A Bayesian Approach To Histogram Comparison

Michael Betancourt
Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

Determining if two histograms are consistent, whether they have been drawn from the same underlying distribution or not, is a common problem in physics. Existing approaches are not only limited in power but also inapplicable to histograms filled with importance weights, a common feature of Monte Carlo simulations. From a Bayesian perspective, the comparison between a single underlying distribution and two underlying distributions is readily solved within the context of model comparison. I introduce an implementation of Bayesian model comparison to the problem, including the extension to importance sampling.

1 Histogram Comparison

The histogram is a fundamental tool in physics, providing a powerful yet accessible means of non-parameteric inference. Often, however, analyses are concerned not with the underlying distribution itself but rather the consistency between multiple histograms. Comparing different analysis techniques, for example, requires the comparison of each respective analysis output given the same input data. Perhaps the most ubiquitous example is the validation of Monte Carlo simulations against data.

Although numerous approaches to the problem exist, none are without their limitations.

2 Orthodox Approaches

2.1 The Komolgorov-Smirnov Test

When comparing the consistency of two one-dimensional samples, $y_1$ and $y_2$, standard frequentist methodologies turn to significance testing, in particular the Kolmogorov-Smirnov (KS) test [1]. Here the null hypothesis that the two samples are consistent with draws from the same distribution is accepted or rejected based on the sampling distribution of the test statistic

$$q = \max_x |F_{y_1}(x) - F_{y_2}(x)|,$$

[1]betan@mit.edu
where \( F_z(x) \) is the empirical distribution of the sample \( z \),

\[
F_z(x) = \sum_{i=1}^{n} I(x_i),
\]

with \( I(x_i) \) the indicator function

\[
I(x_i) = \begin{cases} 
0, & x < x_i \\
1, & x \geq x_i 
\end{cases}
\]

In practical applications, the sampling distribution for \( q \) is usually approximated with a large \( n \) limit such that the KS test becomes extremely conservative (in more technical terms, the Type II error of the test is small only when the samples are large). Moreover, the sampling distribution of the weighted samples from importance sampling cannot be approximated so easily and the proper application of the KS test to simulated events is infeasible.

Given histograms instead of the individual samples, the empirical distributions must be approximated by the bin contents themselves and the test statistic \( q \) reduces to the maximum difference between bins. Information in all other bins is ignored, no matter now relevant it might be. Considering also the usual faults of frequentist significance tests \([2, 3, 4]\), the application of the KS test to the problem of histogram comparison leaves much to be desired.

### 2.2 Other Approaches

Various procedures have also been developed in various communities. One approach subtracts one histogram from the other before performing a significance test on the hypothesis of a constant residual across all bins. The sampling distribution of the residuals, however, does not admit familiar tests such as \( \chi^2 \) outside of the limit of large bin contents. Still, the limit is often assumed and the resulting \( \chi^2 \) taken as a measure of consistency between the two histograms.

A similar approach utilizing the quotient of the two histograms fares even worse. The usual linearized Gaussian approximations to the resulting bin quotients are not very accurate, and the subsequent \( \chi^2 \) test is even more misleading than for the bin residuals. More accurate considerations \([5]\) become considerably more involved, and significance testing much more challenging.

Both approaches are further limited by the dependence on frequentist significance testing.
3 The Bayesian Perspective

From the Bayesian perspective, the problem of histogram comparison becomes one of model comparison between $S$, the model where the two histograms are drawn from the same distribution, and $\bar{S}$, the model where they are drawn from distinct distributions. The probability of $S$ follows from an application of Bayes’ Theorem on the model evidences, $p(m, n|S)$ and $p(n, m|\bar{S})$,

\[
p(S|m, n) = \frac{p(m, n|S) p(S)}{p(m, n|S) p(S) + p(m, n|\bar{S}) p(\bar{S})},
\]

where $m$ and $n$ are the bin populations of the two histograms. In practice, a uniform prior is taken and the probability reduces to

\[
p(S|m, n) = \frac{p(m, n|S)}{p(m, n|S) + p(m, n|\bar{S})}.
\]

Subsequently, $S$ best models the data when $p(S|m, n) > 0.5$ while $\bar{S}$ is superior when $p(S|m, n) < 0.5$. When $p(S|m, n) = 0.5$, however, the interpretation is more subtle. The intermediate case can arise not only when each bin is independently ignorant, but also when individual bins are best described by different models and the discrepancies negate the respective contributions to the model evidences.

