Statistical Challenges of Global SUSY Fits

Roberto Trotta[1] *& Kyle Cranmer*[2]
[1]Imperial College London. [2]New York University

Abstract
We present recent results aiming at assessing the coverage properties of Bayesian and frequentist inference methods, as applied to the reconstruction of supersymmetric parameters from simulated LHC data. We discuss the statistical challenges of the reconstruction procedure, and highlight the algorithmic difficulties of obtaining accurate profile likelihood estimates.

arXiv:1105.5244v1 [hep-ph] 26 May 2011

1 Introduction

Experiments at the Large Hadron Collider (LHC) have already started testing many models of particle physics beyond the Standard Model (SM), and particular attention is being paid to the Minimal Supersymmetric SM (MSSM) and to other scenarios involving softly-broken supersymmetry (SUSY).

In the last few years, parameter inference methodologies have been developed, applying both Frequentist and Bayesian statistics (see e.g., [1–6]). While the efficiency of Markov Chain Monte Carlo (MCMC) techniques has allowed for a full exploration of multidimensional models, the likelihood function from present data is multimodal with many narrow features, making the exploration task with conventional MCMC methods challenging. A powerful alternative to classical MCMC has emerged in the form of Nested Sampling [7], a Monte Carlo method whose primary aim is the efficient calculation of the Bayesian evidence, or model likelihood. As a by-product, the algorithm also produces samples from the posterior distribution. Those same samples can also be used to estimate the profile likelihood. MULTINEST [8], a publicly available implementation of the nested sampling algorithm, has been shown to reduce the computational cost of performing Bayesian analysis typically by two orders of magnitude as compared with basic MCMC techniques. MULTINEST has been integrated in the `SuperBayeS` code[1] for fast and efficient exploration of SUSY models.

Having implemented sophisticated statistical and scanning methods, several groups have turned their attention to evaluating the sensitivity to the choice of priors [4, 9, 10] and of scanning algorithms [11]. Those analyses indicate that current constraints are not strong enough to dominate the Bayesian posterior and that the choice of prior does influence the resulting inference. While confidence intervals derived from the profile likelihood or a chi-square have no formal dependence on a prior, there is a sampling artifact when the contours are extracted from samples produced from Bayesian sampling schemes, such as MCMC or MULTINEST [10].

Given the sensitivity to priors and the differences between the intervals obtained from different methods, it is natural to seek out a quantitative assessment of their performance, namely their *coverage*: the probability that an interval will contain (cover) the true value of a parameter. The defining property of a 95% confidence interval is that the procedure adopted for its estimation should produce intervals that cover the true value 95% of the time; thus, it is reasonable to check if the procedures have the properties they claim. While Bayesian techniques are not designed with coverage as a goal, it is still meaningful to investigate their coverage properties. Moreover, the frequentist intervals obtained from the profile likelihood or chi-square functions are based on asymptotic approximations and are not guaranteed to have the claimed coverage properties.

Here we report on recent studies investigating the coverage properties of both Bayesian and Frequentist procedures commonly used in the literature. We also highlight the numerical and sampling challenges that have to be met in order to obtain a sufficienlty high-resolution mapping of the profile

[1]Available from: `www.superbayes.org`

likelihood when adopting Bayesian algorithms (which are typically designed to map out the posterior mass, instead).

For the sake of example, we consider in the following the so-called mSUGRA or Constrained Minimal Supersymmetric Standard Model (CMSSM), a model with fairly strong universality assumptions regarding the SUSY breaking parameters, which reduce the number of free parameters to be estimated to just five, denoted by the symbol $\mathbf{\Theta}$: common scalar (m_0), gaugino ($m_{1/2}$) and tri–linear (A_0) mass parameters (all specified at the GUT scale) plus the ratio of Higgs vacuum expectation values $\tan\beta$ and sign(μ), where μ is the Higgs/higgsino mass parameter whose square is computed from the conditions of radiative electroweak symmetry breaking (EWSB).

2 Coverage study of the CMSSM

2.1 Accelerated inference from neural networks

Coverage studies require extensive computational expenditure, which would be unfeasible with standard analysis techniques. Therefore, in Ref. [12] a class of machine learning devices called Artificial Neural Networks (ANNs) was used to approximate the most computationally intensive sections of the analysis pipeline.

