Supplement to "Affinity-based measures of biomarker performance evaluation"

Miguel de Carvalho[1], Bradley J. Barney[2] and Garritt L. Page[3]

1 Properties of affinity

1.1 Proof of Proposition 1

1. Since $\sqrt{f_D(y)} \geqslant 0$ and $\sqrt{f_{\bar{D}}(y)} \geqslant 0$ it follows that $\kappa \geqslant 0$. Cauchy–Schwarz inequality further implies that $|\kappa| = \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}} \rangle \leqslant \|\sqrt{f_D}\| \|\sqrt{f_{\bar{D}}}\| = 1$.

2. The argument is tantamount to that of Roos and Held,[1] but it is included here for completeness. Let $g(y) = z$ be a monotone increasing function and let

$$f_D^g(z) = f_D(g^{-1}(z))\frac{\mathrm{d}}{\mathrm{d}z}g^{-1}(z), \quad f_{\bar{D}}^g(z) = f_{\bar{D}}(g^{-1}(z))\frac{\mathrm{d}}{\mathrm{d}z}g^{-1}(z),$$

be densities of the transformed data $g(Y_D)$ and $g(Y_{\bar{D}})$,[2] and denote by $\kappa^g = \langle \sqrt{f_D^g}, \sqrt{f_{\bar{D}}^g} \rangle$ the affinity of the transformed data. It thus follows that

$$
\begin{aligned}
\kappa^g &= \int_{-\infty}^{\infty} \sqrt{f_D^g(z)}\sqrt{f_{\bar{D}}^g(z)}\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} \left\{ f_D(g^{-1}(z))\frac{\mathrm{d}}{\mathrm{d}z}g^{-1}(z) \right\}^{1/2} \left\{ f_{\bar{D}}(g^{-1}(z))\frac{\mathrm{d}}{\mathrm{d}z}g^{-1}(z) \right\}^{1/2}\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} \sqrt{f_D(g^{-1}(z))}\sqrt{f_{\bar{D}}(g^{-1}(z))}\frac{\mathrm{d}}{\mathrm{d}z}g^{-1}(z)\,\mathrm{d}z \qquad (\text{set } g^{-1}(z) = y) \\
&= \int_{-\infty}^{\infty} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y = \kappa.
\end{aligned}
$$

[1] School of Mathematics, University of Edinburgh, Scotland, UK
[2] Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA
[3] Department of Statistics, Brigham Young University, Provo, Utah, USA

Corresponding author:
Miguel de Carvalho
Email: miguel.decarvalho@ed.ac.uk

Properties of induced priors

1.2   Proof of Theorem 1

The proof involves manipulations similar to those used for showing continuity of the inner product (cf Hunter and Nachtergaele[3]), along with a result from Lijoi et al.[4] Just note that

$$
\begin{aligned}
|\kappa^\omega - \kappa| &= |\langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}^\omega}\rangle - \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}}\rangle| \\
&= |\langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}^\omega}\rangle - \langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}}\rangle + \langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}}\rangle - \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}}\rangle| \\
&\leqslant |\langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}^\omega}\rangle - \langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}}\rangle| + |\langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}}\rangle - \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}}\rangle| \\
&\leqslant |\langle \sqrt{f_D^\omega}, \sqrt{f_{\bar{D}}^\omega} - \sqrt{f_{\bar{D}}}\rangle| + |\langle \sqrt{f_D^\omega} - \sqrt{f_D}, \sqrt{f_{\bar{D}}}\rangle| \\
&\leqslant \underbrace{\|\sqrt{f_D^\omega}\|}_{1} \|\sqrt{f_{\bar{D}}^\omega} - \sqrt{f_{\bar{D}}}\| + \|\sqrt{f_D^\omega} - \sqrt{f_D}\| \underbrace{\|\sqrt{f_{\bar{D}}}\|}_{1} \\
&= \sqrt{d_H(f_{\bar{D}}^\omega, f_{\bar{D}})} + \sqrt{d_H(f_D^\omega, f_D)},
\end{aligned}
$$

where $d_H(f,g) = \int \{\sqrt{f(y)} - \sqrt{g(y)}\}^2 \, \mathrm{d}y$ is the Hellinger distance. So, as it can be seen from (1.2), to have $|\kappa^\omega - \kappa| < \varepsilon$, with $\varepsilon > 0$, it would suffice having $d_H(f_{\bar{D}}^\omega, f_{\bar{D}}) < \varepsilon^2/4$ and $d_H(f_D^\omega, f_D) < \varepsilon^2/4$. Thus,

