a+pb)}, \qquad r_p = \frac{a+(p-1)b\gamma^2}{a(a+pb)}.$$

The posterior mean of μ_p given $X_1, \dots, X_p$ is

$$\mathrm{E}(\mu_p \mid X_1, \dots, X_p) = \sum_{i=1}^{p} r_i X_i. \tag{5}$$

Proof. To prove the first half, notice that

$$\mu_p \sim N(0,1), \qquad X_p \mid \mu_p \sim N(\mu_p, \sigma^2),$$

and apply Lemma 1.

To prove the second half, note that $\mu_i = \phi + \mu_i'$, where

$$\phi \sim N\left(0, 1-\gamma^2\right), \qquad \mu_i' \overset{i.i.d.}{\sim} N\left(0, \gamma^2\right);$$

and $\epsilon_i = \xi + \epsilon_i'$, where

$$\xi \sim N\{0, (1-\eta^2)\sigma^2\}, \qquad \epsilon_i' \overset{i.i.d.}{\sim} N(0, \eta^2\sigma^2).$$

Consequently we have

$$\phi + \xi \sim N(0, b), \qquad X_i \mid \phi + \xi \overset{i.i.d.}{\sim} N(\phi+\xi, a).$$

On the one hand, by Lemma 1

$$\mathrm{E}(\phi+\xi \mid X_1, \dots, X_p) = \frac{b}{a+pb} \sum_{i=1}^{p} X_i,$$

and

$$\mathrm{E}(\phi \mid \phi+\xi, X_1, \dots, X_p) = \frac{1-\gamma^2}{b} E(\phi+\xi \mid X_1, \dots, X_p).$$

Consequently,

$$\begin{aligned} \mathrm{E}(\phi \mid X_1, \dots, X_p) &= \mathrm{E}\{\mathrm{E}(\phi \mid \phi+\xi, X_1, \dots, X_p) \mid X_1, \dots, X_p\} \\ &= \frac{1-\gamma^2}{b} E(\phi+\xi \mid X_1, \dots, X_p) \\ &= \frac{1-\gamma^2}{a+pb} \sum_{i=1}^{p} X_i. \end{aligned} \tag{6}$$

On the other hand, similarly we have

$$\begin{aligned} \mathrm{E}(\mu_p' \mid X_1, \dots, X_p) &= \frac{\gamma^2}{a} E(\mu_i' + \epsilon_i' \mid X_1, \dots, X_p) \\ &= \frac{\gamma^2}{a} \left\{ X_p - \frac{b}{a+pb} \sum_{i=1}^{p} X_i \right\}. \end{aligned} \tag{7}$$

Combine (6) and (7), we complete the proof. □

It is worth noting that when $\gamma = \eta$, (5) reduces to (4).

3. Closed-form expression for the selection bias

To simplify future notations, we assume that X_p is the largest observation, i.e., $X_p = \max_{1\le i\le p} X_i$. Consequently, the selection bias defined in (1) becomes

$$\Delta = \mathrm{E}(\mu_p \mid X_p) - \mathrm{E}(\mu_p \mid X_p, X_p = \max X_i). \tag{8}$$

To derive its closed-form expression, we rely on the following lemmas.

Lemma 2. *Let* $\boldsymbol{X}_{-p} = (X_1, \ldots, X_{p-1})'$, *and its distribution conditioning on* X_p *is*

$$N\left(\frac{b}{a+b}\mathbf{1}_{p-1}X_p,\ a\boldsymbol{I}_{p-1} + \frac{ab}{a+b}\mathbf{1}_{p-1}\mathbf{1}'_{p-1}\right). \tag{9}$$

Proof. By (2) we have $\boldsymbol{X} \sim N(0, \boldsymbol{\Psi})$, where

$$\boldsymbol{\Psi} = (\psi_{jk})_{1\le j,k\le p} = a\boldsymbol{I}_p + b\mathbf{1}_p\mathbf{1}'_p.$$

