UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXX

Developments in Approximate Bayesian Computation and Statistical Applications in Astrostatistics

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Alessandra Rosalba Brazzale

Co-supervisore: Prof. Jessi Cisewski–Kehe

Dottorando: Umberto Simola

Abstract

The title of this Thesis embraces two topics that have been investigated.

Most of the present work is dedicated to develops and extensions for Approximate Bayesian Computation (ABC). While several algorithms have been proposed to improve the efficiency of the basic ABC algorithm, a number of subjective choices is left to the researcher. Several of these choices have not only an impact on the efficiency of the algorithm but also on its capability to approximate properly the true posterior distribution. We present a first extension of the ABC Population Monte–Carlo (ABC–PMC) algorithm aimed by the goal of minimizing the number of subjective inputs required to the user, improving at the same time the computational efficiency of the algorithm. In the second work we propose extensions of the ABC–PMC algorithm as an alternative framework for inference to work with finite mixture models.

The second topic was initiated from a collaboration between the Statistics and Data Science Department and the Astronomy Department at Yale University and the Department of Physics at the University of Geneve, with the goal of detecting and characterizing "Earth–like" extrasolar planets. We propose a novel statistical tool to better disentangle stellar activity from the pure signal coming from an extrasolar planet, aimed by the goal of detecting and characterizing "Earth–like" planets.

Sommario

Il titolo di questa Tesi vuole abbracciare i due differenti argomenti che sono stati investigati.

La maggior parte del presente lavoro é dedicata a sviluppi ed estensioni dell' algoritmo Approximate Bayesian Computation Population Monte–Carlo (ABC–PMC). Mentre parecchi algoritmi sono stati proposti per migliorare l'efficienza della procedura base ABC, alcune scelte soggettive vengono lasciate al ricercatore. Alcune di queste scelte hanno non solo un impatto sull'efficienza dell'algoritmo, ma anche sulla capacitá del medesimo di approssimare in maniera consona la vera distribuzione a posteriori. Noi presentiamo una prima estensione dell'algoritmo ABC–PMC che vuole minimizzare il numero di scelte soggettive richieste all'utente, con l'obiettivo di migliorare l'efficienza dell'algoritmo preservando al contempo l'ottenimento di una fedele approssimazione della vera distribuzione a posteriori. Come seconda estensione, proponiamo una procedura basata sull'algoritmo ABC–PMC per lavorare con modelli mistura (caso finito).

Il secondo argomento descrive uno dei risultati della collaborazione tra il Dipartimento di Astronomia e il Dipartimento di Statistica e Data Science all' Universitá di Yale ed il Dipartimento di Fisica all'Universitá di Ginevra, dove l'obiettivo consiste nello scovare e caratterizzare pianeti extrasolari. Noi proponiamo una nuova tecnica statistica per meglio separare l'attivitá stellare dal puro segnale proveniente da un pianeta extrasolare, con l'obiettivo di scovare e caratterizzare esopianeti terrestri teoricamente adatti ad ospitare la vita.

"Qualsiasi cosa dica in merito potrá essere male interpretata: ma anche tacendo daró adito agli equivoci. Le parole possono mentire, il silenzio puó mentire. Persino i fatti possono mentire. I bugiardi peggiori sono quelli che raccontano i fatti fingendo di aver raccontato la veritá. Racconteró solo i fatti, non pretendo di sapere la veritá. Ogni cosa accade secondo una certa opportunitá."

Cesare Borgia

Acknowledgements

I would like to thank my supervisor, Prof. Alessandra R. Brazzale, for her support, her advice, the independence in research she gave to me and for giving me the opportunity to work in the field I have always been deeply interested in. I furthermore acknowledge and thank the financial supports from Fondazione CARIPARO and from the Research Project 2015: "Advances in likelihood-based inference in Biostatistics with application to measurement error problems and meta-analysis" (CPDA153257) lead by Prof. Annamaria Guolo. I would like to thank the Yale Center for Research Computing for the resources I have been using to produce part of the material presented in this Thesis.

Once reached the United States for my visiting I could only image the environment I was about to meet. In retrospective I can say my expectations were by far too short and sometimes I find myself questioning if this life experience was real or merely a dream. These acknowledgements will be surely helpful to remind to myself it was not the latter. I have no words for thanking my co–supervisor, Prof. Jessi Cisewski-Kehe, who followed and encouraged me throughout and beyond my staying. I acknowledge her financial support for covering the tuition costs for the entire length of my staying at Yale. I want also to thank all the guys who helped me during my visiting, starting from the OISS and finishing with all the people I met in the Departments of Statistics and Data Science and Astronomy at Yale. Among the others I wanna thank JoAnn, Karen, Liz, David, John, Debra, Eric, Robert, Allen, Bradley and Derek. A special and great thank to JoAnn, my "neighbor" who welcomed me in February 2016 and allowed me to get a "coffe-chatting break" all the mornings in the stunning Dana House.

To thank all the people I knew in the last 19 months is impossible so I hope not disappointing the hopefully few ones I forgot mentioning. First of all I wanna thank all the guys at Gpscy, especially Dmitri and Sam for making me feel more than welcome in their pub during the sometimes empty evenings in New Haven. Among the others, a special thank to the guys of Anna Liffey's and to Brandon, Aly and RBJ for the amazing Football "pick–6" Sundays that I am missing more than I thought. My staying in New Haven would not have been that fun (and painful) without the time spent in the Payne-Whitney Gym with other "gym-rats" such as Dean, Leo, Rick, S.B. and Rachel. Thanks also to Brandan, Jess, Hope and the Cafe Romeo for the time I spent in their place during some winter storm, working on my stuff and enjoying their amazing spiced tea. I was in luck, because I traveled for conferences through the Country more than once,

letting me seeing long time dreamed to visit cities such as Los Angeles (expecially the Staples Center), New York (aka the Big Apple) and New Orleans (aka the Big Easy) and meeting amazing colleagues: Brittany, Andy, Jacob, Ines, Xavier, Elba, Jorge, Rafael and Roberta. I cannot not mentioning and thanking Caroline, McKenzie, Don, Ally, Ula, Aga, Mirjana, Petra and all the people who letting me living incredible emotions during the Brooklyn Nets' games, the Ivy League and the CT–Open 2017.

Getting a PhD can be sometimes tough and maintaining things on perspective is not always easy. I am really grateful for the great colleagues I have found for sharing this 3 years long journey with. In particular I want to thank Leo and Claudio for the emotions that only a "Magnanelli's goal" or a "Capitano Ondrášek" can provide, as well as for the adventures we spent together during our first year as PhD students, making clear that at the end of the day there is something much more important that getting a title or publishing a paper: making human connections and building long term relationships. Special thanks go to the roommates I have been changing throughout this experience, starting from old friends of mine Luca and Jack, passing to Dan and the "invisible nurse" and finishing with Colin and Adrien.

In closing, I would like to thank my friends that have been living in France for a while right now; although more than 5000 miles we kept staying in touch constantly and more than never. I would love to thank my family for the support and for the challenges that being far away for a long time can imply.

Contents

List of Figures **xv**

List of Tables **xxiii**

1 Introduction **3**
1.1 Overview . . . 3
1.2 Main contributions of the Thesis . . . 5

2 Approximate Bayesian Computation Methods **9**
2.1 Motivations for using Approximate Bayesian Computation . . . 9
2.2 Basic Approximate Bayesian Computation algorithm . . . 10
2.2.1 Selection of informative "enough" summary statistics . . . 12
2.2.2 Selection of small "enough" tolerances . . . 13
2.2.3 A first example: the Normal–Normal model . . . 15
2.3 Approximate Bayesian Computation Population Monte–Carlo algorithm . 16
2.4 Concluding remarks . . . 19

3 Adaptive Approximate Bayesian Computation Tolerance Selection **23**
3.1 Introduction . . . 24
3.2 Automatic tolerance selection . . . 25
3.2.1 Stopping rule . . . 27
3.3 Illustrative Examples . . . 28
3.3.1 Beta-Binomial Model . . . 29
3.3.2 Exponential-Gamma Model . . . 33
3.3.3 Gaussian Mixture Model . . . 37
3.3.4 Presence of local modes . . . 41
3.3.5 Lotka–Volterra model . . . 45
3.4 Concluding remarks . . . 46

4 Approximate Bayesian Computation for Finite Mixture Models **51**
4.1 Introduction . . . 51
4.2 Required extensions of the ABC–PMC algorithm . . . 53
4.2.1 Finite Gaussian Mixture Models . . . 53
4.2.2 Perturbation kernel functions . . . 54
4.2.3 Algorithm for addressing the label switching problem . . . 57
4.2.4 Summary statistics . . . 59

4.3 Illustrative Examples 60
4.3.1 Mixture Model with equal group sizes 60
4.3.2 Mixture Model with unequal group size 63
4.3.3 Application to Galaxy Data 64
4.4 Concluding Remarks 66

5 Measuring precise radial velocities and cross-correlation function line-profile variations using a Skew Normal distribution 71
5.1 Introduction 72
5.1.1 The BIS SPAN parameter for measuring stellar activity 74
5.2 The Skew Normal distribution 78
5.3 Fitting the Skew Normal density to the CCF 80
5.4 Radial Velocity correction function for stellar activity 84
5.5 Illustrative Examples 84
5.5.1 Alpha Centauri B 85
5.5.2 HD192310 90
5.5.3 HD10700 94
5.5.4 HD215152 96
5.5.5 Corot-7 99
5.6 Estimation of standard errors for the CCF parameters 102
5.7 Concluding Remarks 105

Appendix 111
A Analytic expression for the ABC posterior distribution under the assumption of Normal distribution 111
B Impact of the desired particle sample size N on the Adaptive Approximate Bayesian Computation Tolerance Selection algorithm 116
C Resampling the Mixture Weights 119

Bibliography 125

List of Figures

2.1 Comparison between the true posterior distribution (black line) and the ABC posterior distribution using as summary statistic the mean (blue line), the first quartile (yellow dots) and the third quartile (green dots). When using the complete minimal sufficient statistic, the ABC posterior distribution is comparable with the true posterior. When picking an insufficient summary statistic, the resulting ABC posterior distribution is not a suitable approximation of the true posterior. The tolerance ϵ is fixed for all the cases equal to 0.01. Smaller values for the tolerance have been tested when the summary statistic is insufficient. The corresponding ABC posteriors have not improved, suggesting that too much information was lost by using respectively the first and the third quartile as summary statistic for the presented model. 17

3.1 Illustration of selection of q_t. (left) The proposal distribution ABC posterior $\hat{\pi}_{t-1}$ and the resulting ABC posterior $\hat{\pi}_t$, with $\hat{c}_t$ is defined in Equation (3.1) and used for setting q_t as defined in Equation (3.2). (right) The (arbitrary) distribution of distances is from the accepted distances at iteration t, $\{d_t^{(J)}\}_{J=1}^N$, with ϵ_t being the largest possible value. The next iterations tolerance, ϵ_{t+1}, is set as the q_t quantile of $\{d_t^{(J)}\}_{J=1}^N$. . . . 26

3.2 aABC–PMC analysis for the discrete Beta-Binomial model with $N = 2000$. (left) Series of 5 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of automatically selected quantiles: $q_{2:5} = (0.22, 0.55, 0.87, 0.88)$, that lead to the series of tolerances $\epsilon_{1:5} = (0.1, 0.02, 0.01, 0.01, 0.01)$. The automatic stopping rule directly based on the behavior of the ABC posterior distribution is satisfied after 5 iterations. 30

3.3 ABC–PMC analysis for the discrete Beta-Binomial model with $N = 2000$, $q^{th} = 0.75$ and $T = 15$. (left) Series of 15 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of efficiencies based on the selection in advance of the quantile used to reduce the tolerance through the 15 iterations. 31

3.4 ABC–PMC analysis for the Beta-Binomial model with $N = 2000$ and maximum number of allowed draws equal to 54889, for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.068, 0.058, 0.091)$. 32

3.5 ABC–PMC analysis for the Beta-Binomial model with $N = 2000$ and time limit equal to 125.990 sec., for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.14, 0.037, 0.16)$. 33
3.6 ABC final posterior distributions for different initial choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. 34
3.7 aABC–PMC analysis for the Exponential-Gamma model with $N = 2000$. (left) Series of 4 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of automatically selected quantiles: : $q_{2:4} = (0.21, 0.61, 0.88)$, that leads to the series of tolerances $\epsilon_{1:4} = (0.35, 0.08, 0.048, 0.042)$. The automatic stopping rule directly based on the behavior of the ABC posterior distribution is satisfied after 4 iterations. 35
3.8 ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$, $q^{th} = 0.75$ and $T = 15$. (left) Series of 15 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of efficiencies based on the selection in advance of the quantile used to reduce the tolerance through the 15 iterations. 35
3.9 ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$ and maximum number of allowed draws equal to 63784, for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.072, 0.082, 0.09)$. 36
3.10 ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$ and time limit equal to 56.193 sec., for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.072, 0.074, 0.14)$. 37
3.11 ABC final posterior distributions for different initial choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. 38
3.12 Gaussian mixture model example. (left) ABC–PMC and aABC–PMC final posterior distributions. The true posterior distribution is plotted with the black line. (right) Sequential quantities computed for the aABC–PMC method. The q_t's (black circles) generally increase through the iterations and the $1/\hat{C}_t$'s (orange pluses) generally decrease until they stabilize. The acceptance rate (blue triangles) decreases throughout the iterations which is why it is desirable to stop the algorithm once the ABC posterior has stabilized. 39
3.13 Gaussian Mixture model example. (left) Average total number of draws required by the aABC–PMC and the ABC–PMC algorithm for different quantiles. (right) Average computational time required by the aABC–PMC and the ABC–PMC algorithm for different quantiles. Because the aABC–PMC algorithm adaptively selects different quantiles for each iteration, the red 'x' is placed at the average quantile, 0.44. 40

3.14 aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. . . 41
3.15 Example from [130] to investigate the performance of the proposed aABC–PMC in the presence of a local optimal value. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 5 iterations. The series of automatically selected quantiles is $q_{2:5} = (0.18, 0.000016, 0.0044, 0.02)$ which leads to the series of tolerances $\epsilon_{1:5} = (51.59, 51.01, 2.81, 0.00058, 1.42 \cdot 10^{-4})$. 43
3.16 Local maximum example with desired particle sample size equal to $N = 1000$ and initial number of draws from the prior equal to 2000. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 6 iterations. In this case the parametric space is not sufficiently explored, hence the achievement of the true posterior distribution is not guaranteed by the aABC–PMC. The series of automatically selected quantiles $q_{2:6} = (0.44, 0.34, 0.31, 0.28, 0.23)$ are too gentle for forcing the algorithm to consider those few particles coming from the true posterior distribution and available at the end of the first iteration. The series of tolerances $\epsilon_{1:5} = (55.82, 51.91, 51.18, 51.02, 51.02, 51.02)$ is coherent with the results found by by [130]. 44
3.17 Extended example from [130], by considering a further local minimum at $\theta = 15$. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 5 iterations.The aABCpmc algorithm provides a series of tolerances $\epsilon_{1:5} = (51.62, 50.91, 11.43, 0.00078, 1.42 \cdot 10^{-4}, 6.28 \cdot 10^{-5})$. which leads to the true posteriors posterior. 45
3.18 Lotka–Volterra results. (left)(middle) comparison between the final posterior distributions for a and b obtained using [140]'s manually selecting the tolerances (cyan), by fixing the quantile equal to .50 (green) and 0.75 (yellow), and by using the aABC–PMC (blue). (right) The q_t's (black circles) generally increase through the iterations, while the acceptance rate (blue points) mildly increases and then decreases after iterations 5 and 6. Once the ABC posterior distribution stops improving as the tolerance decreases, the series of $1/\hat{C}_t$'s defined for stopping the algorithm (orange points) stabilizes. 47

4.1 Comparison between the ABC and the MCMC marginal posterior distributions for the two-component GMM example from [94]. The final ABC posteriors obtained using the label switching (LS) procedure proposed in Section 4.2.3 are the solid black lines (ABC Posterior good LS), and the naive approach that sorts based on the mixture weights are the solid cyan lines (ABC Posterior bad LS). We recall that only for the MCMC analysis the label switching problem has to be addressed. This is done deterministically sorting the parameters according to the means of the mixture model. The number of particles for the ABC analysis and the number of elements kept from the MCMC analysis (after the burn-in) are equal to 5000. 61
4.2 ABC and the MCMC marginal posterior distributions for the three-component GMM example from [94]. The number of particles for the ABC analysis and the number of elements kept from the MCMC analysis (after the burn-in) are equal to 5000. 63
4.3 (left) The log-likelihood surface of the Gaussian mixture model proposed by [89]. There are two modes in the log-likelihood function, one close to the true value, $(0, 2.5)$, and a second local mode. (right) The marginal ABC, PMC, and MCMC posterior distributions; the displayed MCMC posteriors include the results for good initial starting values (MCMC Posterior (good initial choice)) and bad initial starting values (MCMC Posterior (bad initial choice)). 64
4.4 Histogram of the recessional velocity of 82 galaxies and the estimated three-component Gaussian mixture models for each study. The posterior means for the mixture weights, means and variances used are displayed in Table 4.4. It is possible to note that both the results coming from the extended ABC–PMC algorithm (blue and orange lines) and from [94] (green line) allow for a clear third component in the mixture. The results obtained by [89] (red lines) find a third component whose variance is in particular equal to $\sigma_3^2 = 34.1$, making the third cluster not appreciable in the Figure. 67

5.1 (left) The BIS SPAN of the CCF. V_0 is an arbitrary offset. Note the definition of the boundaries for the computation of ($\bar{V}_t$ and $\bar{V}_b$) [112]. (right) The bisector of the CCF for the star HD166435, constructed with a template selecting only the weak and non saturated lines. This profile represents the mean spectral-line profile of the lines selected by the template (i.e. the CCF). The original image was originally shown by [66]. 76
5.2 Density function of a random variable $Y \sim SN(\xi, \omega^2, \alpha)$ with location parameter $\xi = 0$, scale parameter $\omega = 1$ and different values of the skewness parameter α indicated by different colors and line types. Note that the solid black line has an $\alpha = 0$, making it a Normal distribution. 79
5.3 Among the 1812 analyzed CCF's for Alpha Centauri B, 4 CCF's leads to problems when using the SN distribution. For all the CCF's, the shape of the profile is not recognizable and the numerical minimization using the SN distribution leads to γ exceeding its range, forcing the procedure, correctly, to arrest. 82

5.4 Among the 7935 analyzed CCF's for HD10700, 7 CCF's leads to problems when using the SN distribution. It is worth noting that all these CCF's $= \{9158, \ldots, 9164\}$ are consecutive. For all of them, the profile does not follow to the shape of a proper absorption line. 83
5.5 Correlation between γ and the BIS SPAN for Alpha Centauri B. 85
5.6 (top) RV's and (bottom) RV's differences for Alpha Centauri B considering a Normal and a SN fitting. Two location parameters are proposed using the SN density, SN mean RV (black dots) and SN median RV (cyan crosses), while the location parameter for the Normal fit is N mean RV (red triangles). 87
5.7 (top) Set of RV's for Alpha Centauri B estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable. 88
5.8 Correlation between the asymmetry parameters and the RV's for Alpha Centauri B. The last three plots show the correlation between the FWHM's and the RV's for Alpha Centauri B, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is significantly higher, almost twice, than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.741$). The comparison of the correlations between FWHM's and RV's shows that the indicators retrieved by the SN fit have stronger correlations than the one obtained with the common analysis ($R = 0.817$). 89
5.9 The RV's as a function of time (left) and the RV's plotted against γ (right) with colors and plot symbol according to its temporal cluster assignment for Alpha Centauri B. The RV's are expressed in $\mathrm{m\,s^{-1}}$ 90
5.10 The RV's for Alpha Centauri B corrected from stellar activity using the SN fit and accounting for the temporal clusters (left), and the difference between those values in the left plot and the analogous values without accounting for the temporal clusters (right) which are displayed in the lower left plot of Figure 5.7. The RV's are expressed in $\mathrm{m\,s^{-1}}$ 91
5.11 Correlation between γ and the BIS SPAN for HD192310. 91
5.12 (top) Set of RV's for HD192310 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable. 92
5.13 Correlation between the asymmetry parameters and the RV's for HD192310. The last three plots show the correlation between the FWHM's and the RV's using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.669$). The comparison of the correlations between FWHM's and RV's leads to the same conclusion, with the correlation between SN mean RV and SN FWHM to be the strongest ($R = 0.666$). 93
5.14 Correlation between γ and the BIS SPAN for HD10700. 94

5.15 (top) Set of RV's for HD10700 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals for the Normal and SN analyses are comparable. 95
5.16 Correlation between the asymmetry parameters and the RV's for HD10700. The last three plots show the correlation between the FWHM's and the RV's for HD10700, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's (($R = 0.322$)). The comparison of the correlations between the FWHM's and RV's, when using the Normal and the SN fit leads to comparable considerations. However, in this case, the correlation between N mean RV and FWHM is the strongest ($R = 0.529$). 97
5.17 Correlation between γ and the BIS SPAN for HD215152. 98
5.18 (top) Set of RV's for HD215152 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). While the correction from stellar activity leads to similar considerations when SN mean RV or SN median RV are used, using N mean RV leads to residuals 0.062 $\mathrm{m\,s^{-1}}$ higher than the one retrieved with the SN fit. . . . 99
5.19 Correlation between the asymmetry parameters and the RV's for HD215152. The last three plots show the correlation between the FWHM's and the RV's for HD215152, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.571$). Concerning the comparison of the correlations between FWHM's and RV's, the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.21$). 100
5.20 Correlation between γ and the BIS SPAN for Corot-7. 101
5.21 (top) Set of RV's for Corot-7 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). While the correction from stellar activity leads to similar considerations when SN mean RV or SN median RV are used, using N mean RV leads to residuals 0.25 $\mathrm{m\,s^{-1}}$ higher. 102
5.22 Correlation between the asymmetry parameters and the RV's for Corot-7. The last three plots show the correlation between the FWHM's and the RV's for Corot-7, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.537$). Concerning the comparison of the correlations between FWHM's and RV's, the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.73$). 103

5.23 Comparison between the standard errors using the bootstrap analysis for the RV's, the FWHM and the asymmetry parameters. When using SN mean RV (black circles), the standard errors are in average 60% larger than the standard errors retrieved for RV (red triangles). However, if using SN median RV (cyan crosses), the standard errors are on average 10% smaller than the standard errors related to RV. To use as asymmetry parameter the γ of the SN leads to standard errors on average 15% smaller than the standard errors related to the BIS SPAN. Note that for the asymmetry, the error in BIS SPAN is in $\mathrm{km\,s^{-1}}$. To be able to compare the errors in γ and BIS SPAN we multiplied the error in γ by the slope of the linear fit between γ and BIS SPAN, as shown in Figures 5.11, 5.17 and 5.20. 106

B.1 **Beta-Binomial model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$. 117
B.2 **Exponential-Gamma model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$. 118
B.3 **Gaussian Mixture Model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$. 118

List of Tables

3.1 Results for aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. 34
3.2 aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. . . 38
3.3 The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC algorithm (right), obtained by running the procedure 20 times. For aABC–PMC algorithm, the quantile automatically selected through the iterations is displayed under q_t. 39
3.4 The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC (right) algorithm, obtained by running the procedure 20 times. 41
3.5 aABC–PMC algorithm with different choices of the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$. . . 41
3.6 The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC (right) algorithm, obtained by running the procedure 20 times. In the aABC–PMC algorithm also the quantile automatically selected through the iterations is available. 47

4.1 Posterior means (and posterior standard deviations) obtained by using the MCMC and the ABC–PMC algorithm for the two-component GMM example from [94]. The fourth column indicates the Hellinger distance between the final ABC and the MCMC posteriors. The number of ABC particles and the number of elements retained from the MCMC chain (after the burn-in) are both equal to 5000. 62
4.2 Posterior means (and posterior standard deviations) obtained by using the MCMC and the ABC–PMC algorithm for the three-component GMM example from [94]. The fourth column is the Hellinger distance between the final ABC posterior distribution and the MCMC posterior. The number of particles and the number of elements retained from the MCMC chain (after the burn-in) are both equal to 5000. 63
4.3 Mean posteriors (and standard deviations) obtained by using MCMC (with good and poor choices for initializing the procedure), PMC and ABC algorithms in the example by [89]. The last column indicates the Hellinger distance between the final ABC posterior distributions and the PMC posteriors 65

4.4 Comparison between the posterior means obtained by [89], [94] (MCMC algorithm) and the ABC–PMC algorithm for the Galaxy data. The results of the ABC–PMC analysis including measurement errors are displayed in the fourth column. The proposed ABC–PMC estimates are comparable with the ones obtained by [94], while [89] obtained different results, in particular for the third component of the mixture. 66

5.1 CP values, (μ, σ^2, γ), corresponding to the α values from Fig. 5.2 (with location parameter $\xi = 0$ and scale parameter $\omega = 1$) using Equation (5.5). Values are rounded to three decimal places. 80

5.2 **Alpha Centauri B**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equarion (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. All the three parameters are useful in explaining variations in RV's of the star that can be caused by stellar activity. Anyway the evaluation of the R^2 shows that the linear combination better explains variations in RV's due to stellar activity coming from the SN analysis which uses SN mean RV. 86

5.3 **HD192310**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the BIS SPAN is not statistically useful to explain variations in the RV's of the star. On the other hand, concerning the analyses based on the SN density, all the p–values associated with the parameters involved in Equation (5.9) are statistically different from 0. The evaluation of the R^2 shows that the linear combination better explains variations in RV's due to stellar activity coming from the SN analysis which uses SN mean RV. 94

5.4 **HD10700**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. All the three parameters are useful in explaining variations in RV's of the star that can be caused by stellar activity. The R^2 shows that the correction for stellar activity is equally important for the three analyses. 96

5.5 **HD215152**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the intercept and the FWHM are statistically significant to explain the RV's variations at level 0.05 but not at level 0.01. The BIS SPAN is not significant, which explains why the R^2 is only 0.019. On the contrary, for the SN case, γ is statistically significant in explaining the variations in RV's caused by stellar activity. The correction for stellar activity is more useful when using SN mean RV ($R^2 = 0.34$). 98

5.6 **Corot-7**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the BIS SPAN is not significant, as well as γ if using SN median RV. When using SN mean RV, all the parameters are statistically significant, suggesting again that SN mean RV ($R^2 = 0.56$) is more sensible to stellar activity than SN median RV ($R^2 = 0.44$). 104

B.1 **Beta-Binomial model**: aABC-PMC algorithm performances for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown. 117

B.2 **Exponential-Gamma model**: aABC-PMC algorithm for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown. 118

B.3 **Gaussian Mixture Model**: aABC-PMC algorithm for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown. . 118

Chapter 1

Introduction

1.1 Overview

The title of this Thesis embraces two topics that have been investigated. The first topic is some methodological develops for Approximate Bayesian Computation (ABC). The second topic was initiated from a collaboration between the Statistics and Data Science Department and the Astronomy Department at Yale University and the Department of Physics at the University of Geneve, with the goal of detecting and characterizing "Earth–like" extrasolar planets.

In the first part of the Thesis, ABC is considered. ABC provides a framework for inference in situations where the relationship between the data and the parameters is not well-approximated by a tractable likelihood function, but simulation of the data–generating process is possible. In recent years there have been many extensions to the basic ABC algorithm, and in this Thesis we focus on the ABC - Population Monte–Carlo (ABC–PMC) algorithm. Starting from the ABC–PMC algorithm we developed two extensions. In the first extension we present a method for automatically and efficiently selecting the series of tolerances, $\epsilon_{1:T} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)$, along with determining T (i.e. when to stop the algorithm). All the quantities are based on the online performances of the ABC posterior distribution and the number of arbitrary selections required from the researcher is reduced. In the second work we propose extensions of the ABC–PMC algorithm as an alternative framework for inference to work with finite mixture models. Part of the presented work is the results of the collaboration with Prof. Robert Wolpert from the Statistical Sciences Department at Duke University.

In the second topic of this Thesis we outline one of the results of the collaboration between the Department of Statistics and Data Science and the Department of Astronomy at Yale University and the Department of Physics at the University of Geneve. In particular one of the main goals of this collaboration is to detecting and characterizing

"Earth–like" extrasolar planets. When searching for terrestrial exoplanets in the habitable zone using the radial velocity technique, one of the most important challenges consists in properly addressing the impact of stellar activity. The result of using the radial velocity technique for detecting extrasolar planets, when using data coming from stabilized spectrographs, is usually summarized in the cross-correlation function (CCF), which is an average of certain absorption lines of a stellar spectrum. The CCF is used for measuring the radial velocity of the star and also for providing information about stellar activity (evaluating the shape of the CCF). Poorly disentangling the signals coming from the exoplanet and spurious radial velocity perturbations caused by stellar activity can result in a false positive detection. When studying the CCF, the classic analysis consists in two well defined steps. At first, the Normal distribution is used for retrieving the radial velocity of the star and then, as a second and separate operation, the stellar activity is evaluated by retrieving the so–called Bisector Inverse Slope Span in order to measure the asymmetry of the CCF in order to infer stellar activity. We propose to conduct the entire analysis with using the Skew Normal distribution. By using the Skew Normal distribution the barycenter and skewness of a CCF can be retrieved in a single operation, the correlation between the radial velocity of the star and stellar activity can be better understood and finally the uncertainties associated to all the parameters are smaller than the ones estimated with the classic analysis. This latter point is fundamental when searching for rocky exoplanets in the habitable zone using the state–of–the–art spectrographs.

The Thesis is organized as follows: Chapter 1 highlights the main contributions of the Thesis. Chapter 2 introduces the ABC methods, motivating their use and outlining the nowadays challenges arising in this statistical framework. In Chapter 3 we introduce a novel method to improve the computational performances of the ABC–PMC algorithm. Always starting from the ABC–PMC algorithm, in Chapter 4, we develop an ABC based procedure to work with finite mixture models. Chapter 5 takes a different direction, first introducing the state–of–the–art challenges to detecting and characterizing "Earth–like" exoplanets using the radial velocity technique, and then proposing a novel statistical tool based on the Skew Normal distribution to disentangle stellar activity from the pure Doppler signal coming from a hopefully terrestrial planet belonging to the habitable zone.

1.2 Main contributions of the Thesis

The contributions of the Thesis can be summarized as follow:

1. Developments of the ABC–PMC algorithm to adaptively selecting the series of sequential tolerances $\epsilon_{1:T} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)$ in order to improve the efficiency of the sampling, along with an automatic stopping criterion that defines T (i.e. when to stop the algorithm). The proposed adaptive ABC–PMC tolerance selection algorithm can be easily implemented and examples are presented to show how this extension can improve not only the efficiency of the ABC–PMC algorithm but also avoiding to get stuck in local modes. This method, which works by evaluating the online performances of the ABC posterior distribution, is illustrated in Chapter 3.

2. Developments of the ABC–PMC algorithm as an alternative framework for inference to work with finite mixture models. There are several choices to take when implementing an ABC–PMC algorithm to work with finite mixture models, including the selection of a suitable perturbation kernel to move the particles through the iterations (in particular to resample the mixture weights), how to address the label switching problem and the choice of high informative summary statistics. Beyond to discuss and address the required methodological extensions previously summarized, examples are presented to illustrate the performances of the proposed extended ABC–PMC algorithm to work with finite mixture models. This method is discussed in Chapter 4.

3. The CCF is an average of all the absorption lines of a stellar spectrum retrieved by using the radial velocity technique. Stellar activity can be probed by measuring variations in the shape of the CCF as function of time. Those variations are calculated using different parameters of the CCF. To measure with the best precision the necessary parameters is crucial to disentangle exoplanet signals from spurious variations in radial velocity caused by stellar activity. We propose to measure those parameters using a Skew Normal distribution, that compared to the Normal distribution generally used, naturally includes an extra parameter to model the asymmetry of the CCF induced by convective blueshift. By using the Skew Normal distribution the barycenter and skewness of the CCF can be retrieved in a single operation, the correlation between the radial velocity of the star and stellar activity can be better understood and finally the uncertainties associated to all the parameters are smaller than the ones estimated with the classic analysis based on the Normal distribution. This method is presented in Chapter 5.

Chapter 2

Approximate Bayesian Computation Methods

In this Chapter we introduce the ABC framework, the basic ABC algorithm and some of its already available extensions. Section 2.1 motivates the introduction for ABC as a statistical framework for inference. Starting from the basic ABC algorithm, in Section 2.2 we summarize the main challenges that need to be addressed in this framework in order to retrieve a suitable approximation of the true posterior distribution. In Section 2.3 one of the most famous extensions of the basic ABC algorithm, the ABC–PMC algorithm, is introduced. Our final remarks are outlined in Section 2.4.

2.1 Motivations for using Approximate Bayesian Computation

Bayesian inference has become through the last two decades a suitable alternative to the frequentist approach. The relationship between the observed data y_{obs} and parameters $\theta \in \Theta \subseteq \mathbb{R}^p$ (i.e. $p \geq 1$ is the dimension of the parameter space) can be described by the likelihood function $f(y_{\text{obs}} \mid \theta)$. In the Bayesian framework a prior distribution has assigned to the vector of parameters $\theta \sim \pi(\theta)$, representing the subjective belief of the researcher. Bayesian inference is based on the resulting posterior distribution for θ, defined as:

$$\pi(\theta \mid y_{\text{obs}}) = \frac{f(y_{\text{obs}} \mid \theta)\pi(\theta)}{\int_{\Theta} f(y_{\text{obs}} \mid \theta)\pi(\theta)d\theta}, \tag{2.1}$$

where the denominator of Equation (2.1) is known also as the normalizing constant. If the elements of the posterior distribution in Equation (2.1) can be specified, then various techniques can be used to write down the posterior distribution exactly (e.g. if conjugate priors are specified) or approximated using various sampling techniques known

as Markov Chain Monte–Carlo (MCMC) algorithms, such as the Gibbs Sampling [59] and the Metropolis Hastings [65, 95].

Issues arise when the likelihood function cannot be specified. This happens for a variety of reasons such as the relationship between the data and the parameters is highly complex or unknown or if there are features of the data or data collecting procedure that are difficult to incorporate into a likelihood funtion (e.g. complex censuring or truncations). In the cases where it is not feasible to evaluate the likelihood function, ABC provides a framework for inference to obtain an approximation of the true posterior distribution.

In recent years ABC has been applied in many different fields of science, such as biology [137], ecology [7], epidemiology [92], population genetic problems [10, 33, 110, 135] and population modeling [140]. Given the complexity of its models and simulators, Astronomy seems to be a natural field for which an ABC based analysis could result really helpful in order to successfully address a variety of problems. Among the others, ABC has been used in problems such as the simulation of images for weak lensing measurements [1, 25], model analysis of morphological transformation of galaxies [26, 64], TYPE Ia supernovae [73, 75, 150] and for estimating of the luminosity function [128].

2.2 Basic Approximate Bayesian Computation algorithm

ABC provides a framework for inference in situations where the relationship between the data and the parameters is not well-approximated by a tractable likelihood function, but simulation of the data–generating process is possible. The original idea about ABC comes from [42], although its first methodological and philosophical arguments can be found in [125].

Assuming $\theta \in \mathbb{R}^p$ is the inferential target, the basic accept-reject ABC algorithm [110, 135] consists of the four steps outlined in Algorithm 1.

Following the notation of [90], the resulting ABC posterior distribution can be written as:

$$\pi_\epsilon(\theta \mid y_{\text{obs}}) = \int \left[\frac{f(y_{\text{prop}} \mid \theta)\pi(\theta)\mathbb{I}_{A_{\epsilon,y_{\text{obs}}}}(y_{\text{prop}})}{\int_{A_{\epsilon,y_{\text{obs}}}\times\Theta} f(y_{\text{prop}} \mid \theta)\pi(\theta)dy_{\text{prop}}d\theta} \right] dy_{\text{prop}}, \tag{2.2}$$

where $\mathbb{I}_{A_{\epsilon,y_{\text{obs}}}}(\cdot)$ is the indicator function for the set $A_{\epsilon,y_{\text{obs}}} = \{y_{\text{prop}} \mid \rho(y_{\text{obs}}, y_{\text{prop}}) \leq \epsilon\}$.

It then follows that $\pi_\epsilon(\theta \mid y_{\text{obs}}) \approx \pi(\theta \mid y_{\text{obs}})$ for $\epsilon \to 0$, which requires further explanations. The true posterior distribution, as defined in Equation (2.1), is a conditional

Algorithm 1 Basic ABC algorithm for θ

(1) Sample from the prior distribution, $\theta_{\text{prop}} \sim \pi(\theta)$.
(2) Produce a generated sample of the data by using θ_{prop} in the forward simulation model, $y_{\text{prop}} \sim f(y \mid \theta_{\text{prop}})$.
(3) Compare the true data, y_{obs}, with the generated sample, y_{prop}, using a distance function, $\rho(\cdot, \cdot)$, letting $d = \rho(s(y_{\text{obs}}), s(y_{\text{prop}}))$ where $s(\cdot)$ is some (possibly multi-dimensional) summary statistic of the data.
(4) If the distance, d, is smaller than a fixed tolerance, ϵ, then θ_{prop} is retained, otherwise it is discarded. Repeat until the desired particle sample size N is achieved.

distribution given the observed data, y_{obs}. The ABC posterior is not conditioning on the data exactly, but on the data within some tolerance, ϵ. In fact, with continuous data, the probability for the observed dataset y_{obs} to be equal to the simulated dataset y_{prop} is null, making the condition $\epsilon = 0$ unfeasible. Moreover, as outlined in Algorithm 1, comparing the entire observed dataset y_{obs} with the simulated dataset y_{prop} is computationally impractical, implying that a reduction of the parametric space is needed. Reducing the parametric space by selecting a set of suitable summary statistics is not straightforward. Whereas the desirable situation involves working with summary statistics $s(\cdot)$ which also are sufficient, this rarely happens when facing real problems necessitating ABC. Moreover, [90] pointed out that for most situations the summary statistics are usually determined by the problem at hand and chosen by the experimenters in the field, making the implementation of a general procedure for retrieving high informative summary statistics challenging.

It is worth mentioning that, beyond the selection of the summary statistics and the choice for a suitable tolerance, there is a third reason that motivates the approximated nature of the ABC posterior distribution. According to Algorithm 1, once N particles have been accepted, we have samples coming from the true posterior distribution (assuming $s(\cdot)$ high informative and ϵ "small" enough). The ABC posterior distribution is usually retrieved by using some non parametric technique such as the kernel density estimator or other Monte Carlo methods, leading therefore to a third approximation.

In the following of this Section we briefly discuss the first two sources of approximation to deal with in order to obtain an ABC posterior distribution that suitably approximates the true posterior: the selection of high informative summary statistics $s(\cdot)$ and the choice of the suitable tolerance ϵ. An attempt to address the latter challenge will be discussed in Chapter 3.

2.2.1 Selection of informative "enough" summary statistics

The first element that leads to an approximated posterior distribution is the necessary definition for summary statistics $s(\cdot)$, as shown in Equation (2.3).

$$\pi(\theta \mid y_{\text{obs}}) \approx \pi(\theta \mid s(y_{\text{obs}})). \tag{2.3}$$

This approximation is required for computational reasons and thus, rather than using the complete dataset, lower dimensional summary statistics have in general to be defined. To pick suitable summary statistics $s(\cdot)$ is essential to produce useful inference results. In this Section we provide some indication about the challenges related to the selection for $s(\cdot)$, focusing on the importance of using highly informative summary statistics (possibly sufficient) in order to evaluate the quality of the necessary approximation that takes place by using Equation (2.3).

ABC methods suffers of the so called *curse of dimensionality* effect. Hence a suitable reduction of the parametric space has required. Using too many summary statistics or even the entire dataset will result in a too low acceptance rate, forcing the researcher to increase the level of the tolerance ϵ in order to apply Algorithm 1. For this reason, reducing the parametric space by selecting summary statistics $s(\cdot)$ with minimal or none loss of information is one of the most important steps of an ABC based procedure. In particular there is a trade–off between low dimension of the summary statistics and loss of information on the parameters of interest. Balancing this trade–off between selecting a too large number of summary statistics (i.e. fixing then a large tolerance ϵ) and a too small number of summary statistics (i.e. losing information on the parameters of interest) is necessary for successfully retrieving a suitable approximation of the posterior distribution.

Several studies have been done in the attempt to understand how the error in an ABC procedure is related to the dimension of the summary statistics [11, 16, 51]. Among the others, [6] showed that, asymptotically, the rate at which the error decays becomes worse as the dimension of the dataset increases and that, under optimal ABC tuning and regularity conditions, the mean square error of a Monte Carlo estimate produced by using Algorithm 1 is: $O_p(n^{-4}/(q+4))$, where n is the (large) number of simulated datasets and q is the dimension of the summary statistics. It is clear that high dimensional statistics lead in general to an ABC posterior distribution which is a poor approximation of the true one. Further details can be found in [6].

Since a suitable reduction of the parametric space is necessary, ideally the concept of summary statistics in the ABC framework directly joins the one of complete minimal

sufficient statistics. If we were able to define lower dimension complete minimal sufficient statistics as summary statistics, then all the information about the parameters of interest would be preserved using at the same time the best possible reduction of the parametric space. Using the classic definition of sufficient statistics by [35], $s(\cdot)$ is sufficient if $\pi(y_{\text{obs}}|s,\theta) = \pi(y_{\text{obs}}|s)$. An equivalent definition more coherent with the nature of ABC is the *Bayes sufficient* condition. The statistic $s(y_{\text{obs}})$ is said to be *Bayes sufficient* for θ if $\pi(\theta|s(y_{\text{obs}}))$ and $\pi(\theta|y_{\text{obs}})$ have the same distribution for any prior distribution and almost all y_{obs}. If the summary statistics are *Bayes sufficient*, than there is none approximation in Equation (2.3).

However, as pointed out in Section 2.1, the main justification for using ABC is the intractability of the likelihood function. Therefore, for those models requiring ABC, low dimensional sufficient statistics do not generally exist. As noticed in [16], the central question is hence deriving low dimension summary statistics from the observed dataset with minimal loss information. Many methods have been developed for suitably selecting summary statistics. As described in [109], these methods can fall into one of the following three groups: *subset selection*, *projection* and *auxiliary likelihood*. Whereas both subset selection and projection methods require a preliminary step of choosing a set of data features (using respectively some criteria such as AIC and a training set), the auxiliary likelihood method uses an approximated model whose likelihood is more tractable than the model of interest and hence summary statistics are derived from this approximated model. Further details about the already existent methods can be found in [13, 16, 109].

We cannot emphasize more the importance of selecting highly informative summary statistics s$(\cdot)$ in order to retrieve useful inferential results on the parameters of interest. Anyway, since ABC is used when the likelihood function is intractable or computationally too expensive, s$(\cdot)$ will rarely be sufficient. In other words the loss of some amount of information about θ is a first necessary downside to retrieve an approximation of the posterior distribution.

2.2.2 Selection of small "enough" tolerances

The second element that leads to an approximated posterior distribution is caused by the fact that, assuming summary statistics $s(\cdot)$ are used, the ABC posterior distribution as shown in Equation (2.2) is not a conditional distribution given the observed summary statistics $s(y_{\text{obs}})$, but it is given the observed summary statistics within some tolerance. Therefore, the indication function for the set of the accepted particles in Equation (2.2) is of the type $A_{\epsilon,s(y_{\text{obs}})} = \{s(y_{\text{prop}}) \mid \rho(s(y_{\text{obs}}), s(y_{\text{prop}})) \leq \epsilon\}$. This consideration leads to

the following second approximation:

$$\pi(\theta \mid s(y_{\text{obs}})) \approx \pi_\epsilon(\theta \mid s(y_{\text{obs}})). \tag{2.4}$$

From a computational standpoint, fixing the tolerance $\epsilon = 0$ is impractical, since for the vast majority of cases the probability for simulated summary statistics $s(y_{\text{prop}})$ to be equal to the observed summary statistics $s(y_{\text{obs}})$ is null (i.e. if $\epsilon = 0$, the probability for a proposed value θ_{prop} for being accepted is 0, meaning that Algorithm 1 needs an infinitive amount of time for accepting even one single particle among the required N). Moreover, the choice of ϵ depends on the way the distance metric $\rho(\cdot, \cdot)$ has been defined and this definition is not unique (i.e. different distance functions lead to different suitable levels for ϵ). Hence, in its original implementation, to choose in advance the level of the tolerance ϵ is difficult. Usually ϵ is either fixed equal to some small percentile of the simulated distances [10] or defined consistently to the available computational resources [70].