$$
\{\omega \in \Omega : |\kappa^\omega - \kappa| < \varepsilon\} \supseteq \{\omega \in \Omega : d_H(f_{\bar{D}}^\omega, f_{\bar{D}}) < \varepsilon^2/4, d_H(f_D^\omega, f_D) < \varepsilon^2/4\},
$$

from where it finally follows that

$$
P\{\omega \in \Omega : |\kappa^\omega - \kappa| < \varepsilon\} \geqslant \underbrace{P\{\omega \in \Omega : d_H(f_{\bar{D}}^\omega, f_{\bar{D}}) < \varepsilon^2/4\}}_{\pi_{\bar{D}}} \underbrace{P\{\omega \in \Omega : d_H(f_D^\omega, f_D) < \varepsilon^2/4\}}_{\pi_D} > 0,
$$

given that from the equivalence between Hellinger and L_1 support, and from Section 3 in Lijoi et al.[4] it follows that $\pi_{\bar{D}} > 0$ and $\pi_D > 0$.

B.2. Proof of Theorem 2

The proof is along the same lines as that of Theorem 1, but it requires Theorem 4 of Barrientos et al.,[5] which is essentially a covariate-specific version of the result in Section 3 of Lijoi et al.[4] Similar derivations as those in (1.2) yield

$$
|\kappa^\omega(x_i) - \kappa(x_i)| \leqslant \sqrt{d_H(f_{\bar{D}|x_i}^\omega, f_{\bar{D}|x_i})} + \sqrt{d_H(f_{D|x_i}^\omega, f_{D|x_i})}.
$$

Hence, to have $|\kappa^\omega(x_i) - \kappa(x_i)| < \varepsilon$, for $\varepsilon > 0$, it would suffice having

$$
d_H(f_{\bar{D}|x_i}^\omega, f_{\bar{D}|x_i}) < \varepsilon^2/4, \quad d_H(f_{D|x_i}^\omega, f_{D|x_i}) < \varepsilon^2/4, \quad i = 1, \ldots, n,
$$

and thus using similar arguments to the ones in proof of Theorem 1 it follows that

$$
P\{\omega \in \Omega : |\kappa^\omega(x_i) - \kappa(x_i)| < \varepsilon\} \geqslant \underbrace{P\{\omega \in \Omega : d_H(f_{\bar{D}|x_i}^\omega, f_{\bar{D}|x_i}) < \varepsilon^2/4, i = 1, \ldots, n\}}_{\pi_{\bar{D}}}
$$
$$
\times \underbrace{P\{\omega \in \Omega : d_H(f_{D|x_i}^\omega, f_{D|x_i}) < \varepsilon^2/4, i = 1, \ldots, n\}}_{\pi_D} > 0,
$$

given that Theorem 4 of Barrientos et al.,[5] on the Hellinger support of the DDP, implies that $\pi_{\bar{D}} > 0$ and $\pi_D > 0$.

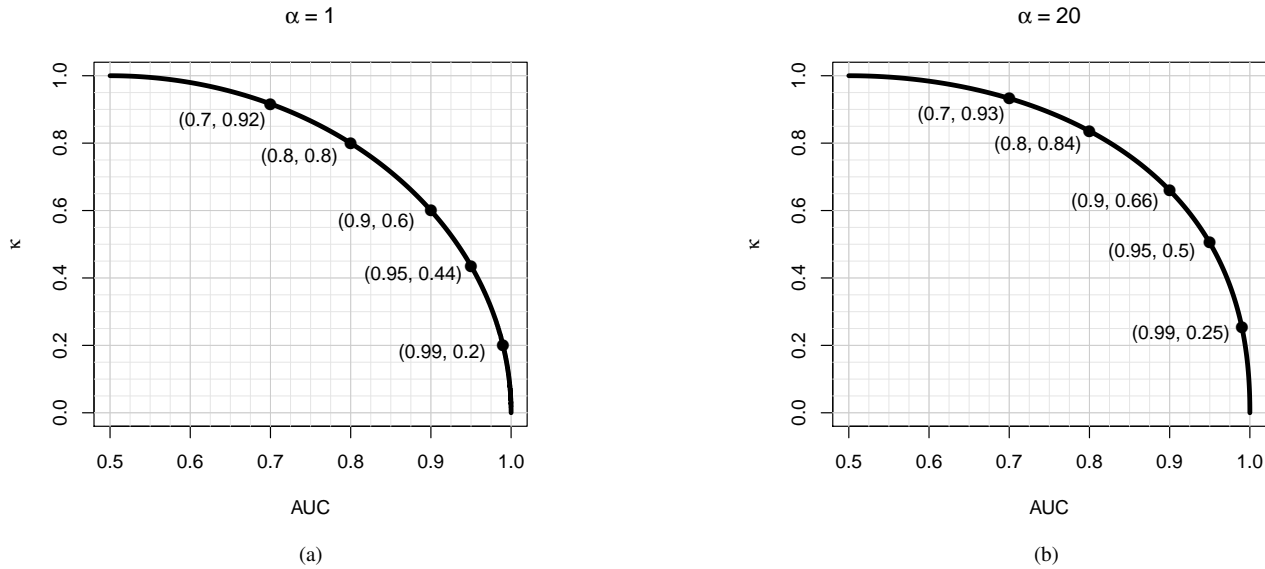**Figure 1.** Relationship between κ and AUC in the proper bigamma model for two different values of the shape parameter, α. a) $\alpha = 1$ (highly skewed densities); b) $\alpha = 20$ (mildly skewed).

