Bayesian model comparison for one-dimensional correlations in 200GeV AuAu collisions

Hans C. Eggers[1,2] Michiel B. de Kock[1] and Thomas A. Trainor[3]

[1] *Department of Physics, Stellenbosch University, 7600 Stellenbosch, South Africa*
[2] *National Institute for Theoretical Physics (NITheP), 7600 Stellenbosch, South Africa*
[3] *CENPA 354290, University of Washington, Seattle, Washington 98195, United States*

Abstract. In the context of data modeling and comparisons between different fit models, Bayesian analysis calls that model best which has the largest evidence, the prior-weighted integral over model parameters of the likelihood function. Evidence calculations automatically take into account both the usual chi-squared measure and an Occam factor which quantifies the price for adding extra parameters. Applying Bayesian analysis to projections onto azimuth of 2D angular correlations from 200 GeV AuAu collisions, we consider typical model choices including Fourier series and a Gaussian plus combinations of individual cosine components. We find that models including a Gaussian component are consistently preferred over pure Fourier-series parametrizations, sometimes strongly so. For 0–5% central collisions the Gaussian-plus-dipole model performs better than Fourier Series models or any other combination of Gaussian-plus-multipoles.

1 Introduction

A significant problem has emerged in the interpretation of high-energy collision data from the relativistic heavy ion collider (RHIC) and the large hadron collider (LHC). Two classes of data models with divergent physics implications are invoked to support competing narratives. One narrative interprets results with concepts from high-energy physics in which the essential phenomenon is dijet production; the other uses quark-gluon plasma and collective flow concepts in which the essential phenomenon is a flowing, dense medium while dijets play only a subordinate role.

To place evaluation and comparison of competing models on a scientific footing we require a formally consistent and statistically sound framework. We suggest that Bayesian inference provides the formal consistency and the necessary mathematical framework for representing prior knowledge, evaluating candidate models for different parameter values and thereby establishing value judgments on models as a whole. In this contribution, we show by example how the Bayesian framework successfully evaluates and compares models for that type of one-dimensional correlation data which has played a large role in claimed support for these narratives. The exercise also serves to pinpoint the weaknesses of the conventional minimum-χ^2 and maximum-likelihood approaches.

2 Data and parametrization

The data analyzed below were published by the STAR Collaboration [1] in the form of two-dimensional binned histograms derived from 1.2 million 200GeV AuAu collision events in eleven centrality classes. The relevant autocorrelations

$$A(\phi_\Delta, \eta_\Delta) = \Delta\rho / \sqrt{\rho_{\text{ref}}} = \rho_0 \left[(\rho/\rho_{\text{ref}}) - 1 \right] \tag{1}$$

are measured on azimuthal and pseudorapidity difference variables $\phi_\Delta = \phi_1 - \phi_2$ and $\eta_\Delta = \eta_1 - \eta_2$, pair densities $\rho(\phi_\Delta, \eta_\Delta)$ and $\rho_{\text{ref}}(\phi_\Delta, \eta_\Delta)$ and the one-particle differential cross section, $\rho_0 = d^2 n_{ch}/d\eta d\phi$.

Here, we consider one-dimensional projections of the 2D autocorrelations onto azimuth difference ϕ_Δ with 24 bins of size $\pi/12$. As the data are symmetrized on both η_Δ and ϕ_Δ, there are only 13 unique points. These are reduced to $N = 11$ data points by excluding from fits the $\phi_\Delta = 0$ bin containing the Bose-Einstein peak and constraining the 1D projections to sum to zero.

