Model-based Rank Aggregation

Based on discussions with Tom Dietterich and Andrew Emmott

Generative Process with Separate λ for each detector

1. Draw $\pi \sim Beta(scale = 0.05, shape = 200)$. This is the proportion of anomalous points
2. Draw $\boldsymbol{\lambda} \triangleq \{\boldsymbol{\alpha}, \boldsymbol{\beta}\} \sim Gamma(.,.)$. This is the score "bonus" for being anomalous
3. For each point x,
    a. Draw $\theta_x \sim Bern(\pi)$
    b. If $\theta_x = 0$ then x is "normal" and $score(x, i) \sim Gamma(\alpha_r, \beta_r)$
    c. Else x is "anomalous", and $score(x, i) \sim Gamma(\alpha_a, \beta_a)$
4. Sort the {x} into descending order and assign ranks such that rank(x, i) is the position of x in the sorted list.

Assume there are *D* detectors. $\boldsymbol{x}_i = \{x_{i1}, \dots, x_{iD}\}$ are the scores reported by detectors $d_1, \dots d_D$. Let $\boldsymbol{\alpha}_r = \{\alpha_{r1}, \dots, \alpha_{rD}\}$ be the mean scores for regular instances and let $\boldsymbol{\alpha}_a = \alpha_{a1}, \dots, \alpha_{aD}$ be the means for anomaly scores. The likelihood of a score under the distribution for 'regular' scores is:

$$f_r(\boldsymbol{x}_i | \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r) \sim Gamma(\alpha_r, \beta_r) = \prod_{d=1}^{D} \frac{{\beta_{rd}}^{\alpha_{rd}}}{\Gamma(\alpha_{rd})} {x_{rd}}^{\alpha_{rd}-1} e^{-\beta_{rd} x_{rd}} \quad (1)$$

And the likelihood of a score under the distribution for 'anomalous' scores is:

$$f_a(\boldsymbol{x}_i | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a) \sim Gamma(\alpha_a, \beta_a) = \prod_{d=1}^{D} \frac{{\beta_{ad}}^{\alpha_{ad}}}{\Gamma(\alpha_{ad})} {x_{ad}}^{\alpha_{ad}-1} e^{-\beta_{ad} x_{ad}} \quad (2)$$

The likelihood of each score assuming we know whether it is normal or anomalous is:

$$f(\boldsymbol{x}_i | \pi, \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a, \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r) = \big((1-\pi) f_r(\boldsymbol{x}_i | \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r)\big)^{1-\theta_i} \big(\pi f_a(\boldsymbol{x}_i | \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a)\big)^{\theta_i} \quad (3)$$

The priors on $\pi, \boldsymbol{\lambda}_r = \{\boldsymbol{\alpha}_r, \boldsymbol{\beta}_r\}, \boldsymbol{\lambda}_a = \{\boldsymbol{\alpha}_a, \boldsymbol{\beta}_a\}$ are:

$$f_\pi(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha, \beta)}, \qquad f_{\lambda_a}(\alpha_{ad}, \beta_{ad} | p_{ad}, q_{ad}, r_{ad}, s_{ad}) \propto \frac{{p_{ad}}^{\alpha_{ad}-1}}{\Gamma(\alpha_{ad})^{r_{ad}} {\beta_{ad}}^{-\alpha_{ad} s_{ad}}} e^{-\beta_{ad} q_{ad}},$$

$$f_{\lambda_r}(\alpha_{rd}, \beta_{rd} | p_{rd}, q_{rd}, r_{rd}, s_{rd}) \propto \frac{{p_{rd}}^{\alpha_{rd}-1}}{\Gamma(\alpha_{rd})^{r_{rd}} {\beta_{rd}}^{-\alpha_{rd} s_{rd}}} e^{-\beta_{rd} q_{rd}}$$

Where we assume that α, β, $\{p_{ad}, q_{ad}, r_{ad}, s_{ad}\}$ and $\{p_{rd}, q_{rd}, r_{rd}, s_{rd}\}$ are known constants; $B(\alpha, \beta)$ is the Beta function and $\Gamma(\alpha_{ad})$, $\Gamma(\alpha_{rd})$ are Gamma functions. We will denote $\lambda_{ad} = \{\alpha_{ad}, \beta_{ad}\}$ and $\lambda_{rd} = \{\alpha_{rd}, \beta_{rd}\}$.