2 Correspondence between κ, AUC in proper bigamma model

To build intuition regarding the relationship between κ and AUC, Figure 3 of the article plotted the relationship between κ and AUC in the proper binormal model. We include Figure 1 here, comprised of corresponding plots for the proper bigamma model, which is characterized by a common shape parameter for each gamma density. The relationship between the two summary measures is affected by the shape parameter. As such, we consider both a shape parameter of 1 (panel a) to represent highly skewed gamma densities and a shape parameter of 20 (panel b) to represent mildly skewed gamma densities. We note that qualitatively the curves are rather similar even for very disparate levels of skewness. We also note that with a moderately large shape parameter, the proper bigamma plot linking kappa and AUC is nearly identical to that of the proper binormal model (cf. Figure 3 of the article).

3 Simulation setting figures

Because of limited space in the article, only the second unconditional setting (70/30 mixtures of normals) from the simulation study was illustrated with a figure. We therefore include Figure 2 to display the densities, κs, and AUCs from the first unconditional setting of Table 2. We also had included three covariate-dependent settings. Due to the challenges inherent in overlaying bivariate densities, we do not display the densities for the three conditional settings of Table 2, instead displaying only $\kappa(x)$ and $\text{AUC}(x)$. These covariate-dependent quantities are plotted in Figure 3.