In order to discriminate between ignorance and disagreement, first note that the underlying multinomial distributions modeling the histogram bin contents are well described by independent Poisson distributions. The resulting evidences conveniently factor,

\[
p(m, n|M) = \prod_{i=1}^{N} p(m_i, n_i|M),
\]

which enables an explicit mixture model,

\[
p(m, n|\pi) = \prod_{i=1}^{N} p(m_i, n_i|\pi)
\]

\[
p(m, n|\pi) = \prod_{i=1}^{N} \left[ \pi p(m_i, n_i|S) + (1 - \pi) p(m_i, n_i|\bar{S}) \right],
\]

with the mixture posterior

\[
p(\pi|m, n) \propto p(m, n|\pi) p(\pi).
\]
In the case of a single bin and a uniform prior, the mixture posterior is linear in $\pi$ and the possible posteriors fall into three classes. Those with a negative slope are best modeled by $\bar{S}$ (Fig 1a), those uniform across $\pi$ are indifferent between the two models (Fig 1b), and those with a positive slope favor $S$ (Fig 1c).

In the general case, each bin is modeled independently. If all bins favor $\bar{S}$ then the combined product quickly concentrates towards small $\pi$ with a mode exactly at $\pi = 0$ (Fig 2a). Conversely, full support of $S$ produces a posterior concentrated towards $\pi = 1$ with a mode exactly at the boundary (Fig 2c). Any bins indifferent between the models contribute little to the final shape, unless all the bins are ignorant in which case the posterior remains uniform (Fig 2b). When bins are best modeled by different models, however, the posterior becomes peaked away from the boundaries with the exact location of the mode indicating the overall superior model (Fig 3).

Diagnosing the best model is then straightforward. When a mode arises at the boundaries, the bins are consistently described by a single model: $\bar{S}$ for $\pi = 0$ and $S$ for $\pi = 1$. A mode away from the boundaries indicates mixed agreement, implying that the two histograms are not entirely consistent.
Note that the non-mixture posterior can be recovered by comparing the boundaries of the mixture posterior,

\[ p(S|m, n) = \frac{p(\pi = 1|m, n)}{p(\pi = 0|m, n) + p(\pi = 1|m, n)} \]

The additional information contained in the shape of mixture posterior underlies the benefit of the extension to a mixture model. Not only does the mixture model break the degeneracy between ignorance and balancing disagreement, but it can also reveal more subtle disagreements that might otherwise be obscured in the model posterior.

### 3.1 Comparing Data

Actually computing the mixture posterior requires the bin by bin evidences, \( p(m_i, n_i|S) \) and \( p(m_i, n_i|\bar{S}) \).

#### 3.1.1 The Same Source Model Evidence

When both bin populations are drawn from the same distribution, the joint likelihood is given by

\[
p(m_i, n_i|\lambda, S) = p(m_i|\lambda) p(n_i|\lambda) = \frac{\lambda^{m_i} e^{-\lambda}}{m_i!} \frac{\lambda^{n_i} e^{-\lambda}}{n_i!} = \frac{\lambda^{m_i+n_i} e^{-2\lambda}}{m_i! n_i!}.
\]
Assigning a uniform prior,

\[ p(\lambda|S) = \begin{cases} (\beta - \alpha)^{-1}, & \alpha \leq \lambda \leq \beta \\ 0, & \text{else} \end{cases} , \]

yields the posterior

\[ p(\lambda|m_i, n_i, S) \propto p(m_i, n_i|\lambda, S) p(\lambda|S) \]

\[ = \frac{\lambda^{m_i+n_i}e^{-2\lambda}}{\int_{\alpha}^{\beta} d\lambda \lambda^{m_i+n_i}e^{-2\lambda}} \]

\[ = \frac{\lambda^{m_i+n_i}e^{-2\lambda}}{\left[ \frac{\Gamma(m_i+n_i+1)}{\Gamma(m_i+n_i+1,2\beta)} - \frac{\Gamma(m_i+n_i+1,2\alpha)}{\Gamma(m_i+n_i+1,2\beta)} \right]}. \]

where \( \tilde{\gamma} \) is the normalized lower incomplete gamma function,

\[ \tilde{\gamma}(n,x) = \frac{\int_{0}^{x} dt t^{n-1}e^{-t}}{\int_{0}^{\infty} dt t^{n-1}e^{-t}}. \]