As limit case, if $\epsilon := 0$, then $\pi_\epsilon(\theta \mid s(y_{\text{obs}})) := \pi(\theta \mid s(y_{\text{obs}}))$ and Equation (2.4) is not approximated. Moreover, assuming that $s(y)$ is *Bayes sufficient*, then:

$$\pi(\theta \mid s(y_{\text{obs}})) := \pi(\theta \mid y_{\text{obs}}), \tag{2.5}$$

implying that in this scenario the ABC posterior distribution is not an approximation of the true posterior (i.e. Equation (2.3) is not the results of a double approximation). In other words, if ϵ is equal to 0 and if the summary statistics $s(y)$ are *Bayes sufficient*, the ABC posterior distribution matches the true posterior (see [90] for details). Unfortunately, for both mathematical and computational reasons, the equivalence presented in Equation (2.5) is not achievable unless having discrete data (i.e. the probability for simulated summary statistics $s(y_{\text{prop}})$ to be equal to the observed summary statistics $s(y_{\text{obs}})$ is not null) and a *Bayes sufficient* summary statistics $s(\cdot)$. We note that once N particles have been accepted the ABC posterior distribution is usually obtained by using some non parametric technique such as the kernel density estimator.

In closing, the ABC posterior distribution is generally the result of three approximations. The first approximation, defined in Equation (2.3), is required to reduce the parametric space and a summary statistics $s(\cdot)$ has to be selected. The second approximation, defined in Equation (2.4), is justified by the fact that some discrepancy between the observed summary statistics $s(y_{\text{obs}})$ and the simulated summary statistics $s(y_{\text{prop}})$ has to be allowed. The third approximation, here only introduced, is the Monte Carlo approximation for non parametrically estimating the ABC posterior distribution once

N particles have been accepted, accordingly with the steps highlighted in Algorithm 1.

The nature of the ABC posterior distribution is summarized in a very informative but nonetheless intuitive way by [90]:

> The basic idea behind ABC is that using a representative (enough) summary statistic $s(\cdot)$ coupled with a small (enough) tolerance ϵ should produce a good (enough) approximation to the posterior distribution.

2.2.3 A first example: the Normal–Normal model

We end the discussions about the approximated nature of the ABC posterior distribution by introducing a simple first example. In the following, we implemented the basic ABC algorithm using a Normal likelihood with unknown mean θ and known variance $\sigma^2 = 1$. The sample size of the observed data y_{obs} is $n = 100$ and as prior distribution we use a Normal distribution having mean $\theta_0 = 0$ and variance $\sigma_0^2 = 100$. We used $\theta = 0$ for generating y_{obs}. Since the prior distribution is a conjugate prior for the likelihood function, in this case the true posterior distribution is analytically available (i.e. a close form for $\pi(\theta \mid y_{\text{obs}})$ is retrievable), providing a benchmark to evaluate the performance of the basic ABC algorithm. For this specific example, to have the exact posterior distribution is also helpful to evaluate the loss of information caused by the definition of insufficient summary statistics. We moreover note that when working with this model the ABC posterior distribution can be analytically retrieved. In Appendix A we present the mathematical details to obtain a close form for the ABC posterior distribution when the model follows a Normal distribution.

Using the notation of Equation (2.1), the likelihood function is

$$f(y_{\text{obs}} \mid \theta) \propto \exp\left(-\frac{n}{2\sigma^2}\left(\bar{y} - \theta\right)^2\right) \sim N(\bar{y}|\theta, \frac{\sigma^2}{n}), \tag{2.6}$$

and the conjugate prior distribution has the form

$$\pi(\theta) \propto \exp\left(-\frac{n}{2\sigma_0^2}\left(\theta - \theta_0\right)^2\right) \sim N(\theta|\theta_0, \sigma_0^2). \tag{2.7}$$

According to Equation (2.1), the posterior distribution is easily retrievable in close form as

$$\pi(\theta \mid y_{\text{obs}}) \sim N\left(\frac{\left(\frac{\theta_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}\right). \tag{2.8}$$

In order to initialize the basic ABC algorithm, we first define the desired particles sample size equal to $N = 1000$. Concerning the selection of a high informative summary statistic, in this case the complete minimal sufficient statistic, $s(y_{\text{obs}}) = \sum_{i=1}^{n} y_i$, is available. To show the consequences for using an insufficient summary statistic, we run the basic ABC algorithm using also the first and the third quartiles as proposed summaries. The definition of the distance metric is $\rho = (s(y_{\text{obs}}), s(y_{\text{prop}})) = \frac{|s(y_{\text{obs}}) - s(y_{\text{prop}})|}{n}$. Since in this example our main goal is to show the consequences on the ABC posterior distribution for using poorly informative summary statistics, the allowed tolerance for all the analyses is $\epsilon = 0.01$ (which is the average distance ρ if y_{prop}'s are 100 samples from a Normal having as input parameter the true parameter $\theta = 0$). The obtained ABC posterior distributions, compared with the true one, are displayed in Figure 2.1. When using as summary statistic the complete minimal sufficient statistic $s(y_{\text{obs}}) = \sum_{i=1}^{n} y_i$, there is none loss of information on θ because of the definition of $s(\cdot)$ and the ABC posterior distribution is comparable with the true one. However, the resulting ABC posterior distribution has two approximations: the first one is the result of using a tolerance $\epsilon = 0.01$ and as second the ABC posterior distribution is obtained by using the kernel density estimator. In Chapter 3 we will discuss in detail the role played by the tolerance ϵ in any ABC based analysis, suggesting ways to properly and automatically define it. When an insufficient statistic is used, the ABC posterior is a poor approximation of the true posterior. The ABC posterior distribution is biased respect the true one and the posterior variance is larger, which reinforces the assumption that some amount of information about θ got lost. Smaller values for the tolerance have been tested when the summary statistic is insufficient; the corresponding ABC posteriors have not improved, suggesting that too much information was lost by using respectively the first and the third quartile as summary statistic for the presented model. Finally, with running the basic ABC algorithm with insufficient summary statistics, we noted that, for an equally fixed tolerance ϵ, the computational time needed in order to accept N particles from the prior distribution drastically increased respect to the case that used as summary statistic the complete minimal sufficient statistic.

2.3 Approximate Bayesian Computation Population Monte–Carlo algorithm

As pointed out in the work by [90], when using non informative priors the basic ABC algorithm can be very inefficient, because simulations from $\pi(\theta)$ do not account for the data at the proposal stage, leading to proposed values located in low posterior

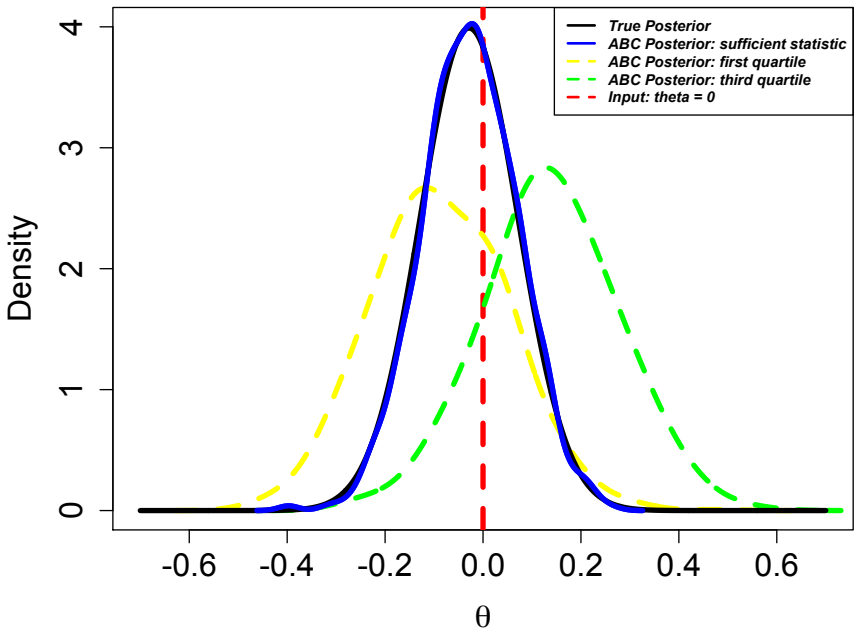

Figure 2.1: Comparison between the true posterior distribution (black line) and the ABC posterior distribution using as summary statistic the mean (blue line), the first quartile (yellow dots) and the third quartile (green dots). When using the complete minimal sufficient statistic, the ABC posterior distribution is comparable with the true posterior. When picking an insufficient summary statistic, the resulting ABC posterior distribution is not a suitable approximation of the true posterior. The tolerance ϵ is fixed for all the cases equal to 0.01. Smaller values for the tolerance have been tested when the summary statistic is insufficient. The corresponding ABC posteriors have not improved, suggesting that too much information was lost by using respectively the first and the third quartile as summary statistic for the presented model.

probability regions. On top of that, picking in advance a suitable tolerance ϵ is unfeasible. For these reasons there have been many extensions to the basic ABC algorithm [12, 13, 36, 44, 51, 62, 76, 90, 98, 116]. In this Thesis we focus on the ABC–PMC approach originally introduced by [8].

The ABC–PMC algorithm is based on importance sampling ideas in order to improve the efficiency of the algorithm by constructing a series of intermediate distributions; the steps are displayed in Algorithm 2. The first iteration of the ABC–PMC algorithm uses tolerance ϵ_1 and draws proposals from the specified prior distribution(s); the resulting ABC posterior is π_{ϵ_1}. Rather than starting the algorithm over from the beginning using a smaller ϵ, the algorithm proceeds sequentially by drawing proposals from the previous iteration's ABC posterior, $\pi_{\epsilon_{t-1}}$. After a particle is selected from the previous iterations particle system, it is moved according to some kernel (e.g. a Gaussian kernel). Since

the proposals are not drawn directly from the prior distribution(s), importance weights are used. The importance weight for particle $J = 1, \ldots, N$ in iteration t is

$$W_t^{(J)} \propto \pi(\theta_t^{(J)}) / \sum_{K=1}^{N} W_{t-1}^{(K)} \phi \left[\tau_{t-1}^{-1} \left(\theta_t^{(J)} - \theta_{t-1}^{(K)} \right) \right],$$

where $\phi(\cdot)$ is a Gaussian kernel with variance τ_{t-1}^2 – twice the (weighted) sample variance of the particles from iteration $t-1$ as recommended in [8]. However, other choices beyond the Gaussian perturbation kernel are possible. For instance [140] proposed an Uniform perturbation kernel in their ABC–PMC based analysis. We note however that, regardless which perturbation transformation kernel has been defined, the importance weights must be accordingly calculated, in order to reflect the fact that to propose new candidates the prior distributions are not directly used. Other proposals for the sequence of intermediate distributions in the ABC–PMC algorithm can be found in [20, 38, 55, 79, 139].

While the proposals are drawn from a more informative distribution, the tolerances also decrease such that $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$, where T is the final iteration. Both the rule to reduce the tolerances and the total number of iterations T are selected in advance from the researcher. In particular, when working with a sequential ABC algorithm, the series of the tolerances sequence $\epsilon_{1:T}$ is selected either by fixing the values in advance [92, 131, 140], or adaptively selecting ϵ_t based on some quantile of $\{d_{t-1}^{(J)}\}_{J=1}^{N}$, the distances of the accepted particles from iteration $t-1$ [8, 70, 85, 150]. After determining the sequence of tolerances, it also has to be determined when to stop an ABC algorithm. An ABC algorithm is often stopped when either a desired (low) tolerance is achieved [131] or once a fixed number of iterations T is reached [8]. [70] showed that once the ABC posterior stabilizes, further reductions of the tolerance lead to low acceptance rates without meaningful improvement in the ABC posterior; they stop the algorithm once the acceptance rate drops below a threshold set by the user.

Both the selection of the series of decreasing tolerances and the stopping rule cover an important role to determine the efficiency of the ABC–PMC algorithm and the achievability of a suitable approximation of the true posterior distribution. Because of the latter reason, beyond the goal of improving the computational efficiency of the ABC–PMC algorithm, another goal of this Thesis is to provide further investigations about the behavior of the ABC–PMC algorithm respect the presence of local modes, as will be discussed in Chapter 3.

Algorithm 2 ABC–PMC algorithm for θ

if $t = 1$ **then**
  for $J = 1, \ldots, N$ **do**
    Set $d_1^{(J)} = \epsilon_1 + 1$
    while $d_1^{(J)} > \epsilon_1$ **do**
      Propose $\theta^{(J)}$ by drawing $\theta_{\text{prop}} \sim \pi(\theta)$,
      Generate $y_{\text{prop}} \sim f\left(y \mid \theta^{(J)}\right)$
      Calculate distance $d_1^{(J)} = \rho(s(y_{\text{obs}}), s(y_{\text{prop}}))$
    end while
    Set weight $W_1^{(J)} = N^{-1}$
  end for
else if $2 \leq t \leq T$ **then**
  Set $\tau_t^2 = 2 \cdot \text{var}\left(\{\theta_{t-1}^{(J)}, W_{t-1}^{(J)}\}_{J=1}^{N}\right)$
  for $J = 1, \ldots, N$ **do**
    Set $\epsilon_t = q^{th}$ quantile of $\{d_{t-1}^{(J)}\}_{J=1}^{N}$ (using an adaptively-selected tolerance sequence)
    Set $d_t^{(J)} = \epsilon_t + 1$
    while $d_t^{(J)} > \epsilon_t$ **do**
      Select θ_t^* from $\theta_{t-1}^{(J)}$ with probabilities $\left\{W_{t-1}^{(J)} / \sum_{K=1}^{N} W_{t-1}^{(K)}\right\}_{J=1}^{N}$
      Propose $\theta_t^{(J)} \sim \mathcal{N}(\theta_t^*, \tau_t^2)$
      Generate $y_{\text{prop}} \sim f\left(y \mid \theta_t^{(J)}\right)$
      Calculate distance $d_t^{(J)} = \rho(s(y_{\text{obs}}), s(y_{\text{prop}}))$
    end while
    Set weight $W_t^{(J)} \propto \pi(\theta_t^{(J)}) / \sum_{K=1}^{N} W_{t-1}^{(K)} \phi\left[\tau_{t-1}^{-1}\left(\theta_t^{(J)} - \theta_{t-1}^{(K)}\right)\right]$
  end for
end if

2.4 Concluding remarks

In this Chapter we introduced ABC methods, motivating their success in addressing modern statistical problems and discussing the state–of–the–art methodological challenges related to this novel framework for statistical inference. In particular the selection for highly informative summary statistics $s(\cdot)$ and the determination for a suitable tolerance ϵ are two choices always mandatory when running an ABC analysis. These two choices have a huge impact on the way the ABC posterior distribution properly approximates the true posterior.

Starting from the ABC–PMC algorithm introduced in Section 2.3 we developed two extensions, whose contents will be provided in Chapter 3 and Chapter 4. In the first extension we present a method for automatically and efficiently selecting the series of

tolerances, $\epsilon_{1:T} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)$, along with determining T (i.e. when to stop the algorithm). All the necessary quantities are based on the online performances of the ABC posterior distribution and the number of arbitrary selections required by the researcher is minimized. In the second work we propose extensions of the ABC–PMC algorithm as an alternative framework for inference for working with finite mixture models.

Chapter 3

Adaptive Approximate Bayesian Computation Tolerance Selection

In this Chapter we extend the ABC–PMC algorithm [8] introduced in Chapter 2, so that the quantile used to update the tolerance in each iteration, q_t, is automatically and efficiently selected (rather than fixed in advance to some quantile that is used for each iteration). Note that efficiency is not only a matter of having a high acceptance rate (as this can easily be accomplished by using larger quantiles), but rather a balance between the acceptance rate and a suitable amount of shrinkage of the tolerance. Moreover the series of tolerances needs to be selected in such a way that the algorithm avoids getting stuck in local modes. Secondly, we propose an automatic stopping rule directly based on the behavior of the ABC posterior distribution. The proposed extensions work best for situations where ABC is required (i.e. the likelihood function is intractable) and the number of parameters is not huge. Among the others, examples of such situations can be found in [1, 62, 73, 75, 139].

In Section 3.1 we highlight the most common used strategies to run the ABC–PMC algorithm. In Section 3.2 we propose a method to adaptively select the sequential tolerances that improves the computational efficiency of the algorithm over other common techniques, while in Section 3.2.1 an automatic stopping rule is proposed. The proposed adaptively ABC–PMC tolerance selection algorithm can be easily implemented and several examples are presented in Section 3.3 to show not only that this extension can improve the efficiency (in terms of computational time and number of draws from the forward model) of the ABC–PMC algorithm but also that a suitable selection of the series of tolerances is necessary to avoid getting stuck in local modes. Our final conclusions are outlined in Section 3.4.

3.1 Introduction

When working with a sequential ABC algorithm, there are two common approaches for selecting the tolerance sequence, $\epsilon_{1:T}$: (i) fixing the values in advance [92, 131, 140], or (ii) adaptively selecting ϵ_t based on some quantile of $\{d_{t-1}^{(J)}\}_{J=1}^{N}$, the distances of the accepted particles from iteration $t-1$ [8, 70, 85, 150]. These approaches can lead to inefficient sampling as discussed below and demonstrated in the simulation study presented in Section 3.3. On top of that, it turns out that selecting tolerances using a predetermined quantile can, if not selected wisely, lead to the particle system getting stuck in local modes [130]. Hence the exact sequence of tolerances has an impact not only on the computational efficiency of the algorithm but also on the achievement of the true posterior. We emphasize that finding the true posterior distribution using ABC is not guaranteed and depends on a number of choices, including the careful selection of summary statistics. However, in the following, we assume that the summary statistics preserves sufficiency, focusing on the approximation caused by the tolerance ϵ.

In a recent work, [130] propose an adaptive approach for selecting the tolerance sequence at each iteration by estimating the threshold-acceptance rate curve (TAR curve), which is used to balance the amount of shrinkage of the tolerance with the acceptance rate. This approach requires, at each time step, the estimation of the TAR curve. The naive, computationally impractical approach to estimate the TAR curve (noted as such in [130]) is to simulate the acceptance rate at a range of difference tolerances; this would have to be repeated at each time step of the ABC algorithm. Instead, they suggest a method for estimating the TAR curve by building an approximation to the ABC simulation model (in their example, using a mixture of Gaussians and the unscented transform of [77]). The TAR curve approach is able to avoid local optima values, but requires the extra step of building a fast approximation of the ABC data-generating model. Our proposed algorithm similarly is able to avoid local modes, but uses quantities that are directly and readily available in the algorithm. More details are presented in Section 3.3.

After determining the sequence of tolerances, it also has to be determined when to stop an ABC algorithm. An ABC algorithm is often stopped when either a desired (low) tolerance is achieved [131] or after a fixed number of iterations T [8]. [70] showed that once the ABC posterior stabilizes, further reduction of the tolerance leads to low acceptance rates without meaningful improvement in the ABC posterior; they stop the algorithm once the acceptance rate drops below a threshold set by the user.

3.2 Automatic tolerance selection

Using the same quantile to update the tolerance at each time step can be computationally inefficient and could result in the particle system getting stuck in local modes (see example in Section 3.3.4). In this Section we introduce a method to adaptively select the quantile such that each iteration will have its own quantile, q_t, set based on the online performance of the algorithm.

In order to initialize the tolerance sequence consider the following. Let N be the desired number of particles to approximate the posterior. The initial tolerance ϵ_1 can be adaptively selected by sampling kN draws from the prior, for some $k \in \mathbb{Z}^+$. Then the N particles of the kN total particles with the smallest distances are retained, and $\epsilon_1 = \max\left(d_1^{(1*)}, \ldots, d_1^{(N*)}\right)$, where $d_1^{(1*)}, \ldots, d_1^{(N*)}$ are the N smallest distances of the kN particles sampled. This initialization procedure effectively selects a quantile for the first step in selecting an appropriate k, but this adaptive first step can be easier to work with than trying to guess at a good ϵ_1 (which can be especially challenging when testing different summary statistics or distance functions because the scale of the distances can be different). We note that k must be large enough for an initial exploration of the parameter space by the ABC algorithm; insufficient initial exploration of the parameter space can lead to getting stuck in local regions of the parameter space. This is true in general for ABC algorithms, including when ϵ_1 is predefined (not set adaptively). The selection of k is discussed in Section 3.3.

For subsequent tolerances, $\epsilon_{2:T}$, the general idea is to gauge the amount of shrinkage for iteration $t+1$ by setting ϵ_{t+1} based on the amount of improvement between $\hat{\pi}_{\epsilon_{t-1}}$ and $\hat{\pi}_{\epsilon_t}$. In particular, we can use the estimated ABC posteriors to select a quantile for updating the tolerance for the next iteration, and adjust the next tolerance based on how slowly or rapidly the sequential ABC posteriors are changing. More specifically, after each iteration $t > 1$, the following can be estimated using the weighted particles

$$\hat{c}_t = \sup_{\theta} \frac{\hat{\pi}_{\epsilon_t}(\theta)}{\hat{\pi}_{\epsilon_{t-1}}(\theta)}. \tag{3.1}$$

Since $\hat{\pi}_{\epsilon_{t-1}}$ and $\hat{\pi}_{\epsilon_t}$ are both proper densities, they will be either exactly the same, making $\hat{c}_t = 1$ or there must be a place where $\hat{\pi}_{\epsilon_t} > \hat{\pi}_{\epsilon_{t-1}}$, making $\hat{c}_t > 1$. We note that $\hat{c}_t$ has a lower limit equal to 1 since $\epsilon_t \leq \epsilon_{t-1}$ and generally the variance of the ABC posterior distribution decreases until convergence to the true posterior has reached, as shown in Appendix A, which is achieved when the ABC posterior is no longer changing sequentially. Then the proposed quantile for iteration t (in order to determine ϵ_{t+1}) is

$$q_t = \frac{1}{\hat{c}_t}, \tag{3.2}$$

which varies between 0 and 1. Small values of q_t imply q_{t-1} lead to a large improvement between $\hat{\pi}_{\epsilon_{t-1}}$ and $\hat{\pi}_{\epsilon_t}$, which then results in a larger percentage reduction of the tolerance for the coming iteration, $t+1$. On the other hand, once the ABC posterior stabilizes, q_t tends to 1 as $\hat{\pi}_{\epsilon_{t-1}}$ gets closer to $\hat{\pi}_{\epsilon_t}$. An illustration of the proposed quantile selection is provided in Figure 3.1. For iteration $t+1$, if $\hat{\pi}_{t-1}$ was still used as the proposal for iteration $t+1$, then q_t would be the percentage decrease in the acceptance rate from iteration t (i.e. if acc_t is the acceptance rate for iteration t, then acc_{t+1} would be approximately $q_t \times \text{acc}_t$). However, we are not proposing from $\hat{\pi}_{t-1}$, but rather $\hat{\pi}_t$ so the *decrease* in the acceptance rate is mitigated by the *improvement* in the proposed particles from iteration t. When there is a large improvement in ABC posteriors from $\hat{\pi}_{t-1}$ to $\hat{\pi}_t$, then q_t is smaller (allowing for a larger drop in tolerance); this larger percentage drop in tolerance does not result in an equal percentage drop in acceptance rate because the new proposal distribution, $\hat{\pi}_t$, is better than $\hat{\pi}_{t-1}$. Conversely, if $\hat{\pi}_{t-1}$ is close to $\hat{\pi}_t$, then the improvement in the ABC posterior is not enough to allow for a large decrease in the acceptance rate so q_t is closer to 1.

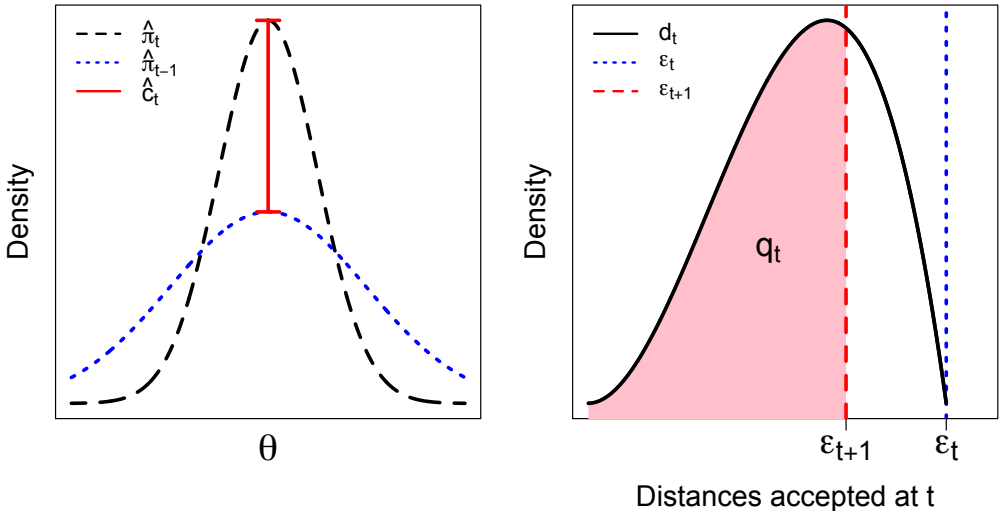

FIGURE 3.1: Illustration of selection of q_t. (left) The proposal distribution ABC posterior $\hat{\pi}_{t-1}$ and the resulting ABC posterior $\hat{\pi}_t$, with $\hat{c}_t$ is defined in Equation (3.1) and used for setting q_t as defined in Equation (3.2). (right) The (arbitrary) distribution of distances is from the accepted distances at iteration t, $\{d_t^{(J)}\}_{J=1}^N$, with ϵ_t being the largest possible value. The next iterations tolerance, ϵ_{t+1}, is set as the q_t quantile of $\{d_t^{(J)}\}_{J=1}^N$.

The acceptance rate is also useful for evaluating the computational effort of the ABC–PMC algorithm, defined as

$$\text{acc}_t = \frac{N}{\text{Draws}_t}, \tag{3.3}$$

where Draws_t is the number of draws taken at iteration t in order to produce N accepted values. We note that Equation (3.3) decreases because, as the tolerance decreases, the number of elements Draws_t required to get N accepted particles generally increases.

3.2.1 Stopping rule

There are several ideas on how to determine the number of iterations to use for an ABC–PMC algorithm. Often one picks some T based on the computational resources available, but this can be needlessly inefficient. [70] proposed to stop the algorithm once the acceptance rate is smaller than some specified, fixed tolerance. We extend this idea, discuss why to stop the algorithm once the acceptance rate is smaller than some arbitrary pre–specified threshold can be inefficient, and suggest a new stopping rule directly based on the estimated sequential ABC posterior distribution.

Again using a similar form as Equation (3.1), we evaluate the stability of the ABC posterior using

$$\hat{C}_t = \sup_{\theta} \frac{\hat{\pi}_{\epsilon_t}(\theta)}{\pi(\theta)}, \tag{3.4}$$

where the denominator is the prior distribution and the numerator is the ABC posterior at iteration t. Once the sequential ABC posteriors stops changing significantly, the series of $\hat{C}_t$ obtained through the iterations stabilizes – small changes are due to the estimation of $\hat{\pi}_{\epsilon_t}$ at the end of each iteration.

The goal is to stop the procedure as soon as the ABC posterior has stabilized, and $1/\hat{C}_t$ generally decreases as the tolerance decreases because the ABC posterior is looking less like the prior. Once the ABC posterior stabilizes, the $1/\hat{C}_t$ will stop monotonically decreasing and further reductions of the tolerance (i.e. further iterations) do not necessarily lead to an improvement by the ABC posterior distribution, but rather fluctuations due to variability of the estimated ABC posterior. This leads to an automatic and simple stopping rule, which is employed starting from the third iteration (i.e. once the transformation kernel was run twice to avoid premature stopping): the algorithm is stopped at time t when

$$\frac{1}{\hat{C}_t} > \frac{1}{\hat{C}_{t-1}}, \text{ for } t \geq 3. \tag{3.5}$$

Hence, the algorithm is stopped once the monotonicity of the updates in the sequential ABC posterior is violated, suggesting that changes are due to its estimation variability.

Using Equation (3.5) as stopping rule and Equation (3.2) as an automatic rule to shrink the tolerance, the ABC–PMC algorithm is stopped once additional time steps with smaller tolerances do not lead to significant changes in the ABC posterior. We also note that this coincides with the stabilization of the particle variance.[1]

Estimation of Equation (3.1) and Equation (3.4) requires estimation of $\hat{\pi}_{\epsilon_t}$ and $\hat{\pi}_{\epsilon_{t-1}}$, which can be done, for example, using kernel density estimation. Though kernel density estimation suffers from the curse of dimensionality, this is a reasonable approach for problems with lower dimensional parameter spaces such as the model considered below in Section 3.3. Other examples in which the ABC approach is employed for addressing problems with low dimensional parameter space are [31, 32, 72, 74]. We tried different kernels when evaluating Equation (3.1) and Equation (3.4), obtaining comparable results. For this reason we used the default Gaussian kernel provided by the function density in **R**. The smoothing bandwidth parameter has not been fixed in advance.

3.3 Illustrative Examples

Next we provide a comparison between the classic ABC–PMC algorithm and our extension proposed in Section 3.2, the adaptive ABC–PMC tolerance selection algorithm (aABC–PMC), with five examples. The first two examples consider discrete (Beta-Binomial) and continuous (Exponential-Gamma) models. We work with Exponential Families because we have the complete minimal sufficient statistics, $s(y)$, which allows the study to focus on the proposed method rather confounding the overall performances with determining reasonable summary statistics. Moreover, by using the conjugate prior distribution, the true posterior distribution is available in closed form, providing a benchmark to evaluate the behavior of both the ABC–PMC and our extension. In the third example the Gaussian mixture model by [131] is used in order to show the efficiency of the proposed aABC–PMC procedure. Then the aABC–PMC algorithm is used for a model from [130], which has local modes, in order to illustrate how the proposed automatic tolerance selection is able avoiding to get stuck in local modes. The final example comes from a population modeling problem with the Lotka–Volterra model by [140] in which the forward model for the analysis is computationally expensive.

[1] The desired sample size N has an impact on the evaluation of Equation (3.5). This problem arises also in the classic MCMC analysis when determining the length of the MCMC chain [58]. An N that is too low leads to more variability of the estimated posterior in Equation (3.4), which could lead to the algorithm stopping prematurely. Further discussions on the role played by the desired particle sample size N are presented in Appendix B.

Expensive forward models are a challenge for ABC methods because the computational cost can be prohibitive. For those cases, selecting an appropriate sequence of tolerances is crucial.

In order to compare the proposed procedure with the classic ABC–PMC algorithm, both the computational time and the total number of draws required to verify the stopping rule are considered. The measure used for evaluating the similarity between the ABC posterior distribution at the iteration t, $\hat{\pi}_{\epsilon_t}$, and the benchmark, π_{true}, is the Hellinger distance, which is defined as

$$H(\hat{\pi}_{\epsilon_t}, \pi_{\text{true}}) = \left(\int \left(\sqrt{\hat{\pi}_{\epsilon_t}(y)} - \sqrt{\pi_{\text{true}}(y)} \right)^2 dy \right)^{\frac{1}{2}}. \tag{3.6}$$

The benchmark π_{true} is the true posterior distribution if it is analytically retrievable (as it is for the first four presented examples). In the last proposed example, the true posterior distributions are not available so the ABC posteriors from [140] are used as benchmarks.

3.3.1 Beta-Binomial Model

In this first example we compare the ABC–PMC algorithm and our proposed extension for the discrete Beta-Binomial model. The parameter of interest is p, the probability of success in $n = 100$ independent replications. The likelihood function is available and follows the law:

$$f(y \mid p) = \binom{n}{s} p^s (1-p)^{n-s},$$

where $s = \sum_{i=1}^{n} y_i$ is the complete minimal sufficient statistics and $y_i \in \{0, 1\}$.

As prior distribution a Beta with hyper parameters $a = 1$ and $b = 1$ is used, since it is conjugate to the likelihood function.

$$\pi(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}.$$

The posterior distribution is available analytically as:

$$\pi(p \mid Y) \propto f(Y \mid p)\pi(p) \propto p^s(1-p)^{n-s} p^{a-1}(1-p)^{b-1} \propto p^{s+a-1}(1-p)^{n-s+b-1}, \tag{3.7}$$

which is the kernel of a Beta distribution with updated hyperparameters $a^* = a + s$ and $b^* = b + n - s$.

The summary $s(y)$ of the data is the complete minimal sufficient statistics $s(y) = \sum_{i=1}^{n} y_i$ and the distance function to evaluate the separation between the real and the

simulated dataset is defined as:

$$\rho\left(y_{\text{obs}}, y_{\text{prop}}\right) = \frac{\mid s(y_{\text{obs}}) - s(y_{\text{prop}}) \mid}{n}. \tag{3.8}$$

In order to start the aABC–PMC algorithm, the only choice concerns the desired particle sample size N, which we fixed to 2000. The quantile to update the tolerance is automatically selected and the stopping rule does not need any specification about a final number of iterations T or a tolerance on the lower limit of the acceptance rate to continue with the procedure, since it is directly based on the behavior of the $\hat{\pi}_{\epsilon_t}$.

Figure 3.2 shows the results of the analysis conducted by updating the quantile according to Equation (3.2) and by evaluating Equation (3.5) to arrest the procedure. The posterior distribution, displayed in Figure 3.2(left), is reached after 5 iterations and the Hellinger distance (H_{dist}) between the true posterior distribution and the final ABC posterior is 0.032. The series of automatically selected quantiles is: $q_{2:5} = (0.22, 0.55, 0.87, 0.88)$, which leads to the series of tolerances $\epsilon_{1:5} = (0.1, 0.02, 0.01, 0.01, 0.01)$. The q_t's retrieved by using Equation (3.2) are displayed in Figure 3.2(right)(black circles), which generally increase until the final iteration. Once the posterior distribution is reached by the algorithm, $1/\hat{C}_t$ stabilizes, as shown in Figure 3.2(right)(orange crosses). Since we are using a discrete model and the tolerance stabilizes starting from the third iteration, the acceptance rate (blue triangles) stabilizes as well.

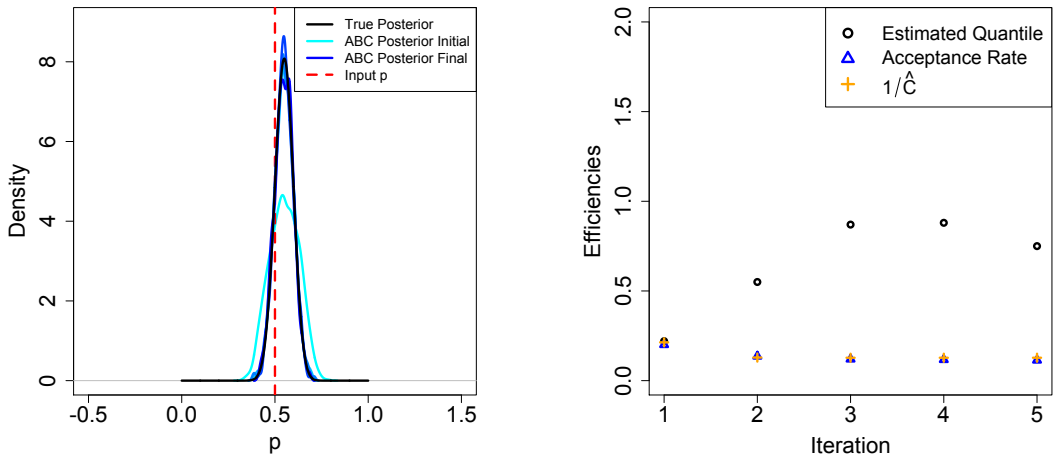

FIGURE 3.2: aABC–PMC analysis for the discrete Beta-Binomial model with $N = 2000$. (left) Series of 5 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of automatically selected quantiles: $q_{2:5} = (0.22, 0.55, 0.87, 0.88)$, that lead to the series of tolerances $\epsilon_{1:5} = (0.1, 0.02, 0.01, 0.01, 0.01)$. The automatic stopping rule directly based on the behavior of the ABC posterior distribution is satisfied after 5 iterations.

For the standard ABC–PMC algorithm the tolerance sequence has to be selected (either adaptively or fixed to particular values). Let us at first to select the quantile used to shrink the tolerance equal to 0.75. The number of iterations T must be selected, which can lead to a poor approximation of the posterior distribution if it is too small or to an inefficient algorithm if it is too large. With the chosen quantile, we fixed $T = 15$. Looking at Figure 3.3(right) the behavior is consistent with the fact that the convergence

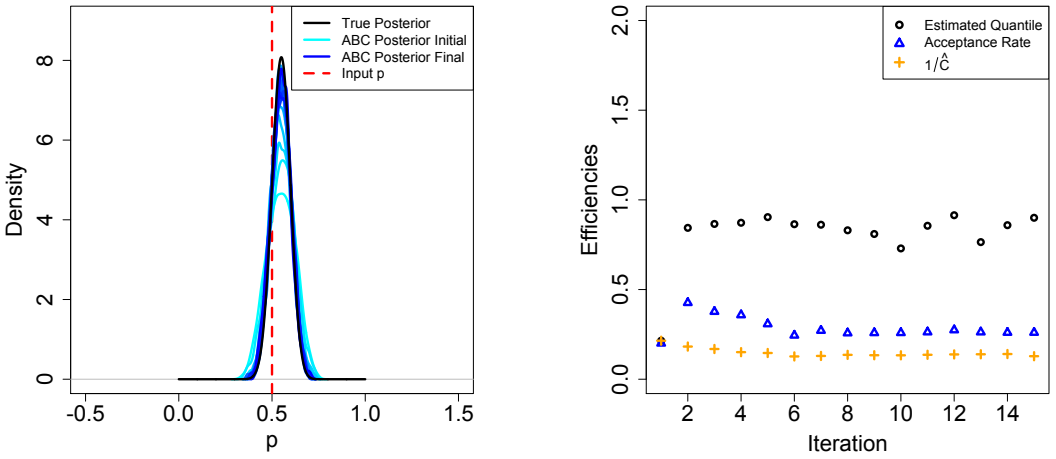

FIGURE 3.3: ABC–PMC analysis for the discrete Beta-Binomial model with $N = 2000$, $q^{th} = 0.75$ and $T = 15$. (left) Series of 15 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of efficiencies based on the selection in advance of the quantile used to reduce the tolerance through the 15 iterations.

has been reached. In fact, the tolerances do not decrease beyond the 6^{th} iteration ($\epsilon_{6:15} = 0.03$). This situation explains why fixing a lower limit for the acceptance rate in order to to arrest the procedure could be inefficient, since in the discrete case that limit could take an excessive amount of time without noticeable benefit. When a continuous model is used, the tolerance continues to decrease through the iterations and the observed acceptance rate decreases as well. Once the posterior distribution is reached by the algorithm, $1/\hat{C}_t$ stabilizes, as shown in Figure 3.3(right)(orange crosses). Figure 3.3(left) shows the evolution of the ABC posterior distribution by the ABC–PMC procedure through all the 15 iterations, suggesting how once the procedure stopped, the true posterior distribution has been reached ($H_{dist} = 0.071$).

To evaluate the reliability of the aABC–PMC, a comparison with the ABC–PMC is done, both in terms of computational time of the entire procedure and total number of draws needed to obtain N accepted values from the final ABC posterior distribution. A simulation study was performed based on 20 independent runs with the same dataset. An average of both the total number of draws needed to accept N values and

the computational time for the aABC–PMC is computed. The average computational time is 125.990 sec. and the average total number of draws is 54889. To compare the performance of aABC–PMC with ABC–PMC, ABC–PMC was run with 3 different fixed quantiles, $q^{th} = 0.25, 0.5, 0.75$. For each quantile, ABC–PMC was run independently 20 times with the total number of draws fixed at the aABC–PMC average of 54889. A second set of 20 independent runs of ABC–PMC was carried out fixing the computational time at the aABC–PMC average of 125.990 sec. Figure 3.4 shows the results for the three quantiles when the total number of draws allowed is 54889. The estimated Hellinger distances are $(0.068, 0.058, 0.091)$ for $q^{th} = 0.25, 0.5, 0.75$ respectively, and only $q^{th} = 0.25$ has not converged based on the proposed stopping rule based on Equation (3.4). The results for the fixed computational time 125.990 sec. are displayed in Figure 3.5. The estimated Hellinger distances are $(0.14, 0.037, 0.16)$ for $q^{th} = 0.25, 0.5, 0.75$ respectively, and only $q^{th} = 0.5$ has converged based on the proposed stopping rule based on Equation (3.4).

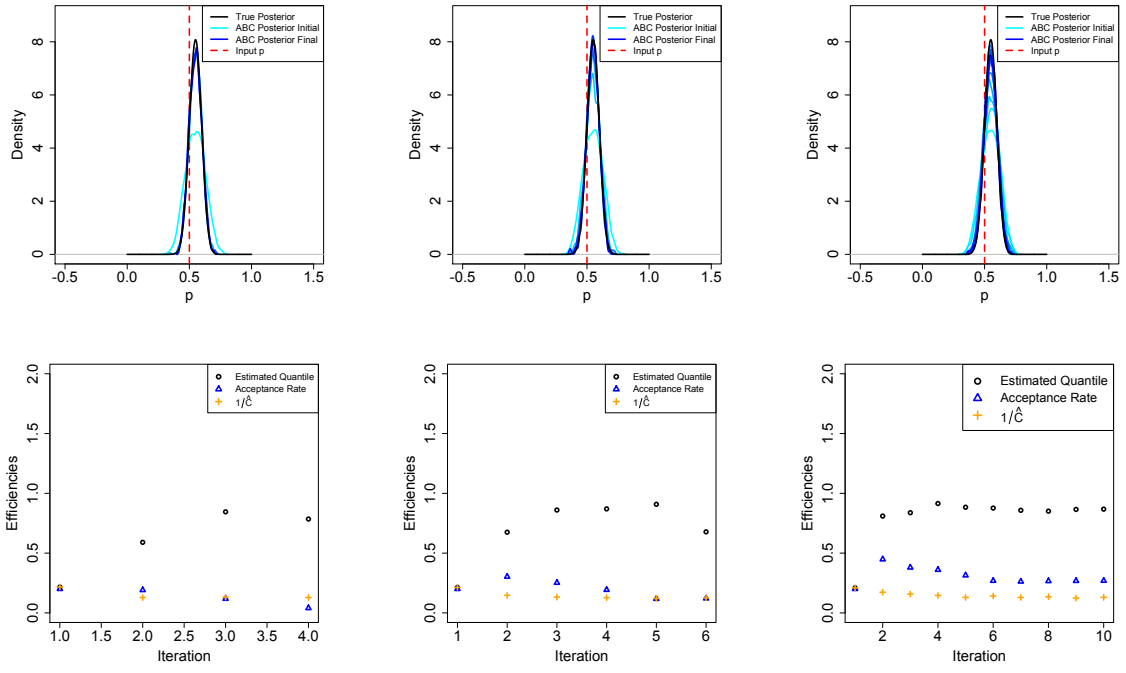

FIGURE 3.4: ABC–PMC analysis for the Beta-Binomial model with $N = 2000$ and maximum number of allowed draws equal to 54889, for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.068, 0.058, 0.091)$.