4 Figures corresponding to simulation results

Because of limited space in the article, only the selected simulation results were depicted. We therefore include Figures 4 and 5 below. Note that as mentioned in text, estimating κ that is close to zero tends to be more biased.
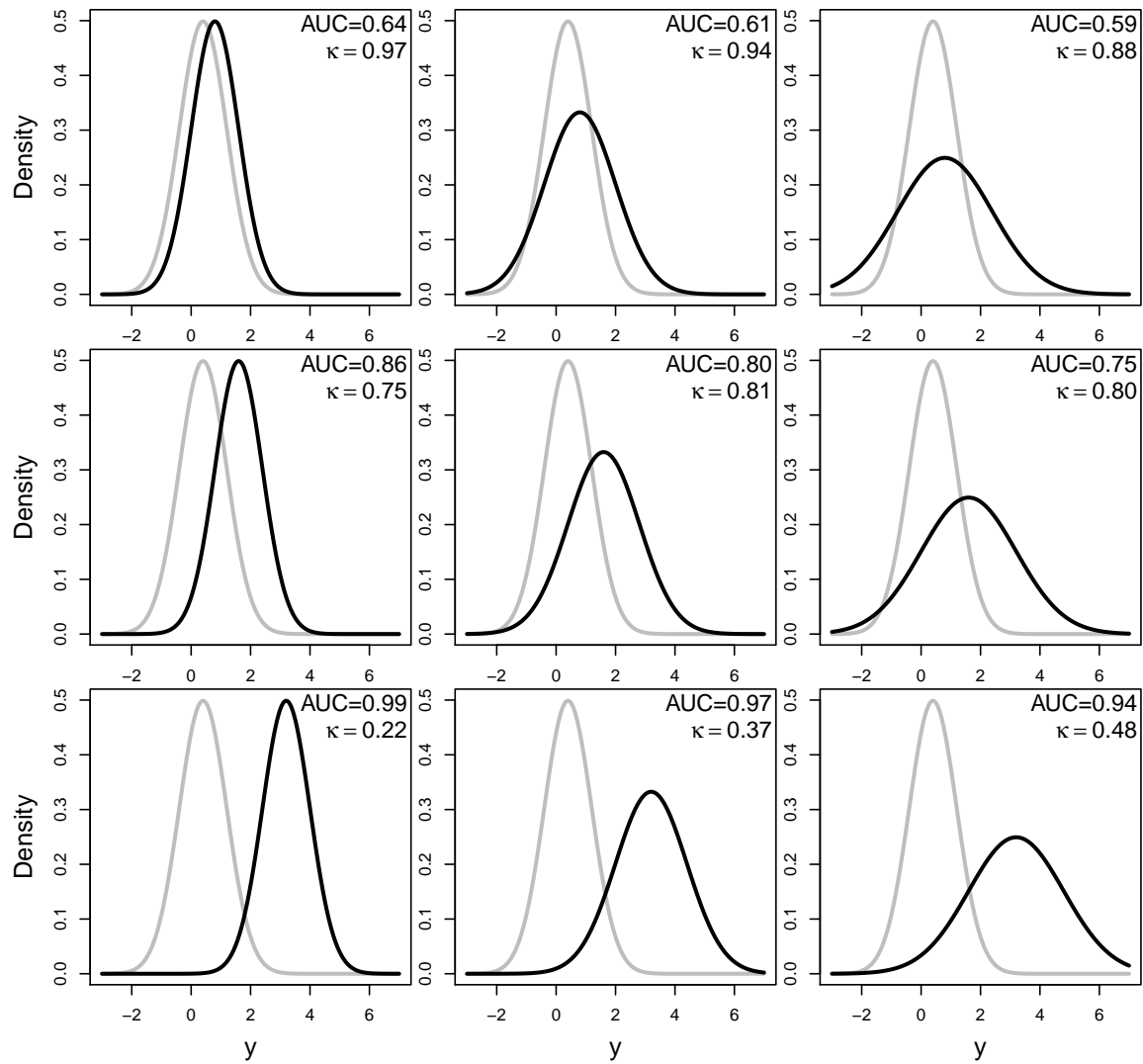
Figure 2. Densities for the first unconditional simulation study setting in Table 2 of the article; the black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects.
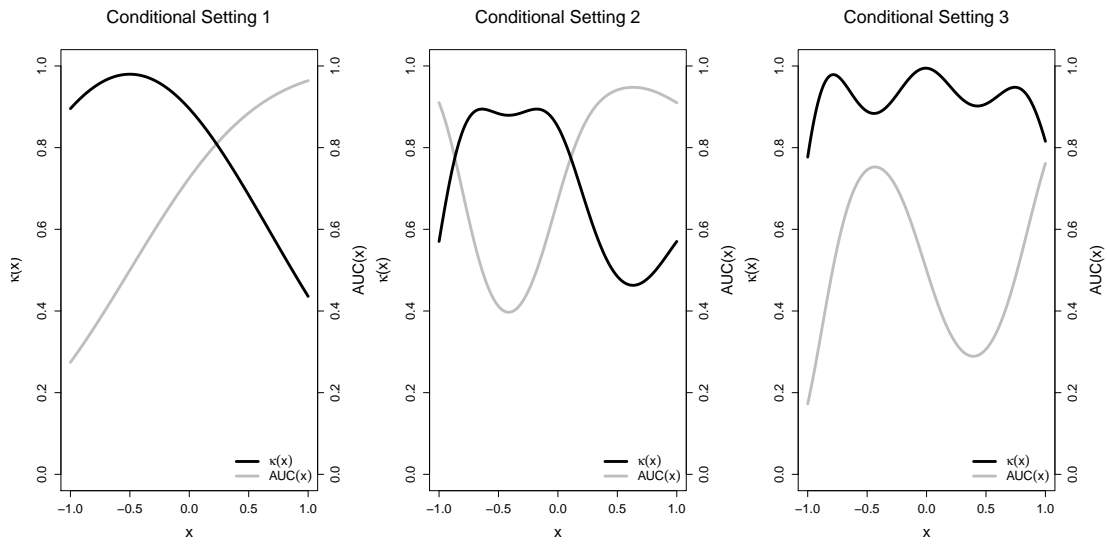
Figure 3. Covariate-dependent affinity and AUC for the three conditional settings in Table 2 of the article; the black and grey lines respectively denote $\kappa(x)$ and $\mathrm{AUC}(x)$.
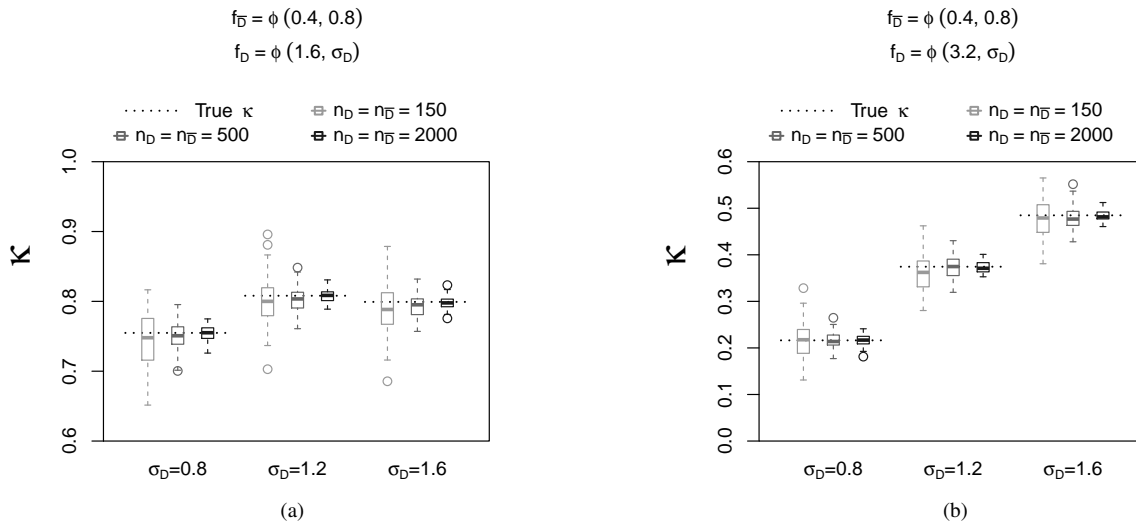