The evidence follows,

\[ p(m_i, n_i|S) = \frac{p(m_i, n_i|\lambda, S) p(\lambda, S)}{p(\lambda|m_i, n_i, S)} \]

\[ = \frac{2^{-(m_i+n_i+1)} \tilde{\gamma}(m_i+n_i+1,2\beta) - \tilde{\gamma}(m_i+n_i+1,2\alpha)}{\beta - \alpha \frac{\Gamma(m_i+n_i+1,2\beta)}{\Gamma(m_i+n_i+1,2\beta)}} \]

where \( \text{Be}(x,y) \) is the Beta function,

\[ \text{Be}(x,y) = \int_{0}^{1} dt t^x (1-t)^{1-y}. \]

In practice the prior support, \( \alpha \leq \lambda \leq \beta \), is chosen to encompass the bulk of the likelihood. Making the support as small as possible while keeping the posterior normalization practically constant implements a crude form of model selection that improves the model without limiting its power.
### 3.1.2 The Two Source Model Evidence

Given two distinct sources, the joint likelihood factors into

\[ p(m_i, n_i | \lambda, \bar{S}) = p(m_i | \lambda) \cdot p(n_i | \mu) \]

\[ = \frac{\lambda^{m_i} e^{-\lambda}}{m_i!} \cdot \frac{\mu^{n_i} e^{-\mu}}{n_i!}. \]

The two sources require two independent priors,

\[ p(\lambda, \mu | \bar{S}) = p(\lambda | \bar{S}) \cdot p(\mu | \bar{S}) \]

\[ = \left\{ \begin{array}{ll}
(\delta - \gamma)^{-1} \cdot (\zeta - \epsilon)^{-1}, & \gamma \leq \lambda \leq \delta, \epsilon \leq \mu \leq \zeta, \\
0, & \text{else}
\end{array} \right. \]

and the joint posterior becomes

\[ p(\lambda, \mu | m_i, n_i, \bar{S}) = p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S}) \cdot p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S}) \cdot p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S}) \]

\[ \propto \frac{\lambda^{m_i} e^{-\lambda}}{m_i!} \cdot \frac{\mu^{n_i} e^{-\mu}}{n_i!} \int_{\gamma}^{\delta} d\lambda \lambda^{m_i} e^{-\lambda} \int_{\epsilon}^{\zeta} d\mu \mu^{n_i} e^{-\mu} \]

\[ = \frac{\Gamma (m_i + 1) [\bar{\gamma} (m_i + 1, \delta) - \bar{\gamma} (m_i + 1, \gamma)] \Gamma (n_i + 1) [\bar{\gamma} (n_i + 1, \zeta) - \bar{\gamma} (n_i + 1, \epsilon)] \Gamma (m_i + 1) [\bar{\gamma} (m_i + 1, \delta) - \bar{\gamma} (m_i + 1, \gamma)] \Gamma (n_i + 1) [\bar{\gamma} (n_i + 1, \zeta) - \bar{\gamma} (n_i + 1, \epsilon)]}{(\delta - \gamma) (\zeta - \epsilon)}. \]

The evidence is then

\[ p(m_i, n_i | \bar{S}) = \frac{p(m_i, n_i | \lambda, \mu, \bar{S}) p(\lambda, \mu | \bar{S})}{p(\lambda | m_i, n_i, \bar{S}) p(\mu | m_i, n_i, \bar{S}) \cdot p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S}) \cdot p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S}) \cdot p(\lambda | m_i, n_i, \bar{S}) \cdot p(\mu | m_i, n_i, \bar{S})} \]

\[ = \frac{\Gamma (m_i + 1) [\bar{\gamma} (m_i + 1, \delta) - \bar{\gamma} (m_i + 1, \gamma)] \Gamma (n_i + 1) [\bar{\gamma} (n_i + 1, \zeta) - \bar{\gamma} (n_i + 1, \epsilon)]}{(\delta - \gamma) (\zeta - \epsilon)}. \]
3.2 Importance Sampling

3.2.1 Sampling as Monte Carlo Estimation

Cross sections defining the physics from which the data are generated are equivalent to the probability distribution

\[ p(x) = \frac{1}{\sigma} \frac{d\sigma(x)}{dx}, \]

where \( x \) defines the phase space of the given physics.