Therefore,

$$f(\boldsymbol{x}_i, \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r) \propto f(\boldsymbol{x}_i | \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r) f_{\lambda_a}(\boldsymbol{\lambda}_a) f_{\lambda_r}(\boldsymbol{\lambda}_r) f_\pi(\pi)$$

$$\propto \left(\pi \prod_{d=1}^{D} f_a(x_{id} | \lambda_{ad})\right)^{\theta_i} \left((1 - \pi) \prod_{d=1}^{D} f_r(x_{id} | \lambda_{rd})\right)^{1-\theta_i} \left(\prod_{d=1}^{D} \frac{{p_{ad}}^{\alpha_{ad}-1}}{\Gamma(\alpha_{ad})^{r_{ad}} {\beta_{ad}}^{-\alpha_{ad} s_{ad}}} e^{-\beta_{ad} q_{ad}} \frac{{p_{rd}}^{\alpha_{rd}-1}}{\Gamma(\alpha_{rd})^{r_{rd}} {\beta_{rd}}^{-\alpha_{rd} s_{rd}}} e^{-\beta_{rd} q_{rd}}\right) \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha, \beta)} \quad (4)$$

The complete-data likelihood of all scores across all detectors is (where *n* is the number of instances):

$$L(\boldsymbol{x}; \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r) \propto \left[\left\{\prod_{i=1}^{n} \prod_{d=1}^{D} f(x_{id} | \pi, \lambda_{ad})\right\}\left\{\prod_{d=1}^{D} f_{\lambda_a}(\lambda_{ad}) f_{\lambda_r}(\lambda_{rd})\right\}\right] f_\pi(\pi) \quad (5)$$

$$= \left[\prod_{i=1}^{n} \left\{ \left(\pi \prod_{d=1}^{D} f_a(x_{id}|\lambda_{ad}) \right)^{\theta_i} \left((1 - \pi) \prod_{d=1}^{D} f_r(x_{id}|\lambda_{rd}) \right)^{1-\theta_i} \right\} \right] \left(\prod_{d=1}^{D} \frac{p_{ad}{}^{\alpha_{ad}-1}}{\Gamma(\alpha_{ad})^{r_{ad}} \beta_{ad}{}^{-\alpha_{ad} s_{ad}}} e^{-\beta_{ad} q_{ad}} \frac{p_{rd}{}^{\alpha_{rd}-1}}{\Gamma(\alpha_{rd})^{r_{rd}} \beta_{rd}{}^{-\alpha_{rd} s_{rd}}} e^{-\beta_{rd} q_{rd}} \right) \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{6}$$

Instead of assigning the hard class labels ($(1-\theta_i), \theta_i$) which will be hard to infer, we will use soft-assignments (i.e., responsibilities) denoted by z_{ai} and z_{ri} which refer to the probability of assigning *i*-th score of *d*-th detector to 'anomaly' and 'normal' classes respectively; $z_{ai} + z_{ri} = 1$. We rewrite $L(\boldsymbol{x}_i; \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r)$:

$$\propto \left[\prod_{i=1}^{n} \left\{ \left(\pi \prod_{d=1}^{D} f_a(x_{id}|\lambda_{ad}) \right)^{z_{ai}} \left((1 - \pi) \prod_{d=1}^{D} f_r(x_{id}|\lambda_{rd}) \right)^{z_{ri}} \right\} \right] \left[\prod_{d=1}^{D} \left\{ \prod_{k \in \{a,r\}} \frac{p_{kd}{}^{\alpha_{kd}-1}}{\Gamma(\alpha_{kd})^{r_{kd}} \beta_{kd}{}^{-\alpha_{kd} s_{kd}}} e^{-\beta_{kd} q_{kd}} \right\} \right] \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{7}$$

The log-likelihood is:

$$l(\boldsymbol{x}; \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r, \sigma^2) = \log\left(L(\boldsymbol{x}; \pi, \boldsymbol{\lambda}_a, \boldsymbol{\lambda}_r, \sigma^2)\right)$$

$$\begin{aligned} &= \sum_{i=1}^{n} z_{ai} \left\{ \log(\pi) + \sum_{d=1}^{D} \log\left(f_a(x_{id}|\lambda_{ad})\right) \right\} + \sum_{i=1}^{n} z_{ri} \left\{ \log(1-\pi) + \sum_{d=1}^{D} \log\left(f_r(x_{id}|\lambda_{rd})\right) \right\} \\ &\quad + \sum_{d=1}^{D} \sum_{k \in \{a,r\}} \left\{ (\alpha_{kd} - 1) \log(p_{kd}) + \alpha_{kd} s_{kd} \log(\beta_{kd}) - r_{kd} \log\left(\Gamma(\alpha_{kd})\right) - \beta_{kd} q_{kd} \right\} \\ &\quad + (\alpha - 1)\log(\pi) + (\beta - 1)\log(1-\pi) - \log\left(B(\alpha,\beta)\right) \end{aligned} \tag{8}$$