We conclude the analysis showing the behavior of the aABC–PMC algorithm for different choices of the number of proposed values from the prior distribution at the first

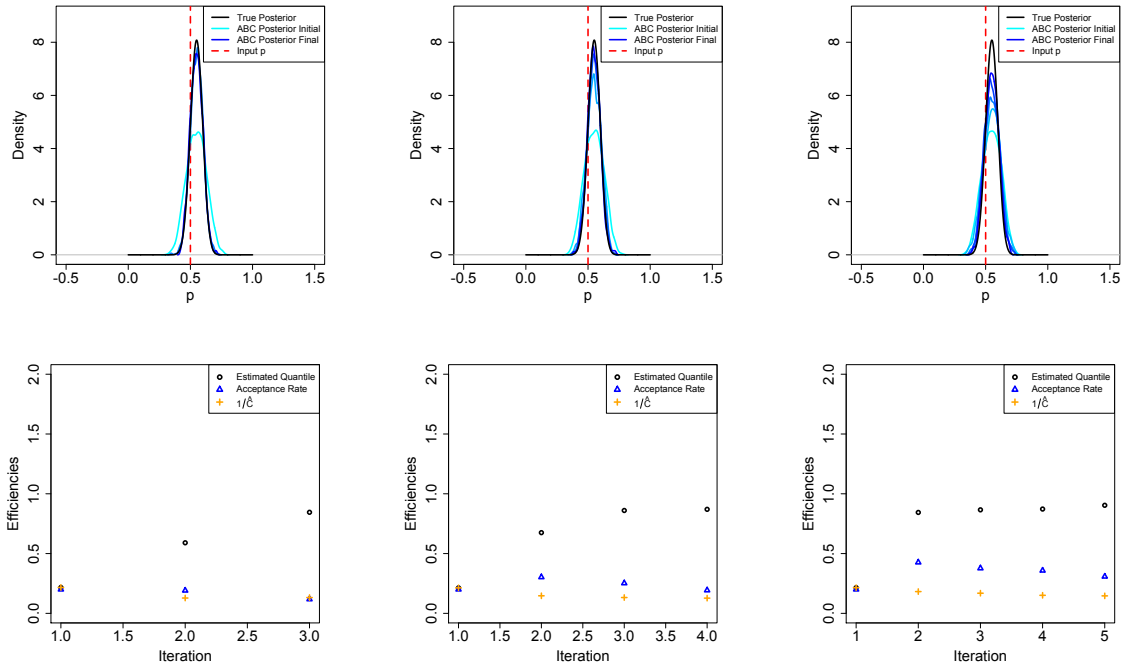

FIGURE 3.5: ABC–PMC analysis for the Beta-Binomial model with $N = 2000$ and time limit equal to 125.990 sec., for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.14, 0.037, 0.16)$.

iteration of the procedure. We recall that the first tolerance ϵ_1 is selected by oversampling from the prior distribution by a factor K. The choice of K is not straightforward, since a K too large implies that the prior distribution is largely used, while a K too small can fail to explore relevant regions of the parametric space. N, 2N, 5N, and 10N initial draws are considered, and the results are displayed in Table 3.1. The first tolerance ϵ_1 and the total number of iterations T naturally decrease as the elements proposed by the prior distribution increases; however, this can have a negative impact on both the total number of draws and the time needed to achieve convergence. Considering the four options for this model, $5N$ seems to performe well with the total computational time and the total number of draws.

3.3.2 Exponential-Gamma Model

Next we investigate the continuous Exponential-Gamma model to compare the performance of the ABC–PMC algorithm with aABC–PMC. The parameter of interest is $\theta \in \mathbb{R}_+$ and the sample size is $n = 100$. The likelihood function is available and follows the law:

$$f(y \mid \theta) = \theta^n \exp\{-s\theta\},$$

	T	$Draws_{tot}$	ϵ_1	ϵ_T	time	H_{dist}
N	13	53967	0.53	0.05	440.792 sec.	0.120
2N	7	45976	0.25	0.03	206.481 sec.	0.092
5N	5	54889	0.10	0.01	125.990 sec.	0.032
10N	4	71767	0.05	0.01	105.280 sec.	0.033

TABLE 3.1: Results for aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

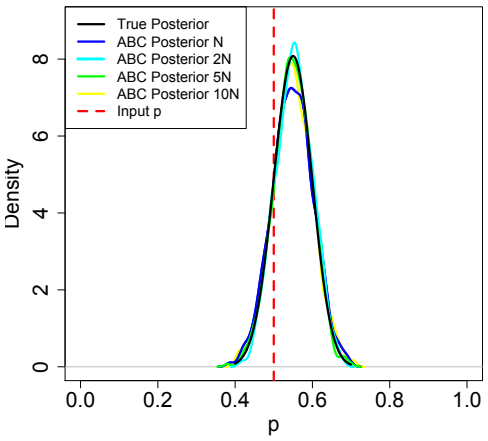

FIGURE 3.6: ABC final posterior distributions for different initial choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

where $s(y) = \sum_{i=1}^{n} y_i$ is again the complete minimal sufficient statistics.

As prior distribution a gamma with hyper parameters $\alpha = 2$ and $\beta = 3$ is used, since it is conjugate to the likelihood function.

$$\pi(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\{-\beta\theta\}.$$

The posterior distribution is available in closed form as:

$$\pi(\theta \mid Y) \propto f(Y \mid \theta)\pi(\theta) \propto \theta^{\alpha+n-1} \exp\{-(\beta + s)\theta\}, \tag{3.9}$$

which is the kernel of a gamma distribution with updated hyper parameters $\alpha^* = \alpha + n$ and $\beta^* = \beta + s$.

With the same distance and summary statistic defined in Equation (3.8) and $N = 2000$, the results of the analysis for both the aABC–PMC and the ABC–PMC are shown in Figure 3.7 and 3.8, respectively. The q_t's retrieved by using Equation (3.2) and displayed in Figure 3.7(right)(black circles), generally increase until the final iteration, while the acceptance rate (blue triangles) decreases. Once the posterior distribution is reached by the algorithm, $1/\hat{C}_t$ stabilizes, as displayed in Figure 3.7(right)(orange crosses). The posterior distribution, displayed in Figure 3.7(left), is reached after 4 iterations and the Hellinger distance (H_{dist}) between the true posterior distribution and the final ABC posterior is 0.07. Stopping the procedure as soon as the convergence is

reached allows for a potentially significant reduction in number of draws.

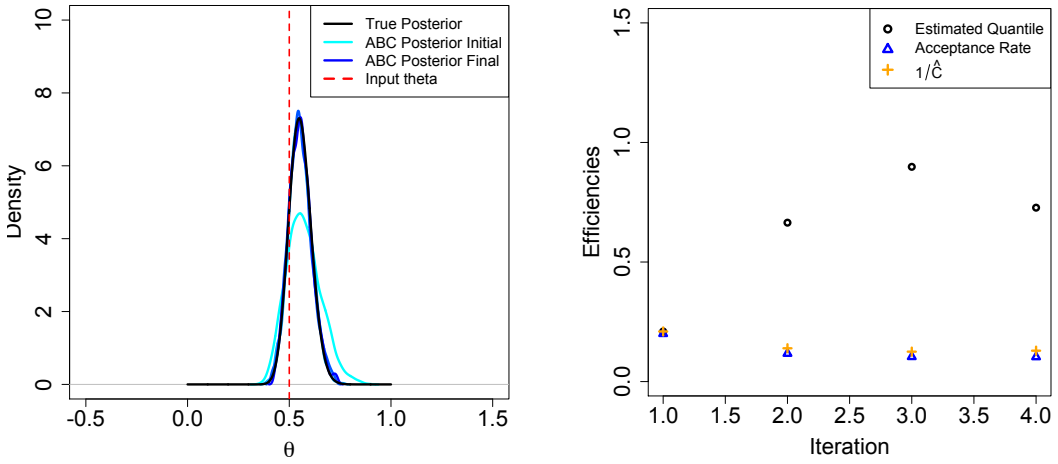

FIGURE 3.7: aABC–PMC analysis for the Exponential-Gamma model with $N = 2000$. (left) Series of 4 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of automatically selected quantiles: : $q_{2:4} = (0.21, 0.61, 0.88)$, that leads to the series of tolerances $\epsilon_{1:4} = (0.35, 0.08, 0.048, 0.042)$. The automatic stopping rule directly based on the behavior of the ABC posterior distribution is satisfied after 4 iterations.

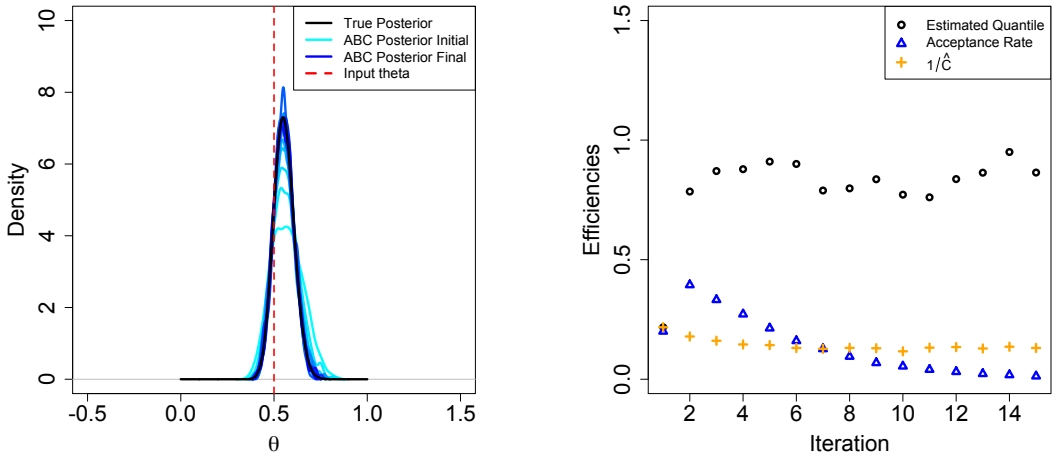

FIGURE 3.8: ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$, $q^{th} = 0.75$ and $T = 15$. (left) Series of 15 ABC posteriors, with the first and final ABC posteriors noted in the legend. (right) Series of efficiencies based on the selection in advance of the quantile used to reduce the tolerance through the 15 iterations.

Concerning the ABC–PMC analysis, Figure 3.8(right) shows how the acceptance rate continues to decrease once $1/\hat{C}_t$ stabilizes (orange crosses) and the Hellinger distance between the true posterior and the final ABC posterior is 0.043. Considering the (high)

estimated quantiles in Figure 3.8(right)(black points), a smaller quantile (i.e. larger reductions in the tolerance) would lead to faster convergence of the ABC posterior. Figure 3.8(left) shows the evolution of the ABC posterior distribution by the ABC–PMC procedure through all the 15 iterations, suggesting how once the procedure stopped, the true posterior distribution has been reached.

As done for the Beta-Binomial Example, next we consider the average total number of draws and average computation time for the Exponential-Gamma model using 20 independent runs. The average total number of draws for aABC–PMC is 63784 and the average computational time is 56.193 sec. These two quantities are, respectively, fixed in 20 independent runs of ABC–PMC with the same 3 quantiles ($q^{th} = 0.25, 0.50, 0.75$). Figure 3.9 displays the results of the ABC–PMC algorithm when the total number of draws is fixed at 63784. The ABC posteriors did not converge based on this threshold for all three quantiles. When the computational time for ABC–PMC is fixed at 56.193 sec., only the $q^{th} = 0.5$ converged, as shown in Figure 3.10.

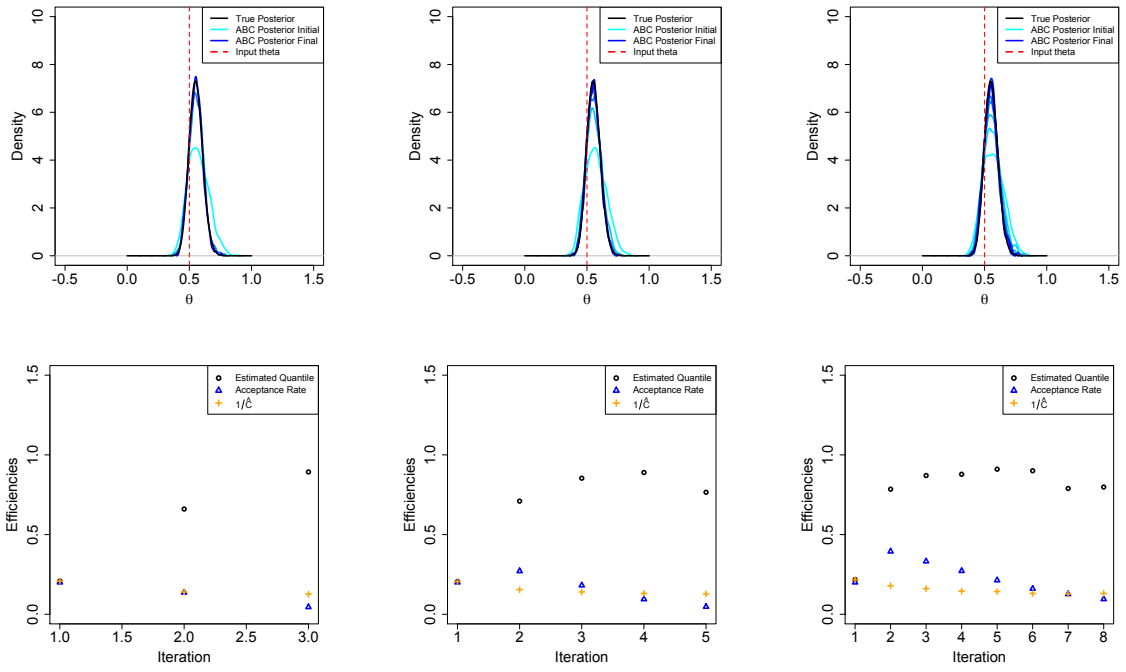

FIGURE 3.9: ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$ and maximum number of allowed draws equal to 63784, for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.072, 0.082, 0.09)$.

We repeated the analysis from the end of Section 3.3.1, where we change the number of draws sampled from the prior distribution in the initial iteration. Table 3.2 lists the average values of the quantities of interest once the experiment has been executed 20

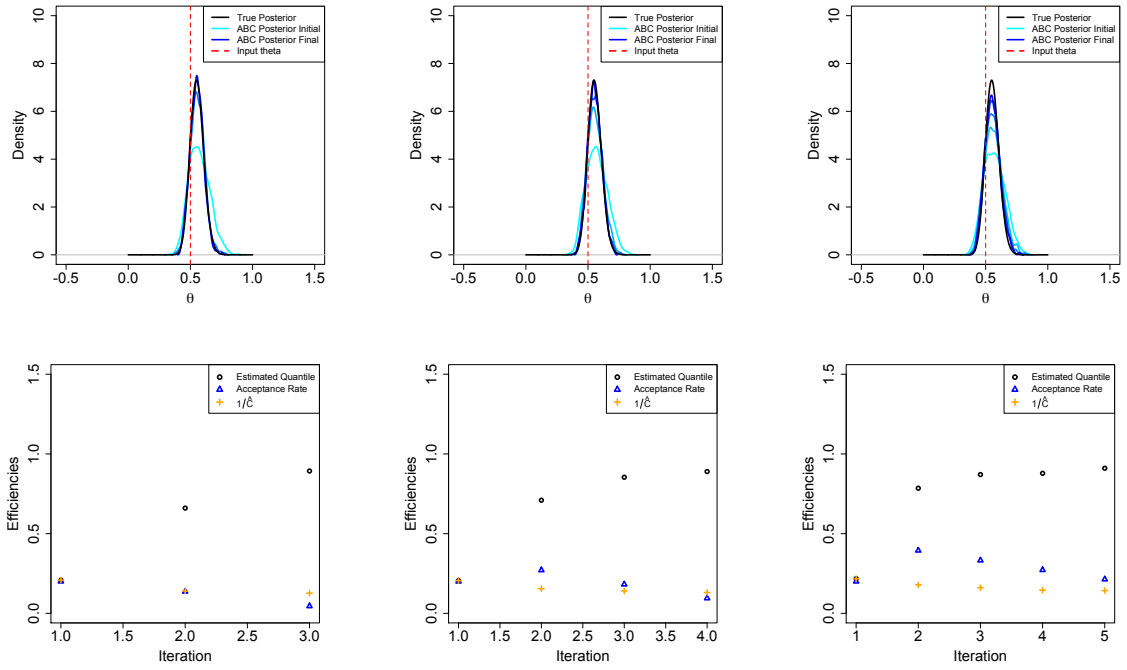

FIGURE 3.10: ABC–PMC analysis for the Exponential-Gamma model with $N = 2000$ and time limit equal to 56.193 sec., for $q^{th} = 0.25$ (first column), $q^{th} = 0.5$ (second column) and $q^{th} = 0.75$ (third column). The final Hellinger distances between the true posterior and the final ABC posteriors are respectively equals to: $(0.072, 0.074, 0.14)$.

times for each of the different choices of the initial number of draws. The final tolerance ϵ_T for each particular choice leads to similar ABC posterior distributions, as shown in Figure 3.11 and confirmed by the evaluation of the Hellinger distance. Also in this case our preference for the choice of the initial number of values directly proposed by the prior distribution is $5N$.

3.3.3 Gaussian Mixture Model

The third application of the aABC–PMC is an example taken from [131], which is also analyzed in [8]. It is a Gaussian mixture model with two components with known variances and mixture weights, but an unknown common mean, $f(y \mid \theta) = 0.5\mathcal{N}(\theta, 1) + 0.5\mathcal{N}(\theta, 0.01)$ and prior distribution $\pi(\theta) \sim Unif(-10, 10)$. With a single observation $y_{\text{obs}} = 0$, the posterior distribution is

$$\pi(\theta \mid y_{\text{obs}}) \sim 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 0.01). \tag{3.10}$$

For consistency with the results presented in [131] and [8], the distance function used is $\rho(y_{\text{obs}}, y_{\text{prop}}) = |y_{\text{obs}} - y_{\text{prop}}|$, $N = 1000$, and a Gaussian kernel for resampling the particles is used. Both [8] and [131] manually define the series of tolerances. In

	T	$Draws_t$	ϵ_1	ϵ_T	time	H_{dist}
N	11	87083	2.17	0.09	163.448 sec.	0.089
2N	7	57434	1.01	0.086	103.456 sec.	0.063
5N	4	63784	0.35	0.042	56.193 sec.	0.07
10N	3	100195	0.22	0.025	46.921 sec.	0.085

TABLE 3.2: aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

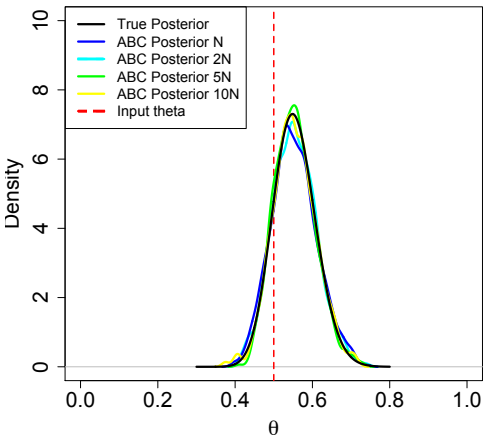

FIGURE 3.11: ABC final posterior distributions for different initial choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

particular, [131] carryout $T = 10$ iterations with a fixed series of tolerances $\epsilon_{1:10}$ displayed in Table 3.3. To evaluate the reliability of the aABC–PMC, a comparison with the ABC–PMC is carried out, both in terms of computational time and total number of draws. The results are based on 20 independent runs with the same dataset. The results of the analysis are shown in Table 3.3, where aABC–PMC outperforms ABC–PMC with total draws (135,373 vs. 1,421,283) and a faster computational time (52.244 seconds vs. 243.531 seconds). The final ABC posteriors for each method are displayed in Figure 3.12. Though the aABC–PMC method is computationally more efficient than the ABC–PMC approach, the final ABC posteriors are very similar. This suggests that after a suitable tolerance is achieved, further decreasing the tolerance does not necessarily lead to a better approximation of the posterior distribution.

From Table 3.3, we note that the final tolerance for [131] is $\epsilon_{10} = 0.0025$ ($H_{\text{dist}} = 0.55$) while the automatic stopping rule of aABC–PMC leads to 6 iterations with a final tolerance of $\epsilon_6 = 0.021$ ($H_{\text{dist}} = 0.54$). In the right plot of Figure 3.12, the q_t's retrieved by using Equation (3.2) are displayed (black circles), which generally increase until the final iteration, while the acceptance rate (blue triangles) decreases. Neglecting to stop the algorithm once the ABC posterior has stabilized can be inefficient since the number of draws needed in order to complete further iterations drastically increases, as evidenced by the increasing Draws_t for later iterations displayed in Table 3.3.

For an additional assessment of the performance of the proposed method, we carryout

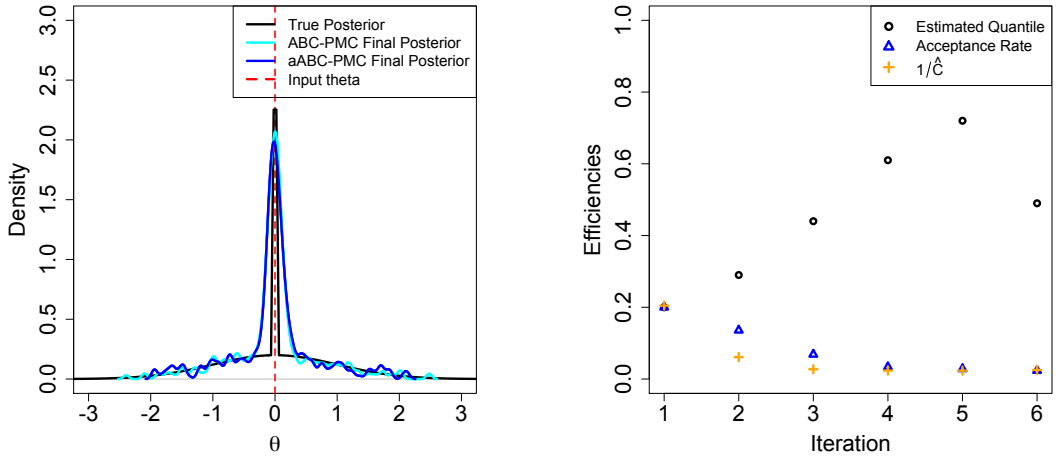

FIGURE 3.12: Gaussian mixture model example. (left) ABC–PMC and aABC–PMC final posterior distributions. The true posterior distribution is plotted with the black line. (right) Sequential quantities computed for the aABC–PMC method. The q_t's (black circles) generally increase through the iterations and the $1/\hat{C}_t$'s (orange pluses) generally decrease until they stabilize. The acceptance rate (blue triangles) decreases throughout the iterations which is why it is desirable to stop the algorithm once the ABC posterior has stabilized.

Sisson et al. (2007)				aABC–PMC				
t	ϵ_t	Draws$_t$	H_{dist}	t	ϵ_t	q_t	Draws$_t$	H_{dist}
1	1.000	2595	0.66	1	2.03		5000	0.76
2	0.5013	8284	0.59	2	0.39	0.20	7365	0.59
3	0.2519	8341	0.57	3	0.11	0.29	14585	0.57
4	0.1272	7432	0.57	4	0.049	0.44	30085	0.55
5	0.0648	10031	0.58	5	0.030	0.61	35584	0.54
6	0.0337	17056	0.53	6	0.021	0.72	42754	0.54
7	0.0181	34178	0.54					
8	0.0102	72704	0.55					
9	0.0064	171656	0.54					
10	0.0025	1089006	0.55					
Total		1421283 (243.531 sec.)					135373 (52.244 sec.)	

TABLE 3.3: The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC algorithm (right), obtained by running the procedure 20 times. For aABC–PMC algorithm, the quantile automatically selected through the iterations is displayed under q_t.

a comparison with the commonly used adaptive tolerance selection method of fixing a quantile using the same Gaussian mixture model example of [131]. For this experiment, we fix the first and final tolerances to the values set using aABC–PMC, $\epsilon_1 = 2.03$ and $\epsilon_T = \epsilon_{suit} = 0.021$, respectively. Then we consider a range of different quantiles and run the usual ABC–PMC algorithm 20 times for each quantile. We evaluate the performance based on computational time and total number of draws required to reach the ϵ_{suit}, at

which time the procedure was stopped. Starting and ending each tolerance sequence with the same tolerance allows for a comparison between the various quantiles, and with the proposed aABC–PMC algorithm.

The results of the experiment are summarized in Figure 3.13; the left plot shows the average total number of draws (for the 20 runs of each quantile) and the right plot shows the average computational time for each quantile. Since the aABC–PMC algorithm uses a different quantile at each iteration, a red 'x' is placed at the average quantile of the series ($\bar{q} = 0.44$); the average total number of draws and the average computational time are among the smallest for the proposed aABC–PMC. In particular, the left plot of Figure 3.13 shows that using a quantile that is too small or too large in the ABC–PMC algorithm results in an increased total number of draws required to reach ϵ_{suit}. In terms of computational time, the right plot of Figure 3.13 reveals that, for this example, quantiles smaller than about 0.45 lead to comparable results.

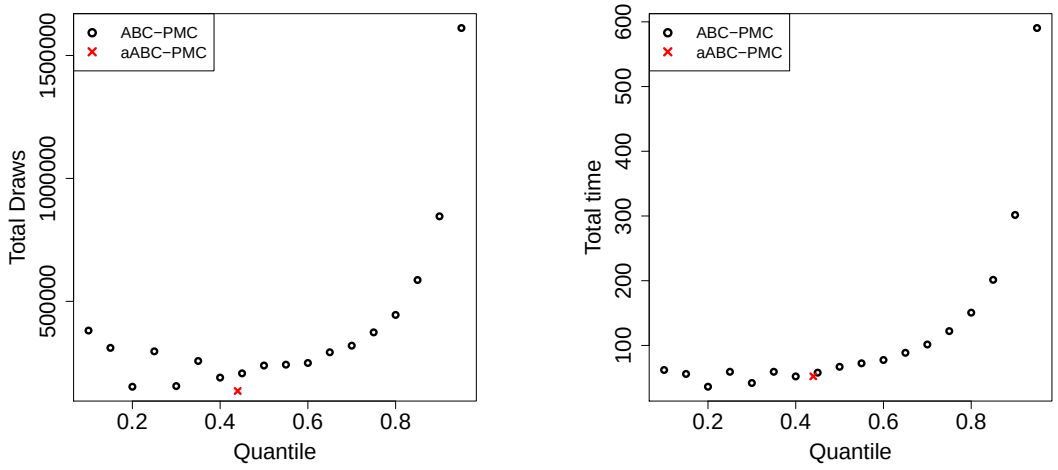

Figure 3.13: Gaussian Mixture model example. (left) Average total number of draws required by the aABC–PMC and the ABC–PMC algorithm for different quantiles. (right) Average computational time required by the aABC–PMC and the ABC–PMC algorithm for different quantiles. Because the aABC–PMC algorithm adaptively selects different quantiles for each iteration, the red 'x' is placed at the average quantile, 0.44.

In [131], the same example is considered but with a series of three fixed tolerances: $\epsilon_{1:3} = (2, 0.5, 0.025)$. We run our aABC–PMC algorithm with a fixed series of tolerances taken by the previous analysis with $\epsilon_{1:3} = (2.03, 0.39, 0.021)$ (corresponding to ϵ_1, ϵ_2, and ϵ_6 from Table 3.3). A comparison, in terms of number of draws and speed of the algorithm is shown in Table 3.4, where again aABC–PMC outperforms ABC–PMC in terms of total number of draws and computational time.

Sisson et al. (2007)			aABC–PMC		
t	ϵ_t	$Draws_t$	t	ϵ_t	$Draws_t$
1	2.000	4907	1	2.03	5508
2	0.500	4899	2	0.39	7215
3	0.025	66089	3	0.021	46495
Total		75895 (22.787 sec.)			59218 (16.710 sec.)

TABLE 3.4: The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC (right) algorithm, obtained by running the procedure 20 times.

	T	$Draws_t$	ϵ_1	ϵ_T	time	H_{dist}
N	16	121439	11.41	0.061	116.697 sec.	0.59
2N	10	136175	4.94	0.036	73.572 sec.	0.59
5N	6	135373	1.95	0.029	52.244 sec.	0.54
10N	4	171385	0.97	0.021	49.969 sec.	0.55

TABLE 3.5: aABC–PMC algorithm with different choices of the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

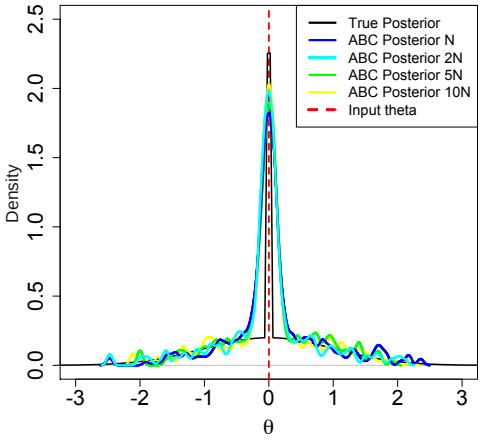

FIGURE 3.14: aABC–PMC algorithm with different choices for the initial number of values directly proposed by the prior distribution: $(N, 2N, 5N, 10N)$.

We conclude the analysis showing the behavior of the aABC–PMC algorithm for different choices of the number of proposed values from the prior distribution at the first iteration of the procedure. Particle sample sizes of N, 2N, 5N, and 10N initial draws are considered (with $N = 1000$ in this example), and the results are displayed in Table 3.5. The initial particle sample size that balances the total number of draws and the time required to satisfy the stopping rule is $5N$. The automatic stopping rule leads to similar ABC posterior distributions based on H_{dist} and visually as displayed in Figure 3.14.

3.3.4 Presence of local modes

The sequence of tolerances has an impact not only on the computational efficiency of the algorithm but also on its ability to find the true posterior [130], noting again

that convergence to the true posterior using ABC is not guaranteed. In fact, when using the ABC–PMC algorithm, perturbation kernels are used rather than the prior distribution for iteration $t > 1$. As a consequence of that, if the previous iteration's accepted particles system does not include relevant regions of the parametric space, then the algorithm can get stuck in local modes. To demonstrate the performance of aABC–PMC in the presence of local modes, we consider an example proposed in [130]. The (deterministic) forward model is $g(\theta) = (\theta - 10)^2 - 100\exp(-100(\theta - 3)^2)$, with the input θ set to $\theta^* = 3$, resulting in true posterior distribution that is a Dirac function at 3. The specifications for both the distance function (L^1 norm) and the desired number of particles ($N = 1000$) are taken from [130].

The distance function is plotted against a range of values for θ in Figure 3.15(left), which highlights the challenge for ABC with this model. There is a broad, local minimum distance around $\theta = 10$, but the global minimum distance occurs at the true value of $\theta = 3$. Initial steps of the ABC algorithm will find the local minimum, and can easily get stuck around $\theta = 10$ if the sequential tolerance is not selected carefully. The series of plots in Figure 3.15 shows the behavior, by iteration, of the aABC–PMC algorithm focusing on the values for θ that were accepted (orange circles). After 5 iterations, the aABC–PMC algorithm has found the global minimum distance around the true θ.

Figure 3.15(left) displays the locations of the accepted particles (orange circles) over the distances for a range of θ's. The third iteration was an important step in which the large reduction of the tolerance allows the algorithm to consider those few particles coming from the global optimal value at $\theta = 3$. Although the raw tolerance hardly decreases between the first and the second iteration ($\epsilon_1 = 51.59$ and $\epsilon_2 = 51.01$), there is a substantial change between the ABC posteriors, from $\hat{\pi}_{\epsilon_1}$ to $\hat{\pi}_{\epsilon_2}$, resulting in a large reduction for the third iteration ($\epsilon_3 = 5.49$). The majority of the accepted values from $t = 2$ are sampled near the local mode at $\theta = 10$, but due to the large reduction resulting in a smaller ϵ_3, values proposed near $\theta = 10$ in iteration 3 are not accepted.

A similar behavior was reported when using the TAR curve, where the tolerance decreased from $\epsilon_2 = 50.94$ to $\epsilon_3 = 5.49 \cdot 10^{-4}$ [130]. We emphasize that this adjustment needs to happen in the first few iterations of the ABC–PMC procedure, since uniformly small reductions in the tolerance sequence (e.g. using a fixed $q = 0.75$) could end up removing those few important particles near the global optimal value.

[130] note that if the particles are sampled from a large region of parameter space that offers negligible or little support for the posterior distribution, there is a risk of getting stuck in this parameter region if the tolerance is not selected carefully. In other words, the parameter space needs to be sufficiently explored in order to get enough

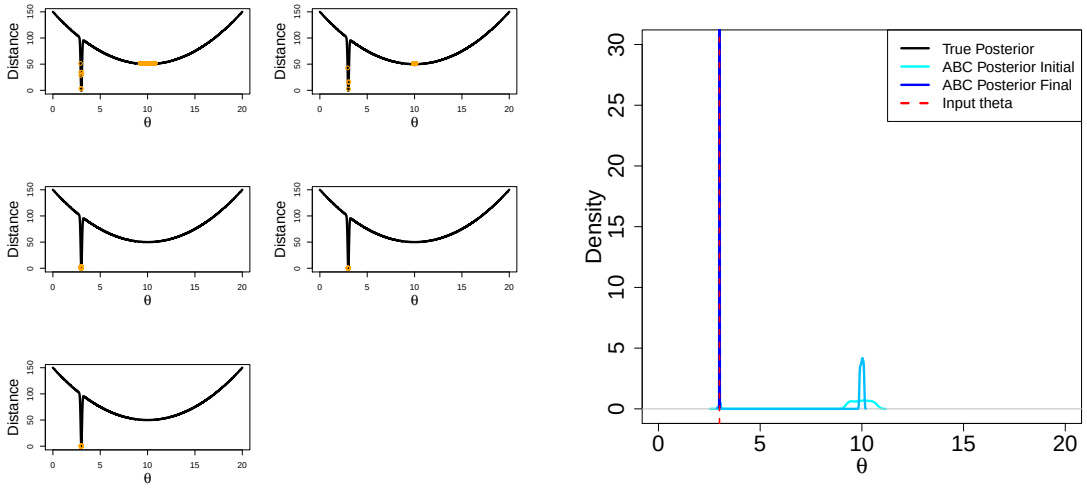

FIGURE 3.15: Example from [130] to investigate the performance of the proposed aABC–PMC in the presence of a local optimal value. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 5 iterations. The series of automatically selected quantiles is $q_{2:5} = (0.18, 0.000016, 0.0044, 0.02)$ which leads to the series of tolerances $\epsilon_{1:5} = (51.59, 51.01, 2.81, 0.00058, 1.42 \cdot 10^{-4})$.

particles in regions near the global optimal value. We suggest that in the first iteration of the aABC–PMC algorithm the number of particles sampled directly from the prior be k times the desired number of particles N, where $k = 5$ seems to work well in the examples considered.

The proposed aABC–PMC algorithm allows for small q_t's early on (when larger improvements occur between sequential ABC posteriors) so that larger reductions in the tolerance sequence can be taken, which results in moving away from local optimal values into better regions of the parameter space. If sufficient reduction of the tolerance is not made early on, achieving a good approximation of the true posterior distribution is unlikely because the distances associated with the local optimal values will overwhelm the particle system so that it gets stuck in the local region.

If the parameter space is not sufficiently explored, the risk of getting stuck in the local maximum increases also if the aABC–PMC algorithm is used, since very few values are coming from the true posterior at the end of the first iteration and their small associated weights rapidly lead to their disappearance in the further iteration. As an example of that, let us sample from the prior, in the first iteration of the aABC–PMC algorithm, a number of particles equal to $2N$. As shown in Figure 3.16(left), very few particles are coming from the true posterior in correspondence to $\theta = 3$. As a result of that, the algorithm rapidly get stuck in the local maximum at $\theta = 10$ and there is no way to retrieve the true posterior distribution. The distance between the simulated and the

observed data does not go under the value of 51 as showed by [130], where a fixed large quantile is used. Hence our suggestion also in this example is to explore the parametric space in the first iteration by sampling from the prior distribution an amount of particles equal to $5N$.

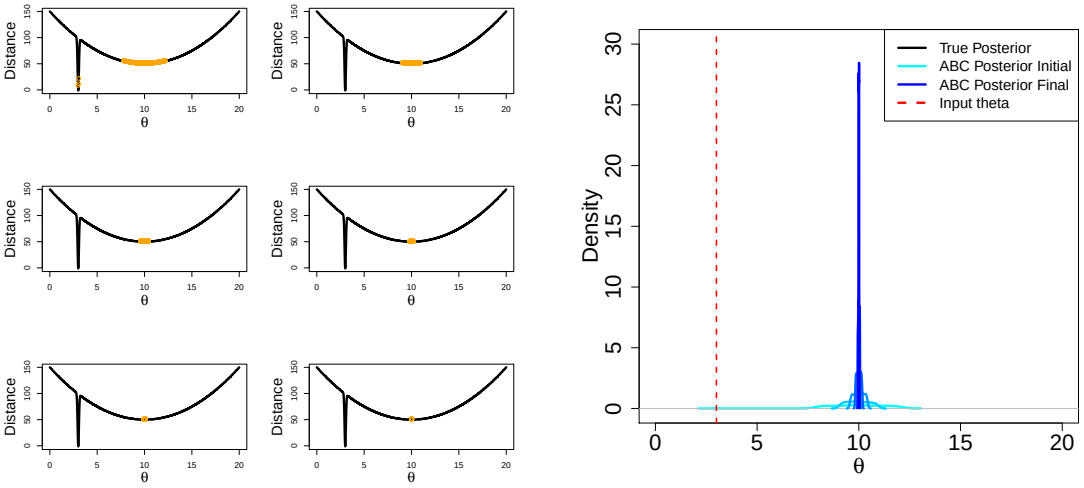

Figure 3.16: Local maximum example with desired particle sample size equal to $N = 1000$ and initial number of draws from the prior equal to 2000. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 6 iterations. In this case the parametric space is not sufficiently explored, hence the achievement of the true posterior distribution is not guaranteed by the aABC–PMC. The series of automatically selected quantiles $q_{2:6} = (0.44, 0.34, 0.31, 0.28, 0.23)$ are too gentle for forcing the algorithm to consider those few particles coming from the true posterior distribution and available at the end of the first iteration. The series of tolerances $\epsilon_{1:5} = (55.82, 51.91, 51.18, 51.02, 51.02, 51.02)$ is coherent with the results found by by [130].

A possible extensions of the model presented by [130] consists in the inclusion of an additional local minimum at $\theta = 15$. We are interested in understand how the aABCpmc behaves in this new setting. The model, which now presents 3 modes, is $g(\theta) = (\theta - 10)^2 - 100\exp(-100(\theta - 3)^2) - 50\exp(-50(\theta - 15)^2)$. The input θ in the simulation is selected to be $\theta^* = 3$, so the true posterior distribution is again a Dirac function at 3. The specifications for both the summary statistics and the desired number of particles are taken from [130] and consistent with the basic aABCpmc introduced at the beginning of this Section. Figure 3.17(left) shows the behavior by iteration of the aABC–PMC algorithm, where it is possible to note that at the beginning the majority of particles are coming from the 2 local minima. Nonetheless the series of automatically selected quantiles leads, after 6 iterations, to accepted values which are coming from the true posterior distribution, as shown in Figure 3.17(right). The third iteration is again

the most important step, because of the large reduction of the tolerance the allows the algorithm to consider those few particles coming from the global optimal value before they are left out for the procedure.

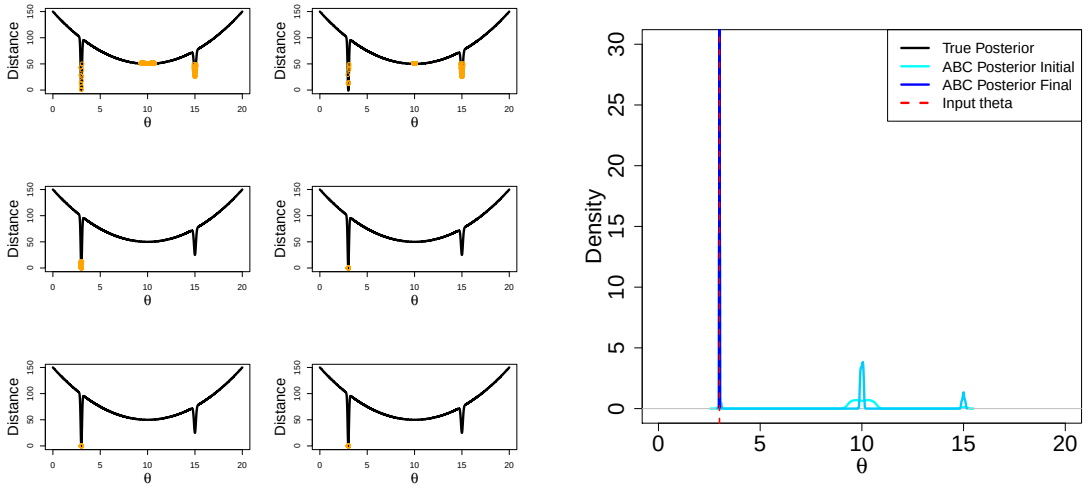

FIGURE 3.17: Extended example from [130], by considering a further local minimum at $\theta = 15$. (left) The accepted θ are plotted in orange against the corresponding distance by iteration, (right) the particle distribution defined with the aABC–PMC for the 5 iterations.The aABCpmc algorithm provides a series of tolerances $\epsilon_{1:5} = (51.62, 50.91, 11.43, 0.00078, 1.42 \cdot 10^{-4}, 6.28 \cdot 10^{-5})$. which leads to the true posteriors posterior.

3.3.5 Lotka–Volterra model

The Lotka-Volterra model [88, 147] describes two interacting populations, and in its original ecological setting these two populations represent, respectively, predators and prey. The interaction between predators (y) and prey (x) is given by the two following differential equations,

$$\frac{dx}{dt} = ax - xy \tag{3.11}$$

$$\frac{dy}{dt} = bxy - y, \tag{3.12}$$

where the parameter of interests are a and b.

Inference on this model using ABC was considered in [140], and we use their same model, dataset, summary statistic and distance function in order to test the performance of the proposed aABC–PMC algorithm. The dataset, $(x_{\text{obs}}, y_{\text{obs}})$, for the analysis was obtained by using (3.11) and (3.12) with input values $a = 1$ and $b = 1$. The sample size

is $n = 8$ for the two species, the distance function for comparing real data (x, y), with the simulated dataset $(x_{\text{sim}}, y_{\text{sim}})$ is defined as:

$$d[(x_{\text{obs}}, y_{\text{obs}}), (x_{\text{sim}}, y_{\text{sim}})] = \sum_{i=1}^{n} \left[(x_{i,\text{obs}} - x_{i,\text{sim}})^2 + (y_{i,\text{obs}} - y_{i,\text{sim}})^2 \right]. \tag{3.13}$$

The forward model solves the deterministic set of differential equations defined above for x and y, then Gaussian noise is added from $N(0, 0.5^2)$ to get the simulated dataset, $(x_{\text{sim}}, y_{\text{sim}})$. The prior distribution for both a and b is an Uniform with range $[-10, 10]$, and the proposed perturbation kernel for $t > 1$ is $K_t = \sigma U(-1, 1)$, with $\sigma = 0.1$. In [140], the series of tolerances are manually selected as listed in Table 3.6.

For the proposed aABC–PMC procedure, the initial number of draws sampled from the prior distributions is set at $N_{\text{init}} = 10 \times 1000$ in order to sufficiently exploring the parametric space. A comparison between the two procedures is done as before, in terms of the computational time and the total number of draws, with the results shown in Table 3.6. Although aABC–PMC requires more iterations, the proposed procedure outperforms [140]'s implementation of ABC–PMC in terms of total number of draws and computational time.

The ABC posteriors for parameters a and b for the manually-selected tolerances of [140] and the proposed aABC–PMC approach are displayed in Figures 3.18. Additionally, ABC posteriors are displayed for two quantile-selected tolerances (0.5 and .75)[2] for comparison. Figures 3.18(right) shows the q_t's (black circles), acceptance rates (blue triangles), and the $1/\hat{C}_t$'s (orange pluses) computed throughout the aABC–PMC run. As the ABC posterior stabilizes, larger q_t's (i.e. smaller reductions of the tolerance) are selected.