Given an integrated luminosity \( \bar{\mathcal{L}} \), the expected number of events falling into a particular histogram bin is then

\[
\lambda = \int_a^b dx \bar{\mathcal{L}} \frac{d\sigma(x)}{dx}
= \int_a^b dx \bar{\mathcal{L}} \sigma p(x)
= \int_{-\infty}^{\infty} dx I_{ab}(x) \bar{\mathcal{L}} \sigma p(x)
= \int_{-\infty}^{\infty} dx N(x) p(x),
\]

where

\[
I_{ab}(x) = \begin{cases} 1, & a \leq x \leq b \\ 0, & \text{else} \end{cases}.
\]

Now consider a Monte Carlo simulation generating \( n = \bar{\mathcal{L}} \sigma \) events from the differential cross section \( d\sigma/dx \). The samples admit the Monte Carlo estimate of the expected counts,

\[
\tilde{N} = \frac{1}{n} \sum_{i=1}^{n} N(x_i)
= \frac{1}{n} \sum_{i=1}^{n} \bar{\mathcal{L}} \sigma I_{ab}(x_i)
= \sum_{i=1}^{n} I_{ab}(x_i),
\]

which is just the number of events falling into the bin \( a \leq x \leq b \).
The estimate is unbiased,

$$E_p[\bar{N}] = \lambda,$$

with the variance

$$\text{Var}_p[\bar{N}] = \frac{1}{n} \text{Var}_p [N(x)]$$

$$= \frac{1}{n} \left( E_p \left[ N^2(x) \right] - E_p [N(x)]^2 \right)$$

$$= \frac{1}{n} \left( \int_{-\infty}^{\infty} I^2_{ab}(x) \left( \hat{L} \sigma \right)^2 p(x) - \lambda^2 \right)$$

$$= \frac{1}{n} \left( \hat{L} \sigma \int_{-\infty}^{\infty} I_{ab}(x) \hat{L} \sigma p(x) - \lambda^2 \right)$$

$$= \frac{1}{n} \left( n\lambda - \lambda^2 \right)$$

$$= \lambda - \lambda^2/n.$$

In the large $n$ limit, the second term becomes negligible and the sampling distribution of the Monte Carlo estimate converges to a Poisson distribution,

$$p(\bar{N} | \lambda) = \frac{\lambda^{\bar{N}} e^{-\lambda}}{\bar{N}!}.$$

Note that this limit does not require large bin contents. In fact, the large $n$ limit is exactly the same rare-events limit in which the multinomial distribution reduces to independent Poisson distributions. The Monte Carlo samples, then, are entirely analogous to the data and inferring the underlying rate $\lambda$ is the same in both cases.

Importance sampling is slightly more subtle. Here events are not drawn from the physics distribution $p(x)$ but instead an auxiliary distribution $q(x)$. The expected number of events becomes

$$\lambda = \int_{-\infty}^{\infty} dx N(x) p(x)$$

$$= \int_{-\infty}^{\infty} dx N(x) \frac{p(x)}{q(x)} q(x)$$

$$\equiv \int_{-\infty}^{\infty} dx N(x) w(x) q(x)$$
with the resulting Monte Carlo estimate

\[
\tilde{N} = \frac{1}{n} \sum_{i=1}^{n} N(x_i) w(x_i) \\
= \frac{1}{n} \sum_{i=1}^{n} \tilde{\sigma} I_{ab}(x_i) w(x_i) \\
= \frac{\tilde{\sigma}}{n} \sum_{i=1}^{n} I_{ab}(x_i) w(x_i) \\
= \sum_{i=1}^{n} I_{ab}(x_i) w(x_i).
\]

Note that this estimate is just the sum of the weights of events falling into the chosen bin.