Inference on the parameters of interest $\mathbf{\Theta}$ requires relating them to observable quantities, such as the sparticle mass spectrum at the LHC, denoted by $\mathbf{m}$, over which the likelihood is defined. This is achieved by evolving numerically the Renormalization Group Equations (RGEs) using publicly available codes, which is however a computationally intensive procedure. One can view the RGEs simply as a mapping from $\mathbf{\Theta} \rightarrow \mathbf{m}$, and attempt to engineer a computationally efficient representation of the function. In [12], an adequate solution was provided by a three-layer perceptron, a type of feed-forward neural network consisting of an input layer (identified with $\mathbf{\Theta}$), a hidden layer and an output layer (identified with the value of $\mathbf{m}(\mathbf{\Theta})$ that we are trying to approximate). The weight and biases defining the network were determined via an appropriate training procedure, involving the minimization of a loss function (here, the discrepancy between the value of $\mathbf{m}(\mathbf{\Theta})$ predicted by the network and its correct value obtained by solving the RGEs) defined over a set of 4000 training samples. A number of tests on the accuracy and noise of the network were performed, showing a correlation in excess of 0.9999 between the approximated value of $\mathbf{m}(\mathbf{\Theta})$ and the value obtained by solving the RGEs for an independent sample. A second classification network was employed to distinguish between physical and un-physical points in parameter space (i.e., values of $\mathbf{\Theta}$ that do not lead to physically viable solutions to the RGEs). The final result of replacing the computationally expensive RGEs with the ANNs is presented in Fig. 1, which shows that the agreement between the two methods is excellent, within numerical noise. By using the neural network, a speed-up factor of about 3×10^4 compared with scans using the explicit spectrum calculator was observed.

2.2 Coverage results for the ATLAS benchmark

We studied the coverage properties of intervals obtained for the so-called "SU3" benchmark point. To this end, we need the ability to generate pseudo-experiments with $\mathbf{\Theta}$ fixed at the value of the benchmark. We adopted a parabolic approximation of the log-likelihood function (as reported in Ref. [13]), based on the measurement of edges and thresholds in the invariant mass distributions for various combinations of leptons and jets in final state of the selected candidate SUSY events, assuming an integrated luminosity of 1 fb^{-1} for ATLAS. Note that the relationship between the sparticle masses and the directly observable mass edges is highly non-linear, so a Gaussian is likely to be a poor approximation to the actual likelihood function. Furthermore, these edges share several sources of systematic uncertainties, such as jet and lepton energy scale uncertainties, which are only approximately communicated in Ref. [13]. Finally, we introduce the additional simplification that the likelihood is also a multivariate Gaussian with the same covariance structure. We constructed 10^4 pseudo-experiments and analyzed them with both MCMC