The series of tolerances are adaptively selected in such a way that the forward model, which is computationally expensive, is drawn from fewer times than the manually-selected approach from [140] and common quantile-selected approaches. Though the final tolerance from [140], $\epsilon_5 = 4.23$, is smaller than the final tolerance of aABC–PMC, $\epsilon_8 = 6.27$, the posteriors for a and b are very similar (Figure 3.18).

3.4 Concluding remarks

The ABC–PMC algorithms of [8] has lead to great improvements over the basic ABC algorithm in terms of sampling efficiency. However, the user is required to set a sequence

[2]The ABC algorithm for the quantile-selected tolerances is stopped once the final number of draws needed by the aABC–PMC is reached.

Toni et al. (2009)			aABC–PMC			
t	ϵ_t	Draws$_t$	t	ϵ_t	q_t	Draws$_t$
1	30	3541	1	24.24		10000
2	16	48402	2	21.65	0.48	5171
3	6	52471	3	18.81	0.47	4472
4	5	25097	4	11.81	0.26	4275
5	4.3	47521	5	8.63	0.36	3875
			6	7.82	0.73	3845
			7	7.18	0.72	4462
			8	6.27	0.59	6949
Total		177032 (1074.842 sec.)				43049 (625.908 sec.)

TABLE 3.6: The mean number of draws needed in each iteration to reach $N = 1000$ accepted values for the ABC–PMC (left) and the aABC–PMC (right) algorithm, obtained by running the procedure 20 times. In the aABC–PMC algorithm also the quantile automatically selected through the iterations is available.

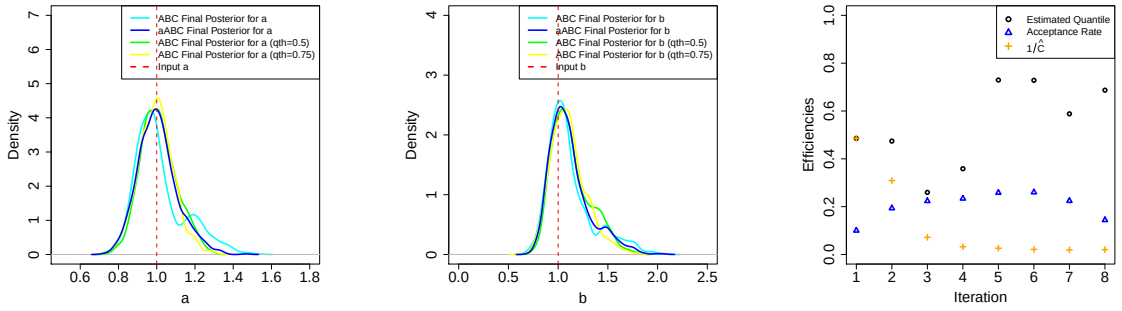

FIGURE 3.18: Lotka–Volterra results. (left)(middle) comparison between the final posterior distributions for a and b obtained using [140]'s manually selecting the tolerances (cyan), by fixing the quantile equal to .50 (green) and 0.75 (yellow), and by using the aABC–PMC (blue). (right) The q_t's (black circles) generally increase through the iterations, while the acceptance rate (blue points) mildly increases and then decreases after iterations 5 and 6. Once the ABC posterior distribution stops improving as the tolerance decreases, the series of $1/\hat{C}_t$'s defined for stopping the algorithm (orange points) stabilizes.

of tolerances along with the total number of iterations. We propose a method for shrinking the tolerances by adaptively selecting a suitable quantile based on the progression of the estimated ABC posterior. The proposed adjustment to the algorithm is able to deal with the possible presence of local optima values, and shrinks the tolerance in such a way that fewer draws are needed from the forward model compared to commonly used techniques for selecting the sequential tolerances. A criterion for stopping the algorithm based on the behavior of the sequential ABC posterior distribution is presented. The empirical performance of the examples considered suggest the proposed aABC–PMC algorithm is superior to the manually-selected tolerance sequences and the fixed-quantile

tolerance sequences in terms of computational time and the number of draws from the forward model. We propose to use the proposed aABC–PMC algorithm for situations where ABC is required and the number of parameters is not huge.

Chapter 4

Approximate Bayesian Computation for Finite Mixture Models

In this Chapter we keep using the ABC–PMC algorithm, but rather than focusing on its efficiency, an extension for working with finite mixture model is proposed.

Finite mixture models are used in statistics and other disciplines, but inference for mixture models is challenging. The multimodality of the likelihood function and the so–called label switching problem contribute to the challenge. When proposing an extension of the ABC–PMC algorithm as an alternative framework for inference on finite mixture models, we need to make several decisions. In Section 4.1 we introduce the basic definitions to work with a finite mixture model. In Section 4.2 we propose the required extensions of the ABC–PMC algorithm, such as the perturbation kernel for moving the mixture weights through the iterations (Section 4.2.2), the deterministic algorithm for addressing the label switching problem (Section 4.2.3) and the definition of informative summary statistics used for comparing the true and the generated dataset into the ABC–PMC algorithm (Section 4.2.4). Examples are presented in Section 4.3 to evaluate the performances of the proposed ABC–PMC algorithm to work with finite mixture models. Concluding remarks are outlined in Section 4.4.

4.1 Introduction

Mixture models have been used in statistics since the late 1800s when Karl Pearson introduced them in an analysis of crab morphometry [101, 102]. Subsequently mixture models have grown in popularity in the statistical community as a powerful framework for modeling clustered data; the book by [93] provides a general overview of mixture modeling while a more Bayesian perspective can be found in [57] or [89]. In recent decades mixture models have become routinely applied in various applications [60, 69,

83, 86, 117, 123, 153]. One reason for the general success of mixture models is due to the opportunity of specifying the number of possibly different component distributions, allowing for flexibility in describing complex systems [89].

The general definition for a finite mixture model with fixed integer $K > 1$ components is:

$$\sum_{i=1}^{K} f_i \cdot p_i(y \mid \theta_i), \tag{4.1}$$

with mixture weights $f_i \in (0, 1)$ such that $\sum_{i=1}^{K} f_i = 1$ and where $p_i(y \mid \theta_i)$ are the component distributions of the mixture, often parametrically specified with a vector of the parameters, θ_i, that are the goal of the statistical inference.

Finite mixture models present computational and methodological challenges due, at least in part, to the complex and possibly multimodal likelihood function, along with the invariance under permutation of the component indices. The EM algorithm [39] provides a method to numerically retrieve the maximum likelihood estimates, although the possible multimodality of the likelihood function makes to find the global maximum challenging [89]. Extensions of the EM algorithm have been proposed in order to improve its speed of convergence and avoiding local optima [97, 99].

Bayesian approaches for mixture modeling have rapidly increased in the last two decades [28, 89, 94, 119, 129]. Bayesian inference for mixture models often relies on MCMC (Markov Chain Monte–Carlo) techniques, which can lead to the so–called label switching problem [41, 71, 121, 133], because the likelihood function is invariant to the re-labelling of the mixture components. Additionally the resulting posterior distribution is multimodal and asymmetric, which makes to summarise the posterior distribution using common statistics such as the posterior mean or the HPD interval unhelpful [94, 133, 134].

A different framework for inference that can be explored in order to address the issues related to the use of mixture models is ABC. ABC is often used in situations where the likelihood function is complex or not available, but simulation of data through a forward model is possible. With mixture models, though the likelihood function is available, working with is difficult. Though it has its own challenges, ABC can be successfully implemented to retrieve the posterior distributions of the parameters of interest, providing an alternative to the MCMC are used. On top of that, ABC allows statistical inference for a much broader class of models respect the ones for which MCMC can be used. In particular, for all those cases in which the likelihood function is intractable but mixture models are used, the following Chapter present a first attempt to suitably extend the ABC–PMC algorithm to successfully run such analyses.

4.2 Required extensions of the ABC–PMC algorithm

One of the main challenges when using the ABC–PMC algorithm for mixture models is related to selecting an appropriate perturbation kernel for the mixture weights, since the individual weights must have range between 0 and 1 and the weights must sum to 1. Consequently, the usual Gaussian perturbation kernel is not a viable option because this kernel can lead to proposed mixture weights that are not consistent with their support.

An additional challenge when using ABC–PMC for mixture models is due to the label switching problem, and for reasons discussed below, this has to be addressed at the end of each iteration.

Finally, in any ABC analysis the definition of both suitable summary statistics, $s(\cdot)$, and a distance function, d, for comparing the true data y_{obs} to the generated sample y_{prop} is needed and is crucial to get useful inference results [9]. The definition of summary statistics is necessary because ABC suffers of the curse of dimensionality and using the entire dataset is computationally unfeasible [13–15, 108]. In this Chapter, beyond extending the ABC–PMC algorithm to work with finite mixture models, we propose the Hellinger distance in order to compare the true data y_{obs} to the generated sample y_{prop}. Details can be found in Section 4.2.4. While the proposed method is valid for $y_{\text{obs}} \in \mathbb{R}^d$, for illustration purposes, we will focus on the one dimensional case where $d = 1$.

For the reasons above introduced, the original version of the ABC–PMC cannot be used to work with mixture models. Our proposed extensions are discussed below, as well as the definition of the finite Gaussian mixture model framework, which is the one used to illustrate the performances of our proposed extended ABC–PMC algorithm. Algorithm 3 summarizes our proposed ABC–PMC algorithm for the special case of a finite Gaussian mixture model, and the details of the steps presented in the Algorithm are discussed in the following subsections.

4.2.1 Finite Gaussian Mixture Models

A common choice for $p_i(\cdot \mid \cdot)$ introduced in Equation (4.1) is the Gaussian distribution. This particular class of models, called Gaussian Mixture Models (GMM's), is very flexible and widely used in various applications [60, 86, 134]. Maintaining the notation of Equation (4.1), a GMM is defined as:

$$\sum_{i=1}^{K} f_i \cdot \phi(y \mid \theta_i), \tag{4.2}$$

where ϕ is the density function of the Normal distribution, $\theta_i = (\mu_i, \sigma_i^2)$ consists of the vector of unknown means and variances for each of the K groups. Hence, the parameters of interest are $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2) \in \mathbb{R}_+^K$ and the mixture weights $f = (f_1, \ldots, f_K)$, that have been previously defined. More specifically, $f = (f_1, \ldots, f_K)$ lies in the unit simplex, $\Delta^{K-1} \equiv \{x \in \mathbb{R}_+^K : \sum_j x_j = 1\}$ inside the unit cube $[0,1]^K$.

A common choice of the prior distribution for $f = (f_1, \ldots, f_K)$ is the Dirichlet distribution of order K with hyperparameter $\delta = (\delta_1, \ldots \delta_K)$, where often $\delta \equiv (1, \ldots, 1)$, as proposed by [149]. Another common choice has been proposed by [124], who defined the hyperparameter $\delta \equiv (1/2, \ldots, 1/2)$; by using latter definition, the prior is marginally a Jeffrey prior distribution.

The priors for the mean and the variance of the GMM can be defined as follows:

$$\mu_i \mid \sigma_i \sim \phi(\xi, \kappa), \qquad \sigma_i^{-2} \sim \text{Gamma}(\alpha, \beta), \tag{4.3}$$

with mean ξ, variance κ, shape parameter α and rate parameter β. There are several methods to select the hyperparameters, $\eta = (\xi, \kappa, \alpha, \beta)$, such as the Empirical Bayes approach [29] and the "weakly informative principle" [117]. Both these options are considered in the simulation study so as to be consistent with the original authors that introduced the examples.

4.2.2 Perturbation kernel functions

As already pointed out in Chapter 2, one of the advantages of ABC–PMC over the basic ABC algorithm is that, starting from the second iteration, rather than drawing proposals from the prior distributions, proposed particles are drawn from the previous step's ABC posterior according to their importance weights. Then, instead of using the actual proposed value that was drawn, it is perturbed according to some kernel. There are a number of possible kernel functions, $K(\cdot \mid \cdot)$, to perturb the proposed particles. [8] suggest a Gaussian kernel having mean on the selected element from the previous iteration and a variance equal to twice the empirical variance of the previous iteration's ABC posterior. This is a reasonable choice if there are no constraints on the support of the parameters of interest. More in detail, the variance of the Gaussian perturbation kernel is defined as equal to twice the empirical variance of the previous iteration's ABC posterior for minimizing the Kullback–Leibler divergence between the proposal distribution at iteration t and the target distribution, as shown in [8]. However, when constraints on the parameter support are present, such as for the variances of the mixtures or the mixture weights, a perturbation kernel should be selected so that it does

Algorithm 3 ABC–PMC for Finite Gaussian Mixture Model

Select the number of components K
Select the desired number of particles N
Select the desired number of particles coming from the prior N_{init}, $N_{init} > N$, for $t = 1$
if $t = 1$ **then**
  for $J = 1, \ldots, N_{init}$ **do**
    Propose $\mu_1^{(J)} = \{\mu_1, \ldots, \mu_K\}_1^{(J)}$ by drawing from prior $\mu_k^* \sim \pi(\mu)$, $k = 1, \ldots, K$
    Propose $\sigma_1^{2^{(J)}} = \{\sigma_1^2, \ldots, \sigma_K^2\}_1^{(J)}$ by drawing from prior $\sigma_k^{2^*} \sim \pi(\sigma^2)$, $k = 1, \ldots, K$
    Propose $f_1^{(J)} = \{f_1, \ldots, f_K\}_1^{(J)}$ by drawing from prior $f_k^* \sim \pi(f)$, $k = 1, \ldots, K$
    Generate y_{prop} from $\sum_{i=1}^K f_1^{(i)} \cdot \phi(y \mid \mu_1^{(i)}, \sigma_1^{2^{(i)}})$
    Calculate distance $d_1^{(J)} = \rho\{y_{\text{obs}}, y_{\text{prop}}\}$
  end for
  Put d_1 in increasing order and set $\epsilon_1 = d_1^{(N)}$, where (N) is the N^{th} smallest distance
  Keep corresponding elements $\mu_1^{(1:N)}, \sigma_1^{2^{(1:N)}}, f_1^{(1:N)}$, the proposed values corresponding to the N smallest distances
  Set weight $W_1^{(J)} = N^{-1}$
  Address the label switching problem (§4.2.3)
else if $2 \leq t \leq T$ **then**
  for $J = 1, \ldots, N$ **do**
    Set $\epsilon_t = q^{th}$ quantile of $\left\{d_{t-1}^{(J)}\right\}_{J=1}^N$
    Set $d_t^{(J)} = \epsilon_t + 1$
    while $d_t^{(J)} > \epsilon_t$ **do**
      Select $\left\{\mu_t^*, \sigma_t^{2^*}\right\}$ from $\left\{\mu_{t-1}^{(J)}, \sigma_{t-1}^{2^{(J)}}\right\}_{J=1}^N$ with probabilities $\left\{W_{t-1}^{(J)} / \sum_{K=1}^N W_{t-1}^{(K)}\right\}_{J=1}^N$
      Propose $f_t^{(J)}$ according to the Dirichlet resampling functions (§4.2.2)
      Propose $\mu_t^{(J)} \sim \phi(\mu_t^*, \tau_{\mu,t-1}^2)$
      Propose $\sigma_t^{2^{(J)}} \sim \text{TruncNormal}(\sigma_t^{2^*}, \tau_{\sigma^2,t-1}^2)$, where TruncNormal is a Normal distribution, centered at $\sigma_t^{2^*}$, with variance $\tau_{\sigma^2,t-1}^2$ and truncated to the positive half-line $\mathbb{R}_+$
      Generate y_{prop} from $\sum_{i=1}^K f_t^{(i)} \cdot \phi(y \mid \mu_t^{(i)}, \sigma_t^{2^{(i)}})$
      Calculate distance $d_t^{(J)} = \rho\{y_{\text{obs}}, y_{\text{prop}}\}$
      Address the label switching problem (§4.2.3)
    end while
    Set weight $W_t^{(J)} \propto \pi(\mu_t^{(J)}, \sigma_t^{2^{(J)}}) / \sum_{K=1}^N W_{t-1}^{(K)} K\left(\mu_t^{(J)}, \sigma_t^{2^{(J)}} \mid \mu_{t-1}^{(K)}, \sigma_{t-1}^{2^{(K)}}\right)$
  end for
end if

not propose values outside the parameter's support. In the following, we defined the perturbation kernels in order to be consistent with the constraints on the parameters the kernel is proposing, keep using as much as possible the original specifications by [8] (i.e. defining the variance of the perturbation kernel equal to twice the empirical variance of the previous iteration's ABC posterior). We note however that, whatever perturbation kernel has been defined, the importance weights are accordingly calculated, in order to reflect the fact that to propose new candidates the prior distributions are not directly used.

When moving the selected values for proposing candidates for the mixture weights, not only is there the constraint that each mixture weight component must be in $[0, 1]$, but it is also required that $\sum_{i=1}^{K} f_i = 1$, making the Gaussian kernel inappropriate.

In the first iteration of the proposed ABC–PMC algorithm, the mixture weights $\{f_1^1 \ldots, f_K^1\}$ are directly sampled from the prior distribution, which is a Dirichlet(δ), where $\delta = (\delta_1, \ldots, \delta_K)$. For $t > 1$, proposals are drawn from the previous step particle system according to their importance weights. After randomly selecting a mixture weight, f^{t-1}, we want to "jitter" or move it in manner that preserves some information coming from the selected particle, but not let it be an identical copy, leading to the resampled mixture weight f^t. This is carried out using Algorithm 4. The mathematical assumptions required to run the proposed algorithm are discussed in the Appendix C.

Algorithm 4 Resampling the mixture weights

1. Draw $Z^t \sim \text{Gamma}(\delta_+, 1)$, with $\delta_+ = \sum_{i=1}^{N} \delta_i$ and set $\xi_i^t = Z^t f^{t-1}$. Then $\{\xi_i^t\} \overset{\text{ind}}{\sim} \text{Gamma}(\delta_i, 1)$
2. Select a real number $p \in [0, 1]$
3. Draw $\{B_i^t\} \sim \text{Beta}(p\delta_i, (1-p)\delta_i)$ independently for $i = 1, \ldots, K$, noticing that $\{\xi_i^t B_i^t\} \overset{\text{ind}}{\sim} \text{Gamma}(p\delta_i, 1)$ are independent gamma-distributed random variables
4. Draw $\{\eta_i^t\} \overset{\text{ind}}{\sim} \text{Gamma}((1-p)\delta_i, 1)$ independently
5. Set $\xi_i^{t^*} = Z^t f_i^{t-1} B_i^t + \eta_i^t$ and $f_i^t = \xi_i^{t^*} / \xi_+^{t^*}$, with $\xi_+^{t^*} = \sum_{i=1}^{K} \xi_i^{t^*}$

From the steps outlined in Algorithm 4, we note that $\xi_i^{t^*}$ is the sum of two independent random variables, with $Z^t f_i^{t-1} B_i^t \sim \text{Gamma}(p\delta_1, 1)$ and $\eta_i^t \sim \text{Gamma}((1-p)\delta_i, 1)$, so that the resampled mixture weight $f^{(t)} \sim \text{Dirichlet}(\delta)$.

The parameter p is a fixed real number with range $[0, 1]$ that determines how much information to retain from f^{t-1}. The choice of p has an impact on both the allowed variability of the marginal ABC posterior distributions for the mixture weights and the efficiency of the entire procedure. In particular fixing a p close to 1 leads to a Dirichlet resampling in which the new set of mixture weights f^t is close to the previous set f^{t-1} (if $p = 1$, then $f^t = f^{t-1}$). On the other hand a choice for p close to 0 implies that

the information coming from f^{t-1} is weakly considered (for $p = 0$ a new set of particles is drawn directly from the prior distribution and no information about f^{t-1} has been retained). We found $p = 0.5$ to be a good choice to balance efficiency and variability (i.e., it allows for some retention of information of the selected particle, but does not lead to nearly identical copies of it). We would love for future studies to provide an extension of Algorithm 4, where the parameter p is not fixed in advance by the researcher but rather automatically retrieved by looking at the sequential performance of the ABC–PMC algorithm. In this way we could speed up the convergence of the mixture weights ABC posterior distributions to the true ones.

4.2.3 Algorithm for addressing the label switching problem

As noted earlier, a common problem arising when dealing with mixture models in the Bayesian framework is the label switching. When drawing a sample from a posterior (for both MCMC and ABC), the sampled values are not necessarily ordered according to their mixture component assignments because the likelihood is exchangeable. For example, suppose a particle $\{(f_1^{(J)}, \ldots, f_K^{(J)}), (\mu_1^{(J)}, \ldots, \mu_K^{(J)}), (\sigma_1^{2^{(J)}}, \ldots, \sigma_K^{2^{(J)}})\}$ is accepted for a K component GMM. This particle was selected with a particular ordering of the $1, \ldots, K$ components with $f_i^{(J)}$, $\mu_i^{(J)}$, and $\sigma_i^{2^{(J)}}$ from the same mixture component, $i = 1, \ldots, K$. However, a new particle that is accepted will not necessarily follow that same ordering of the $i = 1, \ldots, K$ components. Somehow the particles have to be ordered in such a way that aligns different realizations of the $i = 1, \ldots, K$ components in order to eliminate the ambiguity.

Several approaches have been proposed to address the label switching problem and are known as relabeling algorithms. A first group of relabeling algorithms consists of imposing an artificial identifiability constraint in order to arbitrarily pick a parameter (e.g. the mixture weights) and sort all the parameters for each accepted particle according to that parameter's order [41, 117]. However, the majority of the algorithms proposed to address the label switching are deterministic (e.g Stephen's method [133] and the pivotal reordering algorithm [89]). A third class of strategies, called probabilistic relabeling algorithms, uses a probabilistic approach to address the label switching problem [132]. A detailed overview of current methods that try to address label switching is presented in [100]. In Section 4.3.1, we provide an example that illustrates the problems arising with the artificial identifiability constraint approach. Instead, we propose a deterministic strategy for addressing the label switching by selecting a parameter that has at least two well-separated components.

Addressing the label switching problem is especially important for the proposed sequential ABC algorithm because each time step's ABC posterior is used as the proposal in the subsequent step of the algorithm, meaning that the label switching has to be addressed before using it as a proposal distribution. Algorithm 5 outlines the proposed strategy, and is carried out at the end of each iteration. The key aspect of Algorithm 5 is to select the parameter that has at least two well-separated components. To determine this, each set of parameters (e.g. the means, the variances, the mixture weights), is arranged an increasing order. For example, for each particle J, $J = 1, \ldots, N$, $\mu^{(J)}$ would be ordered so that $\mu_{(1)}^{(J)} \leq \mu_{(2)}^{(J)} \leq \cdots \leq \mu_{(K)}^{(J)}$, with $\mu_{(i)}$ as the i^{th} order statistic; let $\mu_{(k)}^{(J)}$ represent the k^{th} smallest mean particle value. This is carried out for each set of parameters with analogous notation.

The next step is to determine which set of parameters has the best separated components values. We propose first shifting and scaling each set of parameters to be supported within the range $[0, 1]$ so that scaling issues are mitigated and the parameter set values are comparable. One option for this adjustment is to use some distribution function, such as a Normal distribution with a mean and standard deviation equal to the sample mean and the sample standard deviation of the considered parameter set (e.g. the sample mean for the μ's is $\bar{\mu} = \sum_{k=1}^{K} \sum_{J=1}^{N} \mu_{(k)}^{(J)}$, and the sample standard deviation for the μ's is $\text{sd}(\mu) = \sqrt{\sum_{k=1}^{K} \sum_{J=1}^{N} (\mu_{(k)}^{(J)} - \bar{\mu})^2}$). The resulting k-smallest standardized value is, for the mixture mean, $\widetilde{\mu}_{(k)}^{(J)}$. This is carried out for each set of parameters with analogous notation.

Then, for each component of each ordered and standardized particle a representative value, such as a mean, is computed (e.g. $\bar{\widetilde{\mu}}_{(k)} = N^{-1} \sum_{J=1}^{N} \widetilde{\mu}_{(k)}^{(J)}$ is the representative value of the k^{th} component of the mean parameter). This is carried out for each set of parameters with analogous notation. The pairwise distances (pdist) between the representative values within each parameter set is determined. The parameter set that has the largest separation between any two of its representative values is selected for the overall ordering of the particle system.

Other methods to address the label switching problem were considered. For example, rather than sorting based on the parameter set with largest separation between any two of its representative values, we considered basing the sorting on the parameter set with the largest separation between its two *closest* representative values (i.e. the maximum of the minimum separations); however, this sorting did not perform well empirically. The issue seemed to be that parameter set with the largest separation between its two closest representative values may actually have all of its components relatively close; after multiple iterations, none of the components would separate out from the other

Algorithm 5 Addressing the label switching problem

1. For each parameter set, obtain the ordered particles $\mu_{(k)}^{(J)}$, $\sigma_{(k)}^{2(J)}$ and $f_{(k)}^{(J)}$, $k = 1, \ldots, K$, $J = 1, \ldots, N$
2. Shift and rescale each set of parameters to be supported within the range $[0, 1]$, retrieving $\widetilde{\mu}_{(k)}^{(J)} = \Phi(\mu_{(k)}^{(J)}, \bar{\mu}, \mathrm{sd}(\mu))$, $\widetilde{\sigma}_{(k)}^{2(J)} = \Phi(\sigma_{(k)}^{2(J)}, \bar{\sigma^2}, \mathrm{sd}(\sigma^2))$ and $\widetilde{f}_{(k)}^{(J)} = \Phi(f_{(k)}^{(J)}, \bar{f}, \mathrm{sd}(f))$, where Φ is the distribution function of a Normal distribution
3. Compute representative values (such as a mean) for each shifted and standardized component, $\bar{\widetilde{\mu}}_{(k)}$, $\bar{\widetilde{\sigma}}_{(k)}^2$ and $\bar{\widetilde{f}}_{(k)}$
4. Compute the pairwise distances (pdist) within each set of representative values, $\bar{\widetilde{\mu}}_{(k)}$, $\bar{\widetilde{\sigma}}_{(k)}^2$ and $\bar{\widetilde{f}}_{(k)}$
5. The overall ordering of the particle system is based on the ordering of the parameter set with the largest separation between any two of its representative values

components. This lead to iteration after iteration of components that remained a blend of components rather than separating out into pure components. Overall, from empirical experiments, the algorithm outlined in Algorithm 5 performed the best and thus is our recommendation. However, we emphasize that alterations to this procedure may be necessary for mixture models that have additional structure or correlations among the parameters of the component distributions or to deal with multi–dimensional mixture models.

4.2.4 Summary statistics

As already pointed out in Chapter 2, to compare the true data y_{obs} with the generated sample y_{prop} in an ABC procedure is not computationally feasible. For this reason the definition of a lower dimensional summary statistic is necessary. For mixture models, due to the multimodality of the data, common summaries such as means or higher order moments do not capture relevant aspects of the distribution. However, an estimate of the density of the data will better account for its key features (e.g. the shape of each component of the mixture).

We suggest using kernel density estimates of the generated sample, $\hat{f}_{y_{\mathrm{prop},n}}$, and the true data, $\hat{f}_{y_{\mathrm{obs},n}}$, to summarize the data, and then the Hellinger distance (H) to carry out the comparison. The Hellinger distance quantifies the similarity between two density functions, f and g, and is defined as:

$$H(f, g) = \left(\int \left(\sqrt{f(y)} - \sqrt{g(y)} \right)^2 dy \right)^{\frac{1}{2}}. \tag{4.4}$$

At each iteration t of the proposed ABC–PMC procedure, a proposed θ is accepted if $H(\hat{f}_{y_{\text{obs},n}}, \hat{f}_{y_{\text{prop},n}}) < \epsilon_t$, where ϵ_t is the tolerance.

4.3 Illustrative Examples

In this Section two simulation studies and a real application are introduced, to evaluate the behavior of the proposed ABC–PMC algorithm presented in Section 4.2. In particular we are interested in evaluating the success of the procedure with respect to the label switching problem and the reliability of the Hellinger distance as summary statistic. To determine the number of iterations, a stopping rule was defined based the Hellinger distance between sequential ABC posteriors; once the sequential Hellinger distance dropped below a tolerance of 0.05 for each of the marginal ABC posteriors, the algorithm was stopped.

4.3.1 Mixture Model with equal group sizes

The first example is taken from [94], which considered a GMM obtained by simulating data coming from $K = 2$ groups of equal size which was designed to evaluate the performance of a method proposed by [94] to address the label switching problem. A total of 40 observations were simulated as follows:

$$Y_{i=1,\ldots,20} \sim \phi(-20, 1), \quad Y_{i=21,\ldots,40} \sim \phi(20, 1). \tag{4.5}$$

The variance is assumed known for both groups and hence the parameters of interest are the mixture weights, f_1 and f_2 (with $f_2 = 1 - f_1$), and the means $\mu = (\mu_1, \mu_2)$. The prior distributions defined to run the analysis are the same as those used by [94], where the values of the hyperparameters are $\eta = (0, 100)$,

$$\mu_i \mid \sigma_i \sim \phi(0, 100).$$

The prior for the mixture weights is the Dirichlet distribution with hyperparameters $\delta = (1, 1)$,

$$f = (f_1, f_2) \sim \text{Dirichlet}(1, 1).$$

The desired particles sample size was set to $N = 5000$ and the quantile used for shrinking the tolerance was $q = 0.5$. The algorithm was stopped after $t = 20$ iterations, since further reduction of the tolerance did not lead to an improvement by the ABC posterior

distributions (evaluated by calculating the Hellinger distance between the sequential ABC posterior distributions, as noted in the introduction to this Section).

Figure 4.1 displays the resulting ABC posteriors (the labels "ABC Posterior good LS" and "ABC Posterior bad LS" are discussed later for illustrating the label switching issue) and the corresponding MCMC posteriors, which are used as a benchmark to access the performances of the proposed extended ABC–PMC. The ABC posteriors, when the label switching is suitably addressed, closely match the MCMC posteriors. The summary of the results presented in Table 4.1 demonstrates that the ABC posterior distributions are a suitable approximation of the MCMC posteriors. The Hellinger distance between the marginal ABC and the MCMC posteriors is also displayed in the last column of Table 4.1; the Hellinger distance between the MCMC and the ABC posterior is 0.032 for the mixture weights and 0.21 for the means.

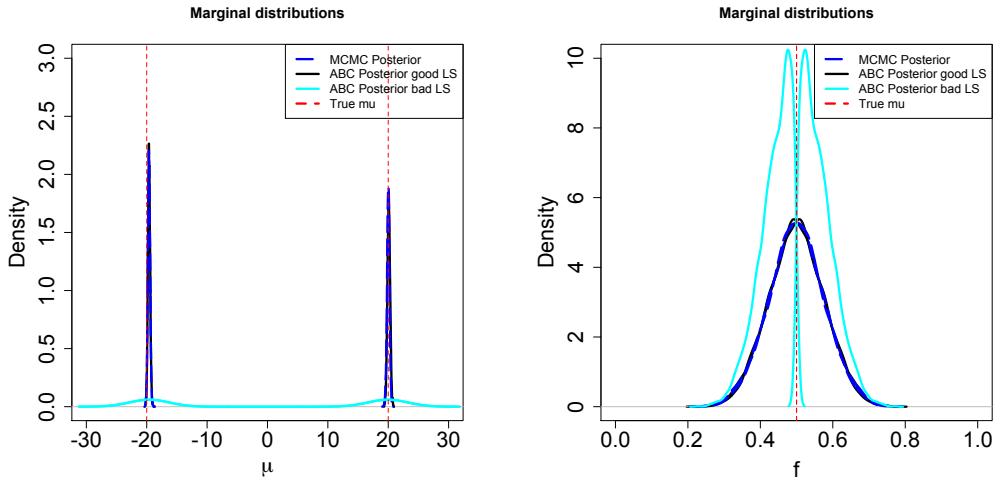

Figure 4.1: Comparison between the ABC and the MCMC marginal posterior distributions for the two-component GMM example from [94]. The final ABC posteriors obtained using the label switching (LS) procedure proposed in Section 4.2.3 are the solid black lines (ABC Posterior good LS), and the naive approach that sorts based on the mixture weights are the solid cyan lines (ABC Posterior bad LS). We recall that only for the MCMC analysis the label switching problem has to be addressed. This is done deterministically sorting the parameters according to the means of the mixture model. The number of particles for the ABC analysis and the number of elements kept from the MCMC analysis (after the burn-in) are equal to 5000.

As noted in Section 4.2.3, the label switching problem has to be carefully addressed when using the ABC–PMC algorithm. For each time step following the initial step, the previous step's ABC posterior is used as the proposal rather than the prior distribution so the procedure for addressing the label switching proposed in Section 4.2.3 is used

Parameter (input)	MCMC (SD)	ABC (SD)	H
$f_1(0.5)$	0.5008(0.076)	0.5003(0.076)	0.032
$f_2(0.5)$	0.4991(0.076)	0.4996(0.076)	0.032
$\mu_1(-20)$	$-19.72(0.18)$	$-19.70(0.18)$	0.21
$\mu_2(20)$	20.05(0.19)	20.10(0.19)	0.21

Table 4.1: Posterior means (and posterior standard deviations) obtained by using the MCMC and the ABC–PMC algorithm for the two-component GMM example from [94]. The fourth column indicates the Hellinger distance between the final ABC and the MCMC posteriors. The number of ABC particles and the number of elements retained from the MCMC chain (after the burn-in) are both equal to 5000.

at the end of each time step. In order to illustrate the consequences of incorrectly addressing the label switching, we ran the proposed ABC algorithm on the example proposed by [94], except rather than using the method proposed in Section 4.2.3, the ordering of the particle system is carried out using the ordering of the mixture weights; the mixture weights are equal in this example making them a poor choice for attempting to separate out the mixture components. The resulting ABC posteriors are displayed in Figure 4.1(cyan lines). The means, $\mu = (\mu_1, \mu_2)$, of the mixture components are shuffled and not close to the MCMC posterior, while the mixture weights are sorted such a way all the elements of f_1 are smaller than 0.5 and all the elements of f_2 are larger than 0.5. We note however that the same undesirable results are retrieved in the MCMC analyses if the label switching is not suitably addressed.

To complete this first example, [94] added a third component to the mixture defined in Equation (4.5), by simulating five additional observations from a standard Normal distribution and obtaining a three-component GMM with known variances. The ABC–PMC algorithm was run with the same specifications as the first part of the example, but required 25 time steps to achieve our stopping rule.

Figure 4.2 shows the MCMC and the ABC posteriors for the weights and the means of the mixture components. The behavior of the ABC posterior distributions is consistent with their MCMC benchmarks. A summary of the results presented in Table 4.2 shows that the posterior means (and the posterior standard deviations) for the ABC posterior distributions are consistent with the ones retrieved using MCMC. Finally, in the third column, the Hellinger distances between the ABC and MCMC posteriors are provided.

Parameter (input)	MCMC (SD)	ABC (SD)	H
$f_1(0.44)$	0.44(0.071)	0.44(0.071)	0.024
$f_2(0.12)$	0.12(0.048)	0.12(0.048)	0.018
$f_3(0.44)$	0.44(0.071)	0.44(0.071)	0.033
$\mu_1(-20)$	−19.61(0.26)	−19.73(0.22)	0.27
$\mu_2(0)$	−0.33(0.45)	−0.30(0.48)	0.17
$\mu_3(20)$	20.06(0.24)	20.19(0.22)	0.29

Table 4.2: Posterior means (and posterior standard deviations) obtained by using the MCMC and the ABC–PMC algorithm for the three-component GMM example from [94]. The fourth column is the Hellinger distance between the final ABC posterior distribution and the MCMC posterior. The number of particles and the number of elements retained from the MCMC chain (after the burn-in) are both equal to 5000.

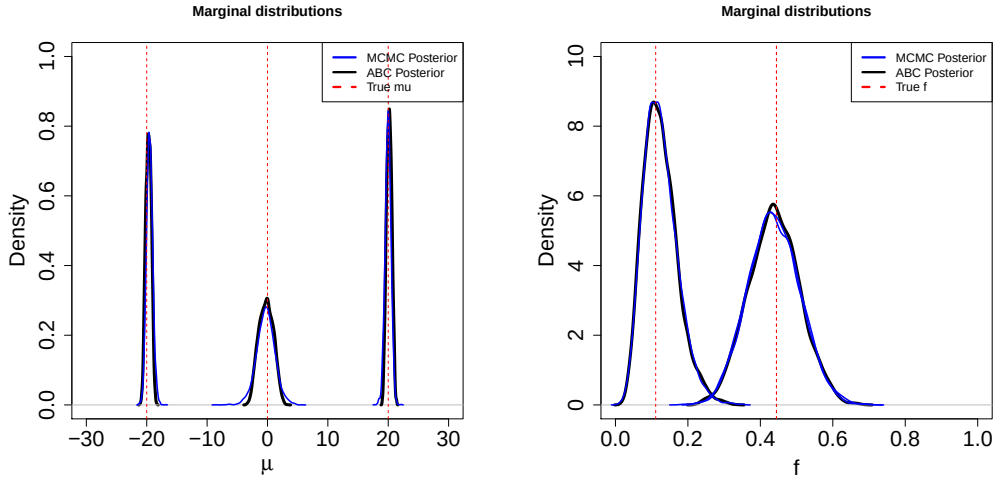

Figure 4.2: ABC and the MCMC marginal posterior distributions for the three-component GMM example from [94]. The number of particles for the ABC analysis and the number of elements kept from the MCMC analysis (after the burn-in) are equal to 5000.

4.3.2 Mixture Model with unequal group size

Even in those cases in which the definition of the mixture model does not lead to the label switching problem, a second category of issues related to the multimodality of the likelihood function is present. This behavior has been studied from both a frequentist and a Bayesian standpoint. In particular, [89] defined the following simple two-component mixture model to illustrate the multimodality issue:

$$f \cdot \phi(\mu_1, 1) + (1 - f) \cdot \phi(\mu_2, 1), \tag{4.6}$$

where the weight f is assumed known and different from 0.5 (avoiding the label switching problem). According to the specifications by [89], $n = 500$ samples were drawn from

the model defined in Equation (4.6), with $\theta = (f, \mu_1, \mu_2) = (0.7, 0, 2.5)$. The bimodality of the likelihood function (Figure 4.3) makes the use of both the EM algorithm [39] and the Gibbs Sampler [41] risky, because their success depends on the set of initial values selected for initiating the algorithms.

The Population Monte Carlo (PMC) sampler [27, 89] is used as a benchmark for the proposed ABC–PMC solution. Figure 4.3 displays the log likelihood function (note the two modes), and the final ABC posteriors with MCMC posteriors using good and bad starting values along with the posteriors using the PMC algorithm. Table 4.3 lists the means for the final ABC, MCMC, and PMC posteriors for μ_1 and μ_2, along with the Hellinger distance between the final ABC–PMC posteriors and the PMC posterior. The ABC–PMC posteriors closely match the PMC posteriors.

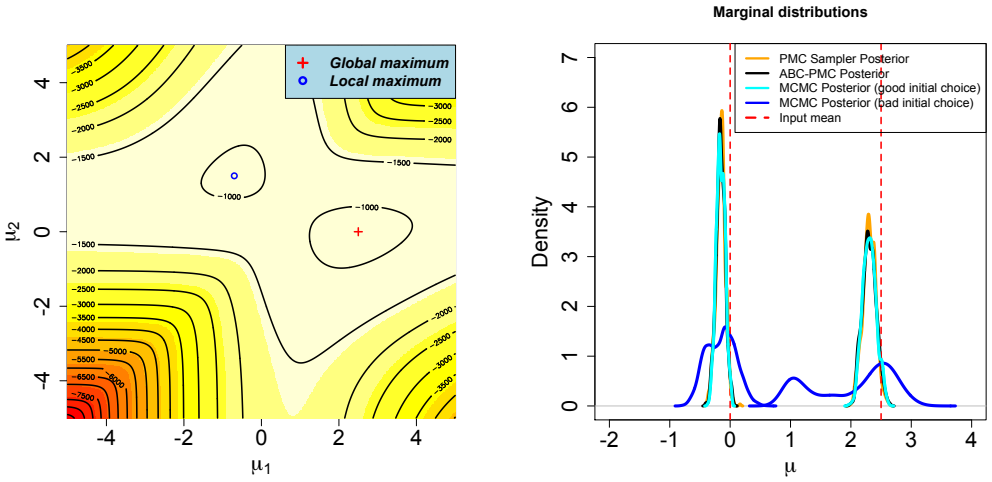

FIGURE 4.3: (left) The log-likelihood surface of the Gaussian mixture model proposed by [89]. There are two modes in the log-likelihood function, one close to the true value, $(0, 2.5)$, and a second local mode. (right) The marginal ABC, PMC, and MCMC posterior distributions; the displayed MCMC posteriors include the results for good initial starting values (MCMC Posterior (good initial choice)) and bad initial starting values (MCMC Posterior (bad initial choice)).

4.3.3 Application to Galaxy Data

The galaxy dataset was introduced to the statistical community in [122] and since than has been commonly used to test clustering methods. The data contain the recessional velocities of 82 galaxies (km/sec) from six well separated sections of the Corona Borealis region. It is worth noticing the analyses have been conducted using the version used in [122] rather than the original dataset introduced in [107]; in the latter case, the

Parameter (input)	$\mu_1(2.5)$	$\mu_2(0)$
PMC (SD)	2.29(0.17)	−0.17(0.11)
ABC–PMC (SD)	2.29(0.17)	−0.16(0.11)
MCMC_{good} (SD)	2.31(0.16)	−0.16(0.11)
MCMC_{bad} (SD)	−0.18(0.29)	1.86(0.80)
H	0.051	0.048

Table 4.3: Mean posteriors (and standard deviations) obtained by using MCMC (with good and poor choices for initializing the procedure), PMC and ABC algorithms in the example by [89]. The last column indicates the Hellinger distance between the final ABC posterior distributions and the PMC posteriors

dataset consists in 83 recessional velocities of galaxies (km/sec) from the same six well separated sections of the Corona Borealis region. We finally note that the dataset used here can be easily found loading the library MASS in the statistical software **R**.

In the last twenty years this dataset has been studied in a number of papers ([84, 117, 123, 148]). The recessional velocities of the galaxies are typically considered realizations of independent and identically distributed Normal random variables, but there is discrepancy in their conclusions about the number of groups in the GMM; estimates vary from three components [123] to six [117].

In this analysis, we focused on the model with three components [123], in order to be consistent with [89] and [94]. For the hyper–parameters, we used the Empirical Bayes approach suggested by [29]. Referring to Equation (4.3) and defining the ordered dataset of the 82 galaxies as $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, $\eta = (\xi = m, \kappa = \frac{r}{2}, \alpha = 2, \beta = 50/r^2)$, where $m = \frac{x_{(1)}+x_{(n)}}{2}$ (i.e. the mid point of the dataset) and $r = (x_{(n)} - x_{(1)})$ (i.e. the range of the dataset). Additionally, since to each recessional velocity was also assigned a measurement error, the ABC forward model has been modified to take into account this information. In order to include the measurement errors in the forward model, each simulated recessional velocity is assigned one of the observed measurement errors. The simulated and observed recessional velocities were matched according to their ranks, and the measurement error of the observations were assigned to the simulated data according to this matching. Then, Gaussian noise was added to each simulated recessional velocity with a standard deviation equal to its assigned measurement error. The goal for presenting this example is not suggesting ABC over MCMC when measurement errors are available, since MCMC is still possible [63, 78, 118]. The goal of this study is to test the proposed extensions of the ABC–PMC algorithm in a more complex and realistic scenario respect the simulation examples presented before in this Section.