As above, the importance sampling estimate is an unbiased estimator,

\[
E_q[\tilde{N}] = \lambda,
\]

but with the variance

\[
\text{Var}_q[\tilde{N}] = \frac{1}{n} \text{Var}_q[N(x) w(x)] \\
= \frac{1}{n} \left( E_q[N^2(x) w^2(x)] - E_q[N(x) w(x)]^2 \right) \\
= \frac{1}{n} \left( \int_{-\infty}^{\infty} I_{ab}(x) (\tilde{\sigma})^2 w^2(x) q(x) - \lambda^2 \right) \\
= \frac{1}{n} \left( \tilde{\sigma} \int_{-\infty}^{\infty} I_{ab}(x) \tilde{\sigma} w(x) p(x) - \lambda^2 \right) \\
= \frac{1}{n} \left( n \int_{-\infty}^{\infty} I_{ab}(x) \tilde{\sigma} w(x) p(x) - \lambda^2 \right) \\
= \int_{-\infty}^{\infty} I_{ab}(x) \tilde{\sigma} w(x) p(x) - \lambda^2 / n.
\]

If the weighting function \( w(x) \) is relatively constant across the bin \( a \leq x \leq b \) then \( w(x) \) can be approximated by its ensemble average,

\[
\bar{w} = \frac{\sum_{i=1}^{n} I_{ab}(x_i) w(x_i)}{\sum_{i=1}^{n} I_{ab}(x_i)}
\]
and the remaining integral becomes

\[
\int_{-\infty}^{\infty} I_{ab}(x) \bar{L} \sigma_w(x)p(x) \approx \int_{-\infty}^{\infty} I_{ab}(x) \bar{L} \sigma \bar{w}p(x) \\
\approx \bar{w} \int_{-\infty}^{\infty} I_{ab}(x) \bar{L} \sigma p(x) \\
\approx \bar{w} \lambda.
\]

The variance is then

\[
\text{Var}_q[\tilde{N}] \approx \bar{w} \lambda - \lambda^2/n,
\]

and in the large \(n\) limit reduces to

\[
\text{Var}_q[\tilde{N}] \approx \bar{w} \lambda.
\]

Note that if the weights are unity, in which case the importance sampling reduces to standard Monte Carlo sampling, then the variance reduces to the previous result as consistency would demand.

The resulting sampling distribution is approximately Gaussian,

\[
p(\tilde{N}|\lambda) = \frac{1}{\sqrt{2\pi\bar{w}\lambda}} \exp\left[ -\frac{1}{2} \left( \frac{\tilde{N} - \lambda}{\bar{w}\lambda} \right)^2 \right].
\]

The above sampling distribution admits a Bayesian approach to simulated histograms and consequently the comparison of simulated histograms. Computation of the evidence ratio follows as in the case of data comparison.

### 3.3 Comparing Data to Simulation

The importance sampling distribution for the simulated histogram bins requires not only the sum of the weights falling into that bin,

\[
n_i = \sum_{j=1}^{n} I_{a_ib_i}(x_j)w(x_j),
\]

but also the ensemble average of the weights,

\[
\bar{w}_i = \frac{\sum_{j=1}^{n} I_{a_ib_i}(x_j)w(x_j)}{\sum_{j=1}^{n} I_{a_ib_i}(x_j)}.
\]

As before, the sampling distribution of the data histograms bins \(m_i\) is taken to be Poisson.
3.3.1 The Same Source Model Evidence

Here the joint likelihood becomes

\[
p (m, n | \lambda, S) = p (m | \lambda) p (n | \lambda)
\]

\[
= \frac{\lambda^m e^{-\lambda}}{m!} \cdot \frac{1}{\sqrt{2\pi \bar{w}\lambda}} \exp \left[ -\frac{1}{2} \left( \frac{n - \lambda}{w\lambda} \right)^2 \right].
\]

A uniform prior,

\[
p (\lambda | S) = \begin{cases} 
(\beta - \alpha)^{-1}, & \alpha \leq \lambda \leq \beta \\
0, & \text{else}
\end{cases},
\]

gives the posterior

\[
p (\lambda | m, n, S) \propto p (m, n | \lambda, S) p (\lambda | S)
\]

\[
= \frac{\lambda^m - \frac{1}{2} \exp \left[ \lambda - \frac{1}{2} \left( \frac{n - \lambda}{\bar{w}\lambda} \right)^2 \right]}{\int_\alpha^\beta d\lambda \lambda^m - \frac{1}{2} \exp \left[ \lambda - \frac{1}{2} \left( \frac{n - \lambda}{\bar{w}\lambda} \right)^2 \right]},
\]

resulting in the evidence

\[
p (m, n | S) = \frac{1}{m! (\beta - \alpha) \sqrt{2\pi \bar{w}}} \int_\alpha^\beta d\lambda \lambda^m - \frac{1}{2} \exp \left[ \lambda - \frac{1}{2} \left( \frac{n - \lambda}{\bar{w}\lambda} \right)^2 \right].
\]

Note that limiting the prior support not only improves the model but also eases the computational demands of the required numerical integration.