Figure 4. AUC and κ estimates (average across 100 simulations) along with true values in the first unconditional scenario of the simulation study (Table 2 from main manuscript): Both populations are normals with a) diseased mean equal to 1.6 and b) diseased mean equal to 3.2.
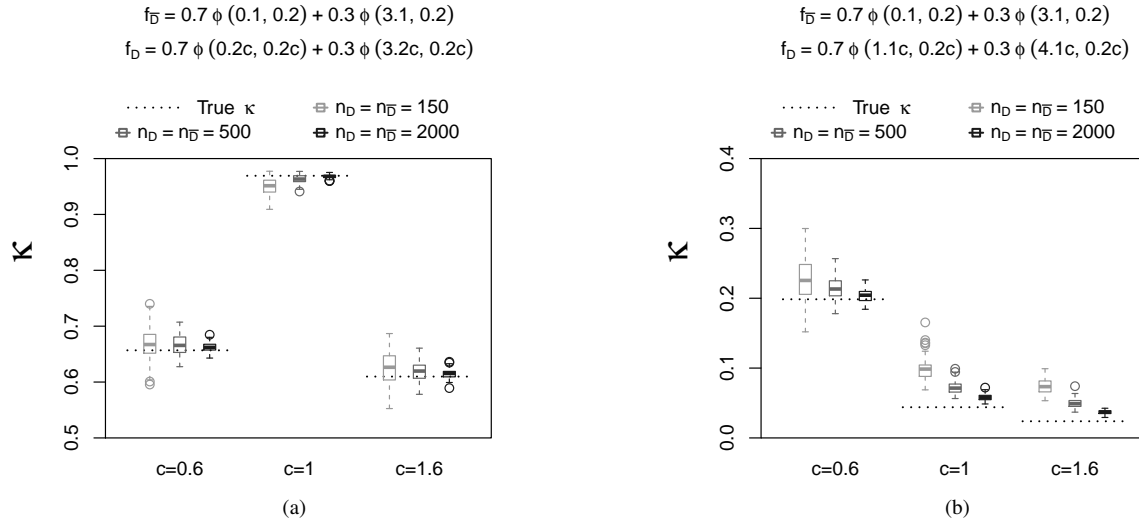
Figure 5. AUC and κ estimates (average across 100 simulations) along with true values in the second unconditional scenario of the simulation study (Table 2 from main manuscript): In both scenarios means and standard deviations of mixture components vary.