After substituting $f_a(x_{id}|\lambda_{ad})$ and $f_r(x_{id}|\lambda_{rd})$ in the above equation:

$$\begin{aligned} l(\boldsymbol{x}; \pi, \lambda_{ad}, \lambda_{rd}) &= \sum_{i=1}^{n} z_{ai} \left\{ \log(\pi) + \sum_{d=1}^{D} \left\{ (\alpha_{ad} - 1)\log(x_{id}) + \alpha_{ad} \log(\beta_{ad}) - \log\left(\Gamma(\alpha_{ad})\right) - \beta_{ad} x_{id} \right\} \right\} \\ &\quad + \sum_{i=1}^{n} z_{ri} \left\{ \log(1-\pi) + \sum_{d=1}^{D} \left\{ (\alpha_{rd} - 1)\log(x_{id}) + \alpha_{rd} \log(\beta_{rd}) - \log\left(\Gamma(\alpha_{rd})\right) - \beta_{rd} x_{id} \right\} \right\} \\ &\quad + \sum_{d=1}^{D} \sum_{k \in \{a,r\}} \left\{ (\alpha_{kd} - 1) \log(p_{kd}) + \alpha_{kd} s_{kd} \log(\beta_{kd}) - r_{kd} \log\left(\Gamma(\alpha_{kd})\right) - \beta_{kd} q_{kd} \right\} \\ &\quad + (\alpha - 1)\log(\pi) + (\beta - 1)\log(1-\pi) - \log\left(B(\alpha,\beta)\right) + (\mathit{some\ constant}) \end{aligned} \tag{9}$$

M-Step: Derive the MLE of parameters by differentiation.

MLE for π:

$$\frac{\partial l(\boldsymbol{x}; \pi, \lambda_{ad}, \lambda_{rd}, \sigma^2)}{\partial \pi} = \frac{1}{\pi} \sum_{i=1}^{n} z_{ai} - \frac{1}{1-\pi} \sum_{i=1}^{n} z_{ri} + \frac{(\alpha - 1)}{\pi} - \frac{(\beta - 1)}{1-\pi} = 0$$

$$\Rightarrow (1-\pi)\sum_{i=1}^{n} z_{ai} - \pi\sum_{i=1}^{n} z_{ri} + (1-\pi)(\alpha-1) - \pi(\beta-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_{ai} + (\alpha-1) = \pi\left(\sum_{i=1}^{n} z_{ai} + \sum_{i=1}^{n} z_{ri} + (\alpha-1) + (\beta-1)\right)$$

$$\Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{n} z_{ai} + (\alpha-1)}{\left(\sum_{i=1}^{n} z_{ai} + \sum_{i=1}^{n} z_{ri} + (\alpha-1) + (\beta-1)\right)}$$

$$\Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{n} z_{ai} + (\alpha-1)}{\left(n + (\alpha-1) + (\beta-1)\right)} \tag{10}$$

The above follows from the observation that: $z_{ai} + z_{ri} = 1$.

MLE for β_{kd} for $k \in \{a, r\}$:

$$\frac{\partial l(\boldsymbol{x}; \pi, \lambda_{ad}, \lambda_{rd})}{\partial \beta_{kd}} = \frac{\partial(\sum_{i=1}^{n} z_{ki}\{\alpha_{kd}\log(\beta_{kd}) - \beta_{kd}x_{id}\} + \alpha_{kd}s_{kd}\log(\beta_{kd}) - \beta_{kd}q_{kd})}{\partial \beta_{kd}} = 0$$

$$\Rightarrow \frac{\partial(\alpha_{kd}(s_{kd} + \sum_{i=1}^{n} z_{ki})\log(\beta_{kd}) - (q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id})\beta_{kd})}{\partial \beta_{kd}} = 0$$

$$\Rightarrow \frac{\alpha_{kd}}{\beta_{kd}}\left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right) = \left(q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}\right)$$

$$\Rightarrow \hat{\beta}_{kd} = \alpha_{kd}\frac{s_{kd} + \sum_{i=1}^{n} z_{ki}}{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}} \tag{11}$$