3.3.2 The Two Source Model Evidence

Given two distinct sources, the joint likelihood factors into

\[
p (m, n | \lambda, \bar{S}) = p (m | \lambda) \cdot p (n | \mu)
\]

\[
= \frac{\lambda^m e^{-\lambda}}{m!} \cdot \frac{1}{\sqrt{2\pi \bar{w}\lambda}} \exp \left[ -\frac{1}{2} \left( \frac{n \mu}{\bar{w}\lambda} \right)^2 \right].
\]
The two sources require two independent priors,

\[
p(\lambda, \mu | S) = p(\lambda | S) \cdot p(\mu | S) = \begin{cases} (\delta - \gamma)^{-1} \cdot (\zeta - \epsilon)^{-1}, & \gamma \leq \lambda \leq \delta, \epsilon \leq \mu \leq \zeta, \\ 0, & \text{else} \end{cases}
\]

and the joint posterior becomes

\[
p(\lambda, \mu | m_i, n_i, S) = p(\lambda | m_i, n_i, S) \cdot p(\mu | m_i, n_i, S)
\approx p(m_i, n_i | \lambda, S) \cdot p(\lambda | S) \cdot p(m_i, n_i | \mu, S) \cdot p(\mu | S)
\]

\[
= \frac{\lambda^m e^{-\lambda}}{\Gamma(m_i + 1)} \cdot \frac{\mu^{i/2} \exp\left[-\frac{1}{2} \frac{(n_i - \mu)^2}{\bar{w}_i \mu}\right]}{\int_\gamma^\delta d\lambda \lambda^{m_i} e^{-\lambda} \int_\epsilon^\zeta d\mu \mu^{i/2} \exp\left[-\frac{1}{2} \frac{(n_i - \mu)^2}{\bar{w}_i \mu}\right]}
\]

yielding the evidence

\[
p(m_i, n_i | S) = \frac{p(m_i, n_i | \lambda, \mu, S) \cdot p(\lambda, \mu, S)}{p(\lambda, \mu | m_i, n_i, S)} = \frac{p(m_i, n_i | \lambda, S) \cdot p(\lambda | S) \cdot p(m_i, n_i | \mu, S) \cdot p(\mu | S)}{p(\lambda | m_i, n_i, S) \cdot p(\mu | m_i, n_i, S)}
\]

\[
= \frac{(\delta - \gamma)^{-1} (\zeta - \epsilon)^{-1}}{m_i! \sqrt{2 \pi \bar{w}_i}} \cdot \Gamma(m_i + 1) \cdot [\bar{\gamma}(m_i + 1, \delta) - \bar{\gamma}(m_i + 1, \gamma)] \cdot \int_\epsilon^\zeta d\mu \mu^{-i/2} \exp\left[-\frac{1}{2} \frac{(n_i - \mu)^2}{\bar{w}_i \mu}\right]
\]

### 3.4 Comparing Simulation to Simulation

Comparing two simulated histogram bins requires the sum of the weights and ensemble average for both the first bin,

\[
m_i = \sum_{j=1}^n I_{a_i b_i}(x_j) \cdot w(x_j)
\]

\[
\bar{w}_i = \frac{\sum_{j=1}^n I_{a_i b_i}(x_j) \cdot w(x_j)}{\sum_{j=1}^n I_{a_i b_i}(x_j)}
\]
and the second bin,
\[ n_i = \sum_{j=1}^{n} I_{a,b_i}(x_j)v(x_j) \]
\[ \bar{v}_i = \frac{\sum_{j=1}^{n} I_{a,b_i}(x_j)v(x_j)}{\sum_{j=1}^{n} I_{a,b_i}(x_j)}. \]