The posterior means for each component's parameters are listed in Table 4.4. The third component was found to have a weight equal to 0.057, and mean and variance equal

Parameter	Marin et al. (2005)	Mena et al. (2015)	ABC–PMC	ABC–PMC$_{\text{(with errors)}}$
f_1	0.09	0.087	0.089	0.087
f_2	0.85	0.868	0.85	0.86
f_3	0.06	0.035	0.061	0.053
μ_1	9.5	9.71	9.36	9.51
μ_2	21.4	21.4	21.32	21.33
μ_3	26.8	32.72	32.94	32.58
σ_1^2	1.9	0.21	0.40	0.20
σ_2^2	6.1	4.76	5.32	4.79
σ_3^2	34.1	0.82	1.16	0.62

Table 4.4: Comparison between the posterior means obtained by [89], [94] (MCMC algorithm) and the ABC–PMC algorithm for the Galaxy data. The results of the ABC–PMC analysis including measurement errors are displayed in the fourth column. The proposed ABC–PMC estimates are comparable with the ones obtained by [94], while [89] obtained different results, in particular for the third component of the mixture.

to 32.94 and 1.16, respectively. The main difference between the proposed ABC–PMC estimates and the estimates found by [89], in which the authors also fixed the number of components equal to $K = 3$, is about the retrieved mean and retrieved variance for the third component. Anyway, the proposed ABC–PMC estimates are comparable with the ones obtained by [94].

By using the additional information about the measurement errors, the proposed ABC–PMC algorithm can provide a more accurate evaluation of the dataset. Including measurement errors in the forward model affects the resulting ABC posterior as reported in Table 4.4 and the relative estimates plotted in Figure 4.4(orange line). In particular, the variance of the estimated posterior means are smaller, which is a positive consequence of appropriately accounting for the extra uncertainty in the data.

4.4 Concluding Remarks

The recent popularity of ABC is, at least in part, due to its capacity to handle complex models. Extensions of the basic algorithm have lead to improved efficiency of the sampling, such as the ABC–PMC algorithm of [8]. We proposed an ABC–PMC algorithm that can successfully handle finite mixture models. Some of the challenges with inference for finite mixture models are due to the complexity of the likelihood function including its possible multimodality and the exchangeability of the mixture component labels, leading to the label switching problem. Fortunately, ABC can handle complicated likelihood functions, but the label switching problem must be addressed. We suggested a procedure to address the label switching problem within the proposed

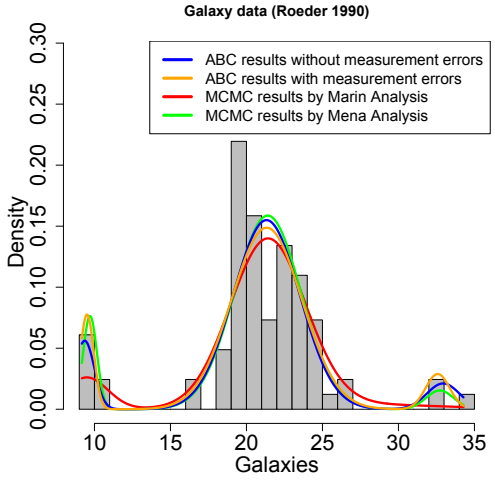

Figure 4.4: Histogram of the recessional velocity of 82 galaxies and the estimated three-component Gaussian mixture models for each study. The posterior means for the mixture weights, means and variances used are displayed in Table 4.4. It is possible to note that both the results coming from the extended ABC–PMC algorithm (blue and orange lines) and from [94] (green line) allow for a clear third component in the mixture. The results obtained by [89] (red lines) find a third component whose variance is in particular equal to $\sigma_3^2 = 34.1$, making the third cluster not appreciable in the Figure.

ABC algorithm that works well empirically. Some additional challenges with using ABC for mixture models include the selection of informative summary statistics and the definition of a kernel to move the mixture weights, since they are constrained to be between 0 and 1 and must sum to 1. For the summary statistics, we proposed to use the Hellinger distance between kernel density estimates of the real and simulated observations; this allows the multimodality of the data to be accounted for and compared between the two sets of data. We proposed a Dirichlet resampling algorithm to move the mixture component weights that preserves some information from the sampled particle, improving as well the efficiency of the ABC–PMC procedure (by not having to draw from the same Dirichlet prior at each time step). The mathematical assumptions required to run the proposed Dirichlet resampling algorithm are discussed in Appendix C.

The proposed ABC algorithm has been explored and tested empirically using popular examples from the literature. The resulting ABC posteriors were compared to the corresponding MCMC posteriors, and in all the considered cases the proposed ABC and MCMC posteriors were very similar. However, we recall that for the MCMC analyses good initial points were chosen to begin with the algorithm and the label switching was addressed by using deterministic approaches. We also considered, as application from real data, the recessional velocities of Galaxies from the Corona Borealis Region [122], which is commonly used to test the performances of a procedure that works

with mixture models. An advantage of ABC over other commonly used methods is that the forward model can be easily expanded to better represent the physical process that is being modeled. For the galaxy dataset, measurement errors are available in the original dataset, but are not generally used when analyzing the data. We slightly extended the proposed ABC–PMC forward model used in this example to include the measurement errors, which provides a more accurate assessment of the uncertainty in the data. Though the presented examples focused on one–dimensional GMM's, the component distributions can be easily changed to other distributions in the ABC forward model. Overall, the proposed ABC–PMC algorithm performs well and is able to recover the benchmark MCMC posteriors suggesting that ABC is a viable approach to carry out inference for finite mixture models.

Chapter 5

Measuring precise radial velocities and cross-correlation function line-profile variations using a Skew Normal distribution

In this Chapter we highlight one of the results of the collaboration between the Statistics and Data Science Department and the Astronomy Department at Yale University and the Department of Physics and the University of Geneve, where in the last years several researches have been done in the attempt to detecting and characterizing "Earth–like" extrasolar planets.

To detecting and characterizing extrasolar planets, direct techniques cannot be used since stars are much brighter than planets and also because the distance between those stars and the Earth is too large in order for us to retrieve a direct image of a planetary system. For these reasons indirect techniques are used where, looking at some particular behavior of the star, the presence of one or more planets is inferred. In particular two aspects of the star have been largely analyzed in recent years: the transit method evaluates photometric variations of the star (e.g. the Kepler project [21]), while the radial velocity method measures spectroscopic variations of the star [143]. Therefore, the presence of an extrasolar planet can be better understood by looking at the impact that its presence leads on photometric and spectroscopic aspects of the star. In this Chapter we focus on the radial velocity technique.

When working with radial velocities (RV's), the main limitation to the detection of small-mass exoplanets is not anymore the precision of the instruments used, but how to properly address the different noises coming from the stars we are observing [47]. In fact stellar oscillations, granulation phenomena and stellar activity introduce spurious RV's

signals [40, 45, 49, 112, 126] that are beyond the precision of $\mathrm{m\,s^{-1}}$ reached by the best state–of–the–art high-resolution spectrographs. It is therefore mandatory to understand as best as possible spurious signals coming from the intrinsic behavior of the star and to find ways to correct from them if in the near future we want to detecting Earth-twin planet using the RV technique. This goal is even more crucial now that instruments like EXPRESS [56] and ESPRESSO [104] should have the stability to theoretically detecting signals coming from Earth-twin planets. However, if solutions are not found to mitigate the impact of stellar activity when estimating RV's, the detection and the confirmation of potential Earth-twins planets will remain extremely challenging.

In Section 5.1 we outline the technological and methodological state–of–the–art challenges to detecting exoplanets, introducing the most common tools used by cosmologists. In particular we rigorously define the bisector inverse slope span, the most common used indicator to capture spurious variations in RV's caused by stellar activity. In Section 5.2 we introduce the Skew Normal distribution, motivating its use to fit the Cross Correlation Function. In Section 5.3 we show that the Skew Normal distribution provides a suitable density of the observed Cross Correlation Function and we study how the parameters of the Skew Normal relate to the radial velocity, the full width at half maximum and the bisector inverse slope span of the Cross Correlation Function. In Section 5.4 we present a simple model to correct the originally estimated set of RV's from spurious variations caused by stellar activity, aimed by the goal to retrieve a set of new RV's having information only about the (possibly) pure doppler shift caused by an exoplanet. In Section 5.5, we compare on real observations the information provided by the Skew Normal distribution with other common indicators. In Section 5.6 we derive error bars for the different parameters of the Cross Correlation Function and finally we discuss our results in Section 5.7.

5.1 Introduction

The RV of a star is defined as the velocity of the center of mass of the star along our line of sight [56]. The information on the RV of a star is contained in the wavelength position of its spectral lines. In particular, the RV of a star can be precisely estimated by measuring the doppler shift variations produced on the spectral lines of the star. This operation has done by using an instrument called spectrograph. For spectrographs that are not stabilized in temperature and pressure, the iodine technique is used, where the light of the star passes through a iodine cell before getting into the spectrograph to imprint the absorption spectrum of iodine on top of the stellar spectrum (the High

Resolution Spectrograph HRS [141], the Tull spectrograph [142], HIRES [145] on the Keck 10-m telescope and the Hamilton spectrograph [144] at Lick Observatory). In this case, if the spectrograph shifts because of changes in the atmospheric conditions, the iodine and the stellar spectrum are shifted in the same way. This leads to complications when reducing the data, because of the operation of "de–correlation" between the iodine spectrum and the stellar spectrum. On the other hand, for spectrographs that are stabilized, the spectrum of a calibration lamp is recorded close to the stellar spectrum on the CCD, which prevents contamination of the stellar spectrum (SOPHIE [22], HARPS-N [34], the High Accuracy Radial Velocity Planet Searcher (HARPS) [106], CORALIE [113], CARMENES [114]). When using this second class of spectrographs, reducing the data is easier since the stellar spectrum is not contaminated with iodine absorption lines.

In the case of spectrographs that are stabilized, the RV of the star is obtained at first by correlating the stellar spectrum with a synthetic [5, 103] or an observed stellar template [2], which gives an average line profile, generally called Cross Correlation Function (CCF). Since the exact regions of the mask associated with the absorption lines depend on the atmospheric properties of the star, there are 3 available different masks for 3 different spectral types (G2, K5 and M5). Further details about the operations required in order to obtain the CCF can be found in [24]. Once the CCF has been retrieved, a Normal density has fitted to this average line profile. The first two estimated parameters of the fitted Normal density are the mean, that defines the RV of the star and the variance, that defines the Full Width at Half Maximum (FWHM) of the line profile. The CCF technique allows for averaging out the RV information of thousands of lines in a stellar spectrum and therefore reaches a very high signal-to-noise ratio (SNR), which is essential to retrieve a precise set of RV's.

Among the different spurious stellar signals we are aware of, the one that is the most difficult to characterize and to correct for is the signal induced by stellar activity. Stellar activity is responsible for creating for instance magnetic regions on the surface of the star; these magnetic regions change locally the temperature and the convection of the star, inducing spurious RV's variations [46, 96]. In theory, it should be easy to differentiate between the pure Doppler-shift induced by a planet, that will shift the entire stellar spectrum and stellar activity, that modifies the shape of spectral lines and by doing that creates a spurious shift of the stellar spectrum [40, 46, 67, 80, 81, 87, 96, 126]. However, on quiet GKM dwarfs, the main target for precise RV's measurements, stellar activity induces spurious variations of few $\mathrm{m\,s^{-1}}$. This corresponds physically to variations smaller than 1/100th of a pixel on the detector. Moreover, the convection in external

layers of solar type stars is responsible for the granulation pattern than can be seen at high spatial resolution on the surface of the Sun. This behavior of the star changes the Normal shape profile of the spectral lines, that becomes asymmetric with a"C"-shaped profile [43]. The strength of the asymmetry depends on the velocity of the convection (approximatively $300\,\mathrm{m\,s^{-1}}$ for the Sun) but also on the depth formation of spectral lines [61]. Stellar activity is responsible also for the appearance of dark spots and bright faculae on the photosphere of the star, breaking the flux balance between the red-shifted and the blue-shifted halves of a rotating star. These active regions induce an asymmetry on the spectral lines and thus on the CCF. As the star rotates, spots and faculae move across the stellar disk, modifying the asymmetries of the line profiles and thus producing an apparent doppler shift [17, 40, 67, 80, 81, 126]. Spots and faculae are also regions where the magnetic field is strong. A strong magnetic fields reduces the stellar convection, modifying as well the asymmetry of the spectral lines [24, 30, 43, 46, 87, 96].

Since the CCF is an average of all the spectral lines, where some of them are strongly asymmetric and other ones are not, its asymmetry is rather small, which motivates the historical use of the Normal density as a reasonable model to fit the CCF. If the star has low levels of activity, the degree of the asymmetry produced by convection is constant as a function of time. However, this asymmetry slightly modifies the estimated RV's of the star, reducing the accuracy of the measurements and potentially leading to a false positive detection [24]. Stellar activity induces spurious variations in the RV's by modifying the asymmetry on the spectral lines, while an orbiting companion only induces a pure doppler shift on the spectral lines without modifying their shapes nor their widths. Therefore, assuming that there are not instrumental systematic errors, stellar activity will induce a variation in the asymmetry and in the width of the CCF. The asymmetry of the CCF is commonly retrieved by calculating at first the bisector of the CCF [146] and then deriving from it further indicators such as the bisector span [66, 138], the curvature of the bisector [66] or the bisector inverse slope span (BIS SPAN)[112].

5.1.1 The BIS SPAN parameter for measuring stellar activity

The bisector of the CCF is defined as a measure of the general asymmetry of the lines of a spectrum [146]. A rigorous definition of the bisector of the CCF can be found in [105]: "[...] the locus of median points midway between equal intensities on either side of a spectral line, thereby dividing it into two halves of equal equivalent width." For each point on the left of the line profile, a matching point on the right is found

with a cubic-spline interpolation [138]. The choice about how many points to consider to calculate the bisector of the CCF is not straightforward. Ad noted in [138]: "They (the points) were chosen on the basis of their strength and relative freedom from blends. Strong lines are needed to get enough points across the profile to define the bisector, and their steeper slopes also mean smaller bisector errors. Weak blends and spurious noise can render a bisector or portion of a bisector unusable. If we were to demand that the lines be completely free of blends, none would remain. By averaging the bisectors for several lines we reduce the effect of unknown blends."

The most used indicator derived from the bisector of the CCF is the BIS SPAN. In particular, according to [24, 105], the BIS SPAN is defined as the difference in bisector velocity between upper and lower regions of the CCF, avoiding wings and cores of the line profile. The BIS SPAN is computed by calculating the so–called mean velocity $B(d)$ at the depth d between both sides of the CCF peak:

$$B(d) = \frac{v_l(d) + v_r(d)}{2}, \tag{5.1}$$

where $v_l(d)$ corresponds to the velocities located on the left side from the minimum of the CCF peak and $v_r(d)$ corresponds to the velocities located on the right side from the minimum of the CCF peak. The mean bisector is then calculated at two depth ranges, identified respectively with the terms BOTTOM, from $d \in (0.1, 0.4)$, and TOP, from $d \in (0.6, 0.85)$, as shown in Figure 5.1. In this way the following two expectation quantities are defined:

$$\bar{V_t} = \mathbb{E}[B(d)], \quad \forall d \in (0.1, 0.4), \quad \bar{V_b} = \mathbb{E}[B(d)], \quad \forall d \in (0.6, 0.85). \tag{5.2}$$

The BIS SPAN is finally defined as:

$$\text{BIS SPAN} = \bar{V_t} - \bar{V_b}. \tag{5.3}$$

Several authors derived indicators that are more sensitive to changes in the asymmetry of the CCF than the BIS SPAN. Recently, [18] proposed a new indicator named V_{span} to calculate the asymmetry of the CCF that is more sensitive than the BIS SPAN in cases having low SNR. In another study, [54] invistigated the use of two new indicators, bi-Gaussian and V_{asy}. The authors were able to show that when using bi-Gaussian, the amplitude in asymmetry is 30% larger, allowing for the detection of smaller-amplitude correlations between this statistics and the estimated RV's. They also demonstrated that V_{asy} seems to be the best indicator to capture changes in the asymmetry of the

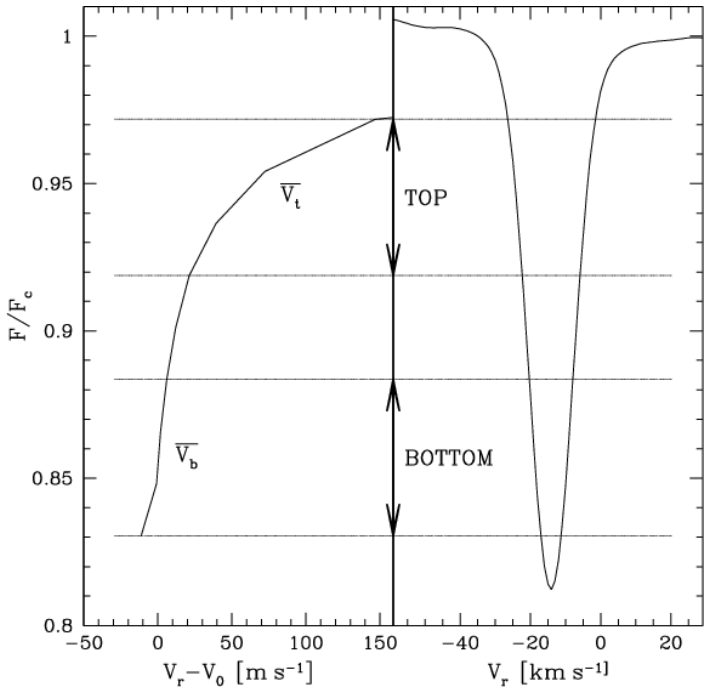

FIGURE 5.1: (left) The BIS SPAN of the CCF. V_0 is an arbitrary offset. Note the definition of the boundaries for the computation of ($\bar{V_t}$ and $\bar{V_b}$) [112]. (right) The bisector of the CCF for the star HD166435, constructed with a template selecting only the weak and non saturated lines. This profile represents the mean spectral-line profile of the lines selected by the template (i.e. the CCF). The original image was originally shown by [66].

CCF at high SNR, as its correlation with the retrieved RV's is stronger than any other asymmetry indicator that has until now been proposed.

A crucial step of the analysis is therefore to retrieve precise and informative statistics about stellar activity able to show strong correlations with the estimated RV's. In fact, as a RV signal is induced by activity, generally a strong correlation will be observed between the RV and chromospheric activity indicators like $\log(R'_{HK})$ or H-α [19, 48, 120], but also between the RV and the FWHM of the CCF or its BIS SPAN [19, 45, 111, 112]. A common strategy is that, when searching for a planetary signal, in addition to a Keplerian function (i.e. the function taking information about the presence of an extrasolar planet) the model includes in addition linear dependancies with the $\log(R'_{HK})$, the FWHM and the BIS SPAN [47, 52]. It is also common to add a Gaussian process to

the model in order to account for the correlated noise induced by stellar activity. The hyperparameters of the Gaussian process are generally trained on the different activity indicators [68, 115].

The major downside for all these different analyses is that the asymmetry of the CCF is retrieved as a second and separated operation with respect to the first operation that allows the researcher to estimate the RV's of the star (i.e. by fitting a Normal density and then retrieving its mean). The latter consideration is the reason why quantities such as BIS SPAN, V_{span} and V_{asy} can miss relevant information about stellar activity. On top of that, the procedure discussed above to retrieve the BIS SPAN and in particular the steps required by Equation (5.2) are not free from uncertainties, as pointed out in [24]. We note also that both selecting the number of points to estimate non parametrically the bisector of the CCF and to fix the levels of the depth d to calculate the mean bisector though Equation (5.1) are not unique straightforward choices. It follows that, consequently, different BIS SPAN's for the same analyzed CCF can be retrieved, leading potentially to different correlations between them and the estimated RV's. Moreover, when analyzing slow rotators stars such as the Sun, because of the limited spectral resolution of the spectrographs and the limited precision in retrieving the corresponding RV's, measuring the asymmetry of the CCF becomes challenging, resulting in complications to detecting very small–mass planets with the RV technique.

As a summary of the discussions introduced in this Section, the estimations respectively of the RV and the FWHM of a CCF have done separately from the evaluation of its asymmetry. The asymmetry of the CCF has usually estimated by retrieving the BIS SPAN. Anyway, all the parameters of interest of the CCF are correlated when stellar activity is dominant and therefore performing a step–by–step approach makes it difficult to correctly estimating the errors on the different parameters. In addition, the Normal density cannot take into account the natural asymmetry of the CCF, leaving correlated noise in the residuals. Finally, as already noted, we know that for solar-type stars and cooler dwarfs, the bisector of the CCF has a "C"-shape due to convective blueshift [43, 61]. Therefore, fitting the CCF using a model that naturally includes an asymmetry, like the Skew Normal density, should in principle lead to more precise results, providing at the same time more information about the spurious variations in RV's caused by stellar activity. For all these reasons we propose to conduct the entire analysis by fitting a Skew Normal density to the CCF, which naturally includes a skewness parameter [3].

5.2 The Skew Normal distribution

The Skew Normal (SN) distribution is a class of probability distributions which includes the Normal distribution as a special case [3]. The SN distribution has, in addition to a location and a scale parameter analogous to the Normal distribution's mean and standard deviation, a third parameter which describes the asymmetry, or the skewness, of the distribution. Considering a random variable $Y \in \mathbb{R}$ (where $\mathbb{R}$ is the real line) which follows a SN distribution with location parameter $\xi \in \mathbb{R}$, scale parameter $\omega \in \mathbb{R}^+$ (i.e., the positive real line), and skewness parameter $\alpha \in \mathbb{R}$, its density at some value $Y = y$ can be written as

$$SN(y; \xi, \omega, \alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\frac{\alpha(y-\xi)}{\omega}\right), \tag{5.4}$$

where ϕ and Φ are respectively the density function and the distribution function of a *standard* Normal distribution and $\alpha \in \mathbb{R}$ is the skewness parameter which quantifies the asymmetry of the SN. We then write $Y \sim SN(\xi, \omega^2, \alpha)$ to mean that the random variable Y follows the noted SN distribution. Examples of SN densities under different skewness parameter values and the same location and scale parameters ($\xi = 0$ and $\omega = 1$) are displayed in Fig. 5.2. The usual Normal distribution is the special case of the SN distribution when the skewness parameter, α, is equal to 0. This can be seen from Equation (5.4): if $\alpha = 0$ then $\Phi\left(\frac{\alpha(y-\xi)}{\omega}\right) = \Phi(0)$ and this is the probability for a standard Normal random variable to be less than or equal to 0, which is 0.5. The 0.5 cancels with the 2 in the denominator and what remains is the usual Normal density, $\frac{1}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)$.

For reasons related to the interpretation of the parameters in Equation (5.4) and well known computational issues with estimating α near 0 [3], a different parametrization is used, which is referred to as the *centered parametrization* (CP). We will be using the CP in this work, which includes a mean parameter μ, a variance parameter σ^2, and a skewness parameter γ. In order to define the CP, we need to express the CP parameters (μ, σ^2, γ) as a function of the ones used in the Equation (5.4) with (ξ, ω^2, α) by

$$\mu = \xi + \omega\beta, \quad \sigma^2 = \omega^2(1-\beta^2), \quad \gamma = \frac{1}{2}(4-\pi)\beta^3\left(1-\beta^2\right)^{-3/2}, \tag{5.5}$$

where $\beta = \sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{\sqrt{1+\alpha^2}}\right)$.

By using Equation (5.5), the new set of parameters (μ, σ^2, γ) provides a more clear interpretation of the behavior of the SN distribution. For the α values used in Figure 5.2, the corresponding values of μ, σ^2, γ are displayed in Table 5.1. In particular, μ and σ^2

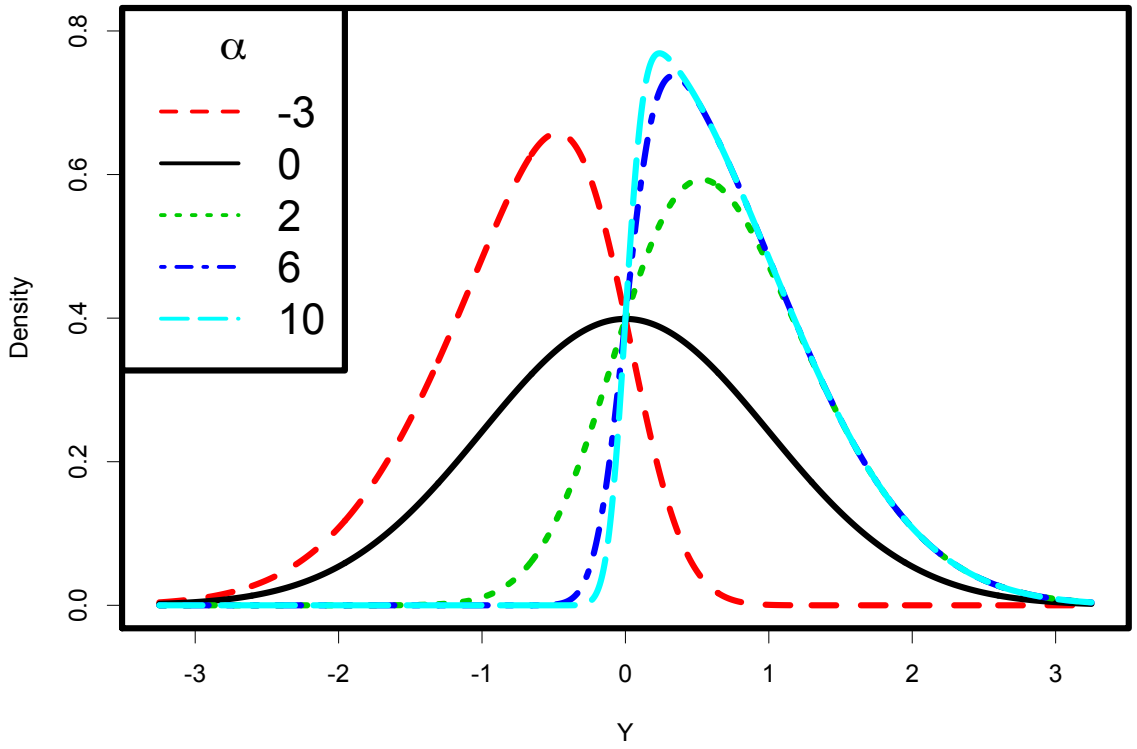

Figure 5.2: Density function of a random variable $Y \sim SN(\xi, \omega^2, \alpha)$ with location parameter $\xi = 0$, scale parameter $\omega = 1$ and different values of the skewness parameter α indicated by different colors and line types. Note that the solid black line has an $\alpha = 0$, making it a Normal distribution.

are the actual mean and variance of the distribution (rather than simply a location and scale parameter), and γ becomes the skewness parameter for evaluating the asymmetry of the SN.

Beyond the mean of the SN, it is convenient to introduce a second location parameter that will be largely used in the analyses: the median. Since the SN is an absolutely continuous random variable, its median is defined as that value m such that

$$\int_{-\infty}^{m} SN(y; \xi, \omega, \alpha) = \frac{1}{2}, \tag{5.6}$$

where $SN(y; \xi, \omega, \alpha)$ follows Equation (5.4).

Further details about the parametrization from Equation (5.4) (called *Direct Parametrization* or DP), the CP and general statistical properties of the SN are treated in rigorous mathematical and statistical viewpoints in the book by [4]. We note however that closed form expressions to estimate the parameters of the SN are not available, either using the maximum likelihood estimate or using the least squares algorithm. Therefore, the estimation must be done numerically. In the present work we used the quasi-Newton numerical optimization method [152].

α	μ	σ^2	γ
-3	-0.757	0.427	-0.667
0	0.000	1.000	0.000
2	0.714	0.491	0.454
6	0.787	0.381	0.891
10	0.794	0.370	0.956

TABLE 5.1: CP values, (μ, σ^2, γ), corresponding to the α values from Fig. 5.2 (with location parameter $\xi = 0$ and scale parameter $\omega = 1$) using Equation (5.5). Values are rounded to three decimal places.

5.3 Fitting the Skew Normal density to the CCF

The SN density shape is used to model the CCF. In particular we define the following function $f_{CCF}(y_i)$ to fit the CCF by using the least-squares algorithm:

$$f_{CCF}(y_i) = \mathrm{C} + \mathrm{A} \times SN(y_i; \mu, \sigma^2, \gamma), \quad i = 1, \ldots, n \tag{5.7}$$

where, beyond the previously defined tern (μ, σ^2, γ), C is an unknown offset fitting the continuum of the CCF, A is an unknown amplitude parameter known also with the name "*contrast*", $y_1, \ldots, y_n$ are the set of RV's considered for the CCF and finally n is equal to the number of points of the CCF. Note that the CCF is expressed in flux as a function of the lag of the cross-correlation template, expressed in RV. We note moreover that only the standard deviation of each data point is reported. The correlated noise is not coming from calibration but from the activity of the star. In order to take into account this information, we tried at first to consider an heteroskedastic function of the variance of the type $var(Y_i) = \sigma^2 + \sigma_i^2,\ i = 1, \ldots, n$, where σ_i represents the measurement error for each available point. By doing that we hoped to interpret the parameter σ^2 as an estimate of the pure variability of the CCF. Unfortunately, the problem we are trying to address consists in fitting a SN distribution to a set of points (namely the CCF); as a consequence of this, the parameter σ^2 is not actually catching the pure variability of the CCF also when considering the heteroskedastic function of the variance as the one defined above. Since there is not a general model to account for stellar activity, deriving the RV's and other parameters from observed spectra is beyond the goals of the present work. Therefore in the presented analyses we did not use the standard deviation of each data point available from the pipeline.

When using the least square estimates to retrieve the best set of parameters using the SN density, for few CCF's the quasi-Newton algorithm leads to an error, caused by the fact that γ exceeds its range, which is $\gamma \in (-0.995, 0.995)$. Analyzing further

this problem, we found out that some of the CCF's selected after a preliminary step of cleaning are probably the result of an error coming from the operation of cross correlation between the stellar spectrum and the template. When using the SN density, having an improperly shaped CCF can lead to problems with estimating γ. In fact, when using the quasi-Newton method (and almost any other numerical method for which the calculation of the Hessian matrix is required), unexpected behaviors of the CCF can lead to issues in properly defining γ, leading to an error. We note moreover that when using the Normal density to fit those particular improperly shaped CCF's, no errors using the least squares are returned, but the estimates for the amplitude parameter A and for the variance σ^2 result much larger than the expected ones (based on CCF's close in time to the studied one). In Figure 5.3 and Figure 5.4 those few CCF's for which the least squares algorithm implemented by using the SN density leads to an error are displayed, respectively for the stars Alpha Centauri B and Tau Ceti. It is straightforward to notice the difference with an expected suitable CCF, comparing these CCF's with the one for example presented in the right plot of Figure 5.1. We note also that a similar problem when using the SN distribution in an optimization problem was found in [53]. In this case the authors suggested that this problem could be caused by having a small sample size n, suggesting to perform the optimization problem by using a grid of starting values, in the attempt to ensure that the true global maximum was reached. In our analyses, depending on the star, the number of points of each CCF varies between $n = 40$ and $n = 50$ points and we have not experienced problems in retrieving the best set of parameters because of the sample size.

In conclusion, since the problems related to the estimation of the parameters of interest of the CCF are related to the operation of cross correlation between the stellar spectrum and the template, we decided to discard those few CCF's that are not actually matching the shape of a line profile. We emphasize on the fact that we discarded these CCF's because their improper shapes clearly suggest an error during the operation of cross correlation between the stellar spectrum and the template. In particular all the CCF's for which the least squares algorithm returned an error have got an extremely low SNR (SNR < 10). Moreover, only few CCF's among the thousands that have been analyzed present an improper shape; therefore discarding those few CCF's does not compromise the achievability of the goals of the analysis.

In the following of the Chapter, we define N mean RV as the mean of the Normal density fitting the CCF. Concerning the fit of the CCF using the SN, we present at first 2 indicators that define the RV of the star: the mean of the SN, defined as SN mean RV and the median of the SN, defined as SN median RV (i.e. looking at Equation

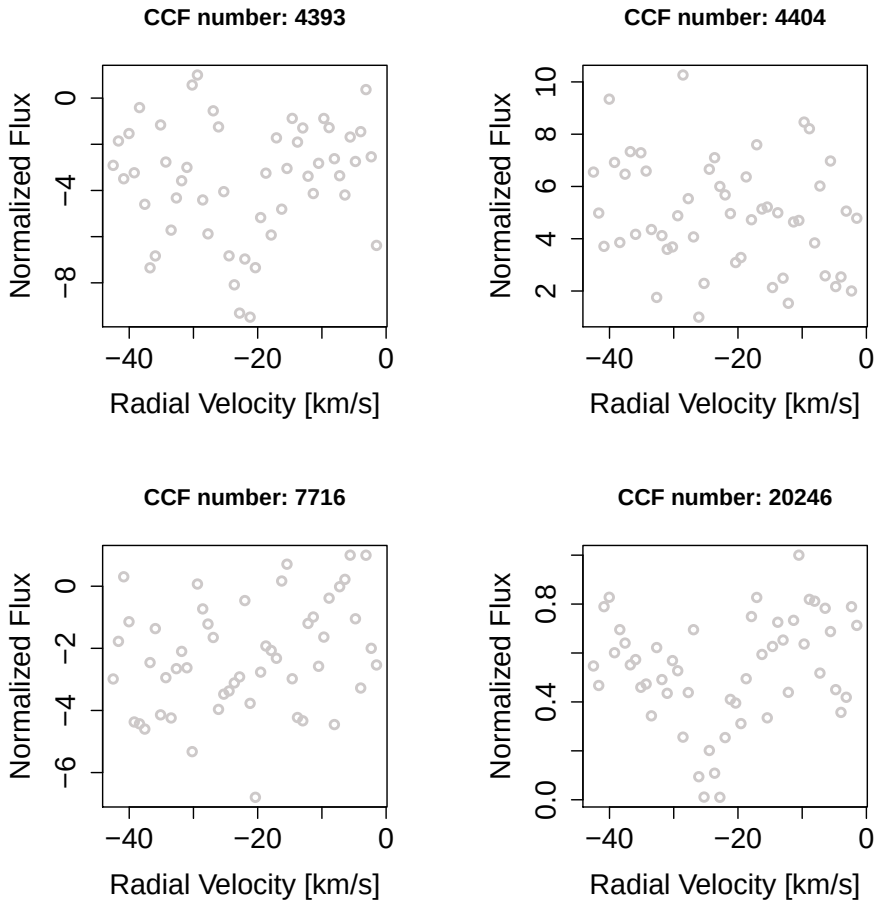

FIGURE 5.3: Among the 1812 analyzed CCF's for Alpha Centauri B, 4 CCF's leads to problems when using the SN distribution. For all the CCF's, the shape of the profile is not recognizable and the numerical minimization using the SN distribution leads to γ exceeding its range, forcing the procedure, correctly, to arrest.

(5.6), m = SN median RV). We will discuss advantages and limits for both these choices through the examples presented in Section 5.5. Concerning the width of the CCF, we use the FWHM of the Normal, which is $2\sqrt{(2\ln 2)}\sigma$. The width of the SN, SN FWHM, is defined in the same way. We note that SN FWHM does not correspond to the width of the SN distribution at half maximum like in the Normal case, but we decided to use this same definition because we have not found any remarkable difference between a numerical estimation of the FWHM of the SN and the approximation derived by $2\sqrt{(2\ln 2)}\sigma$. We also note that, to compare the SN fit to the Normal fit, we are interested mainly in evaluating the ratio between SN FWHM and FWHM, resulting therefore in the ratio between the variances estimated respectively with the SN and Normal densities. Being a Normal density symmetric, there is not such a parameter that evaluates the asymmetry of the distribution, so the BIS SPAN is used. The BIS SPAN is compared with the asymmetry parameter of the SN under CP, γ, defined in

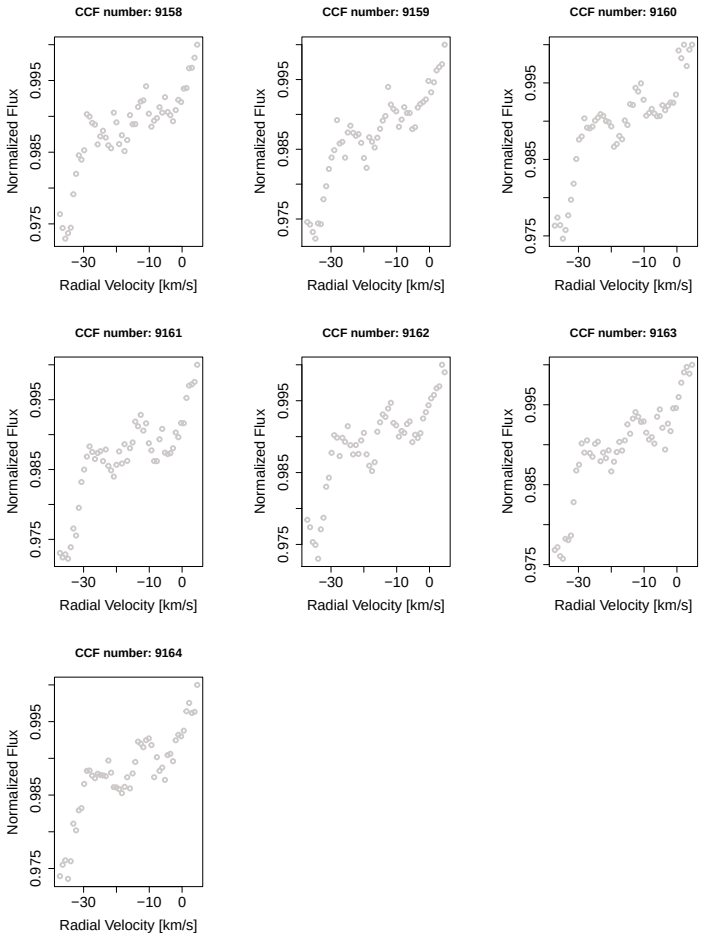

Figure 5.4: Among the 7935 analyzed CCF's for HD10700, 7 CCF's leads to problems when using the SN distribution. It is worth noting that all these CCF's = {9158, ..., 9164} are consecutive. For all of them, the profile does not follow to the shape of a proper absorption line.

the analyses also as SN GAMMA.

To test the strength of the correlation between the estimated RV's and the different indicators used to evaluate stellar activity, we calculated the Pearson correlation coefficient, which in its general form is defined as:

$$R(x, y) = \frac{cov(x, y)}{\sigma(x), \sigma(y)}, \tag{5.8}$$

where x and y are two quantitative variables, $cov(x, y)$ indicates the covariance between x and y, and $\sigma(x)$ and $\sigma(y)$ represent their standard deviations. A $p-$value for the statistical test having null hypothesis $H_0 : R = 0$ is provided, along with a 95 % confidence interval for R.

5.4 Radial Velocity correction function for stellar activity

Exoplanets will produce only variations in RV's induced by a pure doppler shift on the stellar spectra. Stellar activity, on the contrary, does not produce a blueshift or redshift of the spectra, but creates spurious RV's signal by modifying the shape of the spectral lines and therefore of the CCF. To track changes in the shape of the line profile, the general approach consists in using the FWHM, the BIS SPAN or the indicators introduced by [18, 54], which provide an information on the average width and asymmetry of the CCF. A strong correlation between the estimated set of RV's and one or more of these parameters provides an indication that the RV's are affected by stellar activity signals rather than by pure doppler shift variations.

To correct the estimated RV's from spurious variations caused by stellar activity, it is common to consider a linear combination of the RV's with the BIS SPAN and the FWHM (or γ and SN FWHM in the SN case), as shown in [47, 52]. Hence:

$$RV_{\text{stellar activity}} = \beta_0 + \beta_1\gamma + \beta_2\text{SN FWHM} + \epsilon, \tag{5.9}$$

where β_0 is the intercept and ϵ is the vector of the errors with mean equal to 0 and covariance matrix equal to $\sigma^2 I$ (I defined as the identity matrix). When the Normal fit is used, in Equation (5.9) the parameter γ is replaced by the BIS SPAN and the SN FWHM is replaced by FWHM. In order to show the capacity of this function to correct from stellar activity the originally retrieved RV's, a statistical test on the parameters β_0, β_1 and β_2 is presented, where the null hypothesis is $H_0 : \beta_i = 0$, for $i = 0, 1, 2$. The level for not rejecting the null hypothesis is fixed equal to 0.05. The coefficient of determination R^2 is introduced in order to explain how well this linear combination addresses the variability of the RV's of the star as caused by stellar activity.

We note however that a more complex law than the one proposed in Equation (5.9) is probably needed. To find a more realistic way to correct the original set of estimated RV's from stellar activity is something we will evaluate in future analyses.

5.5 Illustrative Examples

In this Section we analyze five stars, showing the advantages of using the SN density over the Normal density to fit the CCF, according to the definitions provided in Section 5.3. We emphasize that by using the SN density all the necessary statistics (i.e. the

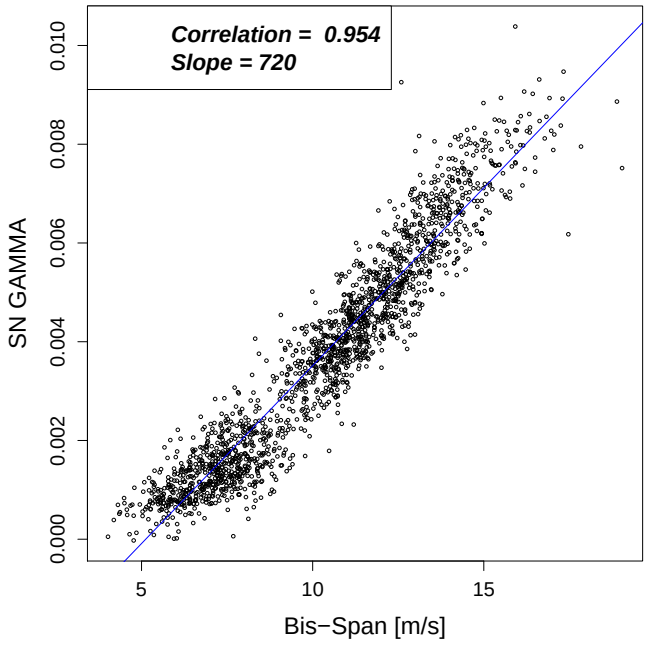

Figure 5.5: Correlation between γ and the BIS SPAN for Alpha Centauri B.

mean, the median, the variance and the skewness of the CCF) are naturally available in one step. A comparison with the results obtained by the classic approach has done, where the RV's of the stars are estimated by calculating the mean of the Normal density used to fit the CCF. When the analyses are performed using the Normal density, the evaluation of the asymmetry of the CCF requires a second step, necessary to retrieve the BIS SPAN or some of the other statistics proposed by [18, 54].

5.5.1 Alpha Centauri B

We first analyze Alpha Centauri B, where 1808 CCF's measured in the year 2010 have been studied. Several measurements taken throughout the year 2010 are contaminated by the presence of the companion star Alpha Centauri A, meaning that as preliminary step to retrieve uncontaminated spectra for Alpha Centauri B, we performed the same selection presented in [48].

We begin our analysis by evaluating the correlation between γ and the classic BIS SPAN. In fact, while the BIS SPAN has got a unit of measure (m/s), the γ parameter of the SN is adimensional. In Figure 5.5 is possible to note that the relationship between γ and the BIS SPAN is linear, with a slope equal to 720 and a correlation of $R = 0.954$. Figure 5.6 shows the comparison between the RV's retrieved using the SN density and the RV's obtained with the Normal density. In particular, as location parameters of the SN density, SN mean RV (black dots) and SN median RV (cyan crosses) are proposed.

Parameter	N mean RV	SN mean RV	SN median RV
β_0	0.0066	$2.29e-11$	$2.29e-11$
β_1	0.022	$2.22e-16$	0.03
β_2	$2.22e-16$	$2.22e-16$	$2.22e-16$
R^2	0.49	0.72	0.57

TABLE 5.2: **Alpha Centauri B**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equarion (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. All the three parameters are useful in explaining variations in RV's of the star that can be caused by stellar activity. Anyway the evaluation of the R^2 shows that the linear combination better explains variations in RV's due to stellar activity coming from the SN analysis which uses SN mean RV.