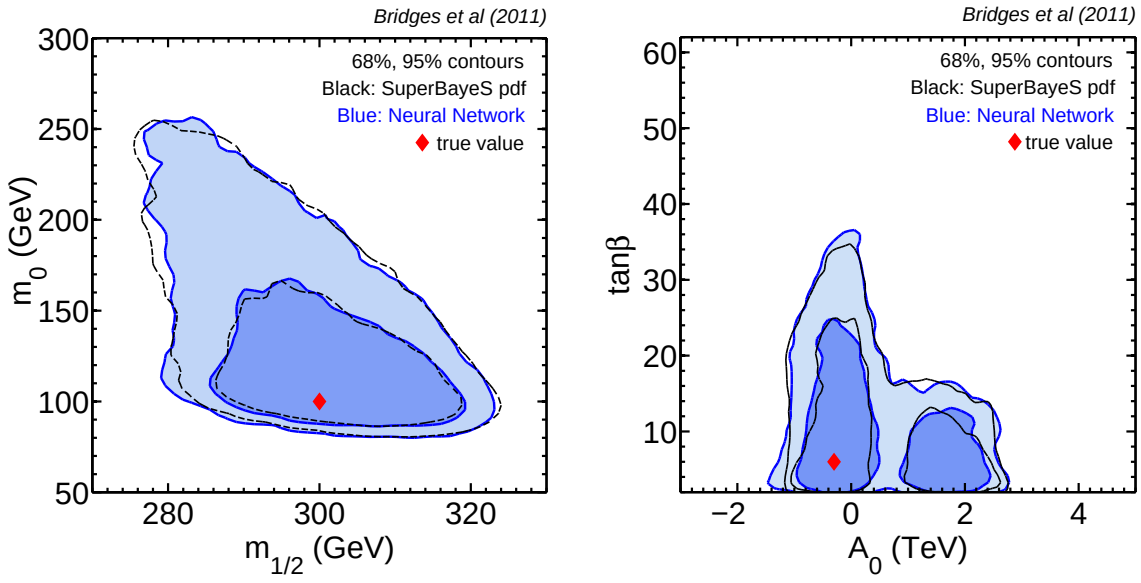


Fig. 1: Comparison of Bayesian posteriors obtained by solving the RGEs fully numerically (black lines, giving 68% and 95% regions) and neural networks (blue lines and corresponding filled regions), from simulated ATLAS data. The red diamond gives the true value for the benchmark point adopted. From [12].

(using a Metropolis-Hastings algorithm) and MULTINEST. Altogether, our neural network MCMC runs have performed a total of 4×10^{10} likelihood evaluations, in a total computational effort of approximately 2×10^4 CPU-minutes. We estimate that the solving the RGEs fully numerically would have taken about 1100-CPU years, which is at the boundary of what is feasible today, even with a massive parallel computing effort.

The results are shown in Fig. 2, where it can be seen that the methods have substantial over-coverage for 1-d intervals, which means that the resulting inferences are conservative. While it is difficult to unambiguously attribute the over-coverage to a specific cause, the most likely cause is the effect of boundary conditions imposed by the CMSSM. When $\mathbf{\Theta}$ is composed of parameters of interest, θ, and nuisance parameters, ψ, the profile likelihood ratio is defined as

$$\lambda(\theta) \equiv \frac{\mathcal{L}(\theta, \hat{\hat{\psi}})}{\mathcal{L}(\hat{\theta}, \hat{\psi})}. \tag{1}$$

where $\hat{\hat{\psi}}$ is the conditional maximum likelihood estimate (MLE) of ψ with θ fixed and $\hat{\theta}, \hat{\psi}$ are the unconditional MLEs. When the fit is performed directly in the space of the weak-scale masses (i.e., without invoking a specific SUSY model and hence bypassing the mapping $\mathbf{\Theta} \rightarrow \mathbf{m}$), there are no boundary effects, and the distribution of $-2\ln\lambda(\mathbf{m})$ (when $\mathbf{m}$ is true) is distributed as a chi-square with a number of degrees of freedom given by the dimensionality of $\mathbf{m}$. Since the likelihood is invariant under reparametrizations, we expect $-2\ln\lambda(\theta)$ to also be distributed as a chi-square. If the boundary is such that $\mathbf{m}(\hat{\theta}, \hat{\psi}) \neq \hat{\mathbf{m}}$ or $\mathbf{m}(\theta, \hat{\hat{\psi}}) \neq \hat{\hat{\mathbf{m}}}$, then the resulting interval will modified. More importantly, one expects the denominator $\mathcal{L}(\hat{\theta}, \hat{\psi}) < \mathcal{L}(\hat{\mathbf{m}})$ since $\mathbf{m}$ is unconstrained, which will lead to $-2\ln\lambda(\theta) < -2\ln\lambda(\mathbf{m})$. In turn, this means more parameter points being included in any given contour, which leads to over-coverage.

The impact of the boundary on the distribution of the profile likelihood ratio is not insurmountable. It is not fundamentally different than several common examples in high-energy physics where an unconstrained MLE would lie outside of the physical parameter space. Examples include downward