3.4.1 The Same Source Model Evidence

In this case the joint likelihood becomes
\[
p(m, n_i|\lambda, S) = p(m_i|\lambda)p(n_i|\lambda) = \frac{1}{\sqrt{2\pi w_i \lambda}} \exp\left[-\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda}\right] \frac{1}{\sqrt{2\pi \bar{v}_i \lambda}} \exp\left[-\frac{1}{2} \frac{(n_i - \lambda)^2}{\bar{v}_i \lambda}\right]
\]
\[= \frac{1}{2\pi \lambda \sqrt{w_i \bar{v}_i}} \exp\left[-\frac{1}{2\lambda} \left(\frac{(m_i - \lambda)^2}{w_i} - \frac{(n_i - \lambda)^2}{\bar{v}_i}\right)\right] \]

A uniform prior,
\[ p(\lambda|S) = \begin{cases} (\beta - \alpha)^{-1}, & \alpha \leq \lambda \leq \beta \\ 0, & \text{else} \end{cases}, \]
gives the posterior
\[
p(\lambda|m, n_i, S) \propto p(m, n_i|\lambda, S)p(\lambda|S) \]
\[= \lambda^{-1} \exp\left[-\frac{1}{2\lambda} \left(\frac{(m_i - \lambda)^2}{w_i} - \frac{(n_i - \lambda)^2}{\bar{v}_i}\right)\right]\]
\[\int_{\alpha}^{\beta} d\lambda \lambda^{-1} \exp\left[-\frac{1}{2\lambda} \left(\frac{(m_i - \lambda)^2}{w_i} - \frac{(n_i - \lambda)^2}{\bar{v}_i}\right)\right],\]
with the evidence
\[
p(m, n_i|S) = \frac{1}{2\pi (\beta - \alpha) \sqrt{w_i \bar{v}_i}} \int_{\alpha}^{\beta} d\lambda \lambda^{-1} \exp\left[-\frac{1}{2\lambda} \left(\frac{(m_i - \lambda)^2}{w_i} - \frac{(n_i - \lambda)^2}{\bar{v}_i}\right)\right]
\]
3.4.2 The Two Source Model Evidence

With two distinct sources the joint likelihood factors into

\[ p(m_i, n_i | \lambda, S) = p(m_i | \lambda) \cdot p(n_i | \mu) \]

\[ = \frac{1}{\sqrt{2\pi w_i \lambda}} \exp \left[ -\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda} \right] \cdot \frac{1}{\sqrt{2\pi v_i \mu}} \exp \left[ -\frac{1}{2} \frac{(n_i - \mu)^2}{v_i \mu} \right] \]

\[ = \frac{1}{2\pi \sqrt{w_i v_i}} \lambda^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda} \right] \cdot \mu^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(n_i - \mu)^2}{v_i \mu} \right]. \]

Taking two independent priors,

\[ p(\lambda, \mu | S) = p(\lambda | S) \cdot p(\mu | S) \]

\[ = \left\{ \begin{array}{ll}
(\delta - \gamma)^{1} \cdot (\zeta - \epsilon)^{-1}, & \gamma \leq \lambda \leq \delta, \epsilon \leq \mu \leq \zeta, \\
0, & \text{else}
\end{array} \right. \]

gives the posterior

\[ p(\lambda, \mu | m_i, n_i, S) = p(\lambda | m_i, n_i, S) \cdot p(\mu | m_i, n_i, S) \]

\[ \propto p(m_i, n_i | \lambda, S) \cdot p(\lambda | S) \cdot p(m_i, n_i | \mu, S) \cdot p(\mu | S) \]

\[ = \lambda^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda} \right] \cdot \mu^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(n_i - \mu)^2}{v_i \mu} \right] \cdot \int_{\gamma}^{\delta} d\lambda \lambda^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda} \right] \cdot \int_{\epsilon}^{\zeta} d\mu \mu^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(n_i - \mu)^2}{v_i \mu} \right], \]

along with the evidence

\[ p(m_i, n_i | S) = \frac{p(m_i, n_i | \lambda, \mu, S) \cdot p(\lambda, \mu, S)}{p(\lambda, \mu | m_i, n_i, S)} \]

\[ = \frac{p(m_i, n_i | \lambda, S) \cdot p(\lambda | S) \cdot p(m_i, n_i | \mu, S) \cdot p(\mu | S)}{p(\lambda | m_i, n_i, S) \cdot p(\mu | m_i, n_i, S)} \]

\[ = \frac{(\delta - \gamma)^{1} \cdot (\zeta - \epsilon)^{-1}}{2\pi \sqrt{w_i v_i}} \int_{\gamma}^{\delta} d\lambda \lambda^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(m_i - \lambda)^2}{w_i \lambda} \right] \cdot \int_{\epsilon}^{\zeta} d\mu \mu^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{(n_i - \mu)^2}{v_i \mu} \right]. \]
4 Examples

Figure 4: Comparison of two histograms drawn from different distributions.