5 Analytical derivations of entries in Table 1

Binormal

Let $Y_D \sim N(\mu_D, \sigma_D)$ and $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}})$. Then,

$$
\begin{aligned}
\kappa &= \int_{-\infty}^{+\infty} \sqrt{f_D(y)} \sqrt{f_{\bar{D}}(y)} \, dy \\
&= \int_{-\infty}^{+\infty} \left[(2\pi\sigma_D^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(y-\mu_D)^2}{\sigma_D^2} \right\} \right]^{1/2} \left[(2\pi\sigma_{\bar{D}}^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(y-\mu_{\bar{D}})^2}{\sigma_{\bar{D}}^2} \right\} \right]^{1/2} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_D\sigma_{\bar{D}}} \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{4}\left(\frac{(y-\mu_D)^2}{\sigma_D^2} + \frac{(y-\mu_{\bar{D}})^2}{\sigma_{\bar{D}}^2} \right) \right\} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_D\sigma_{\bar{D}}} \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{4\sigma_D^2\sigma_{\bar{D}}^2}\left((\sigma_D^2 + \sigma_{\bar{D}}^2)y^2 - 2(\sigma_{\bar{D}}^2\mu_D + \sigma_D^2\mu_{\bar{D}})y + \sigma_{\bar{D}}^2\mu_D^2 + \sigma_D^2\mu_{\bar{D}}^2 \right) \right\} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_D\sigma_{\bar{D}}} \exp\left\{ -\frac{1}{4\sigma_D^2\sigma_{\bar{D}}^2}\left(\sigma_{\bar{D}}^2\mu_D^2 + \sigma_D^2\mu_{\bar{D}}^2 - \frac{(\sigma_{\bar{D}}^2\mu_D + \sigma_D^2\mu_{\bar{D}})^2}{\sigma_{\bar{D}}^2 + \sigma_D^2} \right) \right\} \\
&\qquad \times \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{2(2\sigma_D^2\sigma_{\bar{D}}^2/(\sigma_D^2 + \sigma_{\bar{D}}^2))}\left(y - \left(\frac{\sigma_{\bar{D}}^2\mu_D + \sigma_D^2\mu_{\bar{D}}}{\sigma_{\bar{D}}^2 + \sigma_D^2} \right) \right)^2 \right\} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_D\sigma_{\bar{D}}} \sqrt{2\pi\frac{2\sigma_D^2\sigma_{\bar{D}}^2}{\sigma_D^2 + \sigma_{\bar{D}}^2}} \exp\left\{ -\frac{1}{4\sigma_D^2\sigma_{\bar{D}}^2}\left(\sigma_{\bar{D}}^2\mu_D^2 + \sigma_D^2\mu_{\bar{D}}^2 - \frac{(\sigma_{\bar{D}}^2\mu_D + \sigma_D^2\mu_{\bar{D}})^2}{\sigma_{\bar{D}}^2 + \sigma_D^2} \right) \right\} \\
&= \sqrt{\frac{2\sigma_D\sigma_{\bar{D}}}{\sigma_D^2 + \sigma_{\bar{D}}^2}} \exp\left\{ -\frac{1}{4}\frac{(\mu_D - \mu_{\bar{D}})^2}{\sigma_D^2 + \sigma_{\bar{D}}^2} \right\}.
\end{aligned}
$$

Bibeta

Let $Y_D \sim \text{Beta}(a_D, b_D)$ and $Y_{\bar{D}} \sim \text{Beta}(a_{\bar{D}}, b_{\bar{D}})$. Then,

$$
\begin{aligned}
\kappa &= \int_{-\infty}^{+\infty} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y \\
&= \int_0^1 \left\{ \frac{y^{a_D-1}(1-y)^{b_D-1}}{B(a_D, b_D)} \right\}^{1/2} \left\{ \frac{y^{a_{\bar{D}}-1}(1-y)^{b_{\bar{D}}-1}}{B(a_{\bar{D}}, b_{\bar{D}})} \right\}^{1/2}\,\mathrm{d}y \\
&= \frac{\int_0^1 y^{(a_D+a_{\bar{D}})/2-1}(1-y)^{(b_D+b_{\bar{D}})/2-1}\,\mathrm{d}y}{\{B(a_D, b_D)B(a_{\bar{D}}, b_{\bar{D}})\}^{1/2}} = \frac{B((a_D+a_{\bar{D}})/2, (b_D+b_{\bar{D}})/2)}{\{B(a_D, b_D)B(a_{\bar{D}}, b_{\bar{D}})\}^{1/2}},
\end{aligned}
$$

where $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$.