Concerning the Normal fit, N mean RV is used as location parameter (red triangles). This dataset for Alpha Centauri B, as shown also in [48, 136], presents a strong stellar activity signal. Looking at Figure 5.6 we can see that the RV's retrieved by using as location parameter SN mean RV show more variability than the RV's calculated with the Normal density. SN mean RV seems to be more sensitive to stellar activity. This can be explained by the fact that since the SN includes an asymmetry parameter, SN mean RV gets a shift because of γ. However, when using SN median RV, the variability caused by stellar activity seems smaller, suggesting that this location parameter could be used to define the set of RV's of the star. Both the location parameters derived from the SN fit have desirable properties: SN mean RV seems to capture changes in the asymmetry of the CCF caused by stellar activity and SN median RV can provide in principle a more robust global indicator to define the RV's of the star.

Using Equation (5.9), we provide a new set of RV's corrected from stellar activity. The results are shown in Figure 5.7. We see that, once corrected for stellar activity, the residuals for the Normal and SN analysis are comparable. However, we note that when using SN mean RV, the correction is more important. This is confirmed by the statistical tests on the significance of the parameters β_0, β_1 and β_2, whose results are summarized in Table 5.2. The intercept and both the variables γ (or BIS SPAN) and SN FWHM (or FWHM) are necessary to correct the originally estimated RV's from spurious variations caused by stellar activity for both the SN and the Normal fits. The comparison of R^2 shows as well that the SN fit accounts for a higher percentage of variability in RV's caused by stellar activity (i.e. spurious variations in RV's).

A comparison of the correlation between the different activity indicators and the RV's of the star is presented in Figure 5.8. The correlation between γ and SN mean RV is significantly stronger, almost twice, than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.741$). We note how

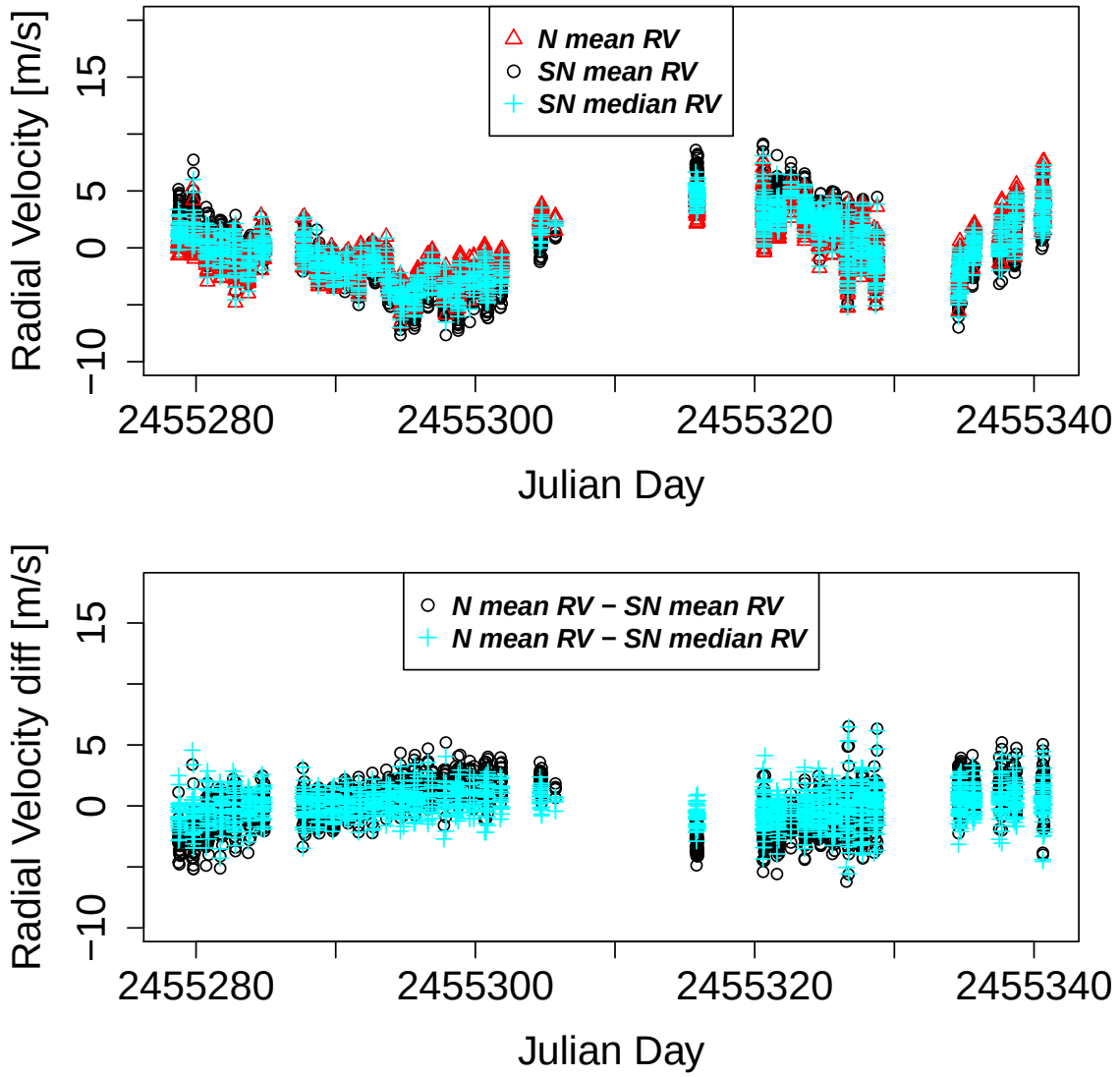

Figure 5.6: (top) RV's and (bottom) RV's differences for Alpha Centauri B considering a Normal and a SN fitting. Two location parameters are proposed using the SN density, SN mean RV (black dots) and SN median RV (cyan crosses), while the location parameter for the Normal fit is N mean RV (red triangles).

the correlation between SN mean RV and γ is higher than the correlation between SN median RV and γ, suggesting again that the first indicator is more sensible to spurious variations in RV's caused by stellar activity. The correlation between SN FWHM's and RV's is as well stronger when fitting a SN density respect using the common Normal density ($R = 0.817$). All the correlations are anyway statistically different from 0.

Analyzing further the data coming from Alpha Centauri B and by looking closer to the data plotted in Figure 5.9, it is straightforward to note that there are three distinct temporal clusters in the Alpha Centauri B measurements and each cluster has a different linear relationship between its asymmetry parameter γ and RV's (and also

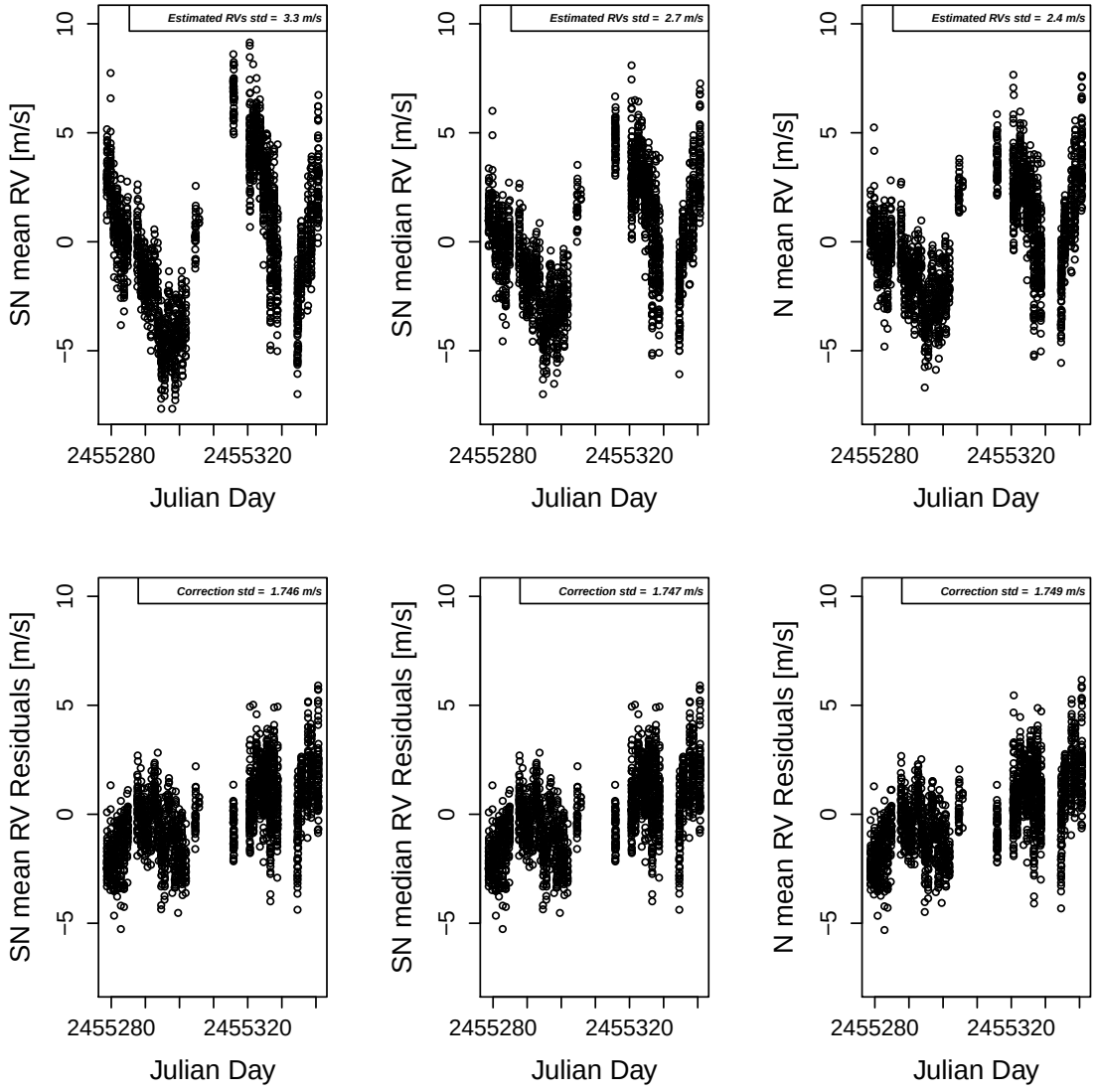

FIGURE 5.7: (top) Set of RV's for Alpha Centauri B estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable.

SN FWHM and RV's, though not displayed here). When considering Equation (5.9) and the subsequent inferences, this clustering is not accounted for in the model. A slightly more general linear model that allows for different intercepts and different slopes for the three clusters for γ and SN FWHM can be considered. Adjusting the RV's for stellar activity using this expanded model produces the corrected RV's displayed in the left plot of Figure 5.10. These corrected RV's are different from those displayed in the lower left plots of Figure 5.7 (which apply the correction derived from Equation (5.9)); the right plot of Figure 5.10 displays the difference between the two sets of corrected RV's. The long-term trend can be explained by the fact that the RV's drifts induced by the companion Alpha Centauri A is not well corrected. However, the shorter-term variations

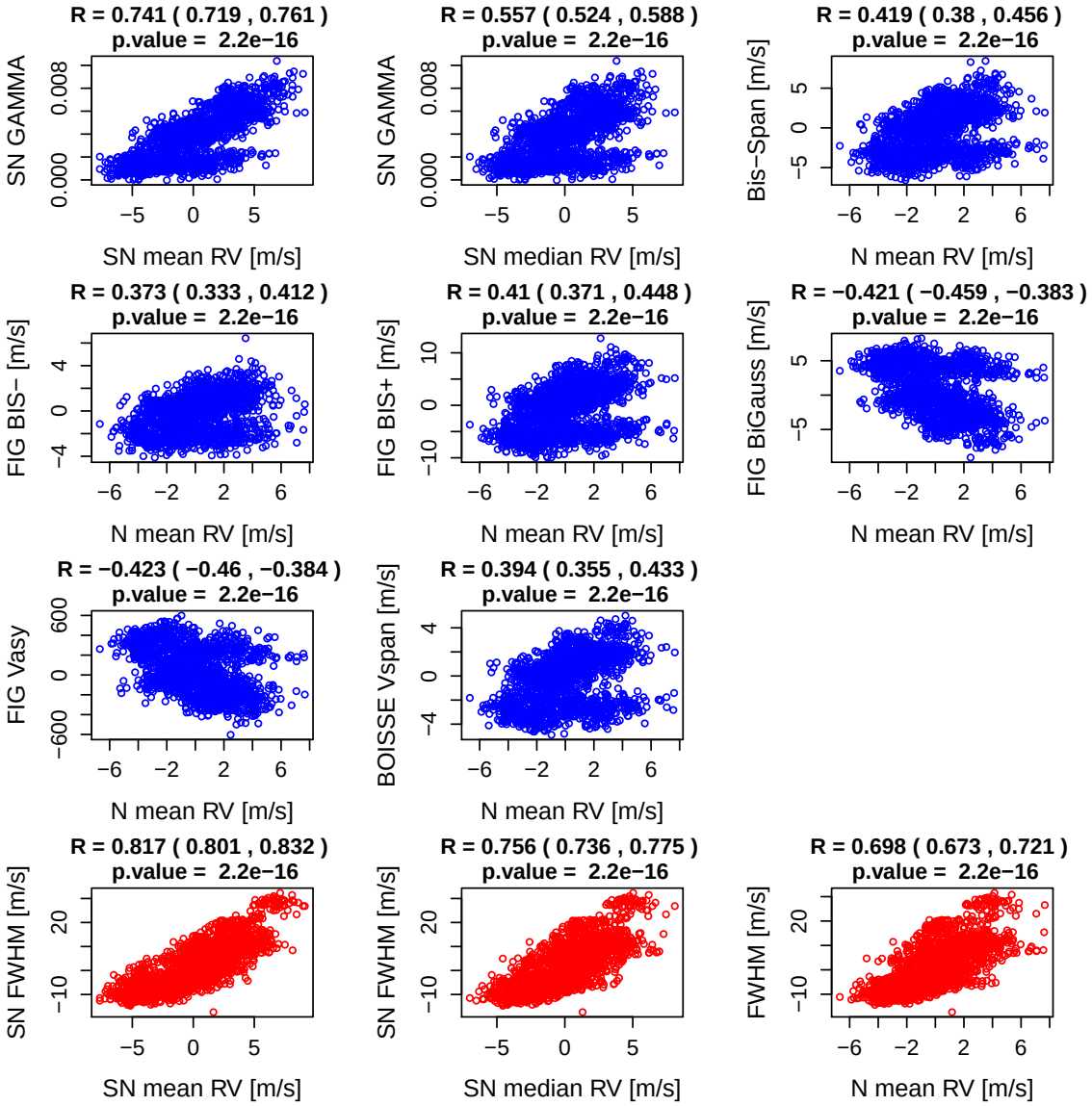

Figure 5.8: Correlation between the asymmetry parameters and the RV's for Alpha Centauri B. The last three plots show the correlation between the FWHM's and the RV's for Alpha Centauri B, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is significantly higher, almost twice, than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.741$). The comparison of the correlations between FWHM's and RV's shows that the indicators retrieved by the SN fit have stronger correlations than the one obtained with the common analysis ($R = 0.817$).

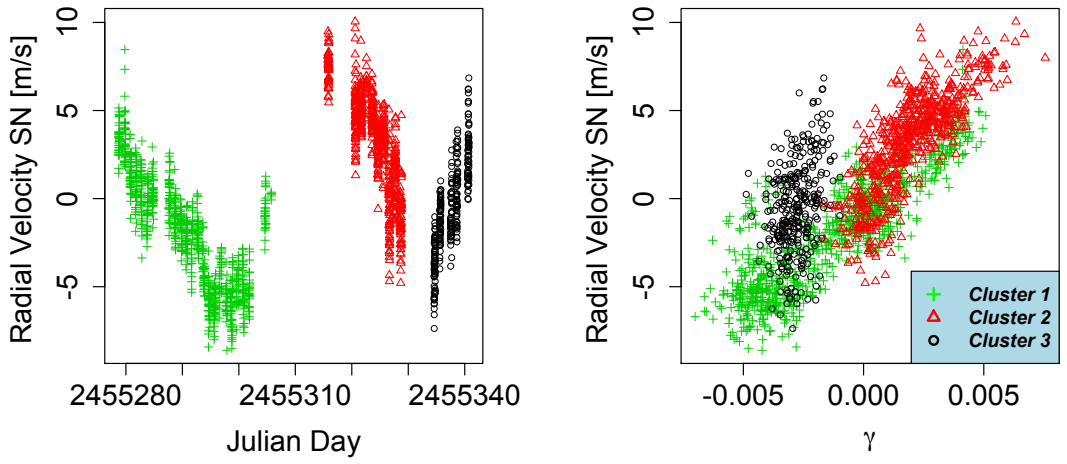

FIGURE 5.9: The RV's as a function of time (left) and the RV's plotted against γ (right) with colors and plot symbol according to its temporal cluster assignment for Alpha Centauri B. The RV's are expressed in $\mathrm{m\,s^{-1}}$.

show that stellar activity varies as a function of time, with spots and faculae evolving, leading to changes in the correlation between the RV's with respect to the asymmetry parameter γ and SN FWHM. These temporal variations are something that we want to explore more, although this Chapter presents already a significant effort in trying to understand and to remove from the data spurious variations in RV's caused by stellar activity.

5.5.2 HD192310

We present now the results of the analysis for the star HD192310 (also known as Gliese785). The dataset consists in 1577 CCF's. The correlation between γ and the BIS SPAN is 0.888 and the slope of the fitted linear regression is 786, as shown in Figure 5.11.

Looking the top three plots of Figure 5.12 it is possible to note that the RV's obtained with the SN analysis, in particular when using SN mean RV, present larger residuals than the RV's obtained with the N mean RV. However, once corrected for stellar activity using the linear combination with γ and SN FWHM (or BIS SPAN and FWHM) presented in Equation (5.9), the results of the three analyses are comparable, as shown in the three bottom plots of Figure 5.12. Table 5.3 summarized the tests conducted to evaluate the role played by stellar activity in introducing spurious signals on the estimated RV's. When using N mean RV, variations in the BIS SPAN are not statistically helpful to explain variations in RV's. On the contrary γ is helpful to explain spurious variations in

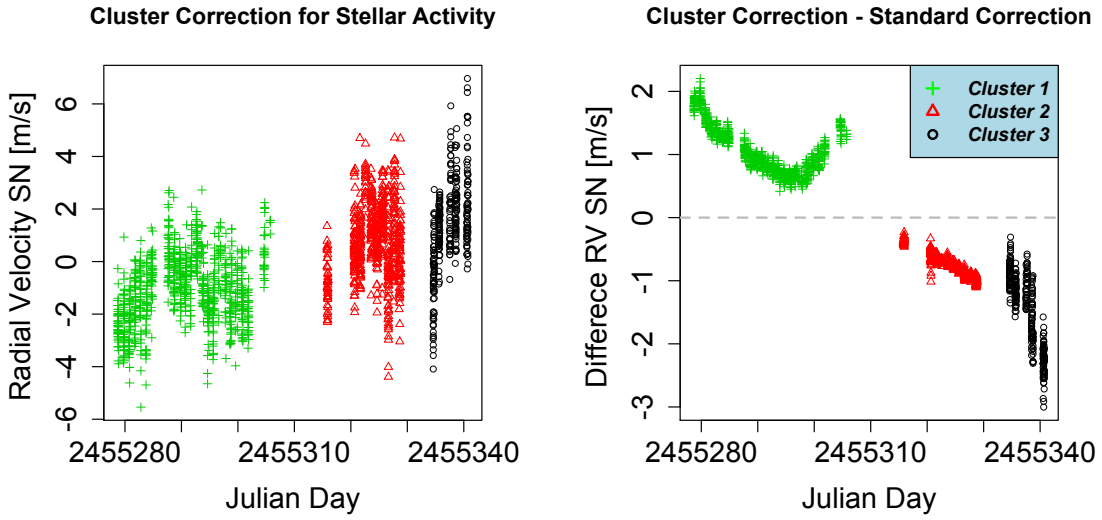

Figure 5.10: The RV's for Alpha Centauri B corrected from stellar activity using the SN fit and accounting for the temporal clusters (left), and the difference between those values in the left plot and the analogous values without accounting for the temporal clusters (right) which are displayed in the lower left plot of Figure 5.7. The RV's are expressed in $\mathrm{m\,s^{-1}}$.

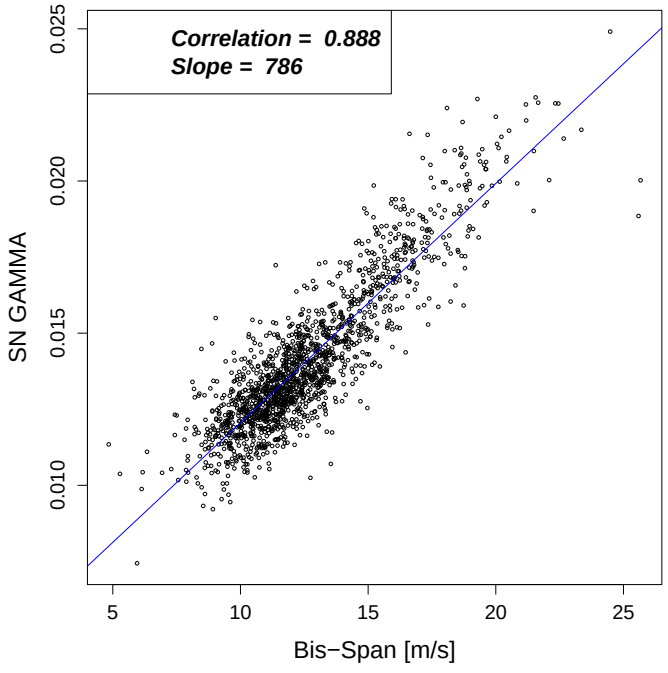

Figure 5.11: Correlation between γ and the BIS SPAN for HD192310.

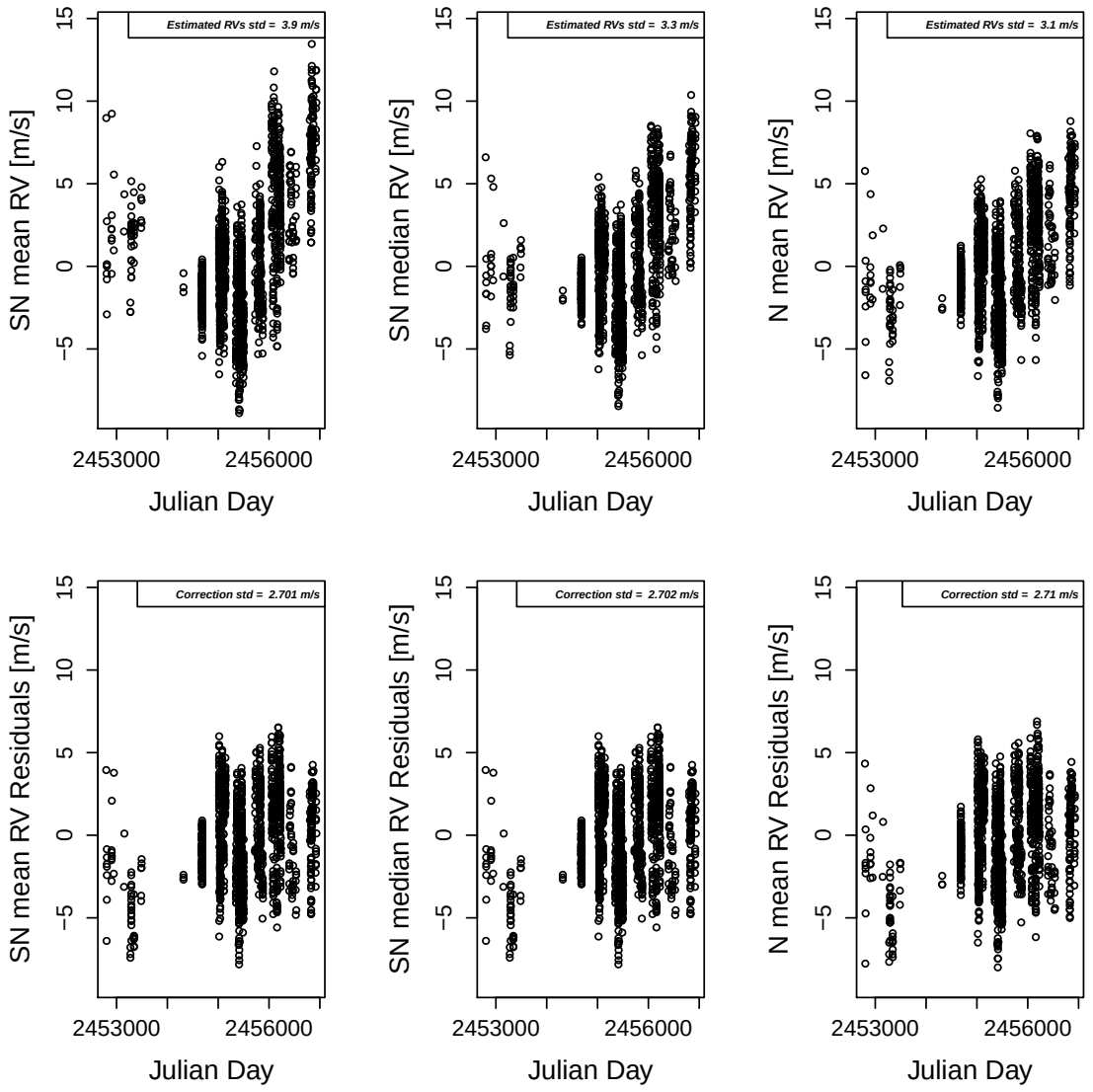

FIGURE 5.12: (top) Set of RV's for HD192310 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable.

RV's when fitting the SN to the CCF. Like for Alpha Centauri B, the Pearson correlation coefficient R^2 shows that the model we used to correct for stellar activity is more useful in the SN case rather than in the Normal one, in particular when using SN mean RV ($R^2 = 0.53$).

The comparison between the asymmetry parameters and the RV's is presented in Figure 5.13. The correlation between γ and SN mean RV is stronger ($R = 0.669$) than the correlation calculated between the other asymmetry statistics and their corresponding RV's. The comparison of the correlations between FWHM's and RV's leads to the same conclusion, with the correlation between SN mean RV and SN FWHM to be the strongest ($R = 0.666$).

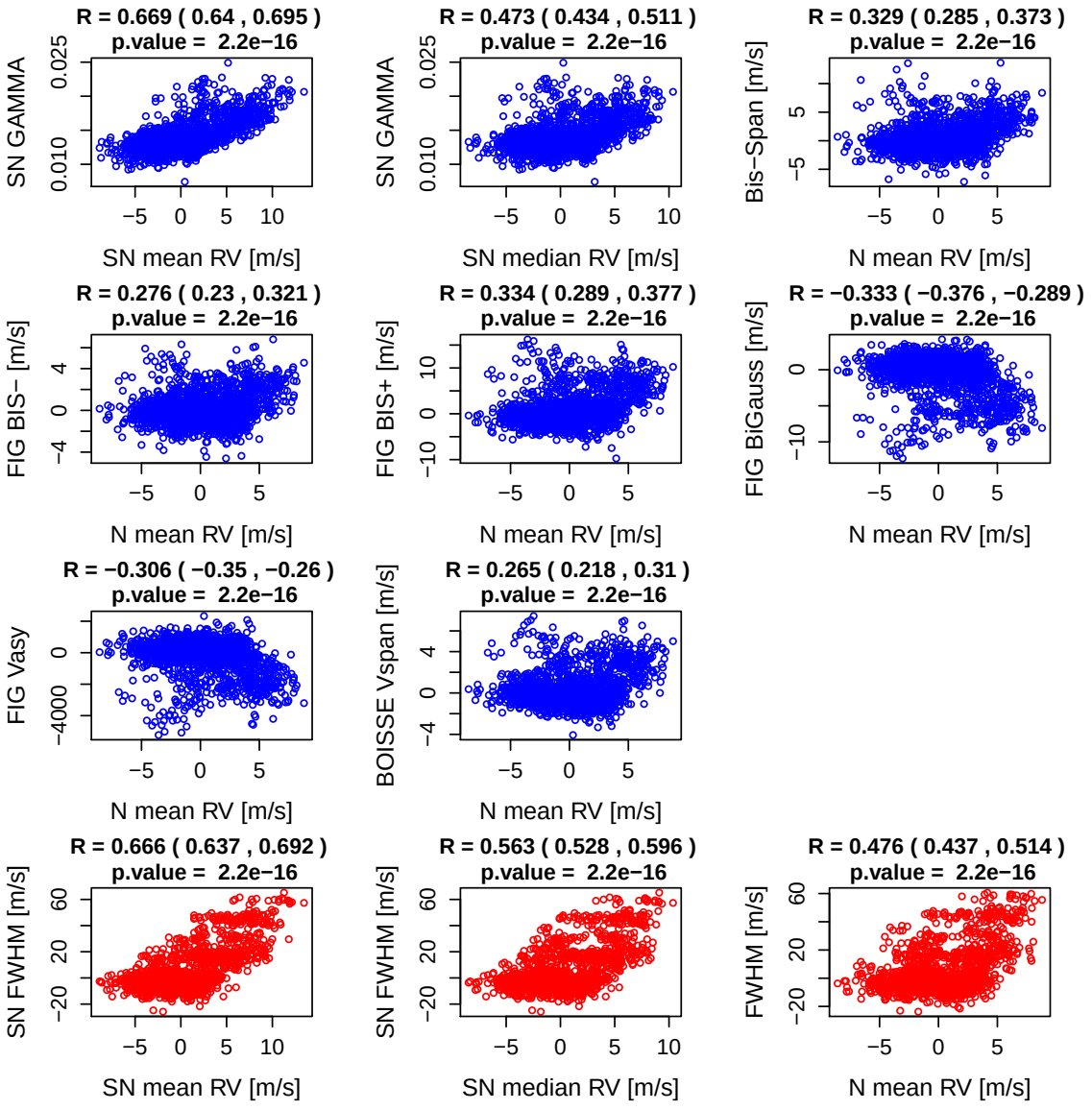

Figure 5.13: Correlation between the asymmetry parameters and the RV's for HD192310. The last three plots show the correlation between the FWHM's and the RV's using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.669$). The comparison of the correlations between FWHM's and RV's leads to the same conclusion, with the correlation between SN mean RV and SN FWHM to be the strongest ($R = 0.666$).

Parameter	N mean RV	SN mean RV	SN median RV
β_0	$2e-16$	$2.22e-16$	$2.22e-16$
β_1	0.24	$2.22e-16$	$2.22e-16$
β_2	$2e-10$	$2.22e-16$	$2.22e-16$
R^2	0.23	0.53	0.33

TABLE 5.3: **HD192310**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the BIS SPAN is not statistically useful to explain variations in the RV's of the star. On the other hand, concerning the analyses based on the SN density, all the p–values associated with the parameters involved in Equation (5.9) are statistically different from 0. The evaluation of the R^2 shows that the linear combination better explains variations in RV's due to stellar activity coming from the SN analysis which uses SN mean RV.

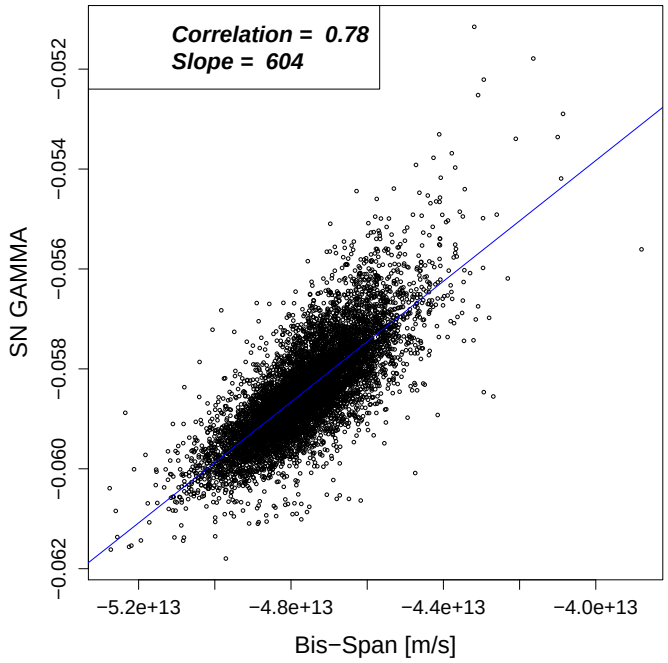

FIGURE 5.14: Correlation between γ and the BIS SPAN for HD10700.

5.5.3 HD10700

The analysis of the star HD10700 (also known as Tau Ceti) consists of 7928 CCF's. Figure 5.14 shows the relation between γ and the BIS SPAN, with a correlation of $R = 0.78$ and a slope of the fitted linear regression equal to 604. These values are smaller with respect to the ones retrieved for the previous analyzed stars, probably because HD10700 is at a very low activity level, similar to the Sun at its minimum phase of activity.

The RV's derived with the SN, using SN mean RV and SN median RV, present slightly larger residuals, as highlighted in Figure 5.15. However, once corrected from

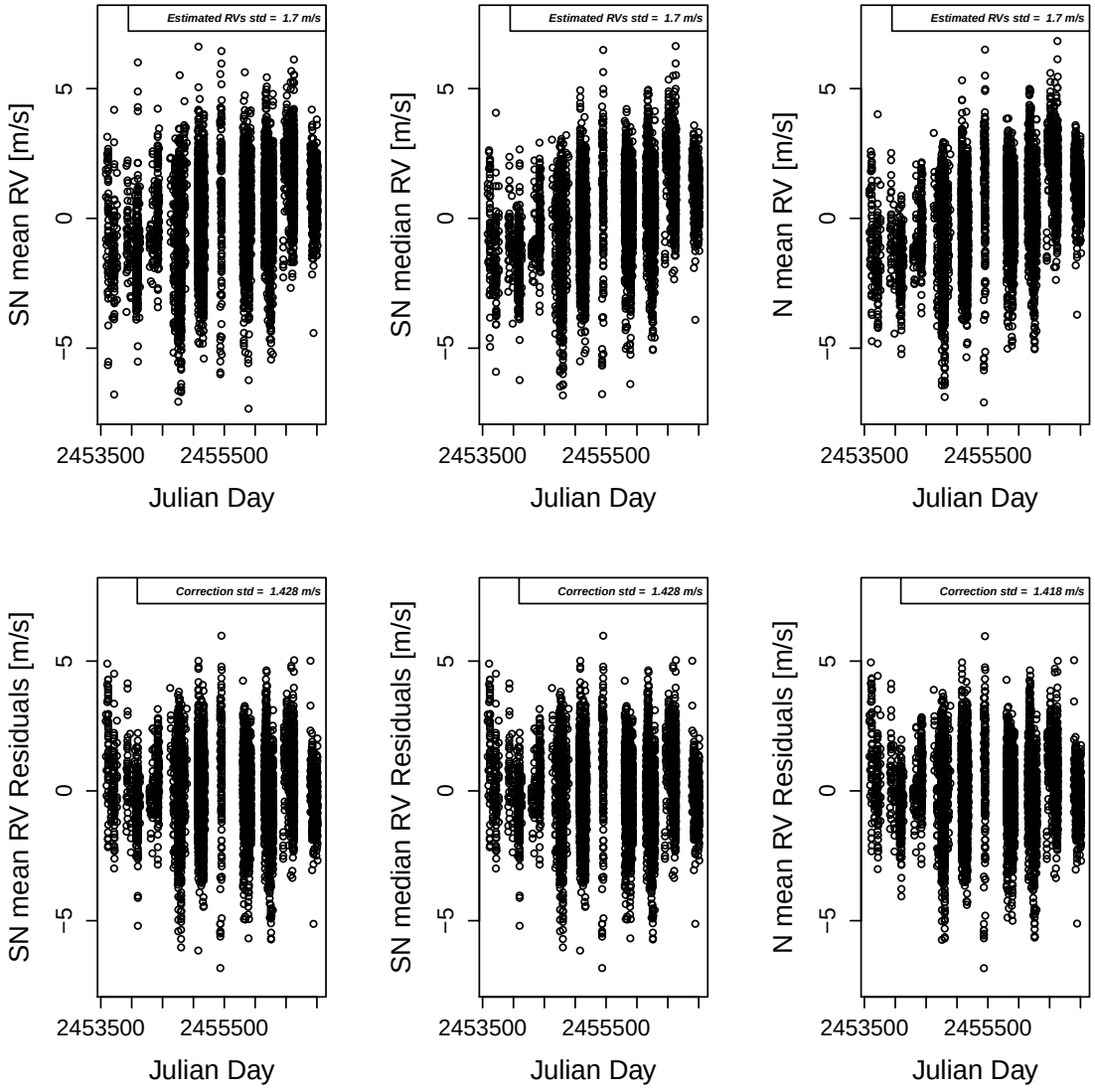

Figure 5.15: (top) Set of RV's for HD10700 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). Once corrected for stellar activity, the residuals for the Normal and SN analyses are comparable.

stellar activity, the residuals of the new set of RV's using the Normal and the SN fit are comparable. Looking at Table 5.4, we see that for both the Normal and SN analyses the intercept, the FWHM and the asymmetry of the CCF (γ or BIS SPAN) can explain part of the variations in RV's as caused by stellar activity. The analyses on the R^2 shows that for this star the correction for stellar activity is equally important for the three analyses.

The comparison between the asymmetry parameters and the RV's is presented in Figure 5.16. The correlation between γ and SN mean RV is stronger ($R = 0.322$) than the correlation calculated between the other asymmetry statistics and their corresponding

Parameter	N mean RV	SN mean RV	SN median RV
β_0	0.00013	$2.22e-16$	$2.22e-16$
β_1	$4.83e-6$	$2.22e-16$	$2.22e-16$
β_2	$2.22e-16$	$2.22e-16$	$2.22e-16$
R^2	0.28	0.33	0.27

TABLE 5.4: **HD10700**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. All the three parameters are useful in explaining variations in RV's of the star that can be caused by stellar activity. The R^2 shows that the correction for stellar activity is equally important for the three analyses.

RV's. In this case also the correlation between SN median RV and γ is weak, suggesting again that the SN mean RV parameter better captures changes in the CCF caused by active regions, in particular for stars having low activity levels, like HD10700. The comparison of the correlations between the FWHM's and RV's, when using the Normal and the SN fit leads, leads to comparable considerations. We note however that only for this star the correlation between N mean RV and FWHM is the strongest ($R = 0.529$).

5.5.4 HD215152

The analysis of the star HD215152 consists in 273 CCF's and Figure 5.17 shows that the slope of the linear regression between γ and the BIS SPAN is 794 and the Pearson correlation coefficient is $R = 0.763$. Figure 5.18 shows the RV's measured with the SN or the Normal density and their corresponding RV's residuals, once corrected for stellar activity. In this case the results are not comparable. While the correction from stellar activity leads to similar considerations when SN mean RV or SN median RV are used, using N mean RV leads to residuals $0.062\,\mathrm{m\,s^{-1}}$ higher. The proposed function that tries to correct from stellar activity seems to be useful in addressing spurious variations in RV's only when SN mean RV is used ($R^2 = 0.34$). This is probably because of the presence of planetary signals in the data [91]. We note however that the information on the orbital phase of the planet is not available in [91] and therefore we cannot remove those pure doppler shift signals. The statistical tests on β_0, β_1 and β_2, summarized in Table 5.5, show that in the Normal case BIS SPAN is not statistically significant, while for the intercept and the FWHM the test with level 0.05 is barely significant. Concerning the analyses based on the SN density, the parameter γ is always useful to explain part of the spurious variations in RV's caused by stellar activity. However, the proposed correction for stellar activity is more useful when using SN mean RV ($R^2 = 0.34$).

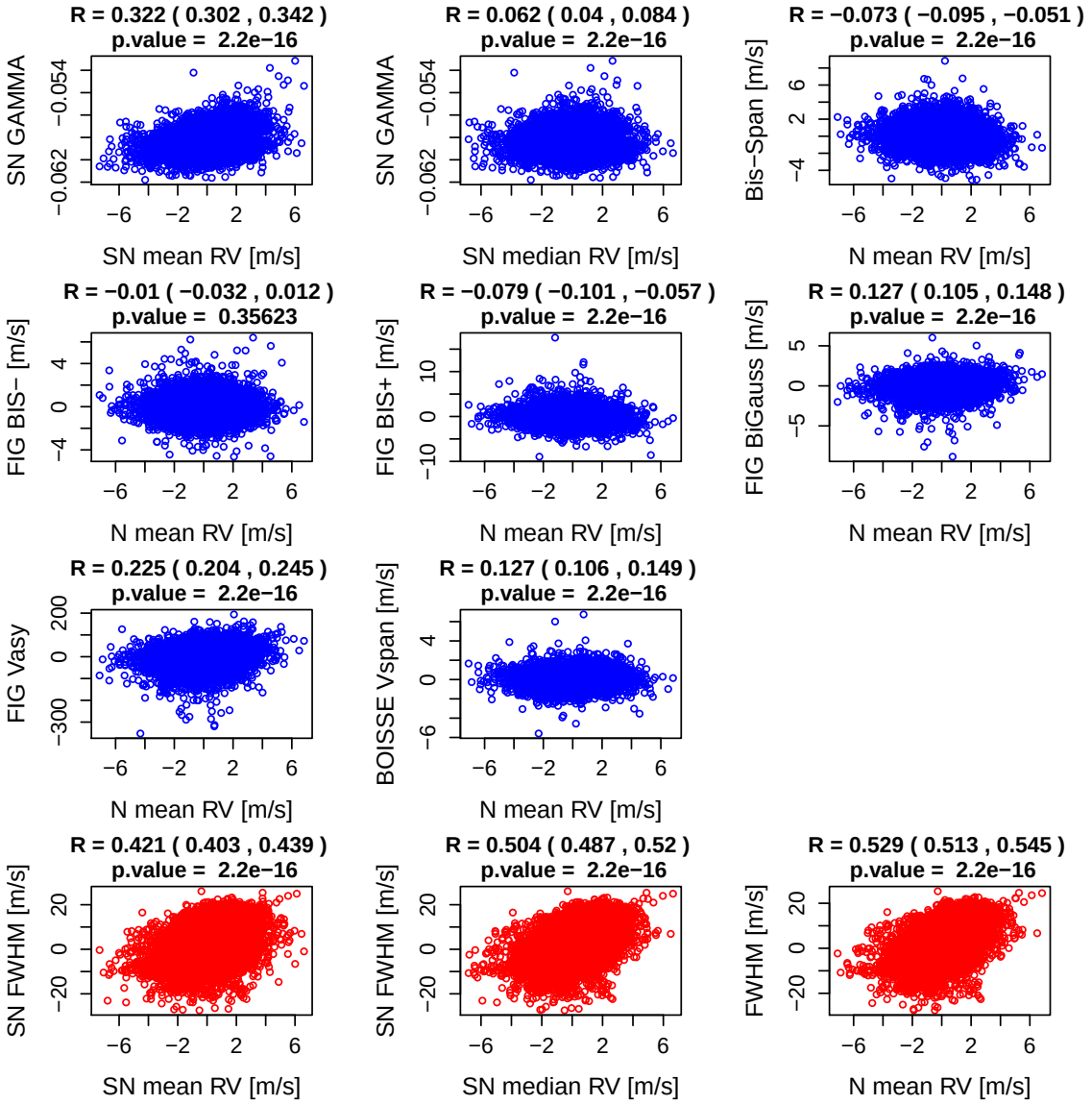

Figure 5.16: Correlation between the asymmetry parameters and the RV's for HD10700. The last three plots show the correlation between the FWHM's and the RV's for HD10700, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's (($R = 0.322$)). The comparison of the correlations between the FWHM's and RV's, when using the Normal and the SN fit leads to comparable considerations. However, in this case, the correlation between N mean RV and FWHM is the strongest ($R = 0.529$).