Since the 11-parameter 2D parametrization (or data model) described in Ref. [1] describes the corresponding 2D data rather well, we adapted it to the present 1D projection. Taking into account the zero-sum constraint, this results in what we call the Basic Model; abbreviating $\phi \equiv \phi_\Delta$,

$$f(\phi, \boldsymbol{w} \,|\, \mathcal{H}_{\text{BM}}) = A_{1D} \left\{ \exp\left[-\frac{1}{2} \left(\frac{\phi}{\sigma_\phi} \right)^2 \right] - \frac{\sigma_\phi}{\sqrt{2\pi}} \right\} - A'_{\text{D}} \, 2\cos(\phi) \tag{2}$$

with three parameters $\boldsymbol{w} = (A_{1D}, \sigma_\phi, A'_{\text{D}})$ reflecting the basic same-side (SS) peak and away-side (AS) dipole structure. The Basic Model $\mathcal{H}_{\text{BM}}$ is supplemented in Augmented Models $\mathcal{H}_{\text{AM}}$ by a quadrupole term $-A'_{\text{Q}} \, 2\cos(2\phi)$, a sextupole $-A'_{\text{S}} \, 2\cos(3\phi)$, and an octupole $-A'_{\text{O}} \, 2\cos(4\phi)$ in various combinations. The factor 2 entering the cosine terms is designed to reflect a corresponding Fourier Series (FS) parametrization popular within the collective-flow paradigm [2],

$$f(\phi, \boldsymbol{w} \,|\, \mathcal{H}_{\text{FS}}) = 2 \sum_{j=1}^{K} w_j \cos(j\phi). \tag{3}$$

In our analysis, the number of free parameters K varies from 1 to 11 in the FS case, while in the Basic and Augmented Models K varies from 3 to 6.

3 Model comparison

The two goals of fitting a parametrization to data are model comparison and, given a particular model $\mathcal{H}$, finding the best-fit values of its parameters. In the present case, the data $\boldsymbol{D} = (y_1, \ldots, y_N)$ are given by binwise autocorrelations $A(\phi)$ at mid-bin sample points ϕ_n while the competing models include $\mathcal{H}_{\text{BM}}$, $\mathcal{H}_{\text{AM}}$ and $\mathcal{H}_{\text{FS}}$ with various values of K.

In the usual **frequentist approach**, model comparison is effected by searching the space of parameters $\boldsymbol{w}$ to find the global minimum χ_0^2 of

$$\chi^2 = \sum_{n=1}^{N} \left(\frac{y_n - f(\phi_n, \boldsymbol{w} \,|\, \mathcal{H})}{\sigma_n} \right)^2 \tag{4}$$

with σ_n the experimental standard errors. Model $\mathcal{H}_1$ with K_1 parameters is deemed to be better than Model $\mathcal{H}_2$ with K_2 parameters if $\min(\chi_1^2)/(N - K_1) < \min(\chi_2^2)/(N - K_2)$. Underlying the χ^2 criterion is the assumption of independent Gaussian fluctuations of deviations between data and parametrization,

$$y_n - f(\phi_n, \boldsymbol{w} \,|\, \mathcal{H}) = \varepsilon_n \qquad \text{with} \qquad p(\varepsilon_n) = \text{Gaussian with standard deviation } \sigma_n \tag{5}$$

which results in a likelihood

$$p(\boldsymbol{D} \,|\, \boldsymbol{w}, \mathcal{H}) = \frac{e^{-\chi^2/2}}{(2\pi)^{N/2} \prod_n \sigma_n}. \tag{6}$$

When parametrizations $f(\phi, \boldsymbol{w} \,|\, \mathcal{H})$ are linear in $\boldsymbol{w}$, the Hessian

$$\mathbb{H}_{k\ell} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial w_k \, \partial w_\ell} \tag{7}$$

is $\boldsymbol{w}$-independent and the likelihood takes on a Gaussian form in parameter space around the mode $\tilde{\boldsymbol{w}}$ in terms of $\mathbb{H}$ or equivalently the covariance matrix $\mathbb{C} = \mathbb{H}^{-1}$.

Bayesian model comparison utilizes the same likelihood (6) and hence also χ^2 and the covariance matrix. Crucially, however, Bayesian model comparison is based not on the maximum likelihood or χ_0^2 but on the so-called *evidence* for data $\boldsymbol{D}$ within model $\mathcal{H}$,

$$p(\boldsymbol{D} \,|\, \mathcal{H}) = \int d\boldsymbol{w} \, p(\boldsymbol{D} \,|\, \boldsymbol{w}, \mathcal{H}) \, p(\boldsymbol{w} \,|\, \mathcal{H}), \tag{8}$$

which averages the likelihood over all values of $\boldsymbol{w}$, weighted by the *prior* $p(\boldsymbol{w} \,|\, \mathcal{H})$ for the parameters.