Furthermore, let

$$\boldsymbol{\Psi}_{11} = (\psi_{jk})_{1\le j,k\le p-1} = a\boldsymbol{I}_{p-1} + b\mathbf{1}_{p-1}\mathbf{1}'_{p-1}, \qquad \boldsymbol{\Psi}_{22} = (\psi_{pp}) = a + b,$$

and

$$\boldsymbol{\Psi}_{12} = (\psi_{1p}, \ldots, \psi_{p-1,p})' = b\mathbf{1}_{p-1}, \qquad \boldsymbol{\Psi}_{21} = (\psi_{p1}, \ldots, \psi_{p,p-1}) = b\mathbf{1}'_{p-1}.$$

Simple probability argument suggests that

$$\boldsymbol{X}_{-p} \mid X_p \sim N\left(\boldsymbol{\Psi}_{12}^{-1}\boldsymbol{\Psi}_{22}X_p,\ \boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21}\right),$$

where

$$\boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}X_p = \frac{b}{a+b}\mathbf{1}_{p-1}X_p$$

and

$$\begin{aligned}\boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21} &= a\boldsymbol{I}_{p-1} + b\mathbf{1}_{p-1}\mathbf{1}'_{p-1} - \frac{b^2}{a+b}\mathbf{1}_{p-1}\mathbf{1}'_{p-1} \\ &= a\boldsymbol{I}_{p-1} + \frac{ab}{a+b}\mathbf{1}_{p-1}\mathbf{1}'_{p-1}.\end{aligned}$$

The proof is complete. □

To state the next lemma, we introduce some notations. First, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ and positive semi-definite matrix $\boldsymbol{\Omega} = (\omega_{jk})_{1\le j,k\le n}$, let

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)' \sim N(\boldsymbol{\theta}, \boldsymbol{\Omega}).$$

Second, let $V_i = Y_i - \theta_i$ for $i = 1, \ldots, n$. Consequently,

$$\boldsymbol{V} = (V_1, \ldots, V_n)' \sim N(\mathbf{0}, \boldsymbol{\Omega}),$$

whose probability density function is

$$f(\boldsymbol{v}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Omega}|^{1/2}}\mathrm{e}^{-\frac{1}{2}\boldsymbol{v}'\boldsymbol{\Omega}^{-1}\boldsymbol{v}}, \quad \boldsymbol{v} = (v_1, \ldots, v_n)'.$$

Third, for constants $b_1, \ldots, b_n$, we let

$$\alpha = \Pr(V_1 \le b_1 - \theta_1, \ldots, V_n \le b_n - \theta_n) = \int_{v_1\le b_1-\theta_1,\ldots,v_n\le b_n-\theta_n} f(\boldsymbol{v})d\boldsymbol{v},$$

and $\boldsymbol{W} = (W_1, \ldots, W_n)'$ be the truncation version of $\boldsymbol{V}$ from above at $(b_1 - \theta_1, \ldots, b_n - \theta_n)'$. Consequently, its probability density function is

$$g(\boldsymbol{w}) = \frac{1}{\alpha(2\pi)^{n/2}|\boldsymbol{\Omega}|^{1/2}}\mathrm{e}^{-\frac{1}{2}\boldsymbol{w}'\boldsymbol{\Omega}^{-1}\boldsymbol{w}} \cdot 1_{\{w_1\le b_1-\theta_1,\ldots,w_n\le b_n-\theta_n\}}, \quad \boldsymbol{w} = (w_1, \ldots, w_n)'.$$

For all $k = 1, \ldots, n$, let the kth marginal density function of $\boldsymbol{W}$ be

$$g_k(w) = \int_{-\infty}^{b_1-\theta_1} \cdots \int_{-\infty}^{b_{k-1}-\theta_{k-1}} \int_{-\infty}^{b_{k+1}-\theta_{k+1}} \cdots \int_{-\infty}^{b_n-\theta_n} g(w_1, \ldots, w_{k-1}, w, w_{k+1}, \ldots, w_n) \prod_{l\ne k} dw_l. \tag{10}$$

For efficient analytical and numerical evaluations of (10), see Cartinhour (1990) and Wilhelm and Manjunath (2010), respectively.