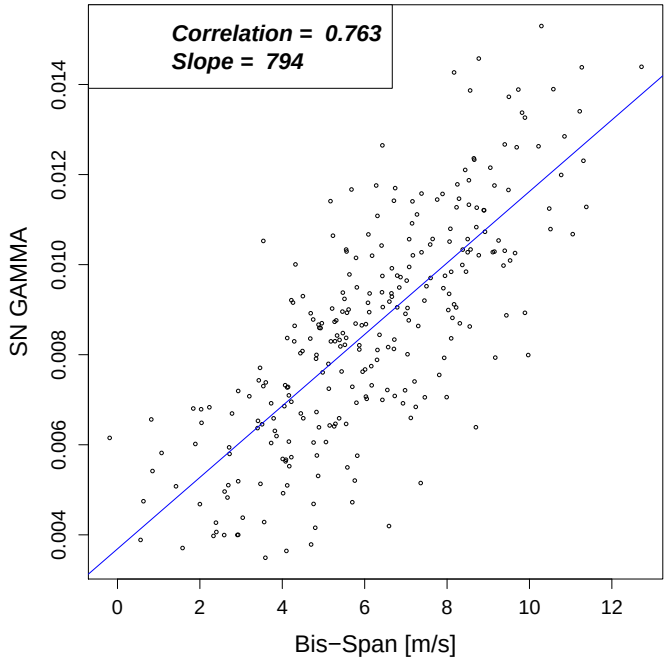

FIGURE 5.17: Correlation between γ and the BIS SPAN for HD215152.

Parameter	N mean RV	SN mean RV	SN median RV
β_0	0.045	0.032	0.025
β_1	0.98	$2.22e-16$	0.0017
β_2	0.046	0.028	0.024
R^2	0.019	0.34	0.0373

TABLE 5.5: **HD215152**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the intercept and the FWHM are statistically significant to explain the RV's variations at level 0.05 but not at level 0.01. The BIS SPAN is not significant, which explains why the R^2 is only 0.019. On the contrary, for the SN case, γ is statistically significant in explaining the variations in RV's caused by stellar activity. The correction for stellar activity is more useful when using SN mean RV ($R^2 = 0.34$).

The comparison between the asymmetry parameters and the RV's is presented in Figure 5.19.The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.571$). The comparison of the correlations between FWHM's and RV's leads to the same conclusion, since the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.21$).

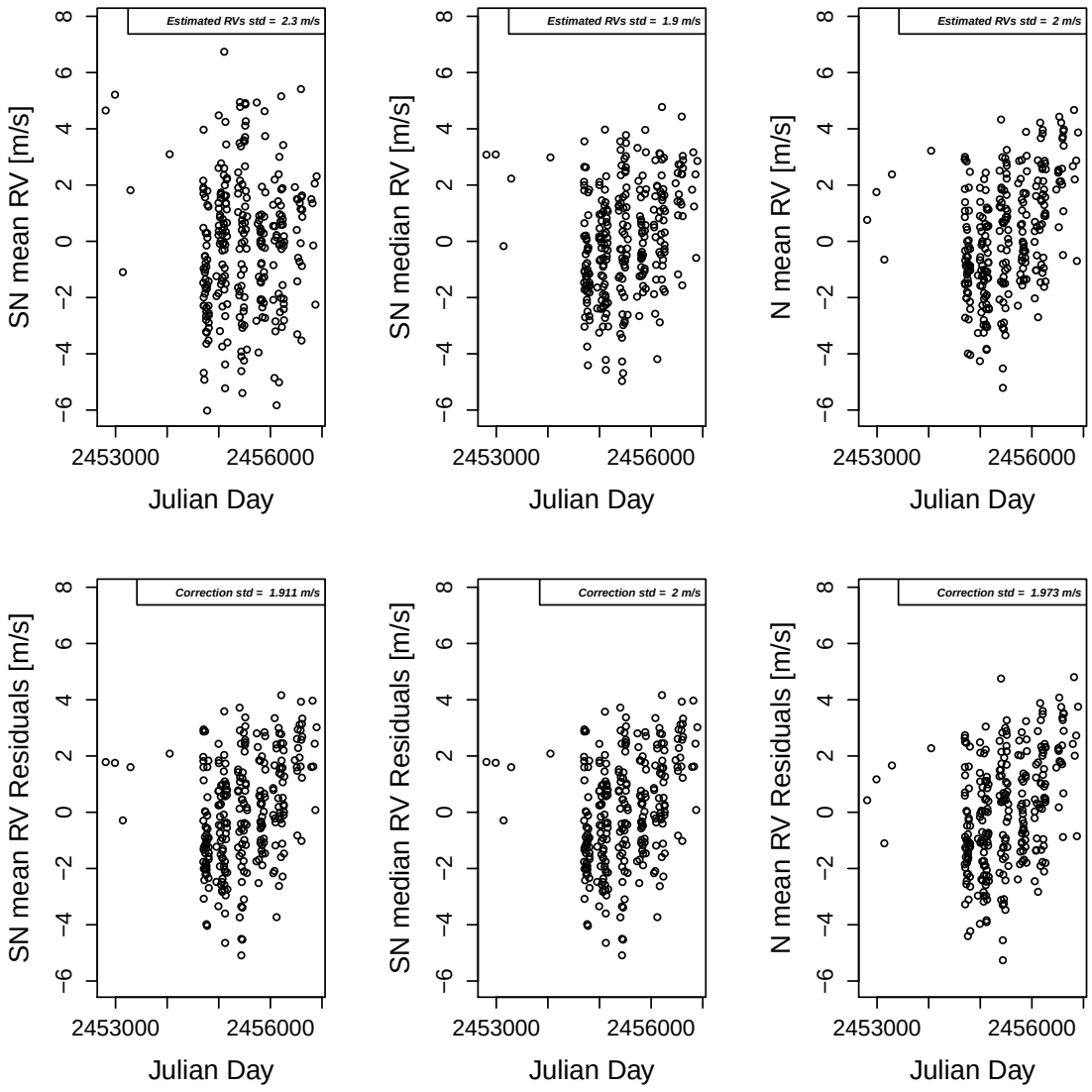

Figure 5.18: (top) Set of RV's for HD215152 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). While the correction from stellar activity leads to similar considerations when SN mean RV or SN median RV are used, using N mean RV leads to residuals 0.062 $\mathrm{m\,s^{-1}}$ higher than the one retrieved with the SN fit.

5.5.5 Corot-7

The final star that has been analyzed is Corot-7, whose CCF's have low SNR (SN50 $<$ 60) . A total of 173 CCF's are analyzed and Figure 5.20 shows the correlation between γ the BIS SPAN, with a linear regression slope of 607 and a Pearson correlation coefficient of $R = 0.814$.

The RV's obtained with the SN density (using SN mean RV or SN median RV) show more variability than the RV's estimated with the Normal density, as shown in the top series of plots in Figure 5.21. Once corrected for stellar activity, using Equation

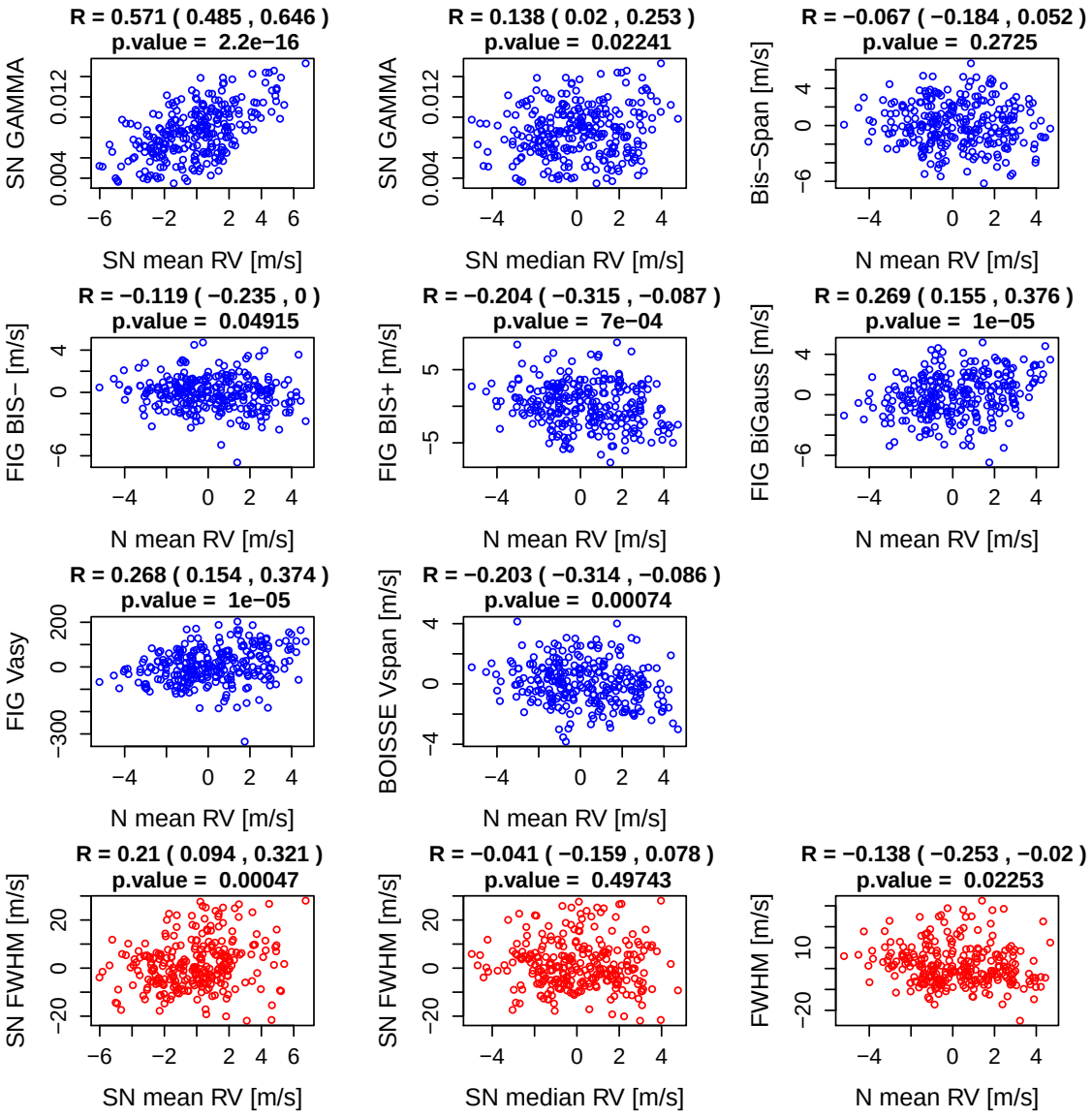

FIGURE 5.19: Correlation between the asymmetry parameters and the RV's for HD215152. The last three plots show the correlation between the FWHM's and the RV's for HD215152, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.571$). Concerning the comparison of the correlations between FWHM's and RV's, the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.21$).

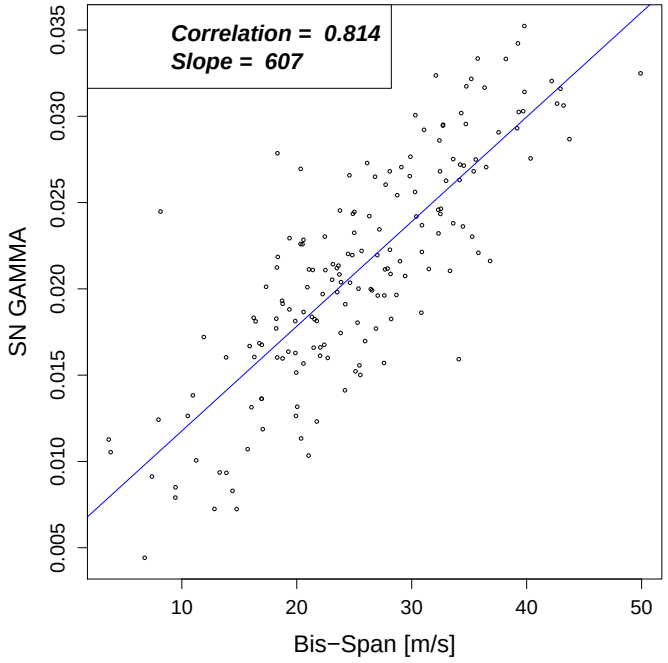

Figure 5.20: Correlation between γ and the BIS SPAN for Corot-7.

(5.9), the RV's residuals obtained with the SN are better than the ones estimated using the Normal, suggesting that the SN fit gains power over the Normal fit as the SNR decreases. In particular using the Normal density to fit the CCF's leads to residuals $0.25\,\mathrm{m\,s^{-1}}$ larger. In Table 5.6 we see that for both the Normal and the SN analysis that uses SN median RV the intercept and the FWHM can explain part of the variations in RV's, but the asymmetry parameter is not statistically helpful. On the other hand, when using SN mean RV, also the asymmetry parameter γ is statistically significant. The latter consideration, combined with the opposite conclusion obtained when SN median RV is used, confirms that the SN median RV is a more robust index and hence a more suitable indicator to define the set of RV's of the star. At the same time, SN mean RV better catches variations in the shape of the CCF because of stellar activity, althoght exactly for this reason the proposed correction for stellar activity is more relevant ($R^2 = 0.56$).

The comparison between the asymmetry parameters and the RV's is presented in Figure 5.22. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.537$). Concerning the comparison of the correlations between FWHM's and RV's, the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.73$). These results suggest that, in particular for low SNR measurements, using the SN to fit the CCF can improve the power in detecting stellar activity signals, which is key to detecting "Earth–like" exoplanets.

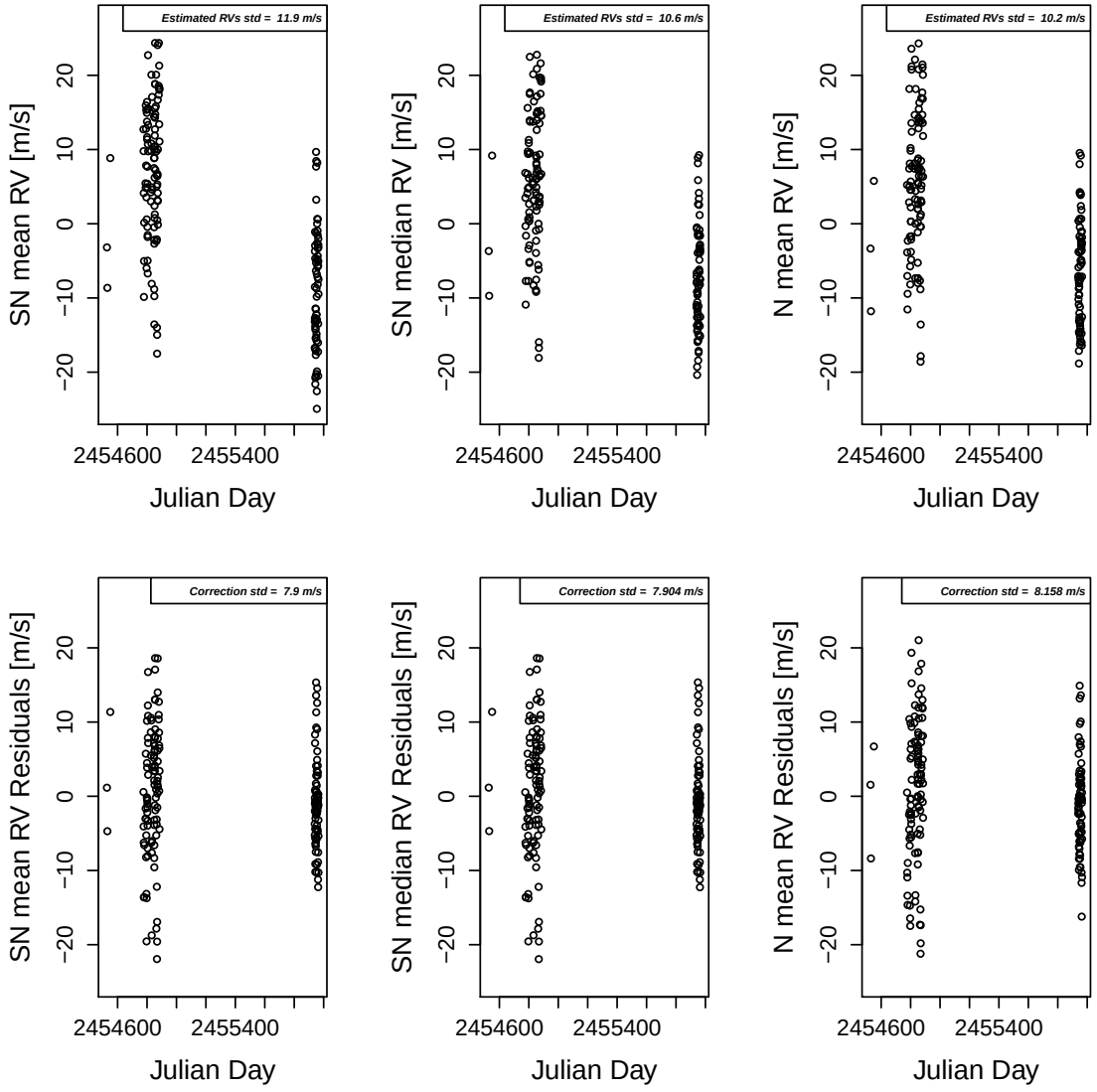

FIGURE 5.21: (top) Set of RV's for Corot-7 estimated using a Normal or a SN fit. (bottom) The residuals from the model fit using Equation (5.9). While the correction from stellar activity leads to similar considerations when SN mean RV or SN median RV are used, using N mean RV leads to residuals $0.25\,\mathrm{m\,s^{-1}}$ higher.

5.6 Estimation of standard errors for the CCF parameters

In this Section, we perform a bootstrap analysis [37, 50] in order to retrieve the standard errors associated to SN mean RV, SN median RV, N mean RV, FWHM, SN FWHM, BIS SPAN and γ. Because a CCF is obtained from a cross-correlation, each point in a CCF is correlated with each other. Therefore, we cannot do a bootstrap analysis on perturbing independently each CCF point with a Gaussian density scaled to the error of each given point. A detailed discussions of the methods nowadays available

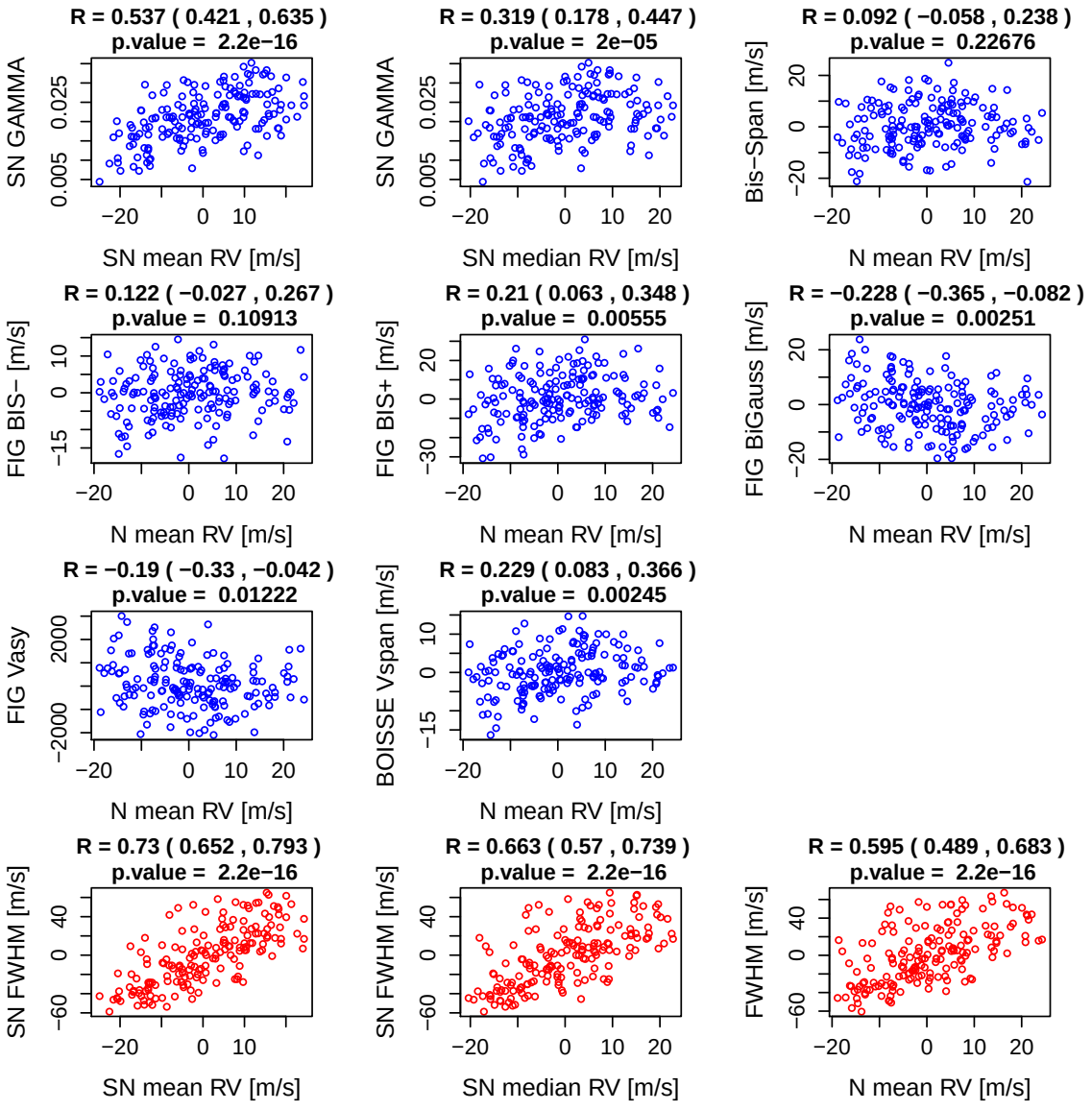

Figure 5.22: Correlation between the asymmetry parameters and the RV's for Corot-7. The last three plots show the correlation between the FWHM's and the RV's for Corot-7, using respectively the SN and the Normal analyses. The correlation between γ and SN mean RV is stronger than the correlation calculated between the other asymmetry statistics and their corresponding RV's ($R = 0.537$). Concerning the comparison of the correlations between FWHM's and RV's, the correlation between SN mean RV and SN FWHM is the strongest ($R = 0.73$).

Parameter	N mean RV	SN mean RV	SN median RV
β_0	$2.22e-16$	$2.22e-16$	$2.22e-16$
β_1	0.12	0.0015	0.36
β_2	$2.22e-16$	$2.22e-16$	$2.22e-16$
R^2	0.36	0.56	0.44

TABLE 5.6: **Corot-7**: Evaluation of the linear combination used for correcting the RV's from stellar activity, according to Equation (5.9). The p–values for the parameters β_0, β_1 and β_2 for all the methodologies are summarized, as well as the R^2. Concerning the Normal fit, the BIS SPAN is not significant, as well as γ if using SN median RV. When using SN mean RV, all the parameters are statistically significant, suggesting again that SN mean RV ($R^2 = 0.56$) is more sensible to stellar activity than SN median RV ($R^2 = 0.44$).

to resampling in situations with dependent data structures is available in [82]. All the bootstrap methods that deal with dependant data structures rely on the so–called Block Bootstrap methods, originally introduced by [151]. In our particular case, since each point in a CCF is correlated with each other, we bootstrap a hundred times the stellar spectrum given the photon-noise error of each wavelength and calculate for each realization a new CCF. We then fit a Normal or a SN to each of these CCF's and then calculate the standard deviations of the density for the location parameters (N mean RV, SN mean RV or SN median RV), the width parameters (FWHM or SN FWHM) and the parameters of asymmetry (BIS SPAN or γ).

In the top plots of Fig. 5.23 we show the different errors for the RV's of the star, defined as N mean RV (red triangles), SN mean RV (black circles) or SN median RV (cyan crosses), FWHM (red triangles), SN FWHM (black circles), BIS SPAN (red triangles) and γ (black circles). The analysis uses information from three real stars, HD215152, HD192310 and Corot-7, whose original CCF's are all at different SNR levels. The parameter SN@550 nm corresponds to the SNR at order 50, which is equal to a wavelength of 550 nm. Looking at the estimates for the RV's, we see that they all follow a similar exponential decay. Although we plot the data for three different stars, we do not see any offsets in this decay, which implies that the parameter SN@550 nm is the main contributor to the precision measured in RV. This is not surprising as the three stars studied here are all main sequence K-dwarfs. In the bottom plots, we show the ratio between the parameters estimated from the bootstrap analysis fitting the SN and the parameters obtained from the bootstrap analysis fitting the Normal density.

When comparing the three different estimates for the RV, we see that SN mean RV presents standard errors that are 60% larger than what N mean RV gives. On the opposite, SN median RV gives errors 10% more precise than N mean RV. Regarding the parameters describing the width of the CCF, FWHM and SN FWHM have the same

standard errors. Finally, for the asymmetry parameters, we see that γ, derived from the SN, is 15% more precise than BIS SPAN.

In closing, using SN median RV leads to uncertainties 10% smaller than usingN mean RV and using γ leads to uncertainties 15% smaller than using BIS SPAN. At the same time the precision on the width parameter of the CCF is preserved. SN mean RV should not be uses to define precise RV's since the precision on this parameter is 60% worse than the precision on the RV's retrieved by using the mean of a Normal density fit to the CCF. We recall moreover that, using the SN density, all the parameters are automatically retrieved in 1 single step, while in the common approach RV and FWHM are calculated separately from the BIS SPAN.

5.7 Concluding Remarks

When searching for small-mass exoplanets using the RV technique, it is crucial to get the best possible precision when retrieving the RV of the star, but also to measure precisely variations in the shape of the CCF, since these variations are induced by stellar activity and not by planets. The correlations between the width of the CCF and the RV and the correlations between the asymmetry of the CCF and the RV are used in order to understand if the estimated RV's are contaminated by stellar activity signals. Therefore, the stronger those correlations are, the better we can probe low level of stellar activity.

In this Chapter we introduced a novel approach based on the SN density to estimate RV's and shape variations in the CCF of stars. The standard approach consists at first to fit a Normal density to the CCF in order to retrieve RV and FWHM. Then, to measure changes in the asymmetry of the CCF, the BIS SPAN or other indicators proposed by [18, 54] are separately retrieved.

We propose to conduct the analysis fitting a SN density to the CCF. Since the CCF presents a natural asymmetry due the convective blueshift, the SN density can in principle better capture spurious variations in RV's caused by stellar activity. On top of that, by using the SN density to fit the CCF, we can retrieve simultaneously the barycenter of the CCF (namely the RV), the width and the asymmetry of the CCF. Using the SN to fit the CCF brings a significant improvement in probing stellar activity. While for the Normal density mean and median are equivalent, using the SN fit different location parameters can be tested. While SN median RV is more robust respect to variations in the shape of the CCF, SN mean RV is more sensible to changes in the asymmetry or width of the CCF.

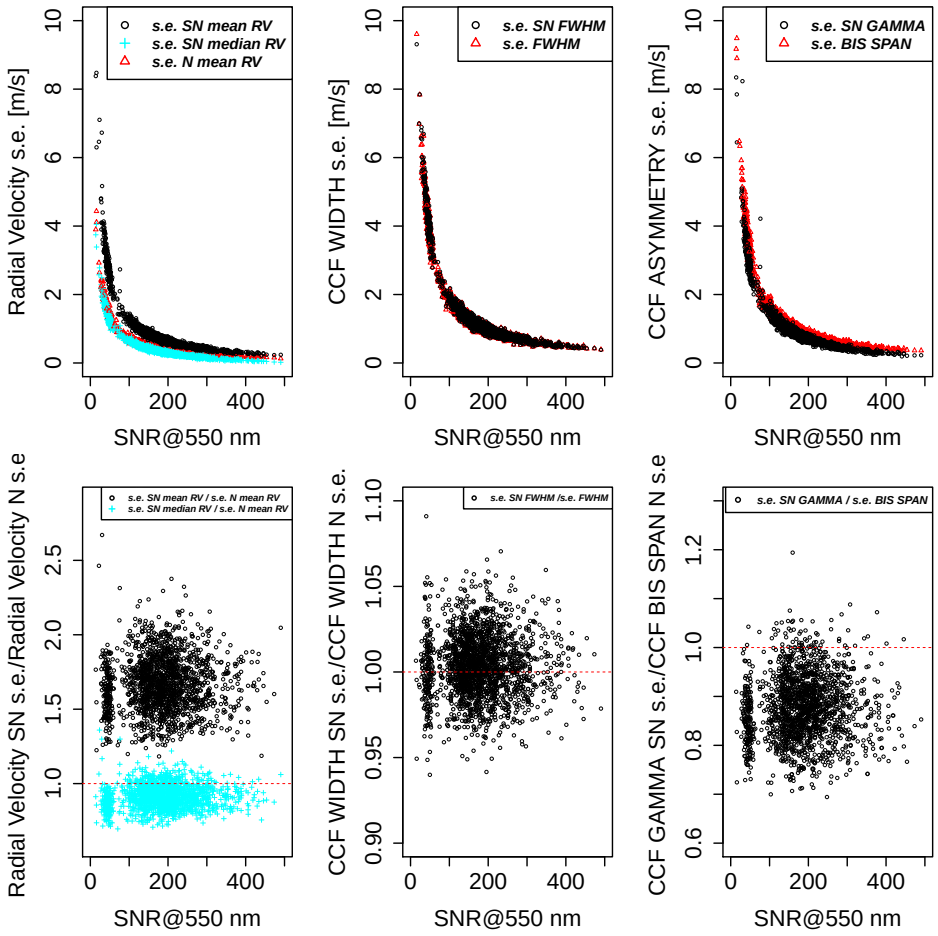

FIGURE 5.23: Comparison between the standard errors using the bootstrap analysis for the RV's, the FWHM and the asymmetry parameters. When using SN mean RV (black circles), the standard errors are in average 60% larger than the standard errors retrieved for RV (red triangles). However, if using SN median RV (cyan crosses), the standard errors are on average 10% smaller than the standard errors related to RV. To use as asymmetry parameter the γ of the SN leads to standard errors on average 15% smaller than the standard errors related to the BIS SPAN. Note that for the asymmetry, the error in BIS SPAN is in $\mathrm{km\,s^{-1}}$. To be able to compare the errors in γ and BIS SPAN we multiplied the error in γ by the slope of the linear fit between γ and BIS SPAN, as shown in Figures 5.11, 5.17 and 5.20.

We suggest to use as parameter that defines the set of RV's of the star SN median RV, since the standard errors related to this parameter are 10% smaller than the standard errors retrieved on N mean RV. In order to evaluate changes in the asymmetry and in the width of the CCF, we suggest to use SN mean RV as location parameter of the CCF. The correlation between SN mean RV and SN FWHM and the correlation between SN mean RV and γ (the asymmetry parameter of the SN) are statistically stronger than the correlations between the equivalent parameters derived using the Normal fit for all the real stars that have been studies. The standard errors related to the asymmetry

parameter γ are on average $\sim$15% smaller than the uncertainties calculated on the BIS SPAN. Therefore, when searching for rotational periods in the data or when applying Gaussian Processes to account for stellar activity signals, the parameters obtained with the SN density should be used.

Appendix

A Analytic expression for the ABC posterior distribution under the assumption of Normal distribution

In the case in which the forward model follows a Normal distribution and the prior is Uniform, the ABC posterior distribution can be retrieved analytically. In order to demonstrate this, let's consider for simplicity a single draw y from $Y \sim N(\theta, \sigma^2)$, with unknown mean θ and known variance σ^2, fixed in this case equal to 1. The results here presented can be easily generalized to the case with n observations. When possible, we will be referring to the same parameterization presented in Chapter 2.

By using Equation (2.1), the true posterior distribution is straightforward to calculate:

$$\pi(\theta \mid y) \sim N(y, 1). \tag{A.1}$$

In order to retrieve the ABC posterior distribution, $\pi_\epsilon(\theta \mid y)$, let's define x as the simulated draw. The distance function used to compare the true observation y with the simulated observation x is the L1 norm: $\rho(x, y) = |x - y|$. A draw $x \sim N(\theta, 1)$ is accepted as an element coming from the true posterior distribution if $|x - y| \leq \epsilon$, with $\epsilon > 0$. Hence, $\pi_\epsilon(\theta \mid y)$ is proportional to:

$$Pr[\theta| \quad |x - y| \leq \epsilon] = Pr[\theta| \quad x - y \leq \epsilon] - Pr[\theta| \quad x - y \leq -\epsilon]. \tag{A.2}$$

In this case Equation (2.2) can be analytically solved. Since y can be seen as a constant, if follows that $x - y \sim N(\theta - y, 1)$. The two probabilities of Equation (A.2) can be respectively wriitten as:

$$Pr[\theta| \quad x - y \leq \epsilon] = Pr\left[\theta| \quad \frac{x - y - \theta + y}{1} \leq \frac{\epsilon - \theta + y}{1}\right] = \Phi\left(\frac{y + \epsilon - \theta}{1}\right) \tag{A.3}$$

and

$$Pr[\theta| \quad x-y \leq -\epsilon] = Pr\left[\theta| \quad \frac{x-y-\theta+y}{1} \leq \frac{-\epsilon-\theta+y}{1}\right] = \Phi\left(\frac{y-\epsilon-\theta}{1}\right), \quad \text{(A.4)}$$

where for both Equations (A.3) and (A.4) we explicitly noted the variance 1 at the denominator, while Φ defines the distribution function of a standard normal distribution.

To completely define the ABC posterior distribution, we need to calculate the normalizing constant of Equation (2.1), defined here as $c(\theta)$. Let's first write down Equation (A.2) as follows:

$$\begin{aligned} Pr[\theta| \quad |x-y| \leq \epsilon] &= \Phi\left(\frac{y+\epsilon-\theta}{1}\right) - \Phi\left(\frac{y-\epsilon-\theta}{1}\right) = \\ &= \frac{1}{\sqrt{(2\pi)}}\left[\int_{-\infty}^{y-\theta+\epsilon} e^{-\frac{t^2}{2}}dt - \int_{-\infty}^{y-\theta-\epsilon} e^{-\frac{t^2}{2}}dt\right] = \\ &= \frac{1}{\sqrt{(2\pi)}}\int_{y-\theta-\epsilon}^{y-\theta+\epsilon} e^{-\frac{t^2}{2}}dt = \\ &= \frac{1}{\sqrt{(2\pi)}}\int_{-\epsilon}^{\epsilon} e^{-\frac{(t'+y-\theta)^2}{2}}dt', \end{aligned} \quad \text{(A.5)}$$

where in the last equivalence we changed variable: $t = t' + y - \theta$, so that for $t = y - \theta + \epsilon$ we get $t' = \epsilon$ and for for $t = y - \theta - \epsilon$ we get $t' = -\epsilon$.

In order to retrieve the normalizing constant $c(\theta)$ we need to integrate Equation (A.5) over θ, that takes values on the entire real line $\mathbb{R}$.

$$\begin{aligned} c(\theta) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}}\int_{-\epsilon}^{\epsilon} e^{-\frac{(t'+y-\theta)^2}{2}}dt'd\theta = \\ &= \int_{-\epsilon}^{-\epsilon}\int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{(t'+y-\theta)^2}{2}}d\theta dt' = \\ &= \int_{-\epsilon}^{-\epsilon} 1dt' = \\ &= 2\epsilon. \end{aligned} \quad \text{(A.6)}$$

Here and in other several parts, we exchanged the order of the integrals using of Fubini's theorem. Moreover we recognized that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{(t'+y-\theta)^2}{2}}d\theta$ is the density of a Normal distribution, hence the result of its integral is 1.

The ABC posterior distribution can be written in compact form as follows:

$$\pi_\epsilon(\theta \mid y) = \frac{\Phi\left(\frac{y+\epsilon-\theta}{1}\right) - \Phi\left(\frac{y-\epsilon-\theta}{1}\right)}{2\epsilon}. \quad \text{(A.7)}$$

Starting from Equation (A.7), we are now interested in retrieving the first two centered moments of $\pi_\epsilon(\theta \mid y)$, and in particular the posterior variance. In fact, the ABC-PMC Algorithm presented in Section 2.3 uses a Gaussian perturbation kernel having variance equal to twice $\text{var}(\pi_\epsilon(\theta \mid y))$.

To retrieve the mean of $\pi_\epsilon(\theta \mid y)$ is straightforward. Following the definition for the mean of a general absolute continuous random variable:

$$\begin{aligned}
\mathbb{E}[\pi_\epsilon(\theta \mid y)] &= \frac{1}{2\epsilon}\int_{-\infty}^{\infty}\theta\frac{1}{\sqrt{(2\pi)}}\int_{-\epsilon}^{\epsilon}e^{-\frac{(t'+y-\theta)^2}{2}}dt'd\theta = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}\int_{-\infty}^{\infty}\frac{1}{\sqrt{(2\pi)}}\theta e^{-\frac{(t'+y-\theta)^2}{2}}d\theta dt' = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}(t'+y)dt' = \\
&= \frac{1}{2\epsilon}\left[\frac{t'^2}{2}+t'y\right]_{-\epsilon}^{\epsilon} = \\
&= \frac{1}{2\epsilon}\left[\frac{\epsilon^2}{2}+\epsilon y-\frac{\epsilon^2}{2}+\epsilon y\right] = \\
&= \frac{1}{2\epsilon}2\epsilon y = \\
&= y.
\end{aligned} \tag{A.8}$$

In order to retrieve the variance of $\pi_\epsilon(\theta \mid y)$, we use the equivalence

$$\text{var}(\pi_\epsilon(\theta \mid y)) = \mathbb{E}[\pi_\epsilon(\theta \mid y)^2] - \mathbb{E}[\pi_\epsilon(\theta \mid y)]^2, \tag{A.9}$$

where clearly, from Equation (A.8), $\mathbb{E}[\pi_\epsilon(\theta \mid y)]^2 = y^2$. We need then to retrieve $\mathbb{E}[\pi_\epsilon(\theta \mid y)^2]$:

$$\begin{aligned}
\mathbb{E}[\pi_\epsilon(\theta \mid y)^2] &= \frac{1}{2\epsilon}\int_{-\infty}^{\infty}\theta^2\frac{1}{\sqrt{(2\pi)}}\int_{-\epsilon}^{\epsilon}e^{-\frac{(t'+y-\theta)^2}{2}}dt'd\theta = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\theta^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta dt'.
\end{aligned} \tag{A.10}$$

As we can see from Equation (A.10), calculating $E[\pi_\epsilon(\theta \mid y)^2]$ is not straightforward, since the inner integral of Equation (A.10) has not an easy solution. We propose here to overcome the direct calculation of this integral by retrieving two auxiliary integrals. For the particular problem we are trying to solve, for which the forward model's variance is equal to 1, we know that, by following the definition of variance for an absolutely

continuous random variable:

$$\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}(\theta-(t'+y))^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta = 1. \tag{A.11}$$

The integral of Equation (A.11) can be written as follows:

$$\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\left[\theta^2-2\theta(t'+y)+(t'+y)^2\right] e^{-\frac{(t'+y-\theta)^2}{2}}d\theta = 1, \tag{A.12}$$

which leads to the three integrals:

$$\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\theta^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta, \tag{A.13}$$

$$\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}-2\theta(t'+y) e^{-\frac{(t'+y-\theta)^2}{2}}d\theta \tag{A.14}$$

and finally

$$\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}(t'+y)^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta. \tag{A.15}$$

By solving the integrals of Equations (A.14) and (A.15) and since the three integrals sum to 1, we will be able to solve Equation (A.13), which actually is the inner integral of Equation (A.10).

Let's start with Equation (A.14):

$$\begin{aligned}\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}-2\theta(t'+y)e^{-\frac{(t'+y-\theta)^2}{2}}d\theta &= -2(t'+y)\int_{-\infty}^{\infty}\theta e^{-\frac{(t'+y-\theta)^2}{2}}d\theta = \\ &= -2(t'+y)(t'+y) = \\ &= -2(t'+y)^2.\end{aligned} \tag{A.16}$$

The solution for the integral of Equation (A.15) is:

$$\begin{aligned}\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}(t'+y)^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta &= \frac{(t'+y)^2}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}e^{-\frac{(t'+y-\theta)^2}{2}}d\theta = \\ &= (t'+y)^2.\end{aligned} \tag{A.17}$$

Therefore, Equation (A.13) has solution:

$$\begin{aligned}\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\theta^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta &= 1+2(t'+y)^2-(t'+y)^2 = \\ &= 1+(t'+y)^2,\end{aligned} \tag{A.18}$$

where the first additive term on the left side of the equivalence is the variance. Starting from Equation (A.10) and knowing the result of its inner integral, we can expand as follows:

$$\begin{aligned}
E[\pi_\epsilon(\theta \mid y)^2] &= \frac{1}{2\epsilon}\int_{-\infty}^{\infty}\theta^2\frac{1}{\sqrt{(2\pi)}}\int_{-\epsilon}^{\epsilon}e^{-\frac{(t'+y-\theta)^2}{2}}dt'd\theta = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\theta^2 e^{-\frac{(t'+y-\theta)^2}{2}}d\theta dt' = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}\left[1+(t'+y)^2\right]dt' = \\
&= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon}\left[1+t'^2+y^2+2t'y\right]dt' = \\
&= \frac{1}{2\epsilon}\left[t'+\frac{t'^3}{3}+t'y^2+\frac{2yt'^2}{2}\right]_{-\epsilon}^{\epsilon} = \\
&= \frac{1}{2\epsilon}\left[\epsilon+\frac{\epsilon^3}{3}+\epsilon y^2+\frac{2y\epsilon^2}{2}+\epsilon+\frac{\epsilon^3}{3}+\epsilon y^2-\frac{2y\epsilon^2}{2}\right] = \\
&= \frac{1}{2\epsilon}\left[2\epsilon+\frac{2\epsilon^3}{3}+2\epsilon y^2\right] = \\
&= \frac{1}{2\epsilon}\left[2\epsilon\left(1+\frac{\epsilon^2}{3}+y^2\right)\right] = \\
&= 1+\frac{\epsilon^2}{3}+y^2.
\end{aligned} \tag{A.19}$$

We can finally retrieve $\text{var}(\pi_\epsilon(\theta \mid y))$ from Equation (A.9):

$$\text{var}(\pi_\epsilon(\theta \mid y)) = 1+\frac{\epsilon^2}{3}+y^2-y^2 = 1+\frac{\epsilon^2}{3}. \tag{A.20}$$

In closing, $E[\pi_\epsilon(\theta \mid y)] = y$ and $\text{var}(\pi_\epsilon(\theta \mid y)) = 1+\frac{\epsilon^2}{3}$. These two equivalences can be easily generalized to the case in which the sample size is $n > 1$ and the forward model's variance σ^2 is known but different from 1. We suggest to use as distance function $|\bar{x}-\bar{y}|$. In this case $E[\pi_\epsilon(\theta \mid \bar{y})] = \bar{y}$ and $\text{var}(\pi_\epsilon(\theta \mid \bar{y})) = \frac{\sigma^2}{n}+\frac{\epsilon^2}{3}$.

The variance of the perturbation kernel used by the ABC-PMC algorithm is twice $\left(\frac{\sigma^2}{n}+\frac{\epsilon^2}{3}\right)$. Using this information, our goal is to determine the best possible choice to reduce the tolerance through the iterations of the algorithm. We note that close forms are also available if the prior distribution for θ follows a Normal distribution having hyperparameters μ_0 and σ_0^2, which is a conjugate prior for the likelihood function.

B Impact of the desired particle sample size N on the Adaptive Approximate Bayesian Computation Tolerance Selection algorithm

In Chapter 3 we proposed a method to adaptively select the sequential tolerances that improves the computational efficiency of the ABC-PMC algorithm. Looking at Equation (3.1) that updates the quantile used to reduce the tolerance, the presence of tail features could lead the ratio of the current posterior estimate to the previous estimate to be large, making the approach too sensitive to those extreme features in the posterior. However, if the posterior distribution presents extreme features on the tails, our task is to take into account of those aspects, in order to properly characterize the posterior distribution. In Section 3.3.4 we presented an example where taking into account of the behavior of the ABC posterior distribution in the tails allows the algorithm for retrieving a suitable approximation of the true posterior distribution.