Figure 5: Comparison of two histograms drawn from the same distribution, with the red sample generated from importance sampling.

Figure 6: Comparison of two histograms drawn from the same distribution, with both samples generated from importance sampling.
5 Model Ensemble Tests

The capability of the Bayesian approach is evident when considering an ensemble of histogram comparisons models. At each iteration in the ensemble a true model is randomly selected between $S$ and $\bar{S}$. In the former case, the two histograms to be compared are drawn from a randomly generated multinomial distribution with 50 bins,

$$p_i = \frac{\tilde{p}_i}{\sum_j \tilde{p}_j},$$
$$\tilde{p}_i \sim U(0, 1).$$

The latter case samples one histogram from the randomly generated multinomial distribution, but draws the second from a perturbed distribution with

$$q_i = \frac{\tilde{q}_i}{\sum_j \tilde{q}_j},$$
$$\tilde{q}_i = \left(1 + \frac{2}{3} \cdot \left( r_i - \frac{1}{2} \right) \right) p_i,$$
$$r_i \sim U(0, 1).$$

The two distributions are kept similar in order to assess the performance of each algorithm when comparison is highly nontrivial.

The two histograms generated for each model are then evaluated with three tests:

- **Non-Mixture Posterior**: $p(S|m, n) > 0.5$
- **Mixture MAP**: $\pi_{MAP} > 0.97$
- **Kolmogorov-Smirnov**: A 95% significance KS test implemented in the ROOT physics library

False acceptances and false rejections were tabulated for four ensembles, each with 2000 independent models but different total bin contents $n$. The respective rate posteriors were calculated assuming a binomial likelihood and beta prior, $B(0.5, 0.5)$; the posterior modes and 68.3% confidence intervals for each algorithm are plotted in Figure 7.

Note that in the two cases with the smallest statistics, $n = 5,000$ and $n = 10,000$ the variance of the bin contents equals or surpasses the expected difference between the two underlying distributions and the large false accept rates are to be expected.

In all cases the Bayesian mixture test, $\pi_{MAP} > 0.97$, is superior to KS. The non-mixture test outperforms KS in almost all cases as well, failing only in the small statistics ensembles where small discrepancies tend to be washed out by the bin content variances.
Figure 7: (a) False reject and false accept rate for each comparison algorithm.

(18)
6 Conclusion

A Bayesian mixture model has been developed to test whether two histograms are consistent, in that they are more likely to have been drawn from a single distribution rather than two distinct distributions. The model is extended to handle histograms generated from importance sampling, resulting in a robust and powerful approach to the comparison of histograms populated by both data and simulation. Said power is evident with studies of a large model ensemble.

A C++ implementation of the Bayesian mixture model utilizing the ROOT [6] data analysis framework is available at http://web.mit.edu/~betan/www/code.html.

7 Acknowledgements

I thank Chris Jones, Steve Voinea, and Matt Walker for helpful discussion and comments.

References

[1] Papoulis, A. and Pillai, S. U. (2002) Probability, Random Variables, and Stochastic Variables. McGraw-Hill, New York

[2] Bernardo, J. M. and Smith, A. F. M. (2000) Bayesian Theory. John Wiley & Sons, Inc., New York

[3] Jaynes, E. T. (2003) Probability Theory: The Logic of Science, Cambridge University Press, New York

[4] MacKay, D. J. C. (2003) Information Theory, Inference, and Learning Algorithms. Cambridge University Press, New York

[5] Hinkley, D. V. (1969) Biometrika 56(3) 635-639

[6] Brun, R. and Radermakers, F. (1997) Nucl. Instr. & Meth. A 389(1-2) 81-86. See also http://root.cern.ch/