The reason why evidence is a good measure for model comparison lies within Bayes' Theorem itself. Starting with the joint data-parameter probability $p(\boldsymbol{D}, \boldsymbol{w} \,|\, \mathcal{H}) = p(\boldsymbol{D} \,|\, \boldsymbol{w}, \mathcal{H}) \, p(\boldsymbol{w} \,|\, \mathcal{H}) = p(\boldsymbol{w} \,|\, \boldsymbol{D}, \mathcal{H}) \, p(\boldsymbol{D} \,|\, \mathcal{H})$, Bayes' Theorem states that, once data $\boldsymbol{D}$ is known, the posterior probability for parameters $\boldsymbol{w}$ can be found from the likelihood, prior and evidence by

$$p(\boldsymbol{w} \,|\, \boldsymbol{D}, \mathcal{H}) = \frac{p(\boldsymbol{D} \,|\, \boldsymbol{w}, \mathcal{H}) \, p(\boldsymbol{w} \,|\, \mathcal{H})}{p(\boldsymbol{D} \,|\, \mathcal{H})} \qquad \text{(posterior)} = \frac{\text{(likelihood)} \, \text{(prior)}}{\text{(evidence)}}. \tag{9}$$

Similarly on a higher level $p(\boldsymbol{D}, \mathcal{H}) = p(\boldsymbol{D} \,|\, \mathcal{H}) \, p(\mathcal{H}) = p(\mathcal{H} \,|\, \boldsymbol{D}) \, p(\boldsymbol{D})$ implies that, given data $\boldsymbol{D}$, the probability for the entire model $\mathcal{H}$ including all possible parameter values is

$$p(\mathcal{H} \,|\, \boldsymbol{D}) = \frac{p(\boldsymbol{D} \,|\, \mathcal{H}) \, p(\mathcal{H})}{p(\boldsymbol{D})} \tag{10}$$

which, used twice for competing models $\mathcal{H}_1$ and $\mathcal{H}_2$, says that model comparison given data $\boldsymbol{D}$ can be expressed as a ratio of evidences,

$$\frac{p(\mathcal{H}_1|\boldsymbol{D})}{p(\mathcal{H}_2|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|\mathcal{H}_1)}{p(\boldsymbol{D}|\mathcal{H}_2)} \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)} = \frac{p(\boldsymbol{D}|\mathcal{H}_1)}{p(\boldsymbol{D}|\mathcal{H}_2)} = \frac{\text{evidence for } \mathcal{H}_1}{\text{evidence for } \mathcal{H}_2} \tag{11}$$

on setting equal *a priori* probabilities for models, $p(\mathcal{H}_1) = p(\mathcal{H}_2) = 1/2$.

The choice of **parameter prior** $p(\boldsymbol{w} \,|\, \mathcal{H})$ is governed by available information or lack thereof. We use uniform priors, implying indifference for any value of $\boldsymbol{w}$ within the bounds set by cutoffs or prior widths Δ_k,