MLE for α_{kd} for $k \in \{a, r\}$:

$$\frac{\partial l(\boldsymbol{x}; \pi, \lambda_{ad}, \lambda_{rd})}{\partial \alpha_{kd}} = \frac{\partial}{\partial \alpha_{kd}}\left\{\sum_{i=1}^{n} z_{ki}\log(\pi) + \sum_{i=1}^{n} z_{ki}\{(\alpha_{kd}-1)\log(x_{id}) + \alpha_{kd}\log(\beta_{kd}) - \log(\Gamma(\alpha_{kd})) - \beta_{ad}x_{id}\}\right.$$
$$\left. + (\alpha_{kd}-1)\log(p_{kd}) + \alpha_{kd}s_{kd}\log(\beta_{kd}) - r_{kd}\log(\Gamma(\alpha_{kd})) - \beta_{kd}q_{kd}\right\} = 0$$

$$\Rightarrow \frac{\partial}{\partial \alpha_{kd}}\left\{\sum_{i=1}^{n} z_{ki}\log(\pi) + \left(\log(p_{kd}) + \sum_{i=1}^{n} z_{ki}\log(x_{id})\right)(\alpha_{kd}-1) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\alpha_{kd}\log(\beta_{kd})\right.$$
$$\left. - \left(r_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log(\Gamma(\alpha_{kd})) - \left(q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}\right)\beta_{kd}\right\} = 0$$

$$\Rightarrow \frac{\partial}{\partial \alpha_{kd}}\left\{\sum_{i=1}^{n} z_{ki}\log(\pi) + \left(\log(p_{kd}) + \sum_{i=1}^{n} z_{ki}\log(x_{id})\right)(\alpha_{kd}-1)\right.$$
$$+ \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\alpha_{kd}\log\left(\alpha_{kd}\frac{s_{kd} + \sum_{i=1}^{n} z_{ki}}{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}}\right) - \left(r_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log(\Gamma(\alpha_{kd}))$$
$$\left. - \left(q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}\right)\alpha_{kd}\frac{s_{kd} + \sum_{i=1}^{n} z_{ki}}{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}}\right\} = 0$$

$$\Rightarrow \left(\log(p_{kd}) + \sum_{i=1}^{n} z_{ki}\log(x_{id})\right) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log(\alpha_{kd}) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)$$
$$+ \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log\left(\frac{s_{kd} + \sum_{i=1}^{n} z_{ki}}{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}}\right) - \left(r_{kd} + \sum_{i=1}^{n} z_{ki}\right)\Psi(\alpha_{kd}) - \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_{ki}\log(x_{id}) + \log(p_{kd}) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log(\alpha_{kd}) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log\left(\frac{s_{kd} + \sum_{i=1}^{n} z_{ki}}{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}}\right) - \left(\sum_{i=1}^{n} z_{ki} + r_{kd}\right)\Psi(\alpha_{kd}) = 0 \quad (12)$$

Equation 12 can be solved for optimal α_{kd} using Newton Raphson. We should note that if this had been the case of a single Gamma distribution, then there would have been only one maxima. However, in the present context we have a mixture, and therefore the problem is no longer convex.

$$f'(\alpha_{kd}) = \sum_{i=1}^{n} z_{ki}\log(x_{id}) + \log(p_{kd}) + \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log(\alpha_{kd}) - \left(s_{kd} + \sum_{i=1}^{n} z_{ki}\right)\log\left(\frac{q_{kd} + \sum_{i=1}^{n} z_{ki}x_{id}}{s_{kd} + \sum_{i=1}^{n} z_{ki}}\right) - \left(\sum_{i=1}^{n} z_{ki} + r_{kd}\right)\Psi(\alpha_{kd})$$

$$f''(\alpha_{kd}) = \left(\sum_{i=1}^{n} z_{ki} + s_{kd}\right)\frac{1}{\alpha_{kd}} - \left(\sum_{i=1}^{n} z_{ki} + r_{kd}\right)\Psi'(\alpha_{kd})$$

$$\alpha_{kd}{}^{new} = \alpha_{kd} - \frac{f'(\alpha_{kd})}{f''(\alpha_{kd})} \quad (13)$$

An alternative to Equation 13 was proposed by Minka using a local approximation [1]:

$$\frac{1}{\alpha_{kd}{}^{new}} = \frac{1}{\alpha_{kd}} + \frac{f'(\alpha_{kd})}{\alpha_{kd}{}^{2} f''(\alpha_{kd})}$$

E-Step: Compute:

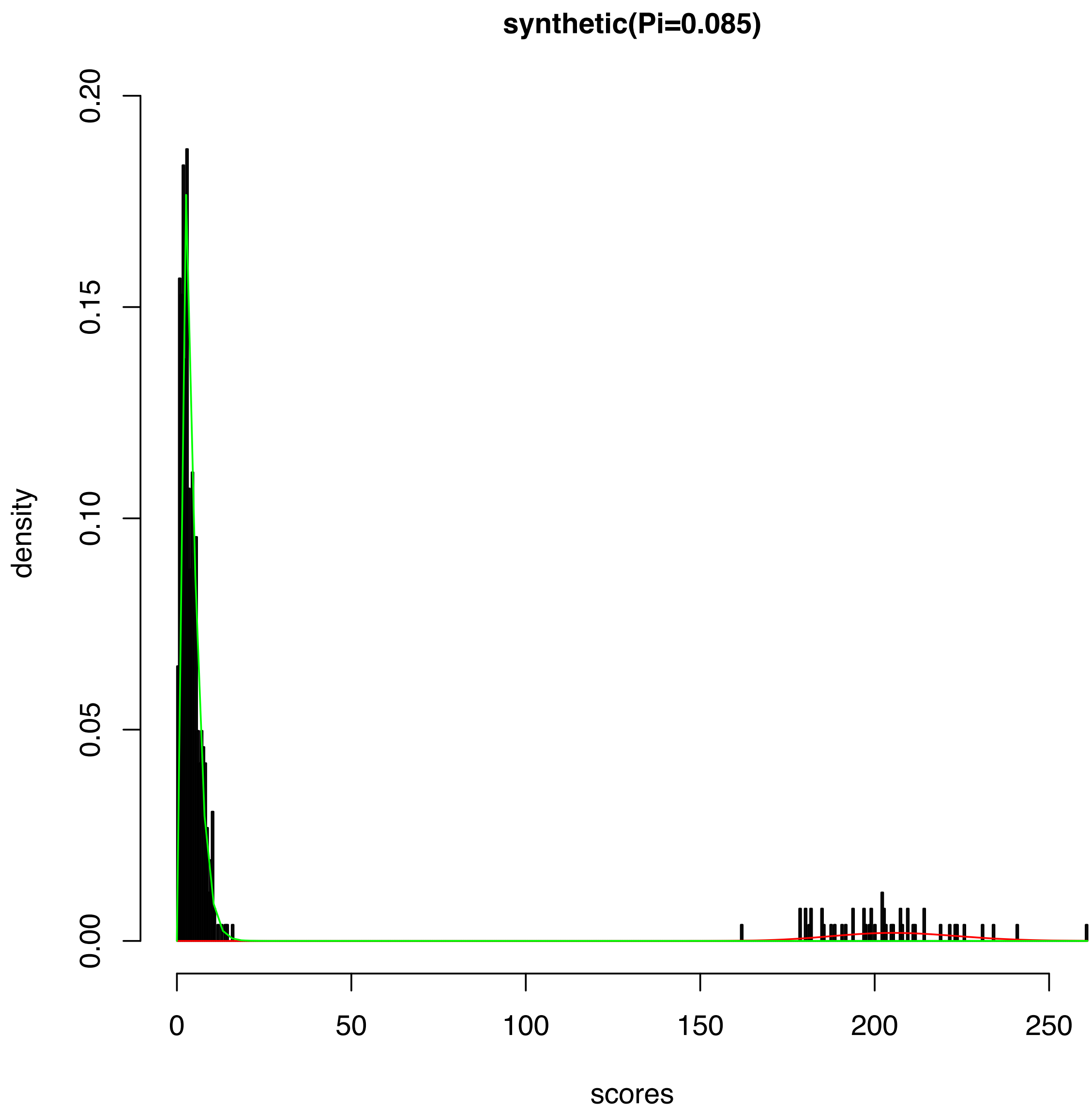
$$E[z_{ai} = 1|\boldsymbol{x}] = P(z_{ai} = 1) = \frac{\pi \prod_{d=1}^{D} f_a(x_{id}|\lambda_{ad})}{\pi \prod_{d=1}^{D} f_a(x_{id}|\lambda_{ad}) + (1-\pi)\prod_{d=1}^{D} f_r(x_{id}|\lambda_{rd})} \quad (14)$$

and,

$$E[z_{ri} = 1|\boldsymbol{x}] = P(z_{ri} = 1) = \frac{(1-\pi)\prod_{d=1}^{D} f_r(x_{id}|\lambda_{rd})}{\pi \prod_{d=1}^{D} f_a(x_{id}|\lambda_{ad}) + (1-\pi)\prod_{d=1}^{D} f_r(x_{id}|\lambda_{rd})} \quad (15)$$

Illustration of a fit on Synthetic Data

In this simple example, we assume that there is only one detector (in a real application we expect there to be more than one detector.) Hence, we are just fitting a mixture of two Gamma distributions on univariate data. The green distribution is the nominal distribution whereas red is the anomaly distribution. The estimated fraction of anomalies is 0.085.



References

[1] Thomas P. Minka. *Beyond newton's method*. research.microsoft.com/~minka/papers/newton.html, 2000.