Lemma 3. *For all* $i = 1, \ldots, n$,

$$\mathrm{E}(Y_i \mid Y_1 \le b_1, \ldots, Y_n \le b_n) = \theta_i - \sum_{k=1}^{n} \omega_{ki} g_k(b_k - \theta_k).$$

Proof. The proof follows Manjunath and Wilhelm (2012). First,

$$\begin{aligned} \mathrm{E}(Y_i \mid Y_1 \le b_1, \ldots, Y_n \le b_n) &= \theta_i + E(V_i \mid V_1 \le b_1 - \theta_1, \ldots, V_n \le b_n - \theta_n) \\ &= \theta_i + E(W_i). \end{aligned} \tag{11}$$

Next, the moment generating function of $\boldsymbol{W}$ at $\boldsymbol{t} = (t_1, \ldots, t_n)'$ is

$$\begin{aligned} m(\boldsymbol{t}) &= \int \mathrm{e}^{\boldsymbol{t}'\boldsymbol{w}} g(\boldsymbol{w}) d\boldsymbol{w} \\ &= \frac{1}{\alpha (2\pi)^{n/2} |\boldsymbol{\Omega}|^{1/2}} \int_{w_1 \le b_1 - \theta_1, \ldots, w_n \le b_n - \theta_n} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{w}'\boldsymbol{\Omega}^{-1}\boldsymbol{w} - 2\boldsymbol{t}'\boldsymbol{w})} d\boldsymbol{w} \\ &= \underbrace{\mathrm{e}^{\frac{1}{2}\boldsymbol{t}'\boldsymbol{\Omega}\boldsymbol{t}}}_{m_1(\boldsymbol{t})} \underbrace{\frac{1}{\alpha (2\pi)^{n/2} |\boldsymbol{\Omega}|^{1/2}} \int_{w_1 \le b_1 - \theta_1, \ldots, w_n \le b_n - \theta_n} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\Omega}\boldsymbol{t})'\boldsymbol{\Omega}^{-1}(\boldsymbol{w} - \boldsymbol{\Omega}\boldsymbol{t})} d\boldsymbol{w}}_{m_2(\boldsymbol{t})}. \end{aligned}$$

On the one hand, by definition

$$\begin{aligned} E(W_i) &= \left. \frac{\partial m(\boldsymbol{t})}{\partial t_i} \right|_{\boldsymbol{t}=\boldsymbol{0}} \\ &= m_1(\boldsymbol{0}) \left. \frac{\partial m_2(\boldsymbol{t})}{\partial t_i} \right|_{\boldsymbol{t}=\boldsymbol{0}} + m_2(\boldsymbol{0}) \left. \frac{\partial m_1(\boldsymbol{t})}{\partial t_i} \right|_{\boldsymbol{t}=\boldsymbol{0}} \\ &= \left. \frac{\partial m_2(\boldsymbol{t})}{\partial t_i} \right|_{\boldsymbol{t}=\boldsymbol{0}}. \end{aligned} \tag{12}$$

On the other hand, let

$$b_i^* = b_i - \theta_i - \sum_{k=1}^{n} \omega_{ik} t_k, \quad i = 1, \ldots, n,$$

and we can rewrite $m_2(\boldsymbol{t})$ as

$$m_2(\boldsymbol{t}) = \int_{-\infty}^{b_1^*} \cdots \int_{-\infty}^{b_n^*} g(\boldsymbol{w}) dw_1 \ldots dw_n.$$