$$p(\boldsymbol{w}|\mathcal{H}) = \prod_{k=1}^{K} \Delta_k^{-1}. \tag{12}$$

We assign the same widths across different models for corresponding quantities. In particular, we set $\Delta_k = 1/3$ for all $k = 1, \dots, K$ cosine coefficients, based on knowledge of typical coefficient magnitude. The interval for σ_ϕ is set to $[0, \pi/2]$ in accordance with its definition as a same-side peak width. To the prior for the Gaussian amplitude parameter A'_{D} we assign width $\Delta = 1$ since correlation-structure amplitudes such as peak-to-peak excursions are generally $< O(1)$.
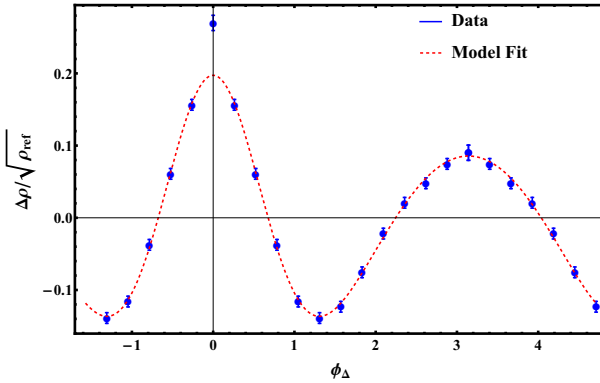
Figure 1. (Color online) Autocorrelation data projected onto one dimension for the most central events in 200 GeV AuAu collisions measured by STAR [1] along with the Basic Model best fit. Corresponding Fourier-Series fits with four or more terms appear identical on the scale of the figure. Eyeball viewing is not a reliable basis for model comparison.

For uniform priors, the evidence integrals (8) can be carried out analytically in terms of the **Laplace approximation** which neglects third- and higher-order terms in the Taylor expansion of the log likelihood and is exact for linear systems. In our case, the log likelihood is nonlinear only in the parameter σ_ϕ. The Laplace Approximation also requires that the effect of finite integration limits set by the Δ_k be negligible. Within these constraints, the evidence takes on a simple form in terms of the maximum likelihood, prior and covariance matrix,

$$p(\boldsymbol{D}|\mathcal{H}) = p(\boldsymbol{D}|\tilde{\boldsymbol{w}}, \mathcal{H}) \, p(\tilde{\boldsymbol{w}}\,|\,\mathcal{H}) \, (2\pi)^{K/2} \, \sqrt{\det \mathbb{C}} \,. \tag{13}$$

Since χ_0^2 is related to the maximum likelihood by $\chi_0^2 = -2 \log p(\boldsymbol{D}\,|\,\tilde{\boldsymbol{w}}, \mathcal{H})$, we can write the corresponding log evidence in terms of χ_0^2 and an *information* $\mathcal{I}$,

$$-2\mathrm{LE} = \chi_0^2 + 2\mathcal{I}, \tag{14}$$

$$\mathcal{I} = -\log\Big[p(\tilde{\boldsymbol{w}}|\mathcal{H}) \, (2\pi)^{K/2} \, \sqrt{\det \mathbb{C}} \, \Big]. \tag{15}$$

The smaller $-2\mathrm{LE}$ for a given model, the larger its evidence. As shown in Eq. (11), $-2\mathrm{LE}$ for different models have no absolute meaning but must always be interpreted in terms of evidence ratios (also called *odds* in the literature) or of log differences.

4 Results

In this conference summary, we can show results only for the STAR "Bin-10" data covering the most-central 0–5% collision events; further results for Bin-0 peripheral collisions (83–94%) and Bin-8 intermediate collisions (9–18%) are presented and discussed in Ref. [3].

In Figure 1, the projected 1D data along with the best-evidence fit for the Basic Model (2) is shown. Experimental error bars were scaled up by a factor 2 to make them visible in the plot. As stated, the point at $\phi = 0$ was excluded since it contains the Bose-Einstein peak and charge-neutral electron pairs from photoconversions. The best-fit parameter values are $A_{1D} = 0.57 \pm 0.007$, $A'_D = 0.115 \pm 0.002$ and $\sigma_\phi = 0.635 \pm 0.007$ while $\chi_0^2 = 12.5$ for $11 - 3$ degrees of freedom. A fit of the Fourier Series model (3) with $K \geq 4$ is indistinguishable to the eye on the scale of the plot. That does not mean that the BM and FS models have the same $-2\mathrm{LE}$.