Bigamma

Let $Y_D \sim \text{Gamma}(\text{shape} = \alpha_D, \text{rate} = \beta_D)$ and $Y_{\bar{D}} \sim \text{Gamma}(\text{shape} = \alpha_{\bar{D}}, \text{rate} = \beta_{\bar{D}})$. Then,

$$
\begin{aligned}
\kappa &= \int_{-\infty}^{+\infty} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y \\
&= \int_0^{+\infty} \left[\frac{\beta_D^{\alpha_D}}{\Gamma(\alpha_D)} y^{\alpha_D-1} \exp\{-\beta_D y\} \right]^{1/2} \left[\frac{\beta_{\bar{D}}^{\alpha_{\bar{D}}}}{\Gamma(\alpha_{\bar{D}})} y^{\alpha_{\bar{D}}-1} \exp\{-\beta_{\bar{D}} y\} \right]^{1/2}\,\mathrm{d}y \\
&= \left[\frac{\beta_D^{\alpha_D}\beta_{\bar{D}}^{\alpha_{\bar{D}}}}{\Gamma(\alpha_D)\Gamma(\alpha_D)} \right]^{1/2} \int_0^{+\infty} y^{(\alpha_D+\alpha_D)/2-1} \exp\left\{ -\frac{\beta_D+\beta_{\bar{D}}}{2} y \right\}\,\mathrm{d}y \\
&= \left[\frac{\beta_D^{\alpha_D}\beta_{\bar{D}}^{\alpha_{\bar{D}}}}{\Gamma(\alpha_D)\Gamma(\alpha_D)} \right]^{1/2} \frac{\Gamma((\alpha_D+\alpha_{\bar{D}})/2)}{((\beta_D+\beta_{\bar{D}})/2)^{(\alpha_D+\alpha_{\bar{D}})/2}}.
\end{aligned}
$$

6 Additional empirical reports

In the PSA data application (as reported in the article), the analysis proceeded as though 683 independent measures were obtained. This was done to enable comparison with an existing analysis of this data set by Rodriguez and Martinez.[6] One natural alternative is to use only the last available observation from each of the $n = 141$ subjects. We made this restriction and repeated the analyses. In doing so, the range of observed ages was slightly reduced from [46.75, 80.83]—the range with the 683 observations—to [51.94, 80.83]. To apply the model in light of this smaller age range, the ages were rescaled so that 51.94 became $x = -1$ and 80.83 became $x = 1$. The results from the unconditional and age-dependent analyses are contained in Figure 6 and Figure 7, respectively. The results when restricting the data to the last available observations are rather similar to the results when ignoring the dependence among all 683 observations. However, not only is there more uncertainty in the affinity and AUC estimates, but the estimates of both κ and AUC are slightly more favorable with the 141 observations than they were with the 683 observations. This might be due to the fact that the last available observations on the PSA biomarkers were the closest observations to the time of diagnosis.

7 Supplemental online resources

The R code for replicating the simulation is available from

http://www.maths.ed.ac.uk/~mdecarv/files/kappasim.zip

A shiny app illustrating how affinity looks like vis-à-vis the AUC of a binormal model is available from

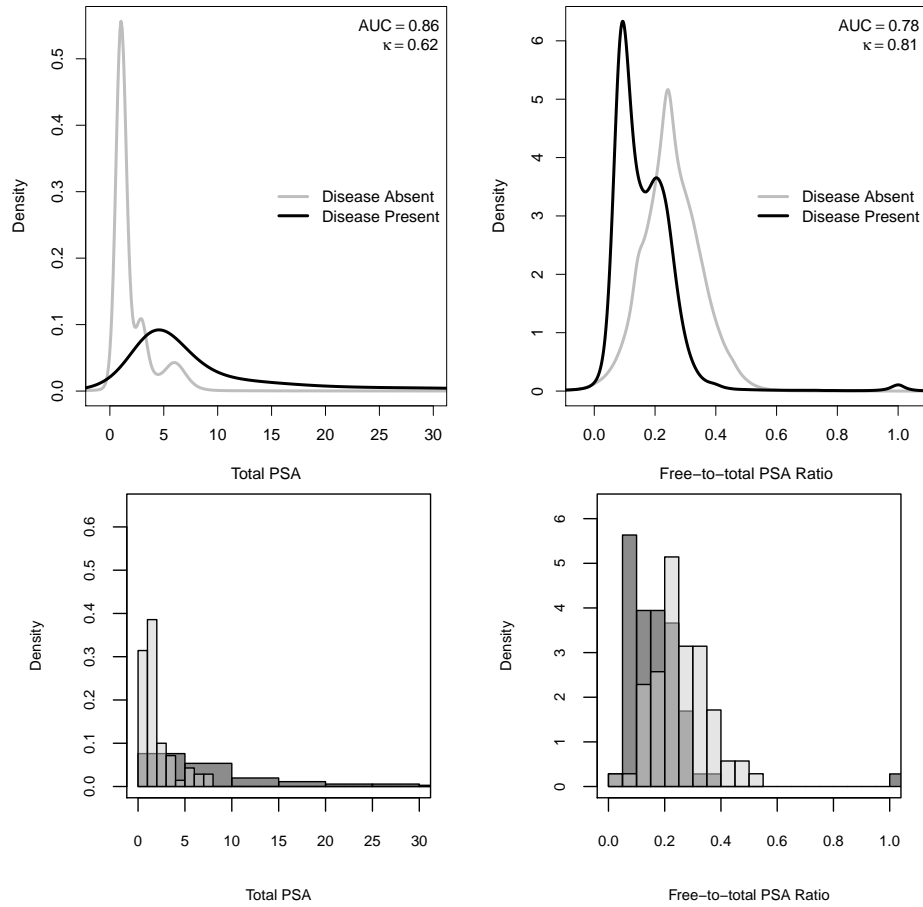https://mdecarvalho.shinyapps.io/decarvalho2018/