Therefore, by chain rule and Leibniz integral rule

$$\begin{aligned} \frac{\partial m_2(\boldsymbol{t})}{\partial t_i} &= \sum_{k=1}^{n} \frac{\partial b_k^*}{\partial t_i} \frac{\partial m_2(\boldsymbol{t})}{\partial b_k^*} \\ &= -\sum_{k=1}^{n} \omega_{ki} \int_{-\infty}^{b_1^*} \cdots \int_{-\infty}^{b_{k-1}^*} \int_{-\infty}^{b_{k+1}^*} \cdots \int_{-\infty}^{b_n^*} g(w_1, \ldots, w_{k-1}, b_k^*, w_{k+1}, \ldots, w_n) \prod_{l \ne k} dw_l, \end{aligned}$$

and consequently

$$\left. \frac{\partial m_2(\boldsymbol{t})}{\partial t_i} \right|_{\boldsymbol{t}=0} = -\sum_{k=1}^{n} \omega_{ki} g_k(b_k - \theta_k). \tag{13}$$

Combine (11)–(13), the proof is complete. □

Proposition 2. *For* $i = 1, \ldots, p-1$, *let* h_i *denote the ith marginal probability density function of the random vector defined by* (9) *truncated from above at* $\mathbf{1}_{p-1} X_p$. *Then the closed-form expression for* (8) *is*

$$\Delta = \frac{\sigma^2(\eta^2 - \gamma^2)}{1 + \sigma^2} \sum_{i=1}^{p-1} h_i \left(\frac{\gamma^2 + \sigma^2 \eta^2}{1 + \sigma^2} X_p \right). \tag{14}$$

Proof of Proposition 2. Apply Lemmas 2 and 3 to (9),

$$E(X_i \mid X_p, X_p = \max X_i) = \frac{a}{a+b}X_p - \underbrace{\left\{ \frac{ab}{a+b}\sum_{j=1}^{p-1} h_j\left(\frac{a}{a+b}X_p\right) + ah_i\left(\frac{a}{a+b}X_p\right)\right\}}_{\delta_i}.$$

Consequently, by (5) we have

$$\begin{aligned}
\mathrm{E}(\mu_p \mid X_p, X_p = \max X_i) &= r_p X_p + \sum_{i=1}^{p-1} r_i \mathrm{E}(X_i \mid X_p, X_p = \max X_i) \\
&= \left(r_p + \frac{a}{a+b}\sum_{i=1}^{p-1} r_i\right) X_p - \sum_{i=1}^{p-1} r_i \delta_i \\
&= \frac{X_p}{a+b} + \left\{\frac{(p-1)ab}{a+b} + a\right\} \sum_{i=1}^{p-1} r_i h_i \left(\frac{a}{a+b}X_p\right) \\
&= \mathrm{E}(\mu_p \mid X_p) - \frac{\sigma^2(\eta^2-\gamma^2)}{1+\sigma^2}\sum_{i=1}^{p-1} h_i\left(\frac{\gamma^2+\sigma^2\eta^2}{1+\sigma^2}X_p\right).
\end{aligned}$$

The proof is complete. □

Proposition 2 confirms the existence of the selection bias in general. Furthermore, it provides the following interesting insights:

1. For fixed σ, p and X_p, the sign of the selection bias is the same as the sign of $\eta^2 - \gamma^2$, i.e., depending on the correlation structures in (3), neglecting the fact that $X_p = \max_{1\le i\le p} X_i$ can either over-estimate or under-estimate μ_{i^*}. In particular, the selection bias is zero when $\gamma = \eta$. This is a generalization of the first main result in Senn (2008), which assumes that $\gamma = \eta = 1$;
2. For fixed γ, η, p and X_p, the selection bias goes to zero as σ goes to zero. This is intuitive because X_p approaches μ_p as σ goes to zero, and therefore the fact that $X_p = \max_{1\le i\le p} X_i$ becomes irrelevant;
3. For fixed σ, γ, η and p, the selection bias disappears for sufficiently large X_p. This is because when X_p goes to infinity,

   $$h_i\left(\frac{\sigma^2+\gamma^2\eta^2}{1+\sigma^2}X_p\right) \to 0, \quad i = 1, \ldots, p-1.$$

   This result is in connection with Dawid (1973).