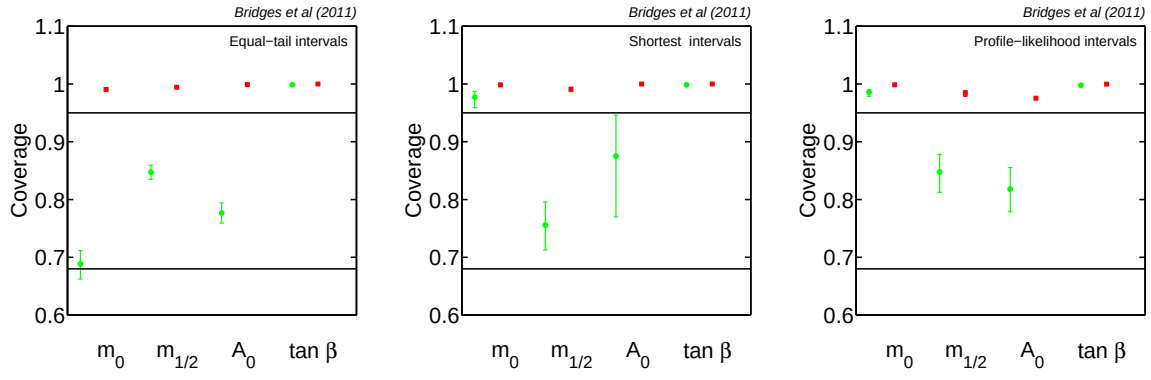


Fig. 2: Coverage for various types of intervals for the CMSSM parameters, from 10^4 realizations, employing MCMC for the reconstruction (each pseudo-experiment is reconstructed with 10^6 samples). Green/circles (red/squares) is for the 68% (95%) error. From [12].

fluctuations in event-counting experiments when the signal rate is bounded to be non-negative. Another common example is the measurement of sines and cosines of mixing angles that are physically bounded between $[-1, 1]$, though an unphysical MLE may lie outside this region. The size of this effect is related to the probability that the MLE is pushed to a physical boundary. If this probability can be estimated, it is possible to estimate a corrected threshold on $-2\ln\lambda$. For a precise threshold with guaranteed coverage, one must resort to a fully frequentist Neyman Construction. A similar coverage study (but without the computational advantage provided by ANNs) has been carried out for a few CMSSM benchmark points for simulated data from future direct detection experiments [14]. Their findings indicate substantial under-coverage for the resulting intervals, especially for certain choices of Bayesian priors. Both works clearly show the timeliness and importance of evaluating the coverage properties of the reconstructed intervals for future data sets.

3 Challenges of profile likelihood evaluation

For highly non-Gaussian problems like supersymmetric parameter determination, inference can depend strongly on whether one chooses to work with the posterior distribution (Bayesian) or profile likelihood (frequentist) [4, 10, 15]. There is a growing consensus that both the posterior and the profile likelihood ought to be explored in order to obtain a fuller picture of the statistical constraints from present-day and future data. This begs the question of the algorithmic solutions available to reliably explore both the posterior and the profile likelihood in the context of SUSY phenomenology.

The profile likelihood ratio defined in Eq. (1) is an attractive choice as a test statistics, for under certain regularity conditions, Wilks [16] showed that the distribution of $-2\ln\lambda(\theta)$ converges to a chi-square distribution with a number of degrees of freedom given by the dimensionality of θ. Clearly, for any given value of θ, evaluation of the profile likelihood requires solving a maximisation problem in many dimensions to determine the conditional MLE $\hat{\hat{\psi}}$. While posterior samples obtained with MULTINEST have been used to estimate the profile likelihood, the accuracy of such an estimate has been questioned [11]. As mentioned above, evaluating profile likelihoods is much more challenging than evaluating posterior distributions. Therefore, one should not expect that a vanilla setup for MULTINEST (which is adequate for an accurate exploration of the posterior distribution) will automatically be optimal for profile likelihoods evaluation. In Ref. [17] the question of the accuracy of profile likelihood evaluation from MULTINEST was investigated in detail. We report below the main results.