Figure 2 illustrates for the example of Fourier-Series models why model comparison based on minimum χ^2 alone is misleading when the number of data points is small, as is the case here since $N = 11$. Plotted are various quantities as a function of the number of free parameters K in Eq. (3).
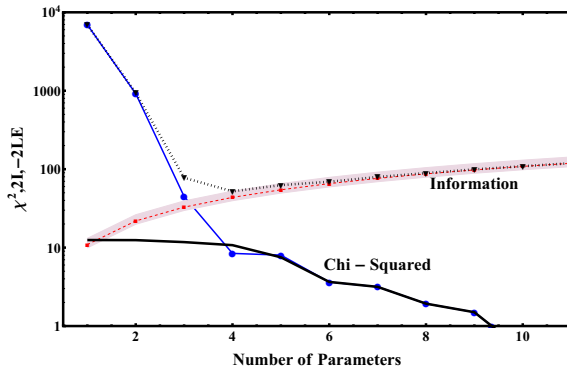
Figure 2. (Color online) Soid blue line: the usual minimum χ_0^2, which becomes identical with the χ_0^2 of the noise for $K \geq 5$ shown as the black solid line. Red dashed line and band: Information $2\mathcal{I}$ plus possible uncertainty due to the choice of prior. The negative log-evidence -2LE, the sum of χ_0^2 and $2\mathcal{I}$, is shown as the dashed black line. It has a clear minimum at $K = 4$.

The solid blue line traces the evolution of χ_0^2 as more cosine terms are added. As expected, there is a dramatic decrease in χ_0^2 up to $K = 4$ but little benefit for adding further terms beyond that as $K \geq 5$ Fourier Series models increasingly fit the noise represented by the lower solid curve showing fits to the residuals (data minus Basic Model). The naive criterion of minimum χ_0^2 is oblivious to this effect: χ_0^2 keeps decreasing and purely χ_0^2-based model comparison would hence select $K = 9$ or higher. Note that the equivalent plot for $\chi_0^2/(N - K)$ (not shown) barely changes the general decrease with K.

The contribution of information $2\mathcal{I}$ of Eq. (15) is shown as the red dashed curve. The shaded band shows the sensitivity to the $\Delta_k = 1/3$ prior width, using $\Delta_k = 1$ and $\Delta_k = 1/5$ as extremes. The resulting log-evidence -2LE of Eq. (14), shown as the black dashed curve, has a clear minimum at $K = 4$, confirming the above conclusion that FS models with $K \geq 5$ are not preferred.

In Figure 3(a), the scope of comparison between pure χ_0^2 and evidence-based model comparison is widened to include FS, Basic Model and Augmented Models with various combinations of higher-order multipoles. Consideration of χ_0^2 alone as in Fig. 3(a) would make the $K \geq 6$ FS models the winner. By contrast, the Occam penalty of $2\mathcal{I}$ forming part of -2LE in Figure 3(b) effects a dramatic re-ordering: the 3-parameter Basic Model has substantially smaller -2LE (larger evidence) than all other combinations. In addition, FS models are universally rejected in favour of Gaussian-plus-multipole models with the same number of parameters.
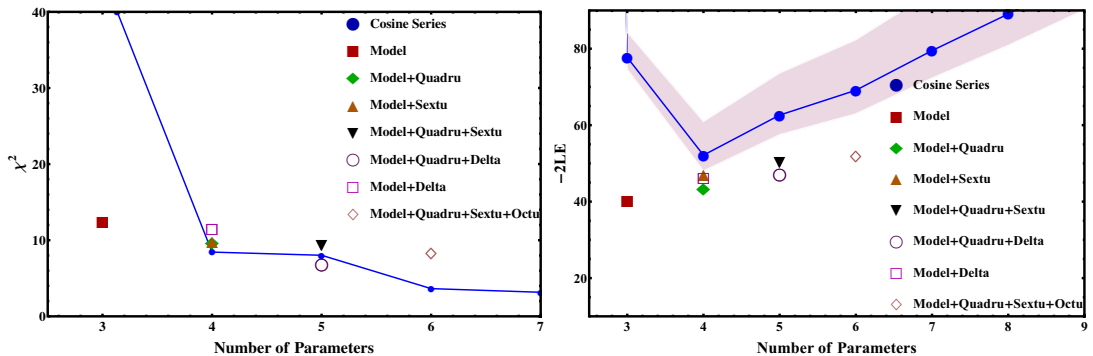

Figure 3. (Color online) (a) Minimum-χ^2 comparison of FS models (solid blue line) with Gaussian plus various combinations of multipoles. (b) The same models assessed with the Bayesian log-evidence criterion. The 3-parameter Basic Model of Eq. (2) (filled square) is the clear winner. Note that -2LE is a logarithmic scale.