4. Numerical and simulated examples

4.1. Numerical examples

Having derived the closed-form expression for the selection bias, we provide some numerical examples for illustration. Let $\sigma = 1$, $p \in \{3, 5, 10\}$ and $X_p \in \{0, 1, \ldots, 6\}$. For fixed p and X_p, we consider two cases. In Case 1, we follow Senn (2008) and let $\gamma^2 = 0.5$ and $\eta = 1$. In Case 2, we let $\gamma = 1$ and $\eta^2 = 0.5$. For both cases we calculate the selection bias by (14). Results are in Fig. 1, which align with the insights discussed in the previous section. Furthermore, it appears that the magnitude of the selection bias increases as p increases.

4.2. Simulated examples

The results in (14) enable us to calculate the "exact post-selection" posterior mean

$$\lambda_{i^*} = \mathrm{E}(\mu_{i^*} \mid X_{i^*}, X_{i^*} = \max X_i). \tag{15}$$

For illustration, we revisit the simulated example in Senn (2008), where $p = 10$, $\sigma = 2$, $\gamma^2 = 0.5$ and $\eta = 1$. Fig. 2 contains 5000 pairs of (μ_{i^*}, X_{i^*}) obtained by repeated sampling, the corresponding linear regression line that Senn (2008) used to approximate (15), and the curve that stands for the closed-form expression for (15).

The results in Fig. 2 suggest that the regression approximation is relatively accurate for non-extreme values of X_{i^*} but not for extreme ones. Therefore our analytical solution has an advantage over the regression approximation in Senn (2008). For further illustration we examine two concrete examples. First, let

$$x_{i^*} = 3.25, \qquad \Pr(X_{i^*} > x_{i^*}) = 0.486.$$

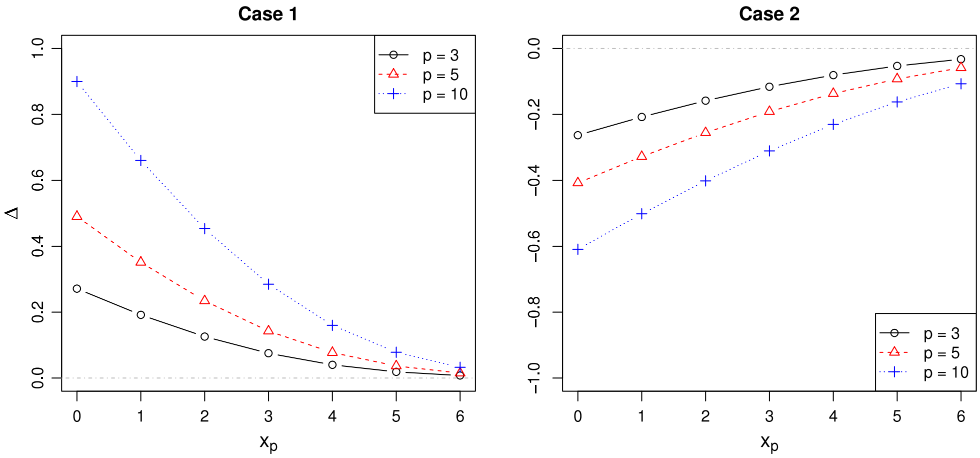

Fig. 1. Numerical examples of selection bias.

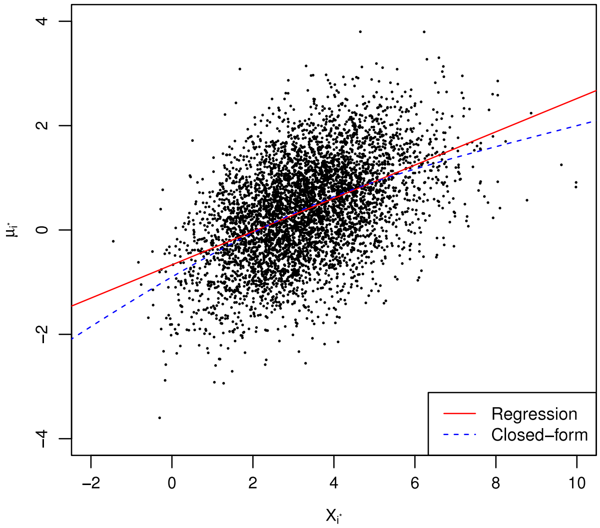

Fig. 2. "Exact post-selection" posterior mean: Regression approximation (solid line) and closed-form expression (dotted line).