The two most important parameters that control the parameter space exploration in MULTINEST are the number of live points $n_{\rm live}$ – which determines the resolution at which the parameter space is

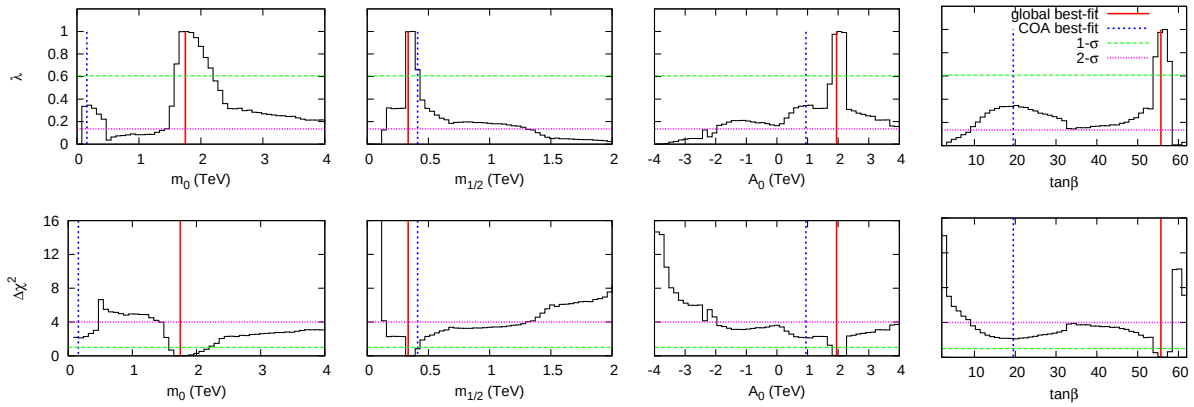


Fig. 3: 1-D profile likelihoods from present-day data for the CMSSM parameters normalized to the global best-fit point. The red solid and blue dotted vertical lines represent the global best-fit point ($\chi^2 = 9.26$, located in the focus point region) and the best-fit point found in the stau co-annihilation region ($\chi^2 = 11.38$) respectively. The upper and lower panel show the profile likelihood and $\Delta\chi^2$ values, respectively. Green (magenta) horizontal lines represent the 1σ (2σ) approximate confidence intervals. MULTINEST was run with 20,000 live points and $\mathrm{tol} = 1 \times 10^{-4}$ (a configuration deemed appropriate for profile likelihood estimation), requiring approximately 11 million likelihood evaluations. From [17].

explored – and a tolerance parameter tol, which defines the termination criterion based on the accuracy of the evidence. Generally, a larger number of live points is necessary to explore profile likelihoods accurately. Moreover, setting tol to a smaller value results in MULTINEST gathering a larger number of samples in the high likelihood regions (as termination is delayed). This is usually not necessary for the posterior distributions, as the prior volume occupied by high likelihood regions is usually very small and therefore these regions have relatively small probability mass. For profile likelihoods, however, getting as close as possible to the true global maximum is crucial and therefore one should set tol to a relatively smaller value. In Ref. [17] it was found that $n_{\mathrm{live}} = 20,000$ and $\mathrm{tol} = 1 \times 10^{-4}$ produce a sufficiently accurate exploration of the profile likelihood in toy models that reproduce the most important features of the CMSSM parameter space.

In principle, the profile likelihood does not depend on the choice of priors. However, in order to explore the parameter space using any Monte Carlo technique, a set of priors needs to be defined. Different choices of priors will generally lead to different regions of the parameter space to be explored in greater or lesser detail, according to their posterior density. As a consequence, the resulting profile likelihoods might be slightly different, purely on numerical grounds. We can obtain more robust profile likelihoods by simply merging samples obtained from scans with different choices of Bayesian priors. This does not come at a greater computational cost, given that a responsible Bayesian analysis would estimate sensitivity to the choice of prior as well. The results of such a scan are shown in Fig. 3, which was obtained by tuning MULTINEST with the above configuration, appropriate for an accurate profile likelihood exploration, and by merging the posterior samples from two different choices of priors (see [17] for details). This high-resolution profile likelihood scan using MULTINEST compares favourably with the results obtained by adopting a dedicated Genetic Algorithm technique [11], although at a slightly higher computational cost (a factor of ~ 4). In general, an accurate profile likelihood evaluation was about an order of magnitude more computationally expensive than mapping out the Bayesian posterior.