5 Conclusions

Physics interpretation: As shown, the evidence for a simple same-side Gaussian plus away-side dipole is larger than for any Fourier Series-based model. While our approach and results must naturally be tested and confirmed with other data, measurement quantities and models, our results do cast serious doubt on the interpretation of azimuthal correlations as indicators of collective flow.

Lessons for data analysis: Physics conclusions must rely on the integrity of the underlying data analysis methods and assumptions. Below, we provide in tabular form a motivation why Bayesian evidence-based model comparison is superior to simple χ^2 fitting. The sometimes very different conclusions shown above suggest that any physical interpretation based on simple χ^2 fitting of one-dimensional correlation data may be debatable or outright wrong. We secondly emphasize that no model or "goodness of fit" has any meaning unless it is compared to another competing model.

Traditional use of χ^2	**Bayesian Inference**
Parameter estimation: Find maximum likelihoods max $p(\boldsymbol{D}\|\boldsymbol{w}_1, \mathcal{H})$ and max $p(\boldsymbol{D}\|\boldsymbol{w}_2, \mathcal{H}_2)$ which determine **single best fit parameters** $\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2$. This is equivalent to $$\min \chi_1^2 = -2 \ln p(\boldsymbol{D}\|\tilde{\boldsymbol{w}}_1, \mathcal{H}_1)$$ $$\min \chi_2^2 = -2 \ln p(\boldsymbol{D}\|\tilde{\boldsymbol{w}}_2, \mathcal{H}_2)$$	**Parameter estimation:** Find evidences $p(\boldsymbol{D}\|\mathcal{H}_1)$ and $p(\boldsymbol{D}\|\mathcal{H}_2)$ which consider **all possible parameter values**. This is equivalent to $$-2\,\mathrm{LE}_1 = -2 \ln p(\boldsymbol{D}\|\mathcal{H}_1)$$ $$-2\,\mathrm{LE}_2 = -2 \ln p(\boldsymbol{D}\|\mathcal{H}_2)$$
Estimating uncertainties: $\sigma(\boldsymbol{w})$ is estimated from parameter covariance matrix $\mathbb{C}$.	**Estimating uncertainties:** Posterior $p(\boldsymbol{w}\|\boldsymbol{D}, \mathcal{H})$ provides comprehensive information on parameters, including $\sigma(\boldsymbol{w})$.
Model comparison: Comparison of $\chi_1^2/(N-K_1)$ to $\chi_2^2(N-K_2)$ uses only maximum likelihood $p(\boldsymbol{D}\|\tilde{\boldsymbol{w}}, \mathcal{H})$ and ignores covariances of $\boldsymbol{w}$.	**Model comparison:** Comparison of $-2\,\mathrm{LE}_1$ to $-2\,\mathrm{LE}_2$ uses covariances of $\boldsymbol{w}$ through $\mathbb{C}$ since $p(\boldsymbol{D}\|\mathcal{H}) = p(\boldsymbol{D}\|\tilde{\boldsymbol{w}}, \mathcal{H})\, p(\tilde{\boldsymbol{w}}\|\mathcal{H})\,(2\pi)^{K/2}\sqrt{\det \mathbb{C}}$.
Model construction: Explore different likelihoods	**Model construction:** Explore different priors and likelihoods
Accurate only for large N	Accurate for all N

Acknowledgements

References

[1] G. Agakishiev, *et al.* (STAR Collaboration), Phys. Rev. C **86**, 064902 (2012).
[2] J. Adams *et al.* (STAR Collaboration), Phys. Rev. C **72**, 014904 (2005).
[3] M.B. de Kock, H.C. Eggers and T.A. Trainor, Phys. Rev. C **92**, 034908 (2015).