Therefore 3.25 is a "common" value of X_{i^*}. In this case the exact value of (15) is $\lambda_{i^*} = 0.400$ and the regression approximation is $\hat{\lambda}_{i^*} = 0.368$. Consequently, although the "absolute discrepancy" $|\hat{\lambda}_{i^*} - \lambda_{i^*}| = 0.032$ seems small, the "relative discrepancy"

$$\frac{|\hat{\lambda}_{i^*} - \lambda_{i^*}|}{|\lambda_{i^*}|} = 8.1\%$$

is moderately large. Second, let

$$x_{i^*} = 1.5, \qquad \Pr(X_{i^*} \le x_{i^*}) = 0.102.$$

Therefore 1.5 is a relatively "uncommon" (but not extreme) value of X_{i^*}. In this case the absolute and relative discrepancies are respectively 0.062% and 24.7%, both moderately large.

5. Concluding remarks

For the treatment selection problem, quantifying the selection bias is important from both theoretical and practical perspectives. In this paper, we extend the heuristic discussion in Senn (2008) and derive the closed-form expression for the selection bias. We illustrate the advantages of our results by numerical and simulated examples.

There are multiple possible future directions based on our current work. First, we can reconcile our Bayesian analysis with Frequentist methods. Second, it is possible to extend our results to more general model specifications by using the Tweedie's formula (Robbins, 1956; Efron, 2011). Third, we need to explore "exact post-selection inference" in multiple hypothesis testing.

Acknowledgments

The authors thank the Co-Editor-in-Chief and a reviewer for their thoughtful comments that substantially improved the presentation of the paper.

References

Cartinhour, J., 1990. One-dimensional marginal density functions of a truncated multivariate normal density function. Commun. Stat.—Theory Methods 17 (1), 197–203.
Dawid, A.P., 1973. Posterior expectations for large observations. Biometrika 60 (3), 664–667.
Dawid, A.P., 1994. Selection paradoxes of Bayesian inference. In: Mvdtiaate Analysis and Its Applications. In: Lecture Notes—Monograph Series, vol. 24. pp. 211–220.
Efron, B., 2011. Tweedie's formula and selection bias. J. Amer. Statist. Assoc. 106 (496), 1602–1614.
G'Sell, M.G., Wager, S., Chouldechova, A., Tibshirani, R., 2016. Sequential selection procedures and false discovery rate control. J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (2), 423–444.
Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., Pohlmann, N., 2013. Online controlled experiments at large scale. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1168–1176.
Lee, J.D., Sun, D.L., Sun, Y., Taylor, J., 2016. Exact post-selection inference, with applications to the Lasso. Ann. Statist. 44 (3), 907–927.
Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the Lasso. Ann. Statist. 42 (2), 413–468.
Manjunath, B.G., Wilhelm, S., 2012. Moments calculation for the doubly truncated multivariate normal density. arXiv:1206.5387.
Reid, S., Tibshirani, R., 2016. Sparse regression and marginal testing using cluster prototypes. Biostatistics 17 (2), 364–376.
Robbins, H., 1956. An empirical bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I. University of California Press, Berkeley, Los Angeles, pp. 157–163.
Senn, S., 2008. A note concerning a selection "paradox" of Dawid's. Amer. Statist. 62 (3), 206–210.
Taylor, J., Tibshirani, R., 2015. Statistical learning and selective inference. Proc. Natl. Acad. Sci. 112 (25), 7629–7634.
Tibshirani, R.J., Taylor, J., Lockhart, R., Tibshirani, R., 2014. Exact post-selection inference for sequential regression procedures. J. Amer. Statist. Assoc., in press, arXiv:1401.3889.
Wilhelm, S., Manjunath, B.G., 2010. tmvtnorm: A package for the truncated multivariate normal distribution. R J. 2 (1), 25–29.