Figure 6. Last available observation analysis I; this figures compares with Fig. 7 on the manuscript. Top: DPM-based estimated densities along with AUC and κ values when age is not considered. The black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects. Bottom: Overlapping histograms. For display purposes, the largest total PSA values ($< 8\%$ of the $n = 141$ observations) are not shown in the leftmost plots. All total PSA values were below 100.
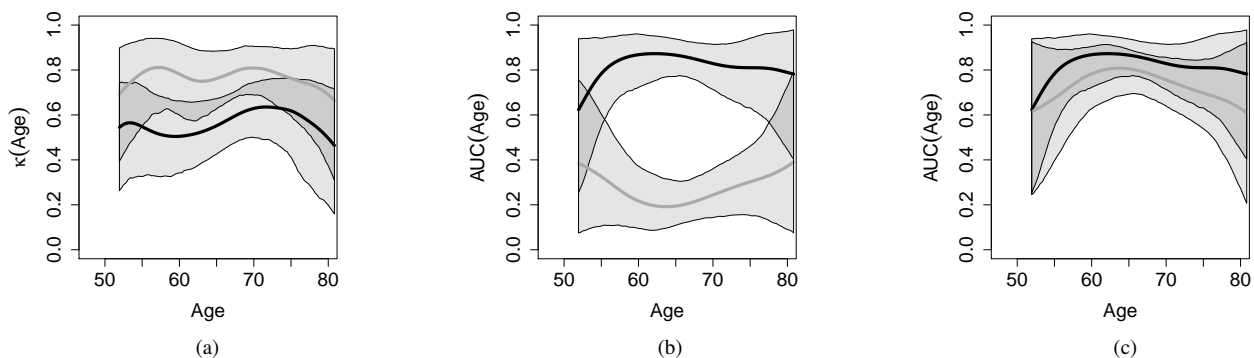

Figure 7. Last available observation analysis II; this figures compares with Fig. 8. Means and 95% pointwise credible intervals for the age-adjusted affinity and AUC of two biomarkers in cases and controls. Only the last available observation per subject was considered. a) is the age-adjusted affinity; b) is the age-adjusted AUC if both biomarkers have upper-tailed diagnostic tests; c) is the age-adjusted AUC if the second biomarker diagnostic test is lower-tailed. In each panel, the black and grey lines respectively denote the first and second biomarkers.

References

1. Roos M and Held L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis* 2011; 6(2): 259–278.

2. Knight K. *Mathematical Statistics*. Boca Raton: Chapman & Hall/CRC Press, 2000. ISBN 978-1-58488-178-0.

3. Hunter J and Nachtergaele B. *Applied Analysis*. London: World Scientific, 2001.

4. Lijoi A, Prünster I and Walker S. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli* 2004; 10(4): 651–663. DOI: 10.3150/bj/1093265634.

5. Barrientos A, Jara A and Quintana F. On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* 2012; 7: 277–310. DOI:10.1214/12-BA709.

6. Rodriguez A and Martinez J. Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics* 2014; 15(2): 353–369. DOI:10.1093/biostatistics/kxt044.