4 Conclusions

As the LHC impinges on the most anticipated regions of SUSY parameter space, the need for statistical techniques that will be able to cope with the complexity of SUSY phenomenology is greater than ever. An intense effort is underway to test the accuracy of parameter inference methods, both in the Frequentist and the Bayesian framework. Coverage studies such as the one presented here require highly-accelerated inference techniques, and neural networks have been demonstrated to provide a speed-up factor of up to $30,000$ with respect to conventional methods. A crucial improvement required for future coverage investigations is the ability to generate pseudo-experiments from an accurate description of the likelihood. Both the representation of the likelihood function and the ability to generate pseudo-experiments are now possible with the workspace technology in RooFit/RooStats [18]. We encourage future experiments to publish their likelihoods using this technology. Finally, an accurate evaluation of the profile likelihood remains a numerically challenging task, much more so than the mapping out of the Bayesian posterior. Particular care needs to be taken in tuning appropriately Bayesian algorithms targeted to the exploration of posterior mass (rather than likelihood maximisation). We have demonstrated that the MULTINEST algorithm can be succesfully employed for approximating the profile likelihood functions, even though it was primarily designed for Bayesian analyses. In particular, it is important to use a termination criterion that allows MULTINEST to explore high-likelihood regions to sufficient resolution.

Acknowledgements: We would like to thank the organizers of PHYSTAT11 for a very interesting workshop. We are grateful to Yashar Akrami, Jan Conrad, Joakim Edsjö, Louis Lyons and Pat Scott for many useful discussions.

References

[1] E. A. Baltz and P. Gondolo, *JHEP* **10** (2004) 052.

[2] B. C. Allanach and C. G. Lester, *Phys. Rev.* **D73** (2006) 015013.

[3] R. Ruiz de Austri, R. Trotta, and L. Roszkowski, *JHEP* **05** (2006) 002.

[4] B. C. Allanach, K. Cranmer, C. G. Lester, and A. M. Weber, *JHEP* **0708** (2007) 023.

[5] O. Buchmueller, R. Cavanaugh, A. De Roeck, J. R. Ellis, H. Flacher, S. Heinemeyer, G. Isidori, K. A. Olive *et al.*, Eur. Phys. J. **C64**, 391-415 (2009).

[6] L. Roszkowski, R. Ruiz de Austri, and R. Trotta, *JHEP* **07** (2007) 075.

[7] J. Skilling, *Nested Sampling*, in *American Institute of Physics Conference Series* (R. Fischer, R. Preuss, and U. V. Toussaint, eds.), pp. 395–405, (2004).

[8] F. Feroz, M. P. Hobson, and M. Bridges, *Mon. Not. R. Astron. Soc.* **398** (2009) 1601–1614.

[9] R. Lafaye, T. Plehn, M. Rauch, and D. Zerwas, *European Physical Journal C* **54** (2008) 617–644.

[10] R. Trotta, F. Feroz, M. Hobson, L. Roszkowski, and R. Ruiz de Austri, *JHEP* **12** (2008) 24.

[11] Y. Akrami, P. Scott, J. Edsjo, J. Conrad, and L. Bergstrom, *JHEP* **04** (2010) 057.

[12] M. Bridges, K. Cranmer, F. Feroz, M. Hobson, R. R. de Austri, R. Trotta, JHEP **1103**, 012 (2011).

[13] The ATLAS Collaboration: G. Aad, E. Abat, B. Abbott, J. Abdallah, A. A. Abdelalim, A. Abdesselam, O. Abdinov, B. Abi, M. Abolins, H. Abramowicz, and et al., *ArXiv e-prints* (Dec., 2009) [http://xxx.lanl.gov/abs/0901.0512].

[14] Y. Akrami, C. Savage, P. Scott, J. Conrad, and J. Edsjö, (2010), pre-print: [http://xxx.lanl.gov/abs/1011.4297].

[15] P. Scott, J. Conrad, J. Edsjö, L. Bergström, C. Farnier, and Y. Akrami, JCAP **1001**, 031 (2010).

[16] S. Wilks, *Ann. Math. Statist.* **9** (1938) 60–2.

[17] F. Feroz, K. Cranmer, M. Hobson, R. Ruiz de Austri, R. Trotta, (2011), pre-print: [http://xxx.lanl.gov/abs/1101.3296].

[18] L. Moneta, K. Belasco, K. Cranmer, A. Lazzaro, D. Piparo, *et al.*, *The RooStats Project*, *Proceed-*

ings of Science (2010) Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, India, [http://xxx.lanl.gov/abs/arXiv:1009.1003].