A possible second concern is the behavior of the presented extensions when the desired particle sample size N is small. In many practical problems that require ABC, either the cost of generating mock data and the number of parameters of interest do not permit a saturation of the parametric space with a high N. As already noted in Section 3.2.1, the desired sample size N has an impact on the evaluation of Equation (3.5) used to automatically arrest the procedure. In particular a too low N leads to more variability of the estimated posterior in Equation (3.4), which could lead to the algorithm stopping prematurely.

We studied the behavior of the Adaptive Approximate Bayesian Computation Tolerance Selection algorithm for the three examples presented in Section 3.3. We performed a simulation study, running the algorithm for 4 different desired particles sample sizes: $N = \{100, 50, 20, 10\}$. For each N, we run the aABC-PMC algorithm 20 times. For the first iteration we explored the parametric space by sampling from the prior distribution an amount of particles equal to $5N$.

The results of the analyses are presented, respectively for the Beta-Binomial model, the Exponential-Gamma model and the Gaussian Mixture Model, in Tables B.1–B.3 and Figures B.1–B.3. In the Tables, we also summarized the results originally presented in Section 3.3, where we used $N = 2000$ to compare the results. The estimates presented in Tables B.1–B.3 are an average of both the final tolerance ϵ_T and the Hellinger distance resulting from the simulation study. We did not plot the ABC posterior distributions for the case $N = 2000$ in Figures B.1–B.3 because, as discussed in Section 3.3, they are

	ϵ_T	H_{dist}
N=2000	0.01	0.032
N=100	0.01	0.11
N=50	0.02	0.17
N=20	0.01	0.22
N=10	0.03	0.23

TABLE B.1: **Beta-Binomial model**: aABC-PMC algorithm performances for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown.

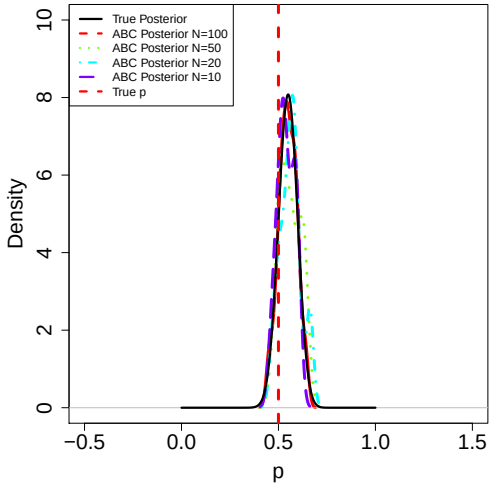

FIGURE B.1: **Beta-Binomial model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$.

comparable with the true posterior distribution.

For all the analyses, we can see that the final tolerances ϵ_T's for which the stopping rule is verified are comparable. We also note that if $N = \{100, 50\}$, the ABC posterior distribution suitably approximates the true one. On the other hand, not surprisingly, the ABC posterior distribution approximates the true posterior when $N = \{20, 10\}$. In particular, focusing on Figure B.3, we can see that $N = \{20, 10\}$ are simply not sufficient to obtain a suitable approximation of the true posterior distribution.

The example presented in Section 3.3.4 has not been discussed, because for that particular model exploring the parametric space is necessary condition to obtain at least few particles coming from the global mode. For this example, if N is small, no particles coming from relevant regions of the parametric space are accepted and a suitable approximation of the true posterior distribution cannot be obtained.

In this Appendix B only examples with 1 parameter have been considered. The stopping rule seems to work properly and the ABC posterior distribution poorly approximates the true posterior because of the extremely low sample size N (e.g. $N = \{20, 10\}$). The evaluation of the behavior of the Adaptive Approximate Bayesian Computation Tolerance Selection algorithm in the multidimensional case is something to explore for future analyses.

	ϵ_T	H_{dist}
N=2000	0.042	0.07
N=100	0.023	0.11
N=50	0.025	0.20
N=20	0.022	0.43
N=10	0.017	0.28

TABLE B.2: **Exponential-Gamma model**: aABC-PMC algorithm for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown.

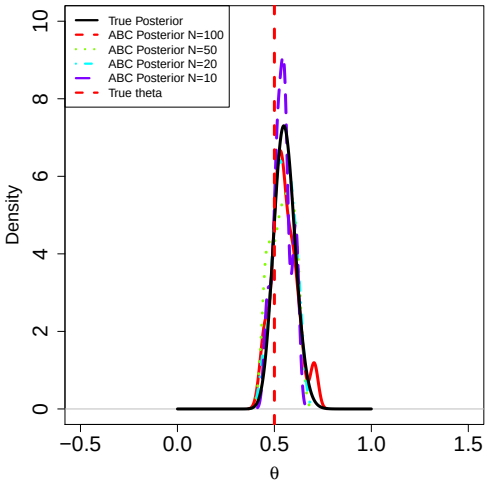

FIGURE B.2: **Exponential-Gamma model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$.

	ϵ_T	H_{dist}
N=2000	0.029	0.54
N=100	0.054	0.57
N=50	0.026	0.63
N=20	0.011	0.88
N=10	0.024	0.94

TABLE B.3: **Gaussian Mixture Model**: aABC-PMC algorithm for different desired particles sample sizes: $N = \{2000, 100, 50, 20, 10\}$. The Hellinger distance between the final aABC-PMC posterior and the true posterior is shown.

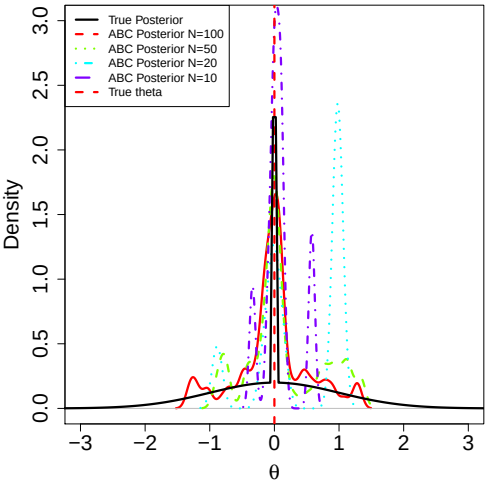

FIGURE B.3: **Gaussian Mixture Model**: aABC-PMC final posterior distribution for different desired particles sample sizes: $N = \{100, 50, 20, 10\}$.

C Resampling the Mixture Weights

In the following we introduce the mathematical definitions required to define Algorithm 4 presented in Section 4.2.2.

When moving the selected values for the mixture weights, not only is there the constraint that each mixture weight component must be in $[0, 1]$, but it is also required that $\sum_{i=1}^{K} f_i = 1$, making the Gaussian kernel inappropriate.

In the first iteration of the proposed ABC-PMC algorithm, the mixture weights $\{f_1^1 \ldots, f_K^1\}$ are directly sampled from the prior distribution, which is a Dirichlet(δ), where $\delta = (\delta_1, \ldots, \delta_K)$. For $t > 1$, proposals are drawn from the previous step particle system according to their importance weights. After randomly selecting a mixture weight, f^{t-1}, we want to "jitter" or move it in manner that preserves some information coming from the selected particle, but not let it be an identical copy, leading to the resampled mixture weight f^t. Overall Algorithm 4 relies on 5 properties, discussed below.

Let's define $k = 2$ independent random variables $X_1 \sim \text{Gamma}(\delta_1, \beta)$ and $X_2 \sim \text{Gamma}(\delta_2, \beta)$, having shape parameter respectively $\delta_1 > 0$ and $\delta_2 > 0$ and the same rate parameter $\beta > 0$.

1. $X_+ := X_1 + X_2 \sim \text{Gamma}(\delta_+, \beta)$, where $\delta_+ := \delta_1 + \delta_2$.

   The moment generating function of a Gamma random variable X_1 is:

$$
\begin{aligned}
M_{X_1}(t; \delta_1, \beta) &= E[e^{tX_1}] \\
&= \int_0^{+\infty} e^{tX_1} \frac{1}{\Gamma(\delta_1)} \beta^{\delta_1} x_1^{\delta_1 - 1} e^{-\beta x_1} dt \\
&= \frac{\beta^{\delta_1}}{\Gamma(\delta_1)} \int_0^{+\infty} x_1^{\delta_1 - 1} e^{-(\beta - t)x_1} dt \\
&= \frac{\beta^{\delta_1}}{\Gamma(\delta_1)} \frac{\Gamma(\delta_1)}{(\beta - t)^{\delta_1}} \\
&= \frac{1}{\left(1 - \frac{t}{\beta}\right)^{\delta_1}}
\end{aligned}
\tag{C.1}
$$

   Since $X_1 \perp\!\!\!\perp X_2$, then

$$
\begin{aligned}
M_{X_+}(t) &= M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \\
&= \frac{1}{\left(1-\frac{t}{\beta}\right)^{\delta_1}} \cdot \frac{1}{\left(1-\frac{t}{\beta}\right)^{\delta_2}} \\
&= \frac{1}{\left(1-\frac{t}{\beta}\right)^{\delta_+}},
\end{aligned}
\tag{C.2}
$$

which is the moment generating function of a Gamma random variable having shape parameter $\delta_+ = \delta_1 + \delta_2$ and rate parameter β.

2. $Y := \left(\frac{X_1}{X_+}, \frac{X_2}{X_+}\right) := \frac{X}{X_+} \sim \text{Dirichlet}(\delta_1, \delta_2)$

   The demonstration consists in the following 4 steps:

   (a) Retrieve the joint distribution of (X_1, X_2):

$$
\begin{aligned}
f_{x_1,x_2}(x_1, x_2; \delta_1, \delta_2, \beta) &:= \prod_{i=1}^{2} \frac{1}{\Gamma(\delta_i)} \beta^{\delta_i} x_i^{\delta_i - 1} e^{-\beta x_i} \\
&:= \frac{1}{\Gamma(\delta_1)\Gamma(\delta_2)} \beta^{\delta_1+\delta_2} x_1^{\delta_1-1} x_2^{\delta_2-1} e^{-\beta(x_1+x_2)}
\end{aligned}
\tag{C.3}
$$

   (b) Set $Y := \frac{X_1}{X_1+X_2}$ and $Z := X_1 + X_2$ and retrieve the joint distribution of $f_{Y,Z}(y, z; \delta_1, \delta_2, \beta)$:

   Given the transformation of the random variables X_1 and X_2, it follows that $X_1 = YZ$ and $X_2 = Z(1-Y)$. The Jacobian is equal to

$$
J(y, z) = \begin{bmatrix} z & y \\ -z & 1-y \end{bmatrix} = z
$$

   (c) The joint distribution $f_{Y,Z}(y, z; \delta_1, \delta_2, \beta)$ is:

$$
\begin{aligned}
f_{Y,Z}(y, z; \delta_1, \delta_2, \beta) &:= z \frac{1}{\Gamma(\delta_1)\Gamma(\delta_2)} (zy)^{\delta_1-1} (z(1-y))^{\delta_2-1} \beta^{\delta_1+\delta_2} e^{-\beta z} \\
&:= \frac{1}{\Gamma(\delta_1)\Gamma(\delta_2)} z^{\delta_1+\delta_2-1} y^{\delta_1-1} (1-y)^{\delta_2-1} \beta^{\delta_1+\delta_2} e^{-\beta z} \\
&:= z^{\delta_1+\delta_2-1} \beta^{\delta_1+\delta_2} e^{-\beta z} \frac{1}{\Gamma(\delta_1)\Gamma(\delta_2)} y^{\delta_1-1} (1-y)^{\delta_2-1} \\
&:= z^{\delta_1+\delta_2-1} \beta^{\delta_1+\delta_2} e^{-\beta z} \frac{1}{\Gamma(\delta_1+\delta_2)} \frac{\Gamma(\delta_1+\delta_2)}{\Gamma(\delta_1)\Gamma(\delta_2)} y^{\delta_1-1} (1-y)^{\delta_2-1}
\end{aligned}
\tag{C.4}
$$

   (d) It is straightforward to see that Z and Y factorize, meaning that $Z \perp\!\!\!\perp Y$. Moreover by just multiplying and dividing for $\Gamma(\delta_1 + \delta_2)$ it follows that:

$$Z \sim \text{Gamma}(\delta_+, \beta), \tag{C.5}$$

as already shown in the previous point.

The random variable $Y \sim \text{Dirichlet}(\delta_1, \delta_2)$:

$$f_Y(y; \delta_1, \delta_2) := \frac{\Gamma(\delta_1 + \delta_2)}{\Gamma(\delta_1)\Gamma(\delta_2)} y^{\delta_1 - 1}(1-y)^{\delta_2 - 1}, \tag{C.6}$$

which in this case, since $k = 2$, corresponds to a $\text{Beta}(\delta_1, \delta_2)$.

3. X_+ and Y are independent.

   This fact comes clearly from Equation (C.4).

4. For any $Z \sim \text{Gamma}(\delta, \beta)$ and independent $B \sim \text{Beta}(p\delta, (1-p)\delta)$ it follows that:

$$BZ \sim \text{Gamma}(p\delta, \beta), \quad (1-B)Z \sim \text{Gamma}((1-p)\delta, \beta) \tag{C.7}$$

   with $BZ \perp\!\!\!\perp (1-B)Z$.

   Again, from Equation (C.4), it is straightforward to show that $BZ \sim \text{Gamma}(p\delta, \beta)$. Just recalling that, given $B \sim \text{Beta}(p\delta, (1-p)\delta)$, $1 - B \sim \text{Beta}((1-p)\delta, p\delta)$ and keep referring to Equation (C.4), it follows that $(1-B)Z \sim \text{Gamma}((1-p)\delta, \beta)$.

5. Let's define $f \sim \textit{Dirichlet}(\delta)$ independent from $Z \sim \text{Gamma}(\delta_+, \beta)$. Then:

$$\{f_i Z\} \overset{\perp\!\!\!\perp}{\sim} \text{Gamma}(\delta_i, \beta) \tag{C.8}$$

   It directly comes from point (b) in the second demonstration introduced here.

From the steps outlined in Algorithm 4, we note that $\xi_i^{t^*}$ is the sum of two independent random variables, with $Z^t f_i^{t-1} B_i^t \sim \text{Gamma}(p\delta_1, 1)$ and $\eta_i^t \sim \text{Gamma}((1-p)\delta_i, 1)$, so that the resampled mixture weight $f^{(t)} \sim \text{Dirichlet}(\delta)$.

The parameter p is a fixed real number with range $[0, 1]$ that determines how much information to retain from f^{t-1}. If p is tiny, then the $\{B_i^t\}$ will be tiny and the new Gamma random variables will be $\xi_i^{t^*} = Z^t f_i^{t-1} B_i^t + \eta_i^t \approx \eta_i^t$, nearly equal to the new $\eta_i^t \overset{\perp\!\!\!\perp}{\sim} \text{Gamma}((1-p)\delta_i, 1)$, so f^t will be almost independent of f^{t-1}. If p is nearly 1 then $\{B_i^t\}$ will be nearly 1, the $\{\eta_i^t\}$ will be nearly 0 and $\xi_i^{t^*} = Z^t f_i^{t-1} B_i^t + \eta_i^t \approx Z^t f_i^{t-1}$, so f^t will be almost identical to f^{t-1}.

Bibliography

[1] Akeret, J., Refregier, A., Amara, A., Seehars, S. and Hasner, C. (2015) Approximate bayesian computation for forward modeling in cosmology. *Journal of Cosmology and Astroparticle Physics* **2015**(08), 043.

[2] Anglada-Escudé, G. and Butler, R. P. (2012) The harps-terra project. i. description of the algorithms, performance, and new measurements on a few remarkable stars observed by harps. *The Astrophysical Journal Supplement Series* **200**(2), 15.

[3] Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scandinavian journal of statistics* pp. 171–178.

[4] Azzalini, A. and Capitanio, A. (2014) The skew-normal and related families. institute of mathematical statistics monographs.

[5] Baranne, A., Queloz, D., Mayor, M., Adrianzyk, G., Knispel, G., Kohler, D., Lacroix, D., Meunier, J.-P., Rimbaud, G. and Vin, A. (1996) Elodie: A spectrograph for accurate radial velocity measurements. *Astronomy and Astrophysics Supplement Series* **119**(2), 373–390.

[6] Barber, S., Voss, J., Webster, M. *et al.* (2015) The rate of convergence for approximate bayesian computation. *Electronic Journal of Statistics* **9**(1), 80–105.

[7] Beaumont, M. A. (2010) Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics 41* **96**, 379 – 406.

[8] Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika* **96**(4), 983 – 990.

[9] Beaumont, M. A., Zhang, W. and Balding, D. J. (2002a) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025 – 2035.

[10] Beaumont, M. A., Zhang, W. and Balding, D. J. (2002b) Approximate bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.

[11] Biau, G., Cérou, F., Guyader, A. *et al.* (2015) New insights into approximate bayesian computation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 51, pp. 376–403.

[12] Blum, M. (2010a) Approximate Bayesian computation: A nonparametric perspective. *Journal of American Statistical Association* **105**(491), 1178 – 1187.

[13] Blum, M., Nunes, M., Prangle, D. and Sisson, S. (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science* **28**(2), 189 – 208.

[14] Blum, M. G. (2010b) Approximate Bayesian computation: A nonparametric perspective. *Journal of American Statistical Association* **105**(491), 1178 – 1187.

[15] Blum, M. G. (2010c) Choosing the summary statistics and the acceptance rate in approximate bayesian computation. *In COMPSTAT 2010: Proceedings in Computational Statistics* p. 47?56.

[16] Blum, M. G. (2010d) Choosing the summary statistics and the acceptance rate in approximate bayesian computation. In *Proceedings of COMPSTAT*, pp. 47–56.

[17] Boisse, I., Bonfils, X. and Santos, N. (2012) Soap-a tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astronomy & Astrophysics* **545**, A109.

[18] Boisse, I., Bouchy, F., Hébrard, G., Bonfils, X., Santos, N. and Vauclair, S. (2011) Disentangling between stellar activity and planetary signals. *Astronomy & Astrophysics* **528**, A4.

[19] Boisse, I., Moutou, C., Vidal-Madjar, A., Bouchy, F., Pont, F., Hébrard, G., Bonfils, X., Croll, B., Delfosse, X., Desort, M. *et al.* (2009) Stellar activity of planetary host star hd 189 733. *Astronomy & Astrophysics* **495**(3), 959–966.

[20] Bonassi, F. V., West, M. *et al.* (2015) Sequential monte carlo with adaptive weights for approximate bayesian computation. *Bayesian Analysis* **10**(1), 171–187.

[21] Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., Caldwell, J., Christensen-Dalsgaard, J., Cochran, W. D., DeVore, E. *et al.* (2010) Kepler planet-detection mission: introduction and first results. *Science* **327**(5968), 977–980.

[22] Bouchy, F., Díaz, R., Hébrard, G., Arnold, L., Boisse, I., Delfosse, X., Perruchot, S. and Santerne, A. (2013) Sophie+: First results of an octagonal-section fiber for high-precision radial velocity measurements. *Astronomy & Astrophysics* **549**, A49.

[23] Bouchy, F., Pepe, F. and Queloz, D. (2001) Fundamental photon noise limit to radial velocity measurements. *Astronomy & Astrophysics* **374**(2), 733–739.

[24] Brahm, R., Jordán, A. and Espinoza, N. (2017) Ceres: A set of automated routines for echelle spectra. *Publications of the Astronomical Society of the Pacific* **129**(973), 034002.

[25] Bruderer, C., Chang, C., Refregier, A., Amara, A., Bergé, J. and Gamper, L. (2016) Calibrated ultra fast image simulations for the dark energy survey. *The Astrophysical Journal* **817**(1), 25.

[26] Cameron, E. and Pettitt, A. N. (2012) Approximate bayesian computation for astronomical model analysis: A case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society* **425**, 44–65.

[27] Cappé, O., Guillin, A., Marin, J.-M. and Robert, C. P. (2004) Population monte carlo. *Journal of Computational and Graphical Statistics* **13**(4), 907–929.

[28] Casella, G., Mengersen, K. L., Robert, C. P. and Titterington, D. M. (2002) Perfect samplers for mixtures of distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 777–790.

[29] Casella, G., Robert, C. P. and Wells, M. T. (2004) Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology* **1**(1), 1–18.

[30] Cavallini, F., Ceppatelli, G. and Righini, A. (1985) Asymmetry and shift of three fe i photospheric lines in solar active regions. *Astronomy and Astrophysics* **143**, 116–121.

[31] Cisewski, J., Weller, G. B., Schafer, C. M. and Hogg, D. W. (2014) Approximate bayesian computation for the stellar initial mass function. Preprint available.

[32] Cornuet, J., Santos, F., Beaumont, M., Robert, C., Marin, J., Balding, D., Guillemaud, T. and Estoup, A. (2008a) Inferring population history with diy abc: a user-friendly approach to approximate bayesian computation. *Bioinformatics* .

[33] Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T. and Estoup, A. (2008b) Inferring population history with diy abc: a user-friendly approach to approximate bayesian computation. *Bioinformatics* **24**(23), 2713–2719.

[34] Cosentino, R., Lovis, C., Pepe, F., Cameron, A. C., Latham, D. W., Molinari, E., Udry, S., Bezawada, N., Black, M., Born, A. *et al.* (2012) Harps-n: the new planet hunter at tng. In *Proc. SPIE*, volume 8446, p. 84461V.

[35] Cox, D. R. and Hinkley, D. V. (1979) *Theoretical statistics*. CRC Press.

[36] Csilléry, K., Blum, M. G., Gaggiotti, O. E. and François, O. (2010) Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution* **25**(7), 410 – 418.

[37] Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap methods and their application*. Volume 1. Cambridge university press.

[38] Del Moral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing* **22**(5), 1009–1020.

[39] Dempster, A. P., Laird, N. and Rubin, D. (1977) Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B* **1**(39), 1 – 38.

[40] Desort, M., Lagrange, A.-M., Galland, F., Udry, S. and Mayor, M. (2007) Search for exoplanets with the radial-velocity technique: quantitative diagnostics of stellar activity. *Astronomy & Astrophysics* **473**(3), 983–993.

[41] Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 363–375.

[42] Diggle, P. J. and Gratton, R. J. (1984) Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 193–227.

[43] Dravins, D., Lindegren, L. and Nordlund, Å. (1981) Solar granulation-influence of convection on spectral line asymmetries and wavelength shifts. *Astronomy and Astrophysics* **96**, 345–364.

[44] Drovandi, C. C. and Pettitt, A. N. (2011) Estimation of parameters for macroparasite population evolution using approximate bayesian computation. biometrics. *Statistics and Computing* **67**(1), 225–233.

[45] Dumusque, X. (2016) Radial velocity fitting challenge-i. simulating the data set including realistic stellar radial-velocity signals. *Astronomy & Astrophysics* **593**, A5.

[46] Dumusque, X., Boisse, I. and Santos, N. (2014) Soap 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. *The Astrophysical Journal* **796**(2), 132.

[47] Dumusque, X., Borsa, F., Damasso, M., Diaz, R. F., Gregory, P., Hara, N., Hatzes, A., Rajpaul, V., Tuomi, M., Aigrain, S. *et al.* (2017) Radial-velocity fitting challenge-ii. first results of the analysis of the data set. *Astronomy & Astrophysics* **598**, A133.

[48] Dumusque, X., Pepe, F., Lovis, C., Ségransan, D., Sahlmann, J., Benz, W., Bouchy, F., Mayor, M., Queloz, D., Santos, N. *et al.* (2012) An earth-mass planet orbiting [agr] centauri b. *Nature* **491**(7423), 207–211.

[49] Dumusque, X., Udry, S., Lovis, C., Santos, N. C. and Monteiro, M. (2011) Planetary detection limits taking into account stellar noise-i. observational strategies to reduce stellar oscillation and granulation effects. *Astronomy & Astrophysics* **525**, A140.

[50] Efron, B. and Tibshirani, R. J. (1994) *An introduction to the bootstrap*. CRC press.

[51] Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B* **74**(3), 419–474.

[52] Feng, F., Tuomi, M. and Jones, H. R. (2017) Evidence for at least three planet candidates orbiting hd 20794. *Astronomy & Astrophysics* **605**, A103.

[53] Fernandes, E., Pacheco, A. and Penha-Gonçalves, C. (2007) Mapping of quantitative trait loci using the skew-normal distribution. *Journal of Zhejiang University-Science B* **8**(11), 792–801.

[54] Figueira, P., Santos, N., Pepe, F., Lovis, C. and Nardetto, N. (2013) Line-profile variations in radial-velocity measurements-two alternative indicators for planetary searches. *Astronomy & Astrophysics* **557**, A93.

[55] Filippi, S., Barnes, C. P., Cornebise, J. and Stumpf, M. P. (2013) On optimality of kernels for approximate bayesian computation using sequential monte carlo. *Statistical applications in genetics and molecular biology* **12**(1), 87–107.

[56] Fischer, D. A., Anglada-Escude, G., Arriagada, P., Baluev, R. V., Bean, J. L., Bouchy, F., Buchhave, L. A., Carroll, T., Chakraborty, A., Crepp, J. R. *et al.* (2016) State of the field: extreme precision radial velocities. *Publications of the Astronomical Society of the Pacific* **128**(964), 066001.

[57] Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer Science & Business Media.

[58] Gelman, A., Carln, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2014) *Bayesian Data Analysis*. Chapman & Hall.

[59] Geman, S. and Geman, D. (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

[60] Gianola, D., Heringstad, B. and Odegaard, J. (2006) On the quantitative genetics of mixture characters. *Genetics* **173**(4), 2247–2255.

[61] Gray, D. F. (2009) The third signature of stellar granulation. *The Astrophysical Journal* **697**(2), 1032.

[62] Gutmann, M. U., Corander, J. *et al.* (2016) Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research* .

[63] Hadfield, J. D. *et al.* (2010) Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software* **33**(2), 1–22.

[64] Hahn, C., Vakili, M., Walsh, K., Hearin, A. P., Hogg, D. W. and Campbell, D. (2017) Approximate bayesian computation in large-scale structure: constraining the galaxy–halo connection. *Monthly Notices of the Royal Astronomical Society* **469**(3), 2791–2805.

[65] Hastings, W. K. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**(1), 97–109.

[66] Hatzes, A. P. (1996) Simulations of stellar radial velocity and spectral line bisector variations: I. nonradial pulsations. *Publications of the Astronomical Society of the Pacific* **108**(728), 839.

[67] Hatzes, A. P. (2002) Starspots and exoplanets. *Astronomische Nachrichten* **323**(3-4), 392–394.

[68] Haywood, R., Collier Cameron, A., Queloz, D., Barros, S., Deleuil, M., Fares, R., Gillon, M., Lanza, A., Lovis, C., Moutou, C. *et al.* (2014) Planets and stellar activity: hide and seek in the corot-7 system? *Monthly notices of the royal astronomical society* **443**(3), 2517–2531.

[69] Henderson, R. and Shimakura, S. (2003) A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**(2), 355–366.

[70] Ishida, E., Vitenti, S., Penna-Lima, M., Cisewski, J., de Souza, R., Trindade, A., Cameron, E. *et al.* (2015) cosmoabc: Likelihood-free inference via population monte carlo approximate bayesian computation. *Astronomy & Computing* **13**, 1–11.

[71] Jasra, A., Stephens, D. and Holmes, C. (2007) On population-based simulation for static inference. *Statistics and Computing* **17**(3), 263–279.

[72] Jennings, E. and Madigan, M. (2016) astroabc : An approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation. *Astronomy and Computing* .

[73] Jennings, E. and Madigan, M. (2017) astroabc: An approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation. *Astronomy and Computing* **19**, 16–22.

[74] Jennings, E., Wolf, R. and Sako, M. (2016a) A new approach for obtaining cosmological constraints from type ia supernovae using approximate bayesian computation. *Astronomy and Computing* .

[75] Jennings, E., Wolf, R. and Sako, M. (2016b) A new approach for obtaining cosmological constraints from type ia supernovae using approximate bayesian computation. *arXiv preprint arXiv:1611.03087* .

[76] Joyce, P. and Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7**(1), 1 – 16.

[77] Julier, S., Uhlmann, J. and Durrant-Whyte, H. F. (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on automatic control* **45**(3), 477–482.

[78] Kaplan, D. L., Swiggum, J. K., Fichtenbauer, T. D. and Vallisneri, M. (2018) A gaussian mixture model for nulling pulsars. *The Astrophysical Journal* **855**(1), 14.

[79] Koutroumpas, K., Ballarini, P., Votsi, I. and Cournède, P.-H. (2016) Bayesian parameter estimation for the wnt pathway: an infinite mixture models approach. *Bioinformatics* **32**(17), i781–i789.

[80] Kurster, M., Endl, M., Rouesnel, F., Els, S., Kaufer, A., Brillant, S., Hatzes, A., Saar, S. and Cochran, W. (2003) The low-level radial velocity variability in barnard's star (= gj 699). secular acceleration, indications for convective redshift, and planet mass limits. *ASTRONOMY AND ASTROPHYSICS-BERLIN-* **403**(3), 1077–1088.

[81] Lagrange, A.-M., Desort, M. and Meunier, N. (2010) Using the sun to estimate earth-like planets detection capabilities-i. impact of cold spots. *Astronomy & Astrophysics* **512**, A38.

[82] Lahiri, S. N. (2013) *Resampling methods for dependent data*. Springer Science & Business Media.

[83] Lancaster, T. and Imbens, G. (1996) Case-control studies with contaminated controls. *Journal of Econometrics* **71**(1), 145–160.

[84] Lau, J. W. and Green, P. J. (2007) Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16**(3), 526–558.

[85] Lenormand, M., Jabot, F. and Deuant, G. (2013) Adaptive approximate bayesian computation for complex models. *Computational Statistics* **6**(28), 2777–2796.

[86] Lin, C.-Y., Lo, Y. and Ye, K. Q. (2012) Genotype copy number variations using gaussian mixture models: Theory and algorithms. *Statistical applications in genetics and molecular biology* **11**(5).

[87] Lindegren, L. and Dravins, D. (2003) The fundamental definition of ?radial velocity? *Astronomy & Astrophysics* **401**(3), 1185–1201.

[88] Lotka, A. (1925) *Elements of physical biology*. Volume 1. Baltimore, MD: Williams & Wilkins Co.

[89] Marin, J.-M., Mengersen, K. and Robert, C. P. (2005) Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics* **25**, 459–507.

[90] Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statistics and Computing* **22**(6), 1167 – 1180.

[91] Mayor, M., Marmier, M., Lovis, C., Udry, S., Ségransan, D., Pepe, F., Benz, W., Bertaux, J.-L., Bouchy, F., Dumusque, X. *et al.* (2011) The harps search for southern extra-solar planets xxxiv. occurrence, mass distribution and orbital properties of super-earths and neptune-mass planets. *arXiv preprint arXiv:1109.2497* .

[92] McKinley, T., Cook, A. and Deardon, R. (2009) Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* **171**(5).

[93] McLachlan, G. and Peel, D. (2004) *Finite mixture models*. John Wiley & Sons.

[94] Mena, R. H. and Walker, S. G. (2015) On the bayesian mixture model and identifiability. *Journal of Computational and Graphical Statistics* **24**(4), 1155–1169.

[95] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**(6), 1087–1092.

[96] Meunier, N., Desort, M. and Lagrange, A.-M. (2010) Using the sun to estimate earth-like planets detection capabilities-ii. impact of plages. *Astronomy & Astrophysics* **512**, A39.

[97] Miyahara, H., Tsumura, K. and Sughiyama, Y. (2016) Relaxation of the em algorithm via quantum annealing for gaussian mixture models. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 4674–4679.

[98] Moral, P. D., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* **22**(5), 1009–1020.

[99] Naim, I. and Gildea, D. (2012) Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427* .

[100] Papastamoulis, P. (2015) label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271* .

[101] Pearson, K. (1894a) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110.

[102] Pearson, K. (1894b) Mathematical contributions to the theory of evolution. ii. skew variation in homogeneous material. *Proceedings of the Royal Society of London* **57**(340-346), 257–260.

[103] Pepe, F., Mayor, M., Galland, F., Naef, D., Queloz, D., Santos, N., Udry, S. and Burnet, M. (2002) The coralie survey for southern extra-solar planets vii-two short-period saturnian companions to hd 108147 and hd 168746. *Astronomy & Astrophysics* **388**(2), 632–638.

[104] Pepe, F., Molaro, P., Cristiani, S., Rebolo, R., Santos, N., Dekker, H., Mégevand, D., Zerbi, F., Cabral, A., Di Marcantonio, P. *et al.* (2014) Espresso: The next european exoplanet hunter. *Astronomische Nachrichten* **335**(1), 8–20.

[105] Perryman, M. (2011) *The exoplanet handbook*. Cambridge University Press.

[106] PHASE, C. (2003) Setting new standards with harps. *The Messenger* **114**, 20.

[107] Postman, M., Huchra, J. and Geller, M. (1986) Probes of large-scale structure in the corona borealis region. *The Astronomical Journal* **92**, 1238–1247.

[108] Prangle, D. (2015a) Summary statistics in approximate bayesian computation. *arXiv preprint arXiv:1512.05633* .

[109] Prangle, D. (2015b) Summary statistics in approximate bayesian computation. *arXiv preprint arXiv:1512.05633* .

[110] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution* **16**(12), 1791–1798.

[111] Queloz, D., Bouchy, F., Moutou, C., Hatzes, A., Hébrard, G., Alonso, R., Auvergne, M., Baglin, A., Barbieri, M., Barge, P. *et al.* (2009) The corot-7 planetary system: two orbiting super-earths. *Astronomy & Astrophysics* **506**(1), 303–319.

[112] Queloz, D., Henry, G., Sivan, J., Baliunas, S., Beuzit, J., Donahue, R., Mayor, M., Naef, D., Perrier, C. and Udry, S. (2001) No planet for hd 166435. *Astronomy & Astrophysics* **379**(1), 279–287.

[113] Queloz, D., Mayor, M., Weber, L., Blécha, A., Burnet, M., Confino, B., Naef, D., Pepe, F., Santos, N. and Udry, S. (2000) The coralie survey for southern extra-solar planets. i. a planet orbiting the star gliese 86. *Astronomy and Astrophysics* **354**, 99–102.

[114] Quirrenbach, A., Amado, P., Caballero, J., Mundt, R., Reiners, A., Ribas, I., Seifert, W., Abril, M., Aceituno, J., Alonso-Floriano, F. *et al.* (2014) Carmenes instrument overview. In *Proc. SPIE*, volume 9147, p. 91471F.

[115] Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S. and Roberts, S. (2015) A gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society* **452**(3), 2269–2291.

[116] Ratmann, O., Camacho, A., Meijer, A. and Donker, G. (2013) Statistical modelling of summary values leads to accurate approximate Bayesian computations. Unpublished.

[117] Richardson, S. and Green, P. J. (1997) On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)* **59**(4), 731–792.

[118] Richardson, S., Leblond, L., Jaussent, I. and Green, P. J. (2002) Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **165**(3), 549–566.

[119] Robert, C. (2007) *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

[120] Robertson, P., Mahadevan, S., Endl, M. and Roy, A. (2014) Stellar activity masquerading as planets in the habitable zone of the m dwarf gliese 581. *Science* p. 1253253.

[121] Rodríguez, C. E. and Walker, S. G. (2014) Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics* **23**(1), 25–45.

[122] Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association* **85**(411), 617–624.

[123] Roeder, K. and Wasserman, L. (1997) Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**(439), 894–902.

[124] Rousseau, J. and Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 689–710.

[125] Rubin, D. B. *et al.* (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**(4), 1151–1172.

[126] Saar, S. H. and Donahue, R. A. (1997) Activity-related radial velocity variation in cool stars. *The Astrophysical Journal* **485**(1), 319.

[127] Santerne, A., Díaz, R., Almenara, J.-M., Bouchy, F., Deleuil, M., Figueira, P., Hébrard, G., Moutou, C., Rodionov, S. and Santos, N. (2015) pastis: Bayesian extrasolar planet validation–ii. constraining exoplanet blend scenarios using spectroscopic diagnoses. *Monthly Notices of the Royal Astronomical Society* **451**(3), 2337–2351.

[128] Schafer, C. M. and Freeman, P. E. (2012) *Statistical Challenges in Modern Astronomy V*, chapter 1, pp. 3 – 19. Lecture Notes in Statistics. Springer.

[129] Shabram, M., Demory, B.-O., Cisewski, J., Ford, E. B. and Rogers, L. (2016) The eccentricity distribution of short-period planet candidates detected by kepler in occultation. *The Astrophysical Journal* **820**(2), 93.

[130] Silk, D., Filippi, S. and Stumpf, M. (2013) Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology* **5**(12), 603–618.

[131] Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Science* **104**(6), 1760 – 1765.

[132] Sperrin, M., Jaki, T. and Wit, E. (2010) Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing* **20**(3), 357–366.

[133] Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 795–809.

[134] Stoneking, C. J. (2014) Bayesian inference of gaussian mixture models with non-informative priors. *arXiv preprint arXiv:1405.4895* .

[135] Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from dna sequence data. *Genetics* **145**(2), 505–518.

[136] Thompson, A., Watson, C., de Mooij, E. and Jess, D. (2017) The changing face of α centauri b: probing plage and stellar activity in k dwarfs. *Monthly Notices of the Royal Astronomical Society: Letters* **468**(1), L16–L20.

[137] Thornton, K. and Andolfatto, P. (2006) Inference in epidemic models without likelihoods. *Genetics* **172**, 1607 – 1619.

[138] Toner, C. and Gray, D. F. (1988) The starpatch on the g8 dwarf xi bootis a. *The Astrophysical Journal* **334**, 1008–1020.

[139] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. (2009a) Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**(31), 187–202.

[140] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. H. (2009b) Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society, Interface / the Royal Society* **6**(31), 187–202.

[141] Tull, R. G. (1998) High-resolution fiber-coupled spectrograph of the hobby-eberly telescope. *Proc. Soc. Photo-opt. Inst. Eng.* **3355**, 387.

[142] Tull, R. G., MacQueen, P. J., Sneden, C. and Lambert, D. L. (1995) The high-resolution cross-dispersed echelle white pupil spectrometer of the mcdonald observatory 2.7-m telescope. *Publications of the Astronomical Society of the Pacific* **107**(709), 251.

[143] Udry, S., Fischer, D. and Queloz, D. (2007) A decade of radial-velocity discoveries in the exoplanet domain. *Protostars and Planets V* **951**, 685–699.

[144] Vogt, S. (1987) *PASP* (99), 1214.

[145] Vogt, S. S., Allen, S. L. and Bigelow, B. C. (1994) *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **2198**(362).

[146] Voigt, H.-H. (1956) " drei-strom-modell" der sonnenphotosphäre und asymmetrie der linien des infraroten sauerstoff-tripletts. mit 12 textabbildungen. *Zeitschrift fur Astrophysik* **40**, 157.

[147] Volterra, V. (1927) *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari.

[148] Wang, L. and Dunson, D. B. (2011) Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **20**(1), 196–216.

[149] Wasserman, L. (2000) Asymptotic inference for mixture models using data dependent priors. *Journal of the Royal Statistical Society* **12**(62), 159–180.

[150] Weyant, A., Schafer, C. and Wood-Vasey, W. M. (2013) Likelihood-free cosmological inference with type Ia supernovae: approximate Bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal* **764**, 116.

[151] Wilks, D. (1997) Resampling hypothesis tests for autocorrelated fields. *Journal of Climate* **10**(1), 65–82.

[152] Wright, S. J. and Nocedal, J. (1999) Numerical optimization. *Springer Science* **35**(67-68), 7.

[153] Wu, G., Holan, S. H., Nilon, C. H., Wikle, C. K. *et al.* (2015) Bayesian binomial mixture models for estimating abundance in ecological monitoring studies. *The Annals of Applied Statistics* **9**(1), 1–26.

Umberto Simola

CURRICULUM VITAE

Contact Information

University of Padova,
Department of Statistical Sciences,
via Cesare Battisti, 241-243,
35121 Padova. Italy.
Phone +39 049 827 4141
e-mail: simola@stat.unipd.it

Current Position

Since November 2014; (expected completion: November 2018)
PhD Candidate in Statistical Sciences
Department of Statistical Sciences, University of Padova.
Thesis title: Developments in Approximate Bayesian Computation and Applications in Astrostatistics
Supervisor: Prof. Alessandra R. Brazzale
Co-supervisor: Prof. Jessi Cisewski-Kehe.

Research interests

- Approximate Bayesian Computation
- Computational Statistics
- Likelihood inference
- Measurement errors
- Statistical applications in Astronomy
- Statistical applications in Sport

Education

October 2012 – July 2014
Master degree (*laurea magistrale***) in Statistical Sciences**.
University of Padova, Department of Statistical Sciences.
Title of dissertation: "THE SEARCH FOR EXOPLANETS: CASE STUDIES IN ASTROSTATISTICS."
Supervisor: Prof. Alessandra R. Brazzale
Final mark: 109/110.

October 2009 – July 2012
Bachelor degree (*laurea triennale***) in Statistics and Informatics**.
University of Padova, Department of Statistical Sciences.
Title of dissertation: "A STUDY ON GENETIC DIFFERENTIATION BETWEEN DIFFUSE LARGE B-CELL LYMPHOMA AND FOLLICOLAR LYMPHOMA."
Supervisor: Prof. Laura Ventura
Final mark: 92/110.

Visiting periods

February 2016 – August 2017
Department of Statistics and Data Science, Yale University, 06510, New Haven, CT, USA
Supervisor: Prof. Jessi Cisewski-Kehe

Computer skills

- Programming languages: R (advanced), SAS (intermediate), JAVA (intermediate), C++ (intermediate), Python (intermediate), Julia (basic).
- Scripting languages: HTML (intermediate), PHP (basic).
- Databases: SQL (intermediate).
- OS environments: Mac OS X (advanced), Windows (advanced), Linux (advanced).
- Packages: LaTeX(advanced), MS Office (advanced), OpenOffice (advanced).

Language skills

Italian (● ● ● ● ●); English (● ● ● ● ●); French (● ● ○ ○ ○); German (● ● ○ ○ ○), Spanish (● ● ○ ○ ○); Finnish (● ● ○ ○ ○).

Publications

Articles in proceedings
U. Simola and J. Cisewski-Kehe., (2018) "Adaptive Approximate Bayesian Computation Tolerance Selection". *under review*

U. Simola, J. Cisewski-Kehe and R. L. Wolpert., (2018) "Approximate Bayesian Computation for Finite Mixture Models". *under review*

U. Simola, X. Dumusque and J. Cisewski-Kehe., (2018) "Measuring precise radial velocities and cross-correlation function line-profile variations using a Skew Normal distribution". *under review*

U. Simola, J. Cisewski-Kehe, M. Shabram and E. Ford. "Inferring the eccentricity distribution of exoplanet populations via Approximate Bayesian Computation". *in preparation*

Conference presentations

U. Simola. (2017) "Modeling Stellar Activity CCF with a Skew Normal distribution". (poster) *ASTRO@STAT2017. Department of Statistical Sciences, University of Padua, Padua, Italy, September* 20^{th} 2017

U. Simola, X. Dumusque and J. Cisewski-Kehe., (2017) "Cross correlation function line-profile variations in radial-velocity measurements by fitting a Skew Normal distribution". (invited) The Yale Summer Program in Astrophysics, Leitner Family Observatory and Planetarium, Yale University, New Haven, CT, July 14^{th} *2017*

U. Simola, A.R. Brazzale and J. Cisewski-Kehe., (2016) "Searching for Extrasolar Planets". (poster)

Sagan Summer Workshop, Is There A Planet In My Data? Statistical Approaches to Finding and Characterizing Planets in Astronomical Data, Caltech, Pasadena, CA, July 2016.

References

Prof. Alessandra R. Brazzale
University of Padova,
Department of Statistical Sciences,
via Cesare Battisti, 241-243,
35121, Padova, Italy.
Phone: +39 049 827 4136
e-mail: brazzale@stat.unipd.it

Prof. Jessi Cisewski-Kehe
Yale University
Department of Statistics and Data Science
24 Hillhouse ave.,
06510, New Haven, CT, USA,
Phone: +1 203 432 5785
e-mail: jessica.cisewski